CS 229 Machine Learning, spring 2020

Homework 4:

Clustering

Due Tuesday April 28, 11:59pm

Submit by the **blackboard system**

**Question1: (3 pts) EM for Gaussian mixture model**

Derive E-step and M-step formulae in a Gaussian mixture model, starting from the definition of log likelihood function.

**Question2: (3 pts) K-means and EM for Gaussian mixture model**

K-means algorithm can be considered as a particular limit of the EM algorithm for Gaussian mixtures. Whereas the K-means algorithm performs a hard assignment of data points to clusters, in which each data point is associated uniquely with one cluster, the EM algorithm makes a soft assignment based on the posterior. Please show how they are relevant, and at what conditions Gaussian mixture model works equivalently like k-means algorithm, by mathematical proof.

**Question3:(4 points) Programming k-means**

Write **your own** code of **k-means** algorithm (submission of the code is required). **Data:** usethe data in: clu\_data.txt

1) Try different settings of parameter **k**. For **each value of k,** compute the **SSE (sum of squared error)** of clustering result. Plot the **SSE curve** w.r.t. the various k values.

2) Does the SSE curve suggest the best clustering results?

3) Plot the best clustering result you think (using different colors to show the different clusters), and answer:

1. How does k-means (with your setting of k) perform on clustering the data?
2. And why? What other methods can perform better for this clustering problem than k-means?