CSCI 567, Spring 2022

Final Project

Due: May 07, 2022, 11:59 pm, PT (Kaggle)
May 11, 2022, 11:59 pm, PT (Report)

1 Introduction

In the project, you will have the chance to explore an interesting machine learning problem by participating in a Kaggle competition, which is being organized by the H&M Group. In this project, you will develop machine learning models for predicting which article of clothing customers will buy in the next 7-day period. Specifically, given data on individual articles of clothing (type, name, color, group), customer's status (shopping activity, age, zip code), and customer's previous transactions (date, price, items purchased), your challenge is to predict what articles each customer will purchase in the 7-day period immediately after the training data ends. For more details, please refer to official competition website on Kaggle (H&M Personalized Fashion Recommendations[1]).

2 Teams

- The project should be done in teams comprising of **up to 3 members**. Note that the competition officially allows teams of up to 5 members, but we are restricting the team size to a maximum of 3 members for all teams from this class.
- Form a team of up to 3 students by **March 29, 11:59pm PST** and create your accounts on the official competition website. Make sure that all members of your team are registered under a single team name which begins with `CSCI567_id<TEAM ID>`, for example, `csci567_id01`.
- Fill out your team information in this google spreadsheet with your team details. Please log in with your USC email. You can find your team ID in column A within the spreadsheet.
- A team can have members from different sections of the class (offline, online, DEN). Please feel free to use Pizza/slack to arrange your team. We will assign you a team if you cannot find one.

3 Grading

- Your grade is bifurcated as follows: 1) your team's relative rank on the leaderboard **(60%)** and 2) the project report and code **(40%)**.
- Note that the leaderboard shows the ranking of all teams, not just those in this class. We will take the relative ranking among the class' teams into consideration. A better ranking will always lead to a better grade.
- Members of the same team will receive the same scores.
- The team that wins the 1st place among all teams in CSCI-567 class will get 5 bonus points in the total score for machine learning class. If the team wins the 1st place among all the teams on the leaderboard, it will get 5 more bonus points in the total score. We will consider the final ranking on the public leaderboard on **May 08** for grading and bonus points. We might announce an additional bonus policy later.

[1] https://www.kaggle.com/c/h-and-m-personalized-fashion-recommendations

4 Deliverables

- Each team needs to write the project report in NeurIPS format. (6 pages maximum, including references; this page limit is strict). The NeurIPS LATEX format can be found here: https://nips.cc/Conferences/2020/PaperInformation/StyleFiles.
  Not following the format will lower your grade.
- In your report, you should cover the details of your solutions, including the general ideas, the way of data processing and cleaning, the learning algorithms and models you have tried, the results you obtained, and any other insightful thoughts during the competition. You should also describe how to run your code to get the results.
- All teams are advised to use Python as the programming language for their code. We also encourage you to use the Google Cloud Platform or Google Collab for running your codes, if needed. The department does not offer free computing resources but Google Collab and Kaggle combined will give you up to 60hrs/week and 40Gb of storage per team member.

5 Important: Data Files

The competition officially provides the data files below:

- `images/` - a folder of images corresponding to each article_id; images are placed in subfolders starting with the first three digits of the article_id; note, not all article_id values have a corresponding image.
- `articles.csv` - detailed metadata for each article_id available for purchase
- `customers.csv` - metadata for each customer_id in dataset
- `sample_submission.csv` - a sample submission file in the correct format
- `transactions_train.csv` - the training data, consisting of the purchases each customer for each date with additional information. Duplicate rows correspond to multiple purchases of the same item.

We note that the data in `images/` is large (∼31GB), and TAs will help with providing a learned image representation (a function f that takes in `article_id` and outputs a vector of dimension $D \in [256, 512]$). You have the option to use the processed image data or use the original image data. In addition, besides using images for an article's features, tabular features describing the article are also in the file `articles.csv`.

TAs will help to point out computational resources and software packages necessary to perform a submission for the competition but will not be responsible for bug-checking code for all teams.

6 Deadlines

- We will grade you based on your final released ranking on the online public leaderboard on **May 08**. Please make your final submissions on time. Note that the official competition deadline is May 09, 11:59pm UTC but we require teams from this class to submit by **May 07, 2022, 11:59 pm, PT**.
- Each team also needs to submit one PDF copy of the project report and all the code via D2L. This needs to be submitted in a single zip file titled `CSCI567_sp22_id<TEAM ID>.zip` (e.g. `CSCI567_sp22_id01.zip`). Your team ID can be found here. The deadline for D2L submission is **May 11, 2022, 11:59 pm, PT**.
- We will not accept any late submissions.

7 Policy on collaboration

In line with the rules of the competition, you are only allowed to share code within your own team. Your code will be analyzed to reproduce the results and compare its similarity to code from other teams. Any violation of the USC Integrity and Plagiarism policy will lead to an immediate "F" grade in this class and you might be subject to harsher penalties.

However, discussion about approaches between each team members and cross-teams are allowed and we encourage you to actively engage in forums, piazza, and discussion with the Kaggle's community. This provides a great learning opportunity for you to explore how to collaborate and researching new approaches to tackle hard challenges.