API 222 Problem Set 1

Liz Masten

9/23/2020

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.

This question is a classification question because the output variable ("finished on time") is discrete. This is discrete because the projects will be either finished on time or they won't be. We are interested in prediction here because we are using existing data (the past data on NYC construction bids) to predict outcomes revolving around new information (the new construction bids).

(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.

This question is a regression question because the outcome variable (how much over budget) is continuous. We are interested in prediction again because we are still using data on past construction bids to predict by how much each new bid will be over budget.

(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.

This question is a classification question because the output variables are categorical or discrete. Here, we are interested in inference because we are using existing data to look at the effects within that data.

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?

FALSE - ideally we should have low bias and low variance.

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

FALSE - linear regression is parametric but KNN is not.

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

TRUE - but it might have a higher variance due to overfitting, depending on the data.

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

FALSE

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

Variance is the amount by which $\hat{f}$ changes if estimated using different training data and bias is the error that is introduced when an overly simplistic model is used on a more complet dataset. A tradeoff occurs when, as we use more flexible models that better fit our data, variance increases while bias decreases.

Data Questions

1. How many observations are in the dataset? (0.5 pts): 420 observations.
2. How many variables are in the dataset? (0.5 pts): 14 variables

```
data("CASchools")

#commented out because this doesn't need to show up in the pdf:

#dim(CASchools)
```

3. Are any of the columns categorical? (0.5 pts): Yes, 4 - District (chr), School (chr), County (factor), and Grades (factor). The rest are numeric.

```
str(CASchools)
```

```
## 'data.frame':    420 obs. of  14 variables:
##  $ district   : chr  "75119" "61499" "61549" "61457" ...
##  $ school     : chr  "Sunol Glen Unified" "Manzanita Elementary" "Thermalito Union Elementary" "Golde
##  $ county     : Factor w/ 45 levels "Alameda","Butte",..: 1 2 2 2 2 6 29 11 6 25 ...
##  $ grades     : Factor w/ 2 levels "KK-06","KK-08": 2 2 2 2 2 2 2 2 2 1 ...
##  $ students   : num  195 240 1550 243 1335 ...
##  $ teachers   : num  10.9 11.1 82.9 14 71.5 ...
##  $ calworks   : num  0.51 15.42 55.03 36.48 33.11 ...
##  $ lunch      : num  2.04 47.92 76.32 77.05 78.43 ...
##  $ computer   : num  67 101 169 85 171 25 28 66 35 0 ...
##  $ expenditure: num  6385 5099 5502 7102 5236 ...
##  $ income     : num  22.69 9.82 8.98 8.98 9.08 ...
##  $ english    : num  0 4.58 30 0 13.86 ...
##  $ read       : num  692 660 636 652 642 ...
##  $ math       : num  690 662 651 644 640 ...
```

4. Are any values missing? (0.5 pts): No.

```
map(CASchools, ~sum(is.na(.)))
```

```
## $district
## [1] 0
##
## $school
## [1] 0
##
## $county
## [1] 0
##
## $grades
## [1] 0
##
## $students
## [1] 0
##
## $teachers
## [1] 0
##
## $calworks
## [1] 0
##
## $lunch
## [1] 0
##
## $computer
## [1] 0
##
## $expenditure
## [1] 0
##
## $income
## [1] 0
##
## $english
## [1] 0
##
## $read
## [1] 0
##
## $math
## [1] 0
```

5. What is the mean number of students in the school? (0.5 pts):

```
mean(CASchools$students) %>% round(2)
```

```
## [1] 2628.79
```

6. What is the standard deviation of number of computers? (0.5 pts):

```
sd(CASchools$computer) %>% round(2)
```

```
## [1] 441.34
```

7. What does the calworks variable measure? Hint: Read the codebook! (0.5 pts): "Percent qualifying for CalWorks (income assistance)"

8. How many observations would it drop if you limited the sample to schools with 500+ students? (0.5 pts): We are left with 281 rows, meaning that we dropped 139 rows.

```
drop <- CASchools %>%
        filter(students >= 500) %>%
        nrow()

drop
```

```
## [1] 281
```

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.

```
CASchools$students <- as.numeric(CASchools$students)

order <-CASchools %>%
        arrange(students) %>%
        select(students, teachers, calworks, lunch, computer, expenditure, income, english, read)

test_CAS <- order[1:80,]

training_CAS <- order[221:420,]
```

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)

Yes, this is not good practice for actual randomization because it's not random. We might be training the data on patterns in the dataset by virtue of observation order rather than a representative sample of our data. If we were doing this for real, we would use the sample() function.

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):

```
#build linear model:

model_10 <- lm(read ~ students + teachers + calworks + lunch + computer + expenditure + income + englis

#find MSE:
```

```r
predict_m10 <- predict(model_10, test_CAS[,-9])

mse_m10 <- mean((predict_m10 - test_CAS[,9])^2) %>% round(2)

mse_m10
```

```
## [1] 141.13
```

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): The intercept is 656.18 and the coefficient of computer is 0.02

```r
model_11 <- lm(read ~ students + teachers + computer, data = training_CAS)

model_11
```

```
##
## Call:
## lm(formula = read ~ students + teachers + computer, data = training_CAS)
##
## Coefficients:
## (Intercept)     students     teachers     computer
##  656.182195    -0.005721     0.065790     0.015936
```

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?: Coefficient on computers is now 0.01 with an intercept of 620.39 • What does that imply about the relationship between computers and income and reading scores and income?:

For each one unit increase in income, reading scores increase by about 2.18. For each one unit increase in number of computers, reading scores increase by 0.01. We can compare this to our model in question 11 which contained the computer variable but not the income variable. In model 11, the effect of number of computers was higher (0.02) which indicates that the income and computer variables are related and the coefficient for computer is lower in model 12 because income may be confounding the computer variable in model 11.

```r
model_12 <- lm(read ~ students + teachers + income + computer, data = training_CAS)

model_12
```

```
##
## Call:
## lm(formula = read ~ students + teachers + income + computer,
##     data = training_CAS)
##
## Coefficients:
## (Intercept)     students     teachers       income     computer
##  620.385081     0.001493    -0.061786     2.174355     0.007850
```

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):

```r
# read scores are the 9th column

knn_m1    <- knn.reg(training_CAS[,-9],
                     test_CAS[,-9],
                     training_CAS[,9],
                     k = 1)

knn_m5   <- knn.reg(training_CAS[,-9],
                    test_CAS[,-9],
                    training_CAS[,9],
                    k = 5)

#find MSE:

mse_knn5 <- mean((knn_m5$pred - test_CAS[,9])^2) %>% round(2)

mse_knn5
```

```
## [1] 383.28
```

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):

```r
knn_m10   <- knn.reg(training_CAS[,-9],
                    test_CAS[,-9],
                    training_CAS[,9],
                    k = 10)

mse_knn10 <- mean((knn_m10$pred - test_CAS[,9])^2) %>% round(2)

mse_knn10
```

```
## [1] 358.18
```