In-Class Kaggle Competition

IEOR E4525 Spring 2021

March 18, 2021

Instructor: Prof. Christian Kroer
TA: Yuan Gao, Tyler Will

Task

Given features of a new ad impression, output its click-through rate (CTR), i.e., the probability that it will be clicked. You should submit your predictions to the following private Kaggle competition: `https://www.kaggle.com/t/7005cd40df944af385049759bdf7235d`

You should also submit a writeup document (using Word, Latex, or any text editing tool), which should include the following:

- Resources you consulted other than documentation of common Python tools: research papers, blogs, other peoples' codes, lecture notes and slides.
- The method you use to produce the final submission.
- Other methods you tried (which may or may not work well).

The writeup should be submitted on gradescope. The writeup can be brief: no need to do anything extensive. We just want to hear what you did. You will not be graded on the quality of your writeup.

Along with your write-up, you must also submit your code as a tar or zip file. Your code will not be graded.

Data

The file `train.gz` contains the training data. All features are categorical. The binary outcome variable is `click`.

All data can be found here: `https://drive.google.com/drive/folders/1NYXsTO23Szv9PzHT4SXk_og8SGdEuUus?usp=sharing`

You can find a description of the data on the original Kaggle competition site (but please remember that you are not allowed to train on the full Kaggle dataset): `https://www.kaggle.com/c/avazu-ctr-prediction/data`

Submission Format and Evaluation

You should submit a `Submission.csv` file with two columns, `id` and `ctr`. Each row of the file contains the id and predicted CTR of the respective test observation. A sample submission is available on the Kaggle webpage. The evaluation metric is the log-loss, that is,

$$\frac{1}{n_t}\sum_{i=1}^{n_t}\left[-y_i \log p_i + (1-y_i)\log(1-p_i)\right],$$

where n_t is the number of test observations, $y_i \in \{0,1\}$ are the true test outcomes (binary variables representing clicks and no-clicks) and p_i are your predicted CTR (a real number between 0 and 1).

A Baseline Solution

The file `baseline.py` illustrates a baseline solution using simple embedding and logistic regression. Note that it only uses a fraction of the training data and does not perform any parameter selection through validation.

Other Notes

This task is a simplified version of the Avazu CTR Prediction challenge. Feel free to search and learn more about the full dataset and winning solutions.

You can use any publicly available algorithms and codes. However, you should cite the resources you used in the writeup and must only use the provided training data `train.gz`.

For this task, cross-validation is not advisable since the dataset is time-dependent. You should split the entire training dataset into training and validation subsets. You should think about what is the right way to do that, however.

Grading

You will get 100% for this assignment as long as you get a score of 0.425 or lower on the private leaderboard (this is easy to achieve, but it is not achieved by the baseline).

Bonus Credit

Bonus credit (applied after grade cutoff computation): top 10% of submissions (measured on Kaggle private leaderboard using held-out data) will receive 50 additional points split proportionally between the submissions according to performance.

The proportional split will look at the performance v_{11} of the 11th ranked submission, and for each $i \in \{1,\ldots,10\}$, give bonus credit of $50\frac{v_{11}-v_i}{\sum_{k=1}^{10}(v_{11}-v_k)}$. Each individual on a winning team will receive $v_i/|T_i|$ where $|T_i|$ is the number of people on the team.