Decision trees: exercises, examples, experiments

Sándor Szabó

April 21, 2026

Contents

Chapter 1

Basic concepts

1.1 A bird's eye point of view

Even a cursory glance at the table of contents of a book on artificial intelligence will convince us that the artificial intelligence has a large number of branches or chapters. Figure 1.1 depicts a few of these chapters. Needless to say that there are many more topics than we mentioned. We do not make the slightest attempt to be exhaustive.

Step by step we will narrow our attention to the subject matter we are concerned with.

1.2 An introductory example

In order to introduce the basic terminology and to illustrate the main problem we will consider an example related to potential buyers.

As a first step we describe the involved variables. The variables are also called as attributes or features. The summary of these information about the variables are summarized in Table 1.1. The variable "BUY" is the target variable and the variables "AGE", "GENDER", "LASTR" are the explanatory variables. We would like to predict the value of the target variable

Table 1.1: The names, descriptions, ranges of the variables in the example.

variable name	variable description	variable range
AGE	age of the customer	1,2,3,...
GENDER	gender of the customer	male, female
LASTR	response to a direct mail	no, yes
BUY	buys the product	NO, YES

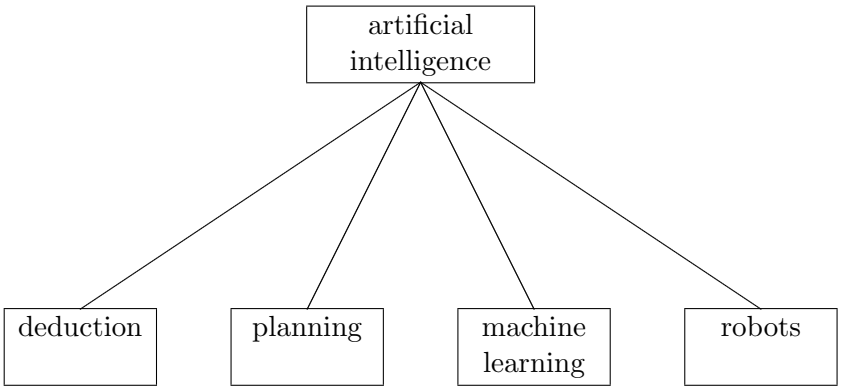


Figure 1.1: Various branches or chapters of artificial intelligence.

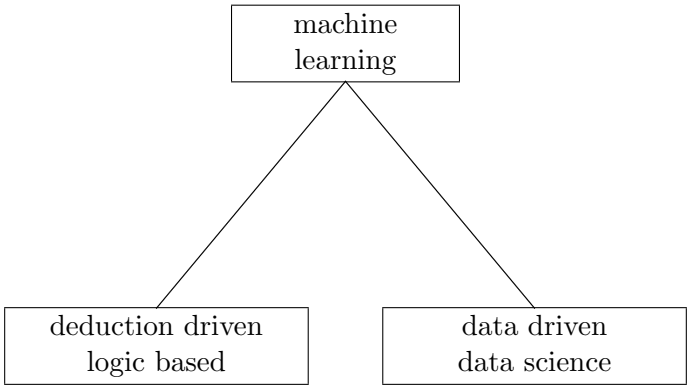


Figure 1.2: Knowledge versus data based learning dichotomy.

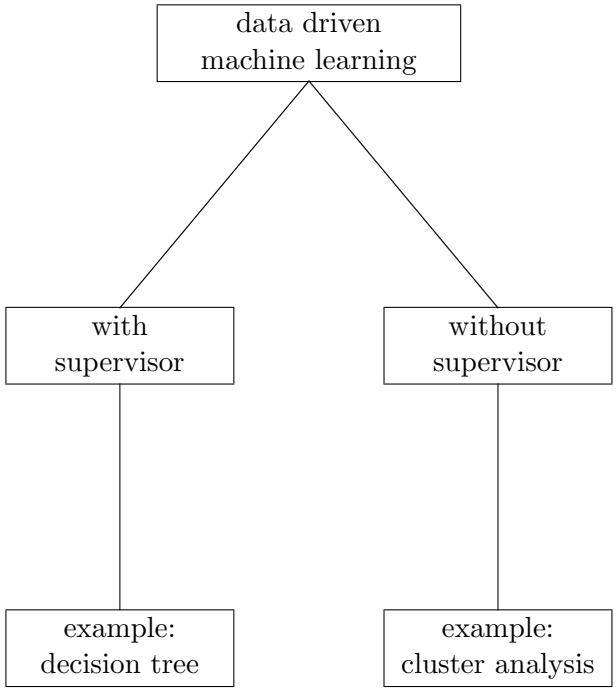


Figure 1.3: Supervised and unsupervised modes of learning.

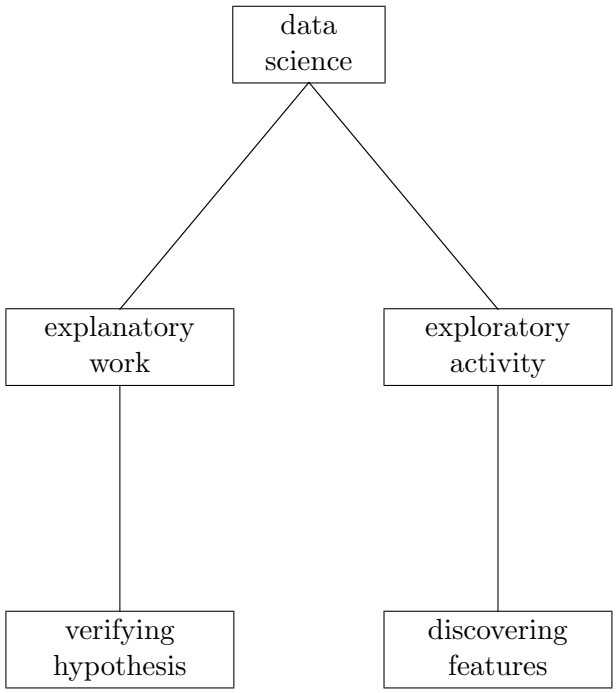


Figure 1.4: Explanatory and exploratory standpoints.

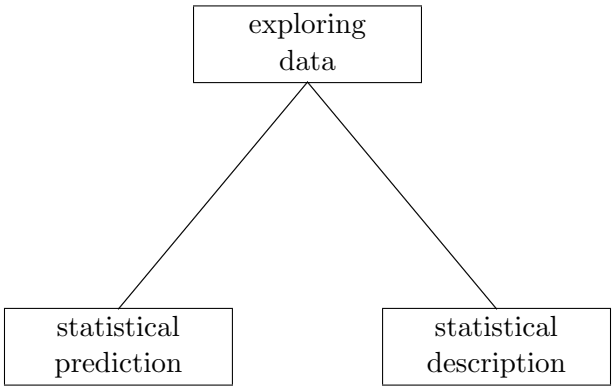


Figure 1.5: Predicting from the data or compiling numerical and visual summaries.

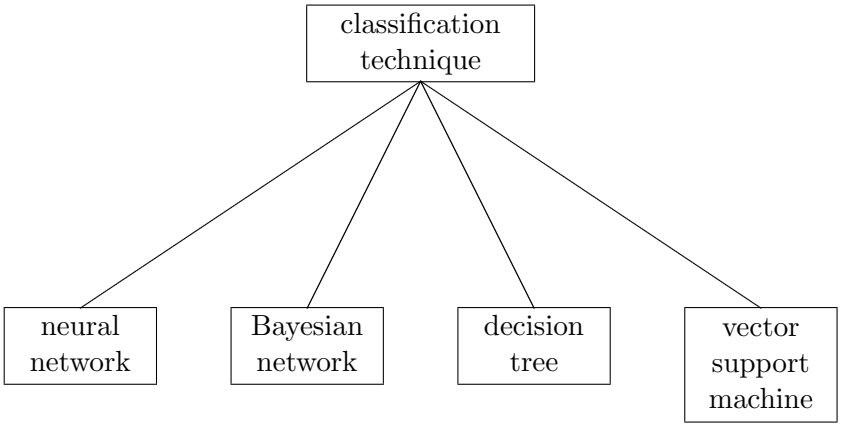


Figure 1.6: Some commonly used classification methods.

provided we know the values of the explanatory variables. The variables "AGE", "GENDER", "LASTR" are also called predictors and the variable "BUY" is termed as predicted variable. One may take the view that the variable "BUY" depends on the variables "AGE", "GENDER", "LASTR". In this case "AGE", "GENDER", "LASTR" are called independend variables and "BUY" is called the dependent variable.

The next figure depicts a small size decision tree related to our example. The tree has 7 nodes and 6 edges. The vertices are labeled and the edges are labeled. The tree is a so-called rooted tree. The node labeled by "AGE" is the root of the tree. The 4 nodes receiving the labels "NO" and "YES" are the leaf nodes. The edges are not only labeled but they are also directed edges. On the picture the edges are not presented as arrows. However, the direction can be identified. The edges are directed downward. From the "AGE" node there are only outward edges. This is why this node is the root node. The "NO" and "YES" nodes have only inward edges. This is why they are leaf nodes. Our tree is a binary tree in the sense that the number of the out going edges at each node is 0, 1, or 2. The values of the target variable are the labels of the leaf nodes. The remaining nodes are labeled by names of the explanatory variables. The possible values of the explanatory variables are the labels of the edges.

For the sake of conceptual clarity, we would like to point out that a decision tree may consists of only one node and it does not have edges at all. In this situation the node is a root node and leaf node of the tree in the same time. When the tree has more than one nodes, then the root node cannot be a leaf node of the tree and a leaf node cannot be a root node of the tree.

From our point of view it is an important property of a rooted tree that there is a unique path along which one can reach a given leaf node starting from the root node. This makes possible to utilize decision trees for classification. Figure 1.7 shows a small decision tree. It is designed in such a way that all the leaf nodes are on the lowest level. When we construct a tree in the middle of the work we most likely draw a very casual geometric representation of the tree. This is completely natural and it is completely all right. However, we should be aware of that a well designed tree comminicates the information better. In Figure 1.8 the leafs are not arranged to be on the lowest level. Figure 1.9 represents the same decision three. This time the edges are oriented. They are depicted as arrows. The nodes can be moved freely on the plane as one can follow the direction of the edges. Labeling the edges may result an overly cluttered picture. We can work with unlabeled edges. A node will be a record that contains the name of the variable together with the possible values of the variable. The arrow points out from a value of the variable and points to the an other node. The picture below illustrates this type of design.

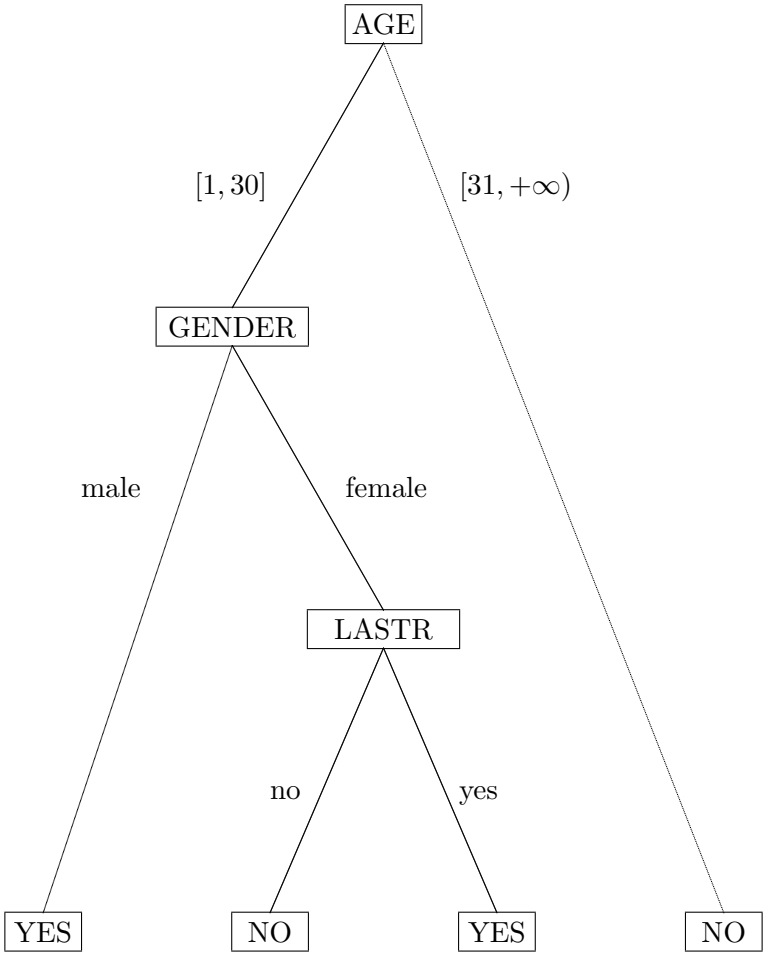


Figure 1.7: An example of a small decision tree.

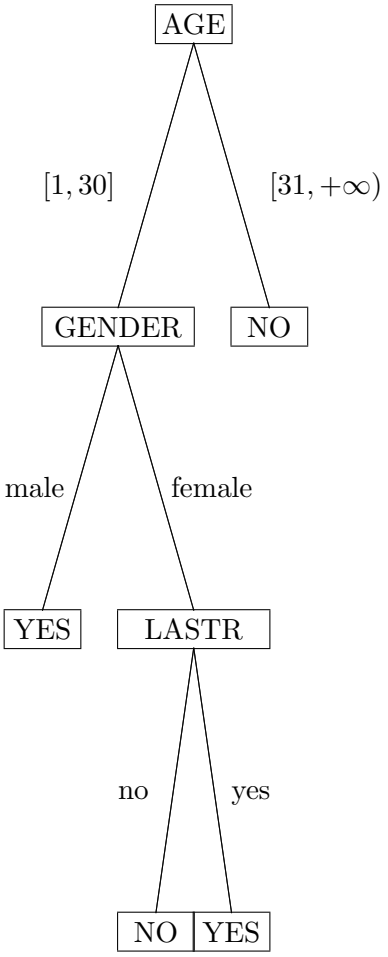


Figure 1.8: The leafs are not on the same level.

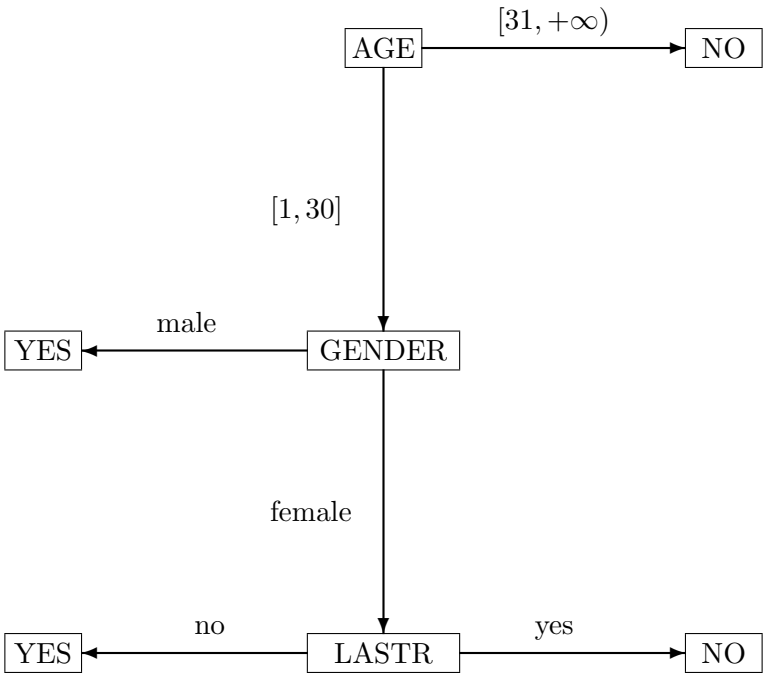


Figure 1.9: An other possible representation of the same decision tree.

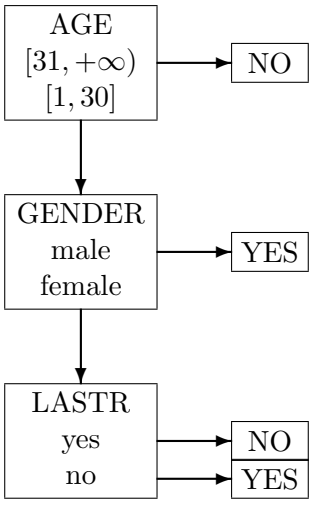


Figure 1.10: A further possible representation of the same decision tree.

Table 1.2: The predicted values of the variable "BUY" in the example.

customer number	AGE	GENDER	LASTR	BUY actual	BUY predicted
1	35	male	yes	—	no
2	26	female	no	—	no
3	22	male	yes	—	yes
4	63	male	no	—	no
5	47	female	no	—	no
6	54	male	no	—	no
7	27	female	yes	—	yes
8	38	female	no	—	no
9	42	female	yes	—	no
10	19	male	no	—	yes

1.3 Using the tree for prediction

In Table 1.2 below the values of the variables "AGE", "GENDER", and "LASTR" belonging to 10 customers are listed. The actual values of the variable "BUY" are not available.

The age of customer 1 is 35 years. At the "AGE" root node of the decision tree given in the figure we choose to move down on the right hand side edge reaching the leaf node "NO". Consequently, we predict that customer 1 is not going to buy the product.

The age of customer 2 is 26 years. At the "AGE" root node of the decision tree we choose to move down on the left hand side edge reaching the node "GENDER". At the "GENDER" node of the decision tree we choose to move down on the right hand side edge reaching the node "LASTR". At the "LASTR" node of the decision tree we choose to move down on the left hand side edge reaching the leaf node "NO". Therefore, we predict that customer 2 is not going to buy the product.

Continuing in this way finally we produce a prediction for all the 10 costumers. The results are listed in the last column of the table.

1.4 Evaluating the quality of prediction

The performance of the decision tree should be evaluated. In other word we should asses how the actual and predicted values of the "BUY" variables compare.

Table 1.3 is almost identical with Table 1.2. The only difference between the tables is that this time beside the predicted values of the variable "BUY" the actual values of the variable "BUY" are also included. The main tool to

Table 1.3: The actual and predicted values of the variable "BUY" in the example.

customer number	AGE	GENDER	LASTR	BUY actual	BUY predicted
1	35	male	yes	no	no
2	26	female	no	no	no
3	22	male	yes	yes	yes
4	63	male	no	yes	no
5	47	female	no	no	no
6	54	male	no	no	no
7	27	female	yes	yes	yes
8	38	female	no	yes	no
9	42	female	yes	yes	no
10	19	male	no	no	yes

Table 1.4: The confusion matrix.

	predicted no	predicted yes
actual no	a	b
actual yes	c	d

assess the performance of the decision tree is the so-called confusion matrix. (A more complete name of this confusion matrix is the confusion matrix for binary classification.) The confusion matrix is a two by two matrix consisting of two rows, two columns and and four cells. (For non-binary classification the confusion matrix has more than two rows and columns.) The rows are labeled as "actual no" and "actual yes". The columns are labeled as "predicted no" and "predicted yes". The cell whose labels are "actual no" and "predicted no" contains the number of customers with "BUY actual" and "BUY predicted" attributes both take on the "no" value. In Table 1.4 this number is denoted by "a". The remaining cells are filled in a similar manner. We have to point out the entries a, b, c, d in the confusion matrix are counts. In other words they are non-negative integers.

A number of quantities defined to measure different aspects of the prediction. We list these numerical indicators together with their names in Table 1.5. Let us construct the confusion matrix and let us compute the various quality indicators in connection with our example. First we filled in a contingency table whose rows and columns are labeled exactly like the confusion matrix. The costumers can be sorted into four categories accord-

Table 1.5: The numerical indicators of performance.

terminology	formula
accuracy	$(a+d)/(a+b+c+d)$
misclassification rate	$(b+c)/(a+b+c+d)$
precision	$d/(b+d)$
true positive rate (recall)	$d/(c+d)$
false positive rate	$b/(a+b)$
true negative rate (specificity)	$a/(a+b)$
false negative rate	$c/(c+d)$

Table 1.6: The confusion matrix whose cells are filled with records of the traing data set.

	predicted no	predicted yes
actual no	1,2,5,6	10
actual yes	4,8,9	3,7

ing to the actual and predicted values of the "BUY" variable. We placed the customer numbers into the appropriate cells. Table **??** is the result of this work. An inspection gives the total numbers of customers in each cell. This provides the confusion matrix. Plugging the contents of the cells into the formulas we get the quality indicators. Table 1.7 contains the details. (The plural of the word formula is formulae as it is a Latin origin word. However, formulas can also be used.) In the remaining part we make some comments on the confusion matrix. (If the comments are not helpful please ignore them.) We may refer to the four cells of the confusion matrix by names. We wrote the names into the cells. The possible outcomes when one carries out a statistical hypotheses H_0 are arranged in an analogous manner. The rows record material facts. While the columns corresponds to actions taken.

Table 1.7: The confusion matrix whose cells are filled with counts.

	predicted no	predicted yes
actual no	4	1
actual yes	3	2

Table 1.8: Numerical illustrations of performance measures.

accuracy	(4+2)/(4+1+3+2)	0.60
misclassification rate	(1+3)/(4+1+3+2)	0.40
precision	2/(1+2)	0.67
true positive rate (recall)	2/(3+2)	0.40
false positive rate	1/(4+1)	0.20
true negative rate (specificity)	4/(4+1)	0.80
false negative rate	3/(3+2)	0.60

Table 1.9: An aid to memorize the names of the performance indicators.

	predicted no	predicted yes
actual no	true negative	false positive
actual yes	false negative	true positive

Table 1.10: A tabular summary of statistical hypothesis testing.

	H_0 rejected no	H_0 rejected yes
H_0 holds no	correct decision	type I error
H_0 holds yes	type II error	correct decision

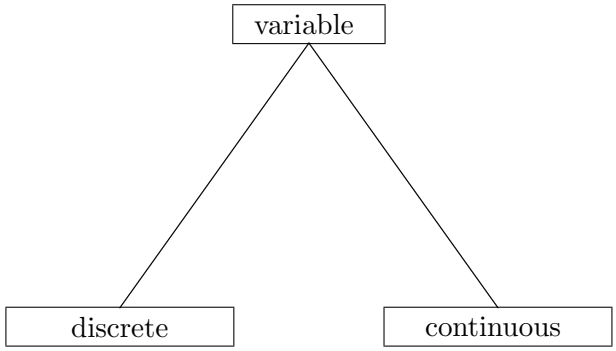


Figure 1.11: The discrete and continuous dihotomy of variables.

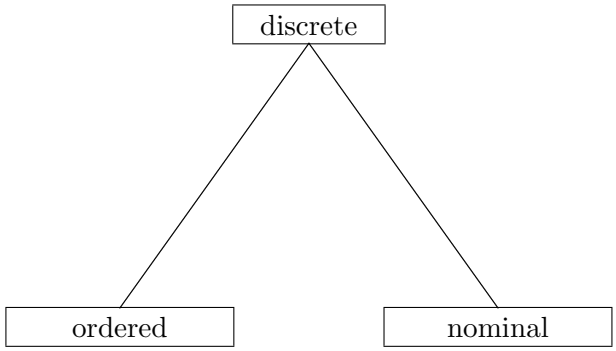


Figure 1.12: The ordered and nominal dihotomy of variables.

1.5 Non-binary classification

In elementary mathematical statistical courses it is customary to distinguish various types of data. We talk about discrete and continuous variables. The discrete variables can be ordered and nominal and also the discrete variables can be binary and non-binary. The binary variables are referred as boolean variables as well. When the target variable is not binary, then the confusion matrix is not a two by two matrix. For example when the target variable takes the on values “dog”, “cat”, “rabbit”, then we we face with a three-fold classification problem. This time the values “dog”, “cat”, “rabbit” not ordered and so the target variable is of nominal type. If the target variable values were “low”, “medium”, “high”, then it would be an ordered discrete variable. In either case the confusion matrix is a 3 by 3 matrix. The rows are labeled by “actual dog”, “actual cat”, “actual rabbit”. The columns are labeled by “predicted dog”, “predicted cat”, “predicted rabbit”. The cells are containing non-negative integers. They are counts. Just to make this point clear we repeat, that the contents of the cells are not negative

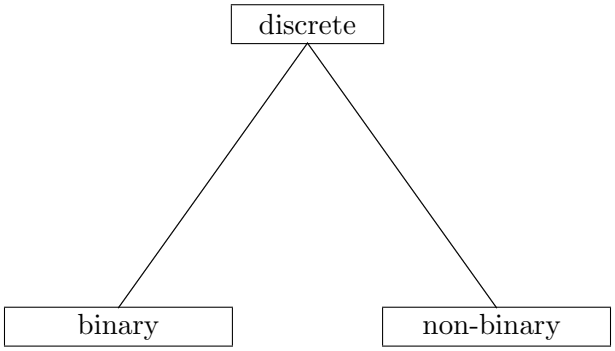


Figure 1.13: The binary and non-binary dihotomy of variables.

Table 1.11: A non-binary confusion matrix.

		predicted dog	predicted cat	predicted rabbit
actual	dog	a	b	c
actual	cat	x	y	z
actual	cat	u	v	w

numbers, not fractions, not percentages. The only performance indicator we will use is the accuracy which is defined to be equal to

$$(a + y + w)/(a + b + c + x + y + z + u + v + w).$$

In plain English, the sum of the main diagonal is divided by the total sum.

Chapter 2

Classification using accuracy

2.1 The training data set

Table 2.1 summarizes the variables their names and ranges. Table 2.2 contains the training data set consisting of data collected on 14 days. The "play tennis" variable is the target variable. The variables "outlook", "temperature", "humidity", "wind" are the explanatory variables. We want to predict the value of "play tennis" using the explanatory variables.

2.2 Assessing the variables one at a time

We summarize the predictive powers of the variable in Table 2.3. The accuracy of predicting the target variable using a given explanatory variable can be interpreted as a predictive power of this given explanatory variable. These are the values we intend to compute in this section. Let us consider the situatiation when none of the explanatory variable is used to predict the values of the "play tennis" target variable. The prediction rule is ( )→yes. This means that we predict yes in all time. The confusion matrix can be filled in without any difficulty.

Table 2.1: Names, descriptions, ranges of variables.

variable name	variable description	variable range
outlook	weather outlook	rain, overcast, sunny
temperature	air temperature	cool, mild, hot
humidity	air humidity	normal, high
wind	wind strength	weak, strong
play tennis	playing tennis	no, yes

Table 2.2: The 14 records of the training data set.

day	outlook	temperature	humidity	wind	play tennis
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rain	mild	high	weak	yes
5	rain	cool	normal	weak	yes
6	rain	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rain	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rain	mild	high	strong	no

Table 2.3: The summary of the predictive power of the variables.

variable attribute feature	accuracy predictive power	false positive rate	true positive rate
none	0.643	1.000	1.000
outlook	0.714	0.400	0.777
temperature	0.643	0.600	0.777
humidity	0.714	0.200	0.666
wind	0.643	0.400	0.666

Table 2.4: The filled in confusion matrix.

	predicted no	predicted yes
actual no	0	5
actual yes	0	9

Table 2.5: A contingency table filled in with the numbers of the days.

	outlook rain	outlook overcast	outlook sunny
play tennis no	6,14		1,2,8
play tennis yes	4,5,10	3,7,12,13	9,11

Table 2.6: A contingency table filled in the counts of the days.

	outlook rain	outlook overcast	outlook sunny
play tennis no	2	0	3
play tennis yes	3	4	2

The accuracy is equal to (0+9)/14=0.643.

The false positive rate is equal to (5)/(0+5)=1.000.

The true positive rate is equal to (9)/(0+9)=1.000.

If the predictive power of an explanatory variable is not exceeding this value, then we can safely ignore this explanatory variable. In other words, prediction without any explanatory variable establishes an absolute minimum base line. The improvement over this base line is what we are looking for when we assess the worth of an explanatory variable. The decision tree associated with this prediction consists of a single node whose label is "yes".

At first we use the explanatory variable "outlook" for the prediction of the "play tennis". We compute contingency tables. In the first contingency table the cells contains days identified by their number. In the second contingency table the cells contains the counts of the days.

The second contingency table suggests the following decision rule.

rain→yes, overcast→yes, sunny→no.

The accuracy is equal to $(3 + 7)/14 = 0.714$.

Table 2.7: A contingency table in which the column labels contains the predictions.

	outlook rain yes	outlook overcast yes	outlook sunny no
play tennis no	2	0	3
play tennis yes	3	4	2

Table 2.8: A contingency table in which the row and column labels are actual and predicted values.

	predicted yes	predicted yes	predicted no
actual no	2	0	3
actual yes	3	4	2

Table 2.9: The confusion matrix.

	predicted no	predicted yes
actual no	3	2
actual yes	2	7

The false positive rate is equal to $(2)/(3+2) = 0.400$.

The true positive rate is equal to $(7)/(2+7) = 0.777$.

The alert reader may notice that we can get the last contingency table (the confusion matrix) from the first contingency table without writing down the intermediate contingency tables.

Here is the suggested short cut. After recording the days in the cells of the first contingency table we count the number of the entries in cell and write the totals in the cells. We put parentheses around the totals in order not to confuse them with the earlier entries.

Using these totals we can identify the decision rules. The decision rule assigns a "yes" or a "no" to the possible values of the explanatory variable. In the second row of the first contingency table we can enter these "yes" or "no" values right after the value of the variable. For the sake of clarity we put parentheses around the "yes" and "no" values.

From this extended contingency table one can construct the confusion matrix in a straight-forward manner. In the future we will use this short cut systematically.

Figure 2.1 is a possible geometric representation of the decision tree based solely on the "outlook" variable. We denote this tree by $\mathrm{T}_{\mathrm{out}}$. Next we consider the case when the "temperature" explanatory variable is used

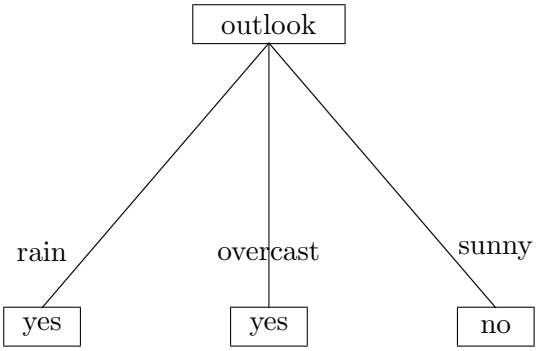


Figure 2.1: The decision tree T_{out} based solely on the "outlook" variable.

to predict the the "play tennis" target variable.

	temperature cool	(yes)	temperature mild	(yes)	temperature hot	(no)
play tennis no	6	(1)	8,14	(2)	1,2	(2)
play tennis yes	5,7,9	(3)	4,10,11,12	(4)	3,13	(2)

	predicted no	predicted yes
actual no	2	3
actual yes	2	7

The accuracy is equal to (2+7)/14=0.643.

The false positive rate is equal to (3)/(2+3)=0.600.

The true positive rate is equal to (7)/(2+7)=0.777.

Figure 2.2 is a possible decision tree based solely on the "temperature" variable. We call this tree T_{temp}.

We turn to predict the "play tennis" variable using the "humidity" variable.

	humidity normal	(yes)	humidity high	(no)
play tennis no	6	(1)	1,2,8,14	(4)
play tennis yes	5,7,9,10,11,13	(6)	3,4,12	(3)

	predicted no	predicted yes
actual no	4	1
actual yes	3	6

The accuracy is equal to (4+6)/14=0.714.

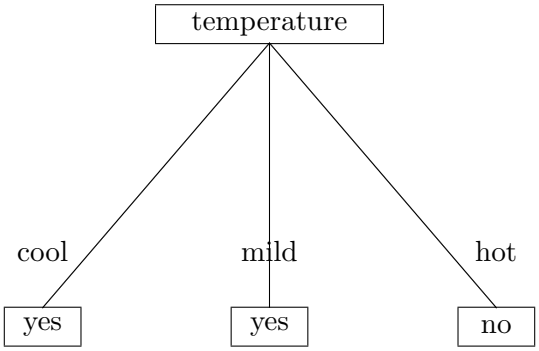


Figure 2.2: The decision tree T_{temp} based solely on the "temperature" variable.

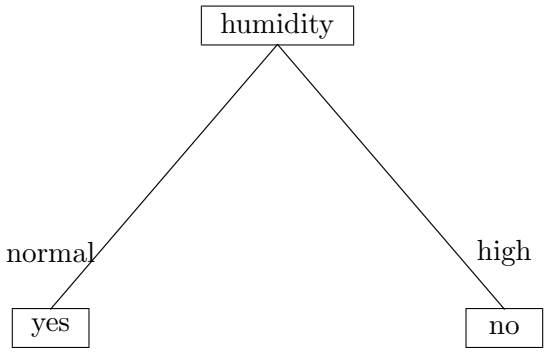


Figure 2.3: The decision tree T_{hum} based solely on the "humidity" variable.

The false positive rate is equal to (1)/(4+1)=0.200.

The true positive rate is equal to (6)/(3+6)=0.666.

Figure 2.3 is a possible geometric representation of the decision tree based solely on the "humidity" variable. It will be denoted by T_{hum}.

Finally, the "wind" variable is used to predict the "play tennis" variable.

	wind strong		wind weak	
play tennis no	2,6,14	(3)	1,8	(2)
play tennis yes	7,11,12	(3)	3,4,5,9,10,13	(6)

	predicted no	predicted yes
actual no	3	2
actual yes	3	6

The accuracy is equal to (3+6)/14=0.714.

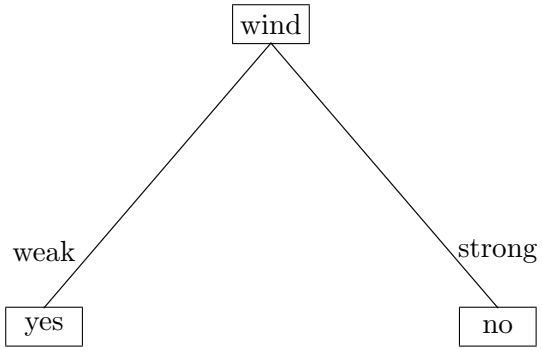


Figure 2.4: The decision tree based T_{wind} on the "wind" variable alone.

The false positive rate is equal to (2)/(3+2)=0.400.

The true positive rate is equal to (6)/(3+6)=0.666.

Figure 2.4 is the decision tree based on the "wind" variable alone. We will refer to the tree as T_{wind}.

2.3 ROC curves

The concept of the receiver operating characteristic (or ROC) is borrowed from mathematical statistics. The main ingredients are a given decision tree and a coordinate system in which the first axis is labeled by the false positive value and the second axis is labeled by the true positive rate. To each decision tree we assign a point. For instance to the decision tree associated with the variable "outlook" and we assign the point $(0.400, 0.777)$. When one is dealing with a large number of decision trees one gets a large number of points. All these points are in a unit square. The coordinates of the vertices of the unit square are $(0, 0)$, $(1, 0)$, $(1, 1)$, $(0, 1)$.

The location of a point in the unit square helps to better understand the related decision tree. For instance if a point is on the diagonal of the unit square connecting the vertices $(0, 0)$ and $(1, 1)$, then the corresponding decision tree is not performing well. The closer is the point to the vertex $(0, 1)$, the better the decision tree is performing.

On the set of the points of the unit square we introduce a dominance relation. Let (a_1, b_1), (a_2, b_2) be points of the unit square. We say, that (a_1, b_1) dominates (a_2, b_2) if $a_1 \leq a_2$ and $b_1 \geq b_2$ hold. Obviously, if the false positive rate is getting smaller and in the same time the true positive rate is getting higher, then the performance of the decision tree is improving. In our example the tree associated with the variable "outlook" dominates the trees associated with the variables "wind" and "temperature". Similarly, the tree associated with variable "humidity" dominates the tree associated with the variable "wind".

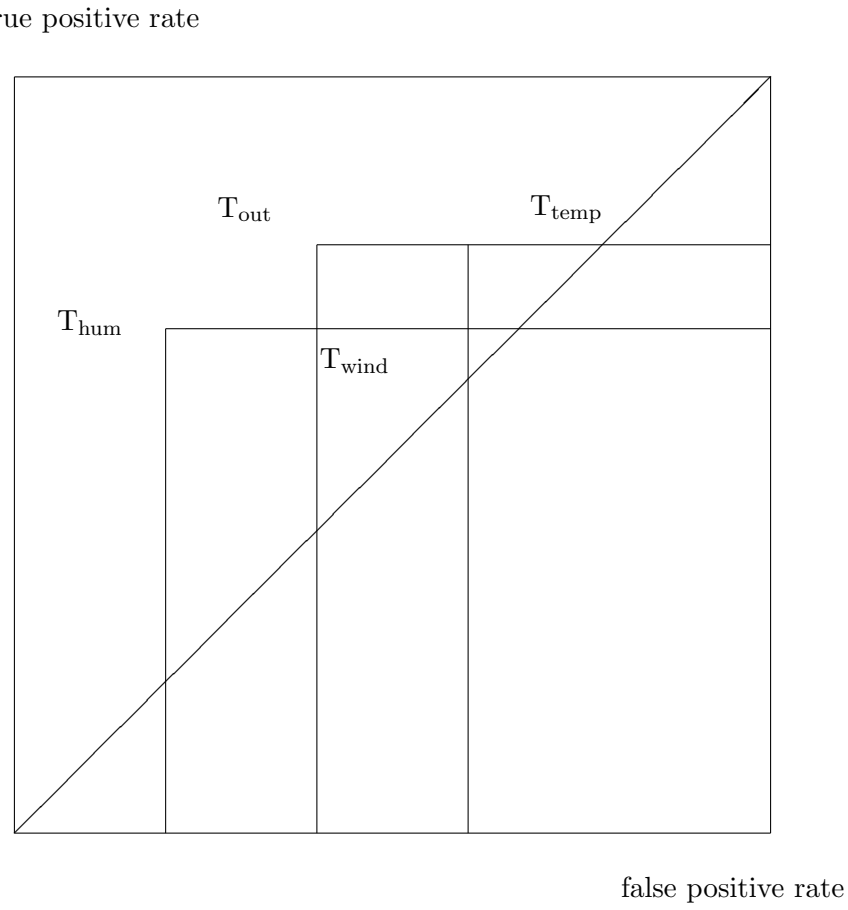


Figure 2.5: The rectangles corresponding to the decision trees.

In connection with the point (a_1, b_1) we can draw a rectangle whose left upper corner is (a_1, b_1) and right lower corner is $(1, 0)$. Similarly, in connection with the point (a_2, b_2) we can draw a rectangle whose left upper corner is (a_2, b_2) and right lower corner is $(1, 0)$. Now, (a_1, b_1) dominates (a_2, b_2) if the rectangle of (a_1, b_1) contains the rectangle of (a_2, b_2).

The set of the points of the unit square forms a partially ordered set with the dominance relation. The maximal elements of this partially ordered sets form the so-called Pareto front. In our example, the points associated with the variables "outlook" and "humidity" form the Pareto front. (Vilfredo Pareto Italian sociologist)

Finite partially ordered sets can be visualized conveniently by drawing Hasse diagrams. In our example the number of the decision trees and consequently the number of the elements of the partially ordered set is equal to 4. We constructed the Hasse diagram of this 4 element partially ordered

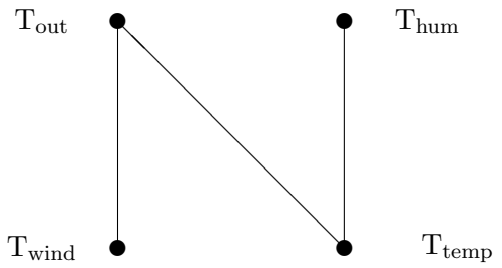


Figure 2.6: The Hasse diagram of the partially ordered set in the worked out example.

set.

