CSE 158/258, Fall 2020: Assignment 1

Instructions

In this assignment you will build **recommender systems** to make predictions related to video game reviews from *Steam*.

Solutions will be graded on Kaggle (see below), with the competition closing at **5pm, Monday November 23** (note that the time reported on the competition webpage is in UTC!).

You will also be graded on a brief report, to be submitted electronically on gradescope by the following day. Your grades will be determined by your performance on the predictive tasks as well as your written report about the approaches you took.

This assignment should be completed **individually**. To begin, download the files for this assignment from: `http://cseweb.ucsd.edu/classes/fa20/cse258-a/files/assignment1.tar.gz`

Files

train.json.gz 175,000 instances to be used for training. This data should be used for the 'play prediction' (both classes) and 'time played prediction' **(CSE258 only)** tasks. It is not necessary to use *all* observations for training, for example if doing so proves too computationally intensive.

userID The ID of the user. This is a hashed user identifier from Steam.

gameID The ID of the game. This is a hashed game identifier from Steam.

text Text of the user's review of the game.

date Date when the review was entered.

hours How many hours the user played the game.

hours_transformed $\log_2(\text{hours}+1)$. **This transformed value is the one we are trying to predict.**

train_Category.json.gz Training data for the category prediction task **(CSE158 only)**. This file is json formatted, and contains the following fields (in addition to most of the fields above):

genreID A numeric label associated with the genre.

genre A string version of the genre.

test_Category.json.gz Test data associated with the category prediction task. This data has the same format as above, with the 'genre' and 'genreID' labels hidden.

pairs_Played.txt Pairs on which you are to predict whether a game was played (both classes).

pairs_Category.txt Pairs (userID and reviewID) on which you are to predict the category of a game **(CSE158 only)**.

pairs_Hours.txt Pairs (userIDs and gameIDs) on which you are to predict time played **(CSE258 only)**.

baselines.py A simple baseline for each task, described below.

Please do not try to collect these reviews from Steam, or to reverse-engineer the hashing function I used to anonymize the data. Doing so will not be easier than successfully completing the assignment. **We will request working code for any solution suspected of violating the competition rules**.

Tasks

You are expected to complete the following tasks:

Play prediction (both classes) Predict given a (user,game) pair from 'pairs_Played.txt' whether the user would play the game (0 or 1). Accuracy will be measured in terms of the *categorization accuracy* (fraction of correct predictions). The test set has been constructed such that exactly 50% of the pairs correspond to played games and the other 50% do not.

Category prediction (CSE158 only) Predict the category of a game from a review. Five categories are used for this task, which can be seen in the baseline program, namely Action, Strategy, RPG, Adventure, and Sport. Performance will be measured in terms of the fraction of correct classifications.

Time played prediction (CSE258 only) Predict how long a person will play a game (transformed as $\log_2(\text{hours} + 1)$, for those (user,game) pairs in 'pairs_Hours.txt'. Accuracy will be measured in terms of the *mean-squared error* (MSE).

A competition page has been set up on Kaggle to keep track of your results compared to those of other members of the class. The leaderboard will show your results on *half of* the test data, but your ultimate score will depend on your predictions across the *whole* dataset.

Grading and Evaluation

This assignment is worth 25% of your grade. You will be graded on the following aspects. Each of the two tasks is worth 10 marks (i.e., 10% of your grade), plus 5 marks for the written report.

- Your ability to obtain a solution which outperforms the leaderboard baselines on *the unseen portion of* the test data (5 marks for each task). Obtaining full marks requires a solution which is substantially better than baseline performance.
- Your ranking for each of the tasks compared to other students in the class (3 marks for each task).
- Obtain a solution which outperforms the baselines on *the seen portion of* the test data (i.e., the leaderboard). This is a consolation prize in case you overfit to the leaderboard. (2 mark for each task).

Finally, your written report should describe the approaches you took to each of the tasks. To obtain good performance, you should not need to invent new approaches (though you are more than welcome to!) but rather you will be graded based on your decision to apply reasonable approaches to each of the given tasks (5 marks total).

Baselines

Simple baselines have been provided for each of the tasks. These are included in 'baselines.py' among the files above. They are mostly intended to demonstrate how the data is processed and prepared for submission to Kaggle. These baselines operate as follows:

Play prediction Find the most popular games that account for 50% of interactions in the training data. Return '1' whenever such a game is seen at test time, '0' otherwise.

Category prediction Look for a few likely words that may appear in reviews of each category (e.g. if the word 'action' appears, classify as Action).

Time played prediction Return the global average time, or the user's average if we have seen them before in the training data.

Running 'baselines.py' produces files containing predicted outputs (these outputs can be uploaded to Kaggle). Your submission files should have the same format.

Kaggle

We have set up a Kaggle page to help you evaluate your solution. You should be able to access the competition via public links. The Kaggle pages for each of the tasks are:
`https://www.kaggle.com/c/cse158258-fa20-play-prediction/`
`https://www.kaggle.com/c/cse158-fa20-category-prediction/`
`https://www.kaggle.com/c/cse258-fa20-time-played-prediction/`

You are welcome to attempt the tasks from either class, but will only be graded on the tasks from your own class.