Laboration 1 (Datamining with R)

Diana Saveh och Linus Sjöbro

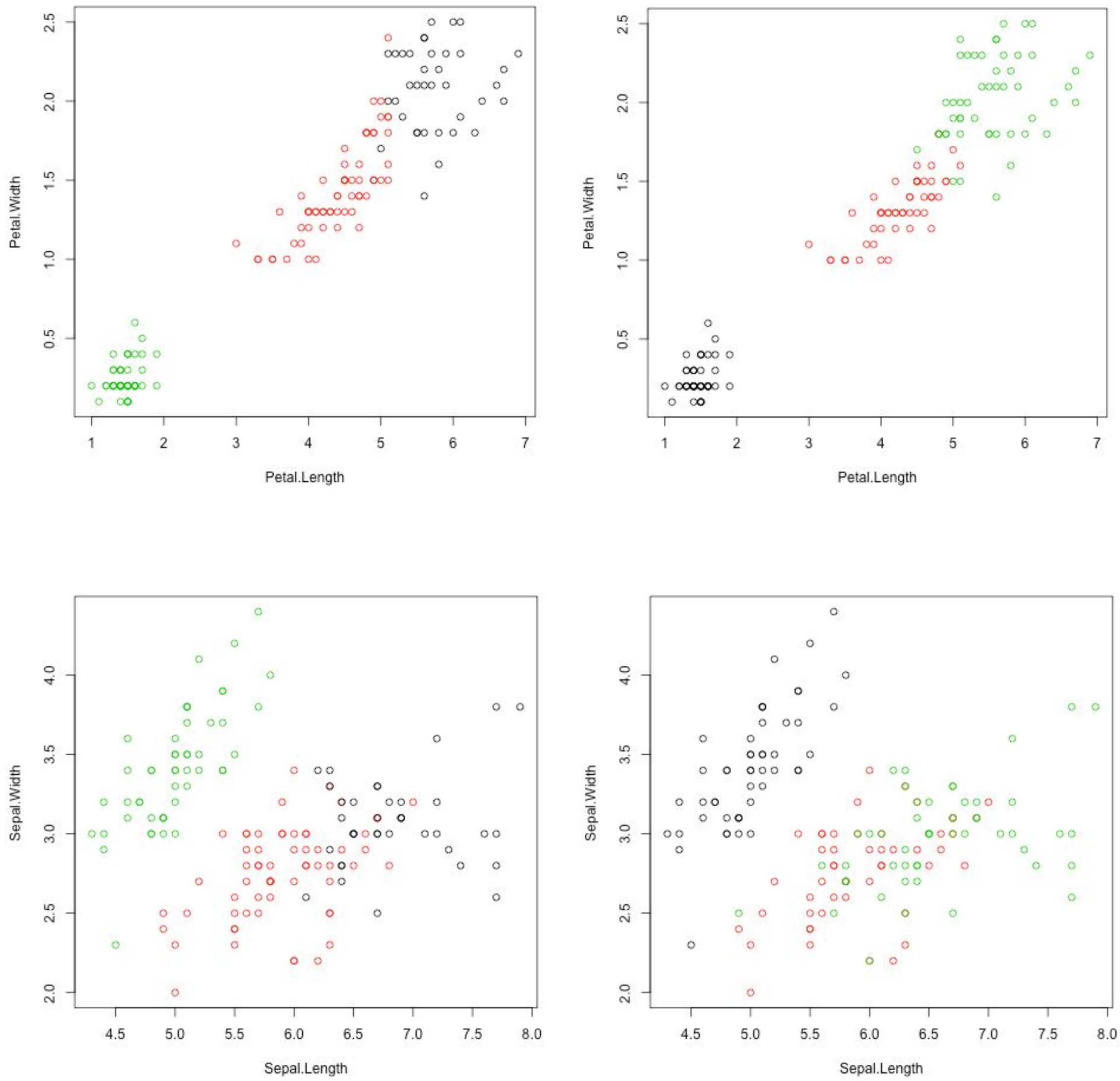
K-means

The k-means is a clustering system which takes in parameters. In the parameter called k you can specify how many clusters are needed, then the k gives different spots unsystematically in different places, and these dots are then the midpoint of the clusters. K-means is an unsupervised algorithm.

In the first task the method K-means is used which is a method for conducting a cluster analysis. The known dataset Iris is used and divided into three clusters.With the K-means method, the flowers are clustered according to the width and length of petals and petals.

Accuracy in solution
To calculate the accuracy of the solution for K-means, the formula was used: the total number of classified instances divided by the total number of incorrect and the total correct classified instances.

Results



Below is the confusion matrix.

	1	**2**	**3**
Iris-setosa	0	0	50
Iris-versicolor	2	48	0
Iris-virginica	36	14	0

Total number of correctly classified instances are: 36 + 48 + 50= 134
Total number of incorrectly classified instances are: 2 + 14= 16
Accuracy = 134/(134+16) = 0.893 i.e our model has achieved 89% accuracy!

K-Nearest-Neighbor (KNN)

Knn is a supervised algorithm , KNN will be trained to recognize the species. it will look at the nearest neighbors and takes the maximum number, but this depends on the K. KNN calculates the distance between test and training objects in order to be able to classify the test objects.

Overfitting/underfitting - none of them are good, k should be in between.

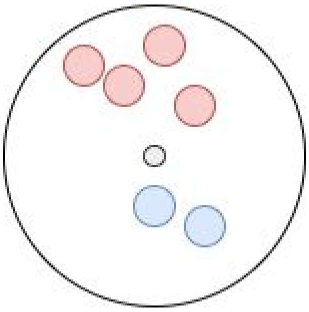
Overfitting

Overfitting is when the k value is too small, which then cause the result to be very sensitive against jitter. If a faulty data point is against these k-values the whole result will be incorrect.

Underfitting

Underfitting is when the value k is too large, which then cause the result to gather too many data points. If, for example, we have a data set with a total of 20 points, then 15 of these points are green and 5 orange. if one green points is near the orange group this points will get discarded and not included in the result.

This figure below demonstrates when there is too many red points compared to blue points, which gives this graph an underfitting result.



Results

We used “set.seed” to get random seeds to use on the algorithm. The total iris dataset contains of 150 rows. We have chosen a 80%/20% ratio between the training and test set. This means that the training set contains of 120 rows and test set of 30 rows.

To get an accurate K , we used the square root of the total amount of rows in the dataset. In this task there was 150 rows , that gave the answer of 12,247 so a good k value should be 12 or 13.In the above mentioned method with square root of the data set size we get that the K value should optimally be 12-13.

To test the impact the K value gives we have tested different spans. When the k value is low the accuracy could be 100% which is not a reliable answer , and this is why it is important to choose a correct K value. Test results is in in figure 1.

K value	Accuracy (%)
3	90
5	100
10	97
13	97
40	83

figure 1