Homework 2 for statistical machine learning

Yi Chen(yc3356)

February 17, 2018

Homework 2

Porblem 3

The zipcode data are high dimensional, and hence linear discriminant analysis suffers from high variance. Using the training and test data for the 3s, 5s, and 8s, compare the following procedures:

```
setwd("C:/Users/cheny/Desktop/study/second term/staistical machine learning/homework/homework two")
train_3 <- read.table("train_3.txt", header=FALSE, sep=",")
train_3$number <- 3
train_5 <- read.table("train_5.txt", header=FALSE, sep=",")
train_5$number <- 5
train_8 <- read.table("train_8.txt", header=FALSE, sep=",")
train_8$number <- 8
train_data <- rbind(train_3,train_5,train_8)
```

```
test <- read.table("zip_test.txt",header = FALSE,sep = " ")
number <- test[,1]
test_data <- test[,-1]
test_data$number <- number
colnames(test_data) <- c("V1",colnames(test_data)[-256])

# we only need the testing data which is represent the number 3,5,8
test_data <- test_data[(test_data$number==3 | test_data$number==5|test_data$number==8),]
```

question 1

LDA on the original 256 dimensional space.

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 3.4.3
```

```
lda.model<-lda(number ~.-number , data=train_data)
lda.pred1 = predict(lda.model, train_data[,-257])
table(train_data$number, lda.pred1$class)
```

```
##
##       3   5   8
##   3 644   5   9
##   5   6 549   1
##   8   2   5 535
```

Let calculate the misclassification error:

the number of the observation is 1756

the number of misclassification is 5 + 9 + 6 + 1 + 2 + 5

Thus: the error rate for training set is: 28/1756 = 0.01594533

now, let test the performance of the model in the test data

```
lda.pred2 = predict(lda.model, test_data[,-257])
table(test_data$number, lda.pred2$class)
```

```
##
##       3   5   8
##   3 148  11   7
##   5  14 145   1
##   8   3   7 156
```

Let calculate the misclassification error:

the number of the observation is 2007

the number of misclassification is 11 + 7 + 14 + 1 + 3 + 7

Thus: the error rate for testing set is: 43/492 = 0.08739837

question 2

LDA on the leading 49 principle components of the features

```
pca <- prcomp(train_data[,1:256])
score <- data.frame(pca$x[, 1:49])
colnames(score) <- colnames(train_data)[1:49]
score$number <- train_data$number
```

Now let's trian the data

```
lda.model2<-lda(number ~.-number , data=score)
lda.pred3 = predict(lda.model2, score[,-50])
table(score$number, lda.pred3$class)
```

```
##
##       3   5   8
##   3 631  16  11
##   5  19 529   8
##   8  12  11 519
```

Let calculate the misclassification error:

the number of the observation is 1756

the number of misclassification is 16 + 11 + 19 + 8 + 12 + 11

Thus: the error rate for training set is: 77/1756 = 0.0438

```
pca2 <- prcomp(test_data[,-257])
score2 <- data.frame(pca2$x[, 1:49])
colnames(score2) <- colnames(test_data)[1:49]
score2$number <- test_data$number
lda.model2<-lda(number ~.-number , data=score2)
lda.pred4 = predict(lda.model2, score2[,-50])
table(score2$number, lda.pred4$class)
```

```
##
##       3   5   8
##   3 154   8   4
##   5   8 151   1
##   8   6   5 155
```

Let calculate the misclassification error:

the number of the observation is 2007

the number of misclassification is 8 + 4 + 8 + 1 + 6 + 5

Thus: the error rate for training set is: 32/492 = 0.06504065

question 3

now, let rebulit the data

```
new_data_transformer <- function(dataset){
        ans <- rep(NA,64)
        for (i in 0:7){
                for (j in 0:7) {
                        ans[i*8 + j + 1] <- mean(c(dataset[32*i + 2*j + 1],
                                                   dataset[32*i + 2*j +2],
                                                   dataset[32*i + 2*j + 17],
                                                   dataset[32*i + 2*j + 17]))
                }
        }
        return(ans)
}

new_train <- matrix(NA,nrow = nrow(train_data),ncol = 64)
for (i in 1:nrow(train_data)) {
        new_train[i,] <- new_data_transformer(as.numeric(train_data[i,1:256]))
}
new_train <- as.data.frame(new_train)
new_train$number <- train_data$number
colnames(new_train)[1:64] <- colnames(train_data)[1:64]

new_test <- matrix(NA,nrow = nrow(test_data),ncol = 64)
for (i in 1:nrow(test_data)) {
        new_test[i,] <- new_data_transformer(as.numeric(test_data[i,1:256]))
}
new_test<- as.data.frame(new_test)
new_test$number <- test_data$number
colnames(new_test)[1:64] <- colnames(test_data)[1:64]
```

```
lda.model3<-lda(number ~.-number , data=new_train)
lda.pred5 = predict(lda.model3, new_train[,-65])
table(new_train$number, lda.pred5$class)
```

```
##
##       3   5   8
##   3 634  13  11
##   5  12 535   9
##   8  12   8 522
```

Let calculate the misclassification error:

the number of the observation is 878

the number of misclassification is 13 + 11 + 12 + 9 + 12 + 8

Thus: the error rate for training set is: 65/1756 = 0.03701595

```
lda.pred6 = predict(lda.model3, new_test[,-65])
table(new_test$number, lda.pred6$class)
```

```
##
##       3   5   8
##   3 150   8   8
##   5   9 149   2
##   8   5   6 155
```

Let calculate the misclassification error:

the number of the observation is 246

the number of misclassification is 8 + 8 + 9 + 2 + 5 + 6

Thus: the error rate for training set is: 38/492 = 0.07723577

question 4

```
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 3.4.3
```

```
## Loading required package: Matrix
```

```
## Loading required package: foreach
```

```
## Warning: package 'foreach' was built under R version 3.4.3
```

```
## Loaded glmnet 2.0-13
```

```
glm.fit=glmnet(as.matrix(new_train[,-65]),y = as.factor(new_train$number), family = "multinomial")
```

```
## Warning: from glmnet Fortran code (error code -95); Convergence for 95th
## lambda value not reached after maxit=100000 iterations; solutions for
## larger lambdas returned
```

```
pred = predict(object = glm.fit, newx=as.matrix(new_train[,-65]),type="class")

table(new_train$number, pred[,94])
```

```
##
##       3   5   8
##   3 656   1   1
##   5   0 556   0
##   8   0   0 542
```

Let calculate the misclassification error:

the number of the observation is 878

the number of misclassification is 2

Thus: the error rate for training set is: 2/1756 = 0.001138952

```
glm_pred = predict(glm.fit,as.matrix(new_test[,-65]),type="class")
table(new_test$number, glm_pred[,94])
```

```
##
##       3   5   8
##   3 147  14   5
##   5   9 146   5
##   8   7   6 153
```

Let calculate the misclassification error:

the number of the observation is 246

the number of misclassification is 14+5+9+5+7+6

Thus: the error rate for training set is: 46/492 = 0.09349593