CPSC 483 - Introduction to Machine Learning

Project 1, Spring 2021

due February 25 (Section 01) / February 26 (Section 02)

Last updated Thursday February 18, 9:15 pm PST

In this project you will familiarize yourself with Python and Jupyter by implementing a rudimentary outlier detection algorithm.

The project may be completed individually, or in a pair of two (2) students as long as both students are enrolled in the same section of the course.

Platforms

For this project (and, in general, for most machine learning or data science projects) you will need a Jupyter notebook with Python 3. Jupyter allows you to create documents mixing text, equations, code, and visualizations.

The Jupyter project itself recommends Anaconda if you intend to run notebooks locally on a laptop or desktop computer. Alternatively you may use a cloud service such as Google Colab that offers Jupyter notebooks online.

Libraries and Code

This project must be implemented in pure Python with the Python Standard Library, without recourse to third-party libraries.

Code from *A Whirlwind Tour of Python* may be reused. All other code and the results of experiments must be your own original work or the original work of other members of your team.

Dataset

The file `participants.csv` contains meeting attendance data reported by Zoom for the first three weeks of a course. Each row contains the name of a student along with the number of minutes that the student was logged in to the course Zoom meeting. (The names of students have been changed to protect the innocent.)

Experiments

Run the following experiments in a Jupyter notebook, performing actions in code cells and reporting results in Markdown cells.

1. Use the `csv` module to load and examine the dataset.

2. Find the quartiles for Week 1. You may wish to consult the Python Sorting HOW TO document for help manipulating the data.

3. In order to record attendance, we want to find the students who logged into the Zoom meeting but did not attend the entire lecture. In order to do this, we can look for outliers in the data.

   Tukey's fences are a simple method to define outliers in terms of the interquartile range. Use this method with *k* = 1.5 to find the outliers in the Week 1 attendance data.

4. Repeat experiments *(2)* and *(3)* for Weeks 2 and 3.

5. Consolidate your code from experiments *(2*) through *(4)* into a Python function named `tardy()`. You may wish to define other functions as well.

   This function should list the name and attendance statistics for any student whose attendance in any week falls below the bottom fence for that week. Tardy students should be listed in the same order as they appear in the original `.csv` file.

Submission

A Markdown cell at the top of the notebook should include project summary information as described in the Syllabus for README files.

Since you may be actively editing and making changes to the code cells in your notebook, be certain that each of your code cells still runs correctly before submission. You may wish to do this by selecting *Run All* from the drop-down menu bar.

Submit your Jupyter `.ipynb` notebook file through Canvas before class on the due date.

If the assignment is completed by a pair, only one submission is required. Be certain to identify the names of both students at the top of the notebook. See the following sections of the Canvas documentation for instructions on group submission:

- How do I join a group as a student?
- How do I submit an assignment on behalf of a group?