Homework 6 Part II

Instructions

Due: 5:00pm on Wednesday, November 3rd

1. Add your name between the quotation marks on the author line in the YAML above.
2. Compose your answer to each problem between the bars of red stars.
3. Commit your changes frequently.
4. Be sure to knit your .Rmd to a .pdf file.
5. Push both the .pdf and the .Rmd file to your repo on the class organization before the deadline.

Theory

Problem 1

Based on ISLR Exercise 4.4

When the number of features p is large, there tends to be a deterioration in the performance of KNN and other *local* approaches that perform prediction using only observations that are *near* the test observation for which a prediction must be made. This phenomenon is known as the *curse of dimensionality*, and it ties into the fact that non-parametric approaches often perform poorly when p is large. We now will investigate this curse.

a. Suppose that we have a set of observations, each with measurements on $p = 1$ feature, X. We assume that X is uniformly distributed on $[0, 1]$. Associated with each observation is a response value. Suppose that we wish to predict a test observation's response using only observations that are within 10% of the range of X closest to that test observation. For instance, in order to predict the response for a test observation with $X = 0.5$, we will use observations in the range $[0.55, 0.65]$. On average, what fraction of the available observations will we use to make the prediction?

b. Now suppose that we have a set of observations, each with measurements on $p = 2$ features X_1 and X_2. We assume that (X_1, X_2) are uniformly distributed on $[0, 1] \times [0, 1]$. We wish to predict a test observation's response using only observations that are within 10% of the range of X_1 and within 10% of the range of X_1. For instance, in order to predict the response for a test observation with $X_1 = 0.6$ and $X_2 = 0.35$, we will use observations in the range $[0.55, 0.65]$ for X_1 and in the range $[0.3, 0.4]$ for X_2. On average, what fraction of the available observations will we use to make the prediction?

c. Now suppose that we have a set of observations on $p = 100$ features. Again the observations are uniformly distributed on each feature, and again each feature ranges in value from 0 to 1. We wish to predict a test observation's response using observations within 10% of each feature's range that is closest to that test observation. What fraction of the available observations will we use to make the prediction?

d. Using your answers to parts (a)-(c), argue that a drawback of KNN when p is large is that there are very few training observations "near" any given test observation.

e. Now suppose that we wish to make a prediction for a test observation by creating a p-dimensional hypercube centered around the test observation that contains, on average, 10% of the training obser-

vations. For $p = 1$, 2, and 100, what is the length of each side of the hypercube. Comment on your answer.

Note: A hypercube is a generalization of a cube to an arbitrary number of dimensions. When $p = 1$, a hypercube is simply a line segment, when $p = 2$ it is a square, and when $p = 100$, it is a 100-dimensional cube.

---

a. With 1 predictor, restricting data to within 10% of the range of that predictor corresponds to 10% of all available observations, on average.

b. Since X_1 and X_2 are uniformly distributed in the unit interval, then we expect 10% of observations to have values of X_1 within a given interval of length 0.1, and similarly, 10% of observations to have values of X_2 within a given interval of length 0.1. Since X_1 and X_2 are independent, then we expect only about 1% of all observations to have values of X_1 and X_2 within their respective intervals.

c. Proceeding as in part b., note that for each $1 \leq i \leq 100$, only 10% of observations will have values of X_i in the appropriate interval of length 0.1. Since all predictors are independent, then only $0.1^{100} \times 100$ percent of observations will be available.

d. In order for two points to be "close" in p-dimensional Euclidean space, they need to be close in each of p-dimensions. As demonstrated above, the fraction of available observations that are "close" to a given point decays exponentially in p. Counteracting this therefore requires exponential increases in data sizes as p increases.

e. Conversely, in order to use 10% of the data in p-dimensional space, we must use a hypercube with sides of length $0.1^{1/p}$. But this quantity tends to 1 as p increases. In particular,

p	length
1	0.1
2	0.316
100	0.977

With 100 predictors, we must use 97.7% of the available range for each predictor!

Problem 2

Based on ISLR Exercise 4.6 and 4.8

Suppose we wish to build a classifier to automatically determine whether an email is spam. We collect data on the following predictors:

- X_1, an indicator for whether the email was addressed to more than 1 recipient.
- X_2, the number of people cc'd on the email
- X_3, the number of times a dollar sign appears in the email.
- X_4, an indicator for whether the word "urgent" appears in the subject line.

