Advanced Statistical Methods Homework 4

Brandon Hosley

University of Illinois - Springfield

Advanced Statistical Methods Homework 4

Introduction to Statistical Learning
Chapter 4.7 : Problem 13

Using the ***Boston*** *data set, fit classification models in order to predict whether a given suburb has a crime rate above or below the median. Explore logistic regression, LDA, and KNN models using various sub-sets of the predictors. Describe your findings.*

Prepare the data set:

```
library(MASS)
attach(Boston)
library(Metrics)

dim(Boston)
cor(Boston[,-14])

summary(crim)
b <- Boston
medCrim = median(b$crim)
b$highCrim <- ifelse(b$crim < medCrim, 0, 1)
summary(b$highCrim)

set.seed(123)
train_ind <- sample(seq_len(nrow(b)), size = floor(0.8 * nrow(b)))
train <- b[train_ind, ]
test <- b[-train_ind, ]
```

(a)

Logistic Regression

```
glm.fits=glm(highCrim~rad+tax+lstat, data=train, family=binomial)
summary(glm.fits)
```

Train the model on the training data. Then we will test it against the test data.

```
glm.pred=predict(glm.fits, test, type="response")
mse(test$highCrim,glm.pred)
[1] 0.1520493
```

The mean-squared error result from the model applied to test data is fairly low. It suggests a model substantially better than random guessing; though it is likely that there is room for improvement.

(b)

LDA

```
lda.fit = lda(highCrim~rad+tax+lstat, data=train)
lda.fit
lda.pred=predict(lda.fit, test)
names(lda.pred)
lda.class=lda.pred$class
table(lda.class, test$highCrim)
```

```
lda.class  0  1
        0 49 20
        1  3 30
```

Based on the same predictors as the Logistic regression, the LDA produces far less accurate results. Where the addition of predictors caused overfit on the logistic regression, adding predictors for LDA improved the results.

```
lda.fit = lda(highCrim~rad+tax+lstat+nox+dis, data=train)
lda.fit
lda.pred=predict(lda.fit, test)
names(lda.pred)
lda.class=lda.pred$class
table(lda.class, test$highCrim)
```

```
lda.class  0  1
        0 50 12
        1  2 38
```

The addition of the next two highest correlated predictors reduces the error to just below that given by the logistic regression. Adding one or more predictors after this begins to increase the error rate, suggesting over fitting. Current results are slightly better than the previous method.

(c)

KNN

```
library(class)

knn.pred=knn(train,test,train$highCrim ,k=1)
table(knn.pred, test$highCrim)
```

```
knn.pred  0  1
        0 51  6
        1  1 44
```

Even with just a single nearest neighbor the results are far better than what has been achieved by the regression or discriminator.

```
knn.pred=knn(train,test,train$highCrim ,k=2)
table(knn.pred, test$highCrim)
```

```
knn.pred  0  1
        0 48  4
        1  4 46
```

Somewhat surprisingly, increasing the number of neighbors above one only decreases the accuracy of the model on the test data.