Problem Set #1

API 222A: Machine Learning and Big Data Analytics

Cian Stryker

9/15/2020

Conceptual Questions

Question 1

a.) This is a classification question and we are interested in prediction.

b.) This is a regression question and we are interested in prediction.

c.) This is a classification question and we are interested in inference.

Question 2

a.) False

b.) False

c.) True

d.) False

Question 3 Models that use few parameters typically have high bias and low variance, but models with many parameters typically have low bias and high variance. A good model typically finds a balance between bias and variance in order to minimize total error.

Data Questions

1.) There are 420 observations in the data set.

2.) There are 14 variables in the data set.

3.) Four of the columns are categorical: District, School, Country, and Grades.

4.) There are no missing values.

5.) The mean number of students is 2628.79.

6.) The standard deviation of the number of computers is 441.34.

7.) Calworks means the percent of stunts who qualify for Calworks or income assistance.

8.) 139 observations would be dropped if you limit the sample to schools with 500+ students.

9.) An issue with splitting the data this way is that your training data is using only large schools while your test data is largely using small schools. Smaller schools and larger schools tend to differ from each other within a few categories. For example, small schools within urban centers are often private schools with radically different socio-economic levels and demographics than larger urban schools. Small schools could also be rural schools that also differ strongly from the typical large school that is urban. Splitting the data based on student size and making the test use small student bodies while the training data uses large student bodies may negatively impact how well an algorithm will be able to translate to the test data because of these differences. Also, in general, you should randomize the process when you split your data into training and test subsets.

10.) The Mean Squared Error for the test data using the linear model from the training data is 141.13.

11.)

```
## (Intercept)    students    teachers    computer
##    "656.18"    " -0.01"    "  0.07"    "  0.02"
```

The coefficient on computers is 0.02.

12.)

```
## (Intercept)    students    teachers      income    computer
##    "620.39"    "  0.00"    " -0.06"    "  2.17"    "  0.01"
```

a.) The coefficient on computers is now 0.01.

b.) This implies that reading scores has a positive correlation with income and that computer scores and income are closely correlated to each other as well. A one unit increase in computer causes a 0.01 increase in reading level, when income is added as a variable. In adding income as a variable to the model, the effect size of computer decreased, which suggests that income captures some of computer's effect. Income has a coefficient of 2.17, which means that for a unit increase in income, reading increases by 2.17. Income has a stronger positive effect on reading than computer does and likely income accounts for some of computer's effect on reading that was attributed to computer in the previous model. When taking income into account, computer's effect decreases.

13.) The Mean Squared Error on the test data with k = 5 is 383.28.

14.) The Mean Squared Error on the test data with k = 10 is 358.18.