Assignment 01 for Machine Learning (2024-25)

Subject Code: CS550/DSL501

August 26, 2024

Deadline: Sept. 02, 2024 (Time = 18:00 hrs) **Total Marks: 80**

1. Given the following data, use PCA (Principal Component Analysis) to reduce the dimension from 2D to 1D **(20 Marks)**

Feature	**Example 01**	**Example 02**	**Example 03**	**Example 04**
x	2	6	10	14
y	5	4	11	14

Table 1: Data for PCA

(Note : You all have to explain each and every steps clearly using mathematical formulas and normalize the values of eigen vectors.)

- *After calculating the first principal component using the given data, proceed with the remaining steps to complete the PCA process. Once all steps have been completed using the eigenvalues obtained from the data, apply the new eigenvalues (*$\lambda_1 = 30.3849$ *and* $\lambda_2 = 6.6151$*) to evaluate the results. Compare these values, and provide an explanation of which principal component is more significant and why it should be selected for further analysis.*

2. In this problem, you will perform K-means clustering manually, with $K = 2$, on a small example with $n = 6$ observations and $p = 2$ features. The observations are as $(14), (13), (04), (51), (62), (40)$. **(10 Marks)**

   (a) Plot the observations.

   (b) Randomly assign a cluster label to each observation. Report the cluster labels for each observation.

   (c) Compute the centroid for each cluster.

   (d) Assign each observation to the centroid to which it is closest, in terms of Euclidean distance. Report the cluster labels for each observation.

   (e) Repeat (c) and (d) until the answers obtained stop changing.

   (f) In your plot from (a), color the observations according to the cluster labels obtained.

3. Consider a 2-D dataset having two types of classes of data points namely $X1$ and $X2$. Given, $X1 = (x1, x2) = (4,1),(2,4),(2,3),(3,6),(4,4)$ and $X2 = (x1, x2) = (9,10),(6,8),(9,5),(8,7),(10,8)$. **(10 Marks)**

   (a) Apply Linear Discriminant Analysis in the view of dimensionality reduction.

   (b) Plot the graphs if required.

   (c) Write advantages, disadvantages and applications of Linear Discriminant Analysis(LDA).

4. For the **Wine Quality Data Set**, convert all the values in the quality attribute to 0 (bad) if the value is less than or equal to 6 and to 1 (good) otherwise. Normalize all the other attributes between 0 and 1 by min-max scaling. Mention why we use min-max scaling. **(10 Marks)**

   (Note: For this question, you have to upload your .ipynb file to your GitHub repository and share the link in the PDF so that we can see the results. The remaining theory answers should be written in this LaTeX (PDF). The GitHub link must be accessible.)

5. What is the difference between model parameters and model hyperparameters? What is meant by hyperparameter tuning? Name some common hyperparameters used in clustering algorithms. **(5 + 2.5 + 2.5 = 10 Marks)**

6. You are given three observed signals $x_1(t)$, $x_2(t)$, and $x_3(t)$ which are linear mixtures of three independent source signals $s_1(t)$, $s_2(t)$, and $s_3(t)$. The mixing process is represented by the following matrix equation: **(20 Marks)**

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$$

   where

$$\mathbf{x}(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{pmatrix}, \quad \mathbf{s}(t) = \begin{pmatrix} s_1(t) \\ s_2(t) \\ s_3(t) \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

   where $\mathbf{A}$ is an unknown 3×3 mixing matrix, and $\mathbf{s}(t)$ represents the independent source signals.

   (a) Explain the steps involved in estimating the mixing matrix $\mathbf{A}$ and the independent source signals $\mathbf{s}(t)$ using ICA.

   (b) Suppose the mixing matrix $\mathbf{A}$ is not full rank (i.e., it is singular or nearly singular). How would this affect the ICA process? Discuss the potential challenges and the implications for recovering the original source signals.