VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY
(Autonomous)

Department of Computer Science and Engineering

IV B.Tech – II Semester (Sections – C & D)

Machine Learning

UNIT-III – Tree Models & Rule Models

VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY

Syllabus

UNIT-III

Tree models:

Decision trees,

Ranking and probability estimation trees,

Tree learning as variance reduction.

Rule models:

Learning ordered rule lists, Learning unordered rule sets,

Descriptive rule learning,

First-order rule learning.

Tree learning as variance reduction – Regression Trees

We will now consider how to adapt decision trees to regression and clustering tasks.

Tree learning as variance reduction.

DI: Gini index $2 \cdot P(1-P)$ expected error rate

Probability head is P

Variance $P(1-P)$

Tree learning as variance reduction.

n

$n_1 \qquad n_2 = n - n_1$

empirical probab $2(P \cdot (1-P))$

$P_1 \qquad P_2$

weighted avg impurity $GI = \frac{n_1}{n} \cdot P_1 + \frac{n_2}{n} \cdot P_2$

$$= \frac{n_1}{n} \cdot 2 \cdot P_1 (1-P_1) + \frac{n_2}{n} \cdot 2 \cdot P_2 (1-P_2)$$

avg

$$= 2 \cdot \frac{n_1}{n} \sigma_1^2 + \frac{n_2}{n} \sigma_2^2$$

$\sigma_j^2 \rightarrow$ Bernouli distribution

Regression Tree:

target values are continuous

variance a set (Y) of target value are average squared distance from the mean.

variance a set (Y) of target value are average squared distance from the mean.

$$\operatorname{Var}(Y) = \frac{1}{|Y|} \sum_{y \in Y} (y - \underbrace{\bar{y}}_{\text{mean}})^2$$

Split $\rightarrow$ $\{Y_1, Y_2, \ldots, Y_l\}$ mutually exclusive

$$\operatorname{Var}(\{Y_1, Y_2, \ldots, Y_l\}) = \sum_{j=1}^{l} \frac{|Y_j|}{|Y|} \operatorname{Var}(Y_j)$$

VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY

Tree learning as variance reduction

- The variance of a Boolean (i.e., Bernoulli) variable with success probability $\dot{p}$ is $\dot{p}(1-\dot{p})$, which is half the Gini index. So we could interpret the goal of tree learning as minimising the class variance (or standard deviation, in case of $\sqrt{\text{Gini}}$) in the leaves.
- In regression problems we can define the variance in the usual way:

$$\text{Var}(Y) = \frac{1}{|Y|} \sum_{y \in Y} (y - \overline{y})^2$$

If a split partitions the set of target values Y into mutually exclusive sets $\{Y_1, \dots, Y_l\}$, the weighted average variance is then

$$\text{Var}(\{Y_1, \dots, Y_l\}) = \sum_{j=1}^{l} \frac{|Y_j|}{|Y|} \text{Var}(Y_j) = \dots = \frac{1}{|Y|} \sum_{y \in Y} y^2 - \sum_{j=1}^{l} \frac{|Y_j|}{|Y|} \overline{y}_j^2$$

The first term is constant for a given set Y and so we want to maximise the weighted average of squared means in the children.

VVIT
VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY

Example 5.4, p.150

Learning a regression tree I

Imagine you are a collector of vintage Hammond tonewheel organs. You have been monitoring an online auction site, from which you collected some data about interesting transactions:

#	Model	Condition	Leslie	Price
1.	B3	excellent	no	4513
2.	T202	fair	yes	625
3.	A100	good	no	1051
4.	T202	good	no	270
5.	M102	good	yes	870
6.	A100	excellent	no	1770
7.	T202	fair	no	99
8.	A100	good	yes	1900
9.	E112	fair	no	77

#	model	condition	Leslie	Price
1	B3	excellent	no	4513
2	T202	fair	yes	625
3	A100	good	no	1051
4	T202	good	no	270
5	M102	good	yes	870
6	A100	excellent	no	1770
7	T202	fair	no	99
8	A100	good	yes	100
9	E112	fair	no	77

model [A100, B3, E112, M102, T202]
[1051, 1770, 100] [4513] [77] [870] [625, 270, 99]

condition [excellent, good, fair]
[4513, 1770] [1051, 270, 870, 100] [625, 99, 77]

Leslie [yes, no]
[625, 870, 100] [4513, 1051, 270, 1770, 99, 77]

VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY

$$\sum \frac{1}{9} (1574-1051)^2 + (1574-1770)^2 + (1574-100)^2$$

$$= \sum \frac{|Y_j|}{|Y|} \operatorname{var}(Y_j)$$

VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY

Tree learning as variance reduction

#	model	condition	Leslie	Price
1	B3	excellent	no	4513
2	T202	fair	yes	625
3	A100	good	no	1051
4	T202	good	no	270
5	M102	good	yes	870
6	A100	excellent	no	1770
7	T202	fair	no	99
8	A100	good	yes	100
9	E112	fair	no	77

model [A100, B3, E112, M102, T202]
[1051, 1770, 100] [4513] [77] [870] [625, 270, 99]

condition [excellent, good, fair]
[4513, 1770] [1051, 270, 870, 100] [625, 99, 77]

Leslie [yes, no]
[625, 870, 100] [4513, 1051, 270, 1770, 99, 77]

A100 T202

Condition [ex, good, fair] [1770] [1051, 100] [

Leslie [yes, NO] [100] [1051, 1770]

Condit T202 [ ] [270] [99]

Leslie [625] [270, 99]

model: A100, B3, E112, M102, T202, ...

Example 5.4, p.150

Learning a regression tree II

From this data, you want to construct a regression tree that will help you determine a reasonable price for your next purchase.
There are three features, hence three possible splits:

$\text{Model} = [\text{A100}, \text{B3}, \text{E112}, \text{M102}, \text{T202}]$

$[1051, 1770, 1900][4513][77][870][99, 270, 625]$

$\text{Condition} = [\text{excellent}, \text{good}, \text{fair}]$

$[1770, 4513][270, 870, 1051, 1900][77, 99, 625]$

$\text{Leslie} = [\text{yes}, \text{no}]$ $[625, 870, 1900][77, 99, 270, 1051, 1770, 4513]$

The means of the first split are 1574, 4513, 77, 870 and 331, and the weighted average of squared means is $3.21 \cdot 10^6$.
The means of the second split are 3142, 1023 and 267, with weighted average of squared means $2.68 \cdot 10^6$;
for the third split the means are 1132 and 1297, with weighted average of squared means $1.55 \cdot 10^6$.
We therefore branch on Model at the top level. This gives us three single-instance leaves, as well as three A100s and three T202s.

Example 5.4, p.150

Learning a regression tree III

For the A100s we obtain the following splits:

$$\text{Condition} = [\text{excellent}, \text{good}, \text{fair}] \qquad [1770][1051, 1900][]$$
$$\text{Leslie} = [\text{yes}, \text{no}] \qquad [1900][1051, 1770]$$

Without going through the calculations we can see that the second split results in less variance (to handle the empty child, it is customary to set its variance equal to that of the parent). For the T202s the splits are as follows:

$$\text{Condition} = [\text{excellent}, \text{good}, \text{fair}] \qquad [][270][99, 625]$$
$$\text{Leslie} = [\text{yes}, \text{no}] \qquad [625][99, 270]$$

Again we see that splitting on Leslie gives tighter clusters of values. The learned regression tree is depicted in Figure 5.8.

