Homework #1

CS 539, Fall 2020

---

100 points total [6% of your final grade]

Due: September, 15, 2020 by 11:59pm
[no submission will be accepted after September 17, 2020 at 11:59pm]

Delivery: Submit via Canvas

For this assignment, you will:
(0 pts) Get started with Python
(60 pts) Implement Decision Tree with Discrete Attributes
(40 pts) Credit Risk Prediction

Part 0: Getting Started with Python
As in all the homeworks this semester, you will be using Python. So let's get started first with our Python installation.

Python Installation and Basic Configuration
First off, you will need to install a recent version of **python 3** in here. Right now, the latest version is 3.8. There are lots of online resources for help in installing Python. Here are several depending on your platform:

Windows: http://docs.python-guide.org/en/latest/starting/install3/win/
https://docs.python.org/3/using/windows.html

Linux: https://www.digitalocean.com/community/tutorials/how-to-install-python-3-and-set-up-a-local-programming-environment-on-ubuntu-16-04

Mac OSX: http://docs.python-guide.org/en/latest/starting/install3/osx/

Alternatively, there is a nice collection called Anaconda, that comes with Python plus tons of helpful packages that we may use down the line in this course:

Anaconda: which has install instructions for Windows, Linux, and Mac OSX.

Installing Python Libraries (optional)
You will may need to install python libraries. To manage your Python installations, we recommend pip. Pip is a tool for installing and keeping track of python packages. It is a replacement for easy_install which is included with python. It's a bit smarter than easy_install, and gives better error messages, so you probably want to use it. You can install pip and the two packages we currently need by running these commands:

> easy_install pip

> pip install -r reqs.pip

Then, you may install other Python libraries such as NumPy by typing ‘pip install numpy’

Part 1: Implement Decision Tree with Discrete Attributes

In this assignment, you will implement the decision tree algorithm for a classification problem in **python 3**. We provide the following three files:

a) data1.csv - You will load the file, build a tree, and evaluate its performance.

The first row of the file is the header (including the names of the attributes). In the remaining rows, each row represents an instance/example. The first column of the file is the target label.

b) part1.py – You will implement several functions. Do not change the input and the output of the functions.

c) test1.py – This file includes unit tests. Run this file by typing 'nosetests -v test1.py' in the terminal to check whether all of the functions are properly implemented. **No modification is required**.

Part 2: Credit Risk Prediction

Let's assume that you work for a credit card company. Given the sample credit dataset (credit.txt) as a training set, your job is to build a decision tree and make risk prediction of individuals. The target/class variable is credit risk described as high or low. Features are debt, income, marital status, property ownership, and gender.

Task 2-1: Draw your decision tree and report it. You may use visualization tools (e.g., Graphviz) or use text. You might find it easier if you turn the decision tree on its side, and use indentation to show levels of the tree as it grows from the left. For example:

```
outlook = sunny
|  humidity = high: no
|  humidity = normal: yes
outlook = overcast: yes
outlook = rainy
|  windy = TRUE: no
|  windy = FALSE: yes
```

Feel free to print out something similarly readable if you think it is easier to code.

Apply the decision tree to determine the credit risk of the following individuals:

Name	Debt	Income	Married?	Owns Property	Gender
Tom	low	low	no	Yes	Male
Ana	low	medium	yes	Yes	female

Report a snapshot of your decision tree, and predicted credit risk of Tom and Ana.

Task 2-2: How does your decision tree change if Sofia's credit risk is high instead of low as recorded in the training data? Given the decision tree constructed from the original dataset, if existing, name any feature not playing a role in the decision tree.

What to turn in:

- Submit to Canvas your part1.py, and a pdf document for part2.
- This is an individual assignment, but you may discuss general strategies and approaches with other members of the class (refer to the syllabus for details of the homework collaboration policy).