Machine Learning Project Report

Bharath Reddy Janumpally

Sundarkumar Ramsundar

Georgia State University

**Question 1 (Missing Value Estimation)**

In this question missing values in gene microarray data had to be predicted.

First we tried using **Neural Network** for this question, but as the sample size was small and the feature size was very large, the neural network did not perform well.

Secondly, the logic behind the solution is that the genes which are closest in terms of distance have more influence on each other. So, we calculate the distance between genes and find the k nearest neighbors and take the weighted mean of the k nearest neighbors to find the missing values. The weights are proportional to inverse of the distances found.

KNN is a famous method in Machine Learning we modified the method to include **weighted KNN. The weights of the neighbors is inversely proportional to the Euclidian distance.**

**Steps in the solution**

1. Initially take the missing values as the mean of that gene
2. Compute Euclidian distance to all the other genes.
3. Ignore the missing value component of the missing genes
4. Find k nearest neighbors for a missing gene
5. Replace missing values for by the weighted average of the k nearest neighbors.

This gives the output with the missing values predicted.

**Simulation result using 10 fold validation**

For Dataset 1

1. For K values in the range 1-5 the error rate is high
2. For K values in the range 6-10 the error rate is medium
3. For K value 11 the error is least.

For Dataset 2

1. For K values in the ranger 1-10 the error rate is medium
2. For K values in the range 11- 20 the error rate is high
3. For K values in the range 21 – 38 the error rate is low. Lowest being at K = 38

Observation

So, we observe that the K values vary from dataset to dataset because of variation in the distance between the data points in the training set.

The accuracy of 10 fold validation varies based on which 10% of the data are we using to test.