**The Chinese University of Hong Kong, Shenzhen**
**School of Data Science**

**DDA 3020 · Homework 4**
Due: 23:59, May 9th, 2024

**Instructions:**

- This assignment accounts for 15/100 of the final score.
- You must independently complete each assignment.
- Late submission will get discounted score: 20 percent discount on $(0, 24]$ hours late; 50 percent discount on $(24, 120]$ hours late; no score on late submission of more than 120 hours.

# 1 Written Problems (50 pts.)

## 1.1 Performance evaluation of Logistic-Regression models (12 pts.)

In this problem, your task is to evaluate the performance of two Logistic-Regression models ($M_1$ and $M_2$) for a binary classification problem. The attributes in the test set you have chosen are represented by $\mathbf{X}$. Table 1 shows the output ($\mathrm{P}(+|M_1)$ and $\mathrm{P}(+|M_2)$) of the models, which represent the posterior probabilities for the positive class. As this is a binary classification problem, $\mathrm{P}(-) = 1 - \mathrm{P}(+)$ and $\mathrm{P}(- \mid \mathbf{X}) = 1 - \mathrm{P}(+ \mid \mathbf{X})$.

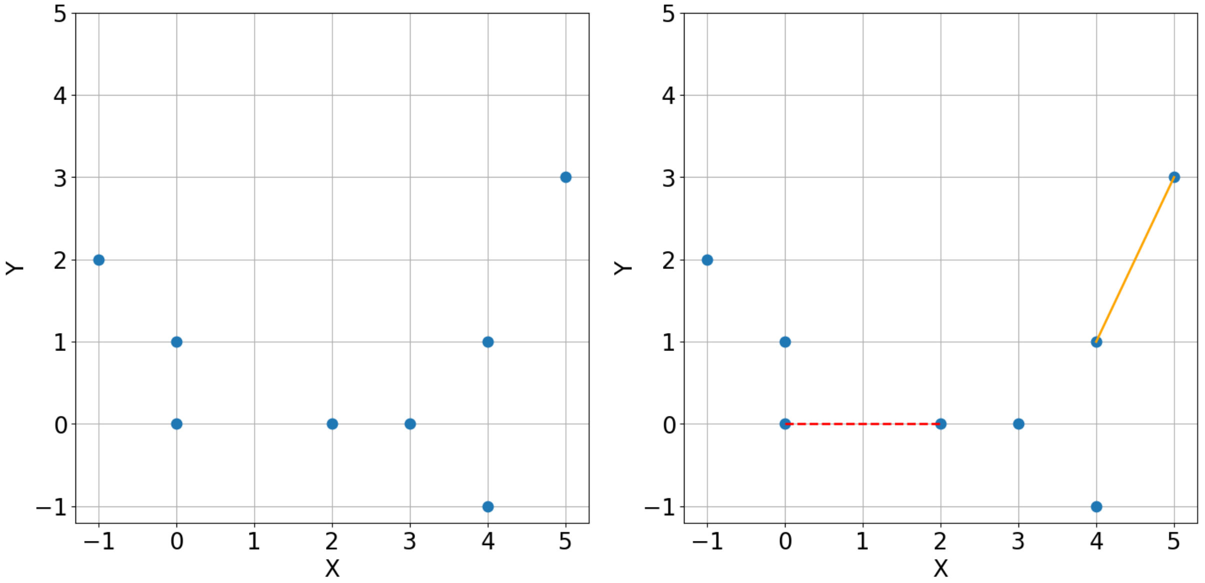
| Instance | True Class | $P(+ \mid \mathbf{X}, M_1)$ | $P(+ \mid \mathbf{X}, M_2)$ |
|---|---|---|---|
| 1 | + | 0.78 | 0.61 |
| 2 | + | 0.62 | 0.08 |
| 3 | - | 0.44 | 0.62 |
| 4 | - | 0.55 | 0.39 |
| 5 | + | 0.61 | 0.48 |
| 6 | + | 0.47 | 0.09 |
| 7 | - | 0.07 | 0.38 |
| 8 | - | 0.14 | 0.09 |
| 9 | + | 0.48 | 0.06 |
| 10 | - | 0.32 | 0.01 |

Table 1: Output of two models

(1) For the two models $M_1$ and $M_2$, suppose you choose the decision threshold to be $\tau = 0.5$. In other words, any test instances whose output of the model is greater than $\tau$ will be classified as a positive example. Write down the confusion matrix and calculate the precision, recall and accuracy of the two models respectively.