To the points associated with decision trees it is customary to assign a convex domain. Namely, we augment the point with the three corner points $(0,0)$, $(1,0)$, $(1,1)$ of the unit square and we form the convex envelope if these points. In the case when the number of the points is finite, the convex domain we get is a convex polygon. (Some of the points may be inside the polygon.) The area of this polygon can be viewed as an area under a curve which is part of the boundary of the polygon. This area is a number which is between 0 and 1. The closer is this number to 1 the more information the decision trees involved carries about the target variable. In machine learning this number is referred as area under the curve (or AUC) measure.

2.4 The ensemble method

We have four decision trees T_{out}, T_{temp}, T_{hum}, T_{wind}. Two of these T_{out}, T_{hum} perform better than the base line prediction. One may suspect that these two prediction trees do better together than each of them alone. The idea is forming a committee consisting of the two decision trees T_{out} and T_{hum} and pool their predictions together. When both vote "yes", then the combined vote is "yes". When both vote "no", then the combined vote is "no". When one votes "yes" and the other votes "no" we face to a tie. In order to avoid ties we form committees consisting of an odd number of members. Thus we add the decision three T_{temp} to the committee.

At this stage it is not clear that the committee performs better than the trees T_{out}, T_{hum} alone. If the committee does better, then we happily employ it. If the committee does not do better, then we simply drop it.

The training data set contains 14 days. We go through the 14 days and compare the predicted and the actual results. This gives an assessment of

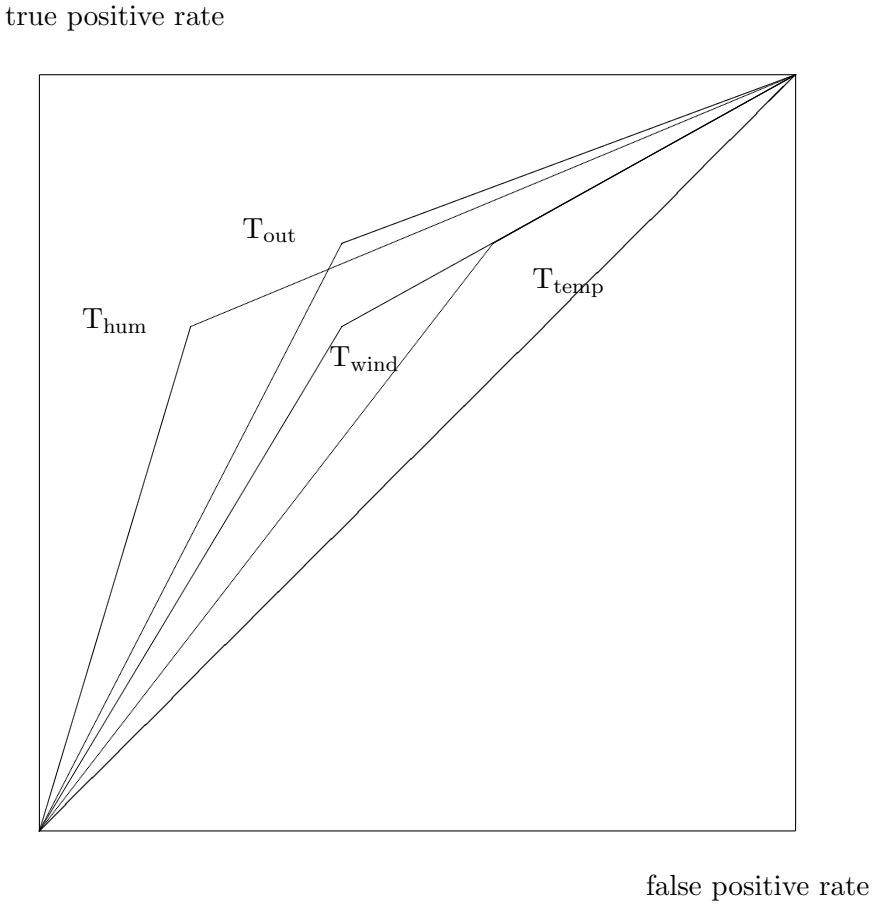


Figure 2.7: The ROC curves of the decision trees associated with the variables “outlook”, “temperature”,“humidity”, “wind”.

the performance of the committee.

day	vote T_{out}	vote T_{temp}	vote T_{hum}	vote committee	play tennis
1	no	no	no	NO	no
2	no	no	no	NO	no
3	yes	no	no	NO	yes
4	yes	yes	no	YES	yes
5	yes	yes	yes	YES	yes
6	no	yes	yes	YES	no
7	yes	yes	yes	YES	yes
8	no	yes	no	NO	no
9	no	yes	yes	YES	yes
10	yes	yes	yes	YES	yes
11	no	yes	yes	YES	yes
12	yes	yes	no	YES	yes
13	yes	no	yes	YES	yes
14	yes	yes	no	YES	no

		predicted NO		predicted YES	
actual	no	1,2,8	(3)	6,14	(2)
actual	yes	3	(1)	4,5,7,9,10,11,12,13	(8)

The accuracy is equal to (3+8)/14=0.786.

The false positive rate is equal to (2)/(3+2)=0.400.

The true positive rate is equal to (8)/(1+8)=0.889.

As it happened the committee over performs both of the trees T_{out}, T_{hum}. It is good to keep in mind that a simple trick of forming committees from existing prediction trees can lead to better prediction. We may try to squeeze more juice out of the orange by giving more weight to trees whose predictive power is larger. At the pooling of the votes we can take a weighted average of the votes.

2.5 Splitting the training data set

The predictive power of the variable "outlook" is equal to 0.714. This is one of the largest among the predictive powers of the variables "outlook", "temperature", "humidity", "wind". The other variable with largest predictive power is the variable "humidity". In such case when there are more than one variable with the same largest predictive power we choose freely any of these variables. This variable will be used to add new branches to the decision tree.

In our example the variable "outlook" is chosen. This variable can take three values "rain", "overcast", "sunny". We split the training data set into three smaller data sets.

day	outlook	temperature	humidity	wind	play tennis
4	rain	mild	high	weak	yes
5	rain	cool	normal	weak	yes
6	rain	cool	normal	strong	no
10	rain	mild	normal	weak	yes
14	rain	mild	high	strong	no
3	overcast	hot	high	weak	yes
7	overcast	cool	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
11	sunny	mild	normal	strong	yes

2.6 The "outlook" variable is equal to "rain"

We are working with the following small training data set.

day	outlook	temperature	humidity	wind	play tennis
6	rain	cool	normal	strong	no
14	rain	mild	high	strong	no
4	rain	mild	high	weak	yes
5	rain	cool	normal	weak	yes
10	rain	mild	normal	weak	yes

We can express our plan in terms of trees. Figure 2.8 shows three small trees and we assess the predictive power of these threes. The trees happen to be paths. We intend the compute the predictive power of these paths of length one. The fact that the length of a path is one is not essential we can extend a longer or a shorter path in a similar way. When we computed the predictive powers of the variables "outlook", "temperature", "humidity", "wind" we have computed the predictive power of paths of length zero. Each of the four paths consisted of a single node.

We summarize the predictive powers of the variables in the following

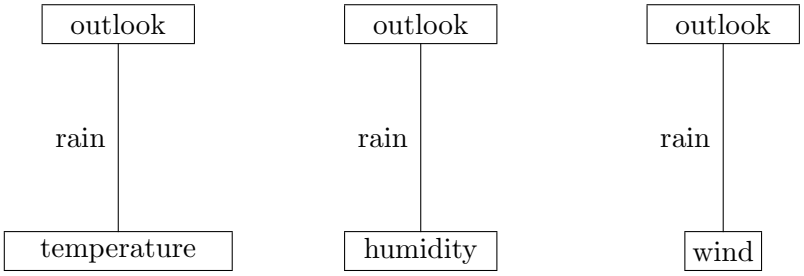


Figure 2.8: The paths whose predictive power we intend to compute.

table. These are the values we intended to compute in this section.

variable attribute feature	accuracy predictive power	false positive rate	true positive rate
none	0.600	1.000	1.000
temperature	0.600	0.500	0.667
humidity	0.600	0.500	0.667
wind	1.000	0.000	1.000

First, we consider the case when the "temperature" explanatory variable is used to predict the the "play tennis" target variable. Using the language of decision trees we compute the predictive power of the tree shown in Figure 2.9.

	temperature cool	(no)	temperature mild	(yes)	temperature hot	(no)
play tennis no	6	(1)	14	(1)		(0)
play tennis yes	5	(1)	4,10	(2)		(0)

	predicted no	predicted yes
actual no	1	1
actual yes	1	2

The accuracy is equal to (1+2)/5=0.600.

The false positive rate is equal to (1)/(1+1)=0.500.

The true positive rate is equal to (2)/(1+2)=0.667.

We turn to predict the "play tennis" variable using the "humidity" variable. This time we compute the predictive power of the decision tree shown in Figure 2.10.

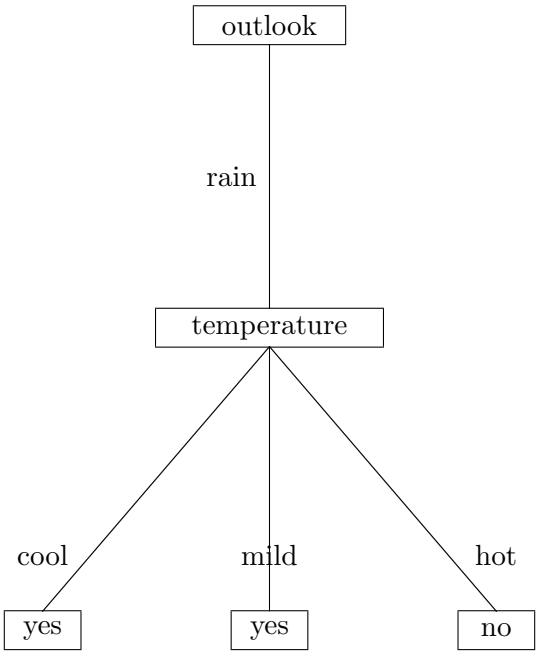


Figure 2.9: We will compute the predictive power of this decision tree.

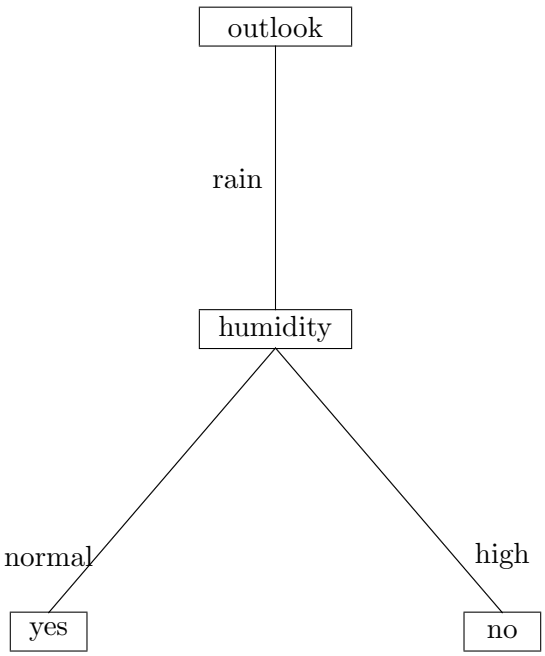


Figure 2.10: The decision whose predictive power we compute.

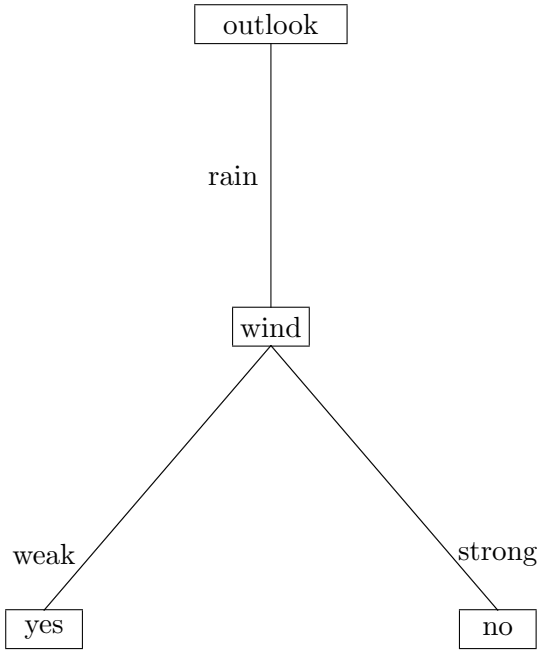


Figure 2.11: The decision tree we use to compute the predictive power of the "wind" variable assume that "outlook" is equal to ´´rain".

	humidity normal	(yes)	humidity high	(no)
play tennis no	6	(1)	14	(1)
play tennis yes	5,10	(2)	4	(1)

	predicted no	predicted yes
actual no	1	1
actual yes	1	2

The accuracy is equal to (1+2)/5=0.600.

The false positive rate is equal to (1)/(1+1)=0.500.

The true positive rate is equal to (2)/(1+2)=0.667.

Finally, the "wind" variable is used to predict the "play tennis" variable. Figure 2.11 exhibits the decision tree whose predictive power we are interested in.

	wind strong	(no)	wind weak	(yes)
play tennis no	6,14	(2)		(0)
play tennis yes		(0)	4,5,10	(3)

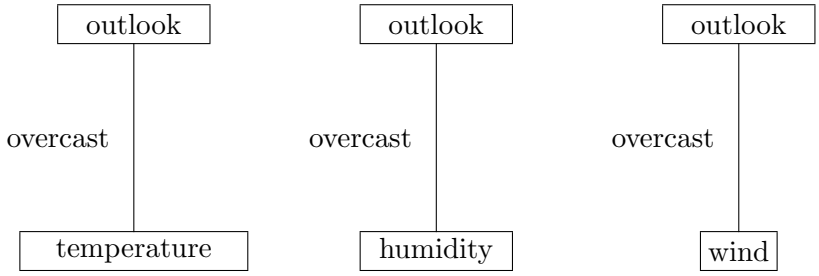


Figure 2.12: The paths whose predictive power we intend to compute.

	predicted no	predicted yes
actual no	2	0
actual yes	0	3

The accuracy is equal to (2+3)/5=1.000.
The false positive rate is equal to (0)/(2+0)=0.000.
The true positive rate is equal to (3)/(0+3)=1.000.

2.7 The "outlook" variable is equal to "overcast"

The small training data set we will be busy with is the following.

day	outlook	temperature	humidity	wind	play tennis
3	overcast	hot	high	weak	yes
7	overcast	cool	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes

Figure 2.12 depicts the paths we would like to work with.

We summarize the predictive powers of the variables in the following table. These are the values we will compute in this section.

variable attribute feature	accuracy predictive power	false positive rate	true positive rate
none	1.000	—	1.000
temperature	1.000	—	1.000
humidity	1.000	—	1.000
wind	1.000	—	1.000

After filling in the row with label "none" we must notice that none of the explanatory variables "temperature", "humidity", "wind" can have larger

predictive power. Therefore filling in the following rows of the table is completely superfluous. However, for the sake of practice we fill in the remaining rows too.

	temperature cool	(yes)	temperature mild	(yes)	temperature hot	(yes)
play tennis no		(0)		(0)		(0)
play tennis yes	7	(1)	12	(1)	3,13	(2)

	predicted no	predicted yes
actual no	0	0
actual yes	0	4

The accuracy is equal to (0+4)/4=1.000.

The false positive rate is equal to (0)/(0+0) which is not defined.

The true positive rate is equal to (4)/(0+4)=1.000.

We turn to predict the "play tennis" variable using the "humidity" variable.

	humidity normal	(yes)	humidity high	(yes)
play tennis no		(0)		(0)
play tennis yes	7,13	(2)	3,12	(2)

	predicted no	predicted yes
actual no	0	0
actual yes	0	4

The accuracy is equal to (0+4)/4=1.000.

The false positive rate is equal to (0)/(0+0) which is not defined.

The true positive rate is equal to (4)/(0+4)=1.000.

Finally, the "wind" variable is used to predict the "play tennis" variable.

	wind strong	(yes)	wind weak	(yes)
play tennis no		(0)		(0)
play tennis yes	7,12	(2)	3,13	(2)

	predicted no	predicted yes
actual no	0	0
actual yes	0	4

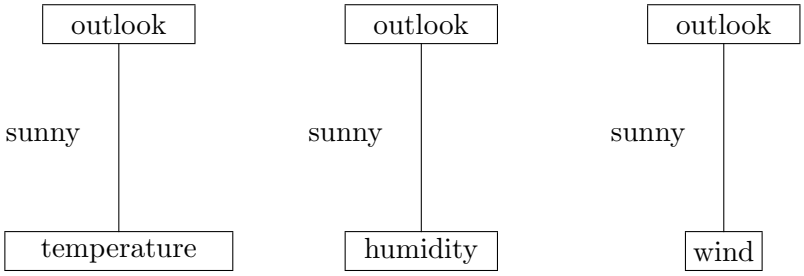


Figure 2.13: The paths whose predictive power we intend to compute.

The accuracy is equal to (0+4)/4=1.000.
The false positive rate is equal to (0)/(0+0) which is not defined.
The true positive rate is equal to (4)/(0+4)=1.000.

2.8 The "outlook" variable is equal to "sunny"

The small training date set we use is contained in the next table.

day	outlook	temperature	humidity	wind	play tennis
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
11	sunny	mild	normal	strong	yes

Figure J.1 depicts the paths we will work with.

We summarize the predictive powers of the variables in the following table. These are the values we plan to compute in this section.

variable attribute feature	accuracy predictive power	false positive rate	true positive rate
none	0.600	0.000	0.000
temperature	0.800	0.333	1.000
humidity	1.000	0.000	1.000
wind	0.600	0.333	0.500

After computing the predictive power of the explanatory variable "humidity" we do not need to fill in the next row of the table. Simply we cannot get a larger predictive power. Once again for the sake of practice we fill in the

next row.

	tempe-rature cool	(yes)	tempe rature mild	(yes)	tempe-rature hot	(no)
play tennis no		(0)	8	(1)	1,2	(2)
play tennis yes	9	(1)	11	(1)		(0)

	predicted no	predicted yes
actual no	2	1
actual yes	0	2

The accuracy is equal to (2+2)/5=0.800.
The false positive rate is equal to (1)/(2+1)=0.333
The true positive rate is equal to (2)/(0+2)=1.000.
We turn to predict the "play tennis" variable using the "humidity" variable.

	humidity normal	(yes)	humidity high	(no)
play tennis no		(0)	1,2,8	(3)
play tennis yes	9,11	(2)		(0)

	predicted no	predicted yes
actual no	3	0
actual yes	0	2

The accuracy is equal to (3+2)/5=1.000.
The false positive rate is equal to (0)/(3+0)=0.000.
The true positive rate is equal to (2)/(0+2)=1.000.
Finally, the "wind" variable is used to predict the "play tennis" variable.

	wind strong	(yes)	wind weak	(no)
play tennis no	2	(1)	1,8	(2)
play tennis yes	11	(1)	9	(1)

	predicted no	predicted yes
actual no	2	1
actual yes	1	1

The accuracy is equal to (2+1)/5=0.600.
The false positive rate is equal to (1)/(2+1)=0.333.
The true positive rate is equal to (1)/(1+1)=0.500.

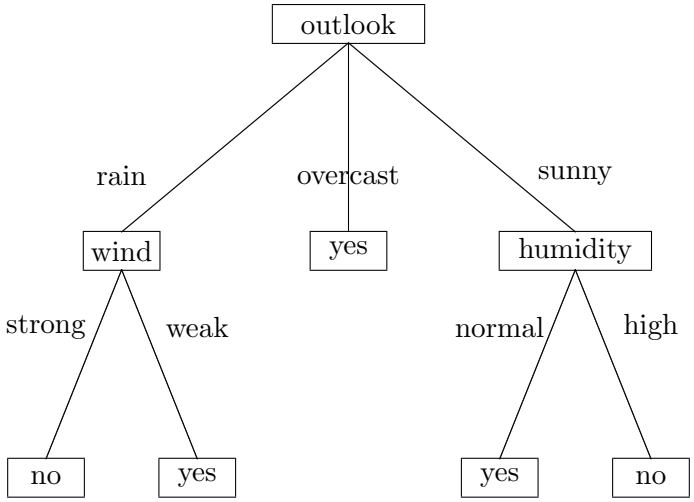


Figure 2.14: The decision tree based on the "outlook", "wind", "humidity" variables.

2.9 An extended decision tree

In the small data set when the "outlook" variable takes on the "rain" value, then the best predictor for the "play tennis" is the "wind" attribute.

In the small data set when the "outlook" variable takes on the "overcast" value, then there is no need for any further predictor (explanatory) variable.

In the small data set when the "outlook" variable takes on the "sunny" value, then the best predictor for the "play tennis" is the "humidity" feature.

Utilizing these facts we construct a new extended decision tree. The decision tree based on the "outlook", "wind", "humidity" variables. Figure 2.14 is a possible geometric representation of the tree.

Naturally, we are curious how this new tree performs the classification

Table 2.10: The decision tree in Figure 2.14 in table from.

node label	edge label	node label	edge label	node label
outlook	rain	wind	strong	no
			weak	yes
	overcast	yes		
	sunny	humidity	normal	yes
			high	no
level 1		level 2		level 3

Table 2.11: The tree in Figure 2.14 in table form.

level 1	outlook					node
	rain		overcast	sunny		edge
level 2	wind		yes	humidity		node
	strong	weak		normal	high	edge
level 3	no	yes		yes	no	node

and so we carry out a performance assessment.

day	outlook	temperature	humidity	wind	play tennis actual	play tennis predicted
1	sunny	hot	high	weak	no	no
2	sunny	hot	high	strong	no	no
3	overcast	hot	high	weak	yes	yes
4	rain	mild	high	weak	yes	yes
5	rain	cool	normal	weak	yes	yes
6	rain	cool	normal	strong	no	no
7	overcast	cool	normal	strong	yes	yes
8	sunny	mild	high	weak	no	no
9	sunny	cool	normal	weak	yes	yes
10	rain	mild	normal	weak	yes	yes
11	sunny	mild	normal	strong	yes	yes
12	overcast	mild	high	strong	yes	yes
13	overcast	hot	normal	weak	yes	yes
14	rain	mild	high	strong	no	no

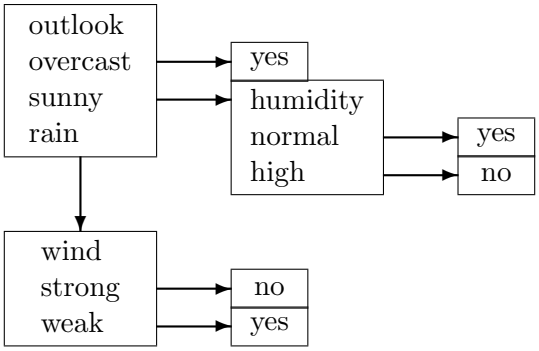


Figure 2.15: A concise decision tree based on the "outlook", "wind", "humidity" variables.

	predicted no	predicted yes
actual no	1,2,6,8,14 (5)	(0)
actual yes	(0)	3,4,5,7,9,10,11,12,13 (9)

The accuracy is equal to (5+9)/14=1.000.

The false positive rate is equal to (0)/(5+0)=0.000.

The true positive rate is equal to (9)/(0+9)=1.000.

Note that the predictive power of the "play tennis" variable alone is equal to 0.643. The predictive power of the "temperature" variable with respect to the "play tennis" target variable is also equal to 0.643. Our computation shows that deleting the "temperature" variable from the training data set the remaining explanatory variables "outlook", "wind", "humidity" still capture all the information contained in the original training data set. In fact, the tree in Figure 2.14 encapsulates all the information included in the original training data set. In short, we can say that the tree has learned all the information contained in the training data set. If our intention is to summarize the information contained by the training data set concisely in the form of a small tree, then we welcome this perfectly fitting decision tree. The decision tree in Figure 2.16 is really a condensed form of the training data set. This is the case of deterministic learning. But when the training data set is only a random sample of a much larger population, then a perfectly fitting tree tells nothing more than the tree has learned this random sample perfectly. It says nothing about how the tree would predict the target variable in an other random sample. This is the case of probabilistic learning. In this situation a perfectly fitting tree is received with suspicion. The name of the effect is over fitting. The thee is overly well fitting to a particular random sample without any hint how it would perform on other

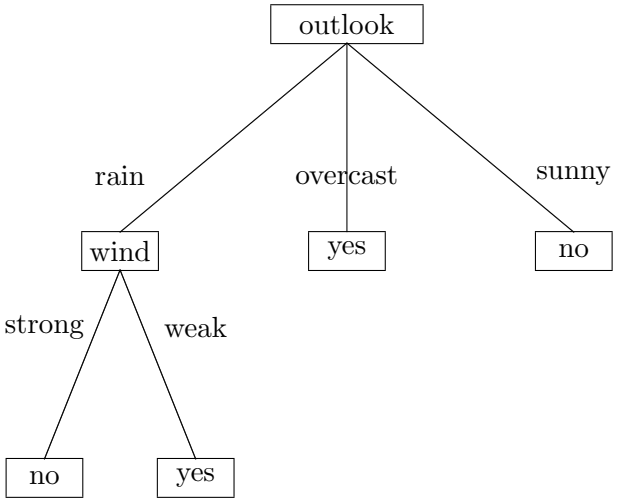


Figure 2.16: The pruned decision tree.

samples.

2.10 Pruning

We can construct further decision trees. One way to do this is to cut of branches from the extended decision tree. We refer to this operation as pruning. We assess the performance of this tree.

day	outlook	temperature	humidity	wind	play tennis actual	play tennis predicted
1	sunny	hot	high	weak	no	no
2	sunny	hot	high	strong	no	no
3	overcast	hot	high	weak	yes	yes
4	rain	mild	high	weak	yes	yes
5	rain	cool	normal	weak	yes	yes
6	rain	cool	normal	strong	no	no
7	overcast	cool	normal	strong	yes	yes
8	sunny	mild	high	weak	no	no
9	sunny	cool	normal	weak	yes	no
10	rain	mild	normal	weak	yes	yes
11	sunny	mild	normal	strong	yes	no
12	overcast	mild	high	strong	yes	yes
13	overcast	hot	normal	weak	yes	yes
14	rain	mild	high	strong	no	no

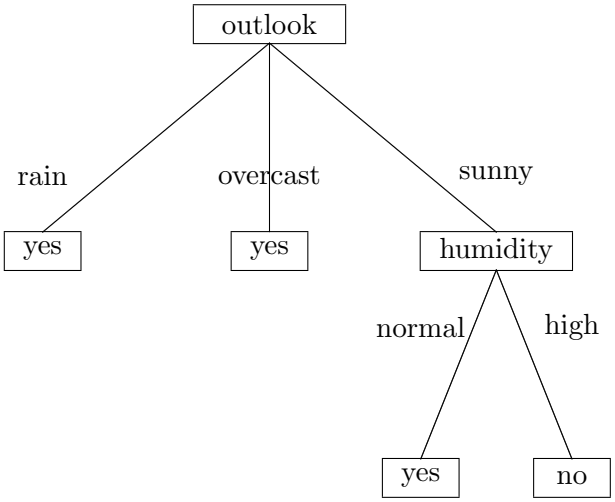


Figure 2.17: An other pruned decision tree.

	predicted no	predicted yes
actual no	1,2,6,8,14 (5)	(0)
actual yes	9,11 (2)	3,4,5,7,10,12,13 (7)

The accuracy is equal to (5+7)/14=0.857.

The false positive rate is equal to (0)/(5+0)=0.000.

The true positive rate is equal to (2)/(0+9)=0.222.

One has an other pruning opportunity as well. The next picture depicts the resulting pruned tree. Just to see how this new tree compares with the

earlier trees we carry out a performance assessment.

day	outlook	temperature	humidity	wind	play tennis actual	play tennis predicted
1	sunny	hot	high	weak	no	no
2	sunny	hot	high	strong	no	no
3	overcast	hot	high	weak	yes	yes
4	rain	mild	high	weak	yes	yes
5	rain	cool	normal	weak	yes	yes
6	rain	cool	normal	strong	no	yes
7	overcast	cool	normal	strong	yes	yes
8	sunny	mild	high	weak	no	no
9	sunny	cool	normal	weak	yes	yes
10	rain	mild	normal	weak	yes	yes
11	sunny	mild	normal	strong	yes	yes
12	overcast	mild	high	strong	yes	yes
13	overcast	hot	normal	weak	yes	yes
14	rain	mild	high	strong	no	yes

		predicted no		predicted yes	
actual	no	1,2,8	(3)		(0)
actual	yes	6,14	(2)	3,4,5,7,9,10,11,12,13	(9)

The accuracy is equal to (3+9)/14=0.857.
The false positive rate is equal to (0)/(5+0)=0.000.
The true positive rate is equal to (2)/(2+9)=0.183.

Chapter 3

Impurity index

3.1 The Gini impurity index

Let us consider a sequence of yes's and no's. The number of yes's is x and the number of no's is y. The total length of the sequence is n. We call the quantity $1-(x/n)^2-(y/n)^2$ the impurity index of the yes-no sequence and we denote it by G.

We claim that the impurity index G varies between 0 and 1/2. In order to verify the claim notice that G is an expression in terms of x and y.

$$\begin{aligned} G &= 1-(x/n)^2-(y/n)^2 \\ &= (n^2)/(n^2)-(x^2)/(n^2)-(y^2)/(n^2) \end{aligned}$$

Using $x+y=n$ we get $y=n-x$. Plugging $n-x$ in place of y in the expression of the impurity index G gives

$$\begin{aligned} G &= (n^2)/(n^2)-(x^2)/(n^2)-((n-x)^2)/(n^2) \\ &= [1/(n^2)][n^2-x^2-(n-x)^2] \\ &= [1/(n^2)][n^2-x^2-(n^2-2nx+x^2)] \\ &= [1/(n^2)][n^2-x^2-n^2+2nx-x^2)] \\ &= [1/(n^2)][-2x^2+2nx] \\ &= [2/(n^2)][-x^2+nx] \\ &= [2/(n^2)][x(n-x)]. \end{aligned}$$

The conclusion of this computation is that G is an expression in terms of x. In the coordinate system whose horizontal axis is labeled by x and whose vertical axis is labeled by G the graph of the equation $G=[2/(n^2)][x(n-x)]$ is a parabola. The parabola is upside down position as the coefficient of x^2 is negative. The parabola intersects the x axis at the points 0 and n. The x-coordinate of the apex of the parabola is $n/2$ and the G-coordinate of the apex of the parabola is

$$[2/(n^2)][(n/2)(n-(n/2))] = [2/(n^2)][(n/2)(n/2)] = 1/2.$$

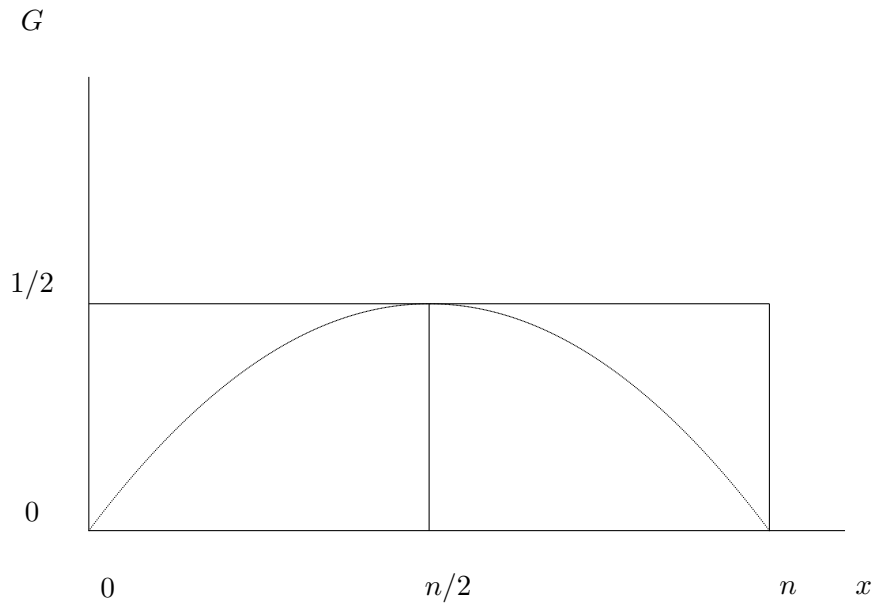


Figure 3.1: The graph of the Gini impurity index G as a function of x.

Therefore, the impurity index G is indeed running between 0 and 1/2.

The Gini impurity measure G reaches its maximum value when the entries in the sequence of yes's and no's are distributed as evenly as possible, that is, when $x = y = n/2$. The impurity measure reaches its minimum values when $x = 0$, $y = n$ or $x = n$, $y = 0$, that is, when the sequence as homogeneous as possible.

At this juncture we may define a purity index P by setting P to be equal to $(1/2) - G$. Clearly, the purity index P varies from 0 to 1/2. The purity measure P reaches its minimum when $x = y = n/2$ and it attains its maximum value when $x = 0$, $y = n$ or when $x = n$, $y = 0$.

It might look overly formal to address the introduced quantity using its full name as Gini impurity index. In case one finds it overly pedantic, then please use the names impurity index or impurity measure. There is an other index introduced by Gini (Corradi Gini an Italian staistician) which is commonly referred as Gini index and we would like to avoid the confusion between these indices.

We close this short section with two remarks. The Gini impurity index is related to probabilities. Without entering into details one may call $p = x/n$, $q = y/n$ probabilities and G will be equal to $1 - p^2 - q^2$.

The next question may occur naturally. What happens if we work with a sequence containing not only two values as yes, no but three values as yes, maybe, no. Let us consider a sequence of yes's, maybe's and no's. Let us assume that the number of yes's is x, the number of maybe's y and the

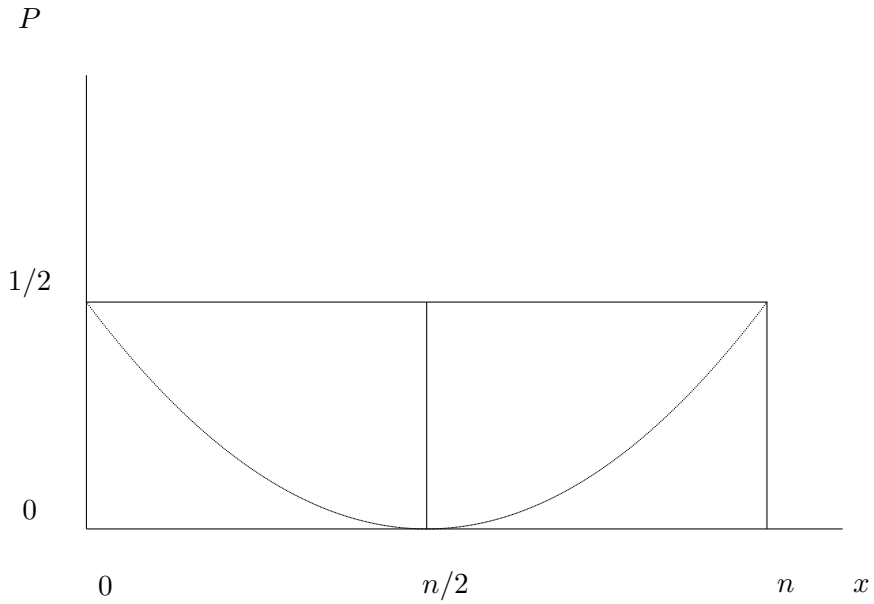


Figure 3.2: The graph of the purity index P as a function of x.

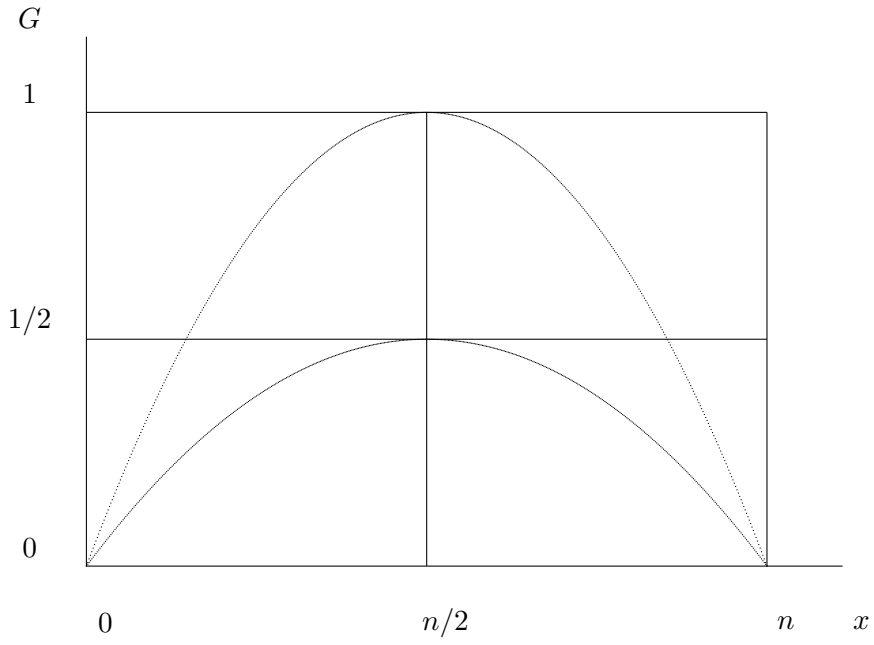


Figure 3.3: The graphs of the Gini impurity index G and the entropy S as a function of x.

number of no's is z. The total length of the sequence is n. We call the quantity $1-(x/n)^2-(y/n)^2-(z/n)^2$ the impurity index of the yes-maybe-no sequence and we denote it by G. This impurity index G varies between 0 and 2/3. It takes its maximum value at $x = y = z = n/3$. It takes its minimum value at three places $x = n$, $y = 0$, $z = 0$ or $x = 0$, $y = n$, $z = 0$ or $x = 0$, $y = 0$, $z = n$.

3.2 The entropy

Let us consider a sequence of yes's and no's. The number of yes's is x and the number of no's is y. The total length of the sequence is n. We call the quantity $(x/n)\log_2(n/x) + (y/n)\log_2(n/y)$ the entropy of the yes-no sequence and we denote it by S. At first glance the formula for S does not defined when $x = 0$ or $y = 0$. However, limit considerations make possible to define S to be 0 at these points. The values of S run between 0 and 1. The entropy S attains its maximum at $x = y = n/2$ and it attains its minimum at two places $x = 0$, $y = n$ or $x = n$, $y = 0$. As $n = x + y$ the entropy S can be viewed as a function of x. The graph of this function is similar to the function of the Gini impurity index G. The maximum values of S and G differ. Clearly, S can be used to detect purities just as well as G.

The entropy is related to probabilities. By setting $p = x/n$, $q = y/n$ the entropy turns to $S = p\log_2(1/p) + q\log_2(1/q)$. The formula

We can work with a sequence containing not only two values as yes, no but three values as yes, maybe, no. Let us consider a sequence of yes's, maybe's and no's. Let us assume that the number of yes's is x, the number of maybe's y and the number of no's is z. The total length of the sequence is n. We call the quantity $(x/n)\log_2(n/x)+(y/n)\log_2(n/y)+(z/n)\log_2(n/z)$ the entropy of the yes-maybe-no sequence and we denote it by S. This entropy S varies between 0 and 1. It takes its maximum value at $x = y = z = n/3$. It takes its minimum value at three places $x = n$, $y = 0$, $z = 0$ or $x = 0$, $y = n$, $z = 0$ or $x = 0$, $y = 0$, $z = n$.

