Assignment 3

1. **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

     Hamming Distance is used denote the number dissimilar bits in binary string. Hamming distance can only be calculated if the length of the two binary strings is same.

     Hamming distance can be calculated by adding one bit from each sting and then adding the results of each one bit addition (carry is discarded).

     0101010001

     \+ 0100011000

     = 0001001001

     After that we sum up the bits of the result

     I.e. 0+0+0+1+0+0+1+0+0+1= 11 which 3 in decimal

     Hence Hamming distance is 3

   - Jacckard Similarity

     Jacckard Similarity for binary stings is given by the following formula:

     Jacckard Similarity= Number of 1's matched in both stings/ Number of 0's matched in both sting.

Thus the Jacckard Similarity compares both the string individual bits calculates similar bits i.e. 1's and 0's

Thus

Jacckard Similarity = 2/5 = 0.40

b) 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:

Hamming distance compares the two binary strings and gives the count between of different bits. Jaccard similarity gives ratio between the number of 1's to number of 0's common in both strings. Thus if the comparison between the organisms in terms of number of gene's they share would be better explained by Jaccard similarity.

c) 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:

The question states that the two human beings under comparison have 99.9 % of the same genes. The Jaccard similarity gives the ratio of similar attributes and hence the application of this mechanism to the data under consideration may not benefit to get result from the given data. Hamming distance gives the number of dissimilarities among the data.

2. 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:

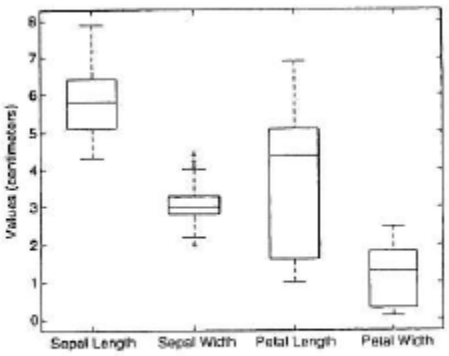

Figure 3.11. Box plot for Iris attributes.

- The above figure shows the Box plot of Iris Attributes like sepal length, sepal width, Petal Length, Petal Width. The information about the symmetry of the data can be given with the help of an attribute like mean or median which is calculated over the data.
- In the given graph, if the line located in the box is assumed to be the line representing median of the data then the conclusion about the distribution of the data.
- The line if located in the exact centre of the box it implies that the data in the box is evenly distributed.
- The application of the above assumption to the given box plot helps us interpret that the attributes Sepal width and Sepal length are evenly distributed.
- The distribution for petal width and petal length is not even some of the points are concentrated towards one end only.

3. 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:

The Entropy is calculated as:

- Total Entropy

  Total class labels are 2 i.e. "+" and "-"with 10 instances.

  Therefore total entropy is

  E = - (4/10) log (4/10) – (6/10) log (6/10)

  = - (0.4)*(-1.32) – (0.6)*(-0.736)

  = 0.528 + 0.438

  = 0.97

- After Splitting on A

A	+	-
T	4	3
F	0	3

Entropy for A=T is

E= - 4/7 log (4/7) - 3/7 log(3/7)

= -(0.57)*(-0.80) – (0.43)*(-1.22)

= 0.98

Entropy for A=F is

E = −3/3 log 3/3 – 0/3 log 0/3

= 1*0 – 0*(-infinity)

= 0

- After Splitting on B

B	+	-
T	3	1
F	1	5

Entropy for B=T is

$E = -3/4 \log(3/4) - 1/4 \log(1/4)$

$= -(0.75)*(-0.41) – (0.25)*(-2)$

$= 0.81$

Entropy for A=F is

$E = -1/6 \log 1/6 – 5/6 \log 5/6$

$= 0.16*(-2.64) – 0.833*(-0.268)$

$= 0.65$

- Information Gain

Information gain for A

$\Delta = E – 7/10(E(A=T) – 3/10(E(A=F)$

$= 0.97 – (0.7*0.99)-(0.3*0)$

$= 0.28$

Information gain for B

$\Delta = E – 4/10(E(B=T) – 6/10(E(B=F)$

$= 0.97 – (0.4*0.81)-(0.6*0.65)$

$= 0.25$

From the above calculation we can say that attribute A has more information gain, hence it can be used as decision tree classifier.

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

Solution:

The Gini Index can be calculated as follows:

$Gini = 1 - \sum p\,(i/t)^2$

Overall Gini Index is

$Gini = 1 – (4/10)^2 – (6/10)^2$

= 1- 0.16 -0.36

=0.48

- Gini Index after splitting on A

Gini (A=T) = $1 – (4/7)^2 – (3/7)^2$

= 0.48

Gini (A=F) = $1 – (1)^2 – (0)^2$

= 0

Information Gain

$\Delta = 1 – 0.7*G(A=T) -0.3*G(B=T)$

= 0.14

- Gini Index after splitting on B

Gini (B=T) = $1 – (1/4)^2 – (3/4)^2$

= 0.37

Gini (A=F) = $1 – (1/6)^2 – (5/6)^2$

= 0.27

Information Gain

$\Delta = 1 – 0.4*G(A=T) -0.6*G(B=T)$

= 0.16

From the above result it can be concluded that attribute B should be selected as decision tree classifier

c) Figure 4.13 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.

