Lab 5: The Sound of Gunfire, Off in the Distance

Alyssa Andrichik

10/13/2019

```
war <- read.csv("http://www.stat.cmu.edu/~cshalizi/uADA/15/hw/06/ch.csv", row.names = 1)
install.packages("ISLR")
library(ISLR)
library(tidyverse)
library(MASS)
nona_war <- na.omit(war)
```

1)Estimate: 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, together with R's p-values. Which ones are found to be significant at the 5% level?

```
logregwar <- glm(start ~ exports^2 + schooling + growth + peace + concentration + lnpop +
                   fractionalization + dominance, data = nona_war, family = binomial)
summary(logregwar)$coef
```

```
##                         Estimate   Std. Error   z value     Pr(>|z|)
## (Intercept)       -9.1752676450 2.260160e+00 -4.059565 4.916412e-05
## exports            3.2800933134 1.223303e+00  2.681341 7.332770e-03
## schooling         -0.0262049915 9.024171e-03 -2.903867 3.685849e-03
## growth            -0.1360349605 4.271830e-02 -3.184466 1.450214e-03
## peace             -0.0040688688 1.098627e-03 -3.703593 2.125673e-04
## concentration     -1.7643918630 9.256921e-01 -1.906025 5.664702e-02
## lnpop              0.5706386893 1.360989e-01  4.192825 2.755025e-05
## fractionalization -0.0001287699 8.353727e-05 -1.541466 1.232034e-01
## dominance          0.6020783978 3.513548e-01  1.713591 8.660387e-02
```

All of these predictors are significant at the 5% level because their p-values are so small.

2)Interpretation: All parts of this question refer to the logistic regression model you just fit.

a)What is the model's predicted probability for a civil war in India...

```
#Probability for civil war in India starting in 1975
pred1 <- nona_war[(nona_war$country=="India") & (nona_war$year==1975),]
signif(prob1 <- predict(logregwar,type="response",newdata=pred1),3)
```

```
##   500
## 0.423
```

```
#Male secondary school enrollment rate 30 points higher
pred2 <- pred1
pred2$schooling <-pred1$schooling+30
signif(prob2 <-predict(logregwar,type="response",newdata=pred2),3)
```

```
##  500
## 0.25
```

```
#Ratio of commodity exports to GDP is 0.1 higher
pred3 <- pred1
pred3$exports <- pred1$exports+0.1
signif(prob3 <-predict(logregwar,type="response",newdata=pred3),3)
```

```
##   500
## 0.504
```

b)What is the model's predicted probability for a civil war...

```
#Probability for civil war in Nigeria starting in 1965
NIGpred1 <- nona_war[(nona_war$country=="Nigeria") & (nona_war$year==1965),]
signif(NIGprob1 <- predict(logregwar,type="response",newdata=NIGpred1),3)
```

```
##   802
## 0.139
```

```
#Male secondary school enrollment rate 30 points higher
NIGpred2 <- NIGpred1
NIGpred2$schooling <-NIGpred1$schooling+30
signif(NIGprob2 <-predict(logregwar,type="response",newdata=NIGpred2),3)
```

```
##    802
## 0.0687
```

```
#Ratio of commodity exports to GDP is 0.1 higher
NIGpred3 <- NIGpred1
NIGpred3$exports <- NIGpred1$exports+0.1
signif(NIGprob3 <-predict(logregwar,type="response",newdata=NIGpred3),3)
```

```
##   802
## 0.184
```

c)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.)

For India, the probability of civil war during the 1975 period was 0.423. The probability of civil war in India, after increasing the male secondary school enrollment rate by 30, decreases to 0.25. The probability of civil war in India, after increasing the ratio of commodity exports to GDP is by 0.1, increase to 0.504.

For Nigeria, the probability of civil war during the 1965 period was 0.139. The probability of civil war in Nigeria, after increasing the male secondary school enrollment rate by 30, decreases to 0.0687. The probability of civil war in India, after increasing the ratio of commodity exports to GDP is by 0.1, increases to 0.184.

Changing the specific predictors mentioned (schooling and exports) caused the likelihood of a civil war breaking out in both India and Nigeria to change differently. The change in the schooling variable caused, in both countries, a decrease in the probability of a civil war starting. The change in the export variable caused, in both countries, an increase in the probability of a civil war. This can mean that increased education for citizens plays a significant role in both countries when it comes to preventing the likelihood of civil war, while increasing the ratio of commodity exports to GDP plays a significant role in increasing the likelihood of a civil war.

During these different time periods in these two different countries, it seems that India was dealing with more political tension overall since the probability of civil war was higher every scenario than Nigeria's probability

of civil war. I say political tension in the sense that there are other factors that are increasing the likelihood of the start of a civil war that I included in my regression, but are not the predictors I focused on. Those other predictors are the reason why India's probability is higher than Nigeria's.

3)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.

a)Build a 2 × 2 confusion matrix (a.k.a. "classification table" or "contigency table") 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.)

```
#Classification Table
my_log_pred <- ifelse(logregwar$fit >= 0.5, "No", "Yes")
conf_log <- table(nona_war$start, my_log_pred)
conf_log
```

```
##    my_log_pred
##      No Yes
##   0   3 639
##   1   3  43
```

1 indicates a civil war has begun, the code of NA means an on-going civil war, 0 means peace. Yes means war, no means peace.

b)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).

```
#Missclassification Rate
(1/nrow(nona_war)) * (conf_log[2, 1] + conf_log[1, 2])
```

```
## [1] 0.9331395
```

c)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?