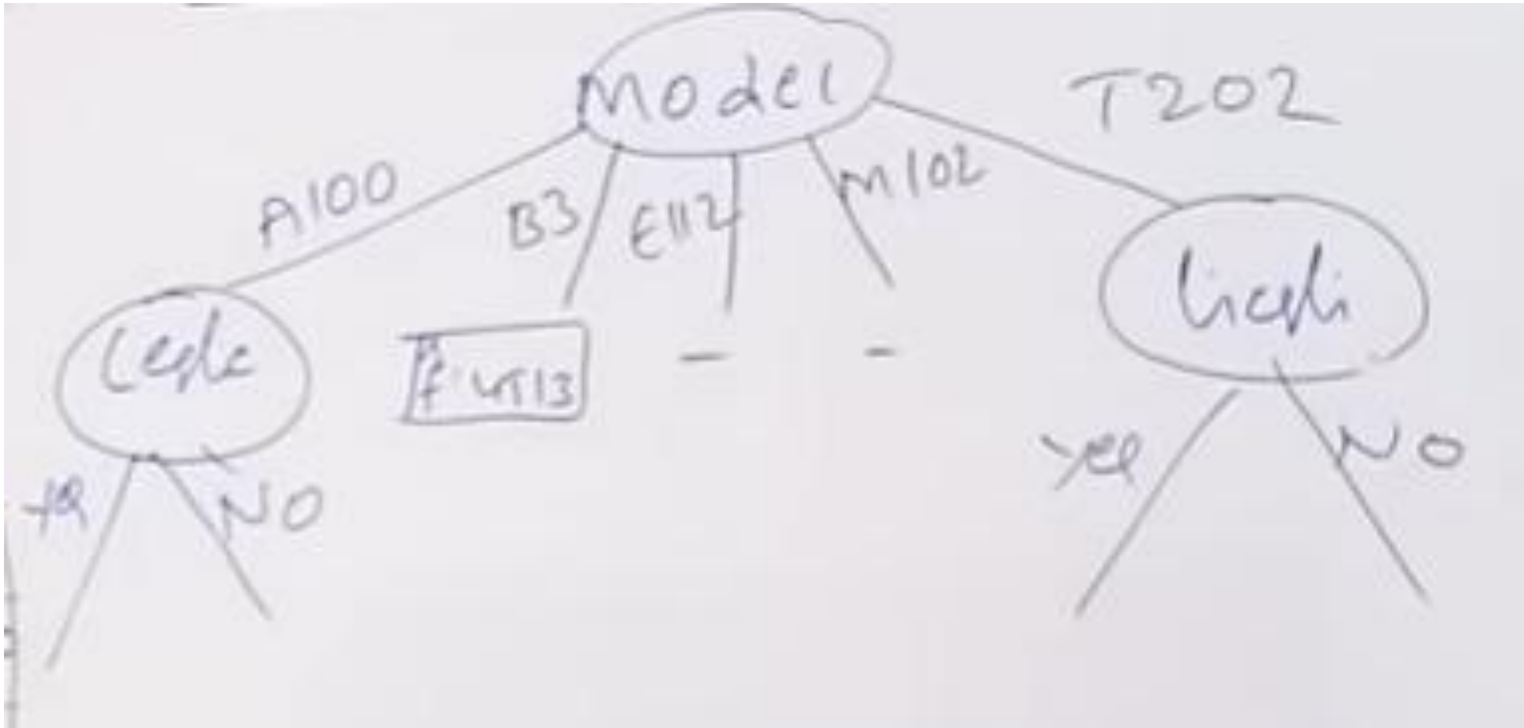
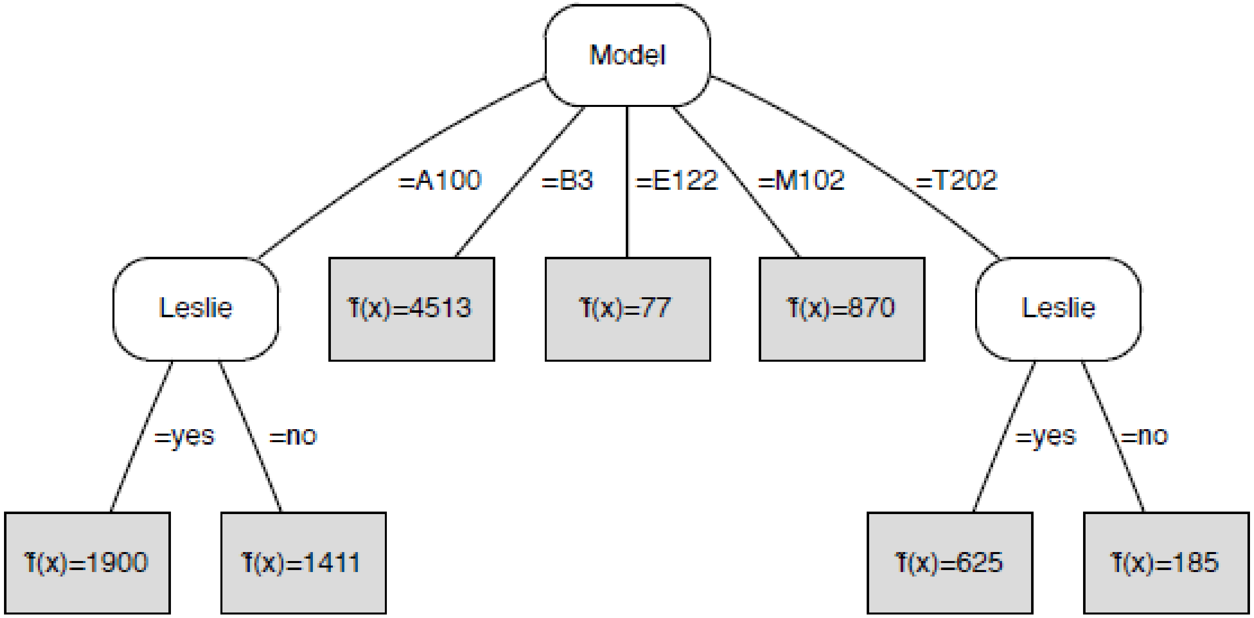

