DDA 3020 Homework 4

December 4, 2022

Homework due: 11:59 pm, December 17, 2022. Note that we have reduced the number of questions in both the written and programming assignments such that you could have more time to prepare for the final exam.

1 Derivation (8 points)

1. (3 points) EM for mixtures of Bernoullis (Exercise 11.3 of Kevin P. Murphy's book).
   (1) Show that the M step for maximum likelihood estimation of a mixture of Bernoullis is given by

$$\mu_{kj} = \frac{\sum_i r_{ik} x_{ij}}{\sum_i r_{ik}}$$

   (2) Show that the M step for MAP estimation of a mixture of Bernoullis with a $\beta(\alpha, \beta)$ prior is given by

$$\mu_{kj} = \frac{\sum_i r_{ik} x_{ij} + \alpha - 1}{\sum_i r_{ik} + \alpha + \beta - 2}$$

2. (3 points) As we showed in class, consider a set of binary samples indexed by $i = 1, ..., m^+$ for positive class and $j = 1, ..., m^-$ for negative class. Let $g(\mathbf{x})$ be a predictor and $e_{ij} = g\left(\mathbf{x}_i^+\right) - g\left(\mathbf{x}_j^-\right)$ is the difference in the value of the predictor between a positive-class sample $\mathbf{x}_i^+$ and a negative-class sample $\mathbf{x}_j^-$.
   (1) Consider also a Heaviside step function given by

$$u(e) = \begin{cases} 1, & \text{if } e > 0 \\ 0.5 & \text{if } e = 0 \\ 0 & \text{if } e < 0 \end{cases}$$

   (2) The Area Under the ROC Curve (AUC) can be expressed as

$$\text{AUC} = \frac{1}{m^+ m^-} \sum_{i=1}^{m^+} \sum_{j=1}^{m-} u\left(e_{ij}\right)$$

The total number of the samples is $n = m^+ + m^-$.
Now we sort the predictor from smallest to largest, and set the rank of the ith sample as $rank_i = 1, 2, ..., n$ (*i.e.*, The sample with the smallest $g(\mathbf{x}_i)$ has $rank_i = 1$ and the sample with the largest $g(\mathbf{x}_i)$ has $rank_i = n$).
Prove:

$$\text{AUC} = \frac{\sum_{i=1}^{m^+} rank_i - (m^+)(m^+ + 1)/2}{m^+ m^-}$$

3. (2 point) 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 Programming using Python (8 points)

Task: Clustering on UCI seed dataset, which can be downloaded from `https://archive.ics.uci.edu/ml/datasets/seeds`. The number of clusters is set as 3.
You need to:

1. Implement **K-means** and **GMM-EM** algorithms **from scratch** (*i.e.*, no third-party or off-the-shelf package or library are allowed). Explain briefly your source codes in the report. (5 points)

2. Implement 2 evaluation metrics including **Silhouette Coefficient** and **Rand Index from scratch** (*i.e.*, not calling off-the-shelf package) to evaluate the performance of above clustering algorithms. (2 points)

3. Analyze the sensitivity to the initialization of each algorithm (*e.g.*, run one clustering algorithm with random initialization multiple times, and calculate the standard deviations of evaluation scores of these clustering results) (1 point)

Note that you should submit A4_StudentID.pdf (report, together with the written answers), and A4_StudentID.ipynb (code). Please zip them into "A4_StudentID.zip". The reference report is in Assignment 1. You can check it on BlackBoard.
(You can submit several files in one submission. Don't submit them in different submissions.) Your report for the programming questions should include necessary formulas, charts, and explanations. The number of pages should be 4-5.