Decision Trees

Abrie Greeff
B.Sc Hons (Computer Science)
Department of Computer Science
University of Stellenbosch

June 26, 2006

Contents

1 Question 1

This question was not answered.

2 Question 2

2.1 Part a

The decision tree application was developed in Java 1.5.0. To execute the application type *java dtree* to see all parameters that can be passed to the program. The application accepts a training file to build a decision tree based on the values found in the training file. Relative information gain was used as the criteria for splitting the tree for every branch. The entropy of a set is defined as a value which represents the distribution of the data. A high entropy means the data is disordered and represented by peaks and valleys. A low entropy means the data is almost a normal distribution. Relative information gain enables us to find the split which will allow the entropy to decrease.

In my application real numbers are split according to the mean value of all the numbers. Other types of categories are split according to the possible values they can obtain. The application starts off with a root node and all the data available. From here all the branching is done by the splitting criteria until a decision can be reached. It is possible to have a holdout set which is used to prune the tree for any errors. This process, reduced error pruning, was added to my application, but because of time constraints was not completed.

When the decision tree has been built the tree can be saved to an output file. The syntax that the tree is saved in is a pre-order traversal method. Which means that it always branches to the left first.

2.2 Part b

This question was not implemented.

3 Question 3

For this question I developed an application which reads a decision tree, that is saved to a file in my chosen syntax, and saves a graphical representation of the tree to a image file. To execute this application type in *java writeImage tree image*, where *tree* is the name of file containing the decision tree and *image* is the name of the jpeg file the image should be saved to.

4 Question 4

4.1 Part a

I obtained the adult data set.

4.2 Part b

The first 28000 points of the adult data set was kept and the rest discarded because I did not implement error pruning.

4.3 Part c

The tree in Fig. 1 was generated The following output was generated with debug output on.

```
java dtree -d -o tree3 adult.data

Reading header

age:num
workclass:cat
fnlwgt:num
education:cat
education-num:num
marital-status:cat
occupation:cat
relationship:cat
race:cat
sex:bool
capital-gain:num
capital-loss:num
hours-per-week:num
```