3.3 The training data set

The first table summarizes the variables their names and ranges. The second table contains the training data set consisting of data collected on 14 days.

variable name	variable description	variable range
outlook	weather outlook	rain, overcast, sunny
temperature	air temperature	cool, mild, hot
humidity	air humidity	normal, high
wind	wind strength	weak, strong
play tennis	playing tennis	no, yes

day	outlook	temperature	humidity	wind	play tennis
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rain	mild	high	weak	yes
5	rain	cool	normal	weak	yes
6	rain	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rain	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rain	mild	high	strong	no

The "play tennis" variable is the target variable. The variables "outlook", "temperature", "humidity", "wind" are the explanatory variables. We want to predict the value of "play tennis" using the explanatory variables.

3.4 Assessing the variables one at a time

We summarize the predictive powers of the variables in the following table. The impurity index gives a hint about predictive power of this given explanatory variable. Namely, the smaller is the impurity measure, the larger is predictive power. These are the values we intend to compute in this

section.

variable	impurity
none	0.45
outlook	0.34
temperature	0.38
humidity	0.36
wind	0.43

At first we do not use any of the explanatory variables the prediction for the "play tennis". We compute a the impurity measure of the target variable.

$$\begin{aligned} G &= 1 - (5/14)^2 - (9/14)^2 \\ &= (196/196) - (25/196) - (81/196) \\ &= (196/196) - (106/196) \\ &= (90/196) \\ &= 0.45 \end{aligned}$$

The Gini impurity measure G varies between 0 and $1/2$. When the impurity measure G is equal to 0, then the purity measure P is $1/2$. In this case the prediction is straight-forward. Namely, if $x = n$, $y = 0$, then the prediction is "yes". When $x = 0$, $y = n$, then the prediction is "no". In general, the closer is the purity measure P to its maximum $1/2$, the easier is to reach a decision. In other words, the closer is the impurity measure to its minimum 0, the easier is to reach a decision.

We use the "outlook" variable to predict the "play tennis" variable. We rearrange the rows of the training data set by grouping then by the possible values of the "outlook".

day	outlook	play tennis
4	rain	yes
5	rain	yes
6	rain	no
10	rain	yes
14	rain	no
3	overcast	yes
7	overcast	yes
12	overcast	yes
13	overcast	yes
1	sunny	no
2	sunny	no
8	sunny	no
9	sunny	yes
11	sunny	yes

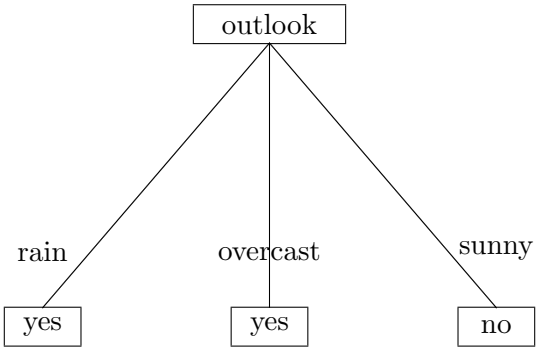


Figure 3.4: The decision tree based solely on the "outlook" variable.

The predictions rule is rain→yes, overcast→yes, sunny→no.
The impurities are 12/25, 0, 12/25.
The combined impurity is

$$(5/14)(12/25) + (4/14)(0) + (5/14)(12/25) = 12/35 = 0.34.$$

The dedicated student may notice that we can arrive to this result using the contingency table we used to construct the confusion matrix earlier.

day	temperature	play tennis
5	cool	yes
6	cool	no
7	cool	yes
9	cool	yes
4	mild	yes
8	mild	no
10	mild	yes
11	mild	yes
12	mild	yes
14	mild	no
1	hot	no
2	hot	no
3	hot	yes
13	hot	yes

The predictions rule cool→yes, mild→yes, hot→no.
The impurities are 6/16, 16/36, 1/2.
The pooled impurity is

$$(4/14)(6/16) + (6/14)(16/36) + (4/14)(1/2) = 0.38$$

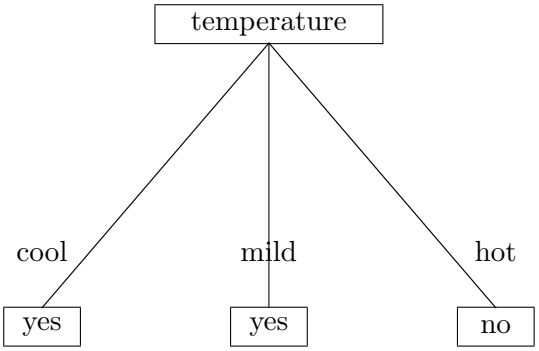


Figure 3.5: The decision tree based solely on the "temperature" variable.

day	humidity	play tennis
5	normal	yes
6	normal	no
7	normal	yes
9	normal	yes
10	normal	yes
11	normal	yes
13	normal	yes
1	high	no
2	high	no
3	high	yes
4	high	yes
8	high	no
12	high	yes
14	high	no

The predictions rule normal→yes, high→no.

The impurities are 12/49, 24/49.

The overall impurity is $(7/14)(12/49) + (7/14)(24/49) = 18/49 = 0.36$.

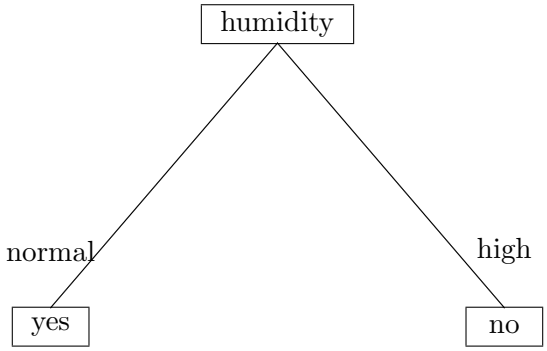


Figure 3.6: The decision tree based solely on the "humidity" variable.

day	wind	play tennis
1	weak	no
3	weak	yes
4	weak	yes
5	weak	yes
8	weak	no
9	weak	yes
10	weak	yes
13	weak	yes
2	strong	no
6	strong	no
7	strong	yes
11	strong	yes
12	strong	yes
14	strong	no

The predictions rule weak→yes, strong→no.
The impurities are 24/64, 1/2.
The overall impurity is

$$(8/14)(24/64) + (6/14)(1/2) = 0.43$$

3.5 A short cut in the book keeping

The dedicated student may notice that we can gather the ingredients for computing the combined impurities by filling in the contingency table we applied to construct the confusion matrix earlier. In the future we will use

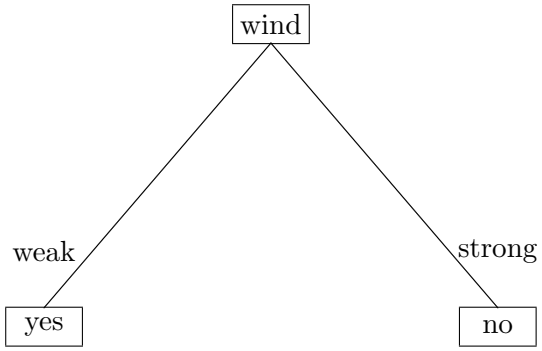


Figure 3.7: The decision tree based on the "wind" variable alone.

this short cut systematically. The long winged way of writing up the rearranged rows of the training data helps us to develop a firmer understanding the details of the techniques we intend to learn. In short, it has some pedagogical importance.

	outlook rain		outlook overcast		outlook sunny	
play tennis no	6,14	(2)		(0)	1,2,8	(3)
play tennis yes	4,5,10	(3)	3,7,12,13	(4)	9,11	(2)
impurity	12/25		0		12/25	

	tempe-rature cool		tempe-rature mild		tempe-rature hot	
play tennis no	6	(1)	8,14	(2)	1,2	(2)
play tennis yes	5,7,9	(3)	4,10,11,12	(4)	3,13	(2)
impurity	6/16		16/36		1/2	

	humidity normal		humidity high	
play tennis no	6	(1)	1,2,8,14	(4)
play tennis yes	5,7,9,10,11,13	(6)	3,4,12	(3)
impurity	12/49		24/49	

	wind strong		wind weak	
play tennis no	2,6,14	(3)	1,8	(2)
play tennis yes	7,11,12	(3)	3,4,5,9,10,13	(6)
impurity	24/64		1/2	

3.6 Splitting the training data set

The predictive power of the variable "outlook" is the largest as its associated impurity is the smallest in comparison with the variables "temperature", "humidity", "wind". This variable will be used to add new branches to the decision tree. It takes on three values "rain", "overcast", "sunny". We split the training data set into three smaller data sets.

day	outlook	temperature	humidity	wind	play tennis
4	rain	mild	high	weak	yes
5	rain	cool	normal	weak	yes
6	rain	cool	normal	strong	no
10	rain	mild	normal	weak	yes
14	rain	mild	high	strong	no
3	overcast	hot	high	weak	yes
7	overcast	cool	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
11	sunny	mild	normal	strong	yes

3.7 The "outlook" variable is equal to "rain"

We are working with the following small training data set.

day	outlook	temperature	humidity	wind	play tennis
4	rain	mild	high	weak	yes
5	rain	cool	normal	weak	yes
6	rain	cool	normal	strong	no
10	rain	mild	normal	weak	yes
14	rain	mild	high	strong	no

We summarize the predictive powers of the variables in the following table. These are the values we intended to compute in this section.

variable	impurity
temperature	0.47
humidity	0.26
wind	0.00

First, we consider the case when the "temperature" explanatory variable is used to predict the the "play tennis" target variable.

	temperature cool (no)	temperature mild (yes)	temperature hot (no)
play tennis no	6 (1)	14 (1)	(0)
play tennis yes	5 (1)	4,10 (2)	(0)

The impurities are 1/2, 4/9, DND (does not defined).

The combined impurity is (2/5)(1/2)+(3/5)(4/9)+(0)(DND)=0.47.

We turn to predict the "play tennis" variable using the "humidity" variable.

	humidity normal (yes)	humidity high (no)
play tennis no	6 (1)	(0)
play tennis yes	5,10 (2)	4,14 (2)

The impurities are 4/9, 0.

The combined impurity is (3/5)(4/9)+(2/5)(0)=0.26.

Finally, the "wind" variable is used to predict the "play tennis" variable.

	wind strong (no)	wind weak (yes)
play tennis no	6,14 (2)	(0)
play tennis yes	(0)	4,5,10 (3)

The impurities are 0, 0. The combined impurity is 0.

3.8 The "outlook" variable is equal to "overcast"

The small training date set we will be busy with is the following.

day	outlook	temperature	humidity	wind	play tennis
3	overcast	hot	high	weak	yes
7	overcast	cool	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes

This is pure node. We do not need to compute the predictive powers of the variables "temperature", "humidity", "wind". No branching occurs at this node.

3.9 The "outlook" variable is equal to "sunny"

The small training date set we use is contained in the next table.

day	outlook	temperature	humidity	wind	play tennis
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
11	sunny	mild	normal	strong	yes

We summarize the predictive powers of the variables in the following table. These are the values we plan to compute in this section.

variable	impurity
temperature	0.20
humidity	0.00
wind	0.26

	temperature cool	(yes)	temperature mild	(yes)	temperature hot	(no)
play tennis no		(0)	8	(1)	1,2	(2)
play tennis yes	9	(1)	11	(1)		(0)

The impurities are 0, 1/2, 0.

The combined impurity is (1/5)(0)+(2/5)(1/2)+(2/5)(0)=1/5=0.20.

We turn to predict the "play tennis" variable using the "humidity" variable.

	humidity normal	(yes)	humidity high	(no)
play tennis no		(0)	1,2,8	(3)
play tennis yes	9,11	(2)		(0)

The impurities are 0, 0.

The combined impurity is 0.

Finally, the "wind" variable is used to predict the "play tennis" variable.

	wind strong	(yes)	wind weak	(no)
play tennis no	2	(1)	1,8	(2)
play tennis yes	11	(1)	9	(1)

The impurities are 0, 4/9.

The combined impurity is (2/5)(0)+(3/5)(4/9)=4/15=0.26.

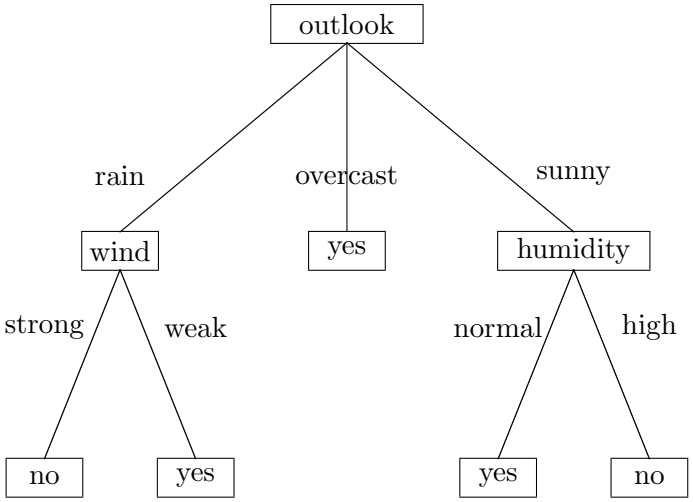


Figure 3.8: The decision tree based on the "outlook", "wind", "humidity" variables.

3.10 The extended decision tree

In the small data set when the "outlook" variable takes the "rain" value the best predictor for the "play tennis" is the "wind" attribute.

In the small data set when the "outlook" variable takes the "overcast" value there is no need for further predictor variable.

In the small data set when the "outlook" variable takes the "sunny" value the best predictor for the "play tennis" is the "humidity" feature. Figure 2.16 is exactly the same as Figure 2.14. We included the the geometric representation of the tree again because we do not expect the reader go through all chapters. The chapters can be studied independently of each other.

Chapter 4

Regression using variance

4.1 The training data set

The first table summarizes the variables their names and ranges. The second table contains the training data set consisting of date collected on 14 days.

variable name	variable description	variable range
outlook	weather outlook	rain, overcast, sunny
temperature	air temperature	cool, mild, hot
humidity	air humidity	normal, high
wind	wind strength	weak, strong
play time	playing time	$0, 1, 2, \ldots$

day	outlook	temperature	humidity	wind	play time
1	sunny	hot	high	weak	25
2	sunny	hot	high	strong	30
3	overcast	hot	high	weak	46
4	rain	mild	high	weak	45
5	rain	cool	normal	weak	52
6	rain	cool	normal	strong	23
7	overcast	cool	normal	strong	43
8	sunny	mild	high	weak	35
9	sunny	cool	normal	weak	38
10	rain	mild	normal	weak	46
11	sunny	mild	normal	strong	48
12	overcast	mild	high	strong	52
13	overcast	hot	normal	weak	44
14	rain	mild	high	strong	30

The "play time" variable is the target variable. The variables "outlook", "temperature", "humidity", "wind" are the explanatory variables. We want to predict the value of "play time" using the explanatory variables.

4.2 The predictive power of the variables

When we do not use any of the explanatory variables the prediction for the "play time" is the mean of the "play time" values. The mean is equal to 39.8. The standard deviation is 9.3 This can be interpreted as an uncertainty of the prediction. In other words, the prediction is 39.8 and the uncertainty is 9.3. The prediction rule is ()→39.8.

day	play time	prediction	deviation	deviation square
1	25	39.8	14.8	219.0
2	30	39.8	9.8	96.0
3	46	39.8	6.2	38.4
4	45	39.8	5.2	27.0
5	52	39.8	12.2	148.8
6	23	39.8	16.8	282.2
7	43	39.8	3.2	10.2
8	35	39.8	4.8	23.0
9	38	39.8	1.8	3.2
10	46	39.8	6.2	38.4
11	48	39.8	8.2	67.2
12	52	39.8	12.2	148.8
13	44	39.8	4.2	17.6
14	30	39.8	9.8	96.0

The table below is a summary of the predictive powers of the variables. The smaller is the listed uncertainty, the larger is the predictive power of the explanatory variable.

variable	uncertainty
baseline	9.3
outlook	8.2
temperature	8.9
humidity	9.1
wind	9.1

We use the "outlook" variable to predict the "play time" target variable.

day	outlook	play time	prediction	deviation	deviation square
4	rain	45	39.2	5.8	33.6
5	rain	52	39.2	12.8	163.8
6	rain	23	39.2	16.2	262.4
10	rain	46	39.2	6.8	46.2
14	rain	30	39.2	9.2	84.6
3	overcast	46	46.2	0.2	0.0
7	overcast	43	46.2	13.2	17.4
12	overcast	52	46.2	5.8	33.6
13	overcast	44	46.2	2.2	4.8
1	sunny	25	35.2	10.2	104.0
2	sunny	30	35.2	5.2	27.0
8	sunny	35	35.2	0.2	0.0
9	sunny	38	35.2	2.8	7.8
11	sunny	48	35.2	12.8	163.8

The prediction rule is rain→39.2, overcast→46.2, sunny→35.2.

The uncertainties are 10.9, 3.7, 7.8.

The total uncertainty is 8.2.

The estimated total uncertainty is

$$\begin{aligned} U &= (5/14)(9.9) + (4/14)(3.7) + (5/14)(7.8) \\ &= (49.5/14) + (14.8/14) + (39.0/14) \\ &= (103.3/14) = 7.4. \end{aligned}$$

Figure 4.1 is a possible geometric representation of the regression tree based solely on the "outlook" variable. We name this tree T_{out}.

We turn our attention to predicting the "play time" target variable using

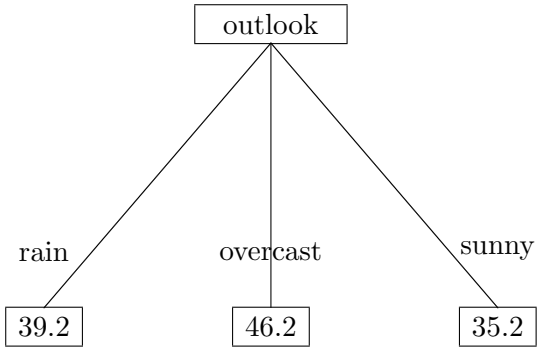


Figure 4.1: The regression tree T_{out} based solely on the "outlook" variable.

the "temperature" explanatory variable.

day	temperature	play time	prediction	deviation	deviation square
5	cool	52	39.0	13.0	169.0
6	cool	23	39.0	16.0	256.0
7	cool	43	39.0	4.0	16.0
9	cool	38	39.0	1.0	1.0
4	mild	45	42.7	2.3	5.3
8	mild	35	42.7	7.7	59.3
10	mild	46	42.7	3.3	10.9
11	mild	48	42.7	5.3	28.1
12	mild	52	42.7	9.3	86.5
14	mild	30	42.7	12.7	161.3
1	hot	25	36.2	11.2	124.4
2	hot	30	36.2	6.2	38.4
3	hot	46	36.2	9.8	96.0
13	hot	44	36.2	7.8	60.8

The prediction rule is cool→39, mild→42.7, hot→36.2.
The uncertainties are 10.5, 7.6, 8.9.
The total uncertainty is 8.9.
The pooled estimated uncertainty is

$$\begin{aligned} U &= (4/14)(10.5) + (6/14)(7.6) + (4/14)(8.9) \\ &= (42/14) + (45.6/14) + (35.6/14) \\ &= (123.2/14) = 8.8. \end{aligned}$$

The regression tree based solely on the "temperature" variable can be seen in Figure **??**. T_{temp} will be the name of this tree. We focus our attention

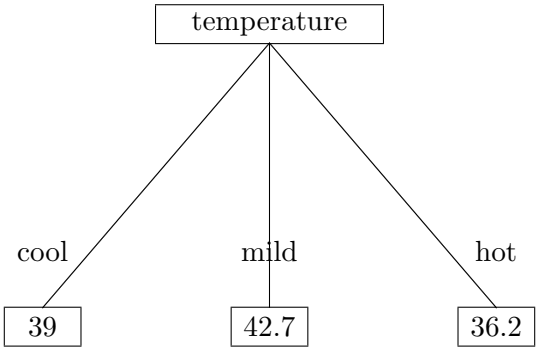


Figure 4.2: The regression tree T_{temp} based solely on the "temperature" variable.

on predicting the "play time" using the "humidity" variable.

day	humidity	play time	predicted	deviation	deviation square
5	normal	52	42.4	9.6	92.2
6	normal	23	42.4	19.4	376.4
7	normal	43	42.4	0.6	0.4
9	normal	38	42.4	4.4	19.4
10	normal	46	42.4	3.6	13.0
11	normal	48	42.4	5.6	31.4
13	normal	44	42.4	1.6	2.6
1	high	25	37.6	12.6	158.8
2	high	30	37.6	7.6	57.8
3	high	46	37.6	8.4	70.6
4	high	45	37.6	7.4	54.8
8	high	35	37.6	2.6	6.7
12	high	52	37.6	14.4	207.4
14	high	30	37.6	7.6	57.8

The prediction rule is normal→42.4, high→37.6.
The uncertainties are 8.7, 9.4.
The total uncertainty is 9.1.
The combined estimated uncertainty is

$$\begin{aligned} U &= (1/2)(8.7) + (1/2)(9.4) \\ &= (18.1/2) = 9.1. \end{aligned}$$

The regression tree based solely on the "humidity" variable is denoted by T_{hum}.

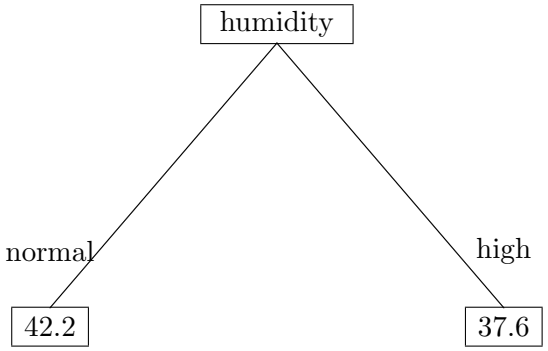


Figure 4.3: The regression tree T_{hum} based solely on the "humidity" variable.

Finally, the "wind" explanatory variable is used to predict the "play time" target variable.

day	wind	play time	prediction	deviation	deviation square
1	weak	25	41.4	16.4	269.0
3	weak	46	41.4	4.6	21.2
4	weak	45	41.4	3.6	13.0
5	weak	52	41.4	10.6	112.4
8	weak	35	41.4	6.4	41.0
9	weak	38	41.4	3.4	11.6
10	weak	46	41.4	4.6	21.2
13	weak	44	41.4	2.6	6.8
2	strong	30	37.7	7.7	59.3
6	strong	23	37.7	14.7	216.0
7	strong	43	37.7	5.3	28.1
11	strong	48	37.7	10.3	106.1
12	strong	52	37.7	14.3	204.5
14	strong	30	37.7	7.7	59.3

The predictions are weak→41.4, strong→37.7.
The uncertainties are 7.9, 10.6.
The total uncertainty is 9.1.
The overall uncertainty is

$$\begin{aligned} U &= (8/14)(7.9) + (6/14)(10.6) \\ &= (4/7)(7.9) + (3/7)(10.6) \\ &= (31.6/7) + (31.8/7) = (63.4/7) = 9.1. \end{aligned}$$

The regression tree based on the "wind" variable alone is depicted in Figure 4.4. We refer to this tree as T_{wind}.

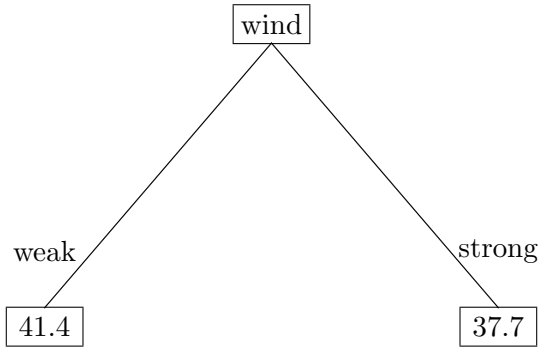


Figure 4.4: The regression tree T_{wind} based on the "wind" variable alone.

4.3 The ensemble method

Let us form a committee consisting of the two best performing regression trees T_{out} and T_{temp}. We compute the predictive power of the committee $\{T_{out}, T_{temp}\}$.

day	outlook	tempe-rature	play time actual	play time predicted	deviation	deviation square
1	sunny	hot	25	35.7	10.7	114.5
2	sunny	hot	30	35.7	5.7	32.5
3	overcast	hot	46	41.2	4.8	23.0
4	rain	mild	45	40.9	4.1	16.8
5	rain	cool	52	39.1	12.9	166.4
6	rain	cool	23	39.1	16.1	259.2
7	overcast	cool	43	42.6	0.4	0.2
8	sunny	mild	35	38.9	3.9	15.2
9	sunny	cool	38	37.1	0.9	0.8
10	rain	mild	46	40.9	5.1	26.0
11	sunny	mild	48	38.9	9.1	82.8
12	overcast	mild	52	44.5	7.5	56.3
13	overcast	hot	44	41.2	2.8	7.8
14	rain	mild	30	40.9	10.9	118.8

The uncertainty of the prediction is 7.6. The predictive power of the committee $\{T_{out}, T_{temp}\}$ is better than the predictive power of the tree T_{out}.

4.4 Splitting the training data set

Among the explanatory variables "outlook", "temperature", "humidity", "wind" the "outlook" variable has the largest predictive power. In other words, among the trees T_{out}, T_{temp}, T_{hum}, T_{wind} the first one has the smallest uncertainty. Therefore, we take the regression tree T_{out} and try to extend it to get a regression tree with a better predictive power. The prediction rule of the tree T_{out} is rain$\rightarrow$39.2, overcast$\rightarrow$46.2, sunny$\rightarrow$35.2. The corresponding associated uncertainties are 10.9, 3.7, 7.8. The largest of these uncertainties is 10.9. Reducing this uncertainty promises a reduction in the uncertainty of the extended tree.

4.5 Outlook takes on the value rain

We focus our attention to that part of training set where the variable "outlook" takes on the value "rain". There are 5 such days. Namely, days 4,5,6,10,14.

day	outlook	temperature	humidity	wind	play time
4	rain	mild	high	weak	45
5	rain	cool	normal	weak	52
6	rain	cool	normal	strong	23
10	rain	mild	normal	weak	46
14	rain	mild	high	strong	30

We compute the predictive power of the variables "temperature", "humidity", "wind". The next table is a summary of the predictive powers of the variables. The row labeled by "baseline" contains the predictive power of the "outcome" variable without using any extra information provided by the any of the "temperature", "humidity", "wind" variable. Basically, we are interested in how much the extra variables can improve on this base line predictive power.

variable	uncertainty
baseline	8.2
temperature	8.2
humidity	8.2
wind	5.4

We use the "outlook" and "temperature" variables to predict the "play time"

target variable. (The "outlook" takes on the value "rain".)

day	temperature	play time	prediction	deviation	deviation square
5	cool	52	37.5	14.5	210.3
6	cool	23	37.5	14.5	210.3
4	mild	45	41.7	3.3	10.9
10	mild	46	41.7	4.3	18.5
14	mild	30	41.7	11.7	136.9

The prediction rules are cool→37.5, mild→41.7. The uncertainty of the prediction is 8.2. Using the last column of the following table we can get that the uncertainty of the prediction is 8.2. This 8.2 should be compared with the baseline uncertainty 9.3.

day	outlook	play time	prediction	deviation	deviation square
4	rain	45	41.7	3.3	10.9
5	rain	52	37.5	14.5	210.3
6	rain	23	37.5	14.5	210.3
10	rain	46	41.7	4.3	18.5
14	rain	30	41.7	11.7	136.9
3	overcast	46	46.2	0.2	0.0
7	overcast	43	46.2	3.2	10.2
12	overcast	52	46.2	5.8	33.6
13	overcast	44	46.2	2.2	4.8
1	sunny	25	35.2	10.2	104.0
2	sunny	30	35.2	5.2	27.0
8	sunny	35	35.2	0.2	0.0
9	sunny	38	35.2	2.8	7.8
11	sunny	48	35.2	12.8	163.8

A possible geometric representation of the regression tree have just constructed is depicted in Figure 4.5.

We use the "outlook" and "humidity" variables to predict the "play time" target variable. (The "outlook" takes on the value "rain".)

day	humidity	play time	prediction	deviation	deviation square
5	normal	52	40.3	11.7	136.9
6	normal	23	40.3	17.3	299.3
10	normal	46	40.3	5.7	32.5
4	high	45	37.5	7.5	56.3
14	high	30	37.5	7.5	56.3

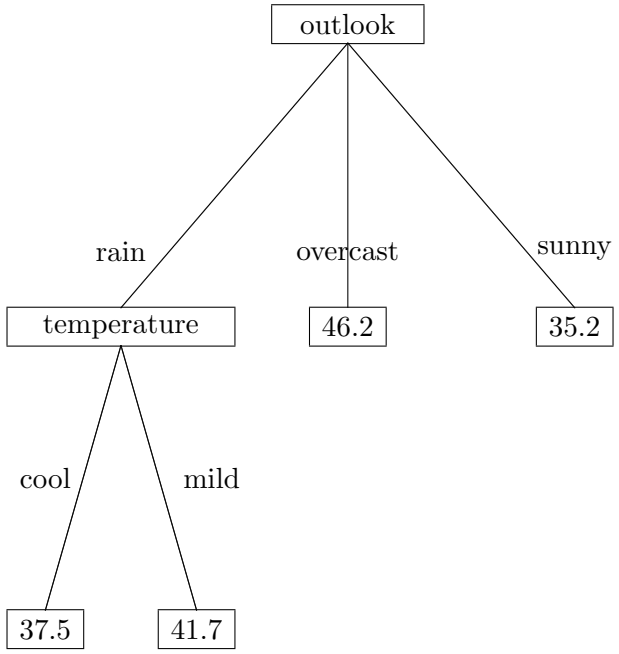


Figure 4.5: The regression tree we have just constructed.

The prediction rules are normal→40.3, high→37.5. The uncertainty of the prediction is 8.2.

day	outlook	play time	prediction	deviation	deviation square
4	rain	45	37.5	7.5	56.3
5	rain	52	40.3	11.7	136.9
6	rain	23	40.3	17.3	299.3
10	rain	46	40.3	5.7	32.5
14	rain	30	37.5	7.5	56.3
3	overcast	46	46.2	0.2	0.0
7	overcast	43	46.2	3.2	10.2
12	overcast	52	46.2	5.8	33.6
13	overcast	44	46.2	2.2	4.8
1	sunny	25	35.2	10.2	104.0
2	sunny	30	35.2	5.2	27.0
8	sunny	35	35.2	0.2	0.0
9	sunny	38	35.2	2.8	7.8
11	sunny	48	35.2	12.8	163.8

Figure 4.5 is a geometric representation of the regression tree based on the variables “outlook” and “humidity”.

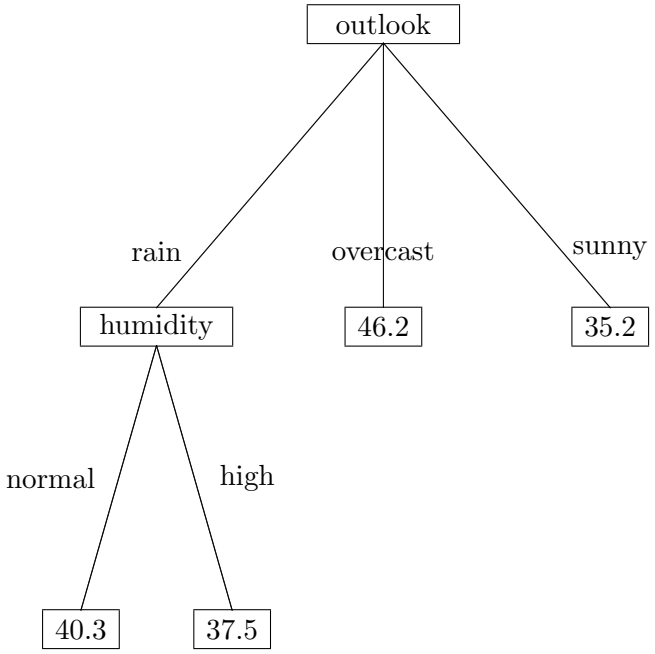


Figure 4.6: The regression tree based on the variables "outlook" and "humidity".

We use the "outlook" and "wind" variables to predict the "play time" target variable. (The "outlook" takes on the value "rain".)

day	wind	play time	prediction	deviation	deviation square
4	weak	45	47.7	2.7	7.3
5	weak	52	47.7	4.3	18.5
10	weak	46	47.7	1.7	2.9
6	strong	23	26.5	3.5	12.3
14	strong	30	26.5	3.5	12.3

The prediction rules are weak→47.7, strong→26.5. The uncertainty of the prediction is 5.4.

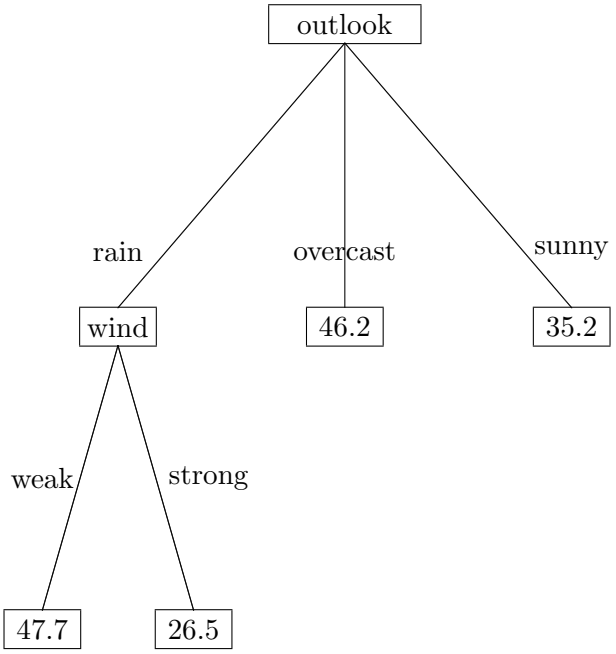


Figure 4.7: The regression tree with the largest predictive power.

day	outlook	play time	prediction	deviation	deviation square
4	rain	45	47.7	2.7	7.3
5	rain	52	47.7	4.3	18.5
6	rain	23	26.5	3.5	12.3
10	rain	46	47.7	1.7	2.9
14	rain	30	26.5	3.5	12.3
3	overcast	46	46.2	0.2	0.0
7	overcast	43	46.2	3.2	10.2
12	overcast	52	46.2	5.8	33.6
13	overcast	44	46.2	2.2	4.8
1	sunny	25	35.2	10.2	104.0
2	sunny	30	35.2	5.2	27.0
8	sunny	35	35.2	0.2	0.0
9	sunny	38	35.2	2.8	7.8
11	sunny	48	35.2	12.8	163.8

The regression tree with the largest predictive power we have located until now is depicted in Figure 4.7.

Chapter 5

Regression via averages

5.1 The training data set

The first table summarizes the variables their names and ranges. The second table contains the training data set consisting of date collected on 14 days.

variable name	variable description	variable range
outlook	weather outlook	rain, overcast, sunny
temperature	air temperature	cool, mild, hot
humidity	air humidity	normal, high
wind	wind strength	weak, strong
play time	playing time	$0, 1, 2, \ldots$

day	outlook	temperature	humidity	wind	play time
1	sunny	hot	high	weak	25
2	sunny	hot	high	strong	30
3	overcast	hot	high	weak	46
4	rain	mild	high	weak	45
5	rain	cool	normal	weak	52
6	rain	cool	normal	strong	23
7	overcast	cool	normal	strong	43
8	sunny	mild	high	weak	35
9	sunny	cool	normal	weak	38
10	rain	mild	normal	weak	46
11	sunny	mild	normal	strong	48
12	overcast	mild	high	strong	52
13	overcast	hot	normal	weak	44
14	rain	mild	high	strong	30

The "play time" variable is the target variable. The variables "outlook",

"temperature", "humidity", "wind" are the explanatory variables. We want to predict the value of "play time" using the explanatory variables.

5.2 No explanatory variable is used for prediction

When we do not use any of the explanatory variables the prediction for the "play time" is the average of the "play time" values. The average is equal to 39.8. The average of the deviations is equal to 8.3. This can be interpreted as an uncertainty of the prediction. The smaller is the uncertainty the more information the explanatory variables provide about the target variable. Stating it differently, the smaller is the uncertainty the larger is the predictive power of the explanatory variable.

day	play time actual	play time predicted	deviation
1	25	39.8	14.8
2	30	39.8	9.8
3	46	39.8	6.2
4	45	39.8	5.2
5	52	39.8	12.2
6	23	39.8	16.8
7	43	39.8	3.2
8	35	39.8	4.8
9	38	39.8	1.8
10	46	39.8	6.2
11	48	39.8	8.2
12	52	39.8	12.2
13	44	39.8	4.2
14	30	39.8	9.8

We summarize the uncertainties associated with the predictions of the regression trees based on the variables "outlook", "temperature", "humidity", "wind". These are the values we intend to compute in the next sections.

variable	uncertainty
none	8.3
outlook	6.8
temperature	7.8
humidity	7.6
wind	8.1

5.3 The "outlook" variable predicts the "play time"

We will use the "outlook" variable to predict the tennis "play time". The prediction rule is rain→39.2, overcast→46.3, sunny→35.2. the corresponding uncertainties are 10.0, 4.2, 6.2. The average of the deviations is 6.8. In other words, the uncertainty of the prediction is 6.8.

Some practicioner computes the weighted average of the deviations

$$\begin{aligned} U &= (5/14)(10.0) + (4/14)(4.2) + (5/14)(6.2) \\ &= (50/14) + (16.8/14) + (31/14) \\ &= (97.8/14) = 6.1. \end{aligned}$$

This may serve as an estimate of the overall uncertainty.

day	outlook	play time actual	play time predicted	deviation
4	rain	45	39.2	5.8
5	rain	52	39.2	12.8
6	rain	23	39.2	16.2
10	rain	46	39.2	6.8
14	rain	30	39.2	9.2
3	overcast	46	46.3	0.3
7	overcast	43	46.3	3.3
12	overcast	52	46.3	5.7
13	overcast	44	46.3	3.3
1	sunny	25	35.2	10.2
2	sunny	30	35.2	5.2
8	sunny	35	35.2	0.2
9	sunny	38	35.2	2.8
11	sunny	48	35.2	12.8

Figure 5.1 illustrates the regression tree based solely on the "outlook" variable.

5.4 The "temperature" variable predicts the "play time"

We will use the "temperature" variable to predict the values of the "play time" variable. The prediction rule is cool→39, mild→42.7, hot→38.8. The corresponding uncertainties are 8.5, 6.8, 8.8. The uncertainty of the prediction is 7.8. The weighted average of the deviations

$$U = (4/14)(8.5) + (6/14)(6.8) + (4/14)(8.8)$$

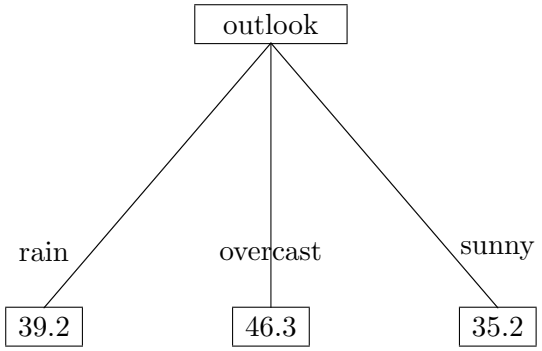


Figure 5.1: The regression tree based solely on the "outlook" variable.

$$
\begin{aligned}
&= (34/14) + (40.8/14) + (35.2/14) \\
&= (110/14) = 7.9.
\end{aligned}
$$

This is an estimate of uncertainty of the prediction.

day	temperature	play time actual	play time predicted	deviation
5	cool	52	39.0	13.0
6	cool	23	39.0	16.0
7	cool	43	39.0	4.0
9	cool	38	39.0	1.0
4	mild	45	42.7	2.3
8	mild	35	42.7	7.7
10	mild	46	42.7	3.3
11	mild	48	42.7	5.3
12	mild	52	42.7	9.3
14	mild	30	42.7	12.7
1	hot	25	38.8	13.8
2	hot	30	38.8	8.8
3	hot	46	38.8	7.2
13	hot	44	38.8	5.2

The end result of our work is the regression tree based solely on the "temperature" variable. This thee is depicted in Figure 5.2.

