CSEC 19h – Past Year Paper Solution AY2018/2019 Semester 1
CE4032/CZ4032 – Data Analytics And Mining

Nathanael S Raj, nathanae001@e.ntu.edu.sg

Question 1:

(a):

i. F, Descriptive
ii. F, Clustering would be better
iii. T
iv. T, OLAP is for quick queries
v. F, Rank is ordinal (has order)
vi. F, Not quantitative when judged by humans
vii. F, Median cannot be computed algebraically
viii. F?, Usually No?
ix. F, Random Sampling does not drop features
x. F?, Phasing weird… Data warehouse can be used to store new data too

(b):

The data mining process would begin with data collection. The burst would first be detected by sensors which would collect information about it. This information would comprise of features. With these features, we could perform clustering algorithms on the data. Given that two classes already exist, we could detect if all the ray data fitted into the distribution of the first two classes. If outliers were detected, clustering algorithms like DBSCAN and CURE could be used to find the new class.

(c):

First, we perform stop-word removal on all articles, i.e. semantically insignificant words like ‘a’, ‘the’, ‘me’ etc. are removed. Next, the news articles are comprised into a matrix, where the ij-th entry represents the count of the jth word in the ith newspaper. Following which we rank the words by frequency and remove all but 10 most popular words from the matrix. The resultant matrix will represent a 10-dimensional vector for each news article.

(d):

$$SMC = \frac{M00 + M11}{M00 + M01 + M10 + M11} = \frac{4 + 2}{10} = \frac{3}{5}$$

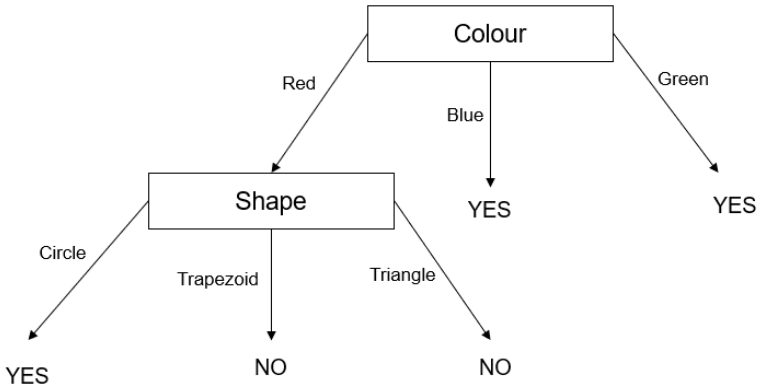
$$Jaccard = \frac{M11}{M01 + M10 + M11} = \frac{2}{6} = \frac{1}{3}$$

$$Euclidean = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2} = \sqrt{1^2 + 1^2 + 1^2 + 1^2} = 2$$

$$cosine = \frac{(p \cdot q)}{\|p\|\|q\|} = \frac{2}{\sqrt{4}\sqrt{4}} = \frac{1}{2}$$

Question 2:

(a):



(b):

$$Entropy(parent) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 2.03$$

$$Gain_{Shape} = 2.03 + triangle + trapezoid + circle + rectangle$$

$$= 2.03 + \frac{2}{7}(0\log_2 0 + 1\log_2 1) + \frac{2}{7}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{2}{7}(0\log_2 0 + 1\log_2 1)$$
$$+ \frac{1}{7}(0\log_2 0 + 1\log_2 1) = 2.03 + 0 - \frac{2}{7} + 0 + 0 = 1.74$$

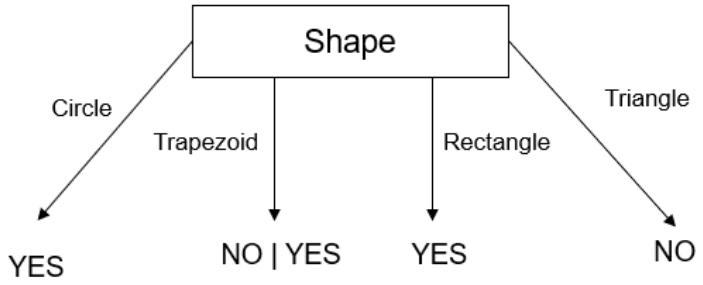
$$Gain_{colour} = 2.03 + red + blue + green$$

$$= 2.03 + \frac{4}{7}\left(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) + \frac{1}{7}(0\log_2 0 + 1\log_2 1) + \frac{2}{7}(0\log_2 0 + 1\log_2 1)$$
$$= 2.03 + \frac{4}{7}(-0.811) + 0 + 0 = 1.57$$