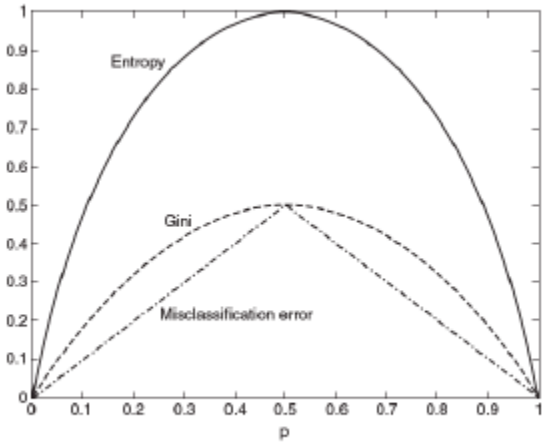

Figure 4.13. Comparison among the impurity measures for binary classification problems.

- The above figure shows classification of based on Gini index and Entropy. The figure shows that for evenly distributed data the Gini index and entropy values are maximum.
- Information gain is may behave similar to entropy and Gini index and entropy because it scaled difference of entropy and Gini index.
- But Information gain may also be dependent on other attributes which may vary the value of the information gain.

4. 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:

Classification error is given as

$$E=1-\max(p(i/t))$$

- Original Error Rate
  E= 1- max (50/100, 50/100)

  = 1 -50/100 = 0.5

- Error rate after partition on A

A	+	-
T	25	0
F	25	50

E (A=T) = 1- max (25/25, 0/50)
= 1 -50/100
= 0

E (A=F) = 1- max (25/75, 50/75)
= 1 -50/75
= 25/75
=0.66

Information Gain

Δ = E – 25/100 E(A=T) – 75/100 E(A=F)

= 1/4 =0.25

- Error rate after partition on B

B	+	-
T	25	25
F	25	25

E (B=T) = 1- max (25/50, 25/50)
= 1 -25/50
= 0.5

E (B=F) = 1- max (25/50, 25/50)
= 0.5

Information Gain

$\Delta = E - 50/100\ E(B=T) - 50/100\ E(B=F)$

$= 0.1$

- Error rate after partition on C

C	+	-
T	25	0
F	25	50

E (C=T) = 1- max (25/25, 0/50)
= 1 -50/100
= 0

E (C=F) = 1- max (25/75, 50/75)
= 1 -50/75
= 25/75
=0.66

Information Gain

$\Delta = E - 50/100\ E(C=T) - 50/100\ E(C=F)$

$= 0$

Attribute A has the highest gain

b) Repeat for the two children of the root node

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

Attribute A was selected as level 1 classifier

Original Error rate

$E = 1 - \max(25/75, 50/75)$

$= 25/75$

- Error rate after partition on B

B	+	-
T	25	20
F	0	30

E(B=T) = 1- max (25/45, 20/45)
= 1 -25/45
= 20/45

E(B=F) = 1- max (0/30, 30/30)
= 0

Information Gain

Δ = E– 45/75 E(B=T) –30/75E(B=F)

= 1/15

- Error rate after partition on C

C	+	-
T	0	25
F	25	25

E (B=T) = 1- max (25/25, 00/25)
= 1 -25/25
= 0

E (B=F) = 1- max (25/50, 25/50)
= 25/50

Information Gain

Δ = E– 25/75 E(B=T) –50/75E(B=F)

= 0

Attribute B has the highest gain

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

Solution:

The number of misclassified instances are 20.

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

When C is used as Splitting attribute, C can have 2 values T and F

i. C=T

Original error rate:

E=1 -25/50

=25/50

- Error rate after partition on A

A	+	-
T	25	0
F	0	25

E (A=T) =0

E (A=F) = 0

Information Gain

$\Delta = 0.5$

- Error rate after partition on B

B	+	-
T	5	20
F	20	5

E(B=T) = 1- max (5/25, 20/25)

= 1 -20/25

=1/5

E (B=F) = 1- max (5/25, 20/25)

= 1/5

Information Gain

$\Delta = 3/10 = 0.3$

Attribute A is the Splitting attribute

ii. C=F

Original error rate

E = 0.5

- Error rate after partition on A

A	+	-
T	0	25
F	0	25

E(A=T) = 1- max (25/25, 0/25)
= 1 -1
= 0

E (A=F) = 1- max (0/25, 25/25)
= 0

Information Gain

$\Delta = 0$

- Error rate after partition on B

B	+	-
T	25	0
F	0	25

E (B=T) = 1- max (0/25, 25/25)
= 0

E (B=F) =0

Information Gain

$\Delta = 0.5 – 0 -0$

= 0.5

Information gain for B is greater than A hence B is chosen as the classification attribute.

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

Solution

The Classification technique is not effective method as seen in (a) example A is chosen then B, but for the (d) example when C is chosen as 1^{st} level classifier the classification is more accurate and tree is well formed. Hence this method is not an accurate way to classify better methods can be used.