Assignment: Clustering and Radial Basis Functions (RBF)

Problem 1 (20 points)

In this problem you will use the data file "QB2022_MLR.csv" and "QB2022_MLR_test.csv".

1. Create a multiple linear regression model that predicts the quarterback ratings based on 'Pass Yds', 'Yds/Att', 'Att', 'Cmp', 'Cmp percent', 'TD', and 'INT'. Normalize your values using MinMaxScaler(). You will use KernelRidge from the sklearn library. Set kernel to 'linear' and alpha to zero for regularization. Use default paramter values for everything else. Keep in mind that you need to add a column of ones to your data for bias. **Report MSE for the training points**.

2. Create an RBF regression model that predicts the quarterback ratings based on 'Pass Yds', 'Yds/Att', 'Att', 'Cmp', 'Cmp percent', 'TD', and 'INT'. Normalize your values using MinMaxScaler().You will use KernelRidge from the sklearn library. Set kernel to 'rbf' and alpha to zero for regularization. Use default paramter values for everything else. Keep in mind that you need to add a column of ones to your data for bias. **Report MSE for the training points**.

3. Find and **report the MSE error for test points in "QB2022_MLR_test.csv" using both models**.

4. Compare both the training and test errors for these two models, and report your observations in a couple of sentences.

Problem 2 (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 3 (20 points)

In this problem you will use the data file "rbfClassification.csv" to create an RBF Network model for classification. 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 Network using $\gamma = 0.5$. Report the correct classification rate of your model.