

| Aug 30, 07 13:27 | README.hw1.f07.txt | Page 1/4 |
|--|--------------------|----------|
| <p>CS578 Fall 2007 Empirical Methods in Machine Learning and Data Mining Homework Assignment #1 Due: Tuesday September 18, 2006</p> <p>The goal of this assignment is to use the IND decision tree package to do simple experiments with decision trees. This will require installing IND on a Unix platform, and running experiments with two data sets included in the distribution.</p> <p>NOTE: Since you might have trouble installing the software, it is important to do Steps 1-5 ASAP to verify that the code works for you. If you have trouble installing the software ask other students for help or contact one of the TAs.</p> <p>STEP 1: Download cs578.hw1.tar to a Unix machine from the link on the course web page. If you don't have access to a unix machine you may wish to use CYGWIN under Windows as your unix environment.</p> <p>STEP 2: Execute "tar -xvf cs578.hw1.tar" to untar the file. This will create a directory called cs578.hw1.</p> <p>STEP 3: cd into the directory and read the README files. This text is README.hw1.f07.txt. After this, read README.install and README.setenv.</p> <p>STEP 4: Run the script install.script by executing "./install.script".</p> <p>The installation is not bulletproof. We tested it on several Unix environments, and modified the code to minimize problems, but it may still fail to compile in some environments.</p> <p>Note that it is normal to get warnings during compilations and while using the code. These warnings do not mean the software is broken. The warnings happen because we are using unmodified code from the 80's and early 90's on modern Unix environments and things have changed.</p> <p>If you get errors and it fails to compile, please do your best to debug and fix the problems yourself. If you can't get it working, ask other students for help, or contact one of the TAs.</p> <p>If things went well, you have installed IND and a set of simple unix utilities collectively called unixstat. Be sure to set your path variables each time you create a new session using the instructions in README.setenv.</p> <p>STEP 5: Execute "inddemo hypo 100 c4 123 more" in the subdirectoy ind/IND/Data/thyroid. You should get results that look like the following. (NOTE: You may get somewhat different results under different Unix environments because the random number generators may differ. Stick to one platform for all of your experiments so that you can reproduce results.)</p> <pre>tgen -e -ir -Pnull -sl,2 -Sfull hypo.attr hypo.bld hypo.treec tprune -fn hypo.attr hypo.treec tclass -e -sl hypo.attr hypo.tree hypo.tst Percentage accuracy for tree 1 = 97.2495 +/- 0.269899 Mean square error for tree 1 = 0.0460837 Expected accuracy for tree 1 = 98.9185</pre> | | |

Thursday August 30, 2007

Aug 30, 07 13:27

README.hw1.f07.txt

Page

Leaf count for tree 1 = 4, expected = 4.000000

+6+92+2+0 negative

TSH < 6.55: +0+87+0+0 negative

TSH >= 6.55: +6+5+2+0 compensated_hypothyroid

TSH_measured = f: +0+4+0+0 negative

TSH_measured = t: +6+1+2+0 compensated_hypothyroid

FTI < 64.5: +0+0+2+0 primary_hypothyroid

FTI >= 64.5: +6+1+0+0 compensated_hypothyroid

| | | | | |
|-------------------------|--------|--------|--------|--------|
| primary_hypothyroid | 0.0000 | 0.0000 | 1.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| compensated_hypothyroid | 0.8571 | 0.1429 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| compensated_hypothyroid | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.0000 | 1.0000 | 0.0000 | 0.0000 |
| negative | 0.000 | | | |

README.hw1.f07.txt

| Aug 30, 07 13:27 | README.hw1.f07.txt | Page 3/4 |
|------------------|--|----------|
| | <p>using the description of the attributes in the glass.attr file and the training cases in glass.bld</p> <ul style="list-style-type: none"> - the line that calls tprint to show the tree and the counts at each node in the tree - the line that calls tclass on the test cases in glass.tst and then massages the output so you can see the predictions side-by-side with the true target values taken from glass.tst with the colex command <p>Note that unixstat utilities such as linex, colex, perm, dm, and stats are incredibly useful things to have around, so you may want to get familiar with these and keep them long after this assignment is finished.</p> <p>If you set up the MANPATH variable, you can learn more about mktree, tprint, tclass, colex, and linex by executing "man mktree" or "man colex". Other programs you might be interested in are tgen, tprune, perm, stats, and dm.</p> <p>STEP 7: Use IND to grow decision trees of types cart, c4, id3, smml, mml, and bayes on different sized training samples using the hwl.dta data set in the ind/IND/Data/hwl subdirectory. There is a sample .attr file in the thyroid directory, but you have to make an hwl.attr file yourself for the probl data set. There is a description of the attributes in the hwl subdirectory. You can run "man attr" to learn more about attribute files, but it's probably easier to look at the example hypo.attr file. Run multiple experiments at each size using different random seeds to get the average performance for each training set size. Report the accuracy of the tree on both the train and test sets, the RMSE (root mean squared error) on both sets, and the tree size. NOTE: Do not use any performance numbers IND reports. Calculate performances on the train and test sets yourself. Calculate both means and variances. Graphs would be a good way to present the results. Briefly comment on or explain the results. Write small scripts or programs to run the experiments and process the results. Include this code in what you hand in. Also include the .attr file you build for hwl.</p> <p>STEP 8: Which tree type(s) are more accurate, larger, more intelligible? Are trees with better RMSE always more accurate?</p> <p>EXTRA CREDIT: do one or more of the following experiments. All of these can be accomplished using different options with tgen or mktree and do not require modifying the IND code. See the man pages (e.g., "man tgen") for a description of the various options.</p> <ul style="list-style-type: none"> - manually control the depth of the decision with tgen -d and see how this affects accuracy for different size train sets - experiment with missing values by replacing a fraction of the case values with missing values (i.e. replacing attribute values with question marks: "?") and using the tgen -U option. Do missing values hurt performance? how much does performance change as the number of missing values is increased? - do experiments to determine how cpu time varies with the size of the training set? explain the results. - use a search engine to find the UC Irvine Machine Learning | |

| Aug 30, 07 13:27 | README.hw1.f07.txt | Page |
|------------------|---|------|
| | <p>Repository. skim the data sets that are available, and pick one. copy the data set, create a stem.attr file for it, and run experiments similar to those in Step 7 with it. For variety, you might want to select a data set which is large, or one which has attribute types different from the ones in the glass and thyroid data sets. Beware of data sets with missing values.</p> <ul style="list-style-type: none"> - experiment with lookahead. does lookahead make the trees more accurate? does lookahead make smaller trees more accurate? - pick something else that looks fun. be creative. <p>Hand in a *brief* summary of your results (5 pages max for results and discussion; attached code can be any length; 1-2 additional pages for each extra credit problem). Give us enough supporting documentation so that we can see what you did and how you did it. Do not write a paper or long report -- this is homework, not a class project. Our goal is to get you to experiment with decision trees, not give you a writing assignment. We will be more impressed with a clear, succinct summary than with a long, rambling report that lists all results in big unintelligible tables. Choose your graphs and tables carefully, and be sure to interpret/explain the results.</p> <p>You'll probably use IND later in the class, so effort spent now to become familiar with it should pay off later. You are allowed to get help from other students installing the software and learning how to use it, but should run the experiments and write supporting code yourself.</p> <p>There is a tutorial for IND in ind/IND/Doc in the files that look like ind0-15.ps. You don't need to read it, but it is there if you want it.</p> <p>Have fun!</p> | |