TPR = TP/(TP+FN)

TPR = 642/(642+46)

TPR = 0.933

4)Comparison: Since this is a classification problem with only two classes, we can compare Logistic Regression right along side Discriminant Analysis.

a)Fit an LDA model using the same predictors that you used for your logistic regression model. What is the training misclassification rate?

```
lda_war <- lda(start ~ exports^2 + schooling + growth + peace + concentration + lnpop +
                 fractionalization + dominance, data = nona_war)
my_lda_pred <- predict(lda_war, nona_war)

#Classification Table
conf_lda <- table(nona_war$start, my_lda_pred$class)
conf_lda
```

```
##
##       0   1
##   0 638   4
##   1  40   6
```

```
#Missclassification Rate
(1/nrow(nona_war)) * (conf_lda[2, 1] + conf_lda[1, 2])
```

```
## [1] 0.06395349
```

b)Fit a QDA model using the very same predictors. What is the training misclassification rate?

```
qda_war <- qda(start ~ exports^2 + schooling + growth + peace + concentration + lnpop +
                 fractionalization + dominance, data = nona_war)
my_qda_pred <- predict(qda_war, nona_war)

#Classification Table
conf_qda <- table(nona_war$start, my_qda_pred$class)
conf_qda
```

```
##
##       0   1
##   0 628  14
##   1  30  16
```

```
#Missclassification Rate
(1/nrow(nona_war)) * (conf_qda[2, 1] + conf_qda[1, 2])
```

```
## [1] 0.06395349
```

c)How does the prediction accuracy of the three models compare? Why do you think this is?

The misclassification rate of the log regression was 0.9331395.

The misclassification rate of the lda regression was 0.06395349.

The misclassification rate of the qda regression was 0.06395349.

Both the qda and the lda models were far more accurate than the log regression since they had smaller misclassification rates. The qda and the lda model have the same misclassification rate, which means that they are equally accurate in predicting. The qda and lda models fit the data better and follow more clear trends, both misclassifying the same amount of data points.

Problem Set

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 will now 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 (evenly) 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=0.6, 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?

It looks like it should be 10% of the available observations because it is between [0.05, 0.95] leaving 0.1 of the observations out which is, again, 10% of the observations.

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]×[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_2 closest to that test observation. 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?

Since X_1 and X_2 are "uniformly distributed," and assuming they are also independent, we can assume that it is $0.1 \times 0.1 = 0.01$ or 1%.

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 the 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?

$(0.1)^p = (0.1)^{100} =$ ~0.0% This means that nearly none of the observations will be available 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.

When p=1, ~10% of the available observations could be used to make the prediction. When p=2, ~1% of the available observations could be used to make the prediction. When p=100, ~0% of the available observations could be used to make the prediction. This clearly shows that when p is large, there are way fewer training observations "near" any given test observation which will lead to a very poor prediction and thus a poor KNN fit.

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

For p = 1, the length = $0.1^{1/1}$. For p = 2, the length = $0.1^{1/2}$. For p = 100, the length = $0.1^{1/100}$. We know that each hypercube has an area of 10% of the training observations, so as p gets larger, and adds more and more dimensions or sides, we are still looking at 10% of the training observations but only at the resulting length of one side. If we multiply the length of all the sides of the hypercube for each p, we would get 0.1

6) Suppose we collect data for a group of students in a statistics class with variables X_1= hours studied, X_2= undergrad GPA, and Y= receive an A. We fit a logistic regression and produce estimated coefficients, β_0 = -6, β_1 = 0.05, β_2 = 1.

a) Estimate the probability that a student who studies for 40 hours and has an undergrad GPA of 3.5 gets an A in the class.

phat(X) =

$(e^{-6+0.05X_1+X_2})/(1+e^{-6+0.05X_1+X_2}) =$

$(e^{-6+0.05(40hrs)+3.5GPA})/(1+e^{-6+0.05(40hrs)+3.5GPA}) =$

0.3775

There is a 37.75% probability that this described student will get an A in the class.

b) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?

$(e^{-6+0.05X_1+3.5GPA})/(1+e^{-6+0.05X_1+3.5GPA}) = 0.5$ ->

$e^{06+0.05X1+3.5GPA} = 1$ ->

$X_1 = (2.5/0.05)$ ->

$X_1 = 50$

The student would need to study 50 hours to have a 50% of getting an A.

7) Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on X, last year's percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was $\bar{X}$ = 10, while the mean for those that didn't was $\bar{X}$ = 0. In addition, the variance of X for these two sets of companies was $\sigma_h at^2$ = 36. Finally, 80% of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage return was X = 4 last year.

$p_k(x) = (\pi_k/(\sqrt{(}2\pi\sigma)e^{-(x-\mu_k)^2/(2\sigma^2)})/(\Sigma_l^K(\pi_l/(\sqrt{(}2\pi\sigma)e^{-(x-\mu_l)^2/(2\sigma^2)})$

$p(4) = 0.8e^{-(1/72)(4-10)^2}/(0.8e^{-(1/72)(4-10)^2} + 0.2e^{-(1/72)(4-0)^2}) = 0.752$

There is a 75.2% chance that a company will issue a dividend this year.