Homework 2: Classification Challenge

COSC 410: Spring 2024, Colgate University

Due February 26 Organizations and researchers often post datasets as public ''competitions'' to see how well machine learning models can perform on associated classification and regression tasks. People participate in these challenges for a variety of reasons. In some subfields of machine learning, beating the state-of- the-art performance on a widely-used dataset can be worth an academic publication alone. Others use these challenges ways to practice ML programming, contribute to citizen science initiatives, or win prizes (this ongoing competition is offering $60,000 in prizes for the models that are able to detect personally identifiable information in student writing).

In this homework, you will gain experience with an ML competition in a low-stakes environment, building on Lab 3. Your task is to try out different models to fit the weather prediction task from Lab 3 (using the same training data, `Lab3_train.csv`). At any point before the homework is due, you may upload your current model to Gradescope, where it will automatically be applied to a held out test set. Its performance will then be posted on an anonymized leaderboard so you can see how your model stacks up against the rest of the class.

You are encouraged to upload your model to the leaderboard early and often. While you won't have access to the test set labels, submitting allows you to check whether changes to your model have improved or reduced test set performance. In real ML competitions, it is common for teams to submit a ''naive'' model at the beginning of the competition to set themselves a baseline for improvement.

Ultimately, you will only be graded on 1) whether your model beats a simple 1-Nearest Neighbor classifier, 2) whether your code demonstrates meaningful effort to improve model performance through data preprocessing, model selection, and hyperparameter optimization, and 3) your answers to open-ended questions.

However, there will be **extra credit** awarded based on the leaderboard! The breakdown of extra credit (per Section to adjust for different class sizes) for winning positions is as follows:

Rank	Points	Section Applicable
1	5	A, B
2	4	A, B
3	3	A, B
4	2	A
5	1	A

ML Task Description

The `Lab3_train.csv` file contains 10 years worth of daily weather observations from locations across Australia, one row per day. It contains a column registering a binary label for each observation (`RainTomorrow`) a `1` if it rained on the following day or a `0` if it did not. Your goal will be to create a ML model that, when given a new weather observations, can predict whether it will rain on the day after the observation. In other words, can you use machine learning to predict if it will rain tomorrow based on the weather today?

Note that you should use the same training and validation data from the Lab. We have held-out test data already prepared that your model will be evaluated on.

Instructions

Try out different data preprocessing, models, and hyperparameters in `HW2.ipynb` using modified versions of your `fit_predict()` and `preprocess()` from `Lab3.ipynb`. Ensure that it uses the exact same parameters and output type as specified in `Lab3.ipynb`. You may add helper functions. You may import any needed modules from `sklearn`, `numpy`, `scipy`, `pandas`, `matplotlib`, or `seaborn`. Do not import any other modules as they will not be available in the leaderboard environment.

Submission and Leaderboard

Once you've settled on a version of `fit_predict()` and `preprocess()` that you want to test, copy them (and any helper functions) to a `.py` file. You will upload this `HW2.py` file. **You must call it `HW2.py`**.

You may submit your `HW2.py` to the Gradescope leaderboard as many times as you like before the due date. Submit your model often to see how small changes affect test set performance. When you submit, it will ask you to supply a ''leaderboard name''. You may enter your real name or choose a pseudonym to stay anonymous.

Once you are satisfied with your model's performance, submit your `HW2.ipynb` to Gradescope (in addition to `HW2.py`).