(2) Plot the ROC for both $M_1$ and $M_2$, and calculate their AUC. Which model performs better based on the AUC value? (**Hint: In this scenario, the ROC is a step line.**)

## 1.2 Problem: Manual implementation of K-means (13 pts.)



(a) Training data set for K-means clustering w/o constraints

(b) Training data set for K-means clustering w/ constraints

Figure 1: Manual implementation of K-means

Given the data set as shown in Figure 1 and assume that points $A1$, $A6$ and $A8$ are chosen to be the initialized cluster centers. The coordinates of the data points are:

$A1 = (0, 0), A2 = (0, 1), A3 = (-1, 2), A4 = (2, 0), A5 = (3, 0), A6 = (4, -1), A7 = (4, 1), A8 = (5, 3)$

(1) Use the K-means algorithm and Euclidean distance to cluster the 8 data points shown in 1a into K = 3 clusters. Show the new clusters (i.e. the examples belonging to each cluster) and cluster centers after the first iterations, does the algorithm converge after the first iteration?
(2) Consider the case that there exist 1 must link (solid orange line) and 1 cannot link (dashed red line) as shown in 1b. Show the new clusters and cluster centers after the first iterations, does the algorithm converge after the first iteration?

## 1.3 Problem: Manual calculation of the M step for a GMM (13 pts.)

In this question we consider clustering 1D data with a mixture of 2 Gaussians using the EM algorithm. A GMM with 1D data represents a distribution as

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(x \mid \mu_k, \sigma_k\right)$$

with $\pi_k$ the mixing coefficients, where: $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k \geq 0, \forall k$. And,

$$\mathcal{N}\left(x \mid \mu_k, \sigma_k\right) = \frac{1}{\sqrt{2\pi\sigma_k^2}} exp(-\frac{(x-\mu_k)^2}{2\sigma_k^2})$$

You are given the l-D data points $X = \{1, 10, 20\}$. Suppose the output of the E step is the following matrix:

$$\mathbf{R} = \begin{bmatrix} 1 & 0 \\ 0.4 & 0.6 \\ 0 & 1 \end{bmatrix}$$

where entry $r_{i,c}$ is the probability of obervation $X_i$ belonging to cluster $c$ (the responsibility of cluster $c$ for data point $i$ ). You just have to compute the M step. You may state the equations for maximum likelihood estimates of these quantities (which you should know from the lecture) and apply to this data set.
(1) Write down the likelihood function you are trying to optimize.
(2) After performing the M step for the mixing weights $\pi_1, \pi_2$, what are the new values?
(3) After performing the M step for the means $\mu_1$ and $\mu_2$, what are the new values?

### 1.4 Problem: PCA (12 pts.)

Consider the following 10 data points: $X = \{(2, 0, 1, -3, -2), (0, 2, -3, -3, -2), (1, 2, 1, 3, -2), (-1, 1, 3, 2, -1), (1, 0, 1, -1, 1), (2, 3, -1, 1, -2), (-2, 3, -3, 3, 2), (-2, -2, 2, 3, -2), (-2, -3, 1, -2, -3), (-3, 2, 0, -1, -2)\}$. Compute the unit-length principle components of $X$ and choose two of them for PCA, then calculate the projection of each data on these two principal components. You could use python or matlab to obtain eigenvectors and eigenvalues.

## 2 Coding Problems (50 pts.)

In the coding part of this assignment, you will implement $k$-means and a GMM from scratch. You are NOT allowed to use off-the-shelf codes to complete the assignment.
You'll complete the following three tasks, each detailed further in the `code.ipynb` file:

- **Task 1**: Implement $k$-means.
- **Task 2**: Implement $k$-means++ initialization.
- **Task 3**: Learn a GMM with EM algorithm.

## Submission Format

Submit a single ZIP file named `YOUR_STUDENT_ID.zip`. The ZIP file should contain two files:

- `code.ipynb`: A Jupyter Notebook file containing your code. Please ensure all code cells are executed, and their outputs (including numbers and plots) are present in the uploaded notebook.
- `StudentID_homework4.pdf`: A PDF file containing your solutions to written problems and a report for the coding problems. The report for the coding problems should include explanations to your code, your choice of hyperparameters, and visualizations of each subtask, as described in detail in `code.ipynb`.