Based on a sample of 2000 emails, we build a logistic regression model with the following coefficients:

Coefficient	Value
β_0	-2.05
β_1	-1.91
β_2	0.02
β_3	-0.07

Coefficient	Value
β_4	2.66

a. Which features make an email more likely to be spam? Which features make an email less likely?

b. Estimate the probability that an email is spam , if it has 1 recipient, 2 people cc'd, 1 dollar sign, and urgent does not appear in the subject line.

c. How many people would need to be cc'd on the email in the previous part for the model to estimate that there is at least 50% probability the email is spam?

d. On average, what fraction of emails with odds of 0.37 of being spam are in fact spam?

e. Suppose that an email has 16% probability of being spam. What are the odds this email is spam?

---

a. Since the coefficients on β_1 and β_3 are negative, then having an email addressed to more than one recipient and having more dollar signs make an email less likely to be spam (in the presence of other variables). Conversely, since the coefficents on β_2 and β_4 are positive, then having a large number of people on the email and having urgent in the subject line makes the email more likely to be spam (again, in the presence of other variables).

b.

```
log_odds <- -2.05-1.91*0+.02*2-.07*1+2.66*0
log_odds
```

```
## [1] -2.08
```

```
prob <- exp(log_odds)/(1 + exp(log_odds))
prob
```

```
## [1] 0.111056
```

The log-odds of the given email being spam are -2.08 and the probability of the email being spam is 0.111056.

c. There is exactly a 50% chane that email is spam if the log-odds are 0. Setting the log-odds linear equation equal to 0 and solving for X_2 gives

$$X_2 = \frac{2.05 + 1.91 \cdot 0 + 0.07 \cdot 1 + 2.66 \cdot 0}{0.02} = \frac{2.12}{.02} = 106$$

d. If the odds of being spam are 0.37, then the probability of an email being spam is

$$\text{prob} = \frac{\text{odds}}{1 + \text{odds}} = \frac{.37}{1.37} = 0.27$$

e. If an email has probability 0.16 of being spam, the odds that it is spam are

$$\text{odds} = \frac{\text{prob}}{1 - \text{prob}} = \frac{.16}{.84} = 0.19$$

Applied

Probelm 3

The data set for this week comes from a study of the causes of civil wars, based on an exercise of Cosmo Shalizi's that uses data from Collier, Paul and Anke Hoeffler (2004). *Greed and Grievance in Civil War.* Oxford Economic Papers, 56: 563–595. URL: http://economics.ouls.ox.ac.uk/12055/1/2002-01text.pdf.

The data can be read into from a csv posted online by using the following command.

```
war <- read.csv("http://www.stat.cmu.edu/~cshalizi/uADA/15/hw/06/ch.csv", row.names = 1)
```

Every row of the data represents a combination of a country and of a five year interval — the first row is Afghanistan, 1960, really meaning Afghanistan, 1960–1965. The variables are:

- The country name;
- The year;
- An indicator for whether a civil war began during that period: 1 indicates a civil war has begun, the code of NA means an on-going civil war, 0 means peace.
- Exports, really a measure of how dependent the country's economy is on commodity exports;
- Secondary school enrollment rate for males, as a percentage;
- Annual growth rate in GDP;
- An index of the geographic concentration of the country's population (which would be 1 if the entire population lives in one city, and 0 if it evenly spread across the territory);
- The number of months since the country's last war or the end of World War II, whichever is more recent;
- The natural logarithm of the country's population;
- An index of social "fractionalization", which tries to measure how much the country is divided along ethnic and/or religious lines;
- An index of ethnic dominance, which tries to measure how much one ethnic group runs affairs in the country.

Some of these variables are NA for some countries.

Estimation

a. Fit a logistic regression model for the start of civil war on all other variables except country and year (yes, this makes some questionable assumptions about independent observations); include a quadratic term for exports. Report the coefficients and their standard errors, along with the p-values. Which ones are found to be significant at the 5% level?

Interpreation

All parts of this question refer to the logistic regression model you just fit.

b. What is the model's predicted probability for a civil war in India in the period beginning 1975? What probability would it predict for a country just like India in 1975, except that its male secondary school enrollment rate was 30 points higher? What probability would it predict for a country just like India in 1975, except that the ratio of commodity exports to GDP was 0.1 higher?

c. What is the model's predicted probability for a civil war in Nigeria in the period beginning 1965? What probability would it predict for a country just like Nigeria in 1965, except that its male secondary school enrollment rate was 30 points higher? What probability would it predict for a country just like Nigeria in 1965, except that the ratio of commodity exports to GDP was 0.1 higher?

d. In the parts above, you changed the same predictor variables by the same amounts. If you did your calculations properly, the changes in predicted probabilities are not equal. Explain why not. (The reasons may or may not be the same for the two variables.)

