CSE 575: Statistical Machine Learning Assignment #3

Instructor: Prof. Jingrui He
Out: Mar 15, 2019; Due: Apr 12, 2019

Submit electronically, using the submission link on Canvas 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 [20 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 [10 points]

What will your parameters be after 1 iteration of EM? Show your major calculations in both the E-step (class membership of all data points) and the M-step (maximum likelihood estimate of the parameters given all data points' class membership). Only giving out the final results will NOT grant you any score. Feel free to initialize your parameters any way you prefer.

2 Principal Component Analysis [20 points]

2.1 Principle Components [10 points]

Given a 2-dimensional data set: $\{(0, 1), (-1, 2), (3, -2), (1, 0), (-3, 4)\}$, what are the first and the second principle components? Show your justification in 1-2 sentences.
Hint: Plotting all the points in the 2-dimensional feature space may greatly help with the analysis, and you don't have to run MATLAB code to get the results.

2.2 Reconstruction Error [10 points]

For an n-dimensional data set consisting of m examples $(m > n)$, in general, how many principle components can you compute? If you were to use the top n principle components to reconstruct the data set, what would your reconstruction error be? Briefly justify your answer.

3 Graphical Models [10 points]

3.1 Joint Distribution [5 points]

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

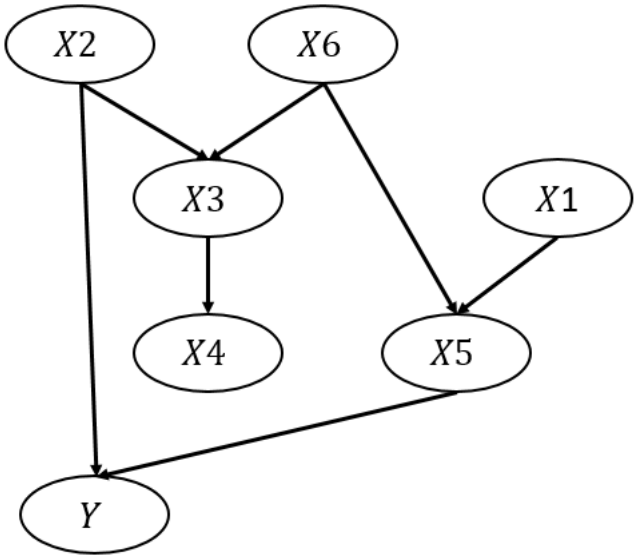

Figure 1: DAG for Question 3

3.2 Conditional Independence [5 points]

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

4 K-Means [50 points]

You are given a data set consisting of 4 examples $a = (3, 3), b = (7, 9), c = (9, 7), d = (5, 3)$ in 2-dimensional space. You will assign the 4 examples into 2 clusters using K-Means algorithm with **Euclidean distance**. To initialize the algorithm, a and c are in one cluster, b and d are in the other cluster.

4.1 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.**

4.2 Potential Function [10 points]

What is the value of the K-Means potential function upon convergence?

4.3 Implementation [30 points]

For this problem, please download the breast-cancer-wisconsin data from the following link:
http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisc
breast-cancer-wisconsin.data
The data set contains 11 columns, separated by 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 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 potential function as follows:

$$\mathcal{L} = \sum_{j=1}^{m} ||\mu_{C(j)} - x_j||^2 \tag{1}$$

where m is the number of examples, x_j denotes the feature vector for j^{th} example and $\mu_{C(j)}$ refers to the centroid of the cluster that x_j belongs to.

- 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.