Analytics 512: Final Exam II

May 16, 2016

Solution Key

The first two problems use artificial data in a data frame with predictors $X_1, \ldots, X_6$ and a response Y. The predictors can have the following values.

$$X_1 \in \{-1, 1\}, \quad X_2,\, X_3, \ldots, X_6 \in \{-1, 0, 1\}$$

All X_i are independent and have uniform distributions, i.e. $P(X_1 = 1) = P(X_1 = -1) = \frac{1}{2}$, $P(X_2 = -1) = P(X_2 = 0) = P(X_2 = 1) = \frac{1}{3}$ and so on. The responses Y are generated by the formula

$$Y = \begin{cases} -3 + 4X_2 - 2X_3 + \varepsilon & \text{if} \quad X_1 = -1 \\ +3 - 3X_4 + X_5 + \varepsilon & \text{if} \quad X_1 = 1 \end{cases}$$

The ε are standard normal random variables and independent of everything else.

The problems also use a test set with $M = 200$ observations that have been generated with the same formula.

1. (40)

A training set with $N = 50$ observations has been created and a linear model has been fitted. The summary is given on the next page. Answer the following questions, with short explanations.

a) If the number of observations in the training set is increased to $N = 200$, how do you expect the F-statistic to change?

The F statistic is expected to increase.

b) If the number of observations in the training set is increased to $N = 200$, how do you expect the residual standard error to change? Strong or modest increase? Strong or modest decrease? No change? Small random change?

The RSS and the number of degrees of freedom will both increase, so the residual standard error is expected to show some modest random change.

c) According to this model, the intercept is not significant. Why is that so? Isn't there an intercept in the formula for Y?

The generating model has an intercept of +3 and -3 each with probability 1/2. These terms cancel out on average, hence the intercept is expected to be zero.

d) Predictions have been made with the test set. The root mean square prediction error is estimated to be 2.702, which is substantially higher than the residual standard error in the summary. What is the reason?

This is most likely due to overfitting.

```
## 
## Call:
## lm(formula = Y ~ ., data = mydf.train)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.7125 -1.1048  0.0128  1.5344  4.4508 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  0.37921    0.33435   1.134   0.2630    
## X1           2.13072    0.29979   7.107 9.00e-09 ***
## X2           1.34983    0.38584   3.498   0.0011 ** 
## X3          -0.74013    0.37457  -1.976   0.0546 .  
## X4          -1.73223    0.38390  -4.512 4.91e-05 ***
## X5          -0.22670    0.39416  -0.575   0.5682    
## X6           0.05222    0.37730   0.138   0.8906    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.047 on 43 degrees of freedom
## Multiple R-squared:  0.6626, Adjusted R-squared:  0.6156 
## F-statistic: 14.08 on 6 and 43 DF,  p-value: 8.682e-09
```

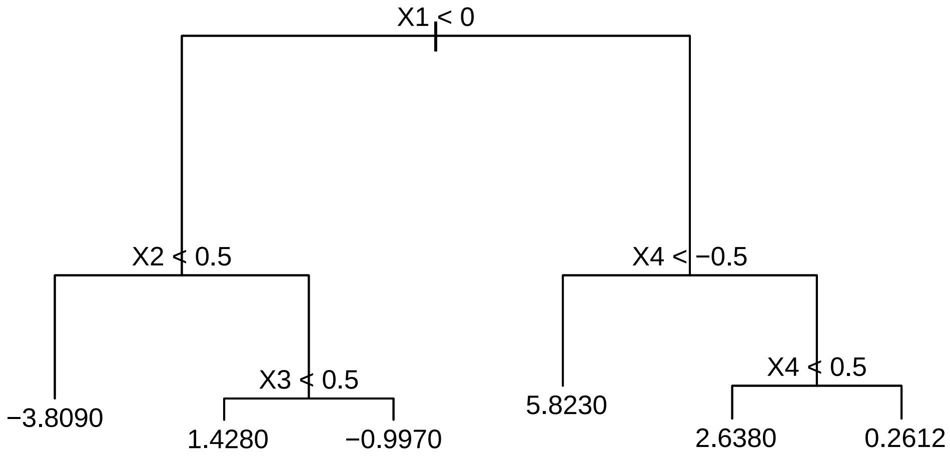
2. (40)

