Homework 2

JEREMY KNOX PSTAT 231, W19

February 10, 2019

K-Nearest Neighbor Method

Problem 1: Cross Validation 10 times with do chunk for each nieghbor (K) to find best K

train.error <dbl>	val.error <dbl>	neighbor <dbl>
0.0003086420	0.09695291	1
0.0000000000	0.10555556	1
0.0003085467	0.08611111	1
0.0003085467	0.13055556	1
0.0000000000	0.09166667	1
0.0000000000	0.11388889	1
0.0003085467	0.08611111	1
0.0006170935	0.10555556	1
0.0006170935	0.09444444	1
0.0006170935	0.10277778	1

1-10 of 60 rows Previous **1** 2 3 4 5 6 Next

neighbor <dbl>	train.error <dbl>	val.error <dbl>
1	0.0003085563	0.1013620
10	0.0835569535	0.0991382
20	0.0950353305	0.1046953
30	0.1035823515	0.1138596
40	0.1137956297	0.1196914

50	0.1180536605	0.1221891

6 rows

```
## [1] 10
```

BEST K = 10

Problem 2: Find error on optimal K

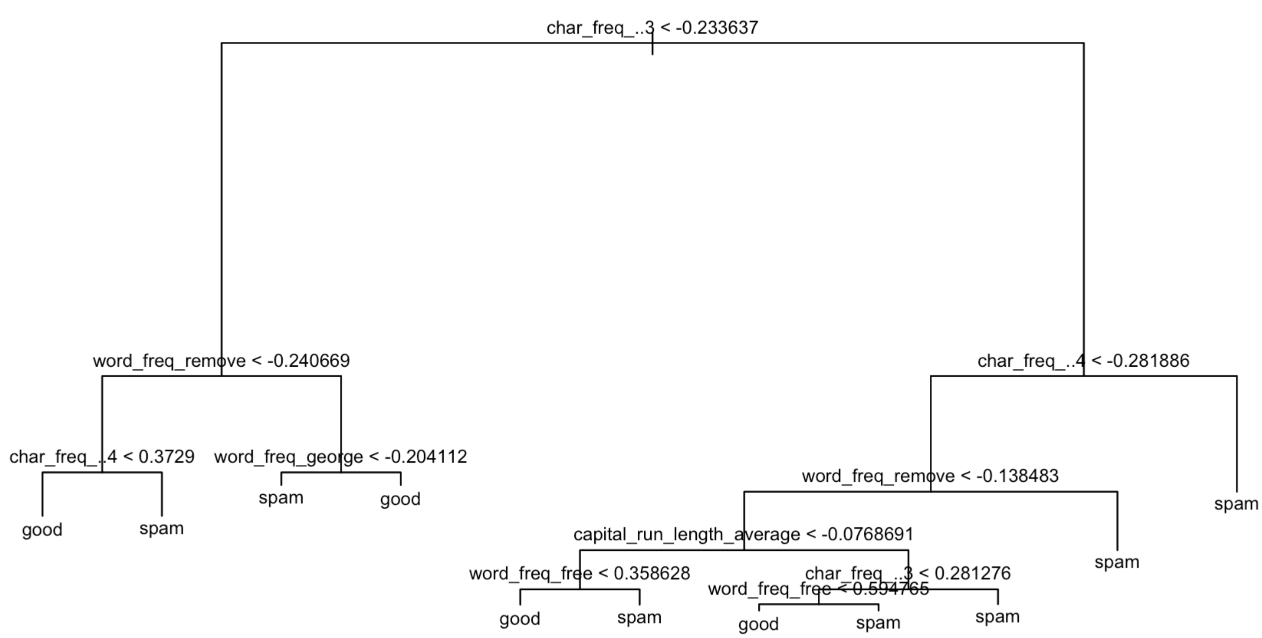
```
##          train.error test.error
## knn       0.07914468      0.094
## tree              NA         NA
## logistic          NA         NA
```

