**Intelligent Data Analysis CS5152/6052**

**Homework #2**

**Due Date(s): Problem #1 is due by March 30th, 9PM; Problem #2 is due by April 4th, 9PM.**

Consider the data in a 2-dimensional space given in the attached file. It includes points belonging to two different classes. A plot of the data points showing their distribution is attached with the assignment. Perform the following tasks with this dataset.

Include your response to each part of a homework question in a single pdf file and upload it on Blackboard. There are two assignments created on Blackboard, one for responses to Question#1 and the other for Question#2. Two separate submissions are required by their respective due dates.

1. Your task is to design a good classifier for this data set. You can use only one of the following two toolboxes: Scikit library (Python), or Matlab. For this question we are going to find a non-linear SVM classifier that fits this data. Perform the following steps and report your work and results as stated below.
   1. Use non-linear SVM with Radial-basis kernel, to design classifier for the two classes of this dataset. Randomly select 75%% of the data points as training data and use the remaining 25% as the test data. Create and test an SVM model for at least 6 different values of the regularization parameter (this parameter is named differently in each toolbox, read the documentation for your toolbox). For each of the six runs, show the chosen parameter value, confusion matrix, and accuracy, and precision values. Plot the accuracy, precision, and recall values for the six parameter values.
   2. Select the model having the highest accuracy value from the six models generate above. For this model consider all the points of the 100X100 grid as the test points. Find the class for each of these test points. Use two different colors for denoting the predicted class label for a point and create the plot with corresponding colors for each point of the full grid. Your output will be a grid in which each point will be plotted using one of the two colors.
   3. The grid generated in (b) above shows the boundaries learned by the classifier to predict the two classes. Comment on the following:
      1. How efficient and effective are these boundaries, given data distribution of the two classes.
      2. On a plot of original data points draw the ideal boundaries that you would like to see from your intuitive point of view. Give reason for your choosing the boundaries that you have drawn.
      3. Which classifier, in your view, can learn the boundaries closest to the intuitive ones you have suggested? Give reasons for your answer.
   4. What is the role of the regularization parameter and how does it work? (Maximum 200 words)
2. Now use the same data set for the following sequence of steps.
   1. For this problem consider only the (x,y) coordinates of the data points and ignore their class labels. Perform k-means clustering of these data points for values of k to be 3, 5, 7, 9, and 11. For each value of k run the clustering algorithm (Scikit or Matlab only) 6 times (with initial cluster centers being selected randomly) and report the following: (i) total SSE for each clustering run, (ii) Average SSE and its standard deviation for each k value, and (iii) the minimum and maximum SSE values for each value of k.
   2. Consider the best clustering obtained for k=5. Each data point is assigned a cluster number by your clustering algorithm. Plot all the given data points on the grid after assigning a different color/symbol to members of each cluster. Write the cluster-specific SSE for each individual cluster on the plot.
   3. Comment on the cluster boundaries obtained by your clustering algorithm, focusing on how they are different from your intuitive idea of five clusters from this dataset.
   4. On the data plot, mark the boundaries of the five clusters that you think are intuitively distinct clusters. Justify the boundaries you have drawn, giving reasons for preferring the boundaries you have drawn.
   5. Compute the Rand Index between the clustering used in parts (b) and (d) above. Show your work and steps performed to arrive at the Rand Index value. How can we interpret the meaning of the Rand index obtained by you?