# Assignment 4: Decision trees

CS 6601

Due November 3 by 9:35 AM

**Abstract**

You will build, train and test several decision tree models to perform basic classification tasks.

## 1 The Challenge

Machine learning offers a number of methods for classifying data into discrete categories, such as k-means clustering. Decision trees provide a structure for such categorization, based on a series of decisions that lead to separate distinct outcomes. Your challenge is to build and to train decision trees capable of solving useful classification problems.

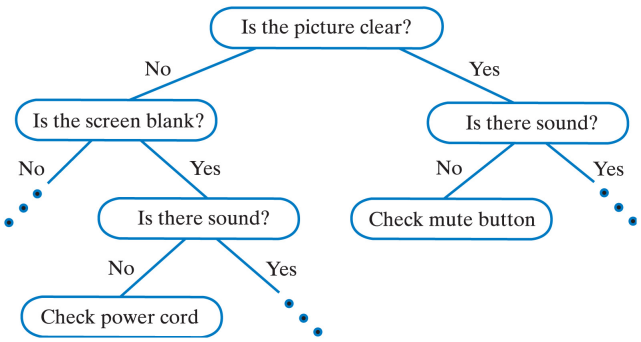


Figure 1: Example decision tree
(representing actions to perform when debugging TV)

## 2 Your Assignment

The structure of this assignment is as follows:

- In Part 1 of the assignment, you'll be given the task of building a decision tree by hand over a toy dataset to familiarize yourself with the way they work.
- In Part 2 of the assignment, you'll be implementing an algorithm to automatically build a decision tree form a list of examples (C4.5).

- In Part 3 of the assignment, you'll be using bagging techniques to build a RandomForest classifier that uses a series of DecisionTrees.
- In Part 4 of the assignment, you'll be competing for classification accuracy on a research dataset.

You will do this in decision_notebook.ipynb, following the instructions therein.

We will provide the following additional files:

| File | Description |
|---|---|
| part2_data.csv | Data to be used in Part2 and Part3 of the assignment |
| challenge_data.pickle | Data to be used to build the classifier you'd like to submit for the challenge portion of the assignment. |

# 3 Grading

Each section of the assignment is associated with a number of points, as follows (out of 100 points total):

Part 1: Build and test a decision tree over a small dataset. (20 points)

Part 2: Implement a variant of the C4.5 algorithm to build decision trees automatically. (40 points)

Part 3: Implement random forest classification to generalize your decision tree performance. (30 points)

Part 4: Improve on one of your classifiers using the challenge dataset (10 points)

# 4 Submission

This assignment is due on T-Square Tuesday November 3rd by the start of class (9:35 AM). The deliverables for the assignment are:

- A filled out version of the iPython notebook provided. (decision_tree_notebook.ipynb)

**Please** submit this in iPython notebook format - it makes grading much easier.

# 5 Resources

These slides provide a good introduction to decision trees and information theory. You can find an overview of the C4.5 algorithm here.

As always, TAs will hold office hours Monday, Tuesday, Thursday and Friday from 2:00 to 4:00 PM outside TSRB 241.