A tree has been fitted to the same training data as in the previous problem. Here is its description.

```
tree1
```

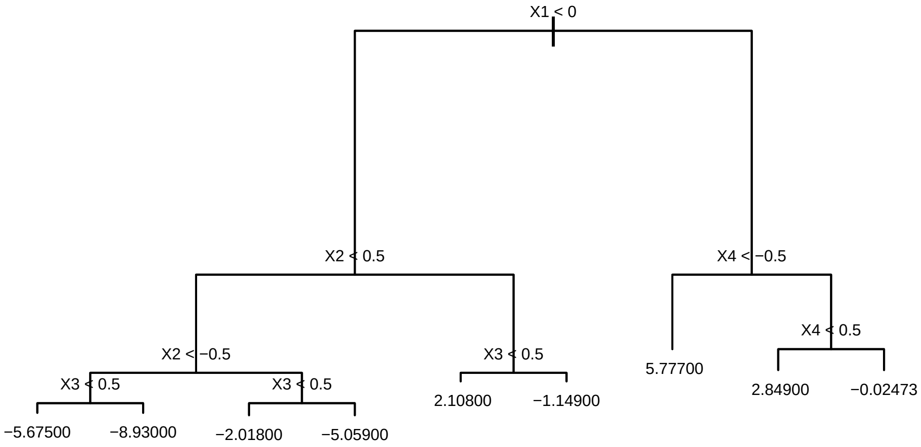
```
## node), split, n, deviance, yval
##       * denotes terminal node
## 
##  1) root 50 534.100  0.8096  
##    2) X1 < 0 23 173.200 -1.3760  
##      4) X2 < 0.5 10  40.310 -3.8090 *
##      5) X2 > 0.5 13  28.180  0.4953  
##       10) X3 < 0.5 8   5.768  1.4280 *
##       11) X3 > 0.5 5   4.320 -0.9970 *
##    3) X1 > 0 27 157.400  2.6720  
##      6) X4 < -0.5 7   7.926  5.8230 *
##      7) X4 > -0.5 20  55.670  1.5690  
##       14) X4 < 0.5 11  10.510  2.6380 *
##       15) X4 > 0.5 9  17.190  0.2612 *
```

a) Sketch the decision tree and label its splits and all terminal nodes.



b) Suppose the size of the training set is increased to $N = 200$. How do you expect the shape of the tree to change? Your answer should be brief.

We expect the tree To grow in complexity (more branches, more leaves). Here's an example.



c) Predictions have been generated with the test set. The root mean square prediction error is estimated to be 1.947. How is that expected to change if the number of observations in the training set is increased to $N = 200$? Strong or modest increase? Strong or modest decrease? No change? Small random change?

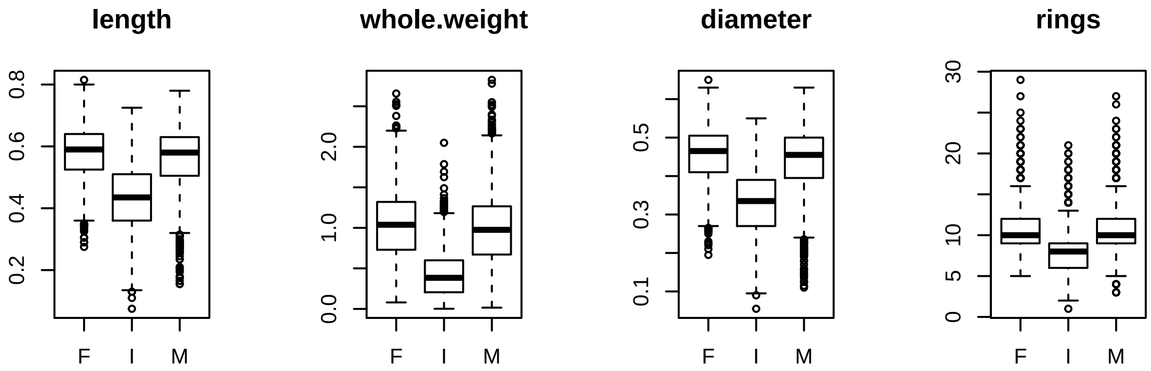
The mean square prediction error will most likely decrease. A tree is suitable for the data that are generated here. For example, for the tree in the picture above, the rms prediction error is estimated to be 1.299.

