Analytics 512: Midterm Exam

March 1, 2016

Print Your Name:

Five problems for 200 points plus a Bonus problem. Write all answers directly on these pages. Continue on the back if necessary.

Closed notes, closed book. Calculators are allowed. 60 minutes.

Give yourself $\frac{x}{4}$ minutes to do a problem with x points.

1. (40 points)

In each of the following supervised learning situations, state whether this is a prediction or classification problem, define a response (as clearly as possible), and list three possible predictors.

(a) Determine if a news article is about politics, sports, the economy, or something else.

(b) Predict the future salary of a new college graduate.

(c) Assess the risk of future heart disease of an individual from self-reported data.

(d) Predict tomorrow's amount of rainfall in Washington, DC.

2. (40 points)

Refer to the explanation and the plots on page 7 and the regression model output on page 8.

(a) Suppose you want to predict the waiting time until the next eruption from the previous waiting time (`previous.waiting`) and nothing else, using simple linear regression. Will this be a significant predictor? Will the slope coefficient be positive or negative? Will its magnitude be > 1 or < 1? Explain your answers.

(b) Consider the regression output on page 8. Write down the form of the model that has been fitted. Then write down the model with estimated coefficients.

(c) One of the predictors could be dropped. Which one? Why? Can you say what will happen to the residual standard error when this is done? Can you say what will happen to the significance levels of the other predictors when this is done?

3. (40 points)

This problem continues with the Old Faithful data (see the previous problem). A logistic regression model has been fitted to the data to predict whether a waiting time is > 68 minutes(`long.wait = T`) or not (`long.wait = F`).

(a) The output is given on page 8. List the predictors and state as clearly as possible what the response is.

(b) Write down the formula for the log-odds that comes from the fitted model.

(c) Explain in words to somebody who knows no statistics what it means that the coefficient for `eruptions` is positive and that the coefficient for `previous.waiting` is negative.

4. (40 points)

This problem continues with the Old Faithful data (see problem 2). A linear discriminant model has been fitted to the data to predict whether a waiting time is > 68 minutes(`long.wait = T`) or not (`long.wait = F`). See page 9 for details.

(a) What are the predictors in this model?

(b) The three confusion matrices on page 9 are obtained from the full model (all data used for training) and from two validation sets. What are the sizes of the two training sets that were used for the validation attempts?

(c) Which of the two validation sets is likely to have the highest variability (i.e. will give a different result each time this is repeated)? Explain your answer.

(d) Which of the three confusion matrices is likely to be most biased towards a low prediction error? Explain your answer.

5. (40 points) True or false? Give a one line explanation for each answer.

(a) A nonsignificant predictor in a multiple linear regression model should always be dropped from the model.

(b) If there are strongly correlated predictors in a multiple linear regression model, the most serious problem will be overfitting.

(c) If leave-one-out cross validation is done to assess a model, the reported error estimates will be very reliable, since the training models use almost all data and the sample size is very large (low bias and low variance).

(d) Suppose you assess the model error with a validation set that has 30% of the full data. Then the estimated error is likely to be larger than the error of the full model for future data points.

(e) Suppose you assess the model error with 5 fold cross validation. Then the estimated error is likely to be smaller than the error of the full model for future data points.

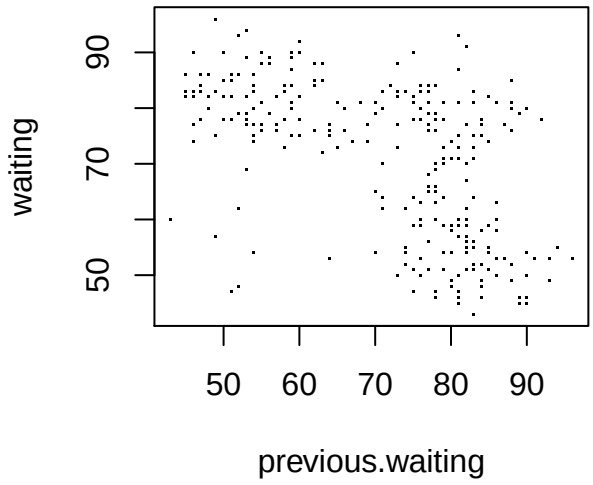
Bonus (20 points)

