CSE 847: Machine Learning Assignment #3

Instructor: Prof. Jun Wu
Out: Oct. 31, 2024; Due: Nov. 27, 2024

Submit electronically for Assignment #3, a file named `yourFirstName-yourLastName.pdf` *containing your solution to this assignment (a .doc or .docx file is also acceptable, but .pdf is preferred).*

1 Gaussian Mixture Model and EM Algorithm [30 points]

Given a 1-dimensional data set: $\{-67, -48, 6, 8, 14, 16, 23, 24\}$, consider using a Gaussian Mixture Model with 2 components $(k = 2)$ to fit your data.

1.1 Parameters [10 points]

How many independent parameters are there in this GMM? Please justify your answer.

1.2 EM Algorithm [20 points]

What will your parameters be after 1 iteration of EM? Show your major calculations in both the E-step and the M-step. Only giving out the final results will NOT grant you any score. Feel free to initialize your parameters in any way you prefer.

2 Graphical Models [20 points]

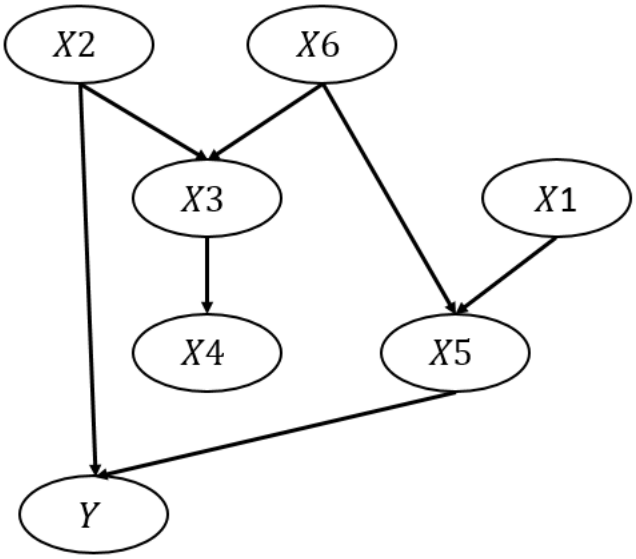

Figure 1: DAG for Question 2

2.1 Joint Distribution [10 points]

Based on the graphical model in Figure 1, what is the joint distribution $P(Y, X1, X2, X3, X4, X5, X6)$?

2.2 Conditional Independence [10 points]

Please justify whether the following variables are conditionally independent or not:
(1) $X2 \perp X5 \mid X3$
(2) $X6 \perp Y \mid X5$

3 K-Means [50 points]

You are given a data set consisting of 4 samples $a = (3, 3), b = (7, 9), c = (9, 7), d = (5, 3)$ in 2-dimensional space. You will assign the 4 samples into 2 clusters using the K-Means algorithm with **Euclidean distance**. The initialization of centroids is given by $c_1 = (6, 5)$ and $c_2 = (6, 6)$.

3.1 K-means Steps [10 points]

Show the steps of the K-Means algorithm until convergence, including each cluster centroid and the cluster membership of each example after each iteration. **Only giving out the final results will NOT grant you any score.**

3.2 Potential Function [10 points]

What is the value of the K-Means loss function $\mathcal{L}(K)$ (see Eq. (1) below) upon convergence?

3.3 Implementation [30 points]

For this problem, please download the breast-cancer-wisconsin data from the following link:
`https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original`
The data set contains 11 columns, separated by a comma. The first column is the example id, and you should ignore it. The **second to tenth columns are the 9 features**, based on which you should run your K-means algorithm. The last column is the class label, and you should ignore it as well.

- Please implement the K-Means algorithm on this data set with $K = 2, 3, 4, 5, 6, 7, 8$. For each K value, you need to first run the K-Means algorithm and then compute the function as follows:

$$\mathcal{L}(K) = \sum_{i=1}^{N} \sum_{j=1}^{K} m_{ij} ||\mathbf{x}_i - \mathbf{c}_j||^2 \qquad (1)$$

where N is the number of samples, $\mathbf{x}_i$ denotes the feature vector for i^{th} sample, $\mathbf{c}_j$ refers to the centroid of the j^{th} cluster, and m_{ij} denotes whether $\mathbf{x}_i$ belongs to the j^{th} cluster.

- Please explain your implementation of K-Means with **pseudo code** and **plot the curve** of $\mathcal{L}(K)$ vs. K value. If you were to pick the optimal value of K based on this curve, would you pick the one with the lowest value of the potential function? Why?

Hint: if you find an empty cluster in a certain iteration, please drop the empty cluster and then randomly split the largest cluster into two clusters to maintain the total number of clusters at K.