CS422 DATA MINING HW - 3

JINYANG LI - A20317851

Chapter 2
- #18 (a,c,d)

(a)
x = 0101010001

y = 0100011000

Hamming distance handled the number of the differences of bits between two binary vectors.

So, for this two vectors, the Hamming distance = 3. The difference bits are underlined.
x = 010<u>1</u>01<u>0</u>00<u>1</u>

y = 010<u>0</u>01<u>1</u>00<u>0</u>

Jaccard Similarity handled the rate '1' and '1' match / (number of bits minus number of '0' and '0' match).. so for this two vectors,

JS = 2/(10-5) = 0.4

(C)

suppose we have animal A 1111111110 ….. (UP TO 100 BITS, ALL LEFT IS RANDOM)

animal B 1111111111 ……(UP TO 100 BITS, ALL LEFT IS RANDOM)

Absolutely, hamming distance can handle a lot of differences, however, it's useless since it's two different classes. so it's bad, plus we know nothing about what DNA is sharing with these two animals.

so, Jaccard Similarity will be my choice since it can let us know how many DNSs are the two animals sharing.

(D)

Using the assume in C, however, change animal to person A and the 'ALL LEFT IS RANDOM' to '1'

it's absolutely hamming distance is better that we can know the difference in the two very similar objects - humans.

Chapter 3
- #8

Since the box plot is creating for presenting the distributing of the data. So, when the line that representing the median of the date is in the middle of the box plot, the data is symmetrically distributed.

For attributes shown in Figure 3.11,

the length, width of sepal are symmetrically distributed.

However, since the line of 50% percentage which points to petal length is in the very top position of the box,

petal length is not symmetrically distributed.

and petal width, the box plot is very like petal length box plot, not symmetrically distributed.

Chapter 4
- # 5 a,b,c
(a)
Entropy before splitting = - (4/10) log (4/10) - (6/10) log (6/10) = 0.971

For A

	T	F
class label +	4	0
class label -	3	3

so the information gain of A after splitting can evaluated as:

GA= (ETPbefore->entropy before splitting) — [7/10 (-4/7log4/7 - 3/7log3/7) + 3/10(0)] = 0.281

For B

	T	F
class label +	3	1
class label -	1	5

GB=(ETPbefore->entropy before splitting) - [4/10 (-3/4 log 3/4 - 1/4 log 1/4) + 6/10 (-1/6 log 1/6 - 5/6 log 5/6)] = 0.257

Since GA > GB, A will be the one that decision tree induction algorithm choose

(b)

gini before splitting = 1 - (0.4^2 + 0.6^2) =0.48

for A

GA = 0.48 - 7/10 (1-(4/7)^2 - (3/7)^2) -0 = 0.137

for B

GB = 0.48 - 4/10 (1- 1/16 - (3/4) ^ 2) - 6/10 (1 - 1/36 - 25/36) = 0.163

Since GA < GB, B will be the one that decision tree induction algorithm choose

(c)

It is possible.

Since the methods of calculation of these two is not the same. So may they have the different attributes.

- #7 a,b,c,d,e

(a)

Error rate before splitting:

Erate = 50/100 =0.5

the gain after splitting for A

Ga = 0.5 - 25/100 * (1- 25/25) - 75/100 * （1-50/75） = 25/100

the gain after splitting for B

Gb = 0.5 - 50/100 * (20/50) - 50/100 * (20/50) = 10/100

the gain after splitting for C

Gc = 0.5 - 50/100 * (25/50) - 50/100 * (25/50) = 0

So, the attribute A will be choose since its gain bigger than other one.

(b)

Since the T value of A node no need splitting, so the

Ga = 25/75

the gain after splitting for B

Gb = 25/75 - 45/75 * (20/45) -20/75(0) = 5/75

the gain after splitting for C

Gc = 25/75 - 25/75 * (0/25) - 50/75 (25/50) = 0

so it will choose B .

(c)

100 * (20/100) = 20

(d)

GA = 25/50
GB=15/50
A will be chosen.

(e)

The greedy nature is not very important of the decision tree induction algorithm.