CPSC 483 - Introduction to Machine Learning

Project 1, Fall 2021

due Friday, September 17 at 9:45 pm PDT

Last updated Wednesday September 8, 2:05 pm PDT

In this project you will familiarize yourself with Python and Jupyter by implementing some rudimentary algorithms for outlier detection.

The project may be completed individually, or in a team of up to three students as long as all 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 the Python documentation and 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 five 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

Begin by adding a Markdown cell introducing your project. Be sure to include the names of the members of your team, the semester, section, and project number.

Run the following experiments in a Jupyter notebook, running Python code in code cells and describing your results in Markdown cells. Be sure to mark clearly in the notebook where each experiment begins.

1. Use the `csv` module to load the dataset and examine the contents of the first few rows.

2. Load the `statistics` module and use it to find the mean and median of Week 1's data.

3. Find the quartiles for Week 1.

4. 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. (In fact, they are usually included as whiskers in box plots in order to visualize outliers).

   Use this method with *k* = 1.5 to find the outliers in the Week 1 attendance data.

5. Recall that in a normal distribution, 99.7% of the values lie within three standard deviations from the mean. If we assume that our data are normally distributed, this gives us another way to find outliers.

   Compute the standard deviation for the Week 1 attendance data, then use this method to find the outliers. Do your results agree with experiment *(4)*?

6. Define a function `tardy_iqr()` to make experiment *(4)* repeatable. This function should take the name of a column (e.g. `'Week 1'`) and return a list of names for whom the number of minutes is below the lower Tukey fence (e.g. `['Alaya Dickinson', 'Owain Emerson']`). Verify that this function returns the same results as experiment *(4)*.

7. Define a second function, `tardy_stdev()`, with the same interface as experiment *(6)* but using the method of experiment *(5)* and verify that its results match that experiment.

8. Compare the results of `tardy_iqr()` and `tardy_stdev()` on Weeks 2-5.

Submission

As described above, the first cell in your notebook should include the names of the members of your team, the semester, section, and project number. Only one submission is required.

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 by selecting *Run All* from the drop-down menu bar.

Submit your Jupyter `.ipynb` notebook file through Canvas before 9:45 pm PDT on the due date.

The Canvas submission deadline includes a grace period of an hour. Canvas will mark submissions after the first submission deadline as late, but your grade will not be penalized. If you miss the second deadline, you will not be able to submit and will not receive credit for the project.

Note: do not attempt to submit projects via email. Projects must be submitted via Canvas, and instructors cannot submit projects on students' behalf.

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?

Grading

The grade for the project will be assigned on the following five-point scale:

Exemplary (5 points)

Results are correct and clearly presented; explanatory text clearly and concisely tells the story with appropriate context and analysis; organization makes it easy to review.

Basically Correct (4 points)

The analysis comes to correct (or defensible) results and conclusions, but the presentation is not easy to follow and/or portions are not clear or lack context.

Right Idea (3 points)

The approach is appropriate, but the work has mistakes in code, analysis, or presentation that undermine the correctness of conclusions.

Solid Start (2 points)

The work makes a good start, but has fundamental conceptual problems in code, analysis, or presentation such that it will not produce legitimate results.

Did Something (1 point)

The solution began an attempt, but is either insufficient complete to assess correctness or is on entirely the wrong track.

Did Nothing (0 points)

Project was not submitted, submitted code belonging to someone other than the members of the team, or submission was of such low quality that there is nothing to assess.

Acknowledgements: this grading scale is drawn from the general rubric used by Professor Michael Ekstrand at Boise State University.