Applications of Data Analysis
Documentation
Exercise 3

Kasperi Keski-Loppi
02.02.2016

Problem

The initial task was to predict total metal concentration (c_total), concentration of Cadmium (Cd) and concentration of Lead (Pb) in tap water. The prediction performance was calculated using the C-index along with the true and predicted values.

Implementation idea

The predictions were done using the k-nearest neighbor algorithm and two different cross validation approaches: leave-one-out and leave-three-out. Also, the data was prepared by normalizing it using the z-score.

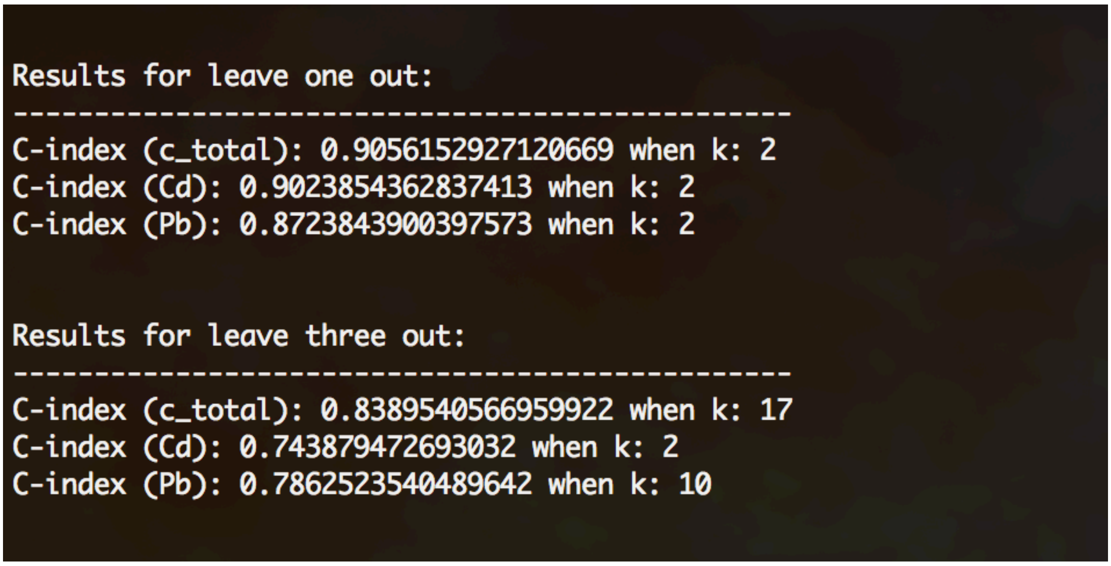
The implementation structures in the following way:

1. Read the data
2. Convert numerical values into floats
3. Normalize the data using the z-score
4. Run the knn algorithm using both cross validation approaches
    a. Determine the test instance / test instance set
    b. Get the neighbor(s) of of those instances
    c. Get the predictions
    d. Calculate the c-index for each output (c_total, Cd, Pb)
    e. Print the best c indexes and their corresponding k values

Description of the data set

Unfortunately, the dataset is not publicly available as it is being used in academic research for the time being. But to introduce it briefly, it included 201 data points obtained from 67 mixtures of Cadmium, Lead and tap water. Each of these data points had six attributes so that the first three were to be predicted and the last three was used for the prediction.

Results



```
Results for leave one out:
------------------------------------------------
C-index (c_total): 0.9056152927120669 when k: 2
C-index (Cd): 0.9023854362837413 when k: 2
C-index (Pb): 0.8723843900397573 when k: 2


Results for leave three out:
------------------------------------------------
C-index (c_total): 0.8389540566959922 when k: 17
C-index (Cd): 0.743879472693032 when k: 2
C-index (Pb): 0.7862523540489642 when k: 10
```

The screenshot represents the results for both the leave-one-out and the leave-three-out approaches. The best C-indexes and their corresponding k values were calculated for each output.