Assignment 4

Due: Nov. 22

1. (40 points) For this question we will explore the k-means clustering algorithm using Weka.We will work with the `segment.arff` dataset distributed with this assignment. This dataset is based on a set of images taken in color around the UMASS campus to which low-level image processing operators were applied. The goal is to find clusters in the data which define different types of objects (buildings, trees, sky etc). But you need not be concerned with understanding the meaning of each cluster.

   (a) Load the `segment.arff` file into Weka. Apply the "SimpleKMeans" algorithm under the clustering algorithms. Vary the k values as $k = 1, 2, \ldots, 10$, and make 5 clustering runs for each k value. For each clustering run, set the "seed" parameter as 1, 2, 3, 4, 5 respectively to obtain different starting points. Report the SSE values of each clustering run for each k value.

   (b) For each $k = 1, 2, \ldots, 10$ compute the mean SSE, which we denote μ_k and the sample standard deviation of SSE, which we denote σ_k, over all 5 clustering runs for that value of k. Produce a table containing the 4 columns: k, μ_k, $\mu_k - 2\sigma_k$ and $\mu_k + 2\sigma_k$ for each of the values of $k = 1, 2, \ldots, 10$.

   (c) As k increases and approaches the total number of examples N, what value does the SSE approach? What problems does this cause in terms of using SSE to choose an optimal k?

2. (20 points) Consider the following dataset:

   `{ 0, 4, 5, 20, 25, 39, 43, 44 }`

   (a) Build a dendrogram for this dataset using the **single-link, bottom-up** approach. Show your work.

   (b) Suppose we want the two top level clusters. List the data points in each cluster.

3. (20 points) Given two clusters

   $$C_1 = \{(1,1),(2,2),(3,3)\} \qquad C_2 = \{(5,2),(6,2),(7,2),(8,2),(9,2)\}$$

   compute the values in (a) - (f). Use the definition for scattering criteria presented in class. Note that tr in the scattering criterion is referring to the trace of the matrix.

(a) The mean vectors m_1 and m_2

(b) The total mean vector m

(c) The scatter matrices S_1 and S_2

(d) The within-cluster scatter matrix S_W

(e) The between-cluster scatter matrix S_B

(f) The scatter criterion $\frac{tr(S_B)}{tr(S_W)}$

4. (20 points) A Naive Bayes classifier gives the predicted probability of each data point belonging to the positive class, sorted in a descending order:

Instance #	True Class Label	Predicted Probability of Positive Class
1	P	0.95
2	N	0.85
3	P	0.78
4	P	0.66
5	N	0.60
6	P	0.55
7	N	0.43
8	N	0.42
9	N	0.41
10	P	0.4

Suppose we use 0.5 as the threshold to assign the predicted class label to each data point, i.e., if the predicted probability ≥ 0.5, the data point is assigned to positive class; otherwise, it is assigned to negative class. Calculate the *Confusion Matrix, Accuracy, Precision, Recall, F1 Score* and *Specificity* of the classifier.