CS 422

DATA MINING

HOME WORK – 2

SUBMITTED BY :

SACHIN KRISHNA MURTHY

CWID : A20354077

Chapter 2

18. This exercise compares and contrasts some similarity and distance measures

a) For binary data, the L1 distance corresponds to the Hamming distance; that is, the number of bits that are different between two binary vectors. The Jaccard similarity is a measure of the similarity between two binary vectors. Compute the Hamming distance and the Jaccard similarity between the following two binary vectors.

x = 0101010001 , y = 0100011000

Solution :

- **Hamming Distance :** It is the distance between two binary strings of equal length.
- Given Strings :

  x = 0101010001
  y = 0100011000

- Both the strings 'x' and 'y' are of equal length of 10.
- The distance between two binary strings are the number of mismatching bits in the string.
- In the given two strings let's consider the first bits of each string.
  - For x it is 1 and for y it is 0, hence it is a mismatch.
  - While second bit for both the strings is 0 and is not a mismatch.
- Let us consider both the strings and record the mismatched bits as 1 and rest as 0.

  x = 0101010001
  y = 0100011000

  Result = 0001001001

- We compute the hamming distance by adding each bit of the resultant string. That is, 0+0+0+1+0+0+1+0+0+1=3

Jaccard Coefficient : It is the ratio of the number of matching presences to the number of attributes expect matching zeroes and it is given by :

$$J = \frac{\text{Number of matching presences}}{\text{Number of attributes not involved in 00 matches}} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

f_{11} : Number of attributes where both x and y are 1.
f_{01} : Number of attributes where x is 0 and y is 1.
f_{10} : Number of attributes where x is 1 and y is 0.
f_{00} : Number of attributes where both x and y are 0.

Here in the given strings,

f_{11} : 2

f_{01} : 1

f_{00} : 5

f_{10} : 2

$$J = \frac{2}{1+2+2} = \frac{2}{5} = 0.4$$

18.(c) Suppose that you are comparing how similar two organisms of different species are in terms of the number of genes they share. Describe which measure; Hamming or Jaccard, you think would be more appropriate for comparing the genetic makeup of two organisms. Explain. (Assume that each animal is represented as a binary vector, where each attribute is 1 if a particular gene is present in the organism and 0 otherwise.)

Solution :

Let us consider two organisms of different species A and B.

	Gene 'p'	Gene 'q'	Gene 'r'	Gene 's'	Gene 't'	Gene 'u'
Species 'A'	1	0	0	1	1	0
Species 'B'	1	1	0	1	0	1

Considering Hamming Distance :

A = 100110
B = 110101

Result = 010011

Hamming Distance = 0+1+0+0+1+1=3

Considering Jaccard Coefficient :

$$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} = \frac{2}{2+1+2} = \frac{2}{5} = 0.4$$

Hamming distances gives us the number of mismatches between two species which is in this case its 3. Jaccard co-efficient gives us the measure of asymmetric information on binary and non-binary variables which has the value of 0.4 in this case.

Hence from the above example considered, we can say that Jaccard Coefficient is more appropriate for considering the genetic makeup of two organisms because it gives us the percentage of genes the organisms share.

18.(d) If you wanted to compare the genetic makeup of two organisms of the same species, e.g., two human beings, would you use the Hamming distance, the Jaccard coefficient, or a different measure of similarity or distance? Explain. (Note that two human beings share > 99.9% of the same genes.)

Solution :

Considering the genetic make up of two organisms of the same species (Humans).

	Gene 'a'	Gene 'b'	Gene 'c'	Gene 'd'	Gene 'e'
Human 'X'	1	1	1	1	1
Human 'Y'	1	1	0	1	1

Considering Hamming Distance :

A = 11111
B = 11011

Result = 00100

Hamming Distance = 0+0+1+0+0=1

Considering Jaccard Coefficient :

$$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} = \frac{4}{0+1+4} = \frac{4}{5} = 0.4$$

Since both X and Y belong to same species Human, we do need the percentage of genes shared but we need the count of genes mismatched. Thus Hamming Distance is more appropriate than the Jaccard coefficient for this situation.

