Homework 2

DSO 530: Applied Modern Statistical Learning Methods

Spring 2022

Deadline. Thursday, March 24th, 5 pm, Los Angeles time.

Submission instruction. To submit your homework, please go to the **assessments** folder on Blackboard and find HW2. The submitted document should be in **pdf** format. Submission of other formats will incur a 20% penalty. Before you submit, please double check if your document is readable. You can submit your homework multiple times before the deadline, and the TA will grade the last submission.

Late submission policy. All late submission should be sent to xint@marshall.usc.edu. Late submission will incur 20% penalty within 24 hours after the deadline, and 40% penalty between 24 and 48 hours. **No submission after 48 hours of the deadline will be accepted**. As we have a large class this year, I wish that everybody respects this policy.

Additional instructions for HW2. When the path to an answer involves coding, please show the **Python codes and proper output**. Answering questions clearly and concisely is better than writing page-long convoluted responses. Discussion among students is strongly encouraged. But everyone should write up their own solution. **If one copies others' homework or lets others to copy the homework, they will receive 0 for this homework and face further penalty in the final grade**.

1. (i) Did you review up to and including lecture 5 and tutorial 6? If you haven't, please do so first. This homework, like every other homework, only covers a small part of the course contents. (ii) How many hours did you spend on preparing for test 1?

2. You have four balls in a box: red, black, blue and yellow. Write python code to sample with replacement from this box. Suppose each ball has an equal chance of getting selected. Start with `np.random.seed(2)` and get three samples of size 5. (You can modify the code in Lecture 5).

3. Suppose random variables $X_1, \ldots, X_n$ are independent and identicially distributed (i.i.d.) $N(0, 1)$. What are the mean and variance of $\bar{X} = (X_1 + \ldots X_n)/n$? You need to show the steps to get the conclusion.

4. Use the `email spam` dataset that appears in Tutorial 6. Split this dataset randomly into training and test sets (use the `train_test_split` function, set `random_state = 5`, and use 20% as test set). Train an LDA model on the training dataset and use 1/2 as threshold for $P(Y = 1|X = x)$ to get a classifier. Report classifcation error on test set.

5. For the LDA model you trained in problem **4**, plot the ROC curve on the test data and calculate the AUC.

6. (optional, not for grading) Using the `Housing` dataset, fit classification models to predict whether a neighborhood has crime rate above ($>$) or below ($\leq$) the median. Explore logistic regression, LDA, and KNN models (e.g., $K = 5$) using the other variables (i.e., not using `crim`) in the dataset as the predictors. (The instruction of this problem is intentionally vague. You have much freedom to explore.)