Decision Tree Method

Problem 3: Make Tree

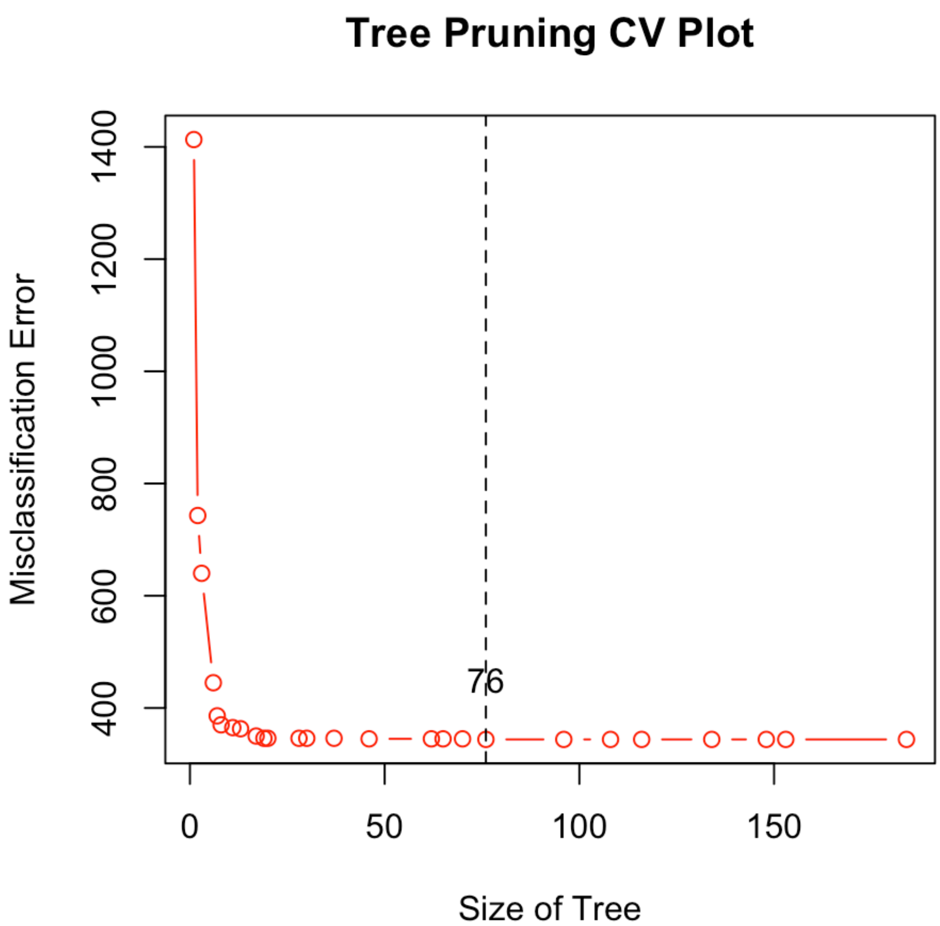
Terminal Nodes 184
Misclassified Training Observations 48

Problem 4: Pruning the Tree



Problem 5: 10-Fold Cross Validation on Tree. Size vs. Misclassification Error

```
## [1] 76
```



Problem 6: Training and Test Errors

```
##            train.error test.error
## knn         0.07914468      0.094
## tree        0.02943627      0.061
## logistic            NA         NA
```

Logisitc Regression

Problem 7:

a. Logit Function

Let $p(z) = \frac{e^z}{1+e^z} = p$, then

$$\begin{aligned} e^z &= p + pe^z \\ e^z(1-p) &= p \\ e^z &= \frac{p}{1-p} \\ z &= ln(\frac{p}{1-p}) \end{aligned}$$

Thus, $ z(p)=ln()$.