Confusion

Logistic regression predicts a probability of civil war for each country and period. Suppose we want to make a definite prediction of civil war or not, that is, to classify each data point. The probability of misclassification is minimized by predicting war if the probability is greater than or equal to 0.5, and peace otherwise.

e. Build a 2×2 *confusion matrix* which counts: the number of outbreaks of civil war correctly predicted by the logistic regression; the number of civil wars not predicted by the model; the number of false

predictions of civil wars; and the number of correctly predicted absences of civil wars. (Note that some entries in the table may be zero.)

f. What fraction of the logistic regression's predictions are incorrect, i.e. what is the misclassification rate? (Note that this is if anything too kind to the model, since it's looking at predictions to the same training data set).

g. Consider a foolish (?) pundit who always predicts "no war". What fraction of the pundit's predictions are correct on the whole data set? What fraction are correct on data points where the logistic regression model also makes a prediction?

h. Construct an ROC curve for your logistic regression model.

---

a.

```
cw_log_reg <- glm(start ~. -country -year + I(exports^2),
                  data = war, family = "binomial")

summary(cw_log_reg)$coefficients
```

```
##                         Estimate   Std. Error   z value     Pr(>|z|)
## (Intercept)       -1.307308e+01 2.795232e+00 -4.676920 2.912151e-06
## exports            1.893704e+01 5.865136e+00  3.228747 1.243336e-03
## schooling         -3.155633e-02 9.784271e-03 -3.225210 1.258804e-03
## growth            -1.152294e-01 4.307150e-02 -2.675305 7.466130e-03
## peace             -3.713408e-03 1.093156e-03 -3.396962 6.813847e-04
## concentration     -2.486984e+00 1.005201e+00 -2.474115 1.335665e-02
## lnpop              7.677375e-01 1.657549e-01  4.631763 3.625655e-06
## fractionalization -2.134524e-04 9.101928e-05 -2.345134 1.902023e-02
## dominance          6.703907e-01 3.535247e-01  1.896305 5.791973e-02
## I(exports^2)      -2.944321e+01 1.178128e+01 -2.499153 1.244907e-02
```

All predictors used in the model except `dominance` appear to be significant at the 5% level.

b.

```
India75 <- war %>% filter(country == "India", year == "1975")

predict(cw_log_reg, India75, type = "response")
```

```
##         1
## 0.3504199
```

The model predicts a probability 0.35 for civil war in India in 1975.

```
India75_school <- war %>% filter(country == "India", year == "1975") %>%
        mutate(schooling = schooling + 30)

predict(cw_log_reg, India75_school, type = "response")
```

```
##       1
## 0.17309
```

The model now predicts a probability of 0.173 for civil war in India in 1975, if schooling were 30 points higher.

```
India75_exports <- war %>% filter(country == "India", year == "1975") %>%
        mutate(schooling = exports + .1)

predict(cw_log_reg, India75_exports, type = "response")
```

```
##         1
## 0.6259445
```

The model now predicts a probability of 0.626 for civil war in India in 1975, if the ratio of exports to GDP was 0.1 higher.

c.

```
Nigeria65 <- war %>% filter(country == "Nigeria", year == "1965")

predict(cw_log_reg, Nigeria65, type = "response")
```

```
##         1
## 0.1709917
```

The model predicts a probability 0.17 for civil war in Nigeria in 1965.

```
Nigeria65_school <- war %>% filter(country == "Nigeria", year == "1965") %>%
        mutate(schooling = schooling + 30)

predict(cw_log_reg, Nigeria65_school, type = "response")
```

```
##          1
## 0.07410315
```

The model now predicts a probability 0.07 for civil war in Nigeria in 1965, if schooling was 30 points higher.

```
Nigeria65_exports <- war %>% filter(country == "Nigeria", year == "1965") %>%
        mutate(schooling = exports + 0.1)

predict(cw_log_reg, Nigeria65_exports, type = "response")
```

```
##         1
## 0.2034681
```

The model now predicts a probability 0.2 for civil war in Nigeria in 1965, if the ratio of exports to GDP was 0.1 higher.

d. The logistic regression model gives a non-linear relationship between predicted values and response. As a result, the amount the response changes when predictors are changed by a fixed amount may depend on the values of the predictors. Moreover, the log-odds are not even linear in `exports`, since we incorporated a quadratic export term in the model.

e.

