DATS 6202 **Machine Learning I** **Quiz 4**
Term 2018-Fall **October 31, 2018**

Quiz 4: Solutions

Full Name: _____

GWID: _____

- DATS 6202, Instructor: Yuxiao Huang

Material Covered

- Decision tree
- Random forest

Note

- The quiz has 100 points.
- The quiz period is 20 minutes.
- The quiz is closed book and closed notes.
- The quiz is closed electronics (e.g., no laptops, netbooks, OLPCs, tablets, iPads, calculators, cellular phones, iPhones, Nexi, iPods, Zunes, Kindles, Nooks).
- There is only one correct answer for each `Multiple Choice Question`.
- For each `Calculation` question (if there is any), you must show the essential steps. **No mark will be given if only the result is provided**.

Table 1: The toy dataset.

Day	Weather	Activity
Weekday	Sunny	Work
Weekday	Cloudy	Work
Weekday	Rainy	Work
Weekend	Sunny	Hike
Weekend	Cloudy	Jog
Weekend	Rainy	Read

Figure 1: Decision tree learning algorithm.

Decision tree learning

Aim: find a small tree consistent with the training examples

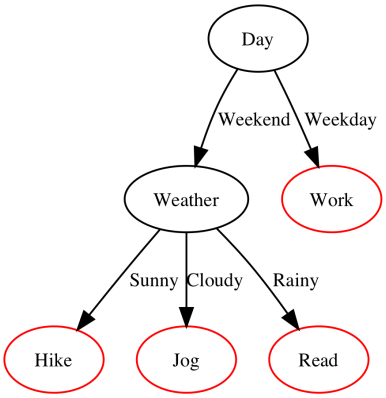
Idea: (recursively) choose "most significant" attribute as root of (sub)tree

```
function DTL(examples, attributes, default) returns a decision tree
    if examples is empty then return default
    else if all examples have the same classification then return the classification
    else if attributes is empty then return MODE(examples)
    else
        best ← CHOOSE-ATTRIBUTE(attributes, examples)
        tree ← a new decision tree with root test best
        for each value v_i of best do
            examples_i ← {elements of examples with best = v_i}
            subtree ← DTL(examples_i, attributes − best, MODE(examples))
            add a branch to tree with label v_i and subtree subtree
        return tree
```

Picture courtesy of the book *Artificial Intelligence: A Modern Approach (Third edition)*

1 Description (100 points)

1. Draw a decision tree learned from the toy dataset (table 1) using Decision tree learning algorithm (fig. 1). Here we assume the best feature is Day. That is, we assume the root of the tree is Day.



2. Suppose there are 1000 features in a dataset and you want to fit your model on 100 features that have the highest predictive power. Briefly describe how this feature selection problem can be addressed by random forest.

   (a) Use random forest to calculate feature importance of each feature

   (b) Sort the features in descending order of their importance

   (c) Select the top 100 features

THIS PAGE INTENTIONALLY LEFT BLANK
(You may use it as scratch paper, but *do* submit it as part of your completed exam.)