Machine Learning & Deep Learning for Data Science
Seoul National University — Spring 2022, Assignment 2, Due: 2022.5.3. 11:59pm

Instructions: Please read the following instructions thoroughly

- For the entire assignment, use `Python` for your analysis. Write your code in a Jupyter Notebook named as `[your-student-ID]_hw2.ipynb` (e.g., `2022-20000_hw2.ipynb`). The use of `R` is not allowed. You are allowed to use any libraries in `Python`. Type up your report and save as PDF named as `[your-student-ID]_hw2.pdf`. We do not allow the submission of a photo or a scanned copy of hand-written reports. A CSV file containing the predictions of test data should be named as `[your-student-ID]_pred.csv`

- Please upload a single zip file named as `[your-student-ID]_hw2.zip` (without the bracket, e.g., `2022-20000_hw2.zip`) on eTL, containing the your report in PDF, the code in Jupyter Notebook, and the CSV file for the test data predictions. (Note there are three files this time!) Submissions via email are not allowed. The violation of the filename or submission instruction will result in the penalty of 5 points.

- You can discuss the assignment with your classmates but each student must write up his or her own solution and write their own code. Explicitly mention your classmate(s) you discussed with or reference you used (e.g., website, Github repo) if there is any. If we detect a copied code without reference, it will be treated as a serious violation of the student code of conduct.

- We will apply a grace period of late submissions with a delay of each hour increment being discounted by 5% after the deadline (i.e., 1-minute to 1-hour delay: 95% of the graded score, 1 to 2-hour delay: 90% of the graded score, 2 to 3-hour delay: 85%, so on). Hence, if you submit after 20 hours post-deadline, you will receive 0 points. No excuses for this policy, so please make sure to submit in time.

1. In this problem, you will use the `OnlineAd` training data set attached in the assignment (`OnlineAd_X_train.csv` and `OnlineAd_Y_train.csv`). The feature matrix `OnlineAd_X_train.csv` contains 1452 observations, where each row represents a different user with 251 features that summarize the user characteristics and previous browsing history. These 251 features have been anonymized and normalized. Hence, each feature's interpretation is hidden for analysis. (i.e., you don't have to worry about the meaning of each feature.) There are two online advertisements, `A` and `B`, with which this experiment was performed. In each observation, a user clicks either `A` or `B`, or she may not click anything. She cannot click on two ads at the same time. This click response is recorded in `OnlineAd_Y_train.csv`, where each row contains information on which ad was clicked (or whether a user clicked nothing). A *no click* response is recorded as 1 in the first column in `OnlineAd_Y_train.csv`, a click on `A` is 1 in the second column, and likewise for a click on `B` is 1 in the third column. Note that 0 means non-chosen options by the user. Hence, each row sum of `OnlineAd_Y_train.csv` is exactly 1. Each row of `OnlineAd_Y_train.csv` is the click response for the corresponding user in the corresponding row of `OnlineAd_X_train.csv`

   (a) [40 pts] Using `OnlineAd_X_train.csv` and `OnlineAd_Y_train.csv`, train multiple models that you learned in class. You may try a model which is modified from the models we covered in the first half of the course. You are allowed to use the existing packages, but make sure to clearly explain what models are used. Report the training results. What metric did you use? How do different models perform on the training data?

(b) [20 pts] Do you think dimension reduction on features (or feature selection) is needed here? If so, provide analysis on which features may be important. If not, please justify your answer.

2. Based on your training results in Problem 1, you now pick your best model that would generalize well to unseen data. Using your best model, we are going to predict on the provided test dataset, `OnlineAd_X_test.csv` which contains 300 observations with 251 features. Here we do not have the corresponding responses for these test observations.

   (a) [20 pts] Report the estimated test performance for your best model. Provide a reason for your choice of a model among the models you considered.

   (b) [20 pts] Predict on the provided test dataset, `OnlineAd_X_test.csv`, and save those predictions as a CSV file named `[your-student-ID]_pred.csv`. The CSV file should only contain the array of dimension [300, 3] in the same format as the `OnlineAd_Y_train.csv` except the number of rows, since there are only 300 observations in the test data, i.e., the first column corresponds to no click, the second column corresponds to the ad `A`, and the third column corresponds to the ad `B`. *A violation of this format guideline will result in 10 point penalty.*[1]

   (c) [*Extra* 10 pts] For those whose submissions achieve the top 5 test performances out of all the submissions, an extra credit will be given.

[1] The score in this part will be based on the test misclassification. However, we will not provide a threshold on how good your prediction should be. (Note that often in real-world problem settings, you may not know a priori how good is good enough...) No worries, we will be generous on this part: we will still give 10 points even for completely wrong predictions as long as the predictions are in the correct format!