Problems 3 and 4 are based on the `Abalone` dataset that was discussed in class. The dataset contains observations on 4177 specimens of abalone (a shellfish), namely

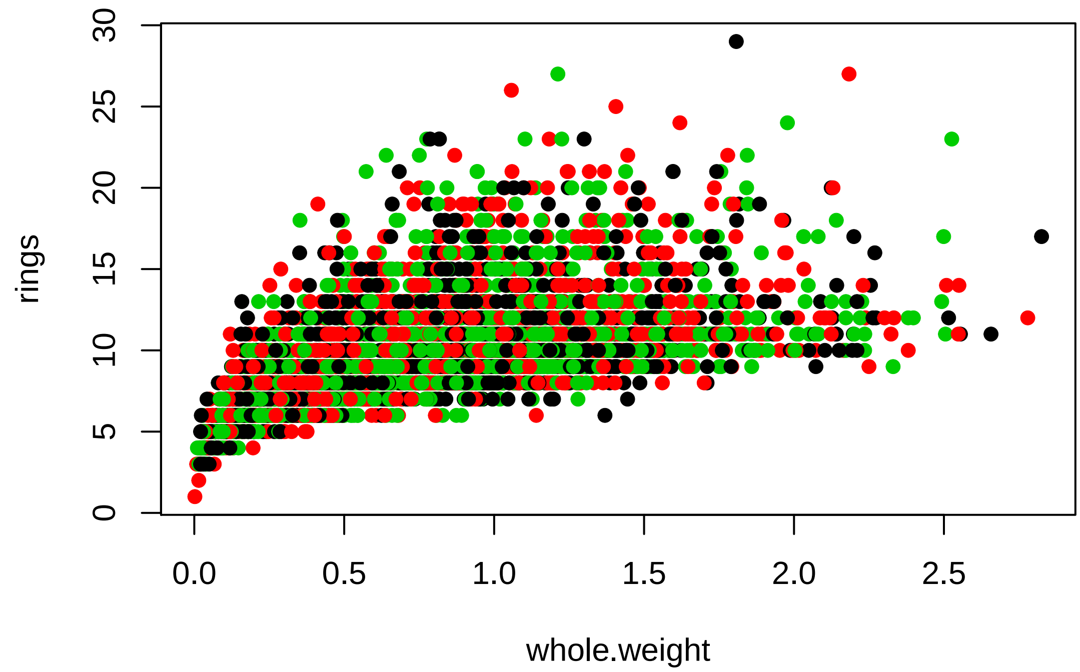
sex, length, diameter, height, whole.weight, shucked.weight, viscera.weight, shell.weight, rings.

The variable sex can have the values *M = male, F = female, I = immature.* The other variables are numerical. There are between 1,300 and 1,400 observations for each of the three sex values. The last variable, *rings*, is essentially the age of the animal and is the response.

Here are boxplots of some of these variables by sex.



Here is a plot of rings against whole.weight, colored by sex (black = F, red = I, green = M).



3. (40)

a) Explain why the boxplots on the previous page imply that length and whole.weight are correlated. Use a sketch if necessary.

The box plots show that for male and female specimens, both length and whole.weight tend to be larger than for immature specimens. This will result in a positive correlation.

b) It has been proposed to predict **rings** by fitting separate linear models to the populations of male, female, and immature abalones. Then a prediction could be made by using the linear model that is appropriate for the sex of this specimen.

