MA598 Homework #1 – Due 9/22
K-means clustering & Gaussian Mixture Model

1) Acquire Data

Diabetes data
The diabetes data set is taken from the UCI machine learning database repository at: http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes.

- 768 samples in the dataset
- 8 quantitative variables

Select the number of clusters between [1, 20]

2) Use K-means in Scikit-learn package or matlab machine learning toolbox

Try to use different input parameters (# of clusters, initial seeds etc.) and compare different results using silhouette score or pdist and select the best result.

Plot Figures to show different clustering results using scatter plot with different clusters in different colors. And use the solid black circles as the centers of the clusters.

3) Use Gaussian mixture model in Scikit-learn package or matlab machine learning toolbox

Try to use different input parameters (# of clusters etc.) and compare different results using silhouette score or pdist and select the best result.

Plot Figures to show different clustering results using scatter plot with different clusters in different colors. And use the solid black circles as the means of the clusters.

4) Use Variational Bayesian Gaussian mixture model (VBGMM) in Scikit-learn package

Try to use different input parameters (# of clusters etc.) and compare different results using silhouette score or pdist and select the best result.

Plot Figures to show different clustering results using scatter plot with different clusters in different colors. And use the solid black circles as the means of the clusters.

5) Use Dirichlet Process Gaussian mixture model (DPGMM) in Scikit-learn package

Try to use different input parameters (alpha etc.) and compare different results using silhouette score or pdist and select the best result.

Plot Figures to show different clustering results using scatter plot with different clusters in different colors. And use the solid black circles as the means of the clusters.

6) Try to compare the clustering results using silhouette score or pdist obtained from K-means, Gaussian mixture model, VBGMM and DPGMM.

Put all Python code, figures and results into one PDF file

Put your name and Purdue student ID in the first page of the PDF file

Submit it to blackboard by 9/22 midnight.