CSE 5520 Fall 2021
Homework 7 (Due 11:59 pm, Sunday, October 30, 2021 at HuskyCT)

Visualization of GMM and EM

This homework is to help you practice with some data visualization fundamentals. You are expected to use these visualization techniques and others in your final project. You are required to do this exercise in Python. All plots/graphs must have titles and x-y coordinate tick labels.

Part 1: K-means and GMM

You need to use your own data to do the problems and you can include your solution or its extended version as part of your proposal/final presentation.

Step 1. Prepare a 2D data from your project and plot it, i.e. as a scatter plot. You can visually inspect and determine *k,* the number of clusters.

Step 2. With your chosen *k*, perform the *k*-means clustering on your data. Plot clustering result so that each color has different color, for example, cluster 1 in red, cluster 2 in blue, etc. You can plot with different *k* to find a better number of clusters. **(Extra credit)** You can use silhouette method or elbow method to find the optimum *k.*

Step 3. Calculate the centroid (mean of x axis, mean of y axis) of each cluster and radii that covers 90 % of data of the cluster.

Step 4. Plot the circles centered at the centroid with radius calculated in Step 3 on top of the plot created from Step 2. Mark the centroid with 'X'.

Step 5. In a markdown cell, discuss the goodness of your chosen *k,* i.e., if you think what you have is the best or close to the best.

Step 6. Repeat steps 2 ~ 5, using GMM with the EM clustering method. In this case, calculate covariance matrix for each cluster instead of radius. Show the values.

Step 7. Using a markdown cell, compare the outcomes from Step 6 with Step 4 and discuss which clustering method is better for your data.

Step 8. (**Extra credit**) Plot 2D Gaussian curve calculated in Step 6 as shown in the 2D GMM with EM examples in the lecture slides.

Part 2: K-means and GMM

This part is to continuously help you do the term project by making you include visualization of k-means and GMM clustering methods in your Dashboard.

Step 1. Publish the visualizations you have done in Part 1 to your Dashboard "privately" for now.

Step 2. Take a screenshot (screen clip) of your Dashboard and include it in a markdown cell so that you succeeded in publishing your Part I plots in your Dashboard publication. The URL should demonstrate that you can publish your Dashboard on the public VM. You should kill your publication after you are done with the screenshot since you do not make your Dashboard public yet. You are only showing that you can publish Dashboard on a third party machine.

You upload your Jupyter notebook in HuskyCT. The file name should be of the following format: HWn_Doe where n is the homework number and Doe denotes the last name.

HWs and Projects, 5% penalty for one day late submission. No acceptance after 5 days late. Extension is allowed only with the supporting medical record.