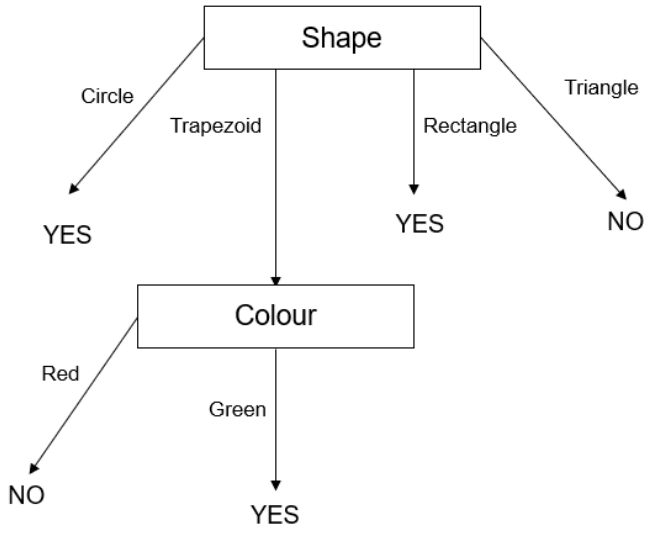
$$Gain_{type} = 2.03 + A + B + C$$

$$= 2.03 + \frac{2}{7}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{1}{7}(0\log_2 0 + 1\log_2 1) + \frac{4}{7}\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right)$$
$$= 2.03 + \frac{2}{7}(-1) + 0 + \frac{4}{7}(-1) = 1.17$$

Shape gives most information gain split on shape



By observation, color gives the most information, next split on color. (Type is C for all trapezoids)



(c):

Using Decision Tree 2(a), it will be classified as YES, since it is blue.

Using Decision Tree 2(b), it is classified as NO, since it is triangle.

CSEC 19h – Past Year Paper Solution AY2018/2019 Semester 1
CE4032/CZ4032 – Data Analytics And Mining

Question 3:

(a):

Since both variables have similar ranges (Variable A is from 0 to 10, while B is from 1 to 10) and similar increments in value, NO NORMALIZATION is necessary.

Sample	Distance from [6,7]
S1	7.81
S2	6.40
S3	5
S4	3.61
S5	3.16
S6	3.61
S7	4.47
S8	6.71
S9	6.32

Using k=3. Classify malignant

Using k=5, classify benign

(b):

Cluster	Initial Centroids
C1	2.5, 2.5
C2	5, 4
C3	9, 9.5
C4	2, 2.5

	Distance from cluster				
Sample	C1	C2	C3	C4	Chosen Cluster
S1	3	7	16.5	2.5	C4
S2	1	5	14.5	0.5	C4
S3	1	3	12.5	1.5	C1
S4	3	1	10.5	3.5	C2
S5	4	0	9.5	4.5	C2
S6	13	9	1.5	13.5	C3
S7	14	10	1.5	14.5	C3
S8	4	5	14.5	3.5	C4
S9	3	4	13.5	3.5	C1

New clusters:

C1: S3, S9

C2: S4, S5

C3: S6, S7

C4: S1, S2, S8

(c):

Jarvic-Patrick Clustering

First, the KNN of all points are found. In graph terms, this can be regarded as breaking Q4 but the K strongest links from each point in a proximity graph. Next, a pair of points is put in the same cluster if any 2 points share more than T neighbors and the two points are in each other's KNN list.

(d):

There will be more outliers.

With more stringent criteria for a core point, there will be fewer core points. Border points that used to be within the neighborhood of a core point would become noise if that core point became a border point. Therefore, more chance of having noise points.

Question 4:

(a):

(i): If an itemset is frequent, then all of its subsets must also be frequent. In other words, support of an itemset never exceeds support of its subsets.

to reach min_sup = 2/9 x 8 = 1.8

(ii):

1-candidates	
Itemset	Support
1	6
2	6
3	6
4	1
5	2

Freq 1-itemsets	
Itemset	Support
1	6
2	6
3	6
5	2

2-candidates	
Itemset	Support
1, 2	4
1, 3	4
1, 5	2
2, 3	4
2, 5	2
3, 5	1

Freq 2-itemsets	
Itemset	Support
1, 2	4
1, 3	4
1, 5	2
2, 3	4
2, 5	2

3-candidates	
Itemset	Support
1, 2, 3	2
1, 2, 5	2

Freq 3-itemsets	
Itemset	Support
1, 2, 3	2
1, 2, 5	2

(iii):

1, 2 →3, confidence = 2/4

1, 3 →2, confidence = 2/4

2, 3 →1, confidence = 2/4

1, 2 →5, confidence = 2/4

1, 5 →2, confidence = 2/2

2, 5 →1, confidence = 2/2