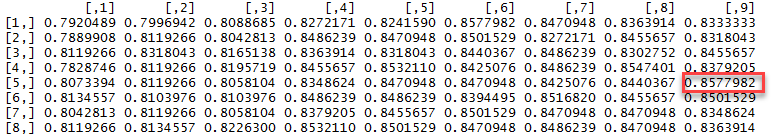
ISYE 6501, Week 2 HW

**Question 1**

Using the same data set (credit\_card\_data.txt or credit\_card\_data-headers.txt) as in Homework 1 Question 2, use the ksvm or kknn function to find a good classifier: (a) using cross-validation for the k-nearest-neighbors model; and (b) splitting the data into training, validation, and test data sets.

**Response –**

1. The function cv.kknn is used for cross-validation of the k-nearest-neighbors model. Multiple k-folds (8 values) and k-values (9 values) are applied to the function using 2 for loops and the accuracy is recorded in a matrix. Out of all k-folds and k-values, k-folds=8 and k-value=20 has the highest accuracy of 85.78%



1. First the data is split 80-20. 80% for training and validation, and the remaining 20% for testing. The training and validation data is again split into 10 partitions and each partition is selected as validation set, while the remaining 9 are used as training it. Using a ‘for loop’ iteration, each partition is used as a validation set.

Cross validation is performed on ksvm and kknn model to compare their performance with the training and validation data. The accuracy of each iteration is stored into a list. The overall accuracy of the model will be average of the accuracies calculated during each iteration. The mean accuracy for kknn model came out to be 85.05% whereas the mean accuracy for ksvm model came out to be 86.6%. Since, ksvm model has higher accuracy, the testing data will be applied on ksvm model to calculate the performance.

Upon running the ksvm model with test data, the performance of the model came out to be 84.73%. This is how better the model predict the data. It is also worth observing that the ksvm model accuracy has gone down from 86.6% to 84.73% upon testing against the test data. The reason is that the model was trained on training+validation data set which might have different randomness than that is present in the testing data.

**Question 2.1**

Describe a situation or problem from your job, everyday life, current events, etc., for which a clustering model would be appropriate. List some (up to 5) predictors that you might use.

**Response –**

For a financial investment company, to extend different kinds of financial services like new investment funds or hedge funds to its customers and to check the viability of the product/service, the firm will have to find the right group of customers, be it individual investors or business investors. Among the huge list of customers to whom the company might be providing financial services, clustering model will help in coming up with understanding their target customer base and to work on providing better curtailed investment products to them. It helps grouping customers based on their similarities (investment type, ROI etc.) related to the product under consideration.

The following can be some of the predictors that the investment firm would like to consider -

* Customer Investment Total
* Customer Type – Individual or Business
* Investment Type – S&P500 Fund, Bonds, Hedge Fund
* Risk appetite of the customer
* Length of Relationship as customer
* Average duration of their investments
* ROI expectation of the customer

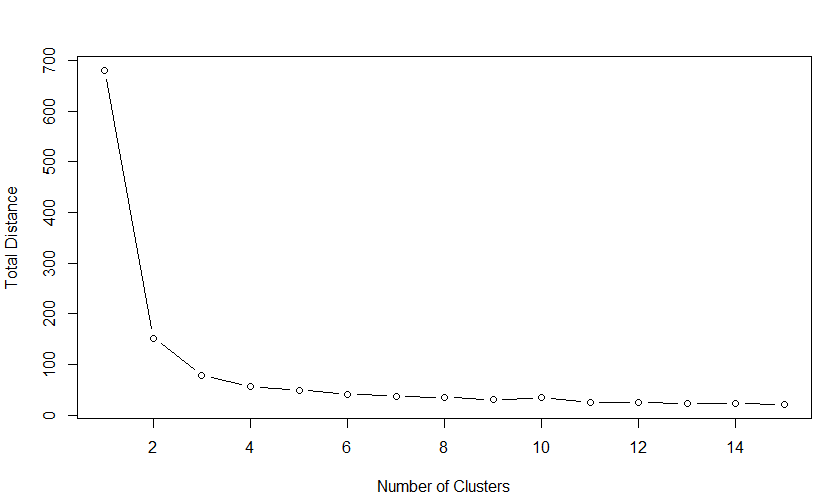
**Question 2.2**

Use the R function kmeans to cluster the points as well as possible. Report the best combination of predictors, your suggested value of k, and how well your best clustering predicts flower type

**Response –**

Upon qualitative review of the data, it can be observed that the petal length and petal width predictors do follow a specific range for each species of flowers and tend to be better predictors than the other 2.

To find the best value of k in the k-means clustering, the total square distance within cluster is plotted against the k value as an elbow curve to find at which value of k the difference of total squares is the least for the first time. From the curve below, it looks like k=7 will be a good choice.



But upon looking at the species data, there should be 3 clusters as there are 3 species of flowers.

**Case 1 – No knowledge of species**

K-means clustering is performed on iris data with k=7 and using all 4 predictors and the accuracy (bss/tss ratio) came out to be 95%. Now k-means is again performed with k=7 but this time using only 2 predictors – petal length and petal width as the parameters for clustering. The accuracy in the latter case is 98.33%. When sepal length and sepal width are used as predictors, the accuracy fell even further down to 89%.

Best combination of predictors – petal length and petal width

Suggested value of k = 7

How well the cluster predicts the flower type – 98.33%

**Case 2 – Knowledge of species available**

K=3 is used as there are 3 type of species available. The accuracy measured in terms of bss/tss ratio is 94.7%.

However, when compared with the species data, only 6 points are misclassified and remaining 146 points are correctly classified.

Best combination of predictors – petal length and petal width

Suggested value of k = 3

Accuracy = 144/150 = 96%

This is where qualitative review of the data makes sense. Just by comparing different k-values, it may appear that k value of 7 produces the best results. But, having observed the response data, it can be said that k value of 3 makes more sense as the iris data has only 3 species of flowers.