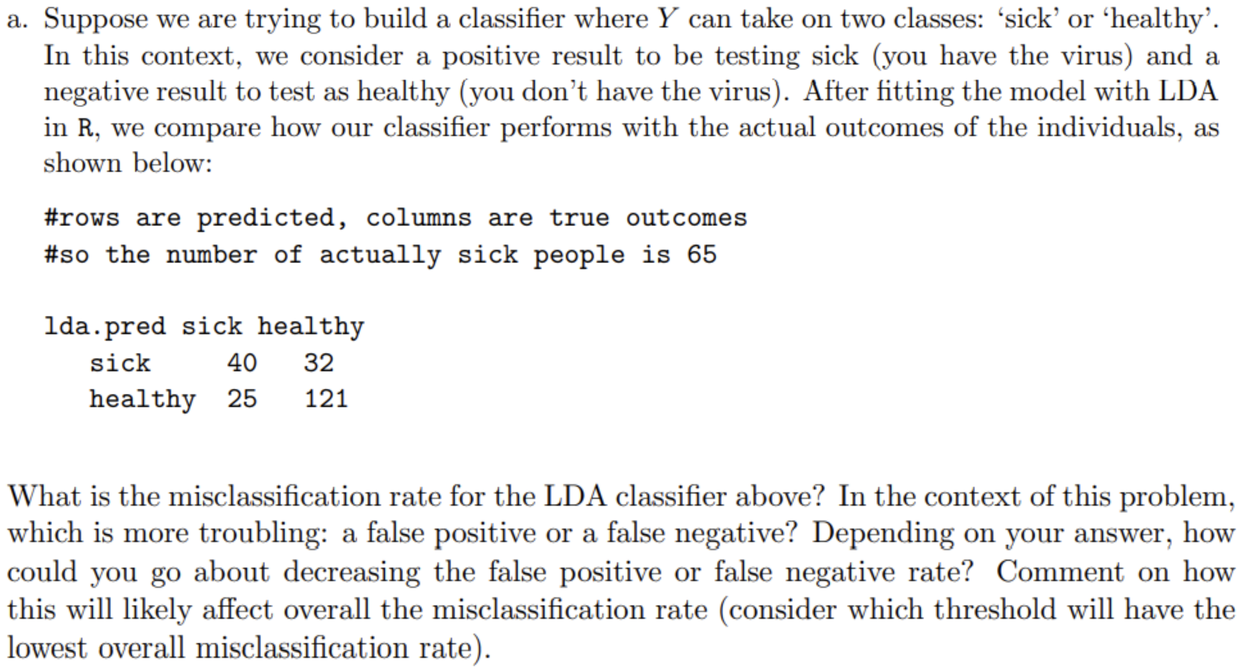
Keven Lin
04/20/2022

Homework 9

Problem 1: Concept Review



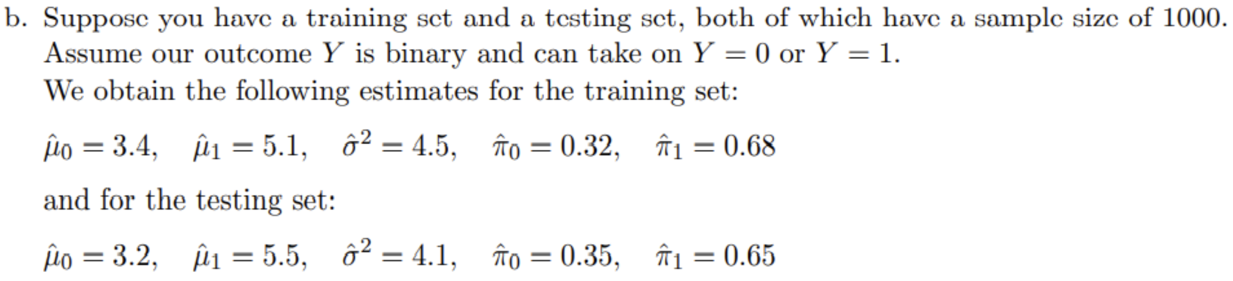
a. Suppose we are trying to build a classifier where Y can take on two classes: 'sick' or 'healthy'. In this context, we consider a positive result to be testing sick (you have the virus) and a negative result to test as healthy (you don't have the virus). After fitting the model with LDA in `R`, we compare how our classifier performs with the actual outcomes of the individuals, as shown below:

```
#rows are predicted, columns are true outcomes
#so the number of actually sick people is 65

lda.pred sick healthy
   sick     40   32
   healthy  25   121
```