Model
A100
B3
E112
M102
T202
NO
NO

Figure 5.8, p.150

A regression tree



A regression tree learned from the data in Example 5.4.

Next Class…..
Clustering Trees

VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY
(Autonomous)

Department of Computer Science and Engineering

IV B.Tech – II Semester (Sections – C & D)

Machine Learning

UNIT-III – Tree Models & Rule Models

Tree learning as variance reduction – Clustering Trees

Dissimilarity measure

Let Dis: $\mathscr{X} \times \mathscr{X} \rightarrow \mathbb{R}$ be an abstract function that measures *dissimularity* of any two instances $x, x' \in \mathscr{X}$, such that the higher $\text{Dis}(x, x')$ is, the less similar x and x' are. The *cluster dissimilarity* of a set of instances D is:

$$\text{Dis}(D) = \frac{1}{|D|^2} \sum_{x \in D} \sum_{x' \in D} \text{Dis}(x, x')$$

VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY

Tree learning as variance reduction:

Clustering Trees.

Dis Similarity — $Dis : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$

The cluster Dis Similarity

$$Dis(D) = \frac{1}{|D|^2} \sum_{x \in D} \sum_{x' \in D} Dis(x, x')$$

Example 5.5, p.152

Learning a clustering tree I

Assessing the nine transactions on the online auction site from Example 5.4, using some additional features such as reserve price and number of bids, you come up with the following dissimilarity matrix:

0	11	6	13	10	3	13	3	12
11	**0**	1	**1**	1	3	**0**	4	0
6	1	**0**	2	1	**1**	2	**2**	1
13	**1**	2	**0**	0	4	**0**	4	0
10	1	1	0	0	3	0	2	0
3	3	**1**	4	3	**0**	4	**1**	3
13	**0**	2	**0**	0	4	**0**	4	0
3	4	**2**	4	2	**1**	4	**0**	4
12	0	1	0	0	3	0	4	0

This shows, for instance, that the first transaction is very different from the other eight. The average pairwise dissimilarity over all nine transactions is 2.94.

How to create a distance matrix

Data can be recorded in a distance matrix at the time of collection. For example, in some studies of perception, people are asked to rate the psychological distance between pairs of objects, and these distances are recorded in a distance matrix.

More commonly, a distance matrix is computed from a *raw data* table. In the example below, we can use high school math (Pythagoras) to work out that the distance between A and B is

$$\sqrt{(24-9)^2+(54-49)^2} = 15.81 \approx 16$$

We can use the same formula with more than two variables, and this is known as the *Euclidean distance*.

Raw Data

	X	Y
A	9	49
B	24	54
C	51	28
D	81	54
E	81	23
F	86	32

Distance Matrix

	A	B	C	D	E	F
A	0	16	47	72	77	79
B	16	0	37	57	65	66
C	47	37	0	40	30	35
D	72	57	40	0	31	23
E	77	65	30	31	0	10
F	79	66	35	23	10	0

