May 20, 2020

CS 530 Spring 2020—Final Exam

The “in-class” part of the final exam includes 2 questions and 1 bonus question.

Sign the below with your full name:

I hereby certify that I worked on this exam by myself and everything written in this exam was conceived and developed by me alone. I did not discuss or share any part of this exam with anyone else and certainly did not use information from anyone else in any part of the exam.

**Your full name: Duc Le.**

Question 1 (3 points)

Multiple linear regression is guaranteed to have no bias (under some assumptions discussed in class). However, there are no guarantees about its variance, which can be very high. One solution to the high variance that we discussed in class is regularization.

You are given a dataset where you need to predict an individual’s height based on a slew of information about each person—weight, gender, shoe size, relevant medical history, yearly income, and much more. All in all, the dataset contains 200 samples, each with 100 features. Domain expertise suggests that the great majority of the variables in this dataset have an influence on the dependent variable (“y”).

1. You are concerned that a simple OLS model would have high variance. Can you think why you might be concerned? (Hint: How does the variance of the error term in multiple linear regression depend on the number of samples and features?)

**Your answer:** If I fit a simple OLS model through a data set with high dimensions, the model will tend to overfit as it tries to fit a hyperplane of best-fit through all those points, while minimizing the RSS. The RSS can have high variance due to the large number of variables in X.

Eq: RSS = (y – XB).T \* (y – XB).

1. Because of your concern, you decide to regularize the model to lower the variance. Your colleague Jordan says that you should take the data (“X”) as is and use either Ridge or LASSO, as both will be similarly successful in this case. Do you agree? Explain why Jordan is correct, if you agree, or describe the correct procedure to regularize the data, if you disagree, and explain why it is correct. (Note that, if you think Jordan is mistaken, there might be more than one problem with what Jordan is suggesting to you.)

**Your answer:**

With domain knowledge, we know the great majority of the variables in the dataset have an influence on the outcome, I agree with Jordan and attempt Ridge or LASSO as a regularization method. They are similar in a way that they both contain a penalty that prevents overfitting. While Ridge reduces the coefficients from non-important variables thus limiting their influence on the model, LASSO completely reduces them to 0, which results in feature selection. Without knowing the rest of the variables, I’d attempt both regularization methods and see which one produces the best results.

1. Working on this dataset, you become worried that it might contain a lot of correlated variables. Jordan says that you must get rid of some of these correlated variables. Why is Jordan claiming that? In other words, what would happen if you kept all the variables? Explain how you would get rid of these variables. Jordan further claims that dimensionality reduction is another method that would assist you in overcoming the shortfalls of correlated variables. Is Jordan correct? If so, explain how you would use dimensionality reduction for this purpose. If you think Jordan is mistaken, explain why you think that.

**Your answer:**

In the case of multi-collinearity, Jordan’s concern is validated since strong collinearity results in high variance in the coefficient estimates. Therefore, if we decide to keep all the variables, our model might be sensitive to future data. To select features, we can use either Forward Stepwise or Backward Stepwise Selection with Adj R2, AIC or BIC as our scoring criteria. Also, we could also compare the variables’ VIF scores and see which ones have high collinearity.

Jordan is correct for most cases. However, in our data set, I am concerned that reducing dimensions from a large data set could be difficult to interpret. For this particular data set with 100 parameters, I don’t think it’s wise to reduce the dimensions as the variance of these variables can be saturated the more dimensions we reduce.

Question 2 (3 point)

Jordan creates a neural network with an input layer, *n* hidden layers (*n* is large), and an output layer. All the neurons in the hidden layers of Jordan’s network use the linear activation function. Show that Jordan’s network is not more powerful than a network with a single hidden layer of linear neurons.

This question may require you to write some math. You might therefore find it easier to write your answer on paper and then take a picture and **upload the photo into the answer space below**. If you do this, make sure to write clearly and legibly. Think of your reader.

**Your answer:**

That is because as you increase the number of hidden layers, the number of weights that need to be updated increase exponentially as well (depends on the number of inputs per hidden layer). Whereas in a single layer of linear neurons, the amount of weights that need to be updated are proportional to the number of inputs that you decide for that layer. Therefore, a network with a single hidden layer is much more powerful and less computationally expensive.

Bonus Question (0.5 points)

What are the 3 most important things that you learned about machine learning in this course?

**Your answer:**

1. Just because you have an excellent training model, it does not guarantee good results in future testing sets.
2. All the different methods and algorithms in Machine Learning.
3. All the coding practices from homework assignments.

Thank you, have a good summer.