BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
HYDERABAD CAMPUS
SECOND SEMESTER 2018 – 2019
Data Mining (CS F415) – COMPREHENSIVE EXAM (REGULAR)

Date: 11.05.2019 **Weightage:45%(90M)**
Duration: 3 hrs **Type: Closed Book**
Note: Answer all parts of a question sequentially
No of pages in the question paper: 3

Q1. Association Rule Mining **[5+2+3+5+4=19M]**

Transaction ID	Items
T1	A B D E
T2	B C E
T3	A B D E
T4	A B C E
T5	A B C D E
T6	B C D

Table -1

a. Use the transactional database given in Table-1 with minimum support count s=4 and build a frequent pattern tree (FP-Tree). Show for each transaction how the tree evolves.

b. Generate the conditional FP Tree for item D.

c. List all the frequent item sets with suffix D.

d. Suppose you are a computational linguist and wish to find commonly used phrases in English literature by Shakespeare. Is it possible to solve this problem using association rule mining? If yes provide your solution, if no argue why it is not possible?

e. Give an example of a Multi-dimensional association rule and Quantitative association rule.

Q2. Clustering **[3+3+3+6+4+8+4=31 M]**

a. The K-Means algorithm usually converge to a local optimum rather than a global one. Given this drawback, how would you increase the chances of finding a good solution?

b. Given a dataset of N Samples, you applied K-means and Gaussian Mixture Model for clustering. Both the algorithms gave you 5 clusters with exactly the same cluster centers. Can 3 points that are assigned to different clusters in the k-means solution be assigned to the same cluster in the Gaussian mixture solution? If no, explain. If so, sketch an example or explain in 1-2 sentences.

c. Consider the dataset {0,4,5,20,25,39,43,44} for which we intend to get two clusters using hierarchical clustering. Suppose the output from single link and complete link give the same clustering as {0,4,5} {20,25,39,43,44}. Now if we run average link is there a chance of getting different clustering?

d. Suppose that we are fitting a Gaussian Mixture Model for data items consisting of a single real value, x, using K = 2 components. We have N = 5 training cases, in which the values of x are as follows: 5, 15, 25, 30, 40. We use the EM algorithm to find the maximum likelihood estimates for the model parameters, which are the mixing proportions for the two components, $\pi 1$ and $\pi 2$, and the means for the two components, $\mu 1$ and $\mu 2$. The standard deviations for the two components are fixed at 10.
Suppose that at some point in the EM algorithm, the E step found that the responsibilities of the two components for the five data items were as follows:

r_{i1}	r_{i2}
0.2	0.8
0.2	0.8
0.8	0.2
0.9	0.1
0.9	0.1

Compute the value of parameters $\pi 1$, $\pi 2$, $\mu 1$, and $\mu 2$ in the next M step of the algorithm?

e. Write the equation that represents the mixture of Gaussians and explain the significance of each component. Are all the estimated parameters free or is there a constraint if so what is it?

f. Given the following dataset in 1-d space as shown in Figure 1, it consists of 3 positive data points {−1, 0, 1} and 3 negative data points {−3, −2, 2}. Answer i-iv

Figure 1

i. Since the data is not linearly separable using the feature map y1=X and y2=X^2 project the data to a 2-d feature space (y1, y2) so that the data becomes linearly-separable. Plot the dataset after mapping in 2-d space.

ii. What will be the size of the gram kernel matrix?

iii. What is the value of the entry [5,6] in the gram kernel matrix?

iv. In the objective function of Kernel K-Means why all the terms are expressed using kernel operations?

g. Table-2 shows the confusion matrix for 3 clusters C1, C2 and C3 for the ground truth T1 to T6

C/T	Entertainment (T1)	Finance(T2)	Foreign(T3)	Metro(T4)	National(T5)	Sports(T6)	Total
C1	1	1	0	11	4	676	693
C2	27	89	33	827	253	33	1562
C3	326	465	8	105	16	29	949
Total	354	555	341	943	273	738	3204

Table-2

i.Compute the purity of the clustering? What the interpretation of this value?

ii. Compute the precision and recall for C1?