VVIT
VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY

Tree learning as variance reduction

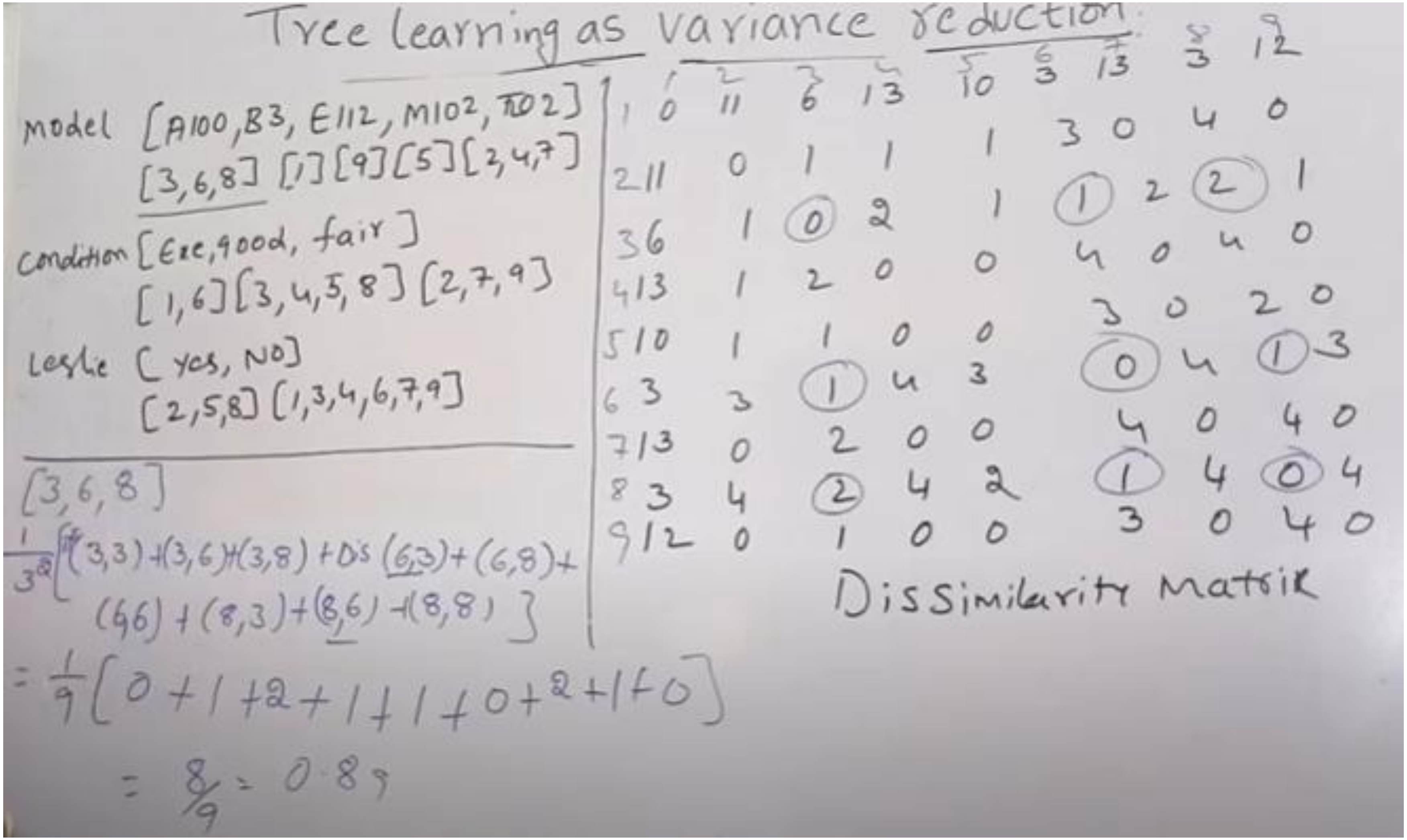
model [A100, B3, E112, M102, T202]
[3,6,8] [1] [9] [5] [2,4,7]

condition [Exc, good, fair]
[1,6] [3,4,5,8] [2,7,9]

Leslie [yes, No]
[2,5,8] [1,3,4,6,7,9]

0	11	6	13	10	3	13	3	12
11	0	1	1	1	3	0	4	0
6	1	0	2	1	1	2	2	1
13	1	2	0	0	4	0	4	0
10	1	1	0	0	3	0	2	0
3	3	1	4	3	0	4	1	3
13	0	2	0	0	4	0	4	0
3	4	2	4	2	1	4	0	4
12	0	1	0	0	3	0	4	0

VVIT
VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY



Avg wt =

$$\frac{3}{9} \times 0.89 + \frac{1}{9} \times 0 + \frac{1}{9} \times 0 + \frac{1}{9} \times 0 + \frac{3}{9} \times 0.22 = 0.37$$

Example 5.5, p.152

Learning a clustering tree II

Using the same features from Example 5.4, the three possible splits are (now with transaction number rather than price):

$\text{Model} = [\text{A100}, \text{B3}, \text{E112}, \text{M102}, \text{T202}]$ $[3,6,8][1][9][5][2,4,7]$

