Lab Course Machine Learning
Exercise Sheet 8

Prof. Dr. Dr. Lars Schmidt-Thieme, Hadi Samer Jomaa
Information Systems and Machine Learning Lab
University of Hildesheim

December 18th, 2017
Submission on December 25th, 2017 at 8:00 am, (on moodle, course code 3113)

Instructions

Please read the lab related instructions, i.e. submission, report format and policies, at `https://www.ismll.uni-hildesheim.de/lehre/prakAIML-16w/exercises/ml_lab_instructions.pdf`

Datasets

1. **Classification Datasets:** You can use one of the two datasets (or optionally, both datasets).
   (a) Car Evaluation dataset D_1: Target attribute **safety**:{low, med, high}. `https://archive.ics.uci.edu/ml/datasets/Car+Evaluation`
   (b) Iris dataset D_2: Target attribute **class**:{Iris Setosa, Iris Versicolour, Iris Virginica}. `https://archive.ics.uci.edu/ml/datasets/Iris`

Exercise 1: Implement Decision Tree (12 Points)

In this task you will implement a decision tree. As a starting point you can implement your complete decision tree model for classification tasks. In particular you have to implement *Learn-Decision-Tree* with an appropriate *Quality-criterion* and *Predict-Decision-Tree*.

Part A: (8 Points): Basic working and Cross Entropy: In Part A, you have to split data into two parts train and test (70% and 30% respectively). Using the train data you will build a decision tree. Use **Cross Entropy** as a *Quality-criterion*. You have to provide information on the learning process that includes.

1. Define an appropriate stopping criteria i.e. max depth, gain is too small or reduction in cost is small
2. At each decision step (or split) present the probability of each class using histogram (properly labeled figure)
3. At each decision step, plot the **Cross Entropy** of each attribute.
4. Note down the ***Information Gain*** at each new node created, you can store it in node structure or class. Display it at the end.
5. Print your tree using a breath first tree traversal. (you can also print node hierarchical level, information gain and decision rule, etc)
6. On a test set measure the cross entropy loss (i.e. logloss, note that this time problem is not binary classification).

Part B: (4 Points): Experiment with other *Quality-criterion*: In Part B, you will implement **Gini Index and Gini Gain** as a *Quality-criterion* and compare it with **Cross Entropy**.

1. Use the train and test split from Part A and implementation of Decision tree as well.
2. modify the *Quality-criterion* to **Gini Index**.
3. At each decision step, plot the **Gini Index** and **Cross Entropy** of each attribute. [Hint: reuse **Cross Entropy** values from Part A]
4. At each decision step, plot the **Information Gain** of each new node created.
5. On a test set measure the cross entropy loss. Compare the test results for both *Quality-criterion*

Exercise 2: Pruning The Decision Tree (8 Points)

The complexity of a decision tree depends on the height/depth of the tree. It also influence the over-fitting or under-fitting behavior. Generally, two rules are used to restrict the height/depth of a decision tree, i.e. pre-pruning and post-pruning. In pre-pruning an early stopping criteria is used, whereas in post-pruning a validation dataset is used (through cross validation) to prune the tree. In this task you have to implement post-pruning technique known as ***Reduced Error Pruning***. Create a validation protocol i.e. split data into three splits, i.e. Train(34%), Validation(33%) and Test(33%). Build the decision tree (using implementation from Exercise 1). In particular you have to show during pruning process:

- You have to show for each pruning decision the gain at the parent node and collective gain of the children of that particular node. (Showing stats for the case at which pruning happens is enough.)

Finally using the test set measure the cross entropy loss (i.e. logloss). Did you see any gain in test results over results in Exercise 1?

Annex

1. Following lecture is relevant this exercise `https://www.ismll.uni-hildesheim.de/lehre/ml-16w/script/ml-07-A6-decision-trees.pdf`
2. A video tutorial on decision trees `https://www.youtube.com/watch?v=-dCtJjlEEgM#t=3918.930301`
3. Decision trees explained: `http://www.saedsayad.com/decision_tree.htm`
4. Pruning Decision trees explained: `http://www.saedsayad.com/decision_tree.htm`
5. Tree pruning `https://www.ismll.uni-hildesheim.de/lehre/ml-08w/skript/decision_trees2.pdf`
6. Chapter 16.1 and 16.2: Machine Learning A Probabilistic Prospective K. P. Murphy
7. Decision Trees, Quality Criterion and Pruning: `http://www.ke.tu-darmstadt.de/lehre/archiv/ws0809/mldm/dt.pdf`