Amrit – 1211075109

Vahini Mohan – 1210414124

Data Mining Assignment – 5

1. K-means Clustering is implemented using the standard k-means algorithm from class. A Random sample of k (given) is chosen as centroids for the first iterations. The Euclidean distance is calculated for each of the points and for a given point, the closest one is selected as the Cluster. Function "Sumsqr" is used to calculate the SSE. The centroids are recalculated using "mean" function and substituted. A total of 100 iterations are run unless convergence happens earlier if difference between consecutive SSE values is 0.001 or lesser. For the Data given the SSE values at 100 iterations are as follows for different trials, since the answer varies due to randomization:

   SSE k = 3: [587.31 between 588.91]

   k = 5: [386.15 between 431.35]

   k = 7: [277.8805 between 325.43]

2. Three methods are used, k-nearest neighbor with k = 7, SVM with a polynomial kernel of degree 2 and feedforward neural network with a single hidden layer with 25 neurons to classify the data. The values reported are:

   Knn: 95.23%

   SVM: 89.97%

   Feedforward neural network: 92.61%

   Ensemble: 93.63%

   The ensemble prediction is obtained by taking a majority vote on the predictions (using mode function) of these individual models and comparing it with the actual test label.

Functions Used:

1. Mean – Average of an array of elements
2. Sumsqr - Sum of squared elements of matrix or matrices
3. pdist2 – pairwise distance between two sets of observations
4. Mode – returns the maximum element in a list.
5. Fitcknn() - A nearest-neighbor classification object, where both distance metric ("nearest") and number of neighbors can be altered. The object classifies new observations using the predict method.
6. fitcsvm(___,Name,Value) - returns a support vector machine classifier with additional options specified by one or more Name,Value pair arguments, using any of the previous syntaxes.
7. Predict() - predicts the output of an identified model.
8. patternnet(`hiddenSizes,trainFcn,performFcn`) – pattern recognition networks are feedforward networks that can be trained to classify inputs according to target classes.
9. Train (`net,X,T`) – Trains a neural network according to the topology specified. Net – network, X – network inputs, T – network targets.

References:

1. https://www.mathworks.com/help/matlab/
2. http://matlab.mathworks.com