An alternative is to fit a tree to the entire dataset. The tree would then presumably split along the three values of the variable sex.

Which approach can be expected to give better predictions? Explain your answer.

Fitting separately linear models to the three populations results in more flexibility, since linear models may be more appropriate for each population than trees. This may give better protection.

c) We want to predict rings from whole.weight. Use the plot on the previous page to suggest a nonlinear transformation in order to improve this prediction. Give an explanation.

The shape of the response suggests that $\sqrt{rings}$ *may be better described by a linear model. So the response could be transformed by taking the square. Alternatively a quadratic or cubic model could be used.*

4. (40)

A generalized additive model has been fitted to the data. Here is the summary.

```
##
## Call: gam(formula = rings ~ sex + poly(whole.weight, 2) * poly(length,
##     2), data = abalone)
## Deviance Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7868 -1.6362 -0.5490  0.9248 16.5707
##
## (Dispersion Parameter for gaussian family taken to be 6.562)
##
##     Null Deviance: 43410.63 on 4176 degrees of freedom
## Residual Deviance: 27337.47 on 4166 degrees of freedom
## AIC: 19725
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##                                          Df  Sum Sq Mean Sq  F value
## sex                                       2  8381.1  4190.6 638.6053
## poly(whole.weight, 2)                     2  6616.3  3308.2 504.1350
## poly(length, 2)                           2  1017.5   508.8  77.5319
## poly(whole.weight, 2):poly(length, 2)    4    58.2    14.5   2.2173
## Residuals                              4166 27337.5     6.6
##                                          Pr(>F)
## sex                                    < 2e-16 ***
## poly(whole.weight, 2)                  < 2e-16 ***
## poly(length, 2)                        < 2e-16 ***
## poly(whole.weight, 2):poly(length, 2) 0.06465 .
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a) Should the variable `sex` be included in the model? Explain your answer.

Yes, this variable appears to be strongly significant. This can be seen from the F statistic and also from the sum of squares.

b) Suppose two abalone specimens (one male, one female) have the same values for `whole.weight` and `length`. How do their predictions for `rings` differ? Be as specific as possible.

Since `sex` *enters as a single variable in the additive model, predictions for these two specimens would differ by fixed amount, which does not depend on any of the other variables.*

c) List all interaction terms explicitly, in the form *height* × *diameter*, *viscera.weight* × *shell.weight*^3 etc.

These are all products of `length` *and* `whole.weight` *and of their squares. Here is a complete list:*

length * whole.weight, length^2 * whole.weight, length * whole.weight^2, length^2 * whole.weight^2

There are four such terms altogether (4 degrees of freedom).

d) The summary suggests that the model should be simplified. State and explain what the simplification should be and then write down a formula for the simplified model, with all predictors including powers if appropriate.

Leave out the interaction terms. The R formula for the simplified model is

```
fit = gam(rings ~ sex + poly(whole.weight,2)+ poly(length,2), data = abalone)
```

5. (40)

True or false? Write a one line explanation for each answer.

- Leave one-out cross validation is generally not an effective way of selecting the right k in k-Nearest neighbors prediction.

True, since this is computationally very expensive.

- When the number of variables is a small fraction (say $< 1\%$) of the number of observations, one should use best subset selection to select variables.

False, since the number of variables could still be very large.

- When the predictors in a linear model are scaled, the value of R^2 may increase.

False, the coefficientss will change, but all predictions will remain the same.

- Random forests are obtained by simply applying bagging to a tree method.

False, there is also a random selection of predictors for each random tree.

- The purpose of ridge regression is usually to decrease the bias of a prediction.

False, the purpose is to decrease the variance. The bias may increase in the process.

Bonus (20)

Refer to problems 1 and 2. Approximately how many leaves does an optimal tree have in this situation?

Since X_1 can have two values, $X_2, \ldots, X_5$ can each have three values, and due to the structure of formula for the data, an optimal tree could have about $3^2 + 3^2 = 18$ leaves.