Homework 1

DSO 530: Applied Modern Statistical Learning Methods

Spring 2022

Deadline. Thursday, Feb 10th, 4 pm, Los Angeles time.

Submission instruction. To submit your homework, please go to the **Assessments** folder on Blackboard and find HW1. 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. If you did not generate PDF documents before, please read the companion guide. 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 HW1. 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 his/her own solution. **If one copies others' homework or lets others to copy the homework, he/she will receive 0 for this homework and face further penalty in the final grade**.

1. (i) Did you go over Lectures 1, 2a, 2b, Python tutorials 1 and 2, and Lecture 2Ex (and your notes). This homework, like every other homework, only covers a small part of our course materials. (ii) How many hours outside class time did you spend on DSO 530 each week on average so far? (an honest answer will help the instructor adjust the pace if necessary). (iii) Are you physically in the California for the spring term? If not, what is your location?

2. For a regression equation $\log y = 1 + 70 \log x$, how does changes in y associate with changes in x? (You don't have to type latex math formulas; as long as your derivation is readable by the grader, it is fine; you might need Python or a calculator to help get the answer).

3. Recall that in Lecture 2b, we regress `medv` on `paratio` and `rm` using the Housing data. Repeat this regression, but normalize both features before running the regression. Compared with what you see in the Lecture, do you get a different R^2, or do you get the same one?

4. With the Housing data, use the per capita crime rate (`crim`) as the response, and the other variables (including the `medv` variable) as the predictors. Report the R^2 and adj-R^2. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$ (at the 5% level)? Then, regress `crim` on the significant predictors ONLY; how do you interpret the slopes in the regression?

5. Randomly split the Housing data into two parts with 30% as test data. Use `random_state = 2` in this split. Because this is a regression problem, you don't want to use `stratify = y` part of the code from our Python tutorial. Regress `medv` on `river` and `rm` using the training data. Compute R^2 on both the training data and the test data (i.e., in-sample R^2 and out-of-sample R^2). (Hint: if you use the `sklearn` package, computing the out-of-sample R^2 just needs you to replace the arguments of the `.score(.,.)` in your code to compute the in-sample R^2.)

6. (optional, not for grading) IThe out-of-sample R^2 can be negative even if one uses the least squares method to fit the linear model. Can you design a simulation setting where this phenomenon indeed occurs?