Chapter 3

8. Describe how a box plot can give information about whether the value of an attribute is symmetrically distributed. What can you say about the symmetry of the distributions of the attributes shown in Figure 3.11?

Solution :

- **Box Plot :**

- Box Plot is a way of graphically depicting groups of numerical data through their qualities.
- Box Plot is advantageous in determining whether the distribution in the data set is skewed or not.
- It also displays the shape, central value and variability of the distribution.

- **Information about Symmetrical Distribution :**

- The data is said to be Symmetrically Distributed (equally spaced) if the line representing the median is in the middle of the box and the remaining data which is outside the box is indicated by whiskers and outliers.
- If the data set consists of maximum number of small values with very few large ones, then the distribution is considered as Right Skewed.
- If the data set consists of maximum number of large values with very few small ones, then the distribution is considered as Left Skewed.
- In right skewed distribution, mean will be greater than median while it is lesser than median in case of left skewed distribution.

- **Symmetry of Distribution of attributes from Fig. 3.11 :**

- The attributes Sepal length and Sepal Width are relatively symmetrically distributed with median line at the center of distribution. Here, the values of mean and median are almost same.
- For the attribute petal length, the data seems to be more skewed towards the right (right skewed). Also the value of mean is greater than median.
- For the attribute petal width, the data seems to be somewhat skewed towards right.

Chapter 4

5. Consider the following data set for a binary class problem.

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
T	F	−

a) Calculate the information gain when splitting on A and B. Which attribute would the decision tree induction algorithm choose.

Solution :

5 (a)

Contigency Tables after Splitting :

For A :-

	A = T	A = F
+	4	0
−	3	3

For B :-

	B = T	B = F
+	3	1
−	1	5

Lets consider the entropy

$$\text{Entropy}(t) = -\sum_{i=0}^{c-1} p(i/t) \log_2 p(i/t)$$

where $p(i/t)$ denotes the fraction of records belonging to class i at a given node t.

Before Splitting

Number of '+' = 4

Number of '−' = 6

Total number of frequencies = 10

Therefore,

$$E_0 = -\frac{4}{10}\log_2\frac{4}{10} - \frac{6}{10}\log_2\frac{6}{10}$$

$$= -0.4 \log 0.4 - 0.6 \log 0.6$$

$$E_0 = 0.9710$$

After Splitting

For 'A'

∧ Information Gain:

$$\Delta = E_0 - \sum_{j=1}^{k} \frac{N(v_j)}{N} E(v_j)$$

$$E_{A=T} = -\frac{4}{7}\log_2\frac{4}{7} - \frac{3}{7}\log_2\frac{3}{7}$$

$$E_{A=T} = 0.9852$$

$$E_{A=F} = -\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3}$$

$$E_{A=F} = 0$$

So,

No. of frequencies with A=T is 7

No. of frequencies with A=F is 3

Total no. of frequencies for A is 10

Therefore, $\Delta = E_0 - 7/10\, E_{A=T} - 3/10\, E_{A=F}$

$$\Delta = 0.9710 - \frac{7}{10}(0.9852) - \frac{3}{10}(0)$$

$$\Delta = 0.9710 - 0.7(0.9852) - 0$$

$$\Delta = 0.9710 - 0.6897$$

$$\Delta = 0.2813$$

For 'B'

$$E_{B=T} = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}$$

$$\boxed{E_{B=T} = 0.8113}$$

$$E_{B=F} = -\frac{1}{6}\log\frac{1}{6} - \frac{5}{6}\log\frac{5}{6}$$

$$\boxed{E_{B=F} = 0.6500}$$

Therefore

$$\Delta = E_0 - \frac{4}{10}E_{B=T} - \frac{6}{10}E_{B=F}$$

No. of frequencies with B=T is 4

No. of frequencies with B=F is 6

Total no. of frequencies for B is 10

$$\Delta = 0.9710 - \frac{4}{10}(0.8113) - \frac{6}{10}(0.6500)$$

$$\Delta = 0.9710 - 0.4\,(0.8113) - 0.6\,(0.6500)$$

$$\Delta = 0.9710 - 0.3245 - 0.39$$

$$\boxed{\Delta = 0.2565}$$

Information Gain is a value of how much information we gained by doing the split using a particular feature.