b. Link Function

Assume $z = \beta_0 + \beta_1 x_1$ and $p = logistic(z)$ from above and $\text{odds: } \frac{p}{1-p}$, then

$$\begin{aligned} \frac{p}{1-p} &= \frac{\frac{e^z}{1+e^z}}{1 - \frac{e^z}{1+e^z}} \\ &= \frac{\frac{e^z}{1+e^z}}{\frac{1}{1+e^z}} \\ &= e^z \end{aligned}$$

Which implies $\text{odds } = e^{\beta_0} e^{\beta_1 x_1}$

Let $2x_1$, then we have

$$\begin{aligned} &= \frac{e^{\beta_0} e^{\beta_1 (x_1+2)}}{e^{\beta_0} e^{\beta_1 x_1}} \\ &= e^{2\beta_1} \end{aligned}$$

Thus a two times increase of x_1 gives us $2x_1 \implies e^{2\beta_1}$.

For $\beta_1 < 0$, what does p approach as $x_1 \to \infty$? We have $p = \frac{e^{\beta_0} e^{\beta_1 x_1}}{1+e^{\beta_0} e^{\beta_1 x_1}}$

$$\begin{aligned} \lim_{x_1 \to \infty} p &= \frac{\lim_{x_1 \to \infty} e^{\beta_0} e^{\beta_1 x_1}}{1 + \lim_{x_1 \to \infty} e^{\beta_0} e^{\beta_1 x_1}} \\ &= \frac{0}{1+0} \\ &= 0 \end{aligned}$$

Thus we have, p approaches 0 as $x_1 \to \infty$.

For $\beta_1 < 0$, what does p approach as $x_1 \to -\infty$? We have $p = \frac{e^{\beta_0} e^{\beta_1 x_1}}{1+e^{\beta_0} e^{\beta_1 x_1}}$

$$\lim_{x_1 \to \infty} p = \frac{\lim_{x_1 \to \infty} e^{\beta_0} e^{\beta_1 x_1}}{\lim_{x_1 \to \infty} 1 + e^{\beta_0} e^{\beta_1 x_1}}$$
$$= \frac{\infty}{\infty}$$

Apply L'Hospital $\lim_{x \to c} \frac{f(x)}{g(x)} = \lim_{x \to c} \frac{f'(x)}{g'(x)}$:

$$\lim_{x_1 \to \infty} p = \lim_{x_1 \to \infty} \frac{e^{\beta_0} \beta_1 e^{\beta_1 x_1}}{e^{\beta_0} \beta_1 e^{\beta_1 x_1}}$$
$$= \lim_{x_1 \to \infty} 1$$
$$= 1$$

Thus, p approaches 1 as $x_1 \to -\infty$.

Problem 8: Classify with Logistic and obtain Training and Test Error

```
##                  train.good.real
## train.good.pred FALSE TRUE
##           FALSE  2087  154
##           TRUE    101 1259
```

```
## [1] 0.07081366
```

