Stat 432 Homework 8

Assigned: Mar 15, 2019; Due: 11:59PM Mar 29, 2019

Before starting this homework, you should read the rlab file of Class on the course website carefully.

Question 1 (logistic regression) [5 points]

We consider a logistic regression problem using the South Africa heart data as a demonstration. The goal is to estimate the probability of `chd`, the indicator of coronary heart disease. The following code is used to prepare the data and fit the logistic regression.

```
library(ElemStatLearn)
data(SAheart)

heart = SAheart
heart$famhist = as.numeric(heart$famhist)-1
n = nrow(heart)
p = ncol(heart)

heart.full = glm(chd~., data=heart, family=binomial)

# fitted value
yhat = (heart.full$fitted.values>0.5)
table(yhat, SAheart$chd)
```

```
##
## yhat      0   1
##   FALSE 256  77
##   TRUE   46  83
```

The goal is to replicate the following summary matrix using your own code. You are **not allowed** to use any built-in optimization functions. The only statistical function you might use is `pnorm()`.

```
# the coefficients and significance
round(summary(heart.full)$coef, dig=3)
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -6.151      1.308  -4.701    0.000
## sbp            0.007      0.006   1.135    0.256
## tobacco        0.079      0.027   2.984    0.003
## ldl            0.174      0.060   2.915    0.004
## adiposity      0.019      0.029   0.635    0.526
## famhist        0.925      0.228   4.061    0.000
## typea          0.040      0.012   3.214    0.001
## obesity       -0.063      0.044  -1.422    0.155
## alcohol        0.000      0.004   0.027    0.978
## age            0.045      0.012   3.728    0.000
```

You should proceed with the following steps:

- Write a function that computes the Hessian matrix
- Write a function that uses Newton–Raphson to iteratively update the parameter values and search for the optimal solution. You should use `rep(0, ncol(x))` as your initial value.
- From the solution, replicate the summary matrix displayed above.
- Change the initial value to `rep(1, ncol(x))` and comment on your findings.

Question 2 (LDA and QDA)

Load the handwritten digit data with the following code to generate training and testing data

```
    # a plot of some samples
    findRows <- function(zip, n) {
        # Find n (random) rows with zip representing 0,1,2,...,9
        res <- vector(length=10, mode="list")
        names(res) <- 0:9
        ind <- zip[,1]
        for (j in 0:9) {
        res[[j+1]] <- sample( which(ind==j), n ) }
        return(res)
    }

    set.seed(1)

    # find 100 samples for each digit for both the training and testing data
    train.id <- findRows(zip.train, 100)
    train.id = unlist(train.id)

    test.id <- findRows(zip.test, 100)
    test.id = unlist(test.id)

    X = zip.train[train.id, -1]
    Y = zip.train[train.id, 1]
    dim(X)
```

```
## [1] 1000  256
```

```
    Xtest = zip.test[test.id, -1]
    Ytest = zip.test[test.id, 1]
    dim(Xtest)
```

```
## [1] 1000  256
```

a) [5 points]

We want to write a code to implement the LDA method. Obtain W_k and b_k in the discriminant function defined in page 29 of the lecture notes “Class”. Then use these quantities on the testing dataset to predict the digit. Calculate the confusion matrix and the prediction accuracy. You are **not allowed** to use built-in functions `lda`.

b) [extra 3 points]

We also demonstrated that QDA does not work on this dataset.

- Use one of the regularized approaches provided in the lecture note, implement a method that can produce good prediction accuracy
- Since the problem was caused by $p > n$, how about we reduced the dimension of the dataset first, and apply QDA? Compare this approach with the previous one.

Question 3 [extra 3 points]

On pages 42-44 of lecture `class`, we have a golf example using naive Bayes method for categorical variables. Write a code to implement naiveBayes method for categorical variables. Automate the calculation in this example and output the calssification result for the instance `today=(Sunny, Hot, Normal, False)`. You are **not allowed** to use built-in functions `naiveBayes`, but you can use it to check your answer.