Homework 4 Solutions

1) Read Chapter 4 (all sections) and Chapter 5 (Section 5.7 only).

2)

a) There are four positive examples and five negative examples. Thus, $P(+) = 4/9$ and $P(-) = 5/9$. The entropy of the training examples is $-4/9 \log_2(4/9) - 5/9 \log_2(5/9) = 0.9911$.

b) For attribute a_1, the corresponding counts and probabilities are:

a_1	+	-
T	3	1
F	1	4

The entropy for a_1 is

$$\frac{4}{9}\Big[-(3/4)\log_2(3/4) - (1/4)\log_2(1/4)\Big] + \frac{5}{9}\Big[-(1/5)\log_2(1/5) - (4/5)\log_2(4/5)\Big] = 0.7616.$$

Therefore, the information gain for a_1 is $0.9911 - 0.7616 = 0.2294$.

For attribute a_2, the corresponding counts and probabilities are:

a_2	+	-
T	2	3
F	2	2

The entropy for a_2 is

$$\frac{5}{9}\Big[-(2/5)\log_2(2/5) - (3/5)\log_2(3/5)\Big] + \frac{4}{9}\Big[-(2/4)\log_2(2/4) - (2/4)\log_2(2/4)\Big] = 0.9839.$$

Therefore, the information gain for a_2 is $0.9911 - 0.9839 = 0.0072$.

c)

Split point	Entropy	Info Gain
2.0	0.8484	0.1427
3.5	0.9885	0.0026
4.5	0.9183	0.0728
5.5	0.9839	0.0072
6.5	0.9728	0.0183
7.5	0.8889	0.1022

The best split for a_3 occurs at split point equals to 2.

d) According to information gain, a_1 produces the best split.

e)

For attribute a_1: error rate $= 2/9$.
For attribute a_2: error rate $= 4/9$.
Therefore, according to error rate, a_1 produces the best split.

3)

b) The overall gini before splitting is:

$$G_{orig} = 1 - 0.4^2 - 0.6^2 = 0.48$$

The gain in gini after splitting on A is:

$$\begin{aligned} G_{A=T} &= 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2 = 0.4898 \\ G_{A=F} &= 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2 = 0 \\ \Delta &= G_{orig} - 7/10 G_{A=T} - 3/10 G_{A=F} = 0.1371 \end{aligned}$$

The gain in gini after splitting on B is:

$$\begin{aligned} G_{B=T} &= 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0.3750 \\ G_{B=F} &= 1 - \left(\frac{1}{6}\right)^2 - \left(\frac{5}{6}\right)^2 = 0.2778 \\ \Delta &= G_{orig} - 4/10 G_{B=T} - 6/10 G_{B=F} = 0.1633 \end{aligned}$$

Therefore, attribute B will be chosen to split the node.

c) Yes, even though these measures have similar range and monotonous behavior, their respective gains, Δ, which are scaled differences of the measures, do not necessarily behave in the same way, as illustrated by the results in parts (a) and (b).

4)

b)

Because the $A = T$ child node is pure, no further splitting is needed. For the $A = F$ child node, the distribution of training instances is:

B	C	Class label	
		+	−
T	T	0	20
F	T	0	5
T	F	25	0
F	F	0	25

The classification error of the $A = F$ child node is:

$$E_{orig} = \frac{25}{75}$$

After splitting on attribute B, the gain in error rate is:

	$B = T$	$B = F$
+	25	0
−	20	30

$$E_{B=T} = \frac{20}{45}$$

$$E_{B=F} = 0$$

$$\Delta_B = E_{orig} - \frac{45}{75}E_{B=T} - \frac{30}{75}E_{B=F} = \frac{5}{75}$$

After splitting on attribute C, the gain in error rate is:

	$C = T$	$C = F$
+	0	25
−	25	25

$$E_{C=T} = \frac{0}{25}$$

$$E_{C=F} = \frac{25}{50}$$

$$\Delta_C = E_{orig} - \frac{25}{75}E_{C=T} - \frac{50}{75}E_{C=F} = 0$$

The split will be made on attribute B.

c) 20 instances are misclassified. (The error rate is $\frac{20}{100}$.)

d) For the $C = T$ child node, the error rate before splitting is: $E_{orig} = \frac{25}{50}$.
After splitting on attribute A, the gain in error rate is:

	$A = T$	$A = F$
+	25	0
−	0	25

$$E_{A=T} = 0$$
$$E_{A=F} = 0$$
$$\Delta_A = \frac{25}{50}$$

After splitting on attribute B, the gain in error rate is:

	$B = T$	$B = F$
+	5	20
−	20	5

$$E_{B=T} = \frac{5}{25}$$
$$E_{B=F} = \frac{5}{25}$$
$$\Delta_B = \frac{15}{50}$$

Therefore, A is chosen as the splitting attribute.

