Homework2

Ruhui Shen A20410233

1 Exercises

1.1 Tan, Chapter 3

8.

(a) If the line in the middle of box, or to be exact, in the middle of the 90% and 10%, we could assume the value of the attribute is symmetrically distributed.

(b)The sepal width and sepal length are relatively symmetrically distributed while the other two, petal length and width are not, seem skewed.

9.

For Setosa, sepal length> sepal width> petal length> petal width.

For Versicolour, sepal length> petal length> sepal width> petal width.

For Virginica, same with the Versicolour.

10.

First, we could create the box plots for each attribute to get some information like the distribution etc. Second, we could use the box plots for one attribute to overlap or cross various categories of another attribute. In that case, we could get the information of what's height or weight going on with age increasing by comparing the box plots of age for different categories of ages.

1.2 Tan, Chapter 4

2.

(a) Gini = $1 - (0.5^2 + 0.5^2) = 0.5$

(b) The Gini for each ID is 0, so the overall Gini for Customer ID is 0.

(c) The Gini for Male is $1 - (0.5^2 + 0.5^2) = 0.5$, same for female. Therefore, overall Gini for Gender is $0.5^2 + 0.5^2 = 0.5$.

(d) The Gini for family car is 1-(1/4)^2-(3/4)^2=3/8=0.375. For Sport: 0. Luxury car: 1-(1/8)^2-(7/8)^2= 14/64=7/32=0.2188

Overall : 0.1625.

(e) Gini for Small Shirt Size is $1 - (3/5)^2 - (2/5)^2 = 0.48$

Medium is 1-(3/7)^2-(4/7)^2= 0.4898

Large is 1-(2/4)^2-(2/4)^2=0.5

Extra is 1-(2/4)^2-(2/4)^2=0.5

The overall Gini is 0.4914

(f) Car Type. Because it's Gini is lowest.

(g) Every customer has their own Customer ID, split it doesn't make sense since no information can be deduced from it. The attribute has no predictive power since new customers are assigned to new Customer IDs.

3.

(a) $-\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9} = 0.9911$

(b) The entropy for a1 is 4/9[-(3/4)log2(3/4)-(1/4)log2(1/4)]+5/9[-(1/5)log2(1/5)-(4/5)log2(4/5)]=0.7616

The information gain for a1 is 0.9911-0.7616=0.2294.

The entropy for a2 is 5/9[-(2/5)log2(2/5)-(3/5)log2(3/5)]+4/9[-(2/4)log2(2/4)-(2/4)log2(2/4)]=0.9839.

The information gain for a1 is 0.9911-0.9839=0.0072.

(c) The entropy for a3 are 0.8484, 0.9885, 0.9183, 0.9839, 0.9728, 0.8889. So the best split is 2.

(d) a1 produce the best split

(e) Error rate for a_1:2/9, a_2:4/9.

since a_1 has smaller error rate, a_1 produce better split

(f) The Gini index for $a_1 = \frac{4}{9}\left(1 - \frac{4}{16} - \frac{1}{5}\right) + \frac{5}{9}\left(1 - \frac{1}{25} - \frac{16}{25}\right) = 0.3444$

The Gini index for $a_2 = \frac{5}{9}\left(1 - \frac{4}{25} - \frac{9}{25}\right) + \frac{4}{9}\left(1 - \frac{4}{16} - \frac{4}{16}\right) = 0.4889$

So a1 produces the best spilt.

5.

(a)

According to the contingency tables, the overall entropy before splitting is:

E(orig) = -0.4 log 0.4 - 0.6 log 0.6 =0.9710

As for A:

E(A=T) = -4/7log(4/7)-3/7log(3/7)=0.9852

E(A=F) = -log(1)-0=0

Information Gain = 0.9710-(7/10)*0.9852-(3/10)*0.9852=0.2813

As for B:

E(B=T) = -3/4log(3/4)-1/4log(1/4)=0.8113

E(B=F) = -1/6log(1/6)-5/6log(5/6)=0.6500

Information Gain = 0.9710-(4/10)*0.8113-(6/10)*0.6500=0.2565

So attribute A will be chosen.

(b)

Before splitting, the overall mini is:

G(orig)=1-0.4^2-0.6^2=0.48

After splitting, the gain in gini on A is:

G(A=T) = 1-(4/7)^2-(3/7)^2=0.4898

G(A=F) = 1-(1)^2-0^2=0

Information Gain = 0.48 - (7/10)*0.4898-(3/10)*0=0.1371

After splitting, the gain in gini on B is:

G(B=T) = 1-(1/4)^2-(3/4)^2=0.3750

G(B=F) = 1-(1/6)^2-(5/6)^2=0.2778

Information Gain = 0.48 - (4/10)*0.3750-(6/10)*0.2778=0.1633

So, attribute B will be chosen.

(c) Yes. Because the information gain and the respective gains, which are scaled differences of measures, don't have to behave in the the same way.

1.3 Tan, Chapter 5

20.

$$\text{Error rate} = \frac{\text{Number of wrong predictions}}{\text{Total number of predictions}} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}. \quad (4.2)$$

(a) 50%

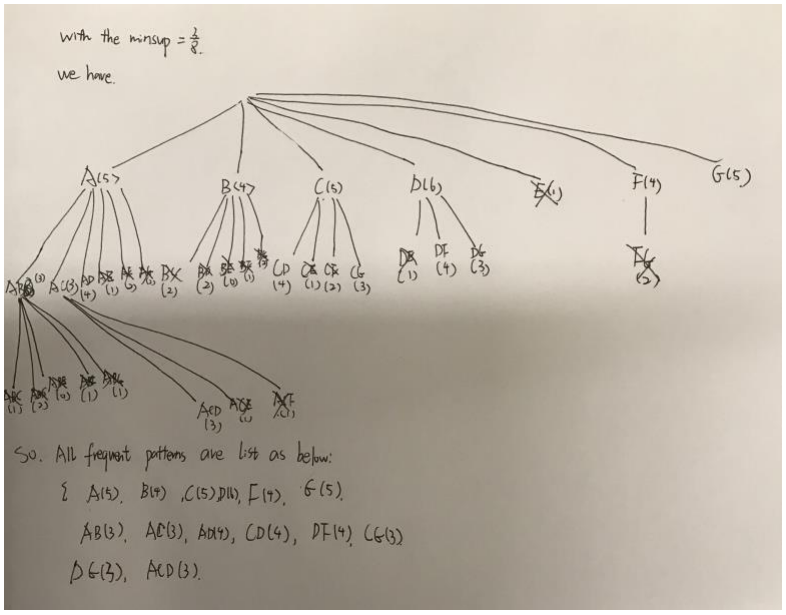
(b) 50%

(c) 30%

(d) 44.4%

1.4 Zaki, Chapter 8

1.



4.

Since all subsets of {ABE} are {A}, {B}, {E}, {A,B}, {A,E},{B,E}, {A,B,E}

As for {A,B} =>{E}, confidence=2/3=0.67

As for {A,E} =>{B}, confidence=2/2=1.0

As for {B,E} =>{A}, confidence=2/4=0.5

As for {A} =>{B,E}, confidence=2/4=0.5

As for {B} =>{A,E}, confidence=2/5=0.4

As for { E} =>{A,B}, confidence=2/4=0.5

6(a). The size of the items search space is 2^11.

6(b). The support of the new items is more than or equal to support of X