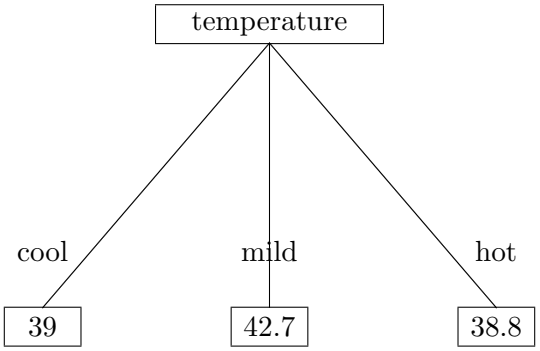


Figure 5.2: The regression tree based solely on the "temperature" variable.

5.5 The "humidity" variable predicts the "play time"

Next we turn to predict the values of the "play time" using the values of the "humidity" variable. The prediction rule is normal→42, high→37.6. The corresponding ancertainties are 6.6, 8.7. The average of the deviations is 7.6 and so the uncertainty of the prediction is 7.6. We compute the weighted average of the deviations

$$\begin{aligned} U &= (7/14)(6.6) + (7/14)(8.7) \\ &= (1/2)(6.6) + (1/2)(8.7) = 7.7. \end{aligned}$$

day	humidity	play time actual	play time predicted	deviation
5	normal	52	42.0	10.0
6	normal	23	42.0	19.0
7	normal	43	42.0	1.0
9	normal	38	42.0	4.0
10	normal	46	42.0	4.0
11	normal	48	42.0	6.0
13	normal	44	42.0	2.0
1	high	25	37.6	12.6
2	high	30	37.6	7.6
3	high	46	37.6	8.4
4	high	45	37.6	7.4
8	high	35	37.6	2.6
12	high	52	37.6	14.4
14	high	30	37.6	7.6

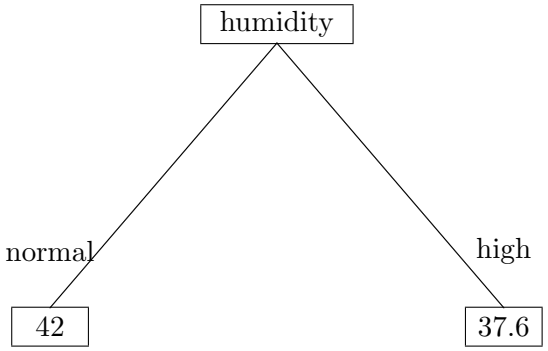


Figure 5.3: The regression tree based solely on the "humidity" variable.

We draw a picture of the regression tree based solely on the "humidity" variable. Figure 5.3 is the result of our drawing.

5.6 The "wind" variable predicts the "play time"

Finally we will predict the values of the "play time" using the values of the "wind" variable. The predictions rules are weak→42.6, strong→37.7. The uncertainty of the prediction is 8.1.

day	wind	play time actual	play time predicted	deviation
1	weak	25	42.6	17.6
3	weak	46	42.6	3.4
4	weak	45	42.6	2.4
5	weak	52	42.6	9.4
8	weak	35	42.6	7.6
9	weak	38	42.6	4.6
10	weak	46	42.6	3.4
13	weak	44	42.6	1.4
2	strong	30	37.7	7.7
6	strong	23	37.7	14.7
7	strong	43	37.7	5.3
11	strong	48	37.7	10.3
12	strong	52	37.7	14.3
14	strong	30	37.7	7.7

We may summarize the result of our considerations by presenting the regression tree based on the "wind" variable alone. See Figure 5.4

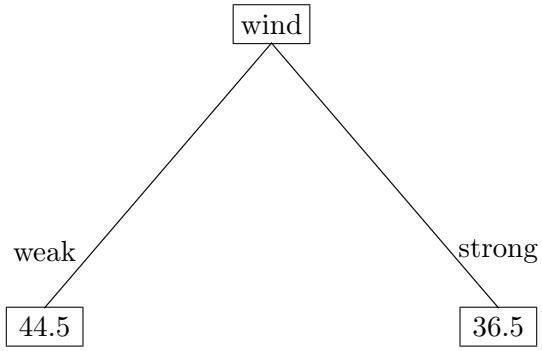


Figure 5.4: The regression tree based on the "wind" variable alone.

5.7 The ensemble method

So far we have constructed four regression trees to predict the "play time" variable. We do not count the tree when none of the explanatory variable was used. Our computation shows that the explanatory variable "outlook" carries the most information about the "play time". Among the remaining explanatory variables the "humidity" is the best explanatory variable.

Let us form committees of consisting of some of the regression trees we already have and see if they perform together better than individually. We use the regression trees bases on the "outlook" and "humidity" explanatory variables.

day	outlook	humidity	play time actual	play time predicted	deviation
1	sunny	high	25	36.4	11.4
2	sunny	high	30	36.4	6.4
3	overcast	high	46	42.0	4.0
4	rain	high	45	38.4	6.6
5	rain	normal	52	40.6	11.4
6	rain	normal	23	40.6	17.6
7	overcast	normal	43	44.2	1.2
8	sunny	high	35	36.4	1.4
9	sunny	normal	38	38.6	0.6
10	rain	normal	46	40.6	5.4
11	sunny	normal	48	38.6	9.4
12	overcast	high	52	42.0	10.0
13	overcast	normal	44	44.2	0.2
14	rain	high	30	38.4	8.4

The average of the deviations is 6.7. The uncertainty of the prediction of

the committee is 6.7. This is a slight improvement over the prediction using the explanatory variable "outlook".

5.8 "Outlook" takes on the value "rain"

We are extending the tree which predicts the "play time" based on the "outlook" variable. There are 5 days when the "outlook" variable takes on the value "rain".

day	outlook	temperature	humidity	wind	play time
4	rain	mild	high	weak	45
5	rain	cool	normal	weak	52
6	rain	cool	normal	strong	23
10	rain	mild	normal	weak	46
14	rain	mild	high	strong	30

We start with using the "temperature" variable to predict "play time".

day	temperature	play time
5	cool	52
6	cool	23
4	mild	45
10	mild	46
14	mild	30

The prediction rule is cool→37.5, mild→40.3. Figure 5.5 combines the decision rules based on the variables "outlook" and "temperature". We assess the predictive power of this regression tree by computing the uncertainty of

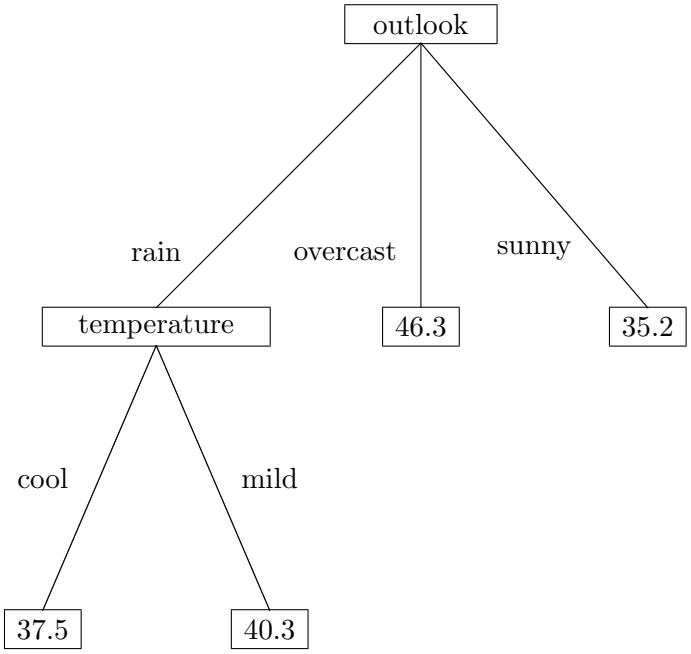


Figure 5.5: The regression tree based on the "outlook" and "temperature" variables.

the prediction.

day	outlook	tempe-rature	play time actual	play time predicted	deviation
5	rain	cool	52	37.5	14.5
6	rain	cool	23	37.5	14.5
4	rain	mild	45	40.3	4.7
10	rain	mild	46	40.3	5.7
14	rain	mild	30	40.3	10.3
3	overcast	—	46	46.3	0.3
7	overcast	—	43	46.3	3.3
12	overcast	—	52	46.3	5.7
13	overcast	—	44	46.3	2.3
1	sunny	—	25	35.2	10.2
2	sunny	—	30	35.2	5.2
8	sunny	—	35	35.2	0.2
9	sunny	—	38	35.2	2.8
11	sunny	—	48	35.2	12.8

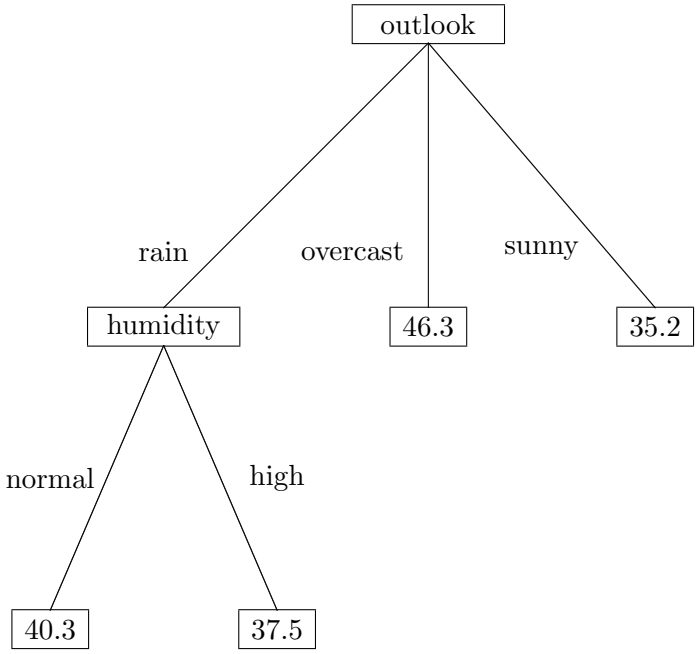


Figure 5.6: The regression tree based on the "outlook" and "humidity" variables.

The average of the deviations is 6.6. Thus the uncertainty of the prediction is 6.6. This is slightly better than the prediction bases solely on the "outlook" variable. However, this larger tree may not be prefereable to the smaller tree as we estimated 3 parameters in the small tree and we have estimated 5 parameters in the larger tree.

Next we use the "humidity" to predict "play time".

day	humidity	play time
5	normal	52
6	normal	23
10	normal	46
4	high	45
14	high	30

The prediction rule is normal→40.3, high→37.5. Figure 5.6 depicts concisely the decision rules based on the variables "outlook" and "humidity". We compute the uncertainty of the prediction of this tree. In this way we can

assess the predictive power of the tree.

day	outlook	humidity	play time actual	play time predicted	deviation
5	rain	normal	52	40.3	11.7
6	rain	normal	23	40.3	17.3
10	rain	normal	46	40.3	5.7
4	rain	high	45	37.5	7.5
14	rain	high	30	37.5	7.5
3	overcast	—	46	46.3	0.3
7	overcast	—	43	46.3	3.3
12	overcast	—	52	46.3	5.7
13	overcast	—	44	46.3	2.3
1	sunny	—	25	35.2	10.2
2	sunny	—	30	35.2	5.2
8	sunny	—	35	35.2	0.2
9	sunny	—	38	35.2	2.8
11	sunny	—	48	35.2	12.8

The average of the deviations is 7. This is the uncertainty of the prediction of the tree. It does not compare favorably to the earlier trees. Therefore, we drop this tree.

Finally, we use the "wind" to predict "play time".

day	wind	play time
4	weak	45
5	weak	52
10	weak	46
6	strong	23
14	strong	30

The prediction rule is weak→47.7, strong→26.5. We summarize the decision rules based on the variables "outlook" and "humidity" in Figure 5.7. We compute the uncertainty of the prediction of this tree. In this way we can

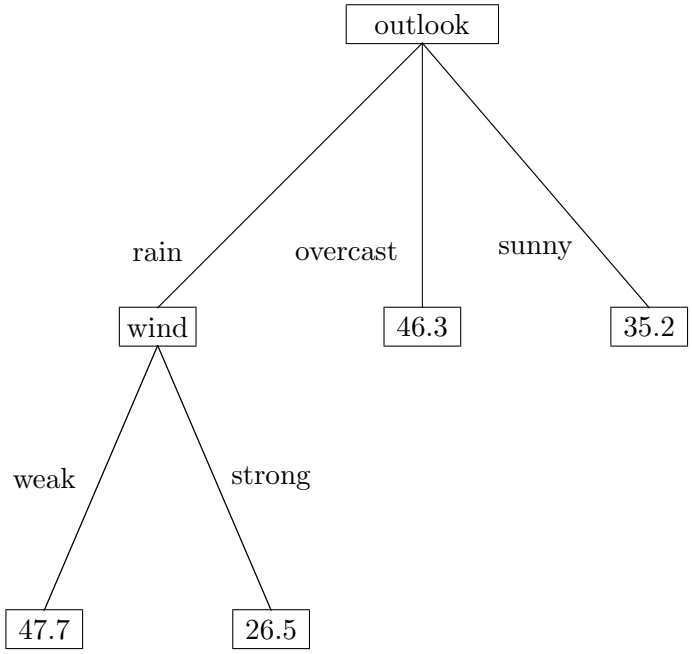


Figure 5.7: The regression tree based on the "outlook" and "wind" variables.

assess the predictive power of the tree.

day	outlook	wind	play time actual	play time predicted	deviation
4	rain	weak	45	47.7	2.7
5	rain	weak	52	47.7	4.3
10	rain	weak	46	47.7	1.7
6	rain	strong	23	26.5	3.5
14	rain	strong	30	26.5	3.5
3	overcast	—	46	46.3	0.3
7	overcast	—	43	46.3	3.3
12	overcast	—	52	46.3	5.7
13	overcast	—	44	46.3	2.3
1	sunny	—	25	35.2	10.2
2	sunny	—	30	35.2	5.2
8	sunny	—	35	35.2	0.2
9	sunny	—	38	35.2	2.8
11	sunny	—	48	35.2	12.8

The average of the deviations is 4.2. The uncertainty of the prediction is 4.2. This tree has the largest predictive power among the trees we have

constructed so far.

Chapter 6

Regression using median

6.1 The training data set

The first table summarizes the variables their names and ranges. The second table contains the training data set consisting of date collected on 14 days.

variable name	variable description	variable range
outlook	weather outlook	rain, overcast, sunny
temperature	air temperature	cool, mild, hot
humidity	air humidity	normal, high
wind	wind strength	weak, strong
play time	playing time	$0, 1, 2, \ldots$

day	outlook	temperature	humidity	wind	play time
1	sunny	hot	high	weak	25
2	sunny	hot	high	strong	30
3	overcast	hot	high	weak	46
4	rain	mild	high	weak	45
5	rain	cool	normal	weak	52
6	rain	cool	normal	strong	23
7	overcast	cool	normal	strong	43
8	sunny	mild	high	weak	35
9	sunny	cool	normal	weak	38
10	rain	mild	normal	weak	46
11	sunny	mild	normal	strong	48
12	overcast	mild	high	strong	52
13	overcast	hot	normal	weak	44
14	rain	mild	high	strong	30

The "play time" variable is the target variable. The variables "outlook",

"temperature", "humidity", "wind" are the explanatory variables. We want to predict the value of "play time" using the explanatory variables.

6.2 No explanatory variable is used for prediction

When we do not use any of the explanatory variables the prediction for the "play time" is the median of the "play time" values. The median is equal to $(43 + 44)/2 = 43.5$. We compute the median of the deviations. This can be interpreted as an uncertainty of the prediction. The smaller is the uncertainty the more information the explanatory variable provides about the target variable. Stating it differently, the smaller is the uncertainty the larger is the predictive power of the explanatory variable.

day	play time	ordered time
1	25	23
2	30	25
3	46	30
4	45	30
5	52	35
6	23	38
7	43	43
8	35	44
9	38	45
10	46	46
11	48	46
12	52	48
13	44	52
14	30	52

The median of the "play time" is equal to $(43+44)/2 = 43.4$. Consequently, the prediction for the "play time" is equal to 43.5.

day	play time actual	play time predicted	deviation	ordered deviation
1	25	43.5	18.5	0.5
2	30	43.5	13.5	0.5
3	46	43.5	2.5	1.5
4	45	43.5	1.5	2.5
5	52	43.5	8.5	2.5
6	23	43.5	20.5	4.5
7	43	43.5	0.5	5.5
8	35	43.5	8.5	8.5
9	38	43.5	5.5	8.5
10	46	43.5	2.5	8.5
11	48	43.5	4.5	13.5
12	52	43.5	8.5	13.5
13	44	43.5	0.5	18.5
14	30	43.5	13.5	20.5

The median of the deviation is equal to $(5.5 + 8.5)/2 = 7$. Therefore, the measure of the uncertainty of the prediction is 7.

We summarize the uncertainties associated with the predictions of the regression trees based on the variables "outlook", "temperature", "humidity", "wind". These are the values we intend to compute in the next sections.

variable	uncertainty
none	7.00
outlook	4.00
temperature	7.00
humidity	5.50
wind	6.50

6.3 The "outlook" variable predicts the "play time"

We will use the "outlook" variable to predict the tennis "play time".

day	outlook	play time	ordered time
4	rain	45	23
5	rain	52	30
6	rain	23	45
10	rain	46	46
14	rain	30	52
3	overcast	46	43
7	overcast	43	44
12	overcast	52	46
13	overcast	44	52
1	sunny	25	25
2	sunny	30	30
8	sunny	35	35
9	sunny	38	38
11	sunny	48	48

The prediction rule is rain→45, overcast→45, sunny→35. The uncertainties of these predictions are 7, 1.5, 5. The weighted average of the uncertainties

$$\begin{aligned} U &= (5/14)(7) + (4/14)(1.5) + (5/14)(5) \\ &= (35/14) + (6/14) + (25/14) \\ &= (66/14) = 4.7. \end{aligned}$$

In some algorithm it is used to estimate the uncertainty of the prediction. We do not see any point to use this estimate of the uncertainty and we are not going to use this number in our arguments. We included the weighted

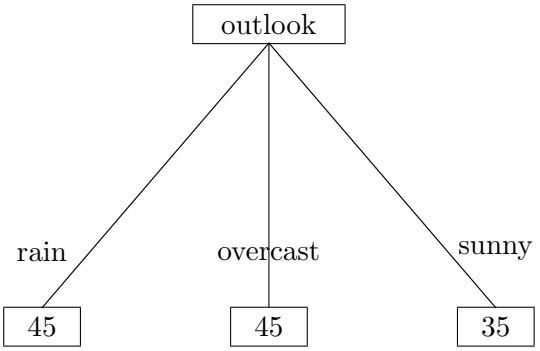


Figure 6.1: The regression tree based solely on the "outlook" variable.

average since we do not wish to divorce from the living practice.

day	outlook	play time actual	play time predicted	deviation	ordered deviation
4	rain	45	45	0	0
5	rain	52	45	7	0
6	rain	23	45	22	1
10	rain	46	45	1	1
14	rain	30	45	15	1
3	overcast	46	45	1	2
7	overcast	43	45	2	3
12	overcast	52	45	7	5
13	overcast	44	45	1	7
1	sunny	25	35	10	7
2	sunny	30	35	5	10
8	sunny	35	35	0	13
9	sunny	38	35	3	15
11	sunny	48	35	13	22

The mean of the deviation is $(3 + 5)/2 = 4$, that is, the uncertainty of the prediction is 4. Figure 6.1 below illustrates the regression tree based solely on the "outlook" variable.

6.4 The "temperature" variable predicts the "play time"

We will use the "temperature" variable to predict the values of the "play time" variable.

day	temperature	play time	ordered time
5	cool	52	23
6	cool	23	38
7	cool	43	43
9	cool	38	52
4	mild	45	30
8	mild	35	35
10	mild	46	45
11	mild	48	46
12	mild	52	48
14	mild	30	52
1	hot	25	25
2	hot	30	30
3	hot	46	44
13	hot	44	46

The prediction rule is cool→40.5, mild→45.5, hot→37. The uncertainties of these predictions are 7, 4.5, 8. The weighted average of the uncertainties is

$$\begin{aligned} U &= (4/14)(7) + (6/14)(4.5) + (4/14)(8) \\ &= (28/14) + (27/14) + (32/14) \\ &= (87/14) = 6.2. \end{aligned}$$

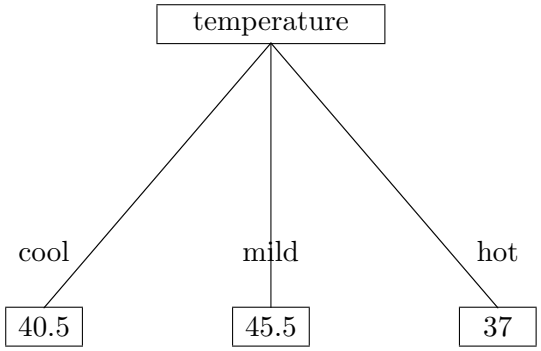


Figure 6.2: The regression tree based solely on the "temperature" variable.

This is an estimate of the uncertainty of the prediction.

day	temperature	play time actual	play time predicted	deviation	ordered deviation
5	cool	52	40.5	11.5	0.5
6	cool	23	40.5	17.5	0.5
7	cool	43	40.5	2.5	2.5
9	cool	38	40.5	2.5	2.5
4	mild	45	45.5	0.5	2.5
8	mild	35	45.5	10.5	6.5
10	mild	46	45.5	0.5	7.0
11	mild	48	45.5	2.5	7.0
12	mild	52	45.5	6.5	9.0
14	mild	30	45.5	15.5	10.5
1	hot	25	37.0	12.0	11.5
2	hot	30	37.0	7.0	12.0
3	hot	46	37.0	9.0	15.5
13	hot	44	37.0	7.0	17.5

The medium of the deviations is $(7 + 7)/2 = 7$ and so the uncertainty is 7. The end result of our work is the regression tree based solely on the "temperature" variable. This thee is depicted in Figure **??**.

6.5 The "humidity" variable predicts the "play time"

Next we turn to predict the values of the "play time" using the values of the "humidity" variable.

day	humidity	play time	ordered time
5	normal	52	23
6	normal	23	38
7	normal	43	43
9	normal	38	44
10	normal	46	46
11	normal	48	48
13	normal	44	52
1	high	25	25
2	high	30	30
3	high	46	30
4	high	45	35
8	high	35	45
12	high	52	46
14	high	30	52

The prediction rule is normal→44, high→35. The uncertainties associated with these predictions are 4, 10. The weighted average of the uncertainties is

$$\begin{aligned} U &= (7/14)(4) + (7/14)(10) \\ &= (1/2)(4) + (1/2) = 7. \end{aligned}$$

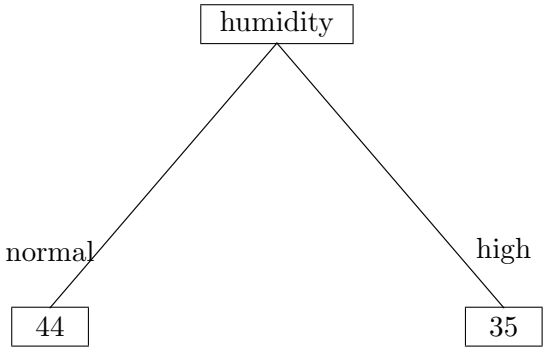


Figure 6.3: The regression tree based solely on the "humidity" variable.

day	humidity	play time actual	play time predicted	deviation	ordered deviation
5	normal	52	44	8	0
6	normal	23	44	21	0
7	normal	43	44	1	1
9	normal	38	44	6	2
10	normal	46	44	2	4
11	normal	48	44	4	5
13	normal	44	44	0	5
1	high	25	35	10	6
2	high	30	35	5	8
3	high	46	35	11	10
4	high	45	35	10	10
8	high	35	35	0	11
12	high	52	35	17	17
14	high	30	35	5	21

The median of the deviations is $(5 + 6)/2 = 5.5$ and so the uncertainty of the prediction is 5.5. (This example shows that the weighted average of the partial uncertainties are not estimating the uncertainty well.) We draw a picture of the regression tree based solely on the "humidity" variable. It is depicted in Figure 6.3.

6.6 The "wind" variable predicts the "play time"

Finally we will predict the values of the "play time" using the values of the "wind" variable.

day	wind	play time	ordered time
1	weak	25	25
3	weak	46	35
4	weak	45	38
5	weak	52	44
8	weak	35	45
9	weak	38	46
10	weak	46	46
13	weak	44	52
2	strong	30	23
6	strong	23	30
7	strong	43	30
11	strong	48	43
12	strong	52	48
14	strong	30	52

The prediction rules are weak→44.5, strong→36.5. The uncertainties of these predictions are 4, 9. Their weighted average is

$$\begin{aligned} U &= (8/14)(4) + (6/14)(9) \\ &= (32/14) + (54/14) \\ &= (86/14) = 6.2. \end{aligned}$$

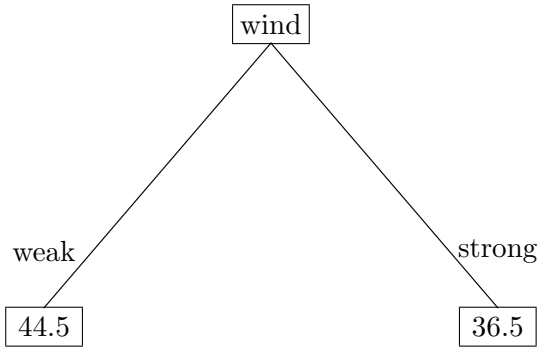


Figure 6.4: The regression tree based on the "wind" variable alone.

day	wind	play time actual	play time predicted	deviation	ordered deviation
1	weak	25	44.5	19.5	0.5
3	weak	46	44.5	1.5	0.5
4	weak	45	44.5	0.5	1.5
5	weak	52	44.5	7.5	1.5
8	weak	35	44.5	9.5	6.5
9	weak	38	44.5	6.5	6.5
10	weak	46	44.5	1.5	6.5
13	weak	44	44.5	0.5	6.5
2	strong	30	36.5	6.5	7.5
6	strong	23	36.5	13.5	9.5
7	strong	43	36.5	6.5	11.5
11	strong	48	36.5	11.5	13.5
12	strong	52	36.5	15.5	15.5
14	strong	30	36.5	6.5	19.5

The median of the deviations is $(6.5+6.5)/2 = 6.5$ and so the uncertainty of the prediction is 6.5. We may summarize the result of our considerations by presenting the regression tree based on the "wind" variable alone in Figure 6.4.

6.7 The ensemble method

So far we have constructed four regression trees to predict the "play time" variable. We do not count the tree when none of the explanatory variables was used. Our computation shows that the explanatory variable "outlook" carries the most information about the "play time". Among the remaining explanatory variables the "humidity" is the best explanatory variable.

Let us form committees of consisting of some of the regression trees we already have and let us see if together they perform better than they perform individually. We use the regression trees based on the "outlook" and "humidity" explanatory variables.

day	outlook	humidity	play time actual	play time predicted	deviation	ordered deviation
1	sunny	high	25	35.0	10.0	0.0
2	sunny	high	30	35.0	5.0	0.5
3	overcast	high	46	40.0	6.0	1.5
4	rain	high	45	40.0	5.0	1.5
5	rain	normal	52	44.5	7.5	1.5
6	rain	normal	23	44.5	21.5	5.0
7	overcast	normal	43	44.5	1.5	5.0
8	sunny	high	35	35.0	0.0	6.0
9	sunny	normal	38	39.5	1.5	7.5
10	rain	normal	46	44.5	1.5	8.5
11	sunny	normal	48	39.5	8.5	10.0
12	overcast	high	52	40.0	12.0	10.0
13	overcast	normal	44	44.5	0.5	12.0
14	rain	high	30	40.0	10.0	21.5

The median of the deviations is (5+6)/2=5.5. The uncertainty of the prediction of the committee is 5.5. This is not better than the prediction using the explanatory variable "outlook" alone. In other words, the committee is not over performing the existing trees.

6.8 "Outlook" takes on the value "rain"

We are extending the tree which predicts the "play time" based on the "outlook" variable. There are 5 days when the "outlook" variable takes on the value "rain".

day	outlook	temperature	humidity	wind	play time
4	rain	mild	high	weak	45
5	rain	cool	normal	weak	52
6	rain	cool	normal	strong	23
10	rain	mild	normal	weak	46
14	rain	mild	high	strong	30

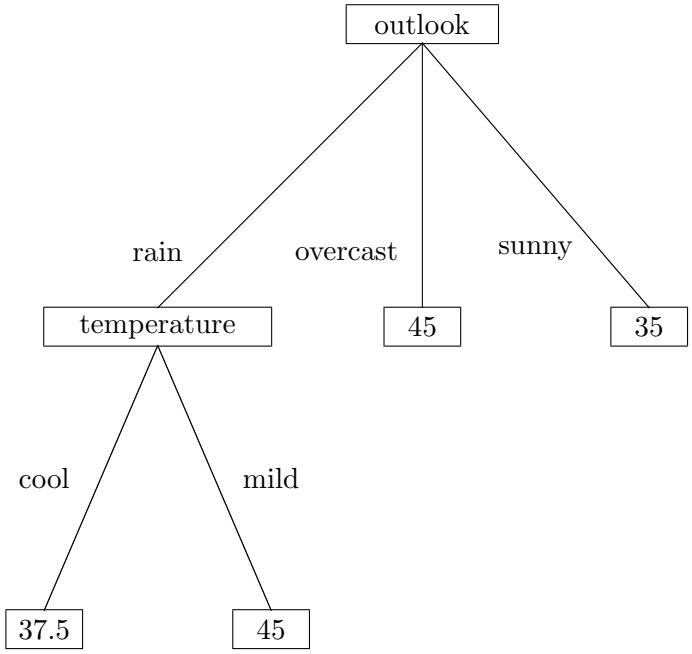


Figure 6.5: The regression tree based on the "outlook" and "temperature" variables.

We start with using the "temperature" variable to predict "play time".

day	temperature	play time	ordered play time
5	cool	52	23
6	cool	23	52
4	mild	45	30
10	mild	46	45
14	mild	30	46

The prediction rule is cool→37.5, mild→45. We assess the predictive power

of this decision tree by computing the uncertainty of the prediction.

day	outlook	tempe-rature	play time actual	play time predicted	deviation	ordered deviation
6	rain	cool	23	37.5	14.5	0.0
5	rain	cool	52	37.5	14.5	0.0
14	rain	mild	30	45.0	15.0	1.0
4	rain	mild	45	45.0	0.0	1.0
10	rain	mild	46	45.0	1.0	1.0
7	overcast	—	43	45.0	2.0	2.0
13	overcast	—	44	45.0	1.0	3.0
3	overcast	—	46	45.0	1.0	5.0
12	overcast	—	52	45.0	7.0	7.0
1	sunny	—	25	35.0	10.0	10.0
2	sunny	—	30	35.0	5.0	13.0
8	sunny	—	35	35.0	0.0	14.5
9	sunny	—	38	35.0	3.0	14.5
11	sunny	—	48	35.0	13.0	15.0

The median of the deviation is (3+5)/2=4. Thus the uncertainty of the prediction is 4. This is not better than the prediction bases solely on the "outlook" variable. In fact, this larger is not preferable to the smaller tree as we estimated 3 parameters in the small tree and we have estimated 5 parameters in the larger tree.

Next we use the "humidity" to predict "play time".

day	humidity	play time	ordered play time
5	normal	52	23
6	normal	23	46
10	normal	46	52
4	high	45	30
14	high	30	45

The prediction rule is normal→46, high→37.5. We compute the uncertainty of the prediction of this tree. In this way we can assess the predictive power

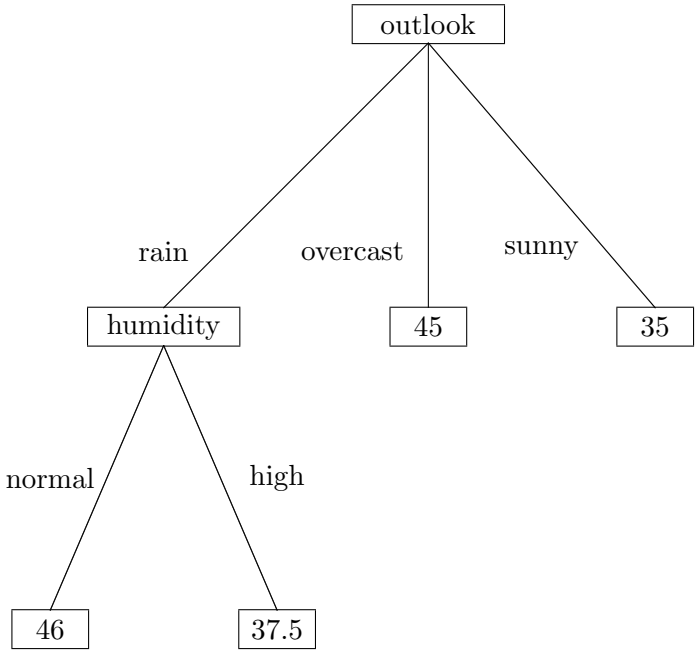


Figure 6.6: The regression tree based on the "outlook" and "humidity" variables.

of the tree.

day	outlook	humidity	play time actual	play time predicted	deviation	ordered deviation
6	rain	normal	23	46.0	23.0	0.0
10	rain	normal	46	46.0	0.0	0.0
5	rain	normal	52	46.0	6.0	1.0
14	rain	high	30	37.5	7.5	1.0
4	rain	high	45	37.5	7.5	2.0
7	overcast	—	43	45.0	2.0	3.0
13	overcast	—	44	45.0	1.0	5.0
3	overcast	—	46	45.0	1.0	6.0
12	overcast	—	52	45.0	7.0	7.0
1	sunny	—	25	35.0	10.0	7.5
2	sunny	—	30	35.0	5.0	7.5
8	sunny	—	35	35.0	0.0	10.0
9	sunny	—	38	35.0	3.0	13.0
11	sunny	—	48	35.0	13.0	13.0

The median of the deviations is (5+6)/2=5.5. This is the uncertainty of the

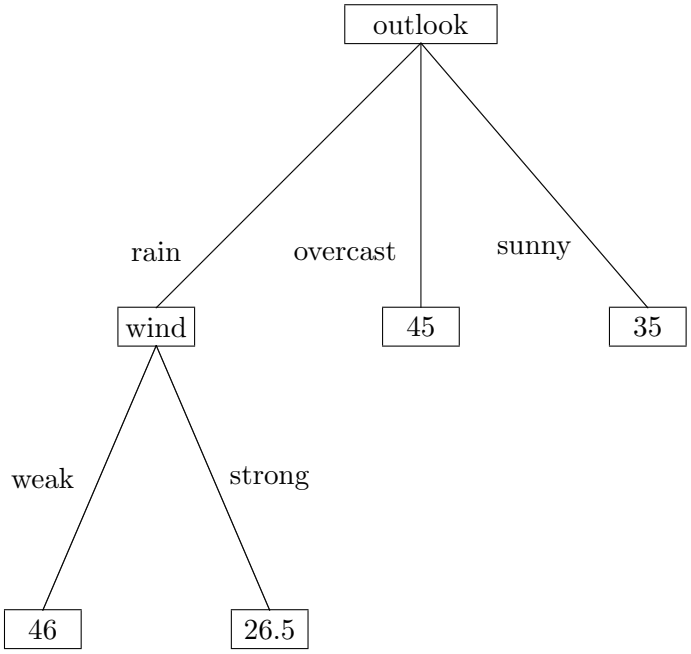


Figure 6.7: The regression tree based on the "outlook" and "wind" variables.

prediction of the tree. It does not compare favorably to the earlier trees. Therefore, we drop this tree.

Finally, we use the "wind" to predict "play time".

day	wind	play time	ordered play time
4	weak	45	45
5	weak	52	46
10	weak	46	52
6	strong	23	23
14	strong	30	30

The prediction rule is weak→46, strong→26.5. We compute the uncertainty of the prediction of this tree. In this way we can assess the predictive power

of the tree.

day	outlook	wind	play time actual	play time predicted	deviation	ordered deviation
4	rain	weak	45	46.0	1.0	0.0
10	rain	weak	46	46.0	0.0	0.0
5	rain	weak	52	46.0	6.0	1.0
6	rain	strong	23	26.5	3.5	1.0
14	rain	strong	30	26.5	3.5	1.0
7	overcast	—	43	45.0	2.0	2.0
13	overcast	—	44	45.0	1.0	3.0
3	overcast	—	46	45.0	1.0	3.5
12	overcast	—	52	45.0	7.0	3.5
1	sunny	—	25	35.0	10.0	5.0
2	sunny	—	30	35.0	5.0	6.0
8	sunny	—	35	35.0	0.0	7.0
9	sunny	—	38	35.0	3.0	10.0
11	sunny	—	48	35.0	13.0	23.0

The median of the deviations is (3+3.5)/2=3.25. The uncertainty of the prediction is 3.25. This tree has the largest predictive power among the trees we have constructed so far.

Chapter 7

Midrange based regression

7.1 The training data set

The first table summarizes the variables their names and ranges. The second table contains the training data set consisting of date collected on 14 days.

variable name	variable description	variable range
outlook	weather outlook	rain, overcast, sunny
temperature	air temperature	cool, mild, hot
humidity	air humidity	normal, high
wind	wind strength	weak, strong
play time	playing time	$0, 1, 2, \ldots$

day	outlook	temperature	humidity	wind	play time
1	sunny	hot	high	weak	25
2	sunny	hot	high	strong	30
3	overcast	hot	high	weak	46
4	rain	mild	high	weak	45
5	rain	cool	normal	weak	52
6	rain	cool	normal	strong	23
7	overcast	cool	normal	strong	43
8	sunny	mild	high	weak	35
9	sunny	cool	normal	weak	38
10	rain	mild	normal	weak	46
11	sunny	mild	normal	strong	48
12	overcast	mild	high	strong	52
13	overcast	hot	normal	weak	44
14	rain	mild	high	strong	30

The "play time" variable is the target variable. The variables "outlook", "temperature", "humidity", "wind" are the explanatory variables. We want to predict the value of "play time" using the explanatory variables.

7.2 No explanatory variable is used for prediction

When we do not use any of the explanatory variables the prediction for the "play time" is the midrange of the "play time" values. The midrange is equal to $(23 + 52)/2 = 37.5$. The midrange of the deviations is equal to $(0.5+14.5)/2 = 7.5$. This can be interpreted as an uncertainty of the prediction. The smaller is the uncertainty the more information the explanatory variables provide about the target variable. Stating it differently, the smaller is the uncertainty the larger is the predictive power of the explanatory variable.

day	play time actual	play time predicted	deviation
1	25	37.5	12.5
2	30	37.5	7.5
3	46	37.5	8.5
4	45	37.5	7.5
5	52	37.5	14.5
6	23	37.5	14.5
7	43	37.5	5.5
8	35	37.5	2.5
9	38	37.5	0.5
10	46	37.5	8.5
11	48	37.5	10.5
12	52	37.5	14.5
13	44	37.5	6.5
14	30	37.5	7.5

We summarize the uncertainties associated with the predictions of the regression trees based on the variables "outlook", "temperature", "humidity", "wind". These are the values we intend to compute in the next sections.

variable	uncertainty
none	7.5
outlook	7.5
temperature	7.5
humidity	7.5
wind	7.5

7.3 The "outlook" variable predicts the "play time"

We will use the "outlook" variable to predict the tennis "play time".

