Homework 3

Instructions

Due: 1:35pm on Wednesday, October 7th

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 2.7 The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Obs.	X1	X2	X3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

a. Compute the Euclidean distance (i.e. distance in 3d space) between each observation and the test point.

b. What is our prediction with $K = 1$? Why?

c. What is our prediction with $K = 3$? Why?

d. If the Bayes decision boundary in this problem is highly non-linear, then would we expect the *best* value for K to be large or small? Why?

Problem 2

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 observations. 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.

Problem 3

Based on ISLR Exercise 4.6 and 4.8

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 = recieve an A. We fit a logistic regression and produce estimated coefficients $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$.

a. Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets 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?

c. On average, what fraction of students with odds of 0.37 of passing will in fact pass?

d. Suppose that a student has a 16% chance of passing the class. What are the odds this student will pass?

Applied

Problem 4

Based on ISLR Exercise 4.10

Load the `Weekly` data set from the `ISLR` package:

This data consists of percentage weekly returns for the S&P 500 stock index over 1089 weeks from the beginning of 1990 to the end of 2010. For each week, the percentage returns were recorded for the 5 previous weeks (`Lag 1` through `Lag 5`). Additionally, the following variables are recorded: `Volume` (number of shares traded the previous week (in billions), `Today` (the percentage return on the week in question), and `Direction` (whether the market was `Up` or `Down` on this week).

a. Produce some numerical and graphical summaries of the `Weekly` data. What patterns do you see?

b. Use the full data set to perform a logistic regression with `Direction` as the response and the five lag variables and `Volume` as predictors. Print the results using the `summary` function. Do any of the predictors appear to be statistically significant? If so, which ones?

c. Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by this logistic regression.

d. Now fit the logistic regression model using a training data period from 1990 to 2008, with `Lag2` as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the test data (from 2009 and 2010).

e. Repeat part d using KNN with K = 1.

f. Experiment with different combinations of predictors, including possible transformations and interactions Logistic Regression, as well as with different values of K for KNN. Report the variables, method, and associated confusion matrix that appears to provide the best results on the test data.

Problem 5

The second 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, together with R's 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.