Here, information gain is higher when splitting is done on attribute 'A'. Therefore attribute 'A' will be chosen to split the node.

5(b) Calculate the gain in the Gini index when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

Solution :

5(b) Contigency Table after Splitting

For A :

	A=T	A=F
+	4	0
−	3	3

For B :

	B=T	B=F
+	3	1
−	1	5

$$\text{Gini}(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2$$

Before Splitting

Number of '+' = 4

Number of '−' = 6

Total no. of frequencies = 10

Therefore,

$$G_0 = 1 - \left(\frac{4}{10}\right)^2 - \left(\frac{6}{10}\right)^2$$

$$= 1 - (0.4)^2 - (0.6)^2$$

$$= 1 - 0.16 - 0.36$$

$$G_0 = 0.48$$

After Splitting

For 'A'

Information Gain: $\Delta = G_0 - \sum_{j=1}^{k} \frac{N(V_j)}{N} G(V_j)$

$$G_{A=\bar{T}} = 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2$$

$$= 1 - (0.5714)^2 - (0.4285)^2$$

$$= 1 - 0.3265 - 0.1837$$

$$\boxed{G_{A=\bar{T}} = 0.4898}$$

$$G_{A=F} = 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2$$

$$\boxed{G_{A=F} = 0}$$

So,

No. of frequencies with $A=\bar{T}$ is 7

No. of frequencies with $A=F$ is 3

Total no. of frequencies with A is 10

Therefore,

$$\Delta = G_0 - \frac{7}{10} G_{A=\bar{T}} - \frac{3}{10} G_{A=F}$$

$$\Delta = 0.48 - \frac{7}{10}(0.4898) - \frac{3}{10}(0)$$

$$\Delta = 0.48 - 0.7\,(0.4898) - 0$$

$$\Delta = 0.48 - 0.3429$$

$$\boxed{\Delta = 0.1371}$$

After Splitting

For 'B'

$$G_{B=T} = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2$$

$$= 1 - (0.25)^2 - (0.75)^2$$

$$= 1 - 0.0625 - 0.5625$$

$$\boxed{G_{B=T} = 0.3750}$$

~~$G_{B=F} = 1 - (1/4)^2 - (3/4)^2$~~

$$G_{B=F} = 1 - \left(\frac{1}{6}\right)^2 - \left(\frac{5}{6}\right)^2$$

$$= 1 - (0.1667)^2 - (0.8333)^2$$

$$= 1 - 0.0277 - 0.6944$$

$$\boxed{G_{B=F} = 0.2778}$$

Therefore,

No. of frequencies with B=T is 4

No. of frequencies with B=F is 6

Total no. of frequencies for B is 10.

$$\Delta = G_0 - \frac{4}{10} G_{B=T} - \frac{6}{10} G_{B=F}$$

$$= 0.48 - \frac{4}{10}(0.3750) - \frac{6}{10}(0.2778)$$

$$= 0.48 - (0.4)(0.3750) - (0.6)(0.2778)$$

$$\boxed{\Delta = 0.1633}$$

In this, attribute 'B' will be chosen to split, since it has high information gain than attribute 'A'.

(c) Figure4.13 shows that entropy and the Gini index are both monotonously increasing on the range[0, 0.5] and they are both monotonously decreasing on the range [0.5, 1]. Is it possible that information gain and the gain in the Gini index favor different attributes? Explain.

Solution :

Yes, it is possible that information gain and gain in the Gini index favor different attributes.

As seen in results of part a and b, even though measures have similar range and monotonous behavior their information gain does not necessarily behave in the same way as the information gain is scaled differences of different measures.

---

7. The following table summarizes a data set with three attributes A, B, C and two class labels +, –. Build a two-level decision tree.

A	B	C	Number of Instances	
			+	–
T	T	T	5	0
F	T	T	0	20
T	F	T	20	0
F	F	T	0	5
T	T	F	0	0
F	T	F	25	0
T	F	F	0	0
F	F	F	0	25

a) **According to the classification error rate, which attribute would be chosen as the first splitting attribute? For each attribute, show the contingency table and the gains in classification error rate.**

Solution :

7 (a) The error rate is given by:

$$\text{classification error}(t) = 1 - \max_i [p(i/t)]$$

where $p(i/t)$ denotes the fraction of records belonging to class i at a given node t.

