Machine Learning & Deep Learning
for Data Science
Seoul National University

Fall 2022
Assignment 1
Due: 2022.10.16. 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]_hw1.ipynb` (e.g., `2022-20000_hw1.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]_hw1.pdf`. We do not allow the submission of a photo or a scanned copy of hand-written reports.
- Please upload the two files on eTL **without zipping**. 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. [**30 pts**] In this problem, you will use the `Carseats` data set attached in the assignment (`Carseats.csv`) for linear regression.

   (a) [10 pts] Fit a multiple linear regression model to predict `Sales` using `Price`, `Urban`, and `US`. Report the R^2 of the model.

   (b) [5 pts] Write out the model in equation form, being careful to handle the qualitative variables properly. Provide an interpretation of each coefficient in the model.

   (c) [5 pts] For which predictor variable j can you reject the null hypothesis $H_0 : \beta_j = 0$? for which there is evidence of association with the outcome.

   (d) [10 pts] Obtain 95% confidence intervals for the coefficient(s).

2. [**30 pts**] In class, we used the example of the logistic regression model to predict the probability of `default` using `income` and `balance` on the `Default` data set attached in the assignment (`Default.csv`). In this problem, we will estimate the test error of this logistic regression model using the validation set approach. Do not forget to set a random seed before beginning your analysis.

   (a) [10 pts] Fit a logistic regression model that uses `income` and `balance` to predict `default`. Report the log-likelihood of the model.

   (b) [5 pts] Write out the model in equation form and provide an interpretation of each coefficient in the trained model.

(c) [5 pts] Perform 5-fold cross-validation using the model in Part (a), and estimate the test error of this model.

(d) [10 pts] Now consider a logistic regression model that predicts the probability of `default` using `income`, `balance`, and a dummy variable for `student`. Estimate the test error for this model using the 5-fold cross-validation set approach. Comment on whether or not including a dummy variable for student would lead to a reduction in the test error rate.

3. [**40 pts**] In this problem, you will predict the number of applications received using the other variables in the `College` data set attached in the assignment (`College.csv`).

   Randomly split the data set into a training set and a test set by 90:10 ratio.

   (a) [10 pts] Fit a linear model using least squares on the training set, and report the test error obtained.

   (b) [10 pts] Fit a ridge regression model on the training set, with λ chosen by 10-fold cross-validation. Report the test error obtained.

   (c) [10 pts] Fit a lasso model on the training set, with λ chosen by 10-fold cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates.

   (d) [10 pts] Comment on the results obtained. How accurately can you predict the number of college applications received? Is there much difference among the test errors resulting from these three approaches? Which model would you use?