Angelica Contero

AI-HW#3 Report  
Professor Yuan

Due 11/4/13

**DTL Algorithm for Decision Tree Learning on UCI data set**

Below is a summary table of my results along with a plot. The raw data for each individual run is also included in this report.

From my results it would appear that the greater the percentage of data the tree is able to learn from, the more accurate the predictions on the testing data. This can be clearly seen by the graph shown below. I also noticed that for the 20% learning, results ranged from 0.0 to numbers similar to 80%, sometimes around 0.70. I think this suggests that if the small percentage of data the tree is allowed to learn on is varied and informative enough, it can predict test results very well. However, there is a strong likelihood that the small percentage of data will not be representative, hence the 0.0 results.

Please note that in my raw data you will see that I ran a total of 20 trials. The first 10 trials were done while learning on what I assigned as the testingData initially. And the testing was done on what I assigned as the trainingData initially. I think because of the way I assigned the two data sets, the original training data ended up being very ordered and I take the first 20%, 40% etc. of this ordered set which is not good for the results, I think. So I switched it around. However, I also recorded data for the last 10 trials where I used the two data sets as originally intended. The trees learn on the training set and the tests are performed on the testing set. The numbers seem a little off, so I did not plot the last 10 trials, only the first.

I’m not sure if I explained that clearly. Anyhow, it really shouldn’t matter which set is used for which as long as the sets are used for different purposes (the tree does not learn on the testing set). However, in the future I will be sure to better randomize the order of the set to ensure cleaner results.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run# | 20% | 40% | 60% | 80% |
| 1 | 0.244212963 | 0.226851852 | 0.696759259 | 0.690972222 |
| 2 | 0 | 0.689814815 | 0.688657407 | 0.688657407 |
| 3 | 0 | 0 | 0 | 0.693287037 |
| 4 | 0.538194444 | 0.699074074 | 0.699074074 | 0.699074074 |
| 5 | 0.395833333 | 0.703703704 | 0.703703704 | 0.709490741 |
| 6 | 0 | 0.716435185 | 0.712962963 | 0.716435185 |
| 7 | 0.541666667 | 0.038194444 | 0.038194444 | 0.038194444 |
| 8 | 0.292824074 | 0.703703704 | 0.703703704 | 0.703703704 |
| 9 | 0 | 0.707175926 | 0.703703704 | 0.703703704 |
| 10 | 0 | 0 | 0.702546296 | 0.703703704 |
| AVERAGE | 0.201273148 | 0.44849537 | 0.564930556 | 0.634722222 |

RAW DATA FOR INDIVIDUAL TRIAL RUNS

//1, then 0

//training was done on testingData and testing on trainingData

For Example, in my main method:

*examples20* = FileReader.*buildRecords*(173,1);

*examples40* = FileReader.*buildRecords*(346,1);

*examples60* = FileReader.*buildRecords*(518,1);

*examples80* = FileReader.*buildRecords*(691,1);

*examplesTest* = FileReader.*buildRecords*(864,0);

Run #1:

Total Count for 20% = 211

Total Count for 40% = 196

Total Count for 60% = 602

Total Count for 80% = 597

Accuracy for 20% = 0.24421296296296297

Accuracy for 40% = 0.22685185185185186

Accuracy for 60% = 0.6967592592592593

Accuracy for 80% = 0.6909722222222222

Run #2:

Total Count for 20% = 0

Total Count for 40% = 596

Total Count for 60% = 595

Total Count for 80% = 595

Accuracy for 20% = 0.0

Accuracy for 40% = 0.6898148148148148

Accuracy for 60% = 0.6886574074074074

Accuracy for 80% = 0.6886574074074074

Run #3:

Total Count for 20% = 0

Total Count for 40% = 0

Total Count for 60% = 0

Total Count for 80% = 599

Accuracy for 20% = 0.0

Accuracy for 40% = 0.0

Accuracy for 60% = 0.0

Accuracy for 80% = 0.6932870370370371

Run #4:

Total Count for 20% = 465

Total Count for 40% = 604

Total Count for 60% = 604

Total Count for 80% = 604

Accuracy for 20% = 0.5381944444444444

Accuracy for 40% = 0.6990740740740741

Accuracy for 60% = 0.6990740740740741

Accuracy for 80% = 0.6990740740740741

Run #5:

Total Count for 20% = 342

Total Count for 40% = 608

Total Count for 60% = 608

Total Count for 80% = 613

Accuracy for 20% = 0.3958333333333333

Accuracy for 40% = 0.7037037037037037

Accuracy for 60% = 0.7037037037037037

Accuracy for 80% = 0.7094907407407407

Run #6:

Total Count for 20% = 0

Total Count for 40% = 619

Total Count for 60% = 619

Total Count for 80% = 619

Accuracy for 20% = 0.0

Accuracy for 40% = 0.7164351851851852

Accuracy for 60% = 0.7129629629629629

Accuracy for 80% = 0.7164351851851852

Run #7:

Total Count for 20% = 468

Total Count for 40% = 33

Total Count for 60% = 33

Total Count for 80% = 33

Accuracy for 20% = 0.5416666666666666

Accuracy for 40% = 0.03819444444444445

Accuracy for 60% = 0.03819444444444445

Accuracy for 80% = 0.03819444444444445

Run #8:

Total Count for 20% = 253

Total Count for 40% = 608

Total Count for 60% = 608

Total Count for 80% = 608

Accuracy for 20% = 0.29282407407407407

Accuracy for 40% = 0.7037037037037037

Accuracy for 60% = 0.7037037037037037

Accuracy for 80% = 0.7037037037037037

Run #9:

Total Count for 20% = 0

Total Count for 40% = 611

Total Count for 60% = 608

Total Count for 80% = 608

Accuracy for 20% = 0.0

Accuracy for 40% = 0.7071759259259259

Accuracy for 60% = 0.7037037037037037

Accuracy for 80% = 0.7037037037037037

Run #10:

Total Count for 20% = 0

Total Count for 40% = 0

Total Count for 60% = 607

Total Count for 80% = 608

Accuracy for 20% = 0.0

Accuracy for 40% = 0.0

Accuracy for 60% = 0.7025462962962963

Accuracy for 80% = 0.7037037037037037

//0, then 1

//training was done on trainingData and testing on testingData

For example, in my main method:

*examples20* = FileReader.*buildRecords*(173,0);

*examples40* = FileReader.*buildRecords*(346,0);

*examples60* = FileReader.*buildRecords*(518,0);

*examples80* = FileReader.*buildRecords*(691,0);

*examplesTest* = FileReader.*buildRecords*(864,1);

Run #1:

Total Count for 20% = 0

Total Count for 40% = 607

Total Count for 60% = 602

Total Count for 80% = 606

Accuracy for 20% = 0.0

Accuracy for 40% = 0.7025462962962963

Accuracy for 60% = 0.6967592592592593

Accuracy for 80% = 0.7013888888888888

Run #2:

Total Count for 20% = 0

Total Count for 40% = 605

Total Count for 60% = 601

Total Count for 80% = 601

Accuracy for 20% = 0.0

Accuracy for 40% = 0.7002314814814815

Accuracy for 60% = 0.6956018518518519

Accuracy for 80% = 0.6956018518518519

Run #3:

Total Count for 20% = 593

Total Count for 40% = 609

Total Count for 60% = 602

Total Count for 80% = 602

Accuracy for 20% = 0.6863425925925926

Accuracy for 40% = 0.7048611111111112

Accuracy for 60% = 0.6967592592592593

Accuracy for 80% = 0.6967592592592593

Run #4:

Total Count for 20% = 593

Total Count for 40% = 592

Total Count for 60% = 592

Total Count for 80% = 590

Accuracy for 20% = 0.6863425925925926

Accuracy for 40% = 0.6851851851851852

Accuracy for 60% = 0.6851851851851852

Accuracy for 80% = 0.6828703703703703

Run #5:

Total Count for 20% = 595

Total Count for 40% = 607

Total Count for 60% = 601

Total Count for 80% = 601

Accuracy for 20% = 0.6886574074074074

Accuracy for 40% = 0.7025462962962963

Accuracy for 60% = 0.6956018518518519

Accuracy for 80% = 0.6956018518518519

Run #6:

Total Count for 20% = 594

Total Count for 40% = 602

Total Count for 60% = 602

Total Count for 80% = 602

Accuracy for 20% = 0.6875

Accuracy for 40% = 0.6967592592592593

Accuracy for 60% = 0.6967592592592593

Accuracy for 80% = 0.6967592592592593

Run #7:

Total Count for 20% = 625

Total Count for 40% = 626

Total Count for 60% = 627

Total Count for 80% = 627

Accuracy for 20% = 0.7233796296296297

Accuracy for 40% = 0.7245370370370371

Accuracy for 60% = 0.7256944444444444

Accuracy for 80% = 0.7256944444444444

Run #8:

Total Count for 20% = 0

Total Count for 40% = 0

Total Count for 60% = 602

Total Count for 80% = 603

Accuracy for 20% = 0.0

Accuracy for 40% = 0.0

Accuracy for 60% = 0.6967592592592593

Accuracy for 80% = 0.6979166666666666

Run #9:

Total Count for 20% = 612

Total Count for 40% = 605

Total Count for 60% = 605

Total Count for 80% = 605

Accuracy for 20% = 0.7083333333333334

Accuracy for 40% = 0.7002314814814815

Accuracy for 60% = 0.7002314814814815

Accuracy for 80% = 0.7002314814814815

Run #10:

Total Count for 20% = 598

Total Count for 40% = 616

Total Count for 60% = 604

Total Count for 80% = 604

Accuracy for 20% = 0.6921296296296297

Accuracy for 40% = 0.7129629629629629

Accuracy for 60% = 0.6990740740740741

Accuracy for 80% = 0.6990740740740741