Assignment 2

Generating Synthetic data and doing unsupervised analysis

1. If N_i is a 2-D Gaussian Distribution given below:

$$N_1\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}100 & 0\\0 & 3\end{pmatrix}\right] \text{ , } N_2\left[\begin{pmatrix}0\\6\end{pmatrix},\begin{pmatrix}100 & 0\\0 & 3\end{pmatrix}\right] \text{ , } N_3\left[\begin{pmatrix}0\\12\end{pmatrix},\begin{pmatrix}100 & 0\\0 & 3\end{pmatrix}\right]$$

   Generate 1000 random samples using:

   a) $X_1 = N_1$
   b) $X_2 = N_1 + N_2$
   c) $X_3 = N_1 + N_2 + N_3$

2. For the generated X_1, X_2 and X_3 above, calculate:
   a) The k-means clustering of (1, 2, 3, 4, 5, 6) centroids. In each case calculate the average distance among each point and its' centroid, then draw the average distances vs the number of clusters.
   b) Suggest an algorithm to detect the appropriate number of clusters.

3. For the generated X_1, X_2 and X_3 above, calculate:
   a) The Gaussian Mixture Model (GMM) using (1, 2, 3, 4, 5, 6) Gaussians. In each case, calculate the average probability for all the given points, then draw the average probability vs the number of Gaussians.

4. Calculate the PCA for X_1, X_2 and X_3 above, then make data reduction at 90% variance. Write down your comments.

5. Calculate the distance from a point P(0, 0, 0) and the plane: $3X_1 - 3X_2 + X_3 = 4$.