$\text{Condition} = [\text{excellent}, \text{good}, \text{fair}]$ $[1,6][3,4,5,8][2,7,9]$

$\text{Leslie} = [\text{yes}, \text{no}]$ $[2,5,8][1,3,4,6,7,9]$

The cluster dissimilarity among transactions 3, 6 and 8 is $\frac{1}{3^2}(0+1+2+1+0+1+2+1+0) = 0.89$; and among transactions 2, 4 and 7 it is $\frac{1}{3^2}(0+1+0+1+0+0+0+0+0) = 0.22$. The other three children of the first split contain only a single element and so have zero cluster dissimilarity. The weighted average cluster dissimilarity of the split is then $3/9 \cdot 0.89 + 1/9 \cdot 0 + 1/9 \cdot 0 + 1/9 \cdot 0 + 3/9 \cdot 0.22 = 0.37$. For the second split, similar calculations result in a split dissimilarity of $2/9 \cdot 1.5 + 4/9 \cdot 1.19 + 3/9 \cdot 0 = 0.86$, and the third split yields $3/9 \cdot 1.56 + 6/9 \cdot 3.56 = 2.89$. The Model feature thus captures most of the given dissimilarities, while the Leslie feature is virtually unrelated.

Example 5.6, p.154

Clustering with Euclidean distance I

We extend our Hammond organ data with two new numerical features, one indicating the reserve price and the other the number of bids made in the auction.

Model	Condition	Leslie	Price	Reserve	Bids
B3	excellent	no	45	30	22
T202	fair	yes	6	0	9
A100	good	no	11	8	13
T202	good	no	3	0	1
M102	good	yes	9	5	2
A100	excellent	no	18	15	15
T202	fair	no	1	0	3
A100	good	yes	19	19	1
E112	fair	no	1	0	5

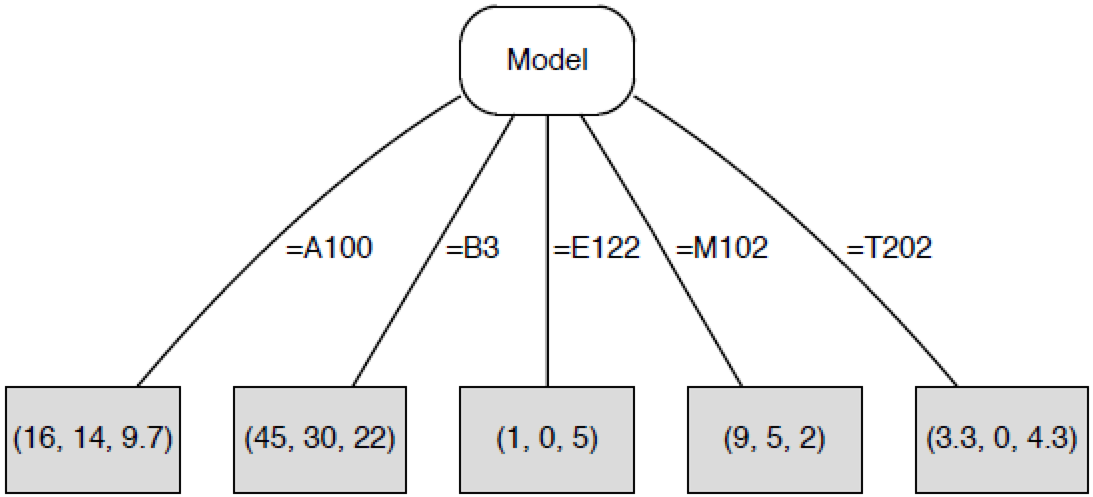
Example 5.6, p.154

Clustering with Euclidean distance II

- The means of the three numerical features are $(13.3, 8.6, 7.9)$ and their variances are $(158, 101.8, 48.8)$. The average squared Euclidean distance to the mean is then the sum of these variances, which is 308.6.
- For the A100 cluster these vectors are $(16, 14, 9.7)$ and $(12.7, 20.7, 38.2)$, with average squared distance to the mean 71.6; for the T202 cluster they are $(3.3, 0, 4.3)$ and $(4.2, 0, 11.6)$, with average squared distance 15.8.
- Using this split we can construct a clustering tree whose leaves are labelled with the mean vectors (Figure 5.9).

Figure 5.9, p.154

A clustering tree



A clustering tree learned from the data in Example 5.6 using Euclidean distance on the numerical features.