Before Splitting

No. of '+' = 50

No. of '-' = 50

Total no. of frequencies = 100

Therefore,

$$E_0 = 1 - \max\left(\frac{50}{100}, \frac{50}{100}\right)$$

$$= 1 - \frac{50}{100}$$

$$= \frac{100 - 50}{100}$$

$$E_0 = \frac{50}{100} = 0.5$$

After Splitting

Contigency Tables

For 'A'

	A=T	A=F
+	25	25
-	0	50

For 'B'

	B=T	B=F
+	30	20
-	20	30

For 'C'

	C=T	C=F
+	25	25
-	25	25

For 'A'

$$E_{A=T} = 1 - \max\left(\frac{25}{25}, \frac{0}{25}\right)$$

$$= 1 - \frac{25}{25}$$

$$= \frac{25 - 25}{25}$$

$$= \frac{0}{25}$$

$$\boxed{E_{A=T} = 0}$$

$$E_{A=F} = 1 - \max\left(\frac{25}{75}, \frac{50}{75}\right)$$

$$= 1 - \frac{50}{75}$$

$$= \frac{75 - 50}{75}$$

$$\boxed{E_{A=F} = \frac{25}{75}}$$

$$E_{A=F} = 1/3$$

$\frac{25}{75} = \frac{1}{3} = 0.333$

$\frac{50}{75} = \frac{2}{3} = 0.667$

Information Gain:

$$\boxed{\Delta = E_0 - \sum_{j=1}^{k} \frac{N(v_j)}{N} E(v_j)}$$

$$\Delta = E_0 - \frac{25}{100} E_{A=T} - \frac{75}{100} E_{A=F}$$

$$= \frac{50}{100} - \frac{25}{100}(0) - \frac{75}{100}\left(\frac{25}{75}\right)$$

$$= \frac{50}{100} - \frac{25}{100} = \frac{25}{100}$$

$$\boxed{\Delta = \frac{25}{100}}$$

For 'B'

$$E_{B=T} = 1 - \max\left(\frac{30}{50}, \frac{20}{50}\right)$$

$$= 1 - \frac{30}{50}$$

$$= \frac{50-30}{50}$$

$$\boxed{E_{B=T} = \frac{20}{50}}$$

$$E_{B=F} = 1 - \max\left(\frac{20}{50}, \frac{30}{50}\right)$$

$$= 1 - \frac{30}{50}$$

$$= \frac{50-30}{50}$$

$$\boxed{E_{B=F} = \frac{20}{50}}$$

Information Gain

$$\Delta = E_0 - \frac{50}{100} E_{B=T} - \frac{50}{100} E_{B=F}$$

$$= \frac{50}{100} - \frac{50}{100}\left(\frac{20}{50}\right) - \frac{50}{100}\left(\frac{20}{50}\right)$$

$$= \frac{50}{100} - \frac{20}{100} - \frac{20}{100}$$

$$= \frac{50-20-20}{100}$$

$$\boxed{\Delta = \frac{10}{100}}$$

For 'C'

$$E_{C=T} = 1 - \max\left(\frac{25}{50}, \frac{25}{50}\right)$$

$$= 1 - \frac{25}{50}$$

$$= \frac{50 - 25}{50}$$

$$\boxed{E_{C=T} = \frac{25}{50}}$$

$$E_{C=F} = 1 - \max\left(\frac{25}{50}, \frac{25}{50}\right)$$

$$= 1 - \frac{25}{50}$$

$$= \frac{50 - 25}{50}$$

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

Information Gain:

$$\Delta = E_0 - \frac{50}{100} E_{C=T} - \frac{50}{100} E_{C=F}$$

$$= \frac{50}{100} - \frac{50}{100}\left(\frac{25}{50}\right) - \frac{50}{100}\left(\frac{25}{50}\right)$$

$$= \frac{50}{100} - \frac{25}{100} - \frac{25}{100}$$

$$\boxed{\Delta = 0}$$

In this case, the information gain

For attribute A = $\frac{25}{100}$

For attribute B = $\frac{10}{100}$

For attribute C = 0

The information gain for attribute 'A' is greater than attribute 'B' and attribute 'C'.