The prediction rule is rain→37.5, overcast→47.5, sunny→36.5. The midrange of the deviations is (1.5+14.5)/2=8. The corresponding uncertainties are 7.5, 7.5, 8. In other words, the uncertainty of the prediction is 7.5. We compute the weighted average of the uncertainties

$$\begin{aligned} U &= (4/14)(7.5) + (6/14)(7.5) + (4/14)(8) \\ &= (30/14) + (45/14) + (32/14) \\ &= (107/14) = 7.6. \end{aligned}$$

This may serve as an estimate of the uncertainty of the prediction.

day	outlook	play time actual	play time predicted	deviation
4	rain	45	37.5	7.5
5	rain	52	37.5	14.5
6	rain	23	37.5	14.5
10	rain	46	37.5	8.5
14	rain	30	37.5	7.5
3	overcast	46	47.5	1.5
7	overcast	43	47.5	4.5
12	overcast	52	47.5	4.5
13	overcast	44	47.5	3.5
1	sunny	25	36.5	11.5
2	sunny	30	36.5	6.5
8	sunny	35	36.5	1.5
9	sunny	38	36.5	1.5
11	sunny	48	36.5	8.5

Figure 7.1 below illustrates the regression tree based solely on the "outlook" variable.

7.4 The "temperature" variable predicts the "play time"

We will use the "temperature" variable to predict the values of the "play time" variable. The prediction rule is cool→37.5, mild→41, hot→35.5. The uncertainties of these uncertainties are 7.5, 7.5, 8. The uncertainty of the prediction is 7.5. We compute the weighted average of the uncertainties

$$U = (4/14)(7.5) + (6/14)(7.5) + (4/14)(8)$$

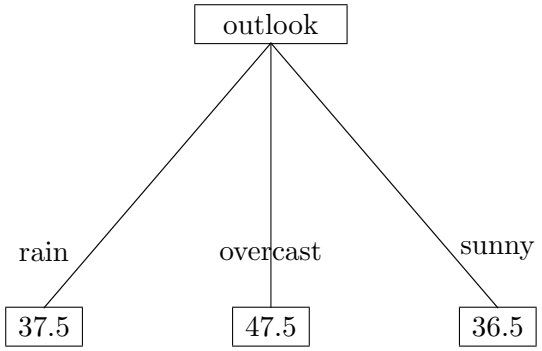


Figure 7.1: The regression tree based solely on the "outlook" variable.

$$= (30/14) + (45/14) + (32/14)$$
$$= (107/14) = 7.6.$$

day	temperature	play time actual	play time predicted	deviation
5	cool	52	37.5	14.5
6	cool	23	37.5	14.5
7	cool	43	37.5	5.5
9	cool	38	37.5	0.5
4	mild	45	41.0	4.0
8	mild	35	41.0	6.0
10	mild	46	41.0	5.0
11	mild	48	41.0	7.0
12	mild	52	41.0	11.0
14	mild	30	41.0	11.0
1	hot	25	35.5	10.5
2	hot	30	35.5	5.5
3	hot	46	35.5	10.5
13	hot	44	35.5	8.5

The end result of our work is the regression tree based solely on the "temperature" variable. This thee is depicted in Figure 7.2.

7.5 The "humidity" variable predicts the "play time"

Next we turn to predict the values of the "play time" using the values of the "humidity" variable. The prediction rule is normal→37.5, high→38.5.

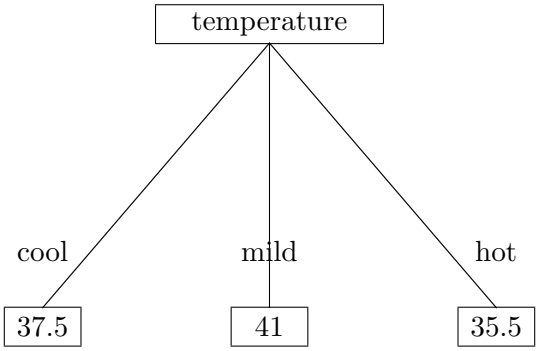


Figure 7.2: The regression tree based solely on the "temperature" variable.

The uncertainties of these predictions are 7.5, 8.5. The midrange of the deviations is 7.5 and so the uncertainty of the prediction is 7.5. We compute the weighted average of the uncertainties

$$\begin{aligned} U &= (7/14)(7.5) + (7/14)(8.5) \\ &= (1/2)(7.5) + (1/2)(8.5) = 8. \end{aligned}$$

day	humidity	play time actual	play time predicted	deviation
5	normal	52	37.5	14.5
6	normal	23	37.5	14.5
7	normal	43	37.5	5.5
9	normal	38	37.5	0.5
10	normal	46	37.5	1.5
11	normal	48	37.5	10.5
13	normal	44	37.5	6.5
1	high	25	38.5	13.5
2	high	30	38.5	8.5
3	high	46	38.5	7.5
4	high	45	38.5	6.5
8	high	35	38.5	3.5
12	high	52	38.5	13.5
14	high	30	38.5	8.5

We draw a picture of the regression tree based solely on the "humidity" variable. See Figure 7.3.

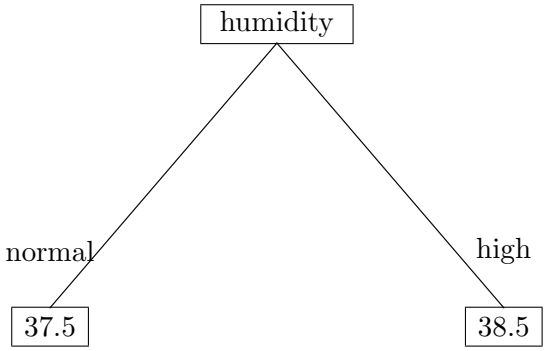


Figure 7.3: The regression tree based solely on the "humidity" variable.

7.6 The "wind" variable predicts the "play time"

Finally we will predict the values of the "play time" using the values of the "wind" variable. The predictions rules are weak→38.5, strong→37.5. The uncertainties of these predictions are 7, 10. The uncertainty of the prediction is 7.5. The weighted average of the uncertainties

$$\begin{aligned} U &= (8/14)(7) + (6/14)(10) \\ &= (56/14) + (60/14) \\ &= (116/14) = 7.6. \end{aligned}$$

day	wind	play time actual	play time predicted	deviation
1	weak	25	38.5	13.5
3	weak	46	38.5	7.5
4	weak	45	38.5	6.5
5	weak	52	38.5	13.5
8	weak	35	38.5	3.5
9	weak	38	38.5	0.5
10	weak	46	38.5	7.5
13	weak	44	38.5	5.5
2	strong	30	37.5	7.5
6	strong	23	37.5	14.5
7	strong	43	37.5	5.5
11	strong	48	37.5	10.5
12	strong	52	37.5	14.5
14	strong	30	37.5	7.5

We may summarize the result of our considerations by presenting the regression tree based on the "wind" variable alone. See Figure 7.4.

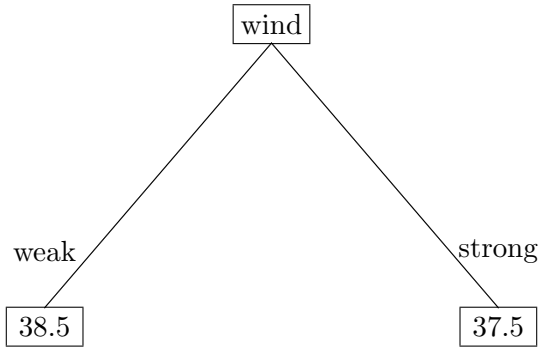


Figure 7.4: The regression tree based on the "wind" variable alone.

7.7 The ensemble method

So far we have constructed four regression trees to predict the "play time" variable. We do not count the tree when none of the explanatory variable was used. Our computation shows that the explanatory variable "outlook" carries the most information about the "play time". Among the remaining explanatory variables the "humidity" is the best explanatory variable.

Let us form committees of consisting of some of the regression trees we already have and see if they perform together better than individually. We use the regression trees bases on the "outlook" and "humidity" explanatory variables.

day	outlook	humidity	play time actual	play time predicted	deviation
1	sunny	high	25	37.5	12.5
2	sunny	high	30	37.5	7.5
3	overcast	high	46.0	43.0	3.0
4	rain	high	45.0	38.0	7.0
5	rain	normal	52	37.5	14.5
6	rain	normal	23	37.5	14.5
7	overcast	normal	43	42.5	0.5
8	sunny	high	35	37.5	2.5
9	sunny	normal	38	37.0	1.0
10	rain	normal	46	37.5	8.5
11	sunny	normal	48	37.0	11.0
12	overcast	high	52	43.0	9.0
13	overcast	normal	44	42.5	1.5
14	rain	high	30	38.0	8 .0

The midrange of the deviations is 7.5. The uncertainty of the prediction of

the committee is 7.5. This is not an improvement over the prediction using only the explanatory variable "outlook".

7.8 "Outlook" takes on the value "rain"

We are extending the tree which predicts the "play time" based on the "outlook" variable. There are 5 days when the "outlook" variable takes on the value "rain".

day	outlook	temperature	humidity	wind	play time
4	rain	mild	high	weak	45
5	rain	cool	normal	weak	52
6	rain	cool	normal	strong	23
10	rain	mild	normal	weak	46
14	rain	mild	high	strong	30

We start with using the "temperature" variable to predict "play time".

day	temperature	play time
5	cool	52
6	cool	23
4	mild	45
10	mild	46
14	mild	30

The prediction rule is cool→37.5, mild→38. We assess the predictive power

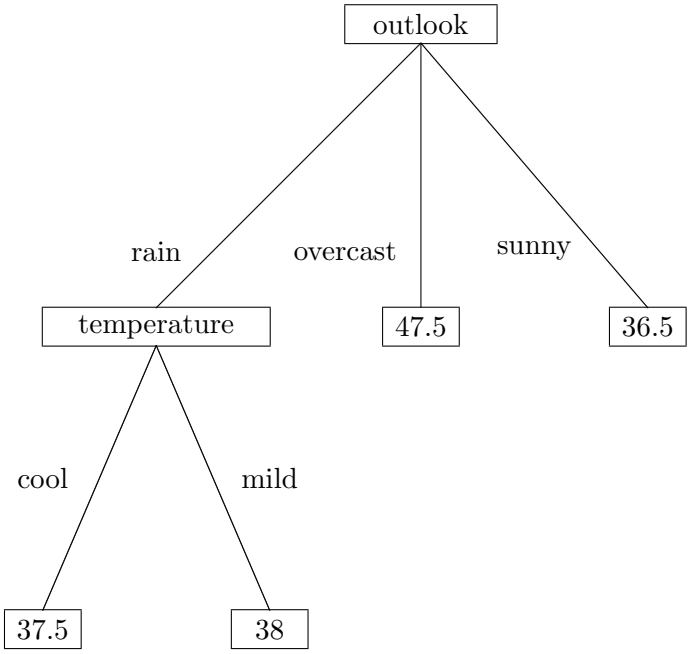


Figure 7.5: The regression tree based on the "outlook" and "temperature" variables.

of this decision tree by computing the uncertainty of the prediction.

day	outlook	tempe-rature	play time actual	play time predicted	deviation
5	rain	cool	52	37.5	14.5
6	rain	cool	23	37.5	14.5
4	rain	mild	45	38.0	7.0
10	rain	mild	46	38.0	8.0
14	rain	mild	30	38.0	8.0
3	overcast	—	46	47.5	1.5
7	overcast	—	43	47.5	4.5
12	overcast	—	52	47.5	4.5
13	overcast	—	44	47.5	3.5
1	sunny	—	25	36.5	11.5
2	sunny	—	30	36.5	6.5
8	sunny	—	35	36.5	1.5
9	sunny	—	38	36.5	1.5
11	sunny	—	48	36.5	8.5

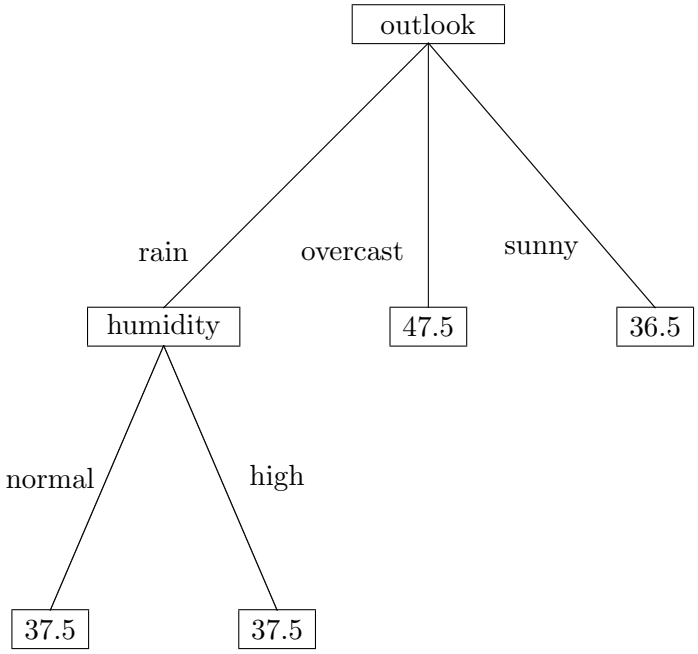


Figure 7.6: The regression tree based on the "outlook" and "humidity" variables.

The midrange of the deviations is (1.5+14.5)/2=8. Thus the uncertainty of the prediction is 8. This is not better than the prediction bases solely on the "outlook" variable.

Next we use the "humidity" to predict "play time".

day	humidity	play time
5	normal	52
6	normal	23
10	normal	46
4	high	45
14	high	30

The prediction rule is normal→37.5, high→37.5. We compute the uncertainty of the prediction of this tree. In this way we can assess the predictive

power of the tree.

day	outlook	humidity	play time actual	play time predicted	deviation
4	rain	high	45	37.5	7.5
14	rain	high	30	37.5	7.5
5	rain	normal	52	37.5	14.5
6	rain	normal	23	37.5	14.5
10	rain	normal	46	37.5	8.5
3	overcast		46	47.5	1.5
7	overcast		43	47.5	4.5
12	overcast		52	47.5	4.5
13	overcast		44	47.5	3.5
1	sunny		25	36.5	11.5
2	sunny		30	36.5	6.5
8	sunny		35	36.5	1.5
9	sunny		38	36.5	1.5
11	sunny		48	36.5	8.5

The midrange of the deviations is (1.5+14.5)/2=8. This is the uncertainty of the prediction of the tree. It does not compare favorably to the earlier trees. Therefore, we drop this tree.

Finally, we use the "wind" to predict "play time".

day	wind	play time
4	weak	45
5	weak	52
10	weak	46
6	strong	23
14	strong	30

The prediction rule is weak→48.5, strong→26.5. We compute the uncertainty of the prediction of this tree. In this way we can assess the predictive

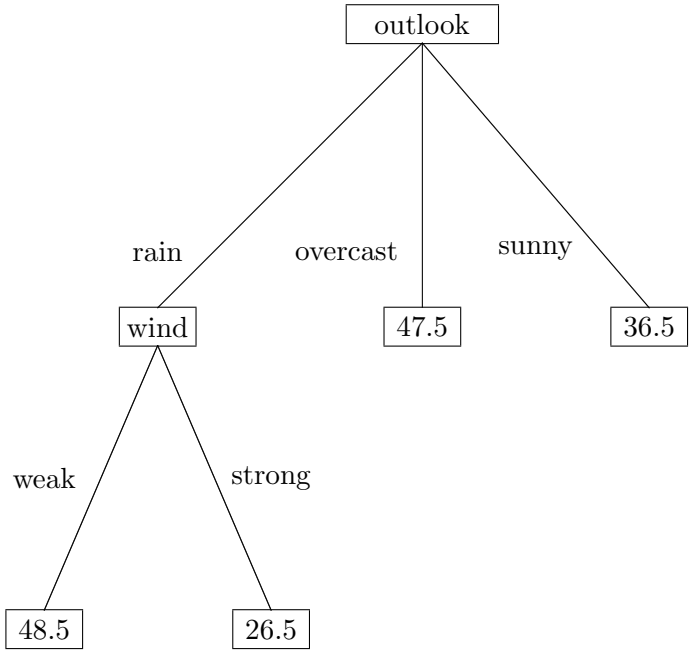


Figure 7.7: The regression tree based on the "outlook" and "wind" variables.

power of the tree.

day	outlook	wind	play time actual	play time predicted	deviation
4	rain	weak	45	48.5	3.5
5	rain	weak	52	48.5	3.5
10	rain	weak	46	48.5	2.5
6	rain	strong	23	26.5	3.5
14	rain	strong	30	26.5	3.5
3	overcast	—	46	47.5	1.5
7	overcast	—	43	47.5	4.5
12	overcast	—	52	47.5	4.5
13	overcast	—	44	47.5	3.5
1	sunny	—	25	36.5	11.5
2	sunny	—	30	36.5	6.5
8	sunny	—	35	36.5	1.5
9	sunny	—	38	36.5	1.5
11	sunny	—	48	36.5	8.5

The midrange of the deviations is (1.5+11.5)/2=6.5. The uncertainty of the prediction is 6.5. This tree has the largest predictive power among the trees

we have constructed so far.

7.9 Comparing regression trees

In Chapter 4 we used variance to construct regression trees. The tree T_{var} with the best predictive power is in Figure 4.7. In Chapter 5 we used averages to construct regression trees. The tree T_{aver} with the best predictive power is in Figure 5.7. In Chapter 6 we used medians to construct regression trees. The tree T_{med} with the best predictive power is in Figure 6.7. In Chapter 7 we used midranges to construct regression trees. The tree T_{midr} with the best predictive power is in Figure 7.7.

day	deviation T_{var}	deviation T_{aver}	deviation T_{med}	deviation T_{midr}
1	10.2	10.2	10	11.5
2	5.2	5.2	5	6.5
3	0.2	0.3	1	1.5
4	2.7	2.7	1	3.5
5	4.3	4.3	6	3.5
6	3.5	3.5	3.5	3.5
7	3.2	3.3	2	4.5
8	0.2	0.2	0	1.5
9	2.8	2.8	3	1.5
10	1.7	1.7	0	2.5
11	12.8	12.8	13	8.5
12	5.8	5.7	7	4.5
13	2.2	2.3	1	3.5
14	3.5	3.5	3.5	3.5

The models provided by the trees do not exactly fit to training set. The deviations between the predictions and the actual values are called residues in the statistical literature. We will inspect these residues.

The first thing we notice is the columns of T_{var} and T_{aver} are almost identical. In fact, they supposed to be identical. The reason of the discrepancies on days 7, 12, 13 is that we do not followed the rules of rounding the last digits properly.

The largest deviation reaches it minimum in the last colum. If from some reason the we wish to avoid large deviation between the predicted and actual values, the we should use the midrange based regression tree.

Chapter 8

Classification using error rate

8.1 The training data set

We have seen that the ideas developed to construct decision trees for classification can be used to construct decision trees for regression. The converse is true if one is able to work with regression trees, then she or he is able to convert this knowledge to work with classification trees. It will shed new light to the classification problems if we view them as regression problems. This will be the subject of this chapters.

As usual we start with our favorite toy data set. We repeat the data set since we would like to keep the chapters as independent as possible. The first table summarizes the variables their names and ranges. The second table contains the training data set consisting of date collected on 14 days.

variable name	variable description	variable range
outlook	weather outlook	rain, overcast, sunny
temperature	air temperature	cool, mild, hot
humidity	air humidity	normal, high
wind	wind strength	weak, strong
play tennis	playing tennis	no, yes

day	outlook	temperature	humidity	wind	play tennis
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rain	mild	high	weak	yes
5	rain	cool	normal	weak	yes
6	rain	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rain	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rain	mild	high	strong	no

The "play tennis" variable is the target variable. The variables "outlook", "temperature", "humidity", "wind" are the explanatory variables. We want to predict the value of "play tennis" using the explanatory variables.

8.2 Assessing the variables one at a time

We summarize the predictive powers of the variable in the following table. The uncertainty of the prediction gives a hint about the predictive power of the given explanatory variable. Namely, the smaller is the uncertainty, the larger is predictive power. We intend to compute the uncertainty values in this section.

variable	uncertainty
none	0.356
outlook	0.286
temperature	0.356
humidity	0.286
wind	0.356

At first we do not use any of the explanatory variables the prediction for the "play tennis". The prediction rule is ( )→yes. In other words, we predict yes all the time. The average of the deviations is 5/14=0.356. The smaller is the average of the deviation the better is prediction. The average of the

deviations can be interpreted as the uncertainty of the prediction.

day	play tennis actual	play tennis predicted	deviation
1	no	yes	1
2	no	yes	1
3	yes	yes	0
4	yes	yes	0
5	yes	yes	0
6	no	yes	1
7	yes	yes	0
8	no	yes	1
9	yes	yes	0
10	yes	yes	0
11	yes	yes	0
12	yes	yes	0
13	yes	yes	0
14	no	yes	1

We use the "outlook" variable to predict the "play tennis" variable. We rearrange the rows of the training data set by grouping them by the possible values of the "outlook" variable. The predictions rule is rain→yes, overcast→yes, sunny→no. The average of the deviations is 4/14=0.286.

day	outlook	play tennis actual	play tennis predicted	deviation
4	rain	yes	yes	0
5	rain	yes	yes	0
6	rain	no	yes	1
10	rain	yes	yes	0
14	rain	no	yes	1
3	overcast	yes	yes	0
7	overcast	yes	yes	0
12	overcast	yes	yes	0
13	overcast	yes	yes	0
1	sunny	no	no	0
2	sunny	no	no	0
8	sunny	no	no	0
9	sunny	yes	no	1
11	sunny	yes	no	1

Figure 8.1 below is a possible geometric representation of the regression tree based solely on the "outlook" variable.

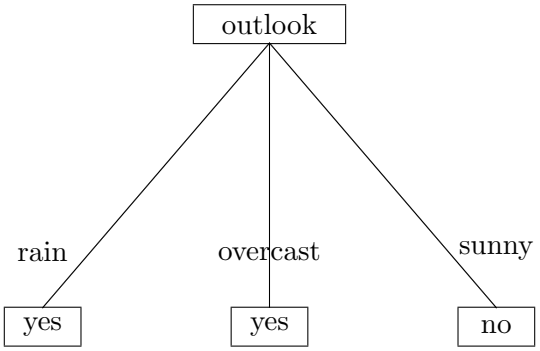


Figure 8.1: The regression tree based solely on the "outlook" variable.

We use the "temperature" explanatory variable to predict the "play tennis" target variable. The predictions rules are cool→yes, mild→yes, hot→no. The average of the deviation is equal to 5/14=0.356.

day	temperature	play tennis actual	play tennis predicted	deviation
5	cool	yes	yes	0
6	cool	no	yes	1
7	cool	yes	yes	0
9	cool	yes	yes	0
4	mild	yes	yes	0
8	mild	no	yes	1
10	mild	yes	yes	0
11	mild	yes	yes	0
12	mild	yes	yes	0
14	mild	no	yes	1
1	hot	no	no	0
2	hot	no	no	0
3	hot	yes	no	1
13	hot	yes	no	1

Figure 8.2 depicts the regression tree bases on the variable "temperature".

We use the "humidity" variable to predict the "play tennis" variable. The prediction rule is normal→yes, high→no. The average of the deviations

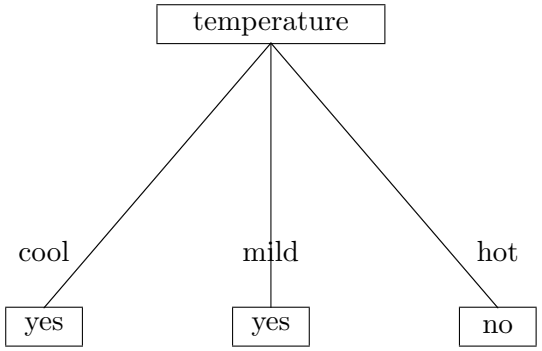


Figure 8.2: The regression tree based solely on the "temperature" variable.

is 4/14=0.286.

day	humidity	play tennis actual	play tennis predicted	deviation
5	normal	yes	yes	0
6	normal	no	yes	1
7	normal	yes	yes	0
9	normal	yes	yes	0
10	normal	yes	yes	0
11	normal	yes	yes	0
13	normal	yes	yes	0
1	high	no	no	0
2	high	no	no	0
3	high	yes	no	1
4	high	yes	no	1
8	high	no	no	0
12	high	yes	no	1
14	high	no	no	0

The result of our consideration is the decision tree based on the variable "humidity" preseted in Figure 8.3.

Finally, we turn to the problem of predicting the "play tennis" target variable using the "wind" explanatory variable. The prediction rules are

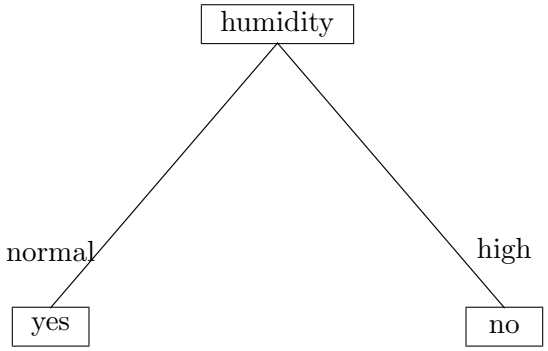


Figure 8.3: The decision tree based solely on the "humidity" variable.

weak→yes, strong→no. The average of the deviations is 5/14=0.356.

day	wind	play tennis actual	play tennis predicted	deviation
1	weak	no	yes	1
3	weak	yes	yes	0
4	weak	yes	yes	0
5	weak	yes	yes	0
8	weak	no	yes	1
9	weak	yes	yes	0
10	weak	yes	yes	0
13	weak	yes	yes	0
2	strong	no	no	0
6	strong	no	no	0
7	strong	yes	no	1
11	strong	yes	no	1
12	strong	yes	no	1
14	strong	no	no	0

We draw the decision tree which predicts the "play tennis" variable using the "wind" variable. The result can be seen in Figure 8.4

8.3 A short cut in the book keeping

The dedicated student may notice that we can gather the ingredients for computing the combined impurities by filling in the contingency table we applied to construct the confusion matrix earlier. In the future we will use this short cut systematically.

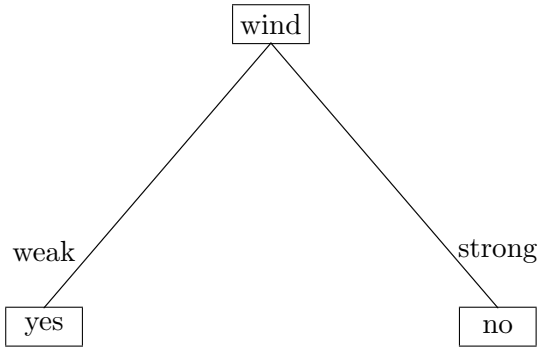


Figure 8.4: The regression tree based on the "wind" variable alone.

The long winged way of writing up the rearranged rows of the training data set helps us to develop a firmer understanding of the details of the techniques we intend to master. It helps us to build up some intuition. In short, it serves as pedagogical tool.

The dedicated student may notice that if we want to solve larger scale problems we can gather the ingredients for computing the combined impurities by filling in the contingency table we applied to construct the confusion matrix earlier. In connection with a larger training data set we would use this short cut systematically. Here we just illustrate a possible way of organizing the work.

The basic observation we should make is the following. The sum of the deviations is equal to the number of the misclassified items. The average of the deviation is equal to misclassification or error rate. To compute the error rate is we can use the information contained by the confusion matrix.

	outlook rain	(yes)	outlook overcast	(yes)	outlook sunny	(no)
play tennis no	6,14	(2)		(0)	1,2,8	(3)
play tennis yes	4,5,10	(3)	3,7,12,13	(4)	9,11	(2)

	temperature cool	(yes)	temperature mild	(yes)	temperature hot	(no)
play tennis no	6	(1)	8,14	(2)	1,2	(2)
play tennis yes	5,7,9	(3)	4,10,11,12	(4)	3,13	(2)

	humidity normal	(yes)	humidity high	(no)
play tennis no	6	(1)	1,2,8,14	(4)
play tennis yes	5,7,9,10,11,13	(6)	3,4,12	(3)

	wind strong	(no)	wind weak	(yes)
play tennis no	2,6,14	(3)	1,8	(2)
play tennis yes	7,11,12	(3)	3,4,5,9,10,13	(6)

8.4 Splitting the training data set

The predictive power of the variable "outlook" is the largest as its associated impurity is the smallest in comparison with the variables "temperature", "humidity", "wind". This variable will be used to add new branches to the decision tree. It takes on three values "rain", "overcast", "sunny". We split the training data set into three smaller data sets.

day	outlook	temperature	humidity	wind	play tennis
4	rain	mild	high	weak	yes
5	rain	cool	normal	weak	yes
6	rain	cool	normal	strong	no
10	rain	mild	normal	weak	yes
14	rain	mild	high	strong	no
3	overcast	hot	high	weak	yes
7	overcast	cool	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
11	sunny	mild	normal	strong	yes

8.5 The "outlook" variable is equal to "rain"

We are working with the following small training data set. We summarize the predictive powers of the variables in the following table. These are the values we intended to compute in this section.

variable	uncertainty
none	0.400
temperature	0.400
humidity	0.400
wind	0.000

We do not use any of the explanatory variables "temperature", "humidity", "wind" to predict the "play tennis" target variable. The prediction rule

is ()→yes. (The prediction is always "yes", that is, the prediction rule is the constant function taking the "yes" value.) The average of the deviation is 2/5=0.400. The average of the deviations of a prediction will be compared against this ground level. If the average of the deviation is not smaller than 0.400, then the new prediction can be sorted out.

day	outlook	play tennis actual	play tennis predicted	deviation
4	rain	yes	yes	0
5	rain	yes	yes	0
6	rain	no	yes	1
10	rain	yes	yes	0
14	rain	no	yes	1

We use the "temperature" explanatory variable to predict the "play tennis" target variable. The prediction rule is cool→no, mild→yes. The average of the deviations is 2/5=0.400.

day	outlook	temperature	play tennis actual	play tennis predicted	deviation
5	rain	cool	yes	no	1
6	rain	cool	no	no	0
10	rain	mild	yes	yes	0
4	rain	mild	yes	yes	0
14	rain	mild	no	yes	1

Figure 8.5 below is the regression tree using the "outlook" and "temperature" explanatory variables to predict the "play tennis" target variable. The predictive power of this tree is not better than the predictive power of the tree without the "temperature" variable.

We use the "humidity" explanatory variable to predict the "play tennis" target variable. The prediction rule is normal→yes, high→no. The average of the deviations is 2/5=0.400.

day	outlook	humidity	play tennis actual	play tennis predicted	deviation
5	rain	normal	yes	yes	0
6	rain	normal	no	yes	1
10	rain	normal	yes	yes	0
4	rain	high	yes	no	1
14	rain	high	no	no	0

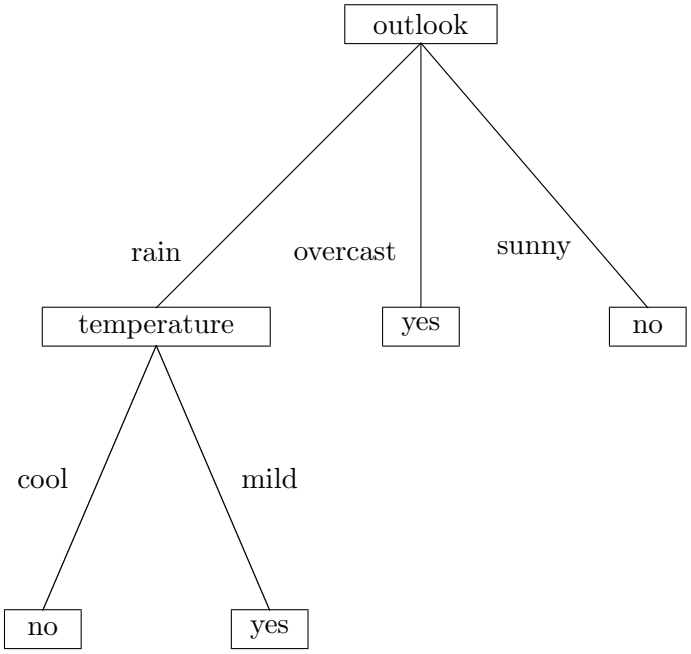


Figure 8.5: The regression tree based on the "outlook" and ´´temperature" variables.

Figure 8.7 is the regression tree using the "outlook" and "humidity" explanatory variables to predict the "play tennis" target variable. The predictive power of this tree is the same as the predictive power of the tree without the "humidity" variable.

We use the "wind" explanatory variable to predict the "play tennis" target variable. The prediction rule is weak→yes, strong→no. The average of the deviations is 0/5=0.000.

day	outlook	wind	play tennis actual	play tennis predicted	deviation
4	rain	weak	yes	yes	0
5	rain	weak	yes	yes	0
10	rain	weak	yes	yes	0
6	rain	strong	no	no	0
14	rain	strong	no	no	0

The following figure is the decision tree using the "outlook" and "wind" explanatory variables to predict the "play tennis" target variable. The predictive power of this tree is better than the predictive power of the tree without the "wind" variable.

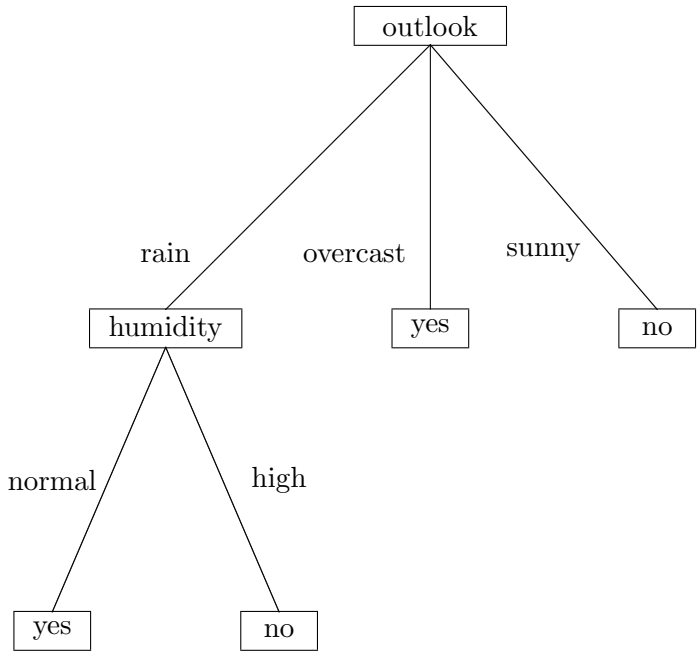
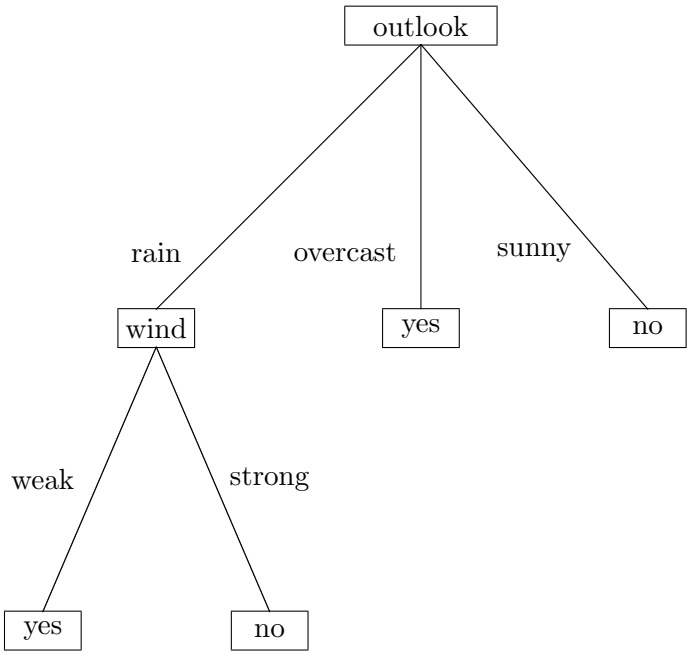


Figure 8.6: The regression tree based on the "outlook" and ´´humidity" variables.



We summarize the predictive powers of the variables in the following table. These are the values we intended to compute in this section.

variable	impurity
temperature	0.47
humidity	0.26
wind	0.00

First, we consider the case when the "temperature" explanatory variable is used to predict the the "play tennis" target variable.

	temperature cool	(no)	temperature mild	(yes)	temperature hot	(no)
play tennis no	6	(1)	14	(1)		(0)
play tennis yes	5	(1)	4,10	(2)		(0)

The impurities are 1/2, 4/9, DND (does not defined).

The combined impurity is (2/5)(1/2)+(3/5)(4/9)+(0)(DND)=0.47.

We turn to predict the "play tennis" variable using the "humidity" vari-

able.

	humidity normal		humidity high	
play tennis no	6	(1)		(0)
play tennis yes	5,10	(2)	4,14	(2)

The impurities are 4/9, 0.
The combined impurity is (3/5)(4/9)+(2/5)(0)=0.26.
Finally, the "wind" variable is used to predict the "play tennis" variable.

	wind strong		wind weak	
play tennis no	6,14	(2)		(0)
play tennis yes		(0)	4,5,10	(3)

The impurities are 0, 0.
The combined impurity is 0.

8.6 The "outlook" variable is equal to "overcast"

The small training date set we will be busy with is the following.

day	outlook	temperature	humidity	wind	play tennis
3	overcast	hot	high	weak	yes
7	overcast	cool	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes

The last column shows that this is pure node and so we do not need to compute the predictive powers of the variables "temperature", "humidity", "wind". No branching occurs at this node.

8.7 The "outlook" variable is equal to "sunny"

The small training date set we use is contained in the next table.

day	outlook	temperature	humidity	wind	play tennis
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
11	sunny	mild	normal	strong	yes

We summarize the uncertainties associated with the variables in the following table. These are the values we intended to compute in this section.

variable	uncertainty
none	0.600
temperature	0.200
humidity	0.000
wind	0.400

We do not use any of the explanatory variables "temperature", "humidity", "wind" to predict the "play tennis" target variable. The prediction rule is ()→yes. (The prediction is always "yes", that is, the prediction rule is the constant function taking the "yes" value.) The average of the deviation is 3/5=0.600. The average of the deviations of a prediction will be compared against this ground level. If the average of the deviation is not smaller than 0.600, then the new prediction can be sorted out.

day	outlook	play tennis actual	play tennis predicted	deviation
1	sunny	no	yes	1
2	sunny	no	yes	1
8	sunny	no	yes	1
9	sunny	yes	yes	0
11	sunny	yes	yes	0

We use the "temperature" explanatory variable to predict the "play tennis" target variable. The prediction rule is cool→yes, mild→yes, hot→no. The average of the deviations is 1/5=0.200.

day	outlook	temperature	play tennis actual	play tennis predicted	play tennis tennis
9	sunny	cool	yes	yes	0
8	sunny	mild	no	yes	1
11	sunny	mild	yes	yes	0
1	sunny	hot	no	no	0
2	sunny	hot	no	no	0

Figure 8.7 is the regression tree using the "outlook" and "temperature" explanatory variables to predict the "play tennis" target variable. The predictive power of this tree is better than the predictive power of the tree without the "temperature" variable.