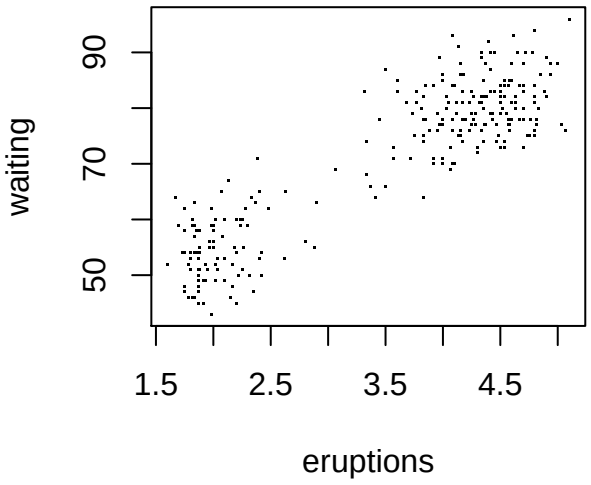
Explain in one paragraph why we do not expect to get information about the significance of predictors from Linear Discriminant Analysis.

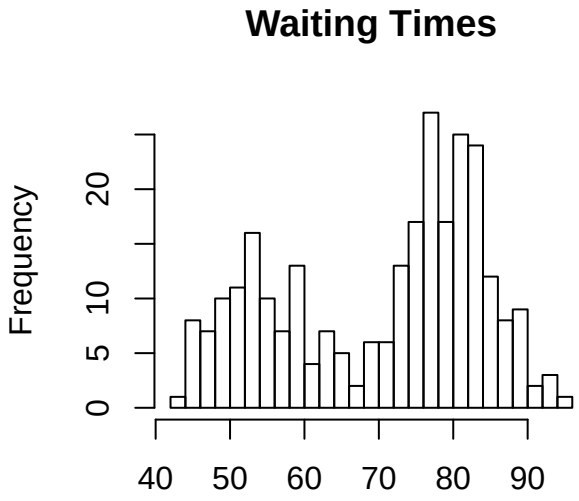
Old Faithful Data

The data for this problem come from the Old Faithful geyser in Yellowstone National Park. The waiting time until the previous eruption, the duration of that eruption, and the waiting time until the next eruption have been recorded in a dataset.

Plots of waiting times against previous waiting time, waiting time against previous eruption length, and a histogram of waiting times.







Output for Problem 2

```
##
## Call:
## lm(formula = waiting ~ eruptions + previous.waiting + eruptions *
##     previous.waiting, data = faithful1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.2200  -4.3259   0.2993   3.6244  16.2678
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                54.34630    8.58375   6.331 1.02e-09 ***
## eruptions                   6.34501    2.09681   3.026  0.00272 **
## previous.waiting           -0.26043    0.10917  -2.386  0.01775 *
## eruptions:previous.waiting  0.05387    0.02778   1.940  0.05348 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.854 on 267 degrees of freedom
## Multiple R-squared:  0.817,  Adjusted R-squared:  0.815
## F-statistic: 397.5 on 3 and 267 DF,  p-value: < 2.2e-16
```

Output for Problem 3

```
##
## Call:
## glm(formula = long.wait ~ . - waiting, family = binomial, data = faithful2)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.40447  -0.04849   0.02902   0.09385   2.94497
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -12.54105    4.72843  -2.652    0.008 **
## eruptions          4.93578    0.89127   5.538 3.06e-08 ***
## previous.waiting  -0.04270    0.04898  -0.872    0.383
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 357.923  on 270  degrees of freedom
## Residual deviance:  35.539  on 268  degrees of freedom
## AIC: 41.539
##
## Number of Fisher Scoring iterations: 8
```

Output for Problem 4

```
## Call:
## lda(long.wait ~ eruptions + previous.waiting + previous.waiting *
##     eruptions, data = faithful2)
##
## Prior probabilities of groups:
##     FALSE      TRUE
## 0.3726937 0.6273063
##
## Group means:
##       eruptions previous.waiting eruptions:previous.waiting
## FALSE  2.106594         79.50495                   167.2994
## TRUE   4.307712         65.76471                   280.9591
##
## Coefficients of linear discriminants:
##                                     LD1
## eruptions                    1.27949867
## previous.waiting            -0.05413774
## eruptions:previous.waiting   0.01681021
```

Confusion Table for Problem 4 (full data set)

```
##
##         FALSE TRUE
##   FALSE    96    5
##   TRUE      2  168
```

Confusion Table for Problem 4 (validation set)

```
##
##         FALSE TRUE
##   FALSE    15    4
##   TRUE      0   36
```

Confusion Table for Problem 4 (another validation set)

```
##
##         FALSE TRUE
##   FALSE    43    4
##   TRUE      2   87
```

Scratch

Scratch