For the $C = F$ child, the error rate before splitting is: $E_{orig} = \frac{25}{50}$.
After splitting on attribute A, the error rate is:

	$A = T$	$A = F$
+	0	25
−	0	25

$$E_{A=T} = 0$$
$$E_{A=F} = \frac{25}{50}$$
$$\Delta_A = 0$$

After splitting on attribute B, the error rate is:

	$B = T$	$B = F$
+	25	0
−	0	25

$$E_{B=T} = 0$$
$$E_{B=F} = 0$$
$$\Delta_B = \frac{25}{50}$$

Therefore, B is used as the splitting attribute.
The overall error rate of the induced tree is 0.

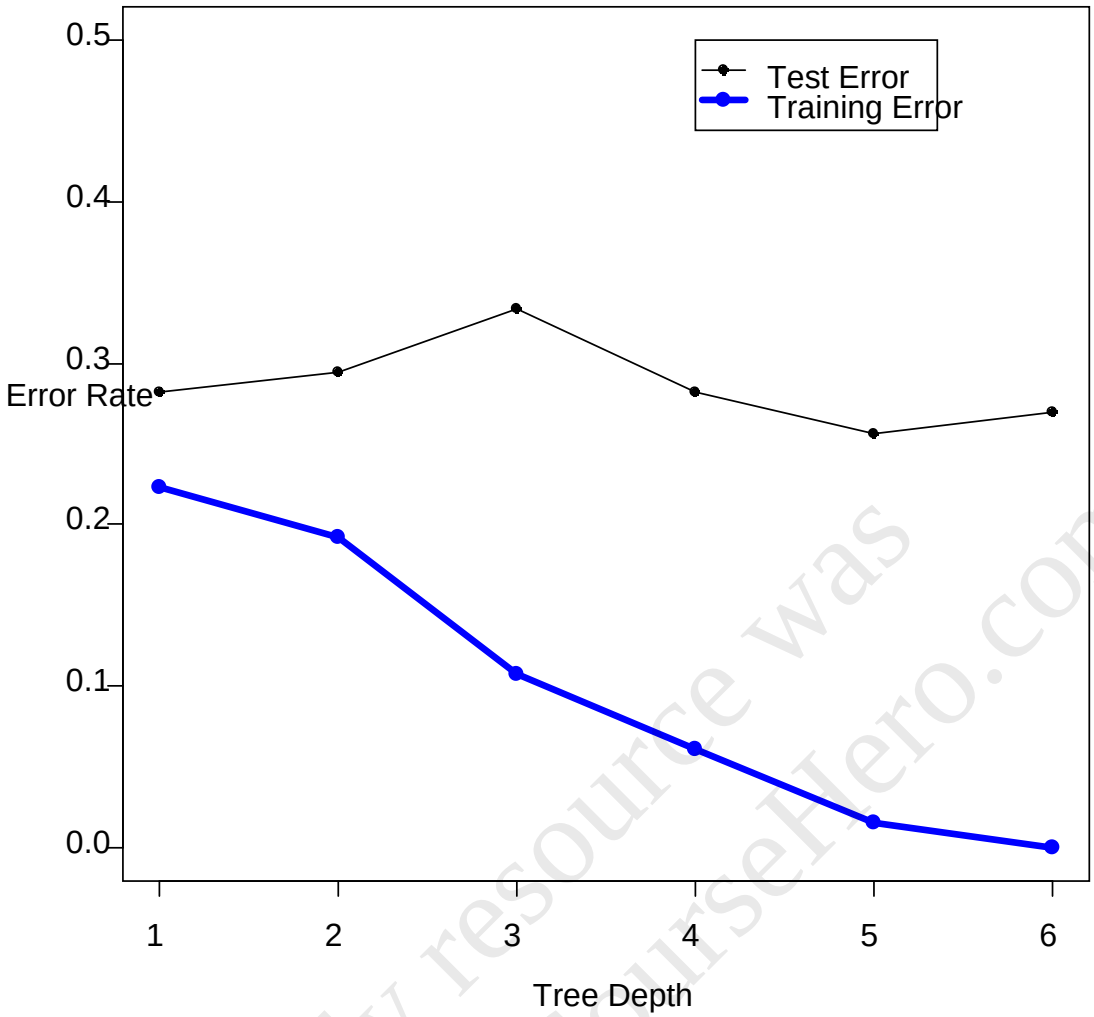
e) The greedy heuristic does not necessarily lead to the best tree.

