CPSC 483 - Introduction to Machine Learning

Project 5, Fall 2020

due November 30 (Section 02) / December 3 (Section 01)

Last updated Monday November 16, 5:10 pm PST

In this project we work with some toy data sets to compare the decision boundaries defined by some of the simpler classifiers.

The project may be completed individually, or in a group of no more than three (3) people. All students on the team must be enrolled in the same section of the course.

Platforms

The platform requirements for this project are the same as for previous projects.

Libraries

You will need pandas to load the datasets, scikit-learn to run the various classifiers, and Matplotlib's pyplot framework to visualize your results.

You may reuse code from the Jupyter notebooks accompanying the textbook and from the documentation for the libraries. All other code and the results of experiments should be your own.

Datasets

Download `dataset1.csv`, `dataset2.csv`, and `dataset3.csv`. These datasets have three columns: features x_1 and x_2 and label $t \in \{0, 1\}$.

Experiments

Run the following experiments in a Jupyter notebook, performing each action in a code cell and answering each question in a Markdown cell.

1. Use `read_csv()` to load and examine each dataset.

2. Use logistic regression to `fit()` and `score()` a binary classifier for dataset 1. How accurate are the model's predictions?

3. Repeat experiment *(2)* for dataset 2. How well does it score?

4. Create scatterplots for datasets 1 and 2, plotting points from class 0 with a different color and marker from points in class 1. What accounts for the discrepancies between experiments *(2)* and *(3)*?

5. Fit and score Gaussian Naive Bayes classifiers for datasets 1 and 2. How well do these classifiers score compared to logistic regression?

6. Repeat experiment *(5)* with K-Nearest Neighbor classifiers.

7. Using the second half of the Python code for Figure 9.2 - Simple Gaussian Naive Bayes Classification from *Statistics, Data Mining, and Machine Learning in Astronomy, 2nd Edition* as a guide, plot the decision boundaries for each classifier and dataset. What differences do you observe?

8. Now repeat experiments *(2)*, *(5)*, *(6)*, and *(7)* with dataset 3.

Submission

Submit your Jupyter `.ipynb` notebook file through Canvas before class on the due date. Your notebook should include the usual identifying information found in a `README.TXT` file.

If the assignment is completed by a team, only one submission is required. Be certain to identify the names of all students on your team at the top of the notebook.