EE 524 Machine Learning Lab

Assignment 5

09 November 2020

K Nearest Neighbours Algorithm: Go through the resources mentioned in the slack work space. We will implement this algorithm from beginning without the help of scikit learn.

1. **Load the Data set** Download and read the iris flower data set as a csv file from here.

2. **Normalize the Data set** Once you read the csv file, try finding the minimum and maximum values of each feature. Use these values to normalize the data. What is Normalization? Use Range normalization aka minmax normalization (Dont use the inbuilt function).

3. **Distance Metric** Use the Euclidean distance metric for calculating distance between 2 arrays. Write a function that calculates the Euclidean distance between 2 input arrays. This will be later used to calculate the closeness of any test sample with rest of the records.

4. **Nearest Neighbours** Use the distance function to get the nearest neighbours for a test record. Write a function that takes in 3 inputs, the dataset, the test example and the number of neighbours (k). The output of this function will be a list of k entries in the dataset that are closest to the test record.

5. **Prediction using KNN** Write a function that takes 3 inputs, first the dataset, second a test record and third the number of neighbours. The output of this function will be to return the label of the test record using KNN.

6. **Test** Use the KNN predictor, to classify the following records:

   1. [5.1, 2.5, 3.2, 4.3]
   2. [4.9, 3, 1.4, 0.2]
   3. [6, 3, 4.8, 1.8]

   Take k as 5.

7. **Accuracy** Use the same KNN predictor to find the labels of the training data. Report the accuracy for each class. You can use confusion matrix to visualize the result (not necessary).

8. **Tuning** Use the KNN predictor on the dataset for different values of k. Calculate the accuracy for different values of k and see which k gives the maximum accuracy. This is known as hyperparameter tuning. Are there different ways of choosing k?

K Means Algorithm: Go through the material on K Means shared in the work space. K Means is a very powerful algorithm for initial analysis of data as well as clustering. There are different types of k means but we will focus on the most basic one.

1. **Dataset** Load the dataset and read it as a csv file from here.

2. **Visualization** Plot the dataset using matplotlib.

3. **K Means Algorithm** Use the inbuilt algorithm for K Means in scikit learn. Vary k from 2 to 10 and plot the clusters in different colours to visualize.

   From the plots, which value of k is the most appropriate for this dataset.