We use the "humidity" explanatory variable to predict the "play tennis" target variable. The prediction rule is normal→yes, high→no. The average

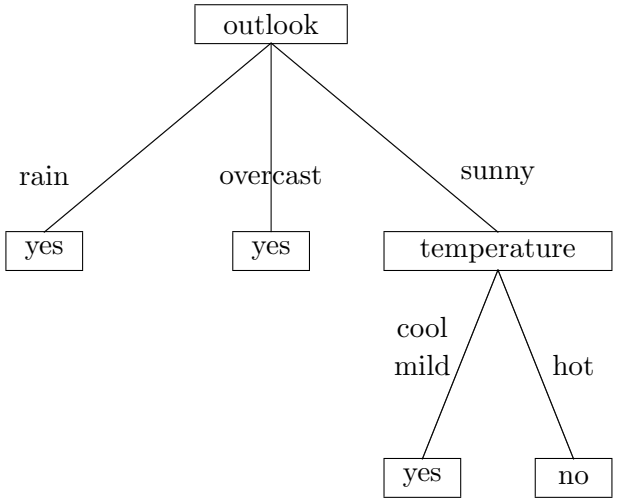


Figure 8.7: The regression tree based on the "outlook" and ´´temperature" variables.

of the deviations is 0/5=0.000.

day	outlook	humidity	play tennis actual	play tennis predicted	deviation
1	sunny	high	no	no	0
2	sunny	high	no	no	0
8	sunny	high	no	no	0
9	sunny	normal	yes	yes	0
11	sunny	normal	yes	yes	0

Figure 8.8 is the regression tree using the "outlook" and "humidity" explanatory variables to predict the "play tennis" target variable. The predictive power of this tree is better than the predictive power of the tree without the "humidity" variable.

We use the "wind" explanatory variable to predict the "play tennis" target variable. The prediction rule is weak→no, strong→yes. The average

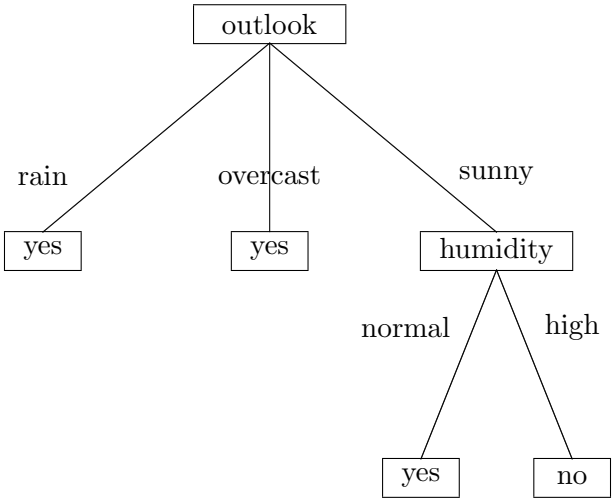


Figure 8.8: The regression tree based on the "outlook" and ´´humidity" variables.

of the deviations is 2/5=0.400.

day	outlook	wind	play tennis actual	play tennis predicted	deviation
1	sunny	weak	no	no	0
8	sunny	weak	no	no	0
9	sunny	weak	yes	no	1
2	sunny	strong	no	yes	1
11	sunny	strong	yes	yes	0

Figure 8.9 below is the regression tree using the "outlook" and "wind" explanatory variables to predict the "play tennis" target variable. The predictive power of this tree is better than the predictive power of the tree without the "wind" variable.

We summarize the predictive powers of the variables in the following table. These are the values we plan to compute in this section.

variable	impurity
temperature	0.20
humidity	0.00
wind	0.26

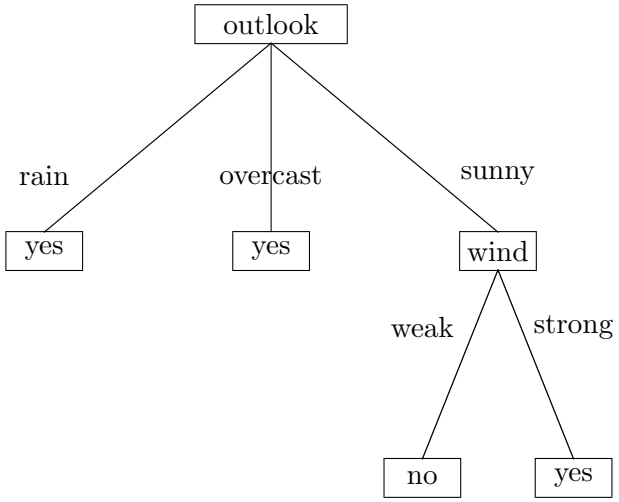


Figure 8.9: The regression tree based on the "outlook" and ´´wind" variables.

	temperature cool	(yes)	temperature mild	(yes)	temperature hot	(no)
play tennis no		(0)	8	(1)	1,2	(2)
play tennis yes	9	(1)	11	(1)		(0)

The impurities are 0, 1/2, 0.

The combined impurity is (1/5)(0)+(2/5)(1/2)+(2/5)(0)=1/5=0.20.

We turn to predict the "play tennis" variable using the "humidity" variable.

	humidity normal	(yes)	humidity high	(no)
play tennis no		(0)	1,2,8	(3)
play tennis yes	9,11	(2)		(0)

The impurities are 0, 0.

The combined impurity is 0.

Finally, the "wind" variable is used to predict the "play tennis" variable.

	wind strong	(yes)	wind weak	(no)
play tennis no	2	(1)	1,8	(2)
play tennis yes	11	(1)	9	(1)

The impurities are 0, 4/9.

The combined impurity is (2/5)(0)+(3/5)(4/9)=4/15=0.26.

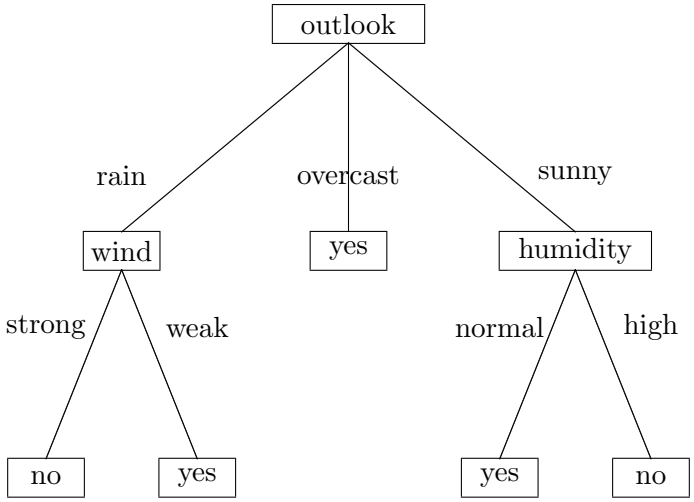


Figure 8.10: The decision tree based on the "outlook", "wind", "humidity" variables.

8.8 The extended decision tree

In the small data set when the "outlook" variable takes the "rain" value the best predictor for the "play tennis" is the "wind" attribute.

In the small data set when the "outlook" variable takes the "overcast" value the is no need for further predictor variable.

In the small data set when the "outlook" variable takes the "sunny" value the best predictor for the "play tennis" is the "humidity" feature. Figure 8.10 is a possible geometric representation of the decision tree we just constructed.

Chapter 9

Continuous variables I

9.1 The training data set

The next table lists the variables, their names, descriptions, and ranges related to a small size toy example.

variable name	variable description	variable range
chest pain	complaint of pain	no,yes
blocked arteries	condition of arteries	no,yes
weight	body weight	$(0, +\infty)$
heart disease	diagnosed illness	no,yes

The "heart disease" is the target variable and "chest pain", "blocked arteries", "weight" are the explanatory variables.

The explanatory variable "weight" is continuous. We will reduce this variable to a discrete variable. Then we apply the machinery of decision trees developed for discrete variables. Discretising a continuous variable can be done in many ways. We will explore some of the naturally occurring

options. The training data set contains information about 16 patients.

patient number	chest pain	blocked arteries	patient weight	heart disease
1	yes	yes	205	yes
2	yes	yes	185	yes
3	no	no	180	yes
4	no	no	182	yes
5	yes	yes	210	yes
6	yes	no	200	yes
7	yes	yes	167	yes
8	yes	yes	170	yes
9	no	yes	156	no
10	no	yes	160	no
11	no	yes	125	no
12	no	yes	110	no
13	yes	no	168	no
14	yes	no	165	no
15	yes	yes	172	no
16	no	yes	175	no

First we reduce the "weight" variable to a binary variable. The median of the "weight" looks a reasonable choice to sort the values of the variable into two classes.

serial number	patient number	ordered weight	binary weight
1	12	110	low
2	11	125	low
3	9	156	low
4	10	160	low
5	14	165	low
6	7	167	low
7	13	168	low
8	8	170	low
9	15	172	high
10	16	175	high
11	3	180	high
12	4	182	high
13	2	185	high
14	6	200	high
15	1	205	high
16	5	210	high

The medium of the weight is equal to $(170 + 172)/2 = 171$. The new

discretised "weight" variable has two values "low" and "high" depending on the original weight is less than 171 or larger than 171. In other words, the cut point for turning the "weight" variable into binary is equal to 171. (The variable "weight" is taking on the value 171 in the training data set.) The new training data set is the following.

patient number	chest pain	blocked arteries	patient weight	heart disease
1	yes	yes	high	yes
2	yes	yes	high	yes
3	no	no	high	yes
4	no	no	high	yes
5	yes	yes	high	yes
6	yes	no	high	yes
7	yes	yes	low	yes
8	yes	yes	low	yes
9	no	yes	low	no
10	no	yes	low	no
11	no	yes	low	no
12	no	yes	low	no
13	yes	no	low	no
14	yes	no	low	no
15	yes	yes	high	no
16	no	yes	high	no

9.2 The predictive power of the variables

We compute the predictive power of the variables.

variable attribute feature	accuracy predictive power	false positive rate	true positive rate
none	0.500	1.000	1.000
chest pain	0.693	0.375	0.750
blocked arteries	0.562	0.250	0.375
weight	0.750	0.250	0.750

	chest pain no	(no)	chest pain yes	(yes)
heart disease no	9,10,11,12,16	(5)	13,14,15	(3)
heart disease yes	3,4	(2)	1,2,5,6,7,8	(6)

		predicted no	predicted yes
actual	no	5	3
actual	yes	2	6

The accuracy is equal to (5+6)/16=0.693.
The false positive rate is equal to (3)/(5+3)=0.375.
The true positive rate is equal to (6)/(2+6)=0.750.

		blocked arteries no	(yes)	blocked arteries yes	(no)
heart disease	no	13,14	(2)	9,10,11,12,15,16	(6)
heart disease	yes	3,4,6	(3)	1,2,5,7,8	(5)

		predicted no	predicted yes
actual	no	6	2
actual	yes	5	3

The accuracy is equal to (6+3)/16=0.562.
The false positive rate is equal to (2)/(6+2)=0.250.
The true positive rate is equal to (3)/(5+3)=0.375.

		weight low	(no)	weight high	(yes)
heart disease	no	9,10,11,12,13,14	(6)	15,16	(2)
heart disease	yes	7,8	(2)	1,2,3,4,5,6	(6)

		predicted no	predicted yes
actual	no	6	2
actual	yes	2	6

The accuracy is equal to (6+6)/16=0.750.
The false positive rate is equal to (2)/(6+2)=0.250.
The true positive rate is equal to (6)/(2+6)=0.750.

The decision trees based on the variables “chest pain”, “blocked arteries”, and the discrete version of “weight” are denoted by T_1, T_2, T_3 presented below.

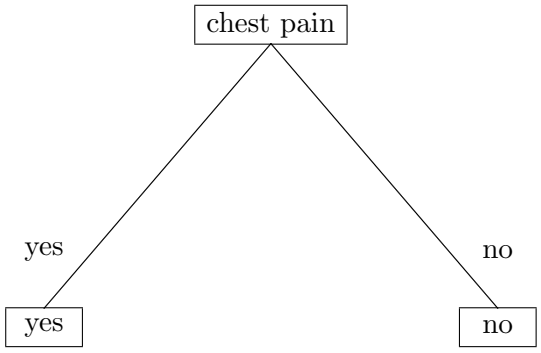


Figure 9.1: The decision tree T_1 based on the variable "chest pain".

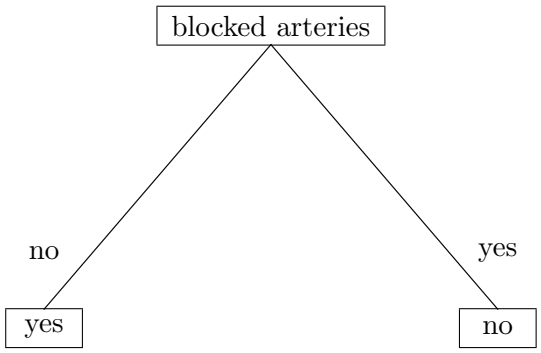


Figure 9.2: The decision tree T_2 based on the variable "blocked arteries".

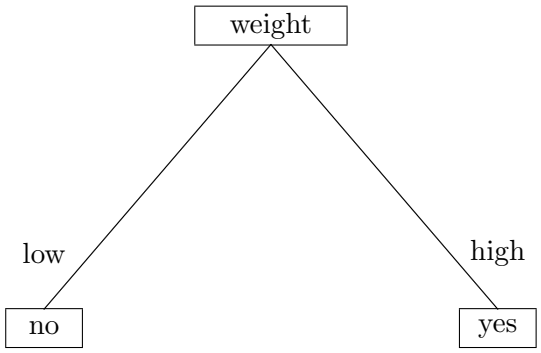


Figure 9.3: The decision tree T_3 based on the variable "weight".

9.3 The ensemble method

We form a committee consisting of the trees T_1, T_2, T_3.

patient number	vote T_1	vote T_2	vote T_3	prediction	heart disease
1	yes	no	yes	YES	yes
2	yes	no	yes	YES	yes
3	no	yes	yes	YES	yes
4	no	yes	yes	YES	yes
5	yes	no	yes	YES	yes
6	yes	yes	yes	YES	yes
7	yes	no	no	NO	yes
8	yes	no	no	NO	yes
9	no	no	no	NO	no
10	no	no	no	NO	no
11	no	no	no	NO	no
12	no	no	no	NO	no
13	yes	yes	no	YES	no
14	yes	yes	no	YES	no
15	yes	no	yes	YES	no
16	no	no	yes	NO	no

We compute the predictive power of this committee.

	predicted NO		predicted YES	
actual no	9,10,11,12,16	(5)	13,14,15	(3)
actual yes	7,8	(2)	1,2,3,4,5,6	(6)

	predicted no	predicted yes
actual no	5	3
actual yes	2	6

The accuracy is equal to (5+6)/16=0.693.

The false positive rate is equal to (3)/(5+3)=0.375.

The true positive rate is equal to (6)/(2+6)=0.750.

Let us reduce the "weight" variable to a ternary variable. Using the cut points 166 and 181 we discretise the "weight" variable to have three values

"low", "medium" and "high".

serial number	patient number	ordered weight	ternary weight
1	12	110	low
2	11	125	low
3	9	156	low
4	10	160	low
5	14	165	low
6	7	167	medium
7	13	168	medium
8	8	170	medium
9	15	172	medium
10	16	175	medium
11	3	180	medium
12	4	182	high
13	2	185	high
14	6	200	high
15	1	205	high
16	5	210	high

The new training data set is the following.

patient number	chest pain	blocked arteries	patient weight	heart disease
1	yes	yes	high	yes
2	yes	yes	high	yes
3	no	no	medium	yes
4	no	no	high	yes
5	yes	yes	high	yes
6	yes	no	high	yes
7	yes	yes	medium	yes
8	yes	yes	medium	yes
9	no	yes	low	no
10	no	yes	low	no
11	no	yes	low	no
12	no	yes	low	no
13	yes	no	medium	no
14	yes	no	low	no
15	yes	yes	medium	no
16	no	yes	medium	no

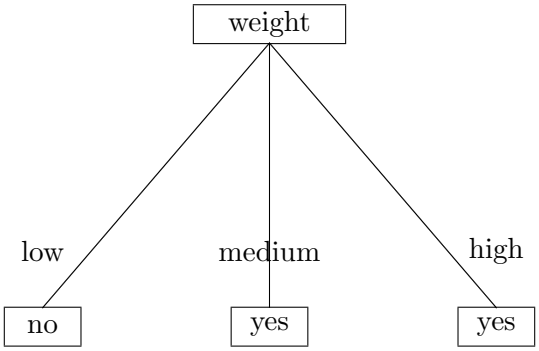


Figure 9.4: The decision tree T_4 based on the variable "weight".

We compute the predictive power of the ternary discrete "weight" variable.

	weight low (no)	weight medium (yes)	weight high (yes)
heart disease no	9,10,11, 12,14 (5)	13,15,16 (3)	(0)
heart disease yes	(0)	3,7,8 (3)	1,2,4, 5,6 (5)

	predicted no	predicted yes
actual no	5	3
actual yes	0	8

The accuracy is equal to (5+8)/16=0.812.

The false positive rate is equal to (3)/(5+3)=0.375.

The true positive rate is equal to (8)/(0+8)=1.000.

Figure 9.4 is a decision tree based solely on the discretised version of the "weight" variable. We denote this tree by T_4. We form a committee

consisting of the trees T_1, T_2, T_4.

patient number	vote T_1	vote T_2	vote T_4	prediction	heart disease
1	yes	no	yes	YES	yes
2	yes	no	yes	YES	yes
3	no	yes	yes	YES	yes
4	no	yes	yes	YES	yes
5	yes	no	yes	YES	yes
6	yes	yes	yes	YES	yes
7	yes	no	yes	YES	yes
8	yes	no	yes	YES	yes
9	no	no	no	NO	no
10	no	no	no	NO	no
11	no	no	no	NO	no
12	no	no	no	NO	no
13	yes	yes	yes	YES	no
14	yes	yes	no	YES	no
15	yes	no	yes	YES	no
16	no	no	yes	NO	no

We compute the predictive power of this committee.

	predicted NO	predicted YES
actual no	9,10,11,12,16 (5)	13,14,15 (3)
actual yes	(0)	1,2,3,4,5,6,7,8 (8)

	predicted no	predicted yes
actual no	5	3
actual yes	0	8

The accuracy is equal to (5+8)/16=0.812.

The false positive rate is equal to (3)/(5+3)=0.375.

The true positive rate is equal to (8)/(0+8)=1.000.

9.4 More tactical discretizations

We order the weighs into an increasing sequence and tag the heart disease status along with the weigh.

serial number	patient number	ordered weight	ternary weight	heart disease
1	12	110	low	no
2	11	125	low	no
3	9	156	low	no
4	10	160	low	no
5	14	165	low	no
6	7	167	medium	yes
7	13	168	medium	no
8	8	170	medium	yes
9	15	172	medium	no
10	16	175	medium	no
11	3	180	high	yes
12	4	182	high	yes
13	2	185	high	yes
14	6	200	high	yes
15	1	205	high	yes
16	5	210	high	yes

Choosing a cut point between 165 and 167 makes the leaf attached to the "low" branch to be pure. Similarly, choosing a cut point between 175 and 180 makes the leaf attached to the "high" edge to be pure. Only the the leaf on the "medium" edges remains impure. This explains why the ternary discretization of the "weight" variable into "low", "medium", "high" works well. We compute the predictive power of the ternary discrete "weight" variable.

	weight low (no)	weight medium (no)	weight high (yes)
heart disease no	9,10,11, 12,14 (5)	13,15,16 (3)	(0)
heart disease yes	(0)	7,8 (2)	1,2,3, 4,5,6 (6)

	predicted no	predicted yes
actual no	8	0
actual yes	2	6

The accuracy is equal to (8+6)/16=0.875.

The false positive rate is equal to (0)/(8+0)=0.000.

The true positive rate is equal to (6)/(2+6)=0.750.

Suppose that we want to use a binary version of the "weight" variable. A cut point between 175 and 180 looks promising. Say, we use the cut point 177.5. The leaf at the "high" edge is a pure node. The leaf at the "low" edge is an impure node. But this impurity is fairly small.

serial number	patient number	ordered weight	binary weight	heart disease
1	12	110	low	no
2	11	125	low	no
3	9	156	low	no
4	10	160	low	no
5	14	165	low	no
6	7	167	low	yes
7	13	168	low	no
8	8	170	low	yes
9	15	172	low	no
10	16	175	low	no
11	3	180	high	yes
12	4	182	high	yes
13	2	185	high	yes
14	6	200	high	yes
15	1	205	high	yes
16	5	210	high	yes

The new training data set is the following.

patient number	chest pain	blocked arteries	patient weight	heart disease
1	yes	yes	high	yes
2	yes	yes	high	yes
3	no	no	high	yes
4	no	no	high	yes
5	yes	yes	high	yes
6	yes	no	high	yes
7	yes	yes	low	yes
8	yes	yes	low	yes
9	no	yes	low	no
10	no	yes	low	no
11	no	yes	low	no
12	no	yes	low	no
13	yes	no	low	no
14	yes	no	low	no
15	yes	yes	low	no
16	no	yes	low	no

We compute the predictive power of the new binary version of the variable "weight".

	weight low		weight high	
heart disease no	9,10,11,12,13,14,15,16	(8)		(0)
heart disease yes	7,8	(2)	1,2,3,4,5,6	(6)

	predicted no	predicted yes
actual no	8	0
actual yes	2	6

The accuracy is equal to (8+6)/16=0.875.
The false positive rate is equal to (0)/(8+0)=0.000.
The true positive rate is equal to (6)/(2+6)=0.750.

9.5 Splitting the training data set

The decision tree based on the "weight" variable using the cut point 177.5 has the largest predictive power. The leaf node on the "high" edge is pure and so we do want to extend the tree at this node. The leaf node on the "low" edge is not pure and so we may to extend the tree at this node. We

split the original training data set and consider the part in which the weigth "variable" takes on the value "low".

patient number	chest pain	blocked arteries	patient weight	heart disease
7	yes	yes	low	yes
8	yes	yes	low	yes
9	no	yes	low	no
10	no	yes	low	no
11	no	yes	low	no
12	no	yes	low	no
13	yes	no	low	no
14	yes	no	low	no
15	yes	yes	low	no
16	no	yes	low	no

We compute the predictive power of the variables "chest pain" and "blocked arteries" in this smaller data set.

variable attribute feature	accuracy predictive power	false positive rate	true positive rate
none	0.800	1.000	1.000
chest pain	0.800	0.000	0.000
blocked arteries	0.800	0.000	0.000

We compute the predictive power of the variable "chest pain".

	chest pain no	(no)	chest pain yes	(no)
heart disease no	9,10,11,12,16	(5)	13,14,15	(3)
heart disease yes		(0)	7,8	(2)

	predicted no	predicted yes
actual no	8	0
actual yes	2	0

The accuracy is equal to (8+0)/10=0.800.
The false positive rate is equal to (0)/(8+0)=0.000.
The true positive rate is equal to (0)/(2+0)=0.000.
We compute the predictive power of the variable "blocked arteries".

	blocked arteries no	(no)	blocked arteries yes	(no)
heart disease no	13,14	(2)	9,10,11,12,15,16	(6)
heart disease yes		(0)	7,8	(2)

		predicted no	predicted yes
actual	no	8	0
actual	yes	2	0

The accuracy is equal to (8+0)/10=0.800.

The false positive rate is equal to (0)/(8+0)=0.000.

The true positive rate is equal to (0)/(2+0)=0.000.

The variables "chest pain" and "blocked arteries" do not improve the predictive power of the decision tree based of the variable "weight" alone.

The decision tree based on the "weight" variable using two cut points 166 and 181 does not have the largest predictive power. However, for the sake of curiosity we make an attempt to extend this tree. The leaf node on the "high" edge is pure and so we do not need to extend the tree at this node. Similarly, the leaf node on the "low" edge is pure and so we do not need to extend the tree at this node. We may try to extend the tree at the leaf node on the edge labeled "medium". We split the original training data set and consider the part in which the weight "variable" takes on the value "medium".

patient number	chest pain	blocked arteries	patient weight	heart disease
3	no	no	medium	yes
7	yes	yes	medium	yes
8	yes	yes	medium	yes
13	yes	no	medium	no
15	yes	yes	medium	no
16	no	yes	medium	no

We compute the predictive power of the variables "chest pain" and "blocked arteries" in this smaller data set.

variable attribute feature	accuracy predictive power	false positive rate	true positive rate
none	0.500	1.000	1.000
chest pain	0.500	0.666	0.666
blocked arteries	0.500	0.666	0.666

We compute the predictive power of the variable "chest pain".

		chest pain no	(no)	chest pain yes	(yes)
heart disease	no	16	(1)	13,15	(2)
heart disease	yes	3	(1)	7,8	(2)

	predicted no	predicted yes
actual no	1	2
actual yes	1	2

The accuracy is equal to (1+2)/6=0.500.
The false positive rate is equal to (2)/(1+2)=0.666.
The true positive rate is equal to (2)/(1+2)=0.666.
We compute the predictive power of the variable "blocked arteries".

	blocked arteries no (no)	blocked arteries yes (yes)
heart disease no	13 (1)	15,16 (2)
heart disease yes	3 (1)	7,8 (2)

	predicted no	predicted yes
actual no	1	2
actual yes	1	2

The accuracy is equal to (1+2)/6=0.500.
The false positive rate is equal to (2)/(1+2)=0.666.
The true positive rate is equal to (2)/(1+2)=0.666.
The variables "chest pain" and "blocked arteries" do not improve the predictive power of the decision tree based of the variable "weight" alone.

The results of our experiments with discretezing continuous explanatory variables can be summed up in a few points. As a first attempt one may discretize a variable by dividing the values using the median (or mean) as a cut point. The result will be a binary variable. One can get a ternary variable by using two cut points such that the three resulting parts contain approximately equal number of values. Our expectation is that the predictive power of the discretized variable improves with the number of the cut points increases. Our computation reinforce this expectation. The predictive power of the binary version of the "weight" variable is equal to 0.750. The predictive power of the ternary version of the "weight" variable is equal to 0.812.

Plainly, trying all possible cut points instead of settling for the median improves the predictive power of the binary version of the variable. In our computation the predictive power of the optimized binary version of the "weight" variable is equal to 0.875. We should keep in mind that carrying out a procedure for computing the predictive power of a discretized variable can be computationally expensive when the training data set is large.

Chapter 10

Continuous variables II

10.1 The training data set

The next table lists the variables, their names, descriptions, and ranges related to a small size toy example.

variable name	variable description	variable range
chest pain	complaint of pain	no,yes
blocked arteries	condition of arteries	no,yes
weight	body weight	$(0, +\infty)$
heart disease	diagnosed illness	no,yes

The "heart disease" is the target variable and "chest pain", "blocked arteries", "weight" are the explanatory variables.

The explanatory variable "weight" is continuous. We will reduce this variable to a discrete variable. Then we apply the machinery of decision trees developed for discrete variables.

10.2 Systematic discretization

Discretising a continuous variable can be done in many ways. We will present here a systematic way. In practice at first we would use the most straightforward technique to discretize a continuous variable. Say, we would use the average as a cut point. In case we are interested in a refinement, then we may apply the method we describe here. However, the refinement has a higher computational cost. The training data set contains information

about 16 patients.

patient number	chest pain	blocked arteries	patient weight	heart disease
1	yes	yes	205	yes
2	yes	yes	185	yes
3	no	no	180	yes
4	no	no	182	yes
5	yes	yes	210	yes
6	yes	no	200	yes
7	yes	yes	167	yes
8	yes	yes	170	yes
9	no	yes	156	no
10	no	yes	160	no
11	no	yes	125	no
12	no	yes	110	no
13	yes	no	168	no
14	yes	no	165	no
15	yes	yes	172	no
16	no	yes	175	no

We would like to reduce the "weight" variable to a binary variable. We rearrange the values of the "weight" into an increasing (non-decreasing) sequence. It starts with the smallest and ends with the largest value. Between each consecutive values we locate a midpoint. These are the candidates for the cut points. In this way we picked 15 candidates as possible cut points. We augment this list with one number which is smaller than the minimal weight. Similarly, we augment the possible cut point list with a further number which is larger than the maximum weight. Altogether, we have $(16-1)+1+1=17$ choices for the optimal cut point.

The second column of the next table contains the ordered values of the "weight" variable. The last column holds the information if the patient is

diagnosed with heart disease.

	ordered weight	heart disease
1	110	no
2	125	no
3	156	no
4	160	no
5	165	no
6	167	yes
7	168	no
8	170	yes
9	172	no
10	175	no
11	180	yes
12	182	yes
13	185	yes
14	200	yes
15	205	yes
16	210	yes

We say that there is a sign change in the no-yes sequence listed in the last column if a "no" is followed by a "yes". Similarly, we say that there is a sign change if a "yes" is followed by a "no". There are 5 sign changes in the no-yes sequence in the table above. The main point is that we need not to check 17 choices in order to select the optimal cut points. It is enough to test the 5 cut points associated with the sign changes in the no-yes sequence. We are turning now to the inspection of these 5 possible cut points.

Checking the cut point 166.

	weight low (no)	weight high (yes)
heart disease no	1,2,3,4,5 (5)	7,9,10 (3)
heart disease yes	(0)	6,8,11,12,13,14,15,16 (8)

	predicted no	predicted yes
actual no	5	3
actual yes	0	8

Testing the cut point 167.5.

	weight low (no)	weight high (yes)
heart disease no	1,2,3,4,5 (5)	7,9,10 (3)
heart disease yes	6 (1)	8,11,12,13,14,15,16 (7)

	predicted no	predicted yes
actual no	5	3
actual yes	1	7

Dealing with the cut point 169.

	weight low	(no)	weight high	(yes)
heart disease no	1,2,3,4,5,7	(6)	9,10	(2)
heart disease yes	7	(1)	8,11,12,13,14,15,16	(7)

	predicted no	predicted yes
actual no	6	2
actual yes	1	7

Evaluating the cut point 171.

	weight low	(no)	weight high	(yes)
heart disease no	1,2,3,4,5,7	(6)	9,10	(2)
heart disease yes	7,8	(2)	11,12,13,14,15,16	(6)

	predicted no	predicted yes
actual no	6	2
actual yes	2	6

Having a closer look at the cut point 177.5.

	weight low	(no)	weight high	(yes)
heart disease no	1,2,3,4,5,7,9,10	(8)		(0)
heart disease yes	7,8	(2)	11,12,13,14,15,16	(6)

	predicted no	predicted yes
actual no	8	0
actual yes	2	6

The following table summarizes the result of our considerations.

cut point	accuracy	false positive rate	true positive rate
166	0.813	0.375	1.000
167.5	0.750	0.375	0.875
169	0.813	0.250	0.875
171	0.750	0.250	0.750
177.5	0.875	0.000	0.750

The massege of this long argument is that 177.5 is the optimal cut point.

10.3 Some reasoning

In this section we reason why we can focus our attention on the sign changes when we locate the optimal cut point. The second column of the next table lists the 17 possible cut points numbered from 0 to 16. The first column contains the serial numbers of the cut points.

	cut point	heart disease
0	109.5	no
1	117.5	no
2	140.5	no
3	158.0	no
4	162.5	no
5	166.0	no
6	167.5	yes
7	169.0	no
8	171.0	yes
9	173.5	no
10	177.5	no
11	181.0	yes
12	183.5	yes
13	192.5	yes
14	202.5	yes
15	207.5	yes
16	210.5	yes

Figure 10.1 below depicts the ROC curve. The top right corner of the unit square whose coordinates are (1,1) corresponds to the cut point 109.5. The serial number of this cut point is 0. At this cut point the false positive rate is equal to 1 and the true positive rate is equal to 1. The point with

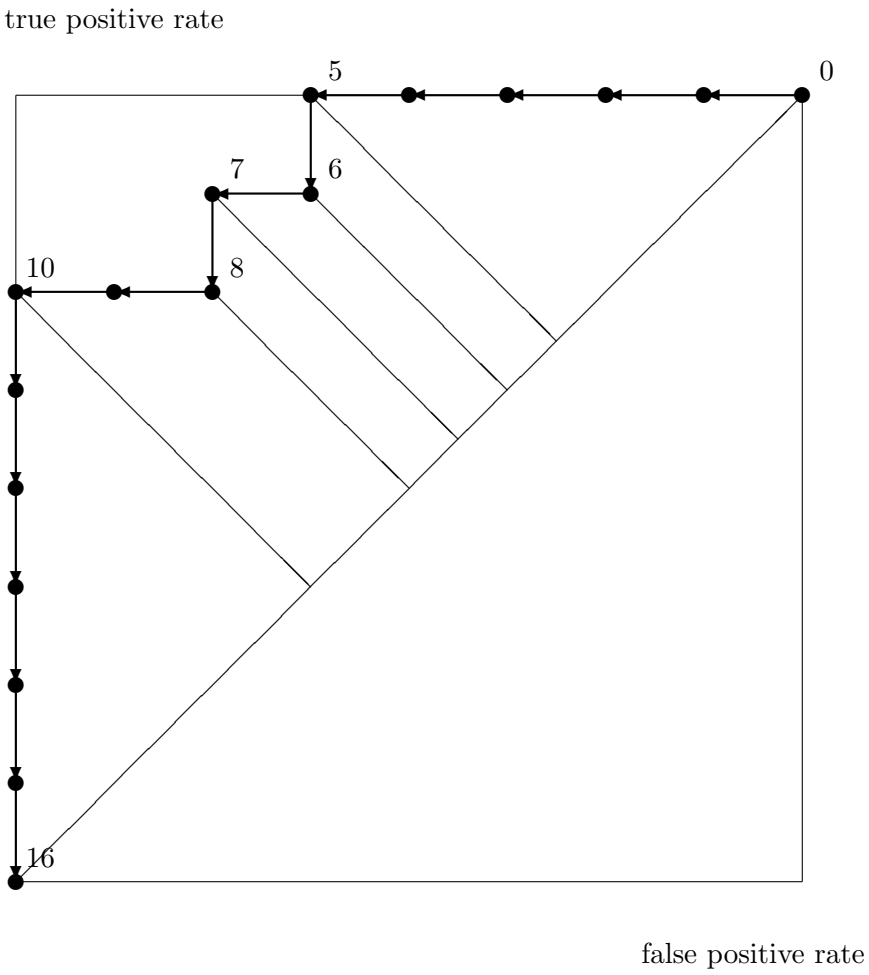


Figure 10.1: The ROC curve.

coordinates (1/8,1) corresponds to the cut point 117.5. The serial number of this cut point is 1. At this cut point the false positive rate is equal to 1/8 and the true positive rate is equal to 1. Continuing in this way finally we reach the bottom left corner of the unit square whose coordinates are (0,0) corresponds to the cut point 210.5. The serial number of this cut point is 16. As we inspect the 17 possible cut points we move along the path marked on the figure. The path starts at the point labeled 0 and ends at the point labeled 16. The accuracy at each cut point is proportional to the distance of the point from the diagonal in the ROC picture. Clearly, the cut point 177.5 with serial number 10 has the largest distance from the diagonal. Further, it is also clear that only points on the path where the path is turning has a chance to be an optimal cut point.

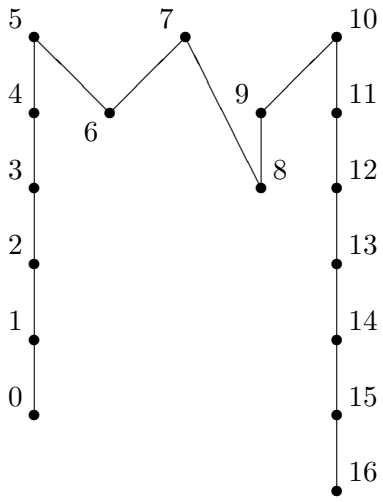


Figure 10.2: The Hasses diagram of the partially ordered set $\{0, 1, \ldots, 16\}$.

Let us find the Pareto front. Point 5 dominates points 0,1,2,3,4,6. Point 7 dominates points 6,8. Point 10 dominates 8,9,1,11,12,13,14,15,16. Points 5,7,10 are pair-wise incomparable. Therefore points 5,7,10 form the Pareto front.

The set of points $\{0, 1, \ldots, 16\}$ equipped with the dominance relation is a partially ordered set. Just to be on the safe side we enclosed the Hasse diagram of this partially ordered set. See Figure 10.2. An inspection of the Hasses diagram reveals that the points 5,7,10 are the maximal elements of the partially ordered set.

	cut x_{L}	heart x_{U}
	8	0
1	8	0
2	8	0
3	8	0
4	8	0
5	8	0
6	7	1
7	7	1
8	6	2
9	6	2
10	6	2
11	5	3
12	4	4
13	3	5
14	2	6
15	1	7
16	0	8

Chapter 11

New variables

11.1 The training data set

When we replaced a continuous variable with a discrete variable we replaced a given variable by a new variable what we constructed using a given variable. One can construct a new variable using two given variables as well. This is the avenue we discover in this chapter.

The next table lists the variables, their names, descriptions, and ranges related to a small size toy example.

variable name	variable description	variable range
chest pain	complaint of pain	no,yes
blocked arteries	condition of arteries	no,yes
weight	body weight	low,high
heart disease	diagnosed illness	no,yes

The "heart disease" is the target variable and "chest pain", "blocked arteries", "weight" are the explanatory variables.

The explanatory variable "weight" is continuous. We will reduce this variable to a discrete variable. Then we apply the machinery of decision trees developed for discrete variables. Discretising a continuous variable can be done in many ways. We will explore some of the naturally occurring

options. The training data set contains information about 16 patients.

patient number	chest pain	blocked arteries	patient weight	heart disease
1	yes	yes	high	yes
2	yes	yes	high	yes
3	no	no	high	yes
4	no	no	high	yes
5	yes	yes	high	yes
6	yes	no	high	yes
7	yes	yes	low	yes
8	yes	yes	low	yes
9	no	yes	low	no
10	no	yes	low	no
11	no	yes	low	no
12	no	yes	low	no
13	yes	no	low	no
14	yes	no	low	no
15	yes	yes	low	no
16	no	yes	low	no

11.2 The predictive power of the variables

We compute the predictive power of the variables.

variable attribute feature	accuracy predictive power	false positive rate	true positive rate
none	0.500	1.000	1.000
chest pain	0.693	0.375	0.750
blocked arteries	0.562	0.250	0.375
weight	0.875	0.000	0.750

We start with the predictive power of "chest pain" explanatory variable.

	chest pain no	(no)	chest pain yes	(yes)
heart disease no	9,10,11,12,16	(5)	13,14,15	(3)
heart disease yes	3,4	(2)	1,2,5,6,7,8	(6)

	predicted no	predicted yes
actual no	5	3
actual yes	2	6

The accuracy is equal to (5+6)/16=0.693.

The false positive rate is equal to (3)/(5+3)=0.375.

The true positive rate is equal to (6)/(2+6)=0.750.

We turn to computing the predictive power of the "blocked arteries" explanatory variable.

	blocked arteries no (yes)	blocked arteries yes (no)
heart disease no	13,14 (2)	9,10,11,12,15,16 (6)
heart disease yes	3,4,6 (3)	1,2,5,7,8 (5)

	predicted no	predicted yes
actual no	6	2
actual yes	5	3

The accuracy is equal to (6+3)/16=0.562.

The false positive rate is equal to (2)/(6+2)=0.250.

The true positive rate is equal to (3)/(5+3)=0.375.

Finally, we compute the predictive power of the variable "weight".

	weight low (no)	weight high (yes)
heart disease no	9,10,11,12,13,14,15,16 (8)	(0)
heart disease yes	7,8 (2)	1,2,3,4,5,6 (6)

	predicted no	predicted yes
actual no	8	0
actual yes	2	6

The accuracy is equal to (8+6)/16=0.875.

