Johannes Gutenberg-Universität Mainz
Institut für Informatik
Data Mining
Prof. Dr. Stefan Kramer

Machine Learning - Sheet 4

28.05.2020
Deadline: 04.06.2020 - 18:00

Task 1: Cross-Validation *(12 Points)*

The goal of this exercise is to implement stratified k-cross-validation and then apply it to optimize the parameters of classifiers.

(a) (*1 point*) Implement the `stratification` method, which splits the entire dataset into a list of datasets, based on the value of the class attribute.

(b) (*1 point*) Implement the `trainCV` and `testCV` methods, which extract training and test subsets of the dataset for a specified fold index and write them into a separate file.

(c) (*2 points*) Implement `stratifiedCrossValidation`, which gets a dataset and the number of folds (k) as input parameters. Shuffles the dataset, calls the previous methods to create train and test files for each fold.

(d) (*2 points*) Implement `CV_learn` that gets the train and test files, number of folds, and a classifier. It trains a classifier on each fold, writes the prediction for the test data in a file, and returns the mean and standard deviation of the accuracy over k folds.

(e) (*1 point*) Run `CV_learn` on the car dataset with $k = 10$ and ID3 classifier. You can use `Id3Estimator` package for ID3 implementation.

(f) (*2 points*) Implement the `CVparameterSelection` method that gets a dataset, one of the `Id3Estimator` options, and the possible range of that parameter (e.g., `start:steps:end` for real-valued options or `{a,b,...}` for nominals). It runs the `stratifiedCrossValidation` method once, then calls `CV_learn` for each possible parameter value in the range. It returns the best parameter value as the output.

(g) (*2 points*) Implement the `OptimalID3` classifier that performs a parameter optimization on the tree depth of the ID3 classifier using `CVparameterSelection` and classifies the instances using a tree with best parameters. Dataset path and parameter range are given as input parameters.

(h) (*1 point*) Evaluate your `OptimalID3` classifier on the car dataset for $k = 10$ and all possible depth values. Plot the accuracy based on the parameter values. Discuss your results.

Task 2: McNemar's test *(6 Points)*

In this task, you are supposed to compare random forests with decision trees on the car and the diabetes datasets (`http://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/diabetes.arff`).

(a) (*2 points*) Use your `stratifiedCrossValidation` method with $k = 10$ to generate different tain and test subsets. Run `sklearn.ensemble.RandomForestClassifier` with 50 trees and `Id3Estimator` on them.

(b) (*2 points*) Implement `McNemarTest`, which compares the output of random forests with decision tree using McNemar's Test.

(c) (*2 points*) Get the results for both datasets. Are random forests performing significantly better than decision trees? Discuss your results.

Task 3: ROC Curve *(2 Points)*

Given the following predictions of a classifier and the true class, give the ROC curve for the classifier. Is the performance of the classifier good? Explain using the curve.

Class	Prediction
P	0.95
N	0.85
P	0.78
P	0.66
N	0.60
P	0.55
N	0.53
N	0.52
N	0.51
P	0.40

Table 1: Prediction of a classifier on new instances.