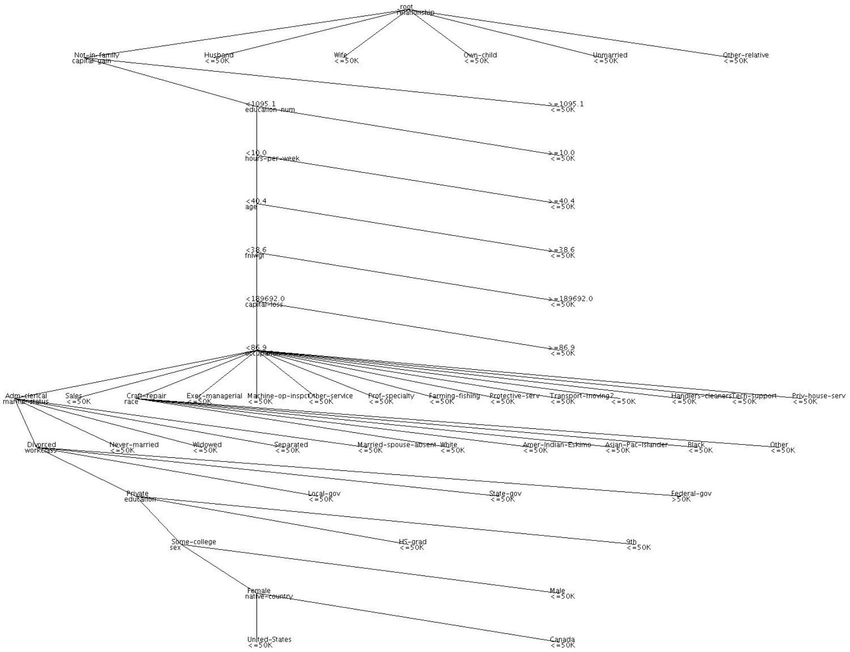

Figure 1: Decision Tree for 4(c)

```
native-country:cat
income:bool

Loading data into hashtable


Building Tree

Entropy: 0.7943268534134382
Info gain for age:num is 0.05247362524336983
Info gain for workclass:cat is 0.027344530055538939
Info gain for fnlwgt:num is 3.032897685141165E-4
Info gain for education:cat is 0.1164788009747424
Info gain for education-num:num is 0.08148483913009445
Info gain for marital-status:cat is 0.19650919674864364
Info gain for occupation:cat is 0.11557295572137143
Info gain for relationship:cat is 0.2083920482581514
Info gain for race:cat is 0.01049173211433773
Info gain for sex:bool is 0.04670461704412059
Info gain for capital-gain:num is 0.05671045769582225
Info gain for capital-loss:num is 0.014357141745399492
Info gain for hours-per-week:num is 0.04955780358233868
Info gain for native-country:cat is 0.011024202913488783
Splitting node on attribute relationship:cat with info gain:0.2083920482581514
Splitting with 6 children
Entropy: 0.47962271851799415
Info gain for age:num is 0.7492764037977332
Info gain for workclass:cat is 0.015264836254411463
Info gain for fnlwgt:num is 0.743409595097107
Info gain for education:cat is 0.12478700515334498
Info gain for education-num:num is 0.7636979686310653
Info gain for marital-status:cat is 0.005100587022183096
Info gain for occupation:cat is 0.10774057785835975
Info gain for race:cat is 0.0027534857388923917
Info gain for sex:bool is 0.01446685721943875
Info gain for capital-gain:num is 0.7665319910075196
Info gain for capital-loss:num is 0.7444295071971625
Info gain for hours-per-week:num is 0.7589891145741615
Info gain for native-country:cat is 0.012566997076712562
Splitting node on attribute capital-gain:num with info gain:0.7665319910075196
Entropy: 0.39757061415091116
Info gain for age:num is 0.7659546420717458
Info gain for workclass:cat is 0.015351719170459164
Info gain for fnlwgt:num is 0.7599852261619879
Info gain for education:cat is 0.12033155636461845
Info gain for education-num:num is 0.7775509808550003
Info gain for marital-status:cat is 0.005743357680761822
Info gain for occupation:cat is 0.11136575715921694
Info gain for race:cat is 0.004459308927860536
Info gain for sex:bool is 0.01695555295040178
Info gain for capital-loss:num is 0.7624899758642233
Info gain for hours-per-week:num is 0.7752813666502336
Info gain for native-country:cat is 0.014407961771521314
Splitting node on attribute education-num:num with info gain:0.7775509808550003
Entropy: 0.2284975997472975
Info gain for age:num is 0.85049507903347
Info gain for workclass:cat is 0.019617895676331955
Info gain for fnlwgt:num is 0.8470045620186168
Info gain for education:cat is 0.018218564697578373
Info gain for marital-status:cat is 0.019198934599617187
Info gain for occupation:cat is 0.07870269406586512
Info gain for race:cat is 0.011303787912002301
Info gain for sex:bool is 0.02427325769901812
Info gain for capital-loss:num is 0.8477007480893093
Info gain for hours-per-week:num is 0.8552661998183919
Info gain for native-country:cat is 0.02266067773014697
Splitting node on attribute hours-per-week:num with info gain:0.8552661998183919
Entropy: 0.14984776646628148
Info gain for age:num is 0.8880657176818849
Info gain for workclass:cat is 0.019912384114547956
Info gain for fnlwgt:num is 0.8851477739808735
Info gain for education:cat is 0.03479453572003036
Info gain for marital-status:cat is 0.024358472345868208
Info gain for occupation:cat is 0.07695279865666019
Info gain for race:cat is 0.014864790133519008
Info gain for sex:bool is 0.012864409139845105
Info gain for capital-loss:num is 0.884195719518572
Info gain for native-country:cat is 0.03978239725654674
Splitting node on attribute age:num with info gain:0.8880657176818849
Entropy: 0.08146202691506
Info gain for workclass:cat is 0.0348251142668829
Info gain for fnlwgt:num is 0.9374324322414234
Info gain for education:cat is 0.06987794423279345
Info gain for marital-status:cat is 0.03577300763865465
Info gain for occupation:cat is 0.09664832229156121
Info gain for race:cat is 0.032077185576033417
Info gain for sex:bool is 0.010037159131425695
Info gain for capital-loss:num is 0.9371154423614224
Info gain for native-country:cat is 0.07921440341269431
Splitting node on attribute fnlwgt:num with info gain:0.9374324322414234
Entropy: 0.10811918331462697
Info gain for workclass:cat is 0.06910093877943092
Info gain for education:cat is 0.08942071758080196