The false positive rate is equal to (0)/(8+0)=0.000.

The true positive rate is equal to (6)/(2+6)=0.750.

The decision trees based on the variables "chest pain", "blocked arteries", and the discrete version of "weight" are denoted by T_1, T_2, T_3 presented below in Figures 11.1, 11.2, 11.3.

11.3 Forming a new variable

Using the "chest pain" and "blocked arteries" explanatory variable let us for a new "agreement" explanatory variable. The "agreement" variable takes on the "yes" value if the values of the variables "chest pain" and "blocked

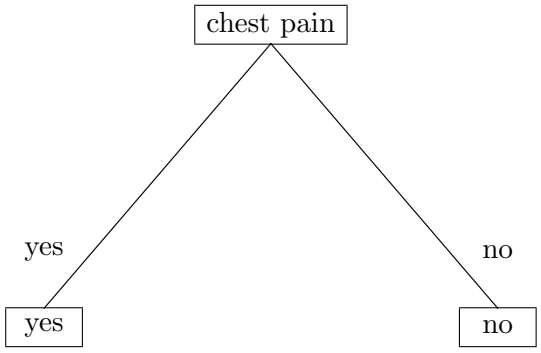


Figure 11.1: The decision tree T_1 based on the variable "chest pain".

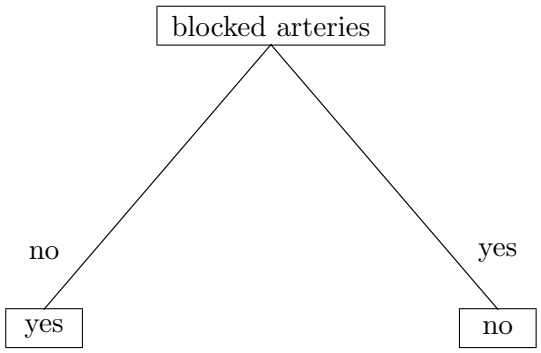


Figure 11.2: The decision tree T_2 based on the variable "blocked arteries".

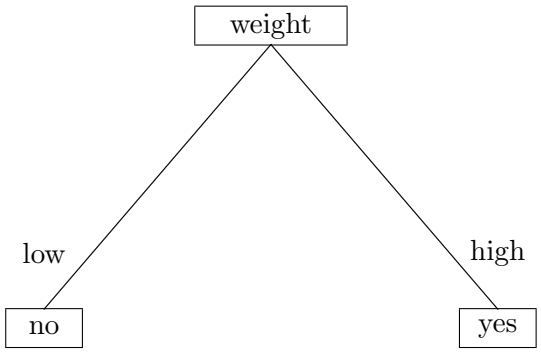


Figure 11.3: The decision tree T_3 based on the variable "weight".

arteries" agree. Further, the "agreement" variable takes on the "no" value if the values of the variables "chest pain" and "blocked arteries" do not agree. The following table gives the new extented training data set.

patient number	chest pain	blocked arteries	agree-ment	patient weight	heart disease
1	yes	yes	yes	high	yes
2	yes	yes	yes	high	yes
3	no	no	yes	high	yes
4	no	no	yes	high	yes
5	yes	yes	yes	high	yes
6	yes	no	no	high	yes
7	yes	yes	yes	low	yes
8	yes	yes	yes	low	yes
9	no	yes	no	low	no
10	no	yes	no	low	no
11	no	yes	no	low	no
12	no	yes	no	low	no
13	yes	no	no	low	no
14	yes	no	no	low	no
15	yes	yes	yes	low	no
16	no	yes	no	low	no

We compute the predictive power of the variables.

variable attribute feature	accuracy predictive power	false positive rate	true positive rate
none	0.500	1.000	1.000
chest pain	0.693	0.375	0.750
blocked arteries	0.562	0.250	0.375
agreement	0.875	0.125	0.875
weight	0.875	0.000	0.750

Only the predictive power of "agreement" explanatory variable need to be computed.

	agreement no (no)	agreement yes (yes)
heart disease no	9,10,11,12,13,14,16 (7)	15 (1)
heart disease yes	6 (1)	1,2,3,4,5,7,8 (7)

	predicted no	predicted yes
actual no	7	1
actual yes	1	7

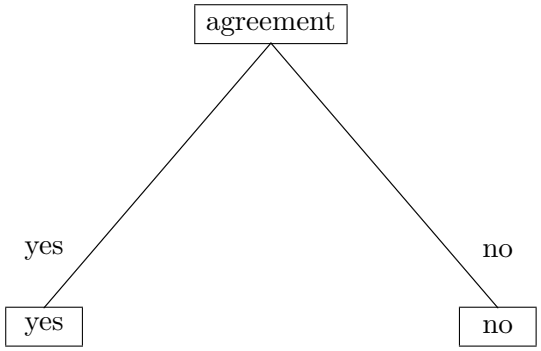


Figure 11.4: The decision tree T_4 based on the variable "agreement".

The accuracy is equal to (7+7)/16=0.875.

The false positive rate is equal to (1)/(7+1)=0.125.

The true positive rate is equal to (7)/(1+7)=0.875.

The predictive power of the new variable is quite high. We call he decision trees based on the "agreement" explanatory variable T_4 and present below. See Figure 11.4

11.4 The ensemble method

We form a committee consisting of the trees T_1, T_4, T_3.

patient number	vote T_1	vote T_4	vote T_3	prediction	heart disease
1	yes	yes	yes	YES	yes
2	yes	yes	yes	YES	yes
3	no	yes	yes	YES	yes
4	no	yes	yes	YES	yes
5	yes	yes	yes	YES	yes
6	yes	no	yes	YES	yes
7	yes	yes	no	YES	yes
8	yes	yes	no	YES	yes
9	no	no	no	NO	no
10	no	no	no	NO	no
11	no	no	no	NO	no
12	no	no	no	NO	no
13	yes	no	no	NO	no
14	yes	no	no	NO	no
15	yes	yes	no	YES	no
16	no	no	no	NO	no

We compute the predictive power of this committee.

	predicted NO	predicted YES
actual no	9,10,11,12,13,14,16 (7)	15 (1)
actual yes	(0)	1,2,3,4,5,6,7,8 (8)

	predicted no	predicted yes
actual no	7	1
actual yes	0	8

The accuracy is equal to (7+8)/16=0.938.
The false positive rate is equal to (1)/(7+1)=0.125.
The true positive rate is equal to (8)/(0+8)=1.000.
The predictive power of the committee $\{T_1, T_4, T_3\}$ is rather high.

11.5 Splitting the training data set

The decision tree based on the "weight" explanatory variable has one of the largest predictive power. The leaf node on the "high" edge is pure and so we do not want to extend the tree at this node. The leaf node on the "low" edge is not pure and so we may to extend the tree at this node. We split the original training data set and consider the part in which the weigth "variable" takes on the value "low".

patient number	chest pain	blocked arteries	agree-ment	patient weight	heart disease
7	yes	yes	yes	low	yes
8	yes	yes	yes	low	yes
9	no	yes	no	low	no
10	no	yes	no	low	no
11	no	yes	no	low	no
12	no	yes	no	low	no
13	yes	no	no	low	no
14	yes	no	no	low	no
15	yes	yes	yes	low	no
16	no	yes	no	low	no

We compute the predictive power of the variables "chest pain", "blocked arteries", "agreement" in this smaller data set. The result of this computation

is summarized in the following table.

variable attribute feature	accuracy predictive power	false positive rate	true positive rate
none	0.800	0.000	0.000
chest pain	0.800	0.000	0.000
blocked arteries	0.800	0.000	0.000
agreement	0.900	0.125	1.000

We turn to the details how the table was filled in and we compute the predictive power of the variable "chest pain".

	chest pain no	(no)	chest pain yes	(no)
heart disease no	9,10,11,12,16	(5)	13,14,15	(3)
heart disease yes		(0)	7,8	(2)

	predicted no	predicted yes
actual no	8	0
actual yes	2	0

The accuracy is equal to (8+0)/10=0.800.
The false positive rate is equal to (0)/(8+0)=0.000.
The true positive rate is equal to (0)/(2+0)=0.000.
Next we compute the predictive power of the variable "blocked arteries".

	blocked arteries no	(no)	blocked arteries yes	(no)
heart disease no	13,14	(2)	9,10,11,12,15,16	(6)
heart disease yes		(0)	7,8	(2)

	predicted no	predicted yes
actual no	8	0
actual yes	2	0

The accuracy is equal to (8+0)/10=0.800.
The false positive rate is equal to (0)/(8+0)=0.000.
The true positive rate is equal to (0)/(2+0)=0.000.
Finally, we compute the predictive power of the variable "agreement".

	agreement no	(no)	agreement yes	(no)
heart disease no	9,10,11,12,13,14,16	(7)	15	(1)
heart disease yes		(0)	7,8	(2)

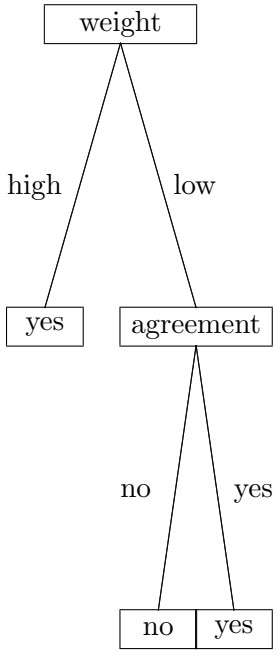


Figure 11.5: The decision tree based on the variables "weight" and "agreement".

	predicted no	predicted yes
actual no	7	1
actual yes	0	2

The accuracy is equal to (7+2)/10=0.900.

The false positive rate is equal to (1)/(7+1)=0.125.

The true positive rate is equal to (2)/(0+2)=1.000.

The explanatory variables "chest pain" and "blocked arteries" are not over performing the base level on this discretised training data set. Therefore, we choose the variable "agreement" for branching. Figure 11.5 is a possible geometric representation of the new classification tree. We assess the predictive power of this tree.

	predicted no		predicted no	
actual no	9,10,11,12,13,14,16	(7)	15	(1)
actual yes		(0)	1,2,3,4,5,6,7,8	(8)

	predicted no	predicted yes
actual no	7	1
actual yes	0	8

The accuracy is equal to (7+8)/16=0.938.

The false positive rate is equal to (1)/(7+1)=0.125.

The true positive rate is equal to (8)/(0+8)=1.000.

The predictive power of this tree is exactly the same as the predictive power of the committee $\{T_1, T_4, T_3\}$ we have constructed earlier.

Appendix A

Complete trees

A.1 Trees in table form

Let us consider the training data set from Chapter 11 section 11.1.

patient number	chest pain	blocked arteries	patient weight	heart disease
1	yes	yes	high	yes
2	yes	yes	high	yes
3	no	no	high	yes
4	no	no	high	yes
5	yes	yes	high	yes
6	yes	no	high	yes
7	yes	yes	low	yes
8	yes	yes	low	yes
9	no	yes	low	no
10	no	yes	low	no
11	no	yes	low	no
12	no	yes	low	no
13	yes	no	low	no
14	yes	no	low	no
15	yes	yes	low	no
16	no	yes	low	no

We construct a complete binary depicted in Figure A.1. Each row of the table containing the training data set corresponds to a path in the complete binary tree. Each path starts at the root node labeled "chest pain" and ends at a leaf node of the tree. For example, the first row of the table corresponds to the path that ends at the rightmost leaf node of the tree. Into the box representing the leaf node we put [1,n]. This means that row 1 ends at this node and patient 1 was diagnosed not to have heart disease. As the table containing the training data set has 16 rows 16 bracketed items are

entered into the boxes representing the leaf nodes of the complete binary tree. Some boxes may hold more than one bracketed items. Some boxes may remain empty. From the table listing the records of the training set one can construct a complete binary tree with filled leaf nodes. Conversely, from the binary tree one can reconstruct the rows of the table describing the records of the training data set. In short, the training data set and the complete binary tree both contain the same information about the 16 patients.

We make an observation. The box of the leaf node on the left of the rightmost leaf node contains information on the patients 7,8,15. Patients 7 and 8 are diagnosed having heart disease. While patient 15 diagnosed not having heart disease. This means that the attributes "chest pain", "blocked arteries", "weight" do not carry enough information to distinguish the "heart disease" status of these patients.

If in a box representing a leaf node there are both "n" and "y" letters, then we call the leaf an impure leaf. Otherwise, we call the leaf a pure leaf. We can see that our complete binary tree has only one impure leaf. We will not be able correctly classify patient 15. But all the other patients we can classify correctly.

Here is the procedure to construct the decision tree. Let us delete temporarily patient 15. Now each leaf node are pure. Each empty leaf node box remains empty. For the non-empty boxes we change the content of each leaf node box putting "no" or "yes". The new tree can be seen in Figure A.7. The empty boxes can be can be cut off. Nodes where there are no branching can be deleted. Nodes where there are branching but the content of the descendants are identical can be deleted. In this way we we reach a decision tree. Figure A.3 exhibits the decision tree we constructed. This decision tree correctly classifies all patients except patient 15. The accuracy of the prediction is 15/16=0.9375.

Needles to say that drawing and manipulating trees is not particularly practical way of constructing decision trees. Here they are used to enhance our understanding

chest pain	blocked arteries	weight	leaf node
no	no	low	
no	no	high	[3,y],[4,y]
no	yes	low	[9,n],[10,n],[11,n],[12,n],[16,n]
no	yes	high	
yes	no	low	[13,n],[14,n]
yes	no	high	[6,y]
yes	yes	low	[7,y],[8,y],[15,n]
yes	yes	high	[1,y],[2,y],[5,y]

The construction of the complete tree corresponding to the training data set

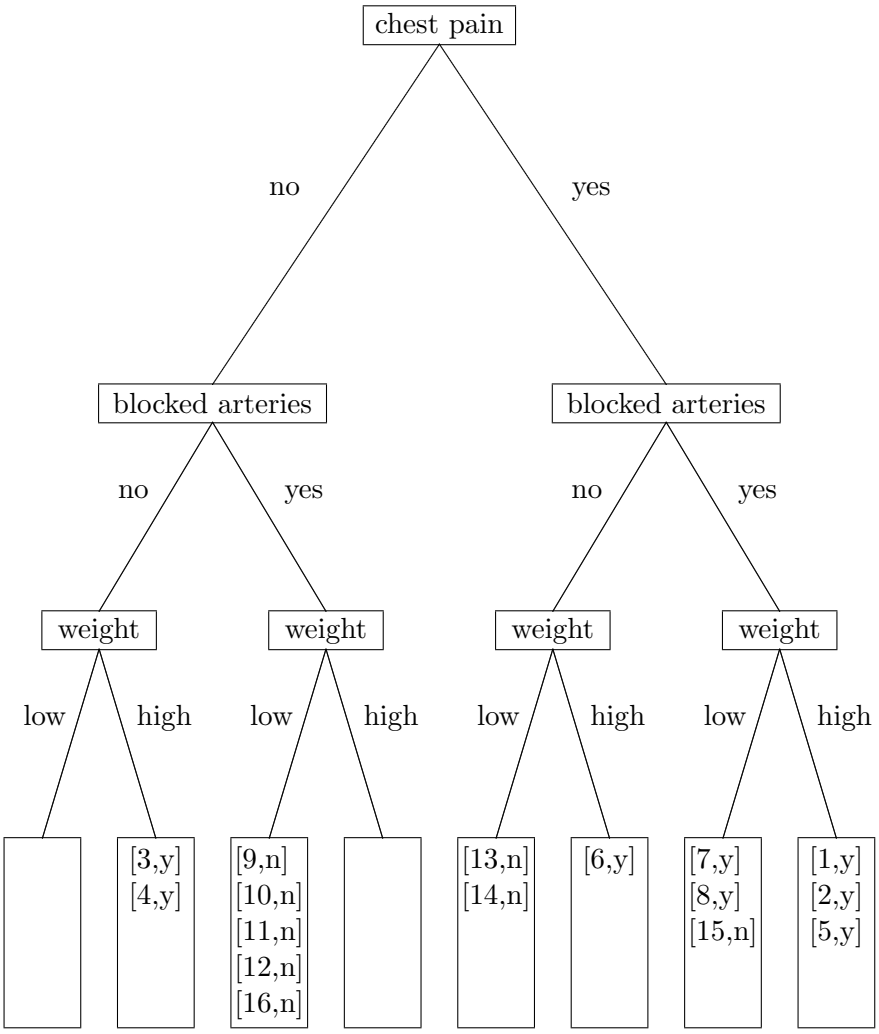


Figure A.1: A complete binary tree representing the training data set.

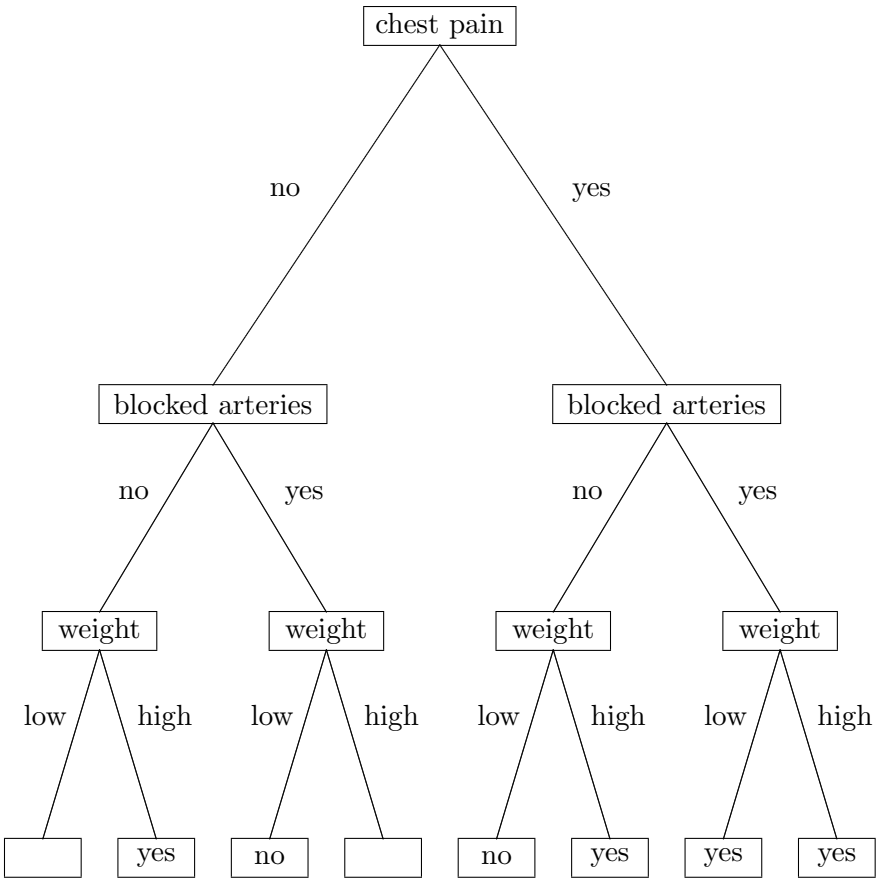


Figure A.2: The complete binary tree serving as a decision tree.

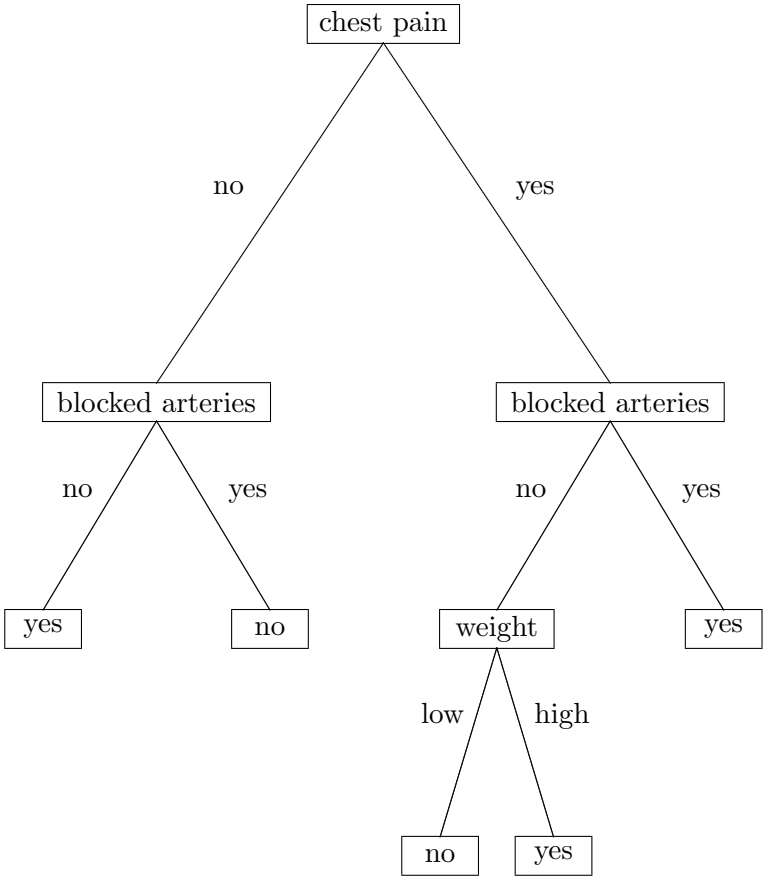


Figure A.3: The decision tree after tidying up.

essentially depends on the way the explanatory variables are listed. In our case there are three independent variables and these variables can be ordered in six ways. Therefore there are six possibilities to construct complete trees. From the six complete trees we can construct six decision trees. It may happen that some of these decision trees are more advantageous from some point of view than the others. Our small scale numerical experiment will show that ordering of the variables "weight","chest pain","blocked arteries" provides a smaller decision three then the ordering "chest pain","blocked arteries","weight".

weight	chest pain	blocked arteries	leaf node
low	no	no	
low	no	yes	[9,n],[10,n],[11,n],[12,n],[16,n]
low	yes	no	[13,n],[14,n]
low	yes	yes	[7,y],[8,y],[15,n]
high	no	no	[3,y],[4,y]
high	no	yes	
high	yes	no	[6,y]
high	yes	yes	[1,y],[2,y],[5,y]

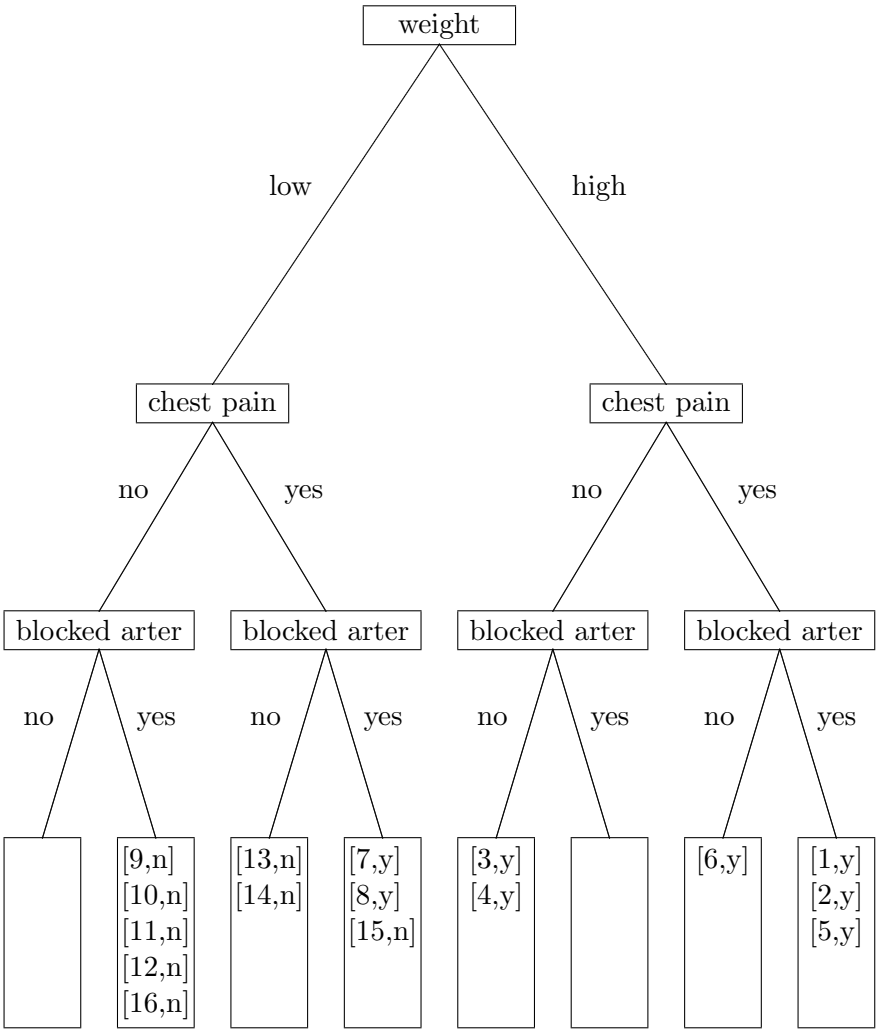


Figure A.4: A complete binary tree.

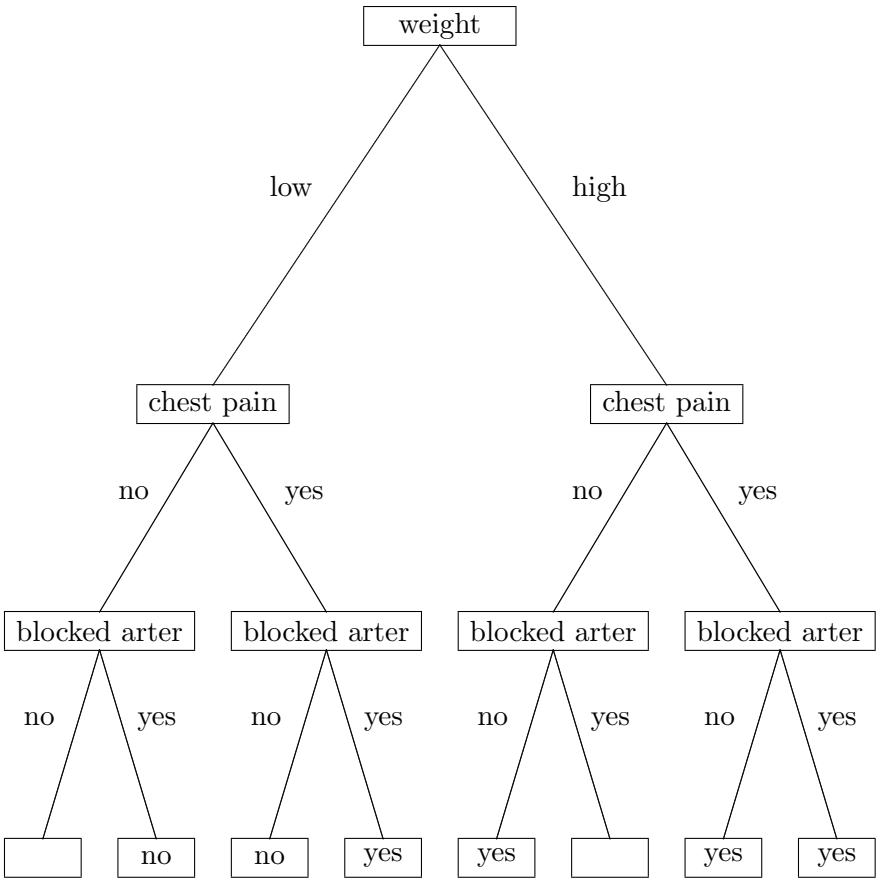


Figure A.5: A complete binary tree acting as a decision tree.

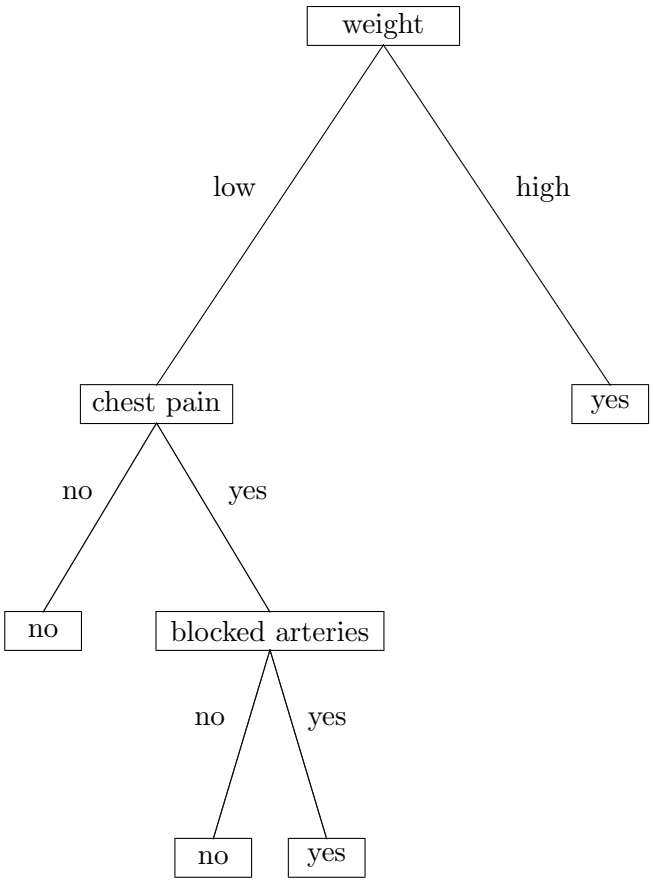


Figure A.6: The tidied up decision tree.

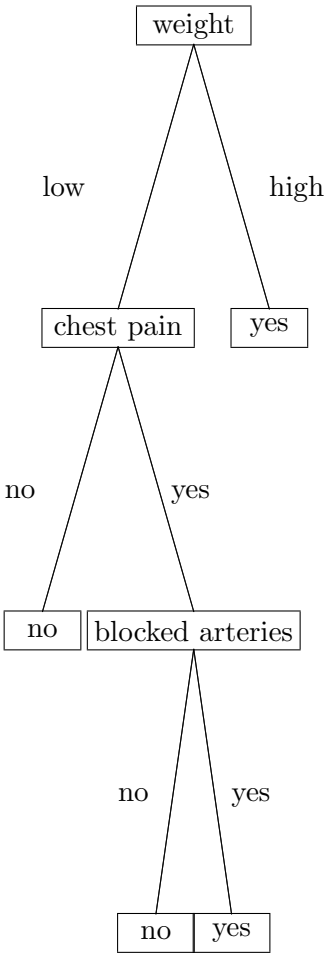


Figure A.7: The decision tree after redesigning.

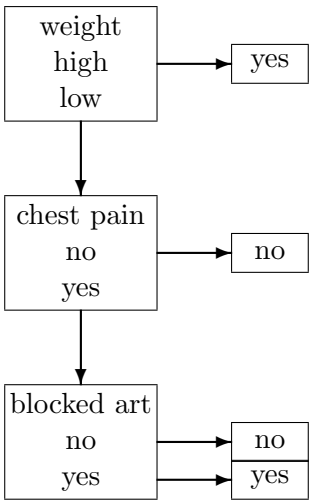


Figure A.8: A further possible representation of the same decision tree.

Appendix B

Grouping variables

B.1 Predictive power of a group of variables

Let us consider the training data set from Chapter 11 section 11.1.

patient number	chest pain	blocked arteries	patient weight	heart disease
1	yes	yes	high	yes
2	yes	yes	high	yes
3	no	no	high	yes
4	no	no	high	yes
5	yes	yes	high	yes
6	yes	no	high	yes
7	yes	yes	low	yes
8	yes	yes	low	yes
9	no	yes	low	no
10	no	yes	low	no
11	no	yes	low	no
12	no	yes	low	no
13	yes	no	low	no
14	yes	no	low	no
15	yes	yes	low	no
16	no	yes	low	no

We will show how to compute the predictive power of a group of explanatory variables with respect to the target variable. Let us see an example. Let

$$\text{pp}[\text{chest pain}, \text{blocked arteries}|\text{heart disease}]$$

be denote the predictive power of the variables "chest pain" and "blocked arteries" with respect to the variable "heart disease". Here is a way to compute this predictive power. First we delete the column "patient weight" and

then rearrange the rows of the table into a lexicographically non-decreasing sequence of the rows of the table.

patient number	chest pain	blocked arteries	heart disease
3	no	no	yes
4	no	no	yes
9	no	yes	no
10	no	yes	no
11	no	yes	no
12	no	yes	no
16	no	yes	no
6	yes	no	yes
13	yes	no	no
14	yes	no	no
15	yes	yes	no
1	yes	yes	yes
2	yes	yes	yes
5	yes	yes	yes
7	yes	yes	yes
8	yes	yes	yes

This table with the rearranged rows provides the following prediction rule.

chest pain	blocked arteries	prediction
no	no	yes
no	yes	no
yes	no	no
yes	yes	yes

The health status of patients 6 and 15 is predicted incorrectly. All the other patients health status is predicted correctly. Therefore,

$$\text{pp}[\text{chest pain}, \text{blocked arteries}|\text{heart disease}] = 14/16 = 0.875.$$

One can construct the confusion matrix and compute various prediction measures. It is not essential that the variables are binary. When we work with a regression tree, that is, when the target variable is continuous, then the way of constructing the prediction rule changes. Say designating the subgroups average (median, midrange) as prediction. The variance (average, median, midrange) of the deviations can play the role of the uncertainty of the prediction of the variables involved.

Appendix C

Grouping variables

Let us consider the following training data set. We would like to predict the "result" target variable using the explanatory variables "assessment", "assignment", "project". The explanatory variables are discrete and the target variable is continuous.

student number	assessment	assignment	project	result
1	good	yes	yes	95
2	average	yes	no	70
3	good	no	yes	75
4	poor	no	no	45
5	good	yes	yes	98
6	average	no	yes	80
7	good	no	no	75
8	poor	yes	yes	65
9	average	no	no	58
10	good	yes	yes	89

We may form eight groups consisting of explanatory variables. We compute the predictive power of these groups. In fact, we will compute the uncertainties of the predictions and we remember that the smaller is the uncertainty the larger is the predictive power.

variables	uncertainty
none	12.4
assessment	8.8
assignment	12.4
project	10.4
assessment,assignment	3.2
assessment,project	8.8
assignment,project	7.7
assessment,assignment,project	1.0

C.1 Constructing trees

The are three independent (explanatory) variables. We consider the possible subsets of these variables. We construct the complete tree in connection with each subset.

assess-ment	assign-ment	project	
good	no	no	[7,75]
godd	no	yes	[3,75]
good	yes	no	
good	yes	yes	[1,95],[5,98],[10,89]
average	no	no	[9,58]
average	no	yes	[6,80]
average	yes	no	[2,70]
average	yes	yes	
poor	no	no	[4,45]
poor	no	yes	
poor	yes	no	
poor	yes	yes	[8,65]

assess-ment	assign-ment	
good	no	[7,75],[3,75]
good	yes	[1,95],[5,98],[10,89]
average	no	[9,58],[6,80]
average	yes	[2,70]
poor	no	[4,45]
poor	yes	[8,65]

assess-ment	project	
good	no	[7,75]
good	yes	[3,75],[1,95],[5,98],[10,89]
average	no	[2,70],[9,58]
average	yes	[6,80]
poor	no	[4,45]
poor	yes	[8,65]

assign-ment	project	
no	no	[7,75],[9,58],[4,45]
no	yes	[3,75],[6,80]
yes	no	[2,70]
yes	yes	[1,95],[5,98],[10,89],[8,65]

assess-ment	
good	[7,75],[3,75],[1,95],[5,98],[10,89]
average	[2,70],[6,80],[9,58]
poor	[4,45],[8,65]

assign-ment	
no	[7,75],[9,58],[4,45],[3,75],[6,80]
yes	[2,70],[1,95],[5,98],[10,89],[8,65]

project	
no	[7,75],[9,58],[4,45],[2,70]
yes	[3,75],[6,80],[1,95],[5,98],[10,89],[8,65]

C.2 Computing the uncertainties

student number	result actual	result predicted	deviation 12.4
1	95	75	20
2	70	75	5
3	75	75	0
4	45	75	30
5	98	75	23
6	80	75	5
7	75	75	0
8	65	75	10
9	58	75	17
10	89	75	14

student number	project	result actual	result predicted	deviation 10.4
2	no	70	62.0	8.0
4	no	45	62.0	17.0
7	no	75	62.0	13.0
9	no	58	62.0	4.0
1	yes	95	83.7	11.3
3	yes	75	83.7	8.7
5	yes	98	83.7	14.3
6	yes	80	83.7	3.7
8	yes	65	83.7	18.7
10	yes	89	83.7	5.3

student number	assign-ment	result actual	result predicted	deviation 12.4
3	no	75	66.7	8.3
4	no	45	66.7	21.7
6	no	80	66.7	13.3
7	no	75	66.7	8.3
9	no	58	66.7	8.7
1	yes	95	83.4	11.6
2	yes	70	83.4	13.4
5	yes	98	83.4	14.7
8	yes	65	83.4	18.4
10	yes	89	83.4	5.6

student number	assess-ment	result actual	result predicted	deviation 8.8
1	good	95	86.4	8.6
3	good	75	86.4	11.4
5	good	98	86.4	11.6
7	good	75	86.4	11.4
10	good	89	86.4	2.7
2	average	70	69.3	0.7
6	average	80	69.3	10.7
9	average	58	69.3	11.3
4	poor	45	55.0	10.0
8	poor	65	55.0	10.0

student number	assign-ment	project	result actual	result predicted	deviation 7.7
4	no	no	45	59.3	14.3
7	no	no	75	59.3	15.7
9	no	no	58	59.3	1.3
3	no	yes	75	77.5	2.5
6	no	yes	80	77.5	2.5
2	yes	no	70	70.0	0.0
1	yes	yes	95	86.8	8.2
5	yes	yes	98	86.8	11.2
8	yes	yes	65	86.8	21.8
10	yes	yes	89	86.8	2.2

student number	assess-ment	project	result actual	result predicted	deviation 4.1
7	good	no	75	75.0	0.0
1	good	yes	95	89.3	5.7
3	good	yes	75	89.3	14.3
5	good	yes	98	89.3	8.7
10	good	yes	89	89.3	0.3
2	average	no	70	64.0	6.0
9	average	no	58	64.0	6.0
6	average	yes	80	80.0	0.0
4	poor	no	45	45.0	0.0
8	poor	yes	65	65.0	0.0

student number	assess-ment	assign-ment	result actual	result predicted	deviation 3.2
3	good	no	75	75	0
7	good	no	75	75	0
1	good	yes	95	94	1
5	good	yes	98	94	4
10	good	yes	89	94	5
6	average	no	80	69	11
9	average	no	58	69	11
2	average	yes	70	70	0
4	poor	no	45	45	0
8	poor	yes	65	65	0

student number	assess-ment	assign-ment	project	result actual	result predicted	deviation 1.0
7	good	no	no	75	75	0
3	good	no	yes	75	75	0
1	good	yes	yes	95	94	1
5	good	yes	yes	98	94	4
10	good	yes	yes	89	94	5
9	average	no	no	58	58	0
6	average	no	yes	80	80	0
2	average	yes	no	70	70	0
4	poor	no	no	45	45	0
8	poor	yes	yes	65	65	0

Appendix D

Mahalanobis distance

D.1 Discretization via z-score

Let us consider the following toy size training data set. The explanatory variables CRP and PCT are continuous. The target variable "infection" is binary. The problem is a classification problem. Figure D.1 displays the training data set in a scatter plot form.

patient number	CRP (mg/L)	PCT (μg/L)	infection type
1	42	33	viral
2	57	29	viral
3	38	39	viral
4	43	40	viral
5	30	45	viral
6	58	39	viral
7	61	67	bacterial
8	70	60	bacterial
9	55	62	bacterial
10	74	75	bacterial
11	56	55	bacterial
12	72	76	bacterial

One way to handle such problem is to discretise the continuous variables and use the machinery of the classification trees. This time we do not discretise the continuous variables. We will modify the construction of decision rules.