Hence, the algorithm chooses attribute 'A' for splitting.

7. (b) Repeat for the two children of the root node

Solution :

7(b) From the given table, it can be noticed that when A=T, the child node is pure i.e whenever A=T, the class label is '+'. Hence, no further splitting is required.

But A=F needs further splitting and is shown below:

B	C	class label +	class label -
T	T	0	20
F	T	0	5
T	F	25	0
F	F	0	25

$$\text{Classification Error}(t) = 1 - \max_i [p(i|t)]$$

$$E_0 = 1 - \max\left(\frac{25}{75}, \frac{50}{75}\right)$$

$$= 1 - \frac{50}{75}$$

$$= \frac{75 - 50}{75}$$

$$\boxed{E_0 = \frac{25}{75}}$$

After Splitting

Contigency Tables

For 'B'

	B=T	B=F
+	25	0
-	20	30

For 'C'

	C=T	C=F
+	0	25
-	25	25

For 'B'

$$E_{B=T} = 1 - \max\left(\frac{25}{45}, \frac{20}{45}\right)$$

$$= 1 - \frac{25}{45}$$

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

$$E_{B=F} = 1 - \max\left(\frac{0}{30}, \frac{30}{30}\right)$$

$$= 1 - \frac{30}{30}$$

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

Information Gain: $\boxed{\Delta = E_0 - \sum_{j=1}^{k} \frac{N(v_j)}{N} E(v_j)}$

$$\Delta = \frac{25}{75} - \frac{45}{75} E_{B=T} - \frac{20}{75} E_{B=F}$$

$$= \frac{25}{75} - \frac{45}{75}\left(\frac{20}{45}\right) - \frac{20}{75}(0)$$

$$= \frac{25}{75} - \frac{20}{75} - 0 = 5/75 \qquad \boxed{\Delta = 5/75}$$

For 'C'

$$E_{C=T} = 1 - \max\left(\frac{0}{25}, \frac{25}{25}\right)$$

$$= 1 - \frac{25}{25}$$

$$= \frac{25 - 25}{25} = 0 \qquad \boxed{E_{C=T} = 0}$$

$$E_{C=F} = 1 - \max\left(\frac{25}{50}, \frac{25}{50}\right)$$

$$= 1 - 25/50$$

$$= \frac{50 - 25}{50} = \frac{25}{50} \qquad \boxed{E_{C=F} = \frac{25}{50}}$$

Information Gain:

$$\Delta = E_0 - \frac{25}{75} E_{C=T} - \frac{50}{75} E_{C=F}$$

$$= \frac{25}{75} - \frac{25}{75}\left(\frac{0}{25}\right) - \frac{50}{75}\left(\frac{25}{50}\right)$$

$$= \frac{25}{75} - 0 - \frac{25}{75}$$

$$\boxed{\Delta = 0}$$

Therefore, information gain for attribute 'B' = 5/75 & for attribute 'C' is 0.

Hence, attribute 'B' which has more information gain than 'C' is used for splitting.

7. (c) How many instances are misclassified by the resulting decision tree?

