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 τ will be classified as a positive example. Write down the confusion matrix and calculate the precision, recall and accuracy of the two models respectively.

Solution

1. For M_1, confusion matrix:

	$\widehat{\mathbf{P}}$ (predicted)	$\widehat{\mathbf{N}}$ (predicted)
P (actual)	$\boldsymbol{TP = 3}$	$\boldsymbol{FN = 2}$
N (actual)	$\boldsymbol{FP = 1}$	$\boldsymbol{TN = 4}$

Precision $= \frac{3}{4} = 0.75$, Recall $= \frac{3}{5} = 0.6$, Accuracy $= \frac{7}{10} = 0.7$

2. For M_2, confusion matrix:

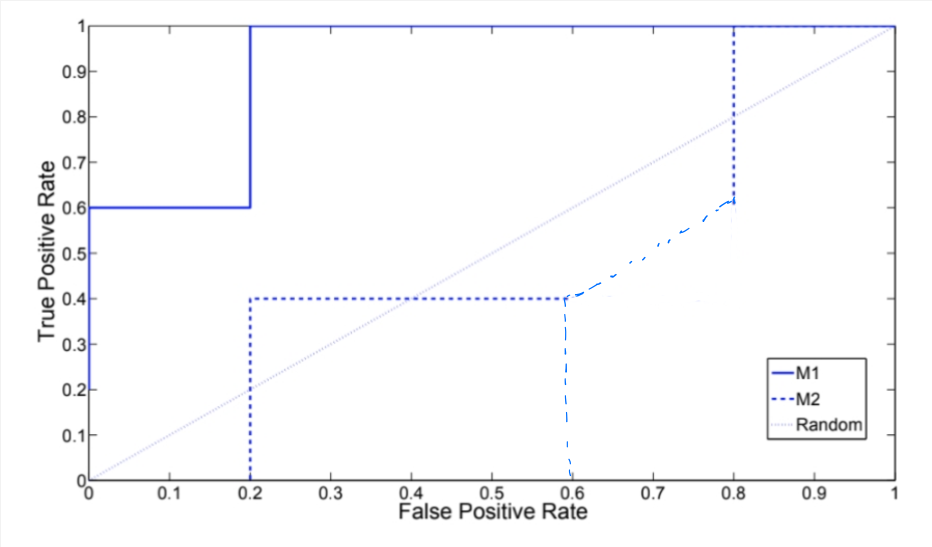
	$\widehat{\mathbf{P}}$ (predicted)	$\widehat{\mathbf{N}}$ (predicted)
P (actual)	$\boldsymbol{TP = 1}$	$\boldsymbol{FN = 4}$
N (actual)	$\boldsymbol{FP = 1}$	$\boldsymbol{TN = 4}$

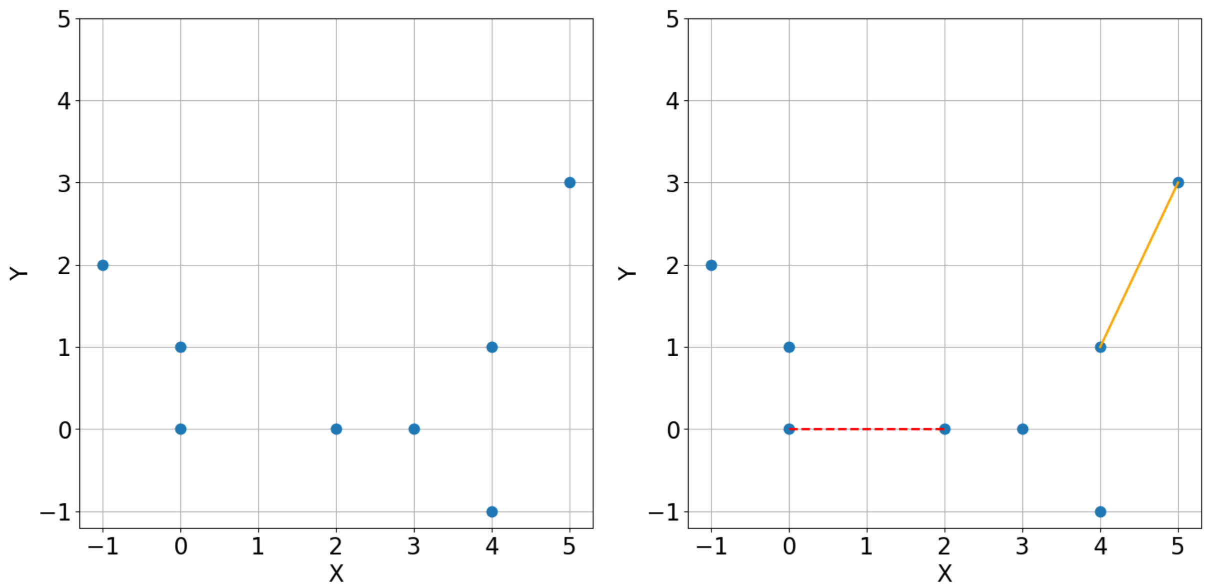
Precision $= \frac{1}{2} = 0.5$, Recall $= \frac{1}{5} = 0.2$, Accuracy $= \frac{5}{10} = 0.5$

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

Solution

The ROC curve for M_1 and M_2 are shown below. AUC = 0.92 for M_1 and AUC = 0.46 for M_2 M_1 is better, since its area under the ROC curve (AUC) is larger than the area under ROC curve (AUC) for M_2.





