CSE 575: Statistical Machine Learning Assignment #2

Instructor: Prof. Jingrui He
Out: Mar 25, 2018; Due: April 22, 2018

Submit electronically, using the submission link on Blackboard for Assignment #1, 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 Guassian Mixture Model with 2 components ($k = 2$) to fit your data.

1.1 **(10 points) Parameters.** How many independent parameters are there in this GMM? What are they?

1.2 **(10 points) EM Updates.** 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 any way you prefer.

2 Principle Component Analysis [20 points]

2.1 **(10 points) Principle Components.**Given a 2-dimensional data set: {(0, 1), (-1, 0), (-3, -2), (1, 2), (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 **(10 points) Reconstruction Error.** 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)

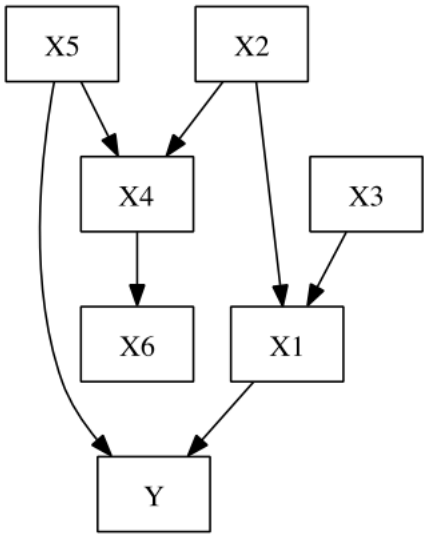

Figure 1: The DAG for Question 3.

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

4 K-means (50 points)

You are given a data set consisting of 4 examples a, b, c, d in two-dimensional space, whose features are shown in the table below.

a	b	c	d
3	7	9	5
3	9	7	3

Table 1: The given data set.

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 a cluster, b and d are in the other cluster.
4.1 **(10 points) K-means Steps.** Show the steps of the K-Means algorithm until convergence, including each cluster centroid and the cluster membership of each example after each iteration.
4.2 **(10 points) Potential Function.** Calculate the value of the K-Means potential function upon convergence.
4.2 **(30 points) Implementation.** Download the breast-cancer-wisconsin.data from http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/ or you can go to blackboard Homework Assignments/Assignment #3.

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 to perform clustering on this dataset 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:

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

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

Please explain your implementation of K-Means with pseudo code and plot the curve of $\mathcal{L}(K)$ vs. K. 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.