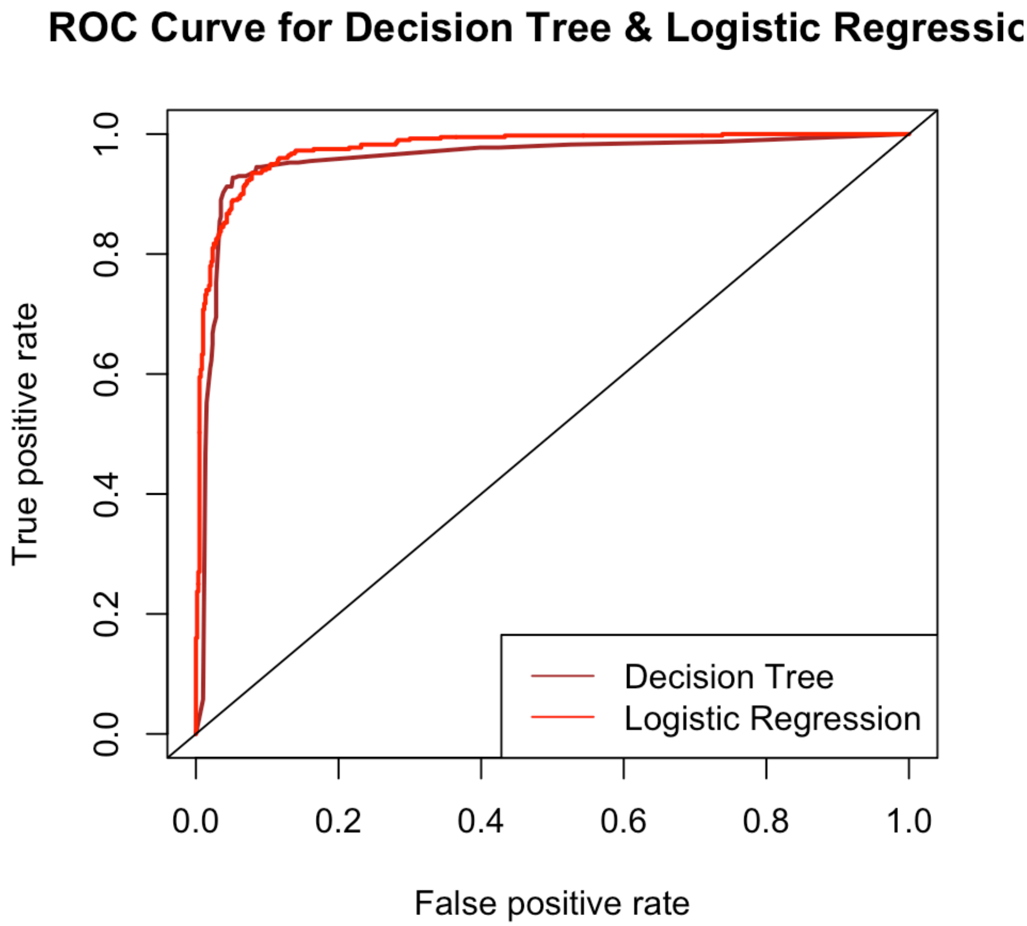
```
##                 test.good.real
## test.good.pred FALSE TRUE
##          FALSE   574   55
##          TRUE     26  345
```

```
## [1] 0.081
```

```
##          train.error test.error
## knn       0.07914468      0.094
## tree      0.02943627      0.061
## logistic  0.07081366      0.081
```

TREE model has lowest test.error of 0.061

Problem 9: ROC Curves for Tree v Logistic




```
## [[1]]
## [1] 0.9578583
```

```
## [[1]]
## [1] 0.9758875
```

Problem 10:

When considering time spent on email, efficieny and accuracy are usually negatively correlated. However, accuracy oftern supersedes efficieny. Thus false positives would be the main concern. Emails that get marked as “spam” that are not can be very bad for the customer of this spam filter.

Problem 11: Multivariate Normal

If $\hat{Y} = 1$, then $P(Y = 1|X = x) > T$

$$\frac{f_1(x)\pi_1}{f_1(x)\pi_1 + f_2(x)\pi_2} > T$$
$$\frac{1}{1 + \frac{f_2(x)\pi_2}{f_1(x)\pi_1}} > T$$
$$\frac{1}{T} > 1 + \frac{f_2(x)\pi_2}{f_1(x)\pi_1}$$
$$\frac{1-T}{T} > \frac{f_2(x)\pi_2}{f_1(x)\pi_1}$$
$$log(\pi_1) + log(f_1(x)) - log(\pi_2) - log(f_2(x)) > log(\frac{T}{1-T})$$
$$\text{expanding the log we get}$$
$$log(f_k(x)) = -\frac{1}{2}log(|\Sigma_k^{-1}|) + log(\pi_k)$$
$$\text{substituting in } log(f_k(x))\ k = 1, 2$$
$$-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - \frac{1}{2}log(|\Sigma_1^{-1}|) + log(\pi_1)$$
$$+\frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2) + \frac{1}{2}log(|\Sigma_2^{-1}|) - log(\pi_2) > log(\frac{T}{1-T})$$
$$\delta_1(x) - \delta_2(x) > log(\frac{T}{1-T})$$

Let $M(T) = log(\frac{T}{1-T})$, then $\delta_1(x) - \delta_2(x) > M(T)$. When the thresshold $p = \frac{1}{2}, \text{then } M(\frac{1}{2}) = log\left(\frac{\frac{1}{2}}{1-\frac{1}{2}}\right) = 0$. When the probability threshold is $\frac{1}{2}$ we have a decision threshold of 0.

