Assignment: Clustering and Radial Basis Functions (RBF)

Problem 1 (20 points)

In this problem you will use the data file "kMeansData.csv" (x_1 and x_2 denote the input features) to create 3 clusters using unsupervised Lloyd's k-means algorithm.

The training should only stop if the difference between the cluster center locations in two consecutive iterations is less than 0.001 or if the number of iterations has reached 1000. For the initial selection of cluster locations choose 3 points from the data set randomly.

After convergence, report the final cluster centers. Plot the 3 clusters in different colors with cluster centers clearly marked on the plot.

Problem 2 (20 points)

In this problem you will use the data file "rbfClassification.csv" to create an RBF classification model. x_1 and x_2 denote the input features and cls denotes the target class of the corresponding data points.

1. Use k-means clustering to determine the location of 2 cluster centers that you will use in your RBF model. Report the coordinate of the cluster centers.

2. Train an RBF model using $\gamma = 0.5$. Report the correct classification rate of your model.

Problem 3 (35 points)

The "stuFile.csv" is a list of students' features at an academic institution that is considering starting a new program. The features include average yearly packaged financial aid, the number of years that the financial aid is awarded for, gender, marital status, marketing code, previous education, admission representative code, program code, citizenship code, ethnicity code, veteran code, and cancel flag code. Some of the students listed in the file cancelled their enrollment after meeting with Financial Aid. These students can be identified by the value of 1 in the cancel flag field.

This institution is considering starting a new program after researching the cannibalization rates, navigating regulatory requirements, and determining the potential market. There is an opportunity to market this program to students who cancelled before starting their originally intended program. The presumption is that the shorter duration and lower tuition rate of this program may be attractive to some students who cancelled.

Your task is to look into the data and provide some insight into who "might" enroll from this population using the clustering schemes discussed in class.

1. Use Value Metric Difference (VDM) to find and report the distance between the different levels of all categorical variables. (10 points)

2. Use VDM along with other distance metrics for continuous features to cluster the data. Justify the number of clusters you have used. Use Elbow plot. (10 points)

3. Report the number of cancels and starts in each cluster in a table. (2 points)

4. Provide a heat map of admission reps, marital status, lead category, and previous education map for your clusters. (8 points)

5. In a few sentences, provide insight into identifying those students that are most likely to be converted. List these students. (5 points)