Assignment 1: PCA & LDA.

Due: Oct. 26, 2023.

The assignment should be submitted in PDF format with the filename "Assignment1-Your Name-Your Student No." to TA (Defang Chen, email: defchern@zju.edu.cn).

1 Principal Component Analysis (PCA)

Given six points $X_i \in \mathbb{R}^2$:

$$\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix}, \begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}1\\2\end{bmatrix}, \begin{bmatrix}2\\1\end{bmatrix}, \begin{bmatrix}2\\2\end{bmatrix},$$

1. Compute the mean of the sample points and write the centered design matrix $\dot{X}$.
2. Find all the principal components of this sample. Write them as unit vectors.
   (a) Which of those two principal components would be preferred if you use only one?
   (b) What information does the PCA algorithm use to decide that one principal components is better than another?
   (c) From an optimization point of view, why do we prefer that one?
3. Compute the vector projection of each of the original sample points (not the centered sample points) onto your preferred principal component. By "vector projection" we mean that the projected points are still in $\mathbb{R}^2$. (Don't just give us the principal coordinate; give us the projected point.)

2 Linear Discriminant Analysis (LDA)

You want to create a model to predict student performance on the Data Mining. You survey several past students and record how many hours they studied for the exam, and whether or not they passed, yielding the two classes.

Passed: [4, 5, 5.5, 6.5, 7, 8]

Failed: [0, 1, 2, 3, 4]

The hours spent studying is the only feature we have for each student (d = 1). Assume that the number of hours is normally distributed for both the passing and failing students.

Consider two ways of modeling this data: Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). Use the 0-1 loss function to define risk.

1. Calculate the sample means μ_p, μ_f and the variances σ_p^2, σ_f^2 computed for QDA. (The subscripts mean "pass" and "fail.") Express your answers as the simplest fractions (not decimals) possible.
2. Calculate the sample means and variances used by **LDA**. Express your answers as the simplest fractions (not decimals) possible.

3. Calculate the decision boundary for **LDA**. Use fractions, not decimals, and express the answer in as simple a form as possible (but expect it to have a logarithm in it).