

Assignment 1

Decision Trees

Due date: Thursday Aug 27 11:59pm

1. Training a Decision Tree (70%)

Write a program in Python to implement the ID3 decision tree algorithm. You should read in a tab delimited dataset, and output to the screen the relevant results in some readable format.

- Name your program *decisiontreeassignment.py*
- It should be written from scratch, no ML libraries to be used
- You may use existing libraries to load the data from file.

There are two sample datasets available from the course blackboard page you can use

- **tennis.txt** - Predict whether or not your tennis partner will join you to play tennis based on weather.
- **titanic2.txt** - Predict the survival status of individual passengers on the Titanic based on their passenger class, age and gender.

For the dataset files

- The first line of the file will contain the name of the fields.
- The last column is the classification attribute, and will always contain the values **yes** or **no**.
- All files are tab delimited.

When you run your program, it should take a command-line parameter that contains the name of the file containing the training data, the maximum tree depth and the name of the file containing the training data (if any). For example:

```
python decisiontreeassignment.py tennis.txt
```

And it should output the training set accuracy in some readable form.

2. Max Tree Depth (15%)

Add to your implementation so that you can limit the maximum tree depth. It should now take an additional command-line parameter that sets the maximum tree depth. For example:

```
python decisiontreeassignment.py tennis.txt 5
```

3. Test Set (15%)

Add to your implementation so that you can also pass a file containing data not in the training data. It should now output the training set accuracy as well as the testing set accuracy in some readable form.

The command-line call should now have a third parameter containing the name of the file containing the testing data. For example:

```
python decisiontreeassignment.py tennis_trainingset.txt 5 tennis_testset.txt
```

You can create training and testing sets by (randomly) splitting the available data appropriately.

Submission

Assignments to be completed individually and submitted through blackboard.

Due date

Thursday Aug 27 11:59pm.