(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

1.2 Problem: 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?

Solution

Define the (squared) distance matrix $D_{ij} = dist_{eucl}(A_i, A_j)^2$
iteration 1 Looking at rows of the distance matrix corresponding to the centers

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
$\mu_1 = A_1$	0	1	5	4	9	17	17	34
$\mu_2 = A_6$	17	20	34	5	2	0	4	17
$\mu_3 = A_8$	34	29	37	18	13	17	5	0

Then the clusters are $C_1 = \{A_1, A_2, A_3, A_4\}$, $C_2 = \{A_5, A_6, A_7\}$ and $C_3 = \{A_8\}$. To find the new cluster centers, we have to compute the means:

$$\mu_1' = \frac{1}{|C_1|}\sum_{A \in C_1} = \frac{1}{4}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \begin{pmatrix} -1 \\ 2 \end{pmatrix} + \begin{pmatrix} 2 \\ 0 \end{pmatrix}\right) = \frac{1}{4}\begin{pmatrix} 1 \\ 3 \end{pmatrix}$$

$$\mu_2' = \frac{1}{|P_2|}\sum_{A \in P_2} = \frac{1}{3}\left(\begin{pmatrix} 3 \\ 0 \end{pmatrix} + \begin{pmatrix} 4 \\ -1 \end{pmatrix} + \begin{pmatrix} 4 \\ 1 \end{pmatrix}\right) = \frac{1}{3}\begin{pmatrix} 11 \\ 0 \end{pmatrix}$$

$$\mu_3' = \mu_3 = \begin{pmatrix} 5 \\ 3 \end{pmatrix}$$

The algorithm does not converge after the first iteration as the assignment of the data points will change in the second 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?

Solution

Define the (squared) distance matrix $D_{ij} = dist_{eucl}(A_i, A_j)^2$

iteration 1 Looking at rows of the distance matrix corresponding to the centers

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
$\mu_1 = A_1$	0	1	5	4	9	17	17	34
$\mu_2 = A_6$	17	20	34	5	2	0	4	17
$\mu_3 = A_8$	34	29	37	18	13	17	5	0

Without constraints $C_1 = \{A_1, A_2, A_3, A_4\}$. However, we have the cannot link for A_3 and A_4. Hence we should assign A_4 to the second cloest clusters C_2, and now $C_1 = \{A_1, A_2, A_3\}, C_2 = \{A_4, A_5, A_6, A_7\}$. Since A_7 must link with A_8, A_7 should be assigned to C_3. Finally, the clusters are $C_1 = \{A_1, A_2, A_3\}$, $C_2 = \{A_4, A_5, A_6\}$ and $C_3 = \{A_7, A_8\}$.

To find the new cluster centers, we have to compute the means:

$$\begin{aligned}
\mu_1' &= \tfrac{1}{|C_1|}\textstyle\sum_{A\in C_1} = \tfrac{1}{3}\left(\begin{pmatrix}0\\0\end{pmatrix} + \begin{pmatrix}0\\1\end{pmatrix} + \begin{pmatrix}-1\\2\end{pmatrix}\right) &&= \tfrac{1}{3}\begin{pmatrix}-1\\3\end{pmatrix}\\
\mu_2' &= \tfrac{1}{|P_2|}\textstyle\sum_{A\in P_2} = \tfrac{1}{3}\left(\begin{pmatrix}2\\0\end{pmatrix} + \begin{pmatrix}3\\0\end{pmatrix} + \begin{pmatrix}4\\-1\end{pmatrix}\right) &&= \tfrac{1}{3}\begin{pmatrix}9\\-1\end{pmatrix}\\
\mu_3' &= \tfrac{1}{|P_2|}\textstyle\sum_{A\in P_2} = \tfrac{1}{2}\left(\begin{pmatrix}4\\1\end{pmatrix} + \begin{pmatrix}5\\3\end{pmatrix}\right) &&= \tfrac{1}{2}\begin{pmatrix}9\\4\end{pmatrix}
\end{aligned}$$

The algorithm converges after the first iteration as the assignment of the data points will not change in the second iteration. You can check:

iteration 2 Looking at rows of the distance matrix corresponding to the centers

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
μ_1'	1.11	0.11	1.44	6.44	12.11	22.78	18.78	32.44
μ_2'	9.11	10.78	21.44	1.11	0.11	1.44	2.78	15.11
μ_3'	24.25	21.25	30.25	10.25	6.25	9.25	1.25	1.25

The clusters are $C_1 = \{A_1, A_2, A_3\}$, $C_2 = \{A_4, A_5, A_6\}$ and $C_3 = \{A_7, A_8\}$, which are the same as the clusters after the iteration 1.

