API 222 Problem Set 1

Machine Learning and Big Data Analytics

Due Sept. 29, 2020 by midnight EST

Conceptual Questions

1. For each of the following questions, state: **(6 pts)**

   (1) Whether it is a regression question or a classification question

   (2) Whether we are interested in inference or prediction

   (a) The New York City Mayor's Office plans to solicit construction bids for a school renovation project. They want to know if each bid will be finished on time. They have past data on NYC construction bids, including information on project type, construction company characteristics, budget estimates and whether the bid was finished on time.

   (b) The mayors office also care about the budget considerations. Officials want to avoid projects that understate their true budget and pick bids that will have final spending close to the proposed budget. The vast majority of projects are over budget, so the office wants to know to how much over budget each potential bid will be.

   (c) The mayor noticed that local contractors seemed to win more construction bids. The mayor was interested in if local contractors are better at assessing the needs of the community or if political connections were driving this trend. The mayor decided to implement a blind submission process where the location and name of the firm are hidden for two years. He wants you to analyze whether blind hiring changes the likelihood of a local bid being chosen. For each proposal for the last 10 years, you have details on year submitted, location, number of employees, firm age, number of previous contracts, total portfolio amount, whether the proposal was accepted, etc.

2. Flexible models versus inflexible models **(1 pt)**

   (a) I have two models, one with high bias and variance and one with low bias and low variance. I should choose the model with high bias. True or False?

   (b) KNN and linear regression are both parametric models, as they both have decision rules. True or False?

   (c) A second order polynomial will always have lower bias than a linear model. True or False?

   (d) We should run our model multiple times and pick the one with the lowest test error. True or False?

3. In two sentences or less, describe the bias variance tradeoff. **(1 pt)**

Data Questions

For this part of the problem set, please load the `CASchools` dataset from the package `AER`. Also, for any non-integer numbers, ***please report your numbers to exactly two decimal places*** for full credit. Working directories are important to help keep your work organized and keep good records. Create a path directory and include this code at the top of your R File.

1. How many observations are in the dataset? **(0.5 pts)**
2. How many variables are in the dataset? **(0.5 pts)**
3. Are any of the columns categorical? **(0.5 pts)**
4. Are any values missing? **(0.5 pts)**
5. What is the mean number of students in the school? **(0.5 pts)**
6. What is the standard deviation of number of computers? **(0.5 pts)**
7. What does the calworks variable measure? Hint: Read the codebook! **(0.5 pts)**
8. How many observations would it drop if you limited the sample to schools with 500+ students? **(0.5 pts)**

   Sort the data by number of students (ascending) and put the first 80 in the test set and the last 200 in the training set. Include only the following variables in the test/training set: students, teachers, calworks, lunch, computer, expenditure, income, english, and read. Our outcome variable is going to be the reading scores.

9. Is there anything wrong with how we split our data into training and test datasets? (Note: do not change your dataset splits, this is purely a theoretical question) **(0.5 pts)**
10. When you use your ***training data*** to build a linear model that regresses reading score on all other variables available in the data (plus an intercept), what is your ***test*** Mean Squared Error? **(0.5 pts)**
11. Now use your ***training data*** to build a linear model that regresses reading scores on number of students, teachers, and computers. What is the coefficient on computers (include an intercept). **(0.5 pts)**
12. Now do the same thing but regress reading scores on number of students, teachers, income, and computers (again, include an intercept). **(0.5 pts)**
    - What is the coefficient on computers now?
    - What does that imply about the relationship between computers and income and reading scores and income?
13. When you use your ***training data*** to build a k-Nearest Neighbors model that regresses reading scores on all other features in the data, what is your ***test*** Mean Squared Error with $k = 5$? **(0.5 pts)**
14. When you use your ***training data*** to build a k-Nearest Neighbors model that regresses reading scores on all other features in the data, what is your ***test*** Mean Squared Error with $k = 10$? **(0.5 pts)**