

Homework 2: Machine Learning 2D5362

Handed out: Tuesday, 21.11.00

Due: Tuesday, 5.12.00 : 13:30

Name:

1. Compare the learning curves (class `LearnCurve`) performance (training and test accuracy (class `CatTestResult`) of the following inducers on the “monk1” dataset.

- a. ID3: class `ID3Inducer`
- b. C4.5: class `C45Inducer` (external binary in `$MLCDIR/external/GNU`)
- c. C4.5rules: class `C45Rinducer` (external binary)
- d. Naïve Bayes : class `NaiveBayesInd`
- e. Nearest Neighbor (PEBLS) : class `PeblsInducer` (external binary)

Plot the results.

2. Compare the following accuracy estimation methods for ID3 on the monk1 training dataset

- a. Holdout method: class `HoldOut`
- b. N-fold cross-validation (for 5, 10 and 20 folds): class `CValidator`
- c. Bootstrap: class `Bootstrap`
- d. Accuracy on the monk1 test data set

Plot the results (accuracy estimates with error-bars).

3. Implement Bagging (class `BaggingInd`) and Boosting (class `BoosterInd`) on the ID3 inducer and the monk1 dataset. For more information on Bagging and Boosting look at the paper “Bagging, Boosting and C4.5” by Quinlan. How much do Bagging and Boosting improve the performance of the ID3 inducer?

4. Implement feature subset selection (class `FSSInducer`) for the ID3 inducer on the monk1 dataset. How much does FSS improve the performance over the test data set? Is FFS able to identify the relevant features (attribute1, attribute2 and attribute 5)?