Info gain for marital-status:cat is 0.05951556048931543
Info gain for occupation:cat is 0.16106778667134689
Info gain for race:cat is 0.04775929340384572
Info gain for sex:bool is 0.0029230576587667275
Info gain for capital-loss:num is 0.9682976987277503
Info gain for native-country:cat is 0.07369050637741627
Splitting node on attribute capital-loss:num with info gain:0.9682976987277503
Entropy: 0.09712405133679194
Info gain for workclass:cat is 0.08280837635121714
Info gain for education:cat is 0.07814532975590899
Info gain for marital-status:cat is 0.0858450844730557
Info gain for occupation:cat is 0.17452764547513847
Info gain for race:cat is 0.06211610357386418
Info gain for sex:bool is 1.0919776234364733E-4
Info gain for native-country:cat is 0.08687311388372618
Splitting node on attribute occupation:cat with info gain:0.17452764547513847
Splitting with 14 children
Entropy: 0.1112410494829923
Info gain for workclass:cat is 0.26341045064619434
Info gain for education:cat is 0.13607245512602428
Info gain for marital-status:cat is 0.34721279156628326
Info gain for race:cat is 0.02120212551452605
Info gain for sex:bool is 0.029599426302886405
Info gain for native-country:cat is 0.014901923293116442
Splitting node on attribute marital-status:cat with info gain:0.34721279156628326
Splitting with 5 children
Entropy: 0.42622865699814483
Info gain for workclass:cat is 0.4316398595974942
Info gain for education:cat is 0.17864371025100662
Info gain for race:cat is 0.0280739277536667
Info gain for sex:bool is 0.1346932128927267
Info gain for native-country:cat is 0.01370211652713505
Splitting node on attribute workclass:cat with info gain:0.4316398595974942
Splitting with 4 children
Entropy: 0.3095434291503252
Info gain for education:cat is 0.13233017820850285
Info gain for race:cat is 0.01523990487164407
Info gain for sex:bool is 0.01523990487164407
Info gain for native-country:cat is 0.01523990487164407
Splitting node on attribute education:cat with info gain:0.13233017820850285
Splitting with 3 children
Entropy: 0.4394969869215134
Info gain for race:cat is 0.0
Info gain for sex:bool is 0.029891800801240345
Info gain for native-country:cat is 0.029891800801240345
Splitting node on attribute sex:bool with info gain:0.029891800801240345
Splitting with 2 children
Entropy: 0.4689955935892812
Info gain for race:cat is 0.0
Info gain for native-country:cat is 0.034249985523887604
Splitting node on attribute native-country:cat with info gain:0.034249985523887604
Splitting with 2 children
Entropy: 0.5032583347756459
Info gain for race:cat is 0.0
Splitting node on attribute age:num with info gain:0.0
Entropy: 0.0
Entropy: 0.0
Entropy: 0.0
Entropy: 0.0
Entropy: 0.0
```

```
Entropy: 0.0
Entropy: 0.0
Entropy: 0.0
Entropy: 0.0
Entropy: 0.0
Entropy: 0.0
Entropy: 0.0
Entropy: 0.07253718299881004
Info gain for race:cat is 0.017785124549207148
Splitting node on attribute race:cat with info gain:0.017785124549207148
Splitting with 5 children
Entropy: 0.07885601377455284
Splitting node on attribute age:num with info gain:0.0
Entropy: 0.0
Entropy: 0.0
Entropy: 0.0
Entropy: 0.0
Entropy: 0.35335933502142136
Splitting node on attribute age:num with info gain:0.0
Entropy: 0.0
Entropy: 0.050080721052405935
Splitting node on attribute age:num with info gain:0.0
Entropy: 0.36505518964028494
Splitting node on attribute age:num with info gain:0.0
Entropy: 0.0
Entropy: 0.0
Entropy: 0.1831220683013728
Splitting node on attribute age:num with info gain:0.0
Entropy: 0.0
Entropy: 0.0
Entropy: 0.0
Entropy: 0.0
Entropy: 0.345117314944953
Splitting node on attribute age:num with info gain:0.0
Entropy: 0.05089613660052264
Splitting node on attribute age:num with info gain:0.0
Entropy: 0.22043582169557271
Splitting node on attribute age:num with info gain:0.0
Entropy: 0.4215312445626803
Splitting node on attribute age:num with info gain:0.0
Entropy: 0.6157861264950467
Splitting node on attribute age:num with info gain:0.0
Entropy: 0.9954280519491803
Splitting node on attribute age:num with info gain:0.0
```

```
Entropy: 0.9918769754695131
Splitting node on attribute age:num with info gain:0.0
Entropy: 0.9988076767835599
Splitting node on attribute age:num with info gain:0.0
Entropy: 0.10310297038359535
Splitting node on attribute age:num with info gain:0.0
Entropy: 0.32585663525820097
Splitting node on attribute age:num with info gain:0.0
Entropy: 0.22391522789457574
Splitting node on attribute age:num with info gain:0.0

Saving tree to tree3
```