5) Here are the correct predictions:

Age	Number	Start	Prediction
middle	5	10	**present**
young	2	17	**absent**
old	10	6	**present**
young	2	17	**absent**
old	4	15	**absent**
middle	5	15	**absent**
young	3	13	**absent**
old	5	8	**present**
young	7	9	**absent**
middle	3	13	**absent**

6)

```
install.packages("rpart")
library(rpart)

train<-read.csv("sonar_train.csv",header=FALSE)
y<-as.factor(train[,61])
x<-train[,1:60]

test<-read.csv("sonar_test.csv",header=FALSE)
y_test<-as.factor(test[,61])
x_test<-test[,1:60]


train_error<-rep(0,6)
test_error<-rep(0,6)

for (dep in 1:6) {
  fit<-rpart(y~.,x,
    control=rpart.control(minsplit=0,minbucket=0,cp=-1,
    maxcompete=0, maxsurrogate=0, usesurrogate=0,
    xval=0,maxdepth=dep))
  train_error[dep]<-
      1-sum(y==predict(fit,x,type="class"))/length(y)
  test_error[dep]<-
      1- sum(y_test==predict(fit,x_test,type="class"))/length(y_test)
}

plot(seq(1,6),test_error,type="o",pch=19,ylim=c(0,.5),
  ylab="Error Rate",
  xlab="Tree Depth",main="Rajan Patel's Tree Error Plot")

points(train_error,type="o",pch=19,lwd=4,col="blue")
  legend(4,.5,c("Test Error","Training Error"),
  col=c("black","blue"),pch=19,lwd=c(1,4))
```



The plot suggests a depth of 5 is optimal.

7)

a)

The ROC curve for $M1$ and $M2$ are shown in the Figure 5.5.

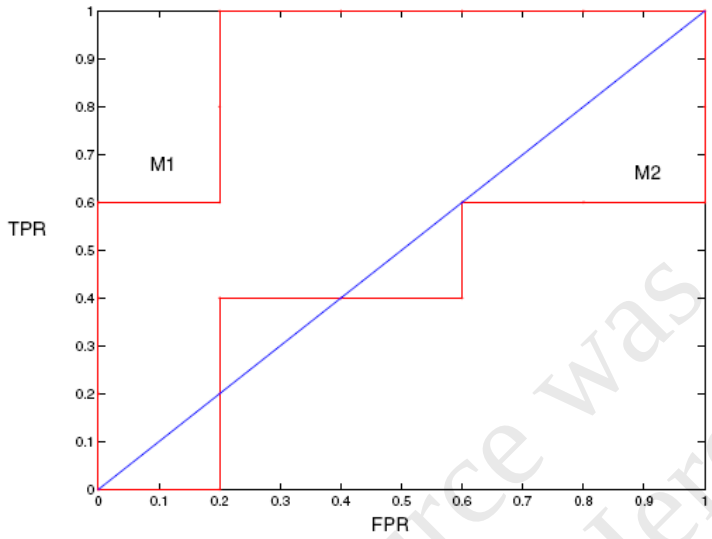


Figure 5.5. ROC curve.

$M1$ is better, since its area under the ROC curve is larger than the area under ROC curve for $M2$.

c)

When $t = 0.5$, the confusion matrix for $M2$ is shown below.

		+	-
Actual	+	1	4
	-	1	4

Precision $= 1/2 = 50\%$.

Recall $= 1/5 = 20\%$.

F-measure $= (2 \times .5 \times .2)/(.5 + .2) = 0.2857$.

Based on F-measure, $M1$ is still better than $M2$. This result is consistent with the ROC plot.