BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI HYDERABAD CAMPUS

**SECOND SEMESTER 2016 – 2017**

**Data Mining (CS F415)**

**TEST-1 REGULAR**

**Date: 23.02.2017 Weightage: 20 %( 40 M) Duration: 60min. Type: Closed Book**

**Q1.** Discuss whether each of the following tasks can be considered as Data Mining Tasks or not. Nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn **[5 M]**

1. Finding the total number of people who listen to Carnatic music on a music portal.
2. Predicting whether a new song album released will be successful or not based on historic data.
3. Finding the number of shoes of size 6 sold in a store.
4. Extracting all the words from a text document.
5. Identifying what type of products are cross sold based on buying patterns.

**Q2.** Classify the data type of each of the following as nominal, ordinal, interval or ratio. **[5 M]**

1. Blood group : A,AB,O,B
2. Socioeconomic status:  poor, middle class, rich.
3. Number of children
4. Blood Pressure
5. IQ level

**Q3.** An analyst collects surveys from different participants about their likes and dislikes. Subsequently, the analyst uploads the data to a database, corrects erroneous or missing entries, and designs a recommendation algorithm on this basis. Which of the following actions represent data collection, data preprocessing, and data analysis? **[3M]**

**a)** Conducting surveys and uploading to database

**b)** Correcting missing entries

**c)** Designing a recommendation algorithm

**Q4.**

1. Given this 2-D dataset as shown in **Figure-1** with two classes represented as + and -, draw a decision boundary to classify the sample using 3 Nearest Neighbors. [**2M]**

+

---

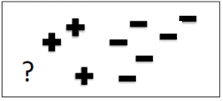
sample

+ +

-

**Figure-1**

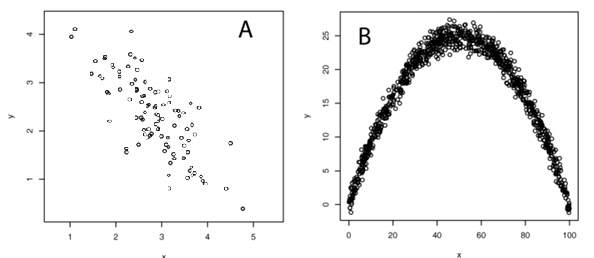
1. Given the following labelled data set as shown in **Figure – 2**. For what minimal value of K will the query point “?” be classified as negative? Assume that the ties are broken randomly when K is an even number. **[2 M]**



**Figure – 2**

**Q5.**

1. Explain how Principal Component Analysis (PCA) can be used for eliminating noise. **[3M]**
2. Given the following data shown in **Figure-2**, answer questions **i-iv [2+2+1+2 = 7M]**

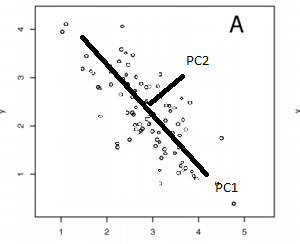


PC2

PC1

**Figure-2**

1. Pictorially represent a pair ofeigenvectors that you would expect to obtain from a Principal Components Analysis (PCA) of the data.

****

1. Indicate which eigenvector you would expect to have the larger corresponding eigenvalue and why?
2. What do the numerical values of the eigenvalues tell you about the data?
3. Do you think it is appropriate to use PCA to reduce the dimensionality of the dataset shown in panel B? Why or why not?
4. Given the following Eigen values and its corresponding Eigen vectors for some data, if you are asked to retain a minimum of 90% of variance from the original data how many Principal Components you will pick and why?

**[Note no marks will be awarded if the reasoning is not mentioned] [3M]**

|  |  |
| --- | --- |
| Eigen value | Eigen Vector |
| 41.5768 | -0.8651  0.5016 |
| 3.3952 | -0.5016  -0.8651 |

1. We have discussed two techniques for feature selection namely PCA and feature subset selection. Justify which of the technique is better in terms of interpreting the output as a user. **[2 M]**

**Q6.** You are asked to identify 3 best features out of 5 and the features are denoted as f1,…,f5 and they are ordered as per the subscript based on the criterion function i.e f1 is more important feature than f2 and so on. **nnnnnnnnnnnnnnnnnnnnnnnnnnnn[3+3+2=8M]**

1. Construct the Branch and Bound tree
2. Explain the process of identifying the optimal feature subset without searching the whole tree.
3. Is it advisable to use correlation between two variables as the criterion function? If yes, the objective function will have to maximized or minimized? If no, why?