CS 445/545

Machine Learning

Spring, 2013

Homework 1: Decision Trees

Due Wednesday April 17, 2pm

In this homework assignment you will experiment with C4.5, an implementation of ID3 written by Ross Quinlan and available on the Internet.

You will be using the optdigits data set that was discussed in class.

The data and executable C4.5 (which runs on the PSU Linux Lab machines) can be downloaded from the course webpage (*Homework1CodeAndData.zip*). This writeup will have all the information you need to do the homework.

Optional: The source code and C4.5 tutorial can be downloaded from http://www2.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tutorial.html

Please address any questions on this homework to either MLSpring2013@cs.pdx.edu or mm@cs.pdx.edu.

Decision Trees

C4.5 is an implementation of the ID3 Decision Tree Induction algorithm written by Ross Quinlan.

For this homework, you need to perform the following steps:

Step 1: Run C4.5 on the optdigits data

C4.5 requires three files: "<base>.names", "<base>.data", and "<base>.test", where "<base>" is some base file name (e.g., "optdigits"). For this step, you will be using optdigits.data as the training set and optdigits.test as the test set.

To run C4.5, type

```
c4.5 -f optdigits -u
```

Create a table for your results (of this and following steps) that lists, for the pruned ("simplified") trees

– attribute at root of tree

– % accuracy on test set

– Number of type 1 errors (false positives) and type 2 errors (false negatives) for each digit

Step 2: Analyze the learned decision tree

A. Consider the first six lines of the "simplified" decision tree obtained in Step 1:

```
f37 <= 0 :
| f43 <= 4 :
| | f22 <= 7 : 5 (11.0/1.3)
| | f22 > 7 :
| | | f26 <= 3 : 9 (17.0/2.5)
| | | f26 > 3 : 5 (2.0/1.0)
```

(this is what I get).

Given the description of the features given in "optdigits.info", show on a 32x32 grid which pixels correspond to f37, f43, f22, and f26, and write a few sentences speculating on why and how these features are used to classify the digits 5 and 9 in the partial tree above.

B. For each digit, 0–9, use the confusion matrix returned by C4.5 to state which other digit it is most frequently confused with by the decision tree. Write a few sentences answering the following: Were these most frequent confusions what you would have expected? Why or why not?

Step 3: Smaller training set

Create a new training set, "small-optdigits.data" that contains half the number of examples as in the original training data. Make sure the new training data contains approximately balanced numbers of examples for each digit class. Use the same test set as in Experiment 1. (You need to copy "optdigits.test" to "small-optdigits.test" and "optdigits.names" to "small-optdigits.names".)

Run

```
c4.5 -f small-optdigits –u
```

and record the results in your table, as in Step 1. Repeat the analysis of Step 2 for the new results, and write a few sentences comparing the results to those for the larger training set.

Step 4: Noisy training set

Create a new training set, "noisy-optdigits.data" into which you introduce some noise. To do this, copy the examples from "opt-digits.data" and change approximately 5% of digits' class to a random class (0-9). (Again, copy "optdigits.test" to "noisy-optdigits.test" and "optdigits.names" to "noisy-optdigits.names".)

Run

```
c4.5 -f noisy-optdigits –u
```

and record the results in your table as in Step 1. Repeat the analysis of Step 2 for these new results, and write a few sentences comparing the results to those for the original (non-noisy) training set.

• Step 5: Summary

Write a few sentences summarizing all your results.

- **Step 6: Hand it in**

Email your completed, formatted report, in PDF format, to mm@cs.pdx.edu, with "Homework 1" in the subject line. No hard copies, please!

Late homework policy: Students must request and be granted an extension on any homework assigment *before* the assignment is due. Otherwise, 5% of the assignment grade will be subtracted for each day the homework is late.