EAS506 - Statistical Data Mining I

Homework 1 – Question 4

Paul M Girdler

09/14/18

Abstract

This is the solution to Question 4 of Homework 1. Linear and KNN models were compared on a classification of the *zipcode dataset*, which is available in the **ElemStatLearn package**. KNN model, at $K = 1$ was the surperior model .

Content

1 Introduction

This report compares the classification performance of linear regression and k-nearest neighbour (KNN) classification on the *zipcode dataset*, which is available in the **ElemStatLearn package**.

For this problem only classification of the 2's and 3's were considered. Both the training and the test error for each choice of $k = 1, 3, 5, 7, 9, 11, 13, 15$ for the KNN model needs to be plotted, as well as the linear model.

2 Method

2.1 Initilization Steps

- Clear the memory
- Install and load all required libraries.
- Data does not need to be divide into test and training as it has already come prepared.
- Briefly examine the data.
- Create a subset of the data where target V1 is equal to 2 or 3 as required by this exercise.

2.2 Linear Regression Model

- Create a simple linear regression model based from the training data to predict the response variable V1.
- Create a predicted response of V1 using both the training and test data and store the results.
- Convert the prediction response into a binary classification to the target.
- Calculate the prediction error rate for the training and test data.

2.3 K-Nearest Neighbours Models

- Build a KNN function that takes K as an argument and produces a model for that K based on the training data.
- Classify the training for each K model and calculate the prediction error rate for each K. Odd numbers of K are used to prevent.

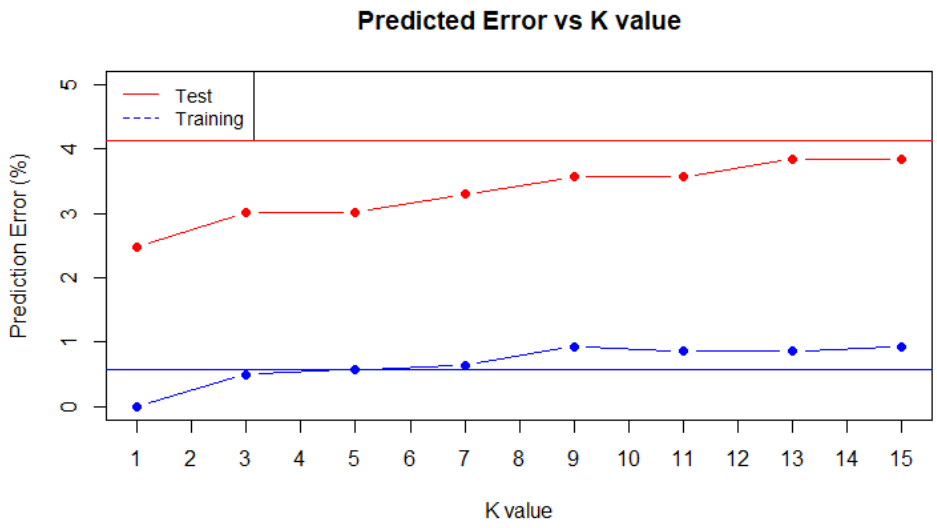
- Build a KNN function that takes K as an argument and produces a model for that K based on the test data.
- Classify the test for each K model and calculate the prediction error rate for each K.
- Store all the prediction error rates.

2.4 Plot the Results

- Plot the results of the prediction error for both training and test models for linear models, and all KNN models.

3 Results

Figure 1 – Predicted Error vs K value and Linear Models



4 Discussion

Both the linear model and the KNN models performed better on the training data than the test data as expected. The KNN models were comparable to the linear models at around K = 15. However, at lower K values, the added model complexity enabled enhanced the prediction rate.

As a rule of thumb one should expect the K value to be best at approximately:

$$K = (N_{samples})^{1/\text{N-dimensions}}$$
$$K = (7{,}291)^{1/256} = 1.035$$

The KNN model performed the best when K was equal to 1.