

Homework 1, Machine Learning

Decision tree learning

1 Task

Your task in this homework is to implement decision tree learning algorithm. You can either:

1. Implement the model and learning algorithm by your self using your favorite programming language or
2. Use some toolkit or library and implement only the code to execute experiments.

You can do both of course if you like to.

2 Toolkits

For python the recommended toolkit is `scikit-learn`: <http://scikit-learn.org>. Look for examples in <http://scikit-learn.org/stable/modules/tree.html> and reference in <http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier>.

You can also use any other toolkit that is available for your favorite programming language.

3 Data

Use the wine dataset from <http://archive.ics.uci.edu/ml/datasets/Wine>. This dataset has 13 features and 3 classes. Class labels are in the first column in each row.

4 Experiments

Experiment with different options:

- cost or score function;
- maximum depth of the tree;

- minimum number of items needed for a split;
- minimum number of samples required in a leaf;
- any other options provided by the toolkit used or you think might prove useful.

If you are using a library tool then the work should contain more experiments in terms of varying different options and hyperparameters. If you are implementing learning algorithm by yourself then the experiments can be less thorough.

5 Evaluation

Evaluate the results of different experiments using k-fold cross-validation. Choose either 5 or 10 folds (or try with both). In `scikit-learn` cross-validation is implemented in http://scikit-learn.org/stable/modules/classes.html#module-sklearn.cross_validation. You can also make your own implementation.

6 Report

The homework submission must include the program code and a short write-up (preferably in \LaTeX). The write-up should include a self-contained description of the task and the solution:

- What is the problem;
- Short description of the used method;
- Details of how to use your program;
- Descriptions of the experiments;
- Results of the experiments.

Preferably include figures that plot the cross-validation accuracy describing the dependence on hyperparameter values. In python the module `matplotlib` can be used to generate plots. These can be saved as pdf and imported easily into \LaTeX .