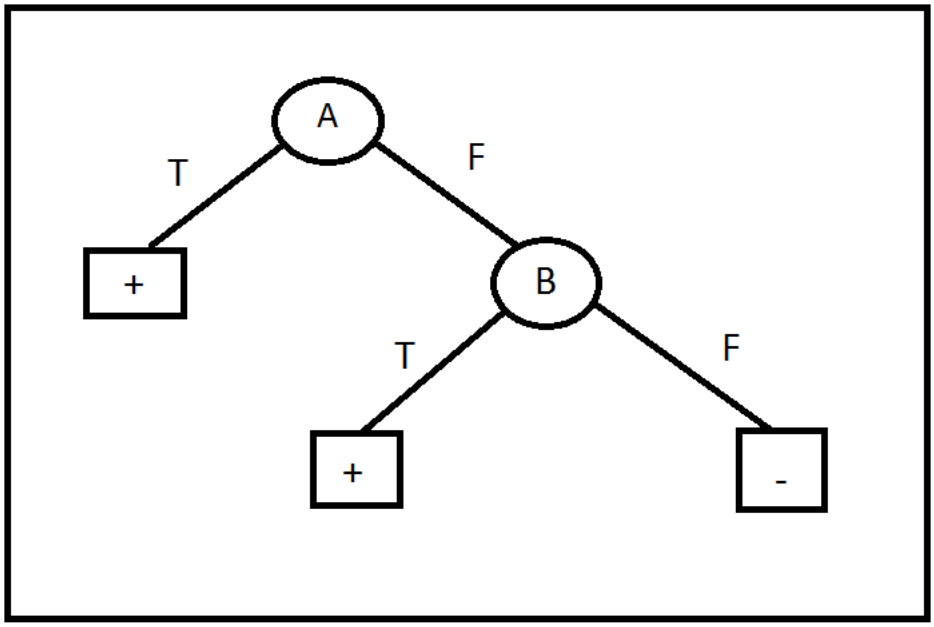
From the solution s provided to the previous questions i,e 7(a) and 7(b) we can notice that attribute 'A' is used for splitting the decision tree at the root.

Since A=T child node is pure, no further splitting is done on this node.

But for A=F child node, attribute 'B' is used for further split.

Decision Tree :

When A=F if we consider that B=T gives class '-' and B=F gives class '+' it can be seen that number of misclassifications will be 50. But if we consider that B=T gives class '+' and B=F gives class '-' as shown in the above decision tree fig. the number of misclassifications will be 20.



Therefore, considering the above decision tree and the given table it can be noticed that 20 instances are misclassified.

Hence error rate : 20/100.

7. (d) Repeat parts (a), (b), and (c) using C as the splitting attribute

Solution :

4(d)

Here the top node for splitting is 'C'.
Both the child nodes i,e C=T and C=F are not pure. Therefore, they need further splitting.

For C=T

A	B	Number of Instances	
		+	-
T	T	5	0
F	T	0	20
T	F	20	0
F	F	0	5

Classification Error(t) = $1 - \max_i [p(i|t)]$

$$E_0 = 1 - \max\left(\frac{25}{50}, \frac{25}{50}\right)$$

$$= 1 - \frac{25}{50}$$

$$= \frac{50 - 25}{50}$$

$$\boxed{E_0 = 25/50}$$

After Splitting

Contigency Tables:

For 'A'

	A=T	A=F
+	25	0
-	0	25

For 'B'

	B=T	B=F
+	5	20
-	20	5

For 'A'

$$E_{A=T} = 1 - \max\left(\frac{25}{25}, \frac{0}{25}\right)$$

$$= 1 - \frac{25}{25}$$

$$= \frac{25 - 25}{25}$$

$$\boxed{E_{A=T} = 0}$$

$$E_{A=F} = 1 - \max\left(\frac{0}{25}, \frac{25}{25}\right)$$

$$= 1 - \frac{25}{25}$$

$$= \frac{25 - 25}{25}$$

$$\boxed{E_{A=F} = 0}$$

Information Gain:

$$\boxed{\Delta = E_0 - \sum_{j=1}^{k} \frac{N(v_j)}{N} E(v_j)}$$

$$\Delta = E_0 - \frac{25}{50} E_{A=T} - \frac{25}{50} E_{A=F}$$

$$\Delta = \frac{25}{50} - \frac{25}{50}(0) - \frac{25}{50}(0)$$

$$\Delta = \frac{25}{50} - 0 - 0$$

$$\boxed{\Delta = \frac{25}{50}}$$

For 'B'

$$E_{B=T} = 1 - \max\left(\frac{5}{25}, \frac{20}{25}\right)$$

$$= 1 - \frac{20}{25}$$

$$= \frac{25 - 20}{25}$$

$$\boxed{E_{B=T} = 5/25}$$

$$E_{B=F} = 1 - \max\left(\frac{20}{25}, \frac{5}{25}\right)$$

$$= 1 - \frac{20}{25}$$

$$= \frac{25-20}{25}$$

$$\boxed{E_{B=F} = 5/25}$$

Information Gain:

$$\Delta = E_0 - \frac{25}{50} E_{B=T} - \frac{25}{50} E_{B=F}$$

$$\Delta = \frac{25}{50} - \frac{25}{50}\left(5/25\right) - \frac{25}{50}\left(5/25\right)$$

$$\Delta = \frac{25}{50} - \frac{5}{50} - \frac{5}{50}$$

$$\Delta = \frac{25}{50} - \frac{10}{50}$$

$$\boxed{\Delta = \frac{15}{50}}$$

Therefore, attribute 'A' has more gain than 'B'. Hence attribute 'A' is used for splitting.

