Applications of Data Analysis
Documentation
Exercise 5

Kasperi Keski-Loppi
11.02.2016

Problem

Our task was to predict if and only if pair members share Gene Ontology annotation in "Molecular function" domain. The output is either 1 or 0, i.e. true or false, respectively.

Implementation idea

The predictions were done using the k-nearest neighbor algorithm. The prediction performance was calculated in C-inedx using two different implementations of cross-validation: modified and unmodified. Modified cross-validation excluded every instance sharing either one or both members with the test instance. Unmodified cross-validation excluded only instances that were either the test instance itself or its symmetrical relative.

The implementation structures in the following way:

1. Read the data and float the data
2. Prepare the data for easier cross-validation implementation (check the description of the dataset)
3. Run the knn algorithm
    a. Determine the test instance
    b. Exclude the necessary instances from the training data
    c. Calculate the euclidean distances
    d. Get the 5-nearest neighbors using the reduced training data
    e. Calculate the prediction by counting the majority vote of the nearest neighbors
    f. Calculate the c-index for each prediction
    g. Print the C-indexes for each k-value and plot the graphs for modified and unmodified cross-validation

Description of the dataset

The dataset included two files: proteins.features and proteins.labels. Proteins.features had 400 rows where each row represented a cell in 20x20 matrix. Each cell represented a protein that had 41 features.

Proteins.labels consisted of labels where 1 meant that the pair members shared Gene Ontology annotation in "Molecular function" domain and 0 that it didn't.

The data was prepared so that each row in proteins.features were appended with the xy position in the 20x20 matrix. This made it possible to easily separate the training and test sets.

