CS 6140: Machine Learning	**Spring 2018**
College of Computer and Information Science	
Northeastern University	
Assignment 5	**April, 6**
Instructor: Bilal Ahmed	**Due Date: April, 20**

Data Clustering

1 Data

We will be working with six datasets in this assignment:

1. Dermatology: 366 instances, 34 features and 6 classes.
2. Vowels: 990 instances, 10 features and 11 classes.
3. Glass: 214 instances, 9 features and 6 classes.
4. Ecoli: 327 instances, 7 features and 5 classes.
5. Yeast: 1479 instances, 8 features and 9 classes.
6. Soybean: 290 instances, 35 features and 15 classes.

These datasets can be downloaded from the course webpage. All the datasets are multi-class classification datasets i.e., have more than two classes.

2 Determining Clustering Quality

2.1 Sum of Squared Errors (SSE)

Assume that we have a clustering for our data X having K clusters. Each cluster is characterized by it's centroid C_k. The sum of squared error criterion for one cluster can be defined as:

$$SSE(X, C = k) = \sum_{x_i \in cluster_k} \|x_i - C_k\|^2$$

then, the overall SSE for the K clusters can be calculated as:

$$SSE(X) = \sum_{k=1}^{K} SSE(X, C = k)$$

SSE provides a notion of how compact the clusters are and we are going to favor clusters that are more compact, hence we will look for clusterings that minimize SSE. This is an internal clustering quality criterion as we can estimate it using just the cluster labels.

2.2 Normalized Mutual Information (NMI)

Assume that we have a clustering for our data X having K clusters. Furthermore, we assume that each instance has a class-label $y_i \in \{1, 2, \ldots, M\}$. In this case we can estimate the quality of a given clustering based on how "pure" each cluster is with respect to the class labels. In this case the optimal clustering would consist of k clusters such that every cluster has instances belonging to only one class. Given a clustering consisting of k clusters for a dataset that has M classes, NMI can be defined as:

$$NMI(Y, C) = \frac{2\ I(Y;C)}{H(Y) + H(C)}$$

where, I(Y;C) is the mutual information between the class labels and cluster labels and H(Y) and H(C) is the Entropy of the class labels and the cluster labels, respectively (See the supplementary material that provides the details of how to calculate NMI given a dataset with class labels and a clustering). NMI quantifies the information we would get about the class labels if we have access to the cluster labels, it is similar to the Information Gain criterion we used to construct decision trees earlier in the course.

3 K-Means Clustering

You will use the k-Means clustering algorithm to cluster all six datasets. In order to have a suitable clustering you would need to find the number of clusters (k) that produce the best clustering. For k-Means this can be achieved by trying different values of k and tracking the SSE criterion.

3.1 Programming Tasks:

1. Implement the k-Means algorithm.
2. For each dataset calculate the SSE for different values of k.
3. Select the optimal number of clusters based on the SSE criterion,and calculate the NMI of the resulting clustering. Briefly explain how you selected the optimal number of clusters.

3.2 Deliverables:

1. For each dataset provide the plot of the SSE vs k (number of cluster).
2. In a table provide the optimal number of clusters for each dataset based on the SSE criterion, and the corresponding NMI.
3. Set the number of clusters equal to the number of classes for each dataset and run the k-means algorithm. List the resulting NMI and SSE for each dataset in a table.

4 Gaussian Mixture Models (GMM)

In this part you will use the GMM clustering algorithm to cluster all six datasets. In order to have a suitable clustering you would need to find the number of clusters (k) that produce the best clustering. Use both NMI and SSE to find the (possibly two different) value(s) of k for each dataset.

4.1 Programming Tasks:

1. Implement the GMM algorithm based on the EM algorithm discussed in class.
2. For each dataset calculate the SSE for different values of k.
3. For each dataset calculate the NMI for different values of k.
4. Briefly explain which criterion i.e., SSE or NMI is better for GMM and why.

4.2 Deliverables:

1. For each dataset provide the plot of the SSE vs k (number of clusters).
2. For each dataset provide the plot of the NMI vs k (number of clusters).
3. In a table provide the optimal number of clusters for each dataset based on the SSE criterion, and the corresponding NMI.
4. In another table provide the optimal number of clusters for each dataset based on the NMI criterion, and the corresponding SSE.
5. Set the number of clusters equal to the number of classes for each dataset and cluster the data using GMM. List the resulting NMI for each dataset in a table.

5 Comparing k-Means and GMM

Please provide brief answers to the following, supporting your answer by providing empirical evidence (when necessary):

1. For each dataset which algorithm would you use to cluster? why?
2. Does the clustering for each dataset gives you any insight about the separability of the classes?
3. *Extra-credit:* Are the k-Means and GMM algorithms sensitive to how the clusters are initialized? What strategies can you formulate for better initialization of each algorithm?
4. *Extra-credit:* Do you think that the distance metric we used to cluster the datasets is appropriate? Provide your opinion about each dataset.