What is the misclassification rate for the LDA classifier above? In the context of this problem, which is more troubling: a false positive or a false negative? Depending on your answer, how could you go about decreasing the false positive or false negative rate? Comment on how this will likely affect overall the misclassification rate (consider which threshold will have the lowest overall misclassification rate).

Misclassification rate: (25+32) / (40+32+25+121) = 26%
False positive: 25 / (25 + 121) = 17%
False negative: 32 / (32 + 40) = 44%

There is an unusually high false negative rate and I think in this situation, say this is Covid-19. This high of a false negative rate is dangerous. A high false negative rate will increase the misclassification rate.



b. Suppose you have a training set and a testing set, both of which have a sample size of 1000. Assume our outcome Y is binary and can take on $Y = 0$ or $Y = 1$.
We obtain the following estimates for the training set:

$$\hat{\mu}_0 = 3.4, \quad \hat{\mu}_1 = 5.1, \quad \hat{\sigma}^2 = 4.5, \quad \hat{\pi}_0 = 0.32, \quad \hat{\pi}_1 = 0.68$$

and for the testing set:

$$\hat{\mu}_0 = 3.2, \quad \hat{\mu}_1 = 5.5, \quad \hat{\sigma}^2 = 4.1, \quad \hat{\pi}_0 = 0.35, \quad \hat{\pi}_1 = 0.65$$

1. Based on the information above, construct the LDA classifier. Explain when test observations will be assigned to $Y = 0$ and when they will be assigned to $Y = 1$. **Show your work for full credit.**

2. What threshold would give us the smallest possible test misclassification rate? Explain why.

c. Suppose you just took on a new consulting client. He tells you he has a large dataset (say 100, 000 observations) and he wants to use this to classify whether or not to invest in a stock based on a set of $p = 10,000$ predictors. He claims KNN will work really well in this case because it is non-parametric and therefore makes no assumptions on the data. Present an argument to your client on why KNN might fail when p is large relative to the sample size.

d. For each of the following classification problems, state whether you would advise a client to use LDA, logistic regression, or KNN and explain why:

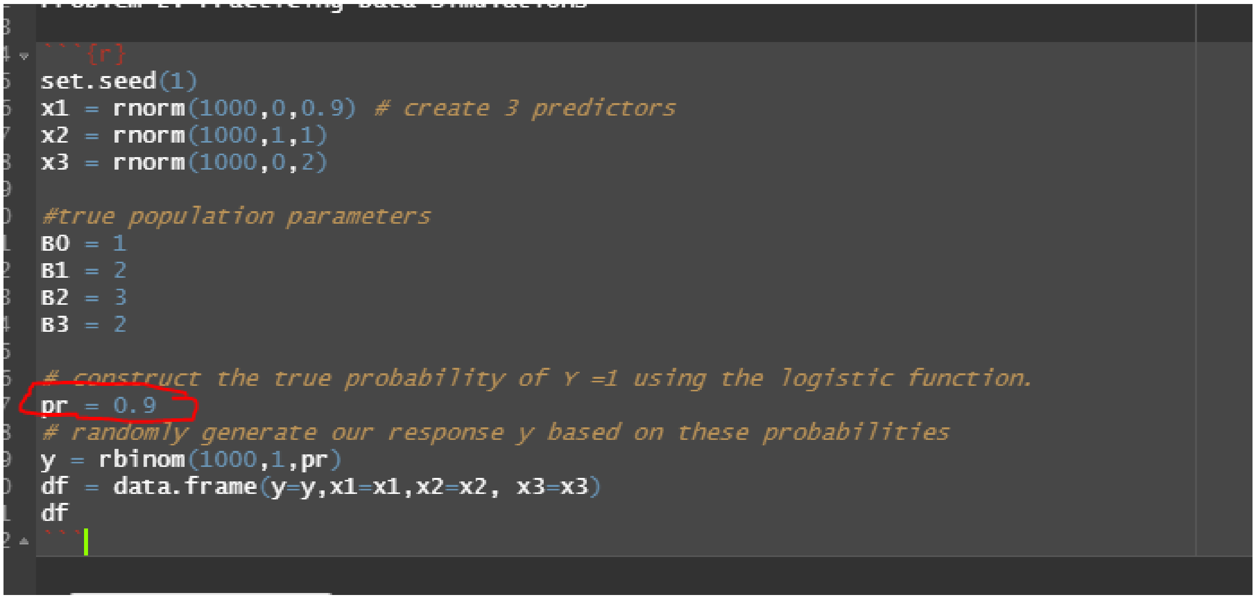
   i. We want to predict gender based on height and weight. The training set consists of heights and weights for 82 men and 63 women.