For C = F

A	B	Number of Instances	
		+	−
T	T	0	0
F	T	25	0
T	F	0	0
F	F	0	25

Classification Error (t) = $1 - \max_i [p(i|t)]$

$$E_0 = 1 - \max\left(\frac{25}{50}, \frac{25}{50}\right)$$

$$= 1 - \frac{25}{50}$$

$$= \frac{50 - 25}{50}$$

$$\boxed{E_0 = 25/50}$$

After Splitting

Contigency Tables

For 'A'

	A = T	A = F
+	0	25
−	0	25

For 'B'

	B = T	B = F
+	25	0
−	0	25

For 'A'

$$E_{A=T} = 1 - \max\left(\frac{0}{0}, \frac{0}{0}\right)$$

$$= 1 - \frac{0}{0}$$

$$\boxed{E_{A=T} = 0}$$

$$E_{A=F} = 1 - \max\left(\frac{25}{50}, \frac{25}{50}\right)$$

$$= 1 - \frac{25}{50}$$

$$\boxed{E_{A=F} = \frac{25}{50}}$$

Information Gain:

$$\Delta = E_0 - \sum_{j=1}^{k} \frac{N(v_j)}{N} E(v_j)$$

$$\Delta = E_0 - \frac{0}{50} E_{A=T} - \frac{50}{50} E_{A=F}$$

$$\Delta = \frac{25}{50} - \frac{0}{50}(0) - \frac{50}{50}\left(\frac{25}{50}\right)$$

$$\Delta = \frac{25}{50} - 0 - \frac{25}{50}$$

$$\boxed{\Delta = 0}$$

For 'B'

$$E_{B=T} = 1 - \max\left(\frac{25}{25}, \frac{0}{25}\right)$$

$$= 1 - \frac{25}{25}$$

$$\boxed{E_{B=T} = 0}$$

$$E_{B=F} = 1 - \max\left(\frac{0}{25}, \frac{25}{25}\right)$$

$$= 1 - \frac{25}{25}$$

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

Information Gain:

$$\Delta = E_0 - \frac{25}{50} E_{B=T} - \frac{25}{50} E_{B=F}$$

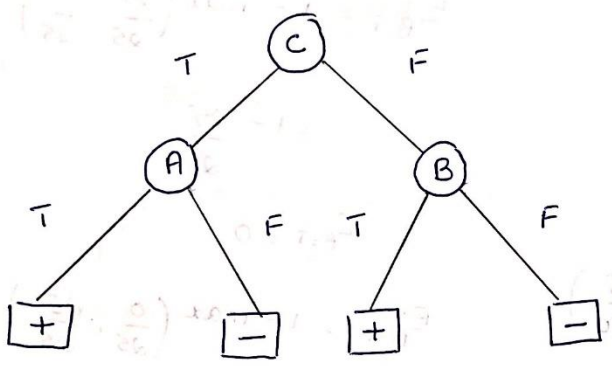
$$\Delta = \frac{25}{50} - \frac{25}{50}(0) - \frac{25}{50}(0)$$

$$\Delta = \frac{25}{50} - 0 - 0$$

$$\boxed{\Delta = \frac{25}{50}}$$

In this case, the information gain for attribute 'B' is more and hence it is considered for splitting.

Decision Tree :



From the above decision tree and the given table, it can be noticed that none of the instances are misclassified.

Hence, we can conclude that overall error rate is 0.

7. (e) Use the results in parts (c) and (d) to conclude about the greedy nature of the decision tree induction algorithm

From the above results we can notice that the error rate is 20 when attribute ‘A’ is used for splitting. But it is 0 when attribute ‘C’ is used.

Considering these results we can infer that the greedy nature of the decision tree induction algorithm does not always lead to the best tree.
