Aniqa Butt

How to run the code?

- Preprocess the training and testing data:
  Execute the code in the file preprocessing.r on the training and testing data to apply the preprocessing steps.

- Run the training and classification:
  python id3.py data/adult.train.preprocessed.txt --testing_file data/adult.test.preprocessed.txt > data/classification.txt

Exploration and preprocessing

- The first step I will take is deleting two variables: fnlwgt, and education_num. The reason for this is they clutter the analysis and does not provide much distinction between the 2 classes.
- Plotting the "Employer type" variable gives us the following:

?	1836
Federal-gov	960
Local-gov	2093
Never-worked	7
Private	22696
Self-emp-inc	1116
Self-emp-not-inc	2541
State-gov	1298
Without-pay	14

Looking through the numbers, we can easily observe that the values "Never-Worked" and "Without-Pay" are very small in number, then we can assume that combining these variables will not much affect our decision tree. Same stands for combining "Federal" and "Local" government employee categories and self-employed which are generally the same.

- Plotting the "Occupation" variable gives us the following:

?	1843
Adm-clerical	3770
Armed-forces	9
Craft-repair	4099
Exec-managerial	4066
Farming-fishing	994

Handlers-cleaners	1370
Machine-op-inspct	2002
Other-service	3295
Priv-house-serv	149
Prof-specialty	4140
Protective-srv	649
Sales	3650
Tech-support	928
Transport-moving	1597

For this variable, one way to reduce the number of options is to create 2 new categories "Blue Collar" and "White Collar" and put the relevant occupations under. This will help us reduce the different values for this variable and therefore could improve the quality of the tree produced.

- Plotting the "Country" variable:

?	583
Cambodia	19
Canada	121
China	75
Columbia	59
Cuba	95
Dominican-Republic	70
Ecuador	28
El-Salvador	106
England	90
France	29
Germany	137
Greece	29
Guatemala	64
Haiti	44
Holand-Netherlands	1
Honduras	13
Hong	20
Hungary	13
India	100
Iran	43
Ireland	24
Italy	73
Jamaica	81
Japan	62
Laos	18
Mexico	643
Nicaragua	34

Outlying-US(Guam-USVI-etc)	14
Peru	13
Philippines	198
Poland	60
Portugal	37
Puerto-Rico	114
Scotland	12
South	80
Taiwan	51
Thailand	18
Trinadad&Tobago	19
United-States	29170
Vietnam	67
Yugoslavia	16

It's obvious that the United States is very dominant in terms of the number of instances, so it will introduce a high class imbalance leading to noise in the tree. One way to overcome this is to combine the relevant countries with respect to their geographical location, political organization and economic zones.

An example of that is to create a value called "Euro_1"; where countries in this zone would be considered more affluent and then could indicate that the income of someone living there should be above 50K.

- Since we are using ID3, then we need to change continuous features into discrete ones. Which what exactly was done on the variables "capital gain", "capital loss", "age" and "hours per week":

```
data[["capital_gain"]] <- ordered(cut(data$capital_gain,c(-Inf, 0,
        median(data[["capital_gain"]][data[["capital_gain"]] >0]),
        Inf)),labels = c("None", "Low", "High"))
data[["capital_loss"]] <- ordered(cut(data$capital_loss,c(-Inf, 0,
        median(data[["capital_loss"]][data[["capital_loss"]] >0]),
        Inf)), labels = c("None", "Low", "High"))

data[["age"]] <- ordered(cut(data[[ "age"]], c(15,25,45,65,100)),
  labels = c("Young", "Middle", "Older", "Senior"))

data[[ "hr_per_week"]] <- ordered(cut(data[[ "hr_per_week"]], c(0,25,40,60,168)),
  labels = c("Part-time", "Full-time", "Over-time", "VeryHigh"))
```

Dealing with missing values

In the technique used, we get rid of the missing values as it does not represent high percentage of the total number of instances (couple of thousands in more than 30 thousands).

Looking at the first 3 levels of the tree

Observing the decision tree, we can tell that the most important attributes would definitely include the “Capital gain”, “Occupation”, “Education” and “Hours per week”.

Here’s the command to display the tree:

c:\python27\python id3.py data/adult.train.preprocessed.txt --rules > data/tree.txt

Here’s an excerpt of the tree:

```
[(('relationship', 'Husband'), ('education', 'Associates'), ('capital_gain', 'High'), '>50K'),
 (('relationship', 'Husband'), ('education', 'Doctorate'), ('capital_gain', 'High'), '>50K'),
 (('relationship', 'Husband'), ('education', 'Masters'), ('occupation', 'Military'), '>50K'),
 (('relationship', 'Husband'), ('education', 'Prof-School'), ('capital_gain', 'High'), '>50K'),
 (('relationship', 'Not-in-family'), ('capital_gain', 'High'), ('occupation', '?'), '>50K'),
 (('relationship', 'Not-in-family'), ('capital_gain', 'High'), ('occupation', 'Blue-Collar'), '>50K'),
 (('relationship', 'Not-in-family'), ('capital_gain', 'High'), ('occupation', 'Military'), '>50K'),
 (('relationship', 'Not-in-family'), ('capital_gain', 'High'), ('occupation', 'Other-Occupations'),
'>50K'),
 (('relationship', 'Not-in-family'), ('capital_gain', 'High'), ('occupation', 'Professional'), '>50K'),
 (('relationship', 'Not-in-family'), ('capital_gain', 'High'), ('occupation', 'Sales'), '>50K'),
 (('relationship', 'Not-in-family'), ('capital_gain', 'High'), ('occupation', 'Service'), '>50K'),
 (('relationship', 'Other-relative'), ('capital_gain', 'High'), ('hr_per_week', 'Full-time'), '>50K'),
 (('relationship', 'Other-relative'), ('capital_gain', 'High'), ('hr_per_week', 'Over-time'), '>50K'),
 (('relationship', 'Other-relative'), ('capital_gain', 'High'), ('hr_per_week', 'Part-time'), '<=50K'),
 (('relationship', 'Other-relative'), ('capital_gain', 'High'), ('hr_per_week', 'VeryHigh'), '>50K'),
 (('relationship', 'Other-relative'), ('capital_gain', 'Low'), ('education', 'Associates'), '<=50K'),
 (('relationship', 'Other-relative'), ('capital_gain', 'Low'), ('education', 'Bachelors'), '<=50K'),
 (('relationship', 'Other-relative'), ('capital_gain', 'Low'), ('education', 'Doctorate'), '<=50K'),
 (('relationship', 'Other-relative'), ('capital_gain', 'Low'), ('education', 'Dropout'), '<=50K'),
 (('relationship', 'Other-relative'), ('capital_gain', 'Low'), ('education', 'HS-grad'), '<=50K'),
 (('relationship', 'Other-relative'), ('capital_gain', 'Low'), ('education', 'Masters'), '>50K'),
```