SSIE 637 Advanced Topics in Healthcare
Assignment 02
Fall 2016

Instructor: *Daehan Won*

Due 10/14/2016

Question 1: Clustering (40 points)
(Part 1, *Revisiting Q3 in HW1*, 20 points)
Run the k-mean clustering again on the given data set "random_data.txt" with various k. Calculate the following performance measure: the sum of square distance J:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} a_{nk} \|x_n - \mu_k\|^2 \tag{1}$$

where the number of data points is N and the number of examined clusteres is K. x_n is n-th data sample whereas μ_k is the centroid of the k-th cluster. a_{nk} is a binary indicator given by:

$$a_{nk} = \begin{cases} 1 \text{ if } x_n \text{ belongs to cluster } k \text{ with the centroid } \mu_k \\ 0 \text{ Otherwise} \end{cases}$$

Once you calculated J, **make the plot** for k versus J.

(Part 2, *Hierarchical clustering and Gaussian Mixture Model*, 20 points)
Apply hierarchical clustering and Gaussian Mixture Model(GMM) to the "random_data.txt" and show graphical results. Compare three methods: k-mean, Hierarchical clustering and Expectation-Maximization and provide graphical results and your conclusion. Note that you need to specify all parameter settings such as distance measure, linkage, etc. you use.

Question 2: Breast Cancer Data Analysis (60 points)
(Part 1, *Finding good k*, 30 points)
'breast_cancer.txt' data set has 699 samples with 9 features. Each sample is classified into 2 classes (index '2' for benign and index '4' for malignant). Note that the class information is represented at **the last column** in the data file
1) Apply k-mean clustering with various $k = (1, 2, 3, ..., 8)$ and its corresponding J. (15 pt)
2) Apply Hierarchical clustering. (15 pt)
Pleae make sure that you **must ignore the last column when you run the clustering methods.**

(Part 2, *Comparison k-mean and GMM*, 30 points)
Since you know the actual class of the data, calculate ground-truth accuracy P with k-mean and GMM. Note that fix the number of clusters as 2.

$$P = \frac{\text{number of true positive}}{\text{number of true positive + number of false positive}} \tag{2}$$

Hint) Use 'predict' function for GMM and 'fit_predict' function for k-mean in scikit-learn package.

Question 3 (Optional, 10 points): BIC score of GMM (10 points)
One of the common crietria to find proper k (i.e., number of clusters or components) in GMM is BIC(Bayesian Information Criterion) score:

$$BIC = -2 \times ln(L) + d \times ln(N) \tag{3}$$

, where L is the maximum likelihood calculated by EM algorithm, d is the number of parameters (if you set the number of components as k, $d = 3k - 1$) and N is the number of samples. Calculate BIC scores with various k on both data sets: 'random_data' and 'breast_cancer'. You can use inherent function in your programming language or your own implementation.