k-Nearest-Neighbours

Lab Session 1
Machine Learning: Pattern Recognition
Master Artificial Intelligence

Camiel Verschoor
StudentID: 10017321
UvAnetID: 6229298
Verschoor@uva.nl

Steven Laan
StudentID: 6036031
UvAnetID: 6036031
S.Laan@uva.nl

September 11, 2012

1 Data Visualization

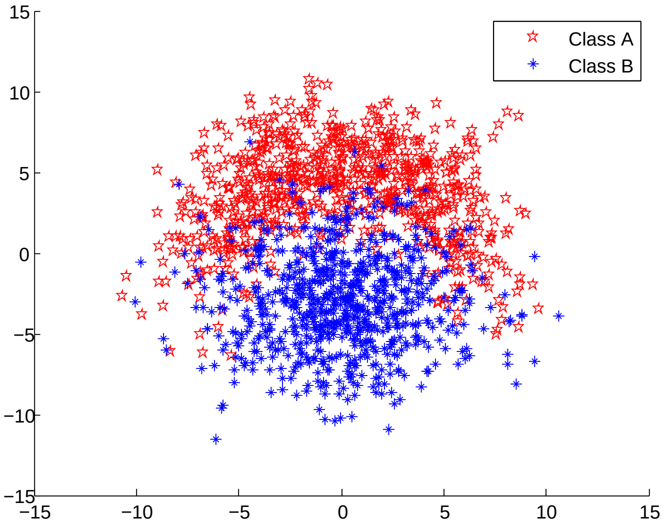

Figure 1: Data visualization.

In figure 1 the data in the training set, which consists of the two classes, A (red) and B (blue).

2 k-Nearest Neighbours

A kNN classifier, with $k = 1$, is trained on the trainings data. The performance is evaluated on the test data. This resulted in the following confusion matrix:

	True	**False**
Positive	206	44
Negative	36	214

From the confusion matrix the error rate is computed by the following formula:

$$1 - accuracy = 1 - \frac{tp + fp}{tp + tn + fp + fn} \tag{1}$$

The error rate and other statistical measures of the classifier are presented in the table below:

Accuracy	84.0%
Precision	82.4%
Recall	85.1%
F-measure	83.7%
Error rate	16.0%

In figure 2 a graph is shown of the test error evolving as a function of k. Notice that test error increases when k is growing, this is caused by overfitting on the training set. Furthermore, it can be noticed that the function is fluctuating, which is caused by the noise of the classifier.

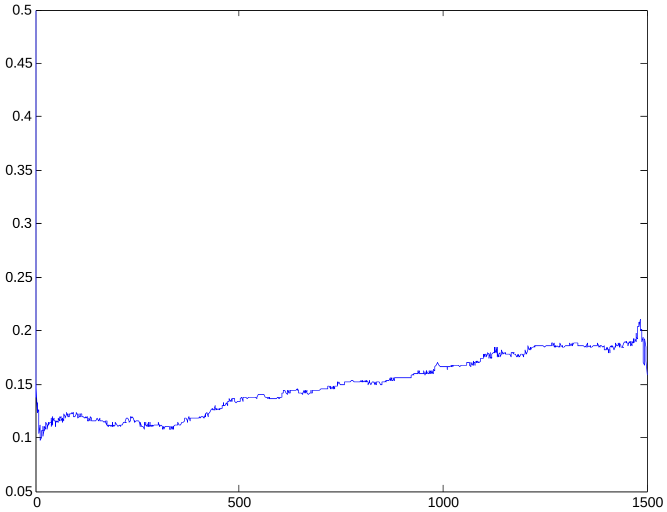

Figure 2: Graph of the error rate with various K

From this single experiment we cannot extract a specific value of k, because it could be the case that the data used for training is not representative of the total distribution. If we must decide on a value for k, given these results, we would choose something between 100 and 250, because in the graph there is clearly a low plateau. In order to determine which value in the range 200-250, statistical measures like the P-test can be employed. This way we can check statistically which values lie on this minimal plateau. Then we will want to choose the highest value for k that statistically still lies on that plateau. This way we generalize the most, while keeping the best performance. By looking at the graph, we guess this k is 240.

3 Cross Validation

The advantage of using an evaluation method like cross-validation and bootstrapping is that this method evaluates how well a hypothesis of the classifier performs on predicting new data.

A validation set is a portion of the dataset used to asses the performance of prediction or classification models that have been fit on a separate set, training data, of the dataset. The validation set is used as a more objective measure of performance of various models that have been fit on the training data as validating the performance with the training set is not likely to be a good guide to the performance of the models on new data.