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

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

with π_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.

Solution

We are to optimize:

$$\sum_{n=1}^{3} \sum_{k=1}^{2} r_{nk} \left(\log \pi_i + \log \mathcal{N}\left(x_n \mid \mu_k, \sigma_k^2\right)\right).$$

(2) After performing the M step for the mixing weights π_1, π_2, what are the new values?

Solution

The new optimal assignment for π_1, π_2 is derived by differentiating the auxiliary function added with a regularizer to ensure $\pi_1 + \pi_2 = 1$:

$$\frac{\partial Q + \lambda\left(\pi_1 + \pi_2 - 1\right)}{\partial \pi_1} = \frac{\sum_{n=1}^{3} r_{n,1}}{\pi_1} + \lambda = \frac{1.4}{\pi_1} + \lambda,$$
$$\frac{\partial Q + \lambda\left(\pi_1 + \pi_2 - 1\right)}{\partial \pi_2} = \frac{\sum_{n=1}^{3} r_{n,2}}{\pi_2} + \lambda = \frac{1.6}{\pi_2} + \lambda.$$

Setting both gradients to zero ends up with:

$$\pi_1 = \frac{7}{15},$$
$$\pi_2 = \frac{8}{15},$$
$$\lambda = -3.$$

(3) After performing the M step for the means μ_1 and μ_2, what are the new values?

Solution

$$\mu_1 = \frac{\sum_{n=1}^{3} r_{n1} \cdot x_n}{\sum_{n=1}^{3} r_{n1}} = \frac{25}{7},$$
$$\mu_2 = \frac{\sum_{n=1}^{3} r_{n2} \cdot x_n}{\sum_{n=1}^{3} r_{n2}} = \frac{65}{4}.$$

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.

Solution

The mean of the data points is

$$\mu = [-0.4, 0.8, 0.2, 0.2, -1.3]^T$$

The covariance matrix:

$$\Sigma = \frac{1}{10} \sum_i \left(X^{(i)} - \mu\right)\left(X^{(i)} - \mu\right)^\top$$
$$= \begin{bmatrix} 3.04 & 0.82 & -0.02 & -0.82 & -0.12 \\ 0.82 & 3.76. & -2.16 & 1.04 & 1.04 \\ -0.02 & -2.16 & 3.56 & 0.76. & -0.84 \\ -0.82 & 1.04 & 0.76 & 5.50 & 1.16 \\ -0.12 & 1.04 & -0.84 & 1.16 & 2.21 \end{bmatrix}$$

Do SVD decomposition of Σ, we get the eigenvalues:

$$\lambda = (0.823, 1.530, 3.076, 5.931, 6.770)$$

Choose the first 2 eigenvalues and calculate their corresponding eigenvectors:

$$u_1 = [0.298, 0.395, -0.580, -0.646, 0.030]^\top$$
$$u_2 = [-0.026, 0.579, -0.329, 0.654, 0.359]^\top$$
$$U = [u_1, u_2]$$

Reconstructed Z can be computed by $Z^{(i)} = U^\top (X^{(i)} - \mu)$,

$$Z = \begin{bmatrix} 1.8 & -3.13 \\ 4.50 & -0.61 \\ -1.40 & 1.97 \\ -2.88 & 0.50 \\ 0.48 & 0.72 \\ 1.74 & 1.88 \\ 0.54 & 5.38 \\ -4.46 & -0.59 \\ -1.07 & -4.47 \\ 0.57 & -0.21 \end{bmatrix}^T$$

The projection can be calculated by $\tilde{X}^{(i)} = \mu + UZ^{(i)}$,

$\tilde{X} = \{(0.27, -0.22, 0.08, -3.13, -2.37), (0.96, 2.22, -2.21, -3.10, -1.38),$
$(-0.87, 1.39, 0.37, 2.40, -0.63), (-1.27, -0.05, 1.71, 2.38, -1.21),$
$(-0.24, 0.57, 0.16, -0.58, -1.54), (0.07, 2.57, -1.43, 0.30, -0.57),$
$(-0.38, 4.13, -1.88, 3.37, 0.65), (-1.71, -1.30, 2.98, 2.69, -1.65),$
$(-0.60, -2.21, 2.29, -2.03, -2.94), (-0.22, 0.90, -0.06, -0.30, -1.36)\}$