Applications of Data Analysis
Documentation
Exercise 1

Kasperi Keski-Loppi
18.01.2016

## Problem

The initial problem was to find the best k for the data in use. When given a new data point, the k value represents the amount of neighbors used for the new data point classification. In another words, the k value is the amount of total votes used to determine the majority vote, i.e. the class, for the new data point.

In order to find the best k, we had to implement the k-nearest neighbor algorithm along with some cross-validation. The cross-validation is used to compare the performance, i.e. the accuracy, of the different values of k.

A tutorial[1] was used as a help for this implementation as the subject was new to me.

## Implementation idea

The best k value can be determined by iterating over k and choosing the k that yields the best accuracy. The accuracy can be calculated by using the leave-on-out method for each k value.

When using the leave-on-out method, the training set and the test set will change for every round. In this implementation, the test set will have only one instance at a time and each instance will be in the test set exactly once. Consequently, the training set will include the rest of the instances.

The implementation structures in the following way:

1. Read the data
2. Split the data into training set and test set
    a. The test set will be the instance determined by the loop iterator and the training set will be the rest of the instances
3. Find the closest neighbors for each k
4. Get the response (i.e. the majority vote) for each k
5. Compare the predicted result (the response) and the data point’s actual value
6. Print the best k value that has the best accuracy

Applications of Data Analysis<br>
Documentation<br>
Exercise 1

Kasperi Keski-Loppi<br>
18.01.2016

## Description of the data set

The dataset can be found from here: http://archive.ics.uci.edu/ml/datasets/Iris. It contains 3 classes of 50 instances each, where each class represents a type of iris plant. In this implementation, the predicted attribute is the class of iris plant.

The attributes of each instances are:

1. Sepal length in cm
2. Sepal width in cm
3. Petal length in cm
4. Petal witdh in cm
5. One of the classes
    a. Iris Setosa
    b. Iris Versicolour
    c. Iris virginica

## Results and observations

This implementation yields the best k value of 19 with the accuracy of 97,987%. In addition, the accuracy drops substantially starting from the k value of 99.

At first, the value of k seemed to be rather high compared to the examples in the lecture slides and elsewhere in the internet. Though, we concluded that the initial dataset is quite small and the data points are packed into smaller areas. Also, we made some small experiments by editing the variables and the initial dataset which seemed to work great resulting in more confidence that the implementation is done correctly.

## Source

1. http://machinelearningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/