Note we need to drop N/A values in order to make predictions.

```
war_no_na <- war %>% drop_na()

cw_probs<-predict(cw_log_reg, war_no_na, type = "response")
cw_preds<-ifelse(cw_probs >= 0.5, "1", "0")

library(yardstick)
```

```
## Warning: package 'yardstick' was built under R version 3.6.2
```

```
## For binary classification, the first factor level is assumed to be the event.
## Use the argument `event_level = "second"` to alter this as needed.
```

```
##
## Attaching package: 'yardstick'
```

```
## The following object is masked from 'package:readr':
##
##     spec
```

```
cw_results <- data.frame(obs = as.factor(war_no_na$start), preds = cw_preds, probs = cw_probs)
conf_mat(cw_results, truth = obs, estimate = preds)
```

```
##           Truth
## Prediction   0   1
##          0 637  43
##          1   5   3
```

f.

```
acc <- cw_results %>% accuracy(truth = obs, estimate = preds) %>% pull(.estimate)
1- acc
```

```
## [1] 0.06976744
```

The model has an error rate of 0.0697674 on the training set.

g. On the **whole** data set, the pundit's predictions have an error rate of 0.067

```
mean(war$start == 1, na.rm = T)
```

```
## [1] 0.06683805
```

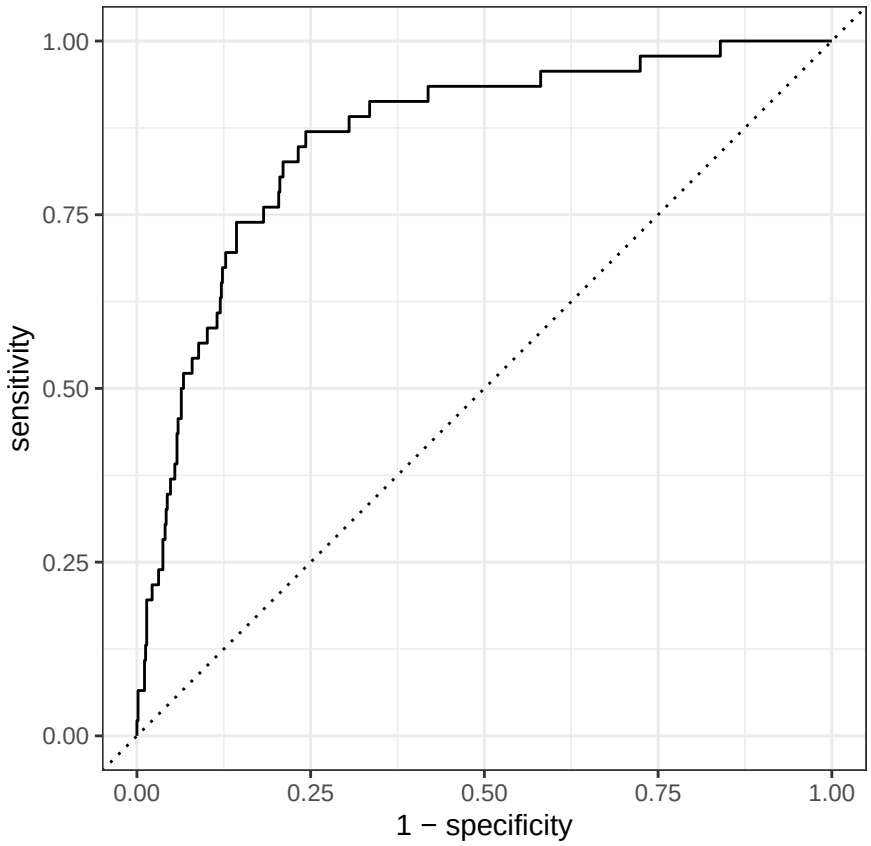
On the data set used to construct the model, the pundit's predictions also have an error rate of 0.067 (albeit slightly larger when rounded to more than 4 digits).

```
mean(war_no_na$start == 1, na.rm = T)
```

```
## [1] 0.06686047
```

h.

```
r <- roc_curve(cw_results, truth = obs, estimate = probs, event_level = "second" )
autoplot(r)
```




```
roc_auc(data = cw_results, truth = obs, probs, event_level = "second")
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.860
```
