Computer Assignment 9 - Model Assumptions and Cross Validation

Machine Learning, Spring 2020

YOUR NAME

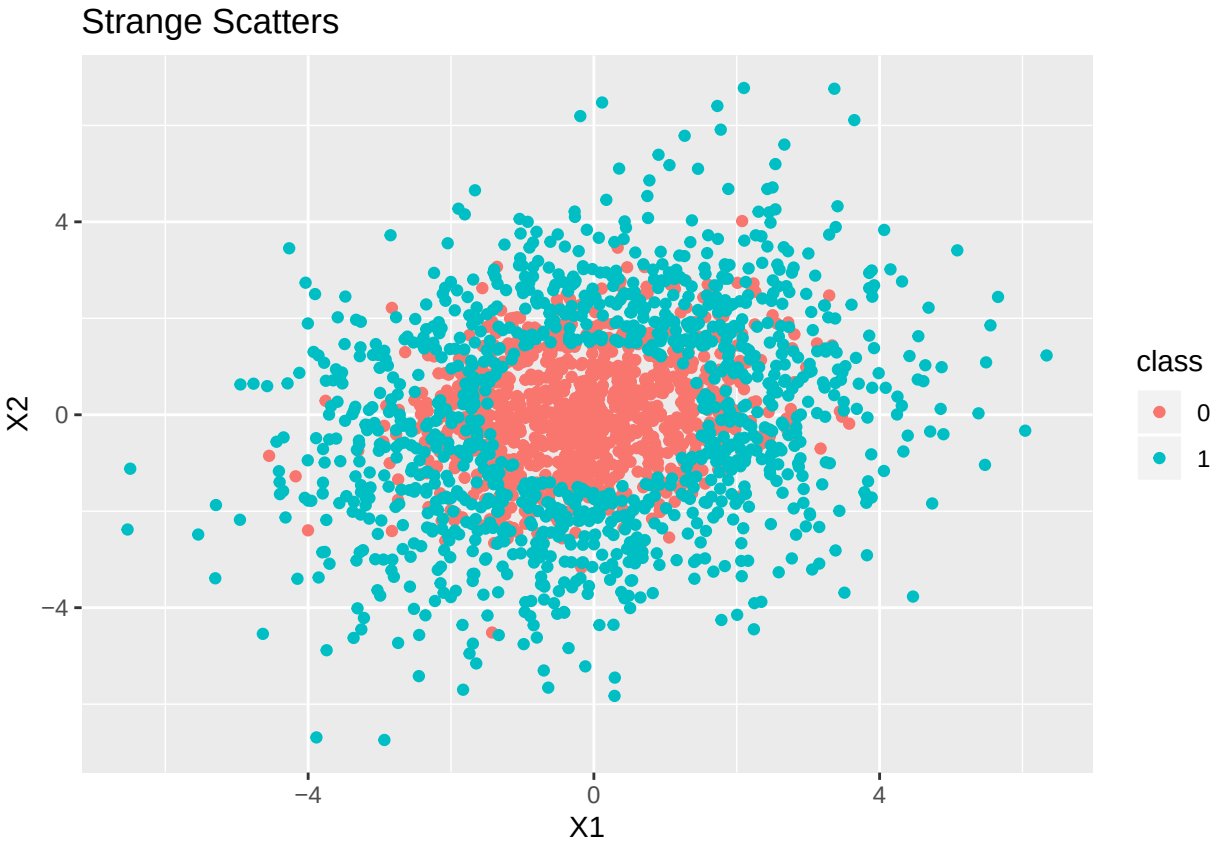
Not-So-Perfect Data for LDA

Recall on the previous CA we performed LDA on a data set that was generated from the LDA model. For that data set the two groups (classified 0 and 1) were generated from two multinormal distributions with the same variance but different means. LDA performed very well on that data because *the data were generated from the underlying LDA model* that we were trying to fit.

In this assignment, we will explore what happens when we apply our methods to a dataset that is generated from a distribution that is *very different* from a LDA model. We will start in two dimensions. Note that while we are still using the multinormal function to generate the following data, the variance of the marginals is different, the means are the same, and we manipulate `data_1` in such a way that it is no longer normal. Indeed,

```
# Make sure to install this package if you do not have it:
library(mvtnorm)
my_sigma_1 = t(matrix(c(2,0.2,0.2,2)*0.6, ncol=2)) %*% matrix(c(2,0.2,0.2,2)*0.6, ncol=2)
my_sigma_2 = t(matrix(c(2,0.2,0.2,2), ncol=2)) %*% matrix(c(2,0.2,0.2,2), ncol=2)
data_0 = rmvnorm(1000, mean = rep(0, times = 2), sigma = my_sigma_1)
data_1 = rmvnorm(1500, mean = rep(0, times = 2), sigma = my_sigma_2)
central_points = ( sqrt( data_1[,1]**2 + data_1[,2]**2) < 1.5 )
data_1 = data_1[!central_points,]
```

```
library(ggplot2)
temp_data = data.frame(rbind(data_0, data_1))
temp_data$class = as.factor(c(rep(0, times = 1000), rep(1, times = nrow(data_1))))
ggplot(temp_data, aes(x = X1, y = X2, color = class)) + geom_point() +
  ggtitle("Strange Scatters")
```



Questions

1. The way these data were created breaks *two* of the assumptions for LDA. What are they? Do you predict LDA or KNN will have better performance on this data? Defend your choice.

YOUR ANSWER HERE

2. If you were to draw a decision boundary on this data, what shape would it be? Why might this not be a good thing for an LDA model?

YOUR ANSWER HERE

3. Split the data into a testing data set and a training data set. Build an LDA model on the training data. Use the training data again to predict the class labels and report the performance of your model. What is the error rate? (the proportion of times your model classified an observation incorrectly)

```
YOUR CODE AND ANALYSIS HERE
```

4. Now predict the class labels using the testing data and report the performance of your model. What is the error rate?

```
YOUR CODE AND ANALYSIS HERE
```

5. Repeat steps 2-3 using a knn model for $k \in \{1, 5, 11\}$. Considering only the cases where you predicted labels on your testing data, compare the error rates between the LDA model and all of the knn models. Was your prediction in question 1 correct or incorrect?

```
YOUR CODