Dr. Kamles Dutta

28/11/2023

108

Name of the student ________________ Roll no 12

NATIONAL INSTITUTE OF TECHNOLOGY HAMIRPUR
Department of Computer Science and Engineering (CSE)

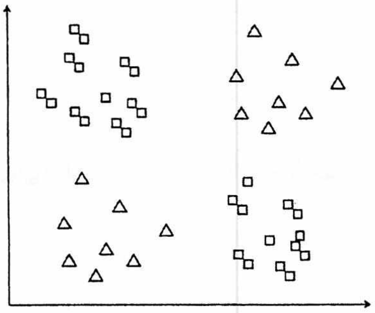
END- TERM EXAMINATION November 2023

Course Dual Degree **Semester – VII** **Branch –** CSE
Subject Name-Machine Learning **Subject Code:** CS-652
Time- 3 hours **Max.Marks- 50**

Note- *All Questions carry equal marks. Answer All Questions. Write all steps of your answers. Assume suitable data. Use suitable diagrams to explain your answers wherever required. Calculator is allowed.*

Questions

Q1. What is the difference between decision tree and Support Vector Machine (SVM)? What are the advantages/ disadvantages of SVM? Give examples to justify your reasoning. What is a kernel function? Why do we need it? Why do we use soft margin in SVM Given the following data samples (square and triangle mean two classes), which kernel can we use in SVM to separate the two classes?



Q2. Why do we need ensemble classification methods? What problems does ensemble classification solve? Compare the Ensemble methods: Bagging, boosting, random forests and Stacking. Use the following dataset to demonstrate ensemble learning method. The dataset consists of 10 data points which consist of two types namely A and B. Each one has been assigned equal weight initially.

A : (1,3), (2,2), (4,4), (5,5), (5,4)
B : (4,1), (4,3), (5,2), (6,1), (6,3)

Q3. Compare latent semantic analysis (LSA) and Probabilistic latent semantic analysis (PLSA). Consider the following set of five documents:

*d1 : **Romeo** and **Juliet**.*
*d2 : **Juliet**: O **happy dagger**!*
*d3 : **Romeo died** by **dagger**.*
*d4 : "**Live free** or **die**", that's the **New-Hampshire**'s motto.*
*d5 : Did you know, **New-Hampshire** is in New-England.*

and a search query: dies, dagger

Demonstrate the application of LSA and PLSA using above example.

Q4. Describe K-mean clustering. How do we measure distances of two clusters? Consider the following dataset:

Example	A	B	C	D	E
Attribute value (X)	0.1	0.6	0.8	2.0	3.0

Apply k-Means Clustering to this data set for k=2. Assuming A as B as the initialized clusters, calculate the following using K-mean algorithm

	Cluster#1	Cluster#2
Cluster assignments that result for C, D and E	A,...	B,...
Recompute the cluster centroids (means) to be the mean (average)		
Reassign the examples to the clusters to which they are closest (i.e., the example is assigned to the closest cluster centroid).		
Recompute the cluster centroids (means) to be the mean (average) of the examples currently assigned to each cluster.		
Reassign the examples to the clusters to which they are closest		

Explain how K-means++ algorithm overcome the limitations of K-mean algorithm. Differentiate between hard clustering, soft clustering and overlapping clustering.

Q5. What is Principal Component Analysis (PCA) and how it is used? What is the relationship between K-means and PCA? What is PC1 and PC2 in PCA?
Given six data points in 5-d space, for Principal component analysis (PCA) : $(1, 1, 1, 0, 0)$, $(-3, -3, -3, 0, 0)$, $(2, 2, 2, 0, 0)$, $(0, 0, 0, -1, -1)$, $(0, 0, 0, 2, 2)$, $(0, 0, 0, -1, -1)$. Calculate

a) sample mean of the data set.
b) first principal component for the original data points.
c) variance of the projected data into 1-d space by principal component
d) reconstruction error in the original 5-d space from the projected 1-d space data obtained in c)