Kasperi Keski-Loppi

Results

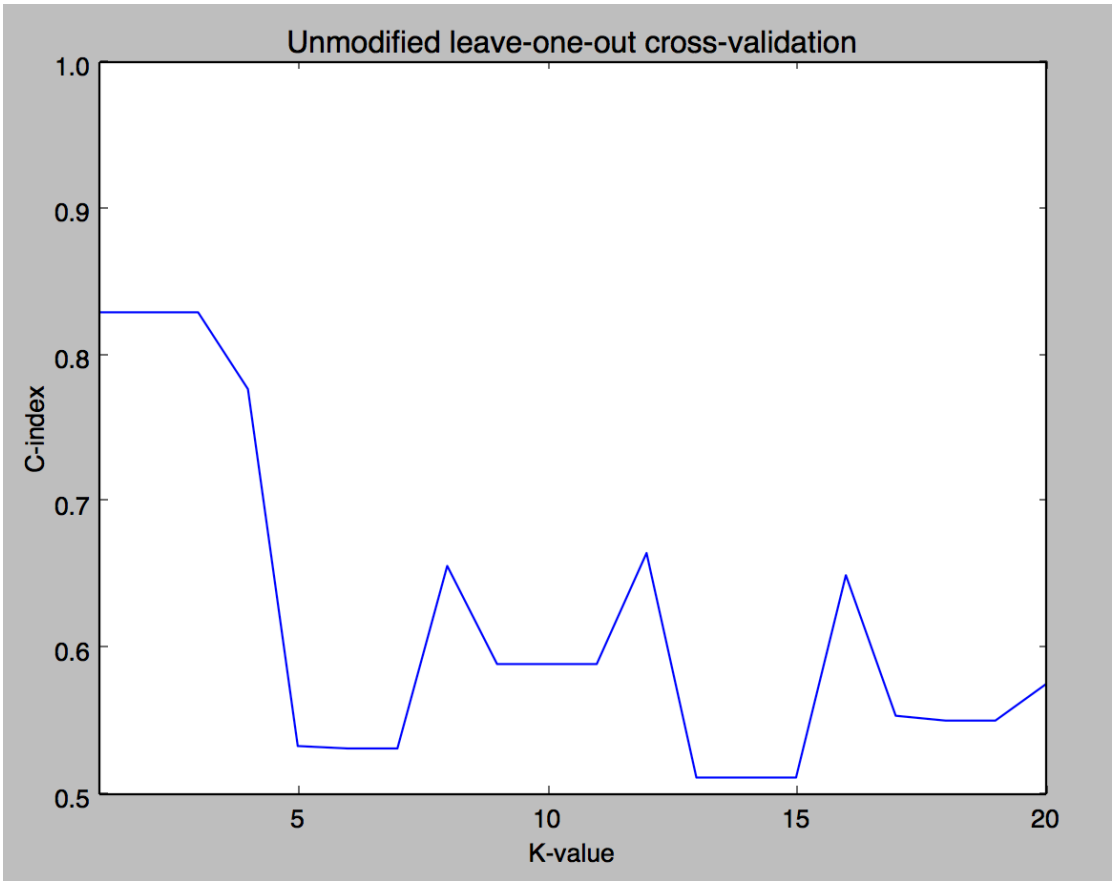

Figure 1. C-index for unmodified leave-one-out cross-validation with K running from 1 to 20.

Kasperi Keski-Loppi

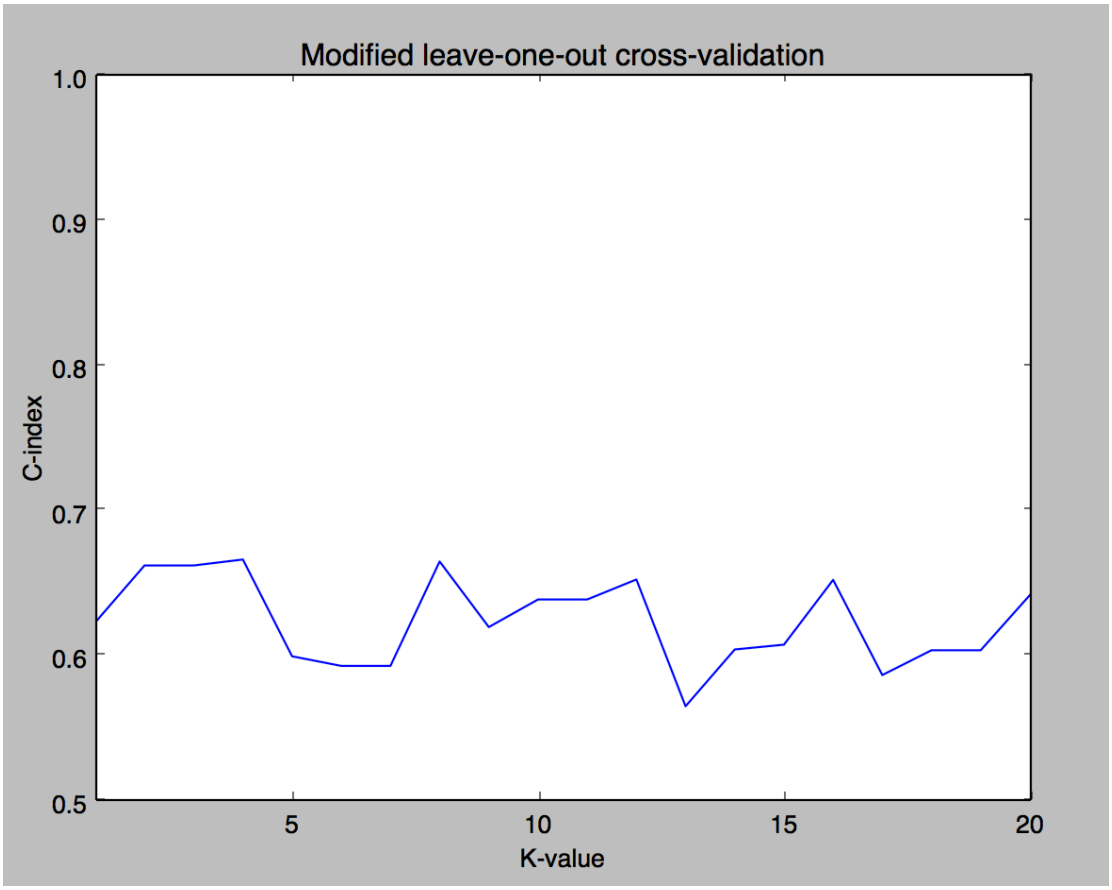

Figure 2. C-index for modified leave-one-out cross-validation with K running from 1 to 20.