Q3. Outlier Analysis **[5+3+3+3+3=17M]**

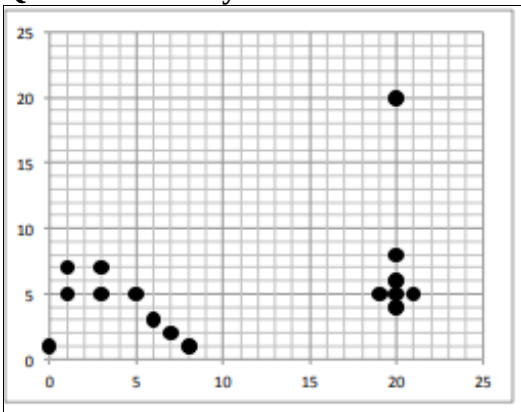

Figure-2

Consider the data shown in the Figure-2 and the following models are applied

- A density based outlier detection algorithm (LOF) with the parameters K=3 and LOF threshold = 1.5
- A KNN distance based outlier detection algorithm with parameter K=1 distance threshold = 3 answer the questions a) and b).

a. If the given density-based approach is applied, will it identify the object with coordinates (20, 8) as an outlier? show your calculation.

b. Assume that the set of outliers that you identified using the given density-based approach is O1, and the set of outliers that you identified using the given distance-based approach is O2. Will O1 and O2 be the same? **(Note: you do not need to calculate O1 and O2).** If they are the same, Explain the reason. If they are different, Identify one object that is in O2 but not in O1 or identify an object that is in O1 but not in O2.

c. The following values represent the potassium levels of patients in a hospital 0.217, 0.224,0.195, 0.221, 0.221, 0.223. Apply Grubbs test to verify if 0.195 is an outlier**. [Note: Use the below values given from the standard G-table for comparison]**

Degrees of freedom	**G table values (95% confidence)**
4	1.463
5	1.672
6	1.822

d. Provide a visual example of a dataset in which LOF fails to detect an outlier.

e. Can we solve the Outlier detection using supervised learning technique? If so what issues do you foresee?

Q4. Classification & Ensembles **[3+3+2+3+5+4+3=23M]**

a. We know that KNN is a lazy classifier since we need to store all training instances until test time. Suppose for a problem we are able to draw a decision boundary for the 1-NN classifier. Suppose we store the decision boundary learnt instead of storing all training data, would that always result in an improvement in terms of storage (memory) requirement for this classifier?

b. Assume that you are designing a classifier where you want to classify a tweet as toxic or not (tweets are continuously streaming). Which classifier would use KNN or Naïve Bayes and why?
Note: Your answer would not be awarded any marks without justification.

c. Given a large training data set what is the major drawback of using a K Nearest Neighbor classifier during testing?

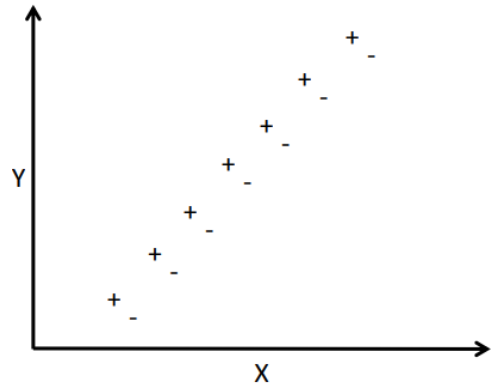

Figure-3

d. Suppose we have the data, represented using two real-valued features (X and Y, as shown in Figure-3) and our goal is to randomly split this data into a training set (90%) and a test set (10%) and to train and evaluate a model.

Which classifier do you think would have a higher chance of doing well in terms of accuracy: KNN (with K=1) or Naïve Bayes? Why?

e. You are given a data set with 3 Boolean input variables, X1, X2 and X3, and one Boolean output, Y .

i. How many parameters must be estimated to train a naïve Bayes classifier? **(you need not list them, just give the total)**

ii. How many parameters would have to be estimated to learn the above classifier if we do not make the naive Bayes conditional independence assumption?

f. You have drawn two random sub samples of size 5000 and 1,00,000 from a very large dataset. The train and test sets are generated by randomly splitting the data 90:10. Draw two curves for training and test error for each dataset (5000 and 1,00,000) with X axis as model complexity and Y-axis as error.
Note: You will totally have 4 curves in a single graph name them as 5000 test, 100k test,5000 train and 100k train.

g. Suppose you are working on a classification problem where the training data is very limited, which ensemble technique, would you prefer, and why?

********** *That's all folks* **********