The mean of the CRP reading of the patients with viral infection is 44.7. The mean of the CRP reading of the patients with bacterial infection is 64.7.

If a patient CRP reading is closer to 44.7 than to 64.7, then we predict viral infection. If the patient CRP reading is closer to 64.7 than to 44.7, then we predict bacterial infection.

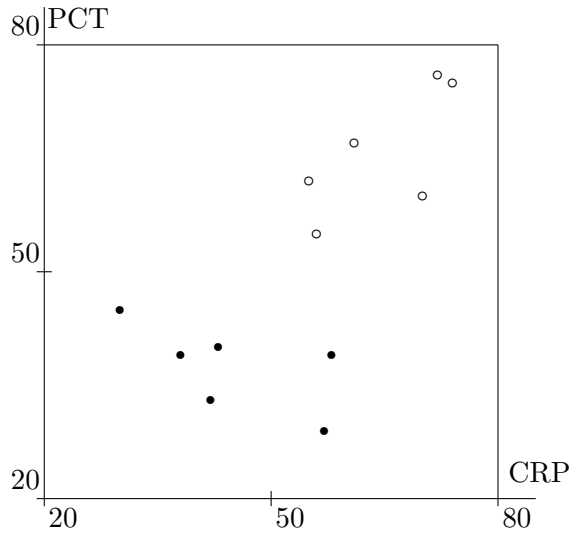


Figure D.1: A scatter plot of the training data set.

The standard deviation of the CRP reading of the patients with viral infection is 9.4. The standard deviation of the CRP reading of the patients with bacterial infection is 7.6.

The Mahalanobis distance of patient 1 from the viral group is equal to $(44.7 - 42)/10 = 0.27$. The Mahalanobis distance of patient 1 from the bacterial group is equal to $(64.7 - 42)/7.6 = 2.99$. The patient is closer to the viral group than to the bacterial group. Our prediction is that patient 1 belongs to the viral group.

We compute the predictive power of the explanatory variable PCT with respect to the target variable "infection". As a first step we list the patient's

Mahalanobis distances from the viral group.

patient number	CRP (mg/L)	viral group mean	deviation	viral group std	distance
1	42	44.7	2.7	10	0.270
2	57	44.7	12.3	10	1.230
3	38	44.7	6.7	10	0.670
4	43	44.7	1.7	10	0.170
5	30	44.7	14.7	10	1.470
6	58	44.7	13.3	10	1.330
7	61	44.7	16.3	10	1.630
8	70	44.7	25.3	10	2.530
9	55	44.7	10.3	10	1.030
10	74	44.7	29.3	10	2.930
11	56	44.7	11.3	10	1.130
12	72	44.7	27.3	10	2.730

We list the patient's Mahalanobis distances from the bacterial group.

patient number	CRP (mg/L)	bacterial group mean	deviation	bacterial group std	distance
1	42	64.7	22.7	7.7	2.948
2	57	64.7	7.7	7.7	1.000
3	38	64.7	26.7	7.7	3.468
4	43	64.7	21.7	7.7	2.818
5	30	64.7	34.7	7.7	4.506
6	58	64.7	6.7	7.7	0.870
7	61	64.7	3.7	7.7	0.481
8	70	64.7	5.3	7.7	0.688
9	55	64.7	9.7	7.7	1.260
10	74	64.7	9.3	7.7	1.208
11	56	64.7	8.7	7.7	1.129
12	72	64.7	7.3	7.7	0.948

We use the Mahalanobis distances to make a prediction.

patient number number	distance from viral	distance from bacterial	infection type predicted	infection type actual
1	0.270	2.948	viral	viral
2	1.230	1.000	bacterial	viral
3	0.670	3.468	viral	viral
4	0.170	2.818	viral	viral
5	1.470	4.506	viral	viral
6	1.330	0.870	bacterial	viral
7	1.630	0.481	bacterial	bacterial
8	2.530	0.688	bacterial	bacterial
9	1.030	1.260	viral	bacterial
10	2.930	1.208	bacterial	bacterial
11	1.130	1.129	bacterial	bacterial
12	2.730	0.948	bacterial	bacterial

The confusion matrix.

		predicted viral	predicted bacterial
actual	viral	4	2
actual	bacterial	1	5

The accuracy is equal to (4+5)/12=0.750.

The false positive rate is equal to (2)/(4+2)=0.333.

The true positive rate is equal to (5)/(1+5)=0.833.

We compute the predictive power of the explanatory variable PCT with respect to the target variable “infection”. As a first step we list the patient’s

Mahalanobis distances from the viral group.

patient number	PCT (μg/L)	viral group mean	deviation	viral group std	distance
1	33	37.5	4.5	5.2	0.9
2	29	37.5	8.5	5.2	1.6
3	39	37.5	1.5	5.2	0.3
4	40	37.5	2.5	5.2	0.48
5	45	37.5	7.5	5.2	1.4
6	39	37.5	1.5	5.2	0.3
7	67	37.5	29.5	5.2	5.7
8	60	37.5	22.5	5.2	4.3
9	62	37.5	24.5	5.2	4.7
10	75	37.5	37.5	5.2	7.2
11	55	37.5	17.5	5.2	3.4
12	76	37.5	38.5	5.2	7.4

Next we list the patient's Mahalanobis distances from the bacterial group.

patient number	PCT (μg/L)	bacterial group mean	deviation	bacterial group std	distance
1	33	65.8	32.8	7.7	4.3
2	29	65.8	36.8	7.7	4.8
3	39	65.8	26.8	7.7	3.5
4	40	65.8	25.8	7.7	3.4
5	45	65.8	20.8	7.7	2.7
6	39	65.8	26.8	7.7	3.5
7	67	65.8	1.2	7.7	0.2
8	60	65.8	5.8	7.7	0.8
9	62	65.8	3.8	7.7	0.5
10	75	65.8	9.2	7.7	1.2
11	55	65.8	10.8	7.7	1.4
12	76	65.8	10.2	7.7	1.3

Using the Mahalanobis distances we make predictions.

patient number	distance from viral	distance from bacterial	infection type predicted	infection type actual
1	0.9	4.3	viral	viral
2	1.6	4.8	viral	viral
3	0.3	3.5	viral	viral
4	0.5	3.4	viral	viral
5	1.4	2.7	viral	viral
6	0.3	3.5	viral	viral
7	5.7	0.2	bacterial	bacterial
8	4.3	0.8	bacterial	bacterial
9	4.7	0.5	bacterial	bacterial
10	7.2	1.2	bacterial	bacterial
11	3.4	1.4	bacterial	bacterial
12	7.4	1.3	bacterial	bacterial

The confusion matrix.

		predicted viral	predicted bacterial
actual	viral	6	0
actual	bacterial	0	6

The accuracy is equal to (6+6)/12=1.000.

The false positive rate is equal to (0)/(6+0)=0.000.

The true positive rate is equal to (6)/(0+6)=1.000.

D.2 The standard discretisation

We discretise the CRP explanatory variable.

patient number	CRP (mg/L)	infection type
5	30	viral
3	38	viral
1	42	viral
4	43	viral
9	55	bacterial
11	56	bacterial
2	57	viral
6	58	viral
7	61	bacterial
8	70	bacterial
12	72	bacterial
10	74	bacterial

There are three sign changes. Namely, they occur at the cut points 49, 56.5, 59.5.

Let us inspect the cut point 49.

	CRP low (yes)	CRP high (no)
viral infection no	(0)	7,8,9,10,11,12 (6)
viral infection yes	1,3,4,5 (4)	2,6 (2)

	predicted no	predicted yes
actual no	6	0
actual yes	2	4

The accuracy is (6+4)/12=0.833.

The false positive rate is 6/(6+0)=1.000.

The true positive rate is 4/(2+4)=0.333.

We check the cut point 56.5.

	CRP low (yes)	CRP high (no)
viral infection no	9,11 (2)	7,8,10,12 (4)
viral infection yes	1,3,4,5 (4)	2,6 (2)

	predicted no	predicted yes
actual no	4	2
actual yes	2	4

The accuracy is (4+4)/12=0.666.
The false positive rate is 2/(4+2)=0.333.
The true positive rate is 4/(2+4)=0.666.
Finally, we inspect the cut point 59.5.

	CRP low (yes)	CRP high (no)
viral infection no	9,11 (2)	7,8,10,12 (4)
viral infection yes	1,2,3,4,5,6 (6)	(0)

	predicted no	predicted yes
actual no	6	2
actual yes	0	4

The accuracy is (6+4)/12=0.833.
The false positive rate is 2/(6+2)=0.250.
The true positive rate is 4/(0+4)=1.000.
We discretise the PCT explanatory variable.

patient number	PCT (μg/L)	infection type
2	29	viral
1	33	viral
3	39	viral
6	39	viral
4	40	viral
5	45	viral
11	55	bacterial
8	60	bacterial
9	62	bacterial
7	67	bacterial
10	75	bacterial
12	76	bacterial

There is only one sign change at the cut point 50. Let us evaluate this cut point 50.

	CRP low (yes)	CRP high (no)
viral infection no	(0)	7,8,9,10,11,12 (6)
viral infection yes	1,2,3,4,5,6 (6)	(0)

	predicted no	predicted yes
actual no	6	0
actual yes	0	6

The accuracy is (6+6)/12=1.000.
The false positive rate is 0/(6+0)=0.000.
The true positive rate is 6/(0+6)=1.000.

Appendix E

Discretization

E.1 Mahalanobis distance

Let us consider the following toy size training data set. The explanatory variables "chest pain" and "blocked arteries" and the target variable "heart disease" are binary. The explanatory variable "patient weight" is continuous. The problem is a classification problem.

patient number	chest pain	blocked arteries	patient weight	heart disease
1	yes	yes	205	yes
2	yes	yes	185	yes
3	no	no	180	yes
4	no	no	182	yes
5	yes	yes	210	yes
6	yes	no	200	yes
7	yes	yes	167	yes
8	yes	yes	170	yes
9	no	yes	156	no
10	no	yes	160	no
11	no	yes	125	no
12	no	yes	110	no
13	yes	no	168	no
14	yes	no	165	no
15	yes	yes	172	no
16	no	yes	175	no

We have already seen techniques to discretise the continuous variables explanatory variables. We will present here yet another way.

We compute the predictive power of the explanatory variable PCT with respect to the target variable "infection". As a first step we list the patients's Mahalanobis distances from the healthy group. (Prasanta Chandra

Mahalanobis Indian statistician)

patient number	patient weight	healthy group mean	deviation	healthy group std	distance
1	205	153.9	51.1	22.1	2.3
2	185	153.9	31.1	22.1	1.4
3	180	153.9	26.1	22.1	1.2
4	182	153.9	28.1	22.1	1.3
5	210	153.9	56.1	22.1	2.5
6	200	153.9	46.1	22.1	2.1
7	167	153.9	13.9	22.1	0.6
8	170	153.9	16.1	22.1	0.7
9	156	153.9	2.1	22.1	0.1
10	160	153.9	6.1	22.1	0.3
11	125	153.9	28.9	22.1	1.3
12	110	153.9	43.9	22.1	2.0
13	168	153.9	14.1	22.1	0.6
14	165	153.9	11.1	22.1	0.5
15	172	153.9	18.1	22.1	0.8
16	175	153.9	21.1	22.1	1.0

We list the patients's Mahalanobis distances from the sick group.

patient number	patient weight	sick group mean	deviation	sick group std	distance
1	205	187.4	17.6	16.5	1.1
2	185	187.4	2.4	16.5	0.1
3	180	187.4	7.4	16.5	0.4
4	182	187.4	5.4	16.5	0.3
5	210	187.4	22.6	16.5	1.4
6	200	187.4	12.6	16.5	0.8
7	167	187.4	20.4	16.5	1.2
8	170	187.4	17.4	16.5	1.1
9	156	187.4	31.4	16.5	1.9
10	160	187.4	27.4	16.5	1.7
11	125	187.4	62.4	16.5	3.8
12	110	187.4	77.4	16.5	4.7
13	168	187.4	19.4	16.5	1.2
14	165	187.4	22.4	16.5	1.4
15	172	187.4	15.4	16.5	0.9
16	175	187.4	12.4	16.5	0.8

We use the Mahalanobis distancies to make a prediction.

patient number number	distance from healthy	distance from sick	heart disease predicted	heart disease actual
1	2.3	1.1	yes	yes
2	1.4	0.1	yes	yes
3	1.2	0.4	yes	yes
4	1.3	0.3	yes	yes
5	2.5	1.4	yes	yes
6	2.1	0.8	yes	yes
7	0.6	1.2	no	yes
8	0.7	1.1	no	yes
9	0.1	1.9	no	no
10	0.3	1.7	no	no
11	1.3	3.8	no	no
12	2.0	4.7	no	no
13	0.6	1.2	no	no
14	0.5	1.4	no	no
15	0.8	0.9	no	no
16	1.0	0.8	yes	no

The confusion matrix.

	predicted no	predicted yes
actual no	7	1
actual yes	2	6

The accuracy is equal to (7+6)/16=0.813.
The false positive rate is equal to (1)/(7+1)=0.125.
The true positive rate is equal to (6)/(2+6)=0.750.

Appendix F

Relative frequences

F.1 Empirical cumulativ probabilities

Let us consider the following toy size training data set. The explanatory variables "chest pain" and "blocked arteries" and the target variable "heart disease" are binary. The explanatory variable "patient weight" is continuous. The problem is a classification problem.

patient number	chest pain	blocked arteries	patient weight	heart disease
1	yes	yes	205	yes
2	yes	yes	185	yes
3	no	no	180	yes
4	no	no	182	yes
5	yes	yes	210	yes
6	yes	no	200	yes
7	yes	yes	167	yes
8	yes	yes	170	yes
9	no	yes	156	no
10	no	yes	160	no
11	no	yes	125	no
12	no	yes	110	no
13	yes	no	168	no
14	yes	no	165	no
15	yes	yes	172	no
16	no	yes	175	no

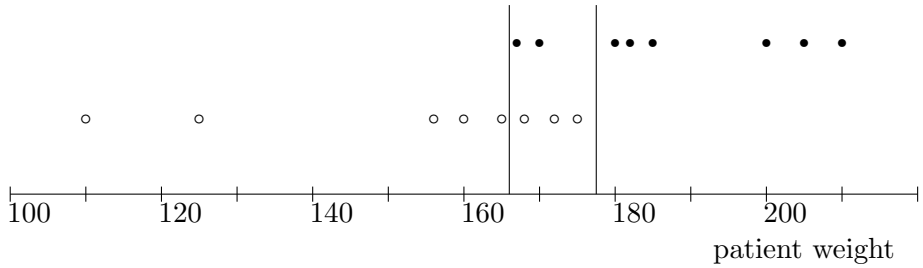


Figure F.1: The pure node heuristic. The cut points are 166 and 177.5.

We have already seen techniques to discretise the continuous variables explanatory variables. We will present here yet another way.

patient number	patient weight	cummulative probability
	175	1/8
	172	2/8
	168	3/8
	165	4/8
	160	5/8
	156	6/8
	125	7/8
	110	8/8

patient number	patient weight	cummulative probability
	167	1/8
	170	2/8
	180	3/8
	182	4/8
	185	5/8
	200	6/8
	205	7/8
	210	8/8

The largest weight in the healthy group is 175. The smallest weight in the sick group is 167. The cut point must be between 167 and 175. The empirical probability the event $\{172 \leq \text{weight} \leq 175\}$ is equal to 1/8. The empirical probability the event $\{167 \leq \text{weight} \leq 170\}$ is equal to 1/8. The cut point must be between 170 and 172.

F.2 Neyman-Pearson lemma

F.3 Missing data entries

F.4 Using a variable more than once

I think this is relevant when the variables are continuous.

Appendix G

Training data sets

G.1 Collection of training data sets

day	outlook	temperature	humidity	wind	play tennis
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rain	mild	high	weak	yes
5	rain	cool	normal	weak	yes
6	rain	cool	normal	strong	no
7	overcast	cool	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rain	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rain	mild	high	strong	no

day	outlook	temperature	humidity	wind	play time
1	sunny	hot	high	weak	25
2	sunny	hot	high	strong	30
3	overcast	hot	high	weak	46
4	rain	mild	high	weak	45
5	rain	cool	normal	weak	52
6	rain	cool	normal	strong	23
7	overcast	cool	normal	strong	43
8	sunny	mild	high	weak	35
9	sunny	cool	normal	weak	38
10	rain	mild	normal	weak	46
11	sunny	mild	normal	strong	48
12	overcast	mild	high	strong	52
13	overcast	hot	normal	weak	44
14	rain	mild	high	strong	30

customer number	gender	car type	shirt size	class type
1	male	family	S	A
2	male	sport	M	A
3	male	sport	M	A
4	male	sport	L	A
5	male	sport	XL	A
6	male	sport	XL	A
7	female	sport	S	A
8	female	sport	S	A
9	female	sport	M	A
10	female	luxury	L	A
11	male	family	L	B
12	male	family	XL	B
13	male	family	M	B
14	male	luxury	XL	B
15	female	luxury	S	B
16	female	luxury	S	B
17	female	luxury	M	B
18	female	luxury	M	B
19	female	luxury	M	B
20	female	luxury	L	B

S=small, M=medium, L=large, XL=extra large

customer number	age	job	home	credit	loan approved
1	young	false	no	fair	no
2	young	false	no	good	no
3	young	true	no	good	yes
4	young	true	yes	fair	yes
5	young	false	no	fair	no
6	middle	false	no	fair	no
7	middle	false	no	good	no
8	middle	true	yes	good	yes
9	middle	false	yes	excellent	yes
10	middle	false	yes	excellent	yes
11	old	false	yes	excellent	yes
12	old	false	yes	good	yes
13	old	true	no	good	yes
14	old	true	no	excellent	yes
15	old	false	no	fair	no

customer number	job	salary	max loan amount
1	no	10	10
2	no	11	20
3	yes	12	11
4	yes	30	15
5	no	29	40
6	no	22	11
7	no	11	23
8	yes	5	34
9	no	11	24
10	no	15	12
11	no	32	60
12	no	33	11
13	yes	22	22
14	yes	11	21
16	no	19	11

student number	assessment	assignment	project	result
1	good	yes	yes	95
2	average	yes	no	70
3	good	no	yes	75
4	poor	no	no	45
5	good	yes	yes	98
6	average	no	yes	80
7	good	no	no	75
8	poor	yes	yes	65
9	average	no	no	58
10	good	yes	yes	89

patient number	CRP (mg/L)	PCT (μg/L)	infection type
1	42	33	viral
2	57	29	viral
3	38	39	viral
4	43	40	viral
5	30	45	viral
6	58	39	viral
7	61	67	bacterial
8	70	60	bacterial
9	55	62	bacterial
10	74	75	bacterial
11	56	55	bacterial
12	72	76	bacterial

record number	x	y	type
1	0.111	0.312	A
2	0.210	0.105	A
3	0.103	0.150	A
4	0.241	0.214	A
5	0.611	0.251	A
6	0.720	0.056	A
7	0.905	0.897	A
8	0.702	0.791	A
9	0.750	0.510	A
10	0.951	0.618	A
11	0.331	0.200	B
12	0.261	0.455	B
13	0.429	0.971	B
14	0.025	0.950	B
15	0.101	0.798	B
16	0.345	0.591	B
17	0.250	0.750	B
18	0.200	0.800	B
19	0.490	0.770	B
20	0.470	0.960	B

patient number	chest pain	blocked arteries	patient weight	heart disease
1	yes	yes	205	yes
2	yes	yes	185	yes
3	no	no	180	yes
4	no	no	182	yes
5	yes	yes	210	yes
6	yes	no	200	yes
7	yes	yes	167	yes
8	yes	yes	170	yes
9	no	yes	156	no
10	no	yes	160	no
11	no	yes	125	no
12	no	yes	110	no
13	yes	no	168	no
14	yes	no	165	no
15	yes	yes	172	no
16	no	yes	175	no

record number	drug dosage (mg)	drug effectiveness (%)
1	2	2
2	26	85
3	4	1
4	24	87
5	22	90
6	6	3
7	8	4
8	20	100
9	18	98
10	10	1
11	12	4
12	17	95
13	38	3
14	13	6
15	14	5
16	36	2
17	34	2
18	15	20
19	16	24
20	30	20
21	32	18

Appendix H

Deterministic learning

H.1 Probabilistic and deterministic learning

The subject of this section can be illuminated with the following problem. We would like to separate a basket of mushrooms we collected in the forest. It is clear that the poisonous and edible mushrooms should be separated with absolute certainty. We are not willing to take any chances. The mushrooms have a large number of attributes. We collect a number of these attributes which make possible to classify mushrooms into edible and poisonous groups. This is a typical case of deterministic learning.

One can view a deterministic decision tree as a way to convert the knowledge given in table form to the form of a tree. This is just a transformation of information from one form to another. How this transformation can be useful? The tree form can be more concise and so more usable than the table form.

Here is further possibility to exploit deterministic decision trees. In the probabilistic case even if the accuracy of the prediction is very high meeting

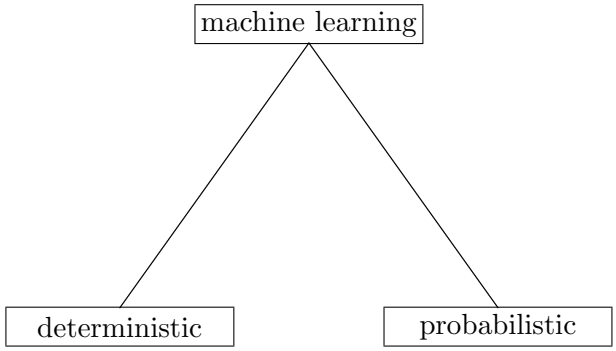


Figure H.1: The deterministic and probabilistic learning dichotomy.

Table H.1: The training data set in table form

		normal	normal	high	high
		weak	strong	weak	strong
rain	cool	[5,y]	[6,n]		
rain	mild	[10,y]		[4,y]	[14,n]
rain	hot				
overcast	cool		[7,y]		
overcast	mild				[12,y]
overcast	hot	[13,y]		[3,y]	
sunny	cool	[9,y]			
sunny	mild		[11,y]	[8,n]	
sunny	hot			[1,n]	[2,n]

with a particular record we cannot be sure if the classification is correct or not. In the deterministic case we may try to arrange things such that when the tree predicts class A then the record is actually from class A and when the tree predicts class B then the record is actually belongs to class B. We should keep open the possibility, that the tree is not able arrive to a definite conclusion. In this case it should classify the record as inconclusive.

If one is able to construct such decision trees, then it has many possible uses. The easily identifiable records can be sorted into classes A or B. The inconclusive cases are left for further checking. If a large portion of the records is easily classifiable, then just a small part of the records requires further checking.

Appendix I

Graphical method

I.1 A graphical method

Let us consider the following small size toy training data set. Here x,y are continuous explanatory variables and the target variable "type" is binary. We describe a graphical method to construct a decision tree for this classification problem.

Each record can be represented by a point using the values of the variables x and y as coordinates. Type A records are plotted as bullets and type B records are plotted as circles. The plotted point are all lying in a square. The vertical line whose equation is $x = 0.550$ cuts the square into two rectangles. The right hand side rectangle contains purely bullets. We express this fact saying that the right hand side rectangle is a pure subset. The horizontal line whose equation is $y = 0.350$ cuts the left hand side rectangles into two parts. The upper side rectangle is pure. With the vertical line whose equation is $x = 0.280$ we divide lower rectangle into two parts. Both of these rectangles are pure. The procedure we described can be followed on Figure I.1. The classification tree is in Figure I.2.

The guiding principles of dividing the training data set are typically greedy considerations. We may try to find a horizontal or vertical cutting lines to produce at least one pure subset with as large cardinality as possible. We stop dividing a subset when it is getting overly small. We call the procedure a graphical procedure because we spot the cutting lines by inspecting a picture depicted in Figure I.1. In this situation the name graphical method is really justified. Of course, the procedure can be mechanized and can be carried out using a computer.

We still can use this graphical method even if there are more than two continuous explanatory variables. For the sake of definiteness suppose that there are four continuous explanatory variables. We can construct a decision tree for each pairs of the continuous explanatory variables. We will end up with six decision trees. Any of these trees can be included into a committee

if we decide to use the ensemble method.

record number	x	y	type
1	0.111	0.312	A
2	0.210	0.105	A
3	0.103	0.150	A
4	0.241	0.214	A
5	0.611	0.251	A
6	0.720	0.056	A
7	0.905	0.897	A
8	0.702	0.791	A
9	0.750	0.510	A
10	0.951	0.618	A
11	0.331	0.200	B
12	0.261	0.455	B
13	0.429	0.971	B
14	0.025	0.950	B
15	0.101	0.798	B
16	0.345	0.591	B
17	0.250	0.750	B
18	0.200	0.800	B
19	0.490	0.770	B
20	0.470	0.960	B

From the variables x and y we may construct a new variable z. The values of z will be A$'$ and B$'$. Namely, when the classification tree above reaches the leaf node labeled by A, then z takes on the value A$'$ and when the tree reaches the leaf node labeled by B, then the variable z takes on the value B$'$.

In this way a visual inspection can help to construct new explanatory variables for each pair of the earlier explanatory variables.

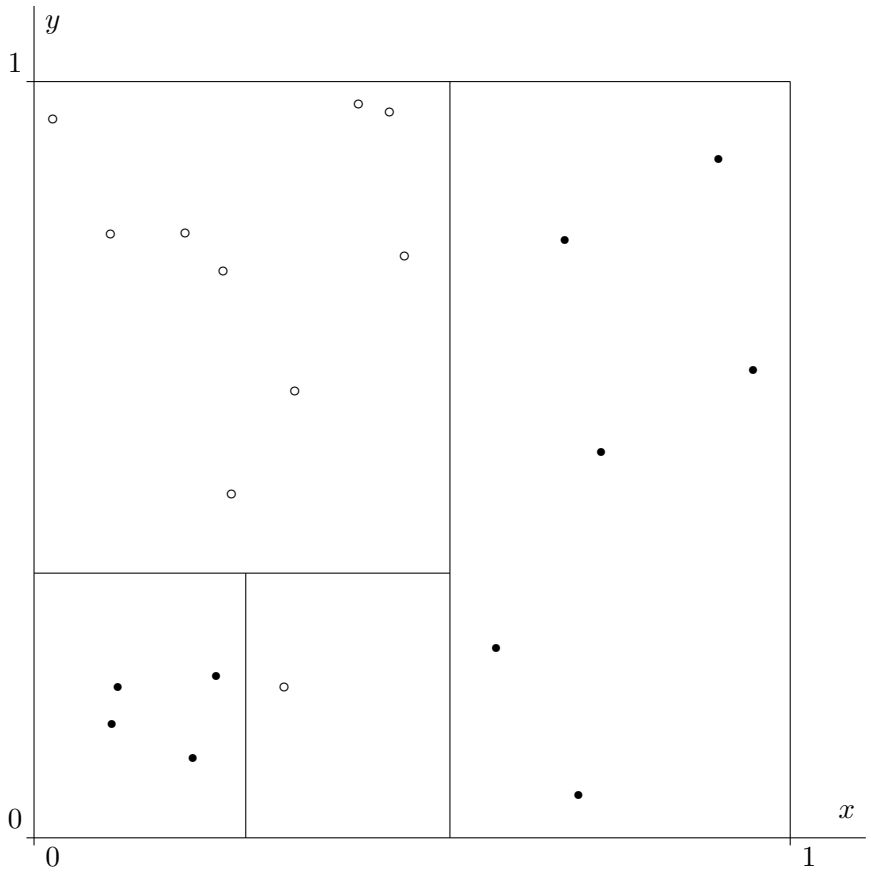


Figure I.1: The rectangles representing the nodes to the decision tree.

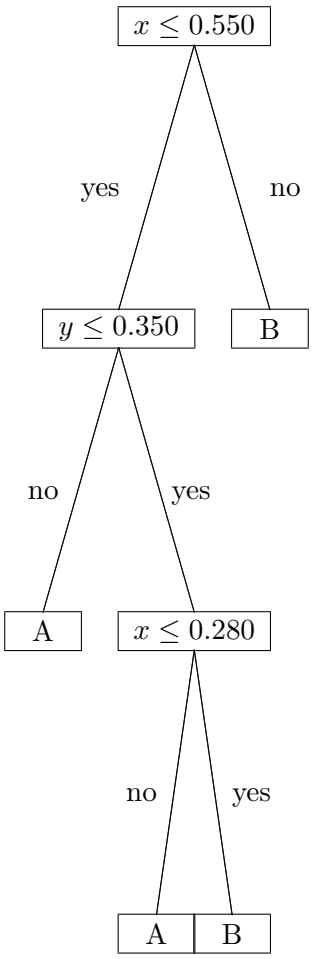


Figure I.2: The decision tree for classification.

Appendix J

Regression

J.1 Continuous variables

Let us consider the following toy size artificial training data set. The variable "drug dosage" is the predictive variable and the variable "drug effectiveness" is the target variable. The novelty of this problem is that both variables are continuous. What we do in such a situation is that replace explanatory variable to a discrete variable. This reduces the problem to an already settled case.

We describe the most commonly used method. This method starts with ordering the records such that the smallest value of the "drug dosage" at the first place and the largest values of the "drug dosage" is the last one. More precisely, the values of the "drug dosage" form an increasing (non-decreasing) sequence.

In our example there are 21 records each with a drug dosage and a drug effectiveness values. Between two consecutive drug dosage values there is a cut point. The cut point are $(2+4)/2 = 3, \ldots, (36+38)/2 = 37$. This gives 20 cut points. For each such cut point we construct a regression tree. Each of theses trees has a root node and two leaf nodes. We denotes these trees by $T_3, \ldots, T_{37}$. In addition there is a cut point before the first drug dosage value and there is a cut point after the last drug dosage value. The cut points are $2 - (1/2) = 1.5$ and $38 + (1/2) = 38.5$. We construct two regression trees $T_{1.5}$, $T_{38.5}$ associated with this cut points. Altogether there are 22 cut points and 22 regression trees. We compute the uncertainty of the regression trees and. This is quite a demanding tusk. This makes possible to locate the cut point with the smallest uncertainty. This optimal cut point is then used to replace drug dosage variable by a binary variable.

The uncertainty of the trees $T_{1.5}$, $T_{38.5}$ are both equal to 33.3. The uncertainty of the tree T_3, is equal to 33.2.

record number	drug dosage (mg)	drug effectiveness (%)
1	2	2
2	26	85
3	4	1
4	24	87
5	22	90
6	6	3
7	8	4
8	20	100
9	18	98
10	10	1
11	12	4
12	17	95
13	38	3
14	13	6
15	14	5
16	36	2
17	34	2
18	15	20
19	16	24
20	30	20
21	32	18

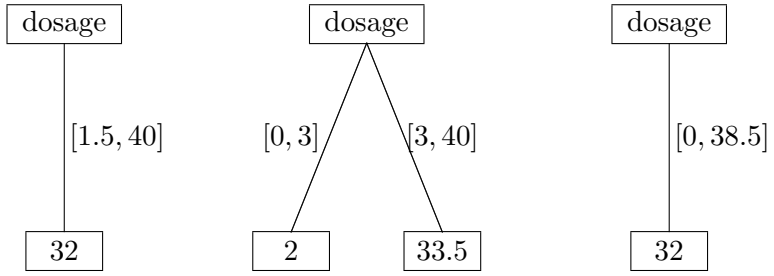


Figure J.1: The regression trees $T_{1.5}$, T_3, $T_{38.5}$ whose uncertainties we compute.

serial number	record number	drug dosage (mg)	drug effectivness (%)
1	1	2	2
2	3	4	1
3	6	6	3
4	7	8	4
5	10	10	1
6	11	12	4
7	14	13	6
8	15	14	5
9	18	15	20
10	19	16	24
11	12	17	95
12	9	18	98
13	8	20	100
14	5	22	90
15	4	24	87
16	2	26	85
17	20	30	20
18	21	32	18
19	17	34	2
20	16	36	2
21	13	38	3

serial number	record number	drug dosage	drug effectiveness actual	drug effectiveness predicted	deviation
1	1	2	2	32	30
2	3	4	1	32	31
3	6	6	3	32	29
4	7	8	4	32	28
5	10	10	1	32	31
6	11	12	4	32	28
7	14	13	6	32	26
8	15	14	5	32	27
9	18	15	20	32	12
10	19	16	24	32	8
11	12	17	95	32	63
12	9	18	98	32	66
13	8	20	100	32	68
14	5	22	90	32	58
15	4	24	87	32	55
16	2	26	85	32	53
17	20	30	20	32	12
18	21	32	18	32	14
19	17	34	2	32	30
20	16	36	2	32	30
21	13	38	3	32	29

serial number	record number	drug dosage	drug effectiveness actual	drug effectiveness predicted	deviation
1	1	2	2	2.0	0.0
2	3	4	1	33.5	32.5
3	6	6	3	33.5	30.5
4	7	8	4	33.5	29.5
5	10	10	1	33.5	32.5
6	11	12	4	33.5	29.5
7	14	13	6	33.5	27.5
8	15	14	5	33.5	28.5
9	18	15	20	33.5	13.5
10	19	16	24	33.5	9.5
11	12	17	95	33.5	61.5
12	9	18	98	33.5	64.5
13	8	20	100	33.5	66.5
14	5	22	90	33.5	56.5
15	4	24	87	33.5	53.5
16	2	26	85	33.5	51.5
17	20	30	20	33.5	13.5
18	21	32	18	33.5	15.5
19	17	34	2	33.5	31.5
20	16	36	2	33.5	31.5
21	13	38	3	33.5	30.5

J.2 A visual aid

In order to gain some insight into the nature of the connection between the variables "dosage" and "effectiveness" we plot the ordered pairs (x, y) in an orthogonal coordinate system. Here x and y are the values of the variables "dosage" and "effectiveness" from a chosen record. We plot 21 points as the number of the records is 21. Figure J.2 shows the points in the coordinate system. The "effectiveness" is not measured in percentages. We use ordinary real numbers instead.

The points are all lying in a rectangle. The horizontal line whose equation is "effectiveness"=0.5 cuts the enveloping rectangle into two parts and it separates the points into two groups. We may say that we discretised the "effectiveness" variable. The new discretised variable takes on "highly effective" and "moderately effective" values. The vertical lines with equations "dosage"=16.5 and "dosage"=28 separate the point into thee groups. We may say that the "dosage" variable takes the "low", "medium", "high"

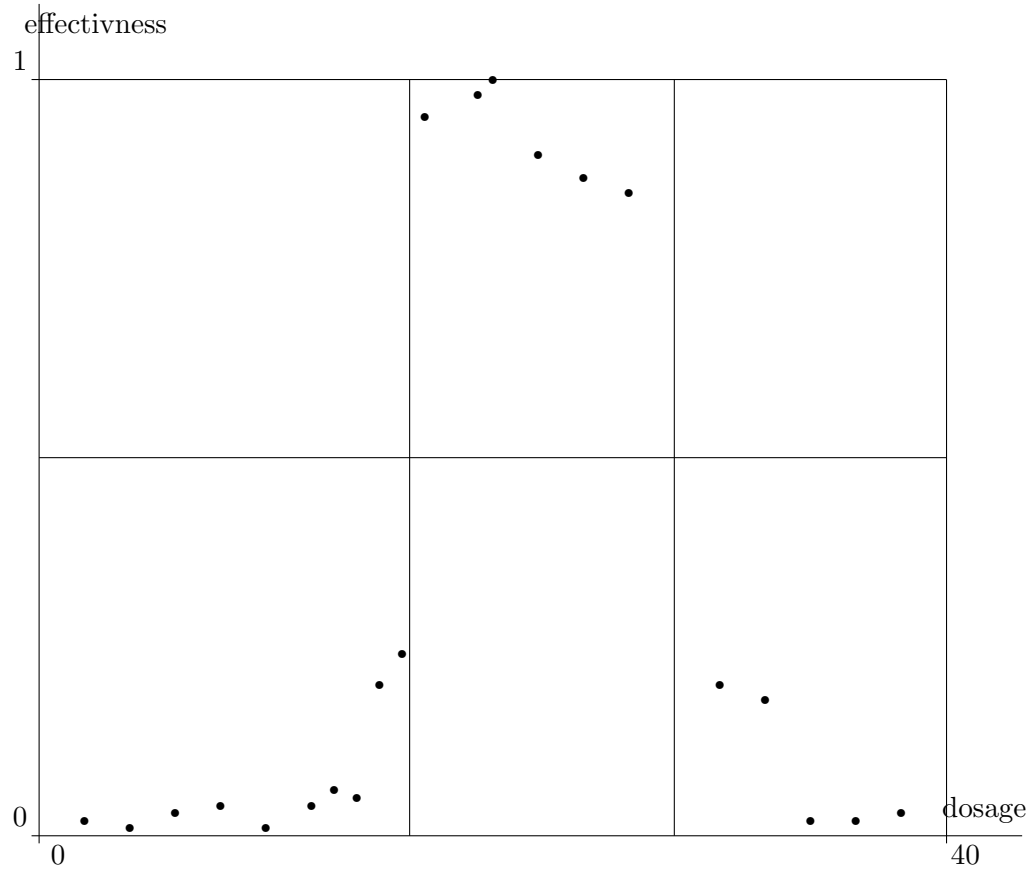


Figure J.2: Construction of a regression tree using a scatter plot.

dosage values. The introduced horizontal line and the two vertical lines cut the enveloping rectangle into 6 smaller rectangles. We can propose a regression tree using the result of this visual inspection. The tree is depicted in Figure J.3.

We assess the predictive power of the tree by computing the uncertainty of the prediction. The uncertainty of the prediction is equal to 6.2. The uncertainty of the prediction reduced (33.3/6.2)=5.37 fold compared with the regression tree $T_{1.5}$. The information gain is more than 2 bits.

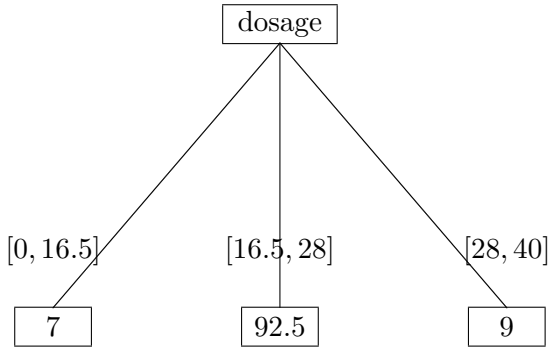


Figure J.3: The regression tree suggested by the visual inspection.

serial number	record number	drug dosage	drug effectiveness actual	drug effectiveness predicted	deviation
1	1	2	2	7.0	5.0
2	3	4	1	7.0	6.0
3	6	6	3	7.0	4.0
4	7	8	4	7.0	3.0
5	10	10	1	7.0	6.0
6	11	12	4	7.0	3.0
7	14	13	6	7.0	1.0
8	15	14	5	7.0	2.0
9	18	15	20	7.0	13.0
10	19	16	24	7.0	17.0
11	12	17	95	92.5	2.5
12	9	18	98	92.5	5.5
13	8	20	100	92.5	7.5
14	5	22	90	92.5	2.5
15	4	24	87	92.5	5.5
16	2	26	85	92.5	7.5
17	20	30	20	9.0	11.0
18	21	32	18	9.0	9.0
19	17	34	2	9.0	7.0
20	16	36	2	9.0	7.0
21	13	38	3	9.0	6.0