   ii. We want to predict gender based on annual income and weekly working hours. The training set consists of 770 mean and 820 women.

   iii. We want to predict gender based on a set of predictors where the decision boundary is complicated and highly non-linear. The training set consists of 960 men and 1040 women.

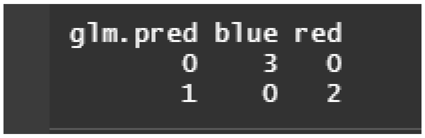
Problem 2: Practicing Data Simulations

a.

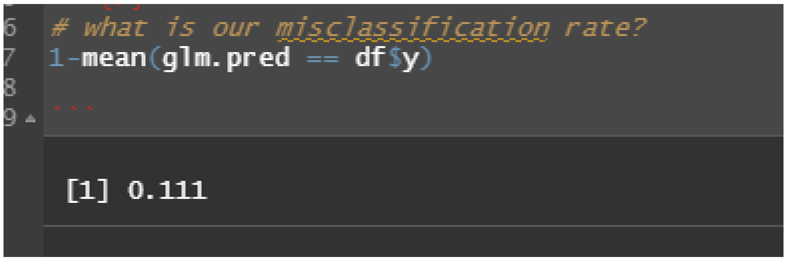


~~~
```{r}
set.seed(1)
x1 = rnorm(1000,0,0.9) # create 3 predictors
x2 = rnorm(1000,1,1)
x3 = rnorm(1000,0,2)

#true population parameters
B0 = 1
B1 = 2
B2 = 3
B3 = 2

# construct the true probability of Y =1 using the logistic function.
pr = 0.9
# randomly generate our response y based on these probabilities
y = rbinom(1000,1,pr)
df = data.frame(y=y,x1=x1,x2=x2, x3=x3)
df
```
~~~

b.



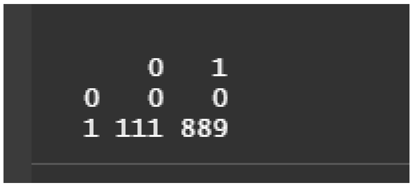
```
glm.pred blue red
       0    3   0
       1    0   2
```



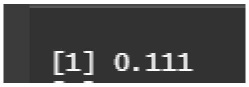
~~~
# what is our misclassification rate?
1-mean(glm.pred == df$y)
```
~~~

```
[1] 0.111
```

c. On the simulated data, apply LDA. Compute the confusion matrix and the misclassification rate.

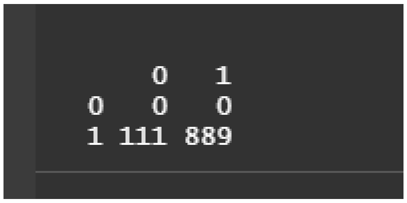


```
     0   1
0    0   0
1 111 889
```

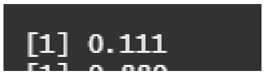


```
[1] 0.111
```

d. On the simulated data, apply K-NN (obtain the optimal K using cross-validation). Remember to standardize your predictors for K-NN. Report the K you obtained. Compute the confusion matrix and the misclassification rate.



```
     0   1
0    0   0
1 111 889
```



```
[1] 0.111
```

e. How do the 3 methods compare?

Basically identical.

Problem 3: Weekly Data