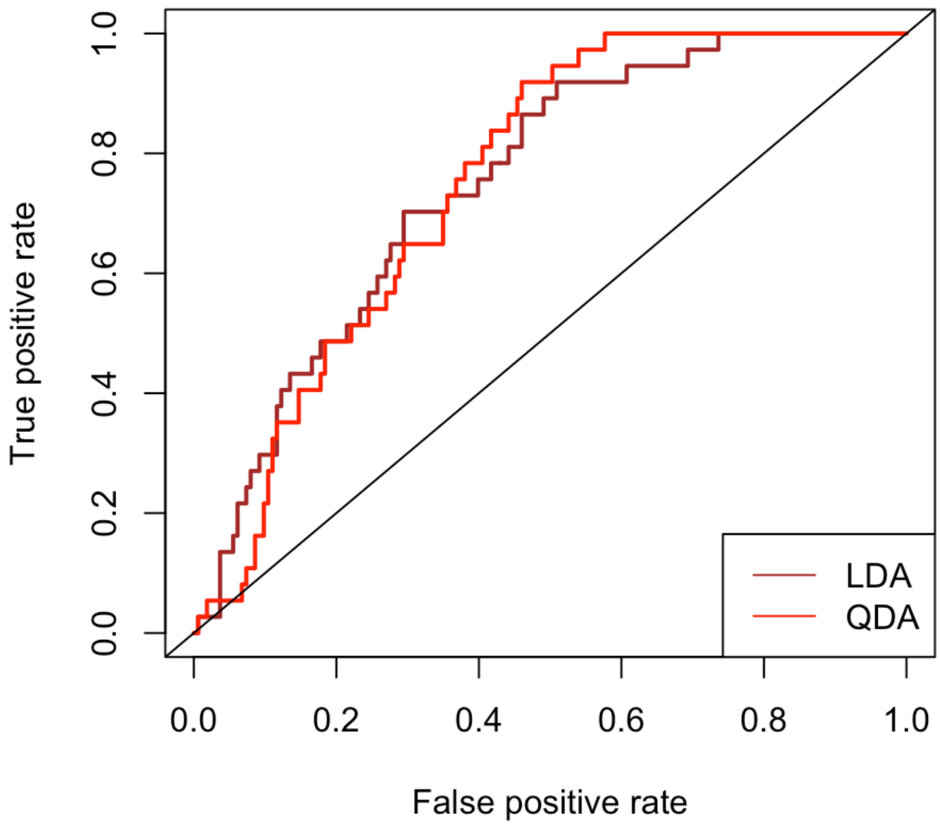
Problem 12: Variable Standardization and Discretization

```
algae = mutate_at(algae, vars(colnames[4:11]),
                  funs(log(.))) # log transform
algae = mutate_at(algae, vars(colnames[4:11]),
                       funs(ifelse(is.na(.), median(algae$.,na.rm=TRUE), .))) # repal
ce NA's with medians
algae = mutate_at(algae, vars(a1), funs(ifelse(.>0.5,"High","Low"))) # a1 as factor
```

Problem 13. Linear and Quadratic Discriminant Analysis

a) LDA

b) QDA




```
## [[1]]
## [1] 0.7517825
```

```
## [[1]]
## [1] 0.7534406
```

AUC of QDA is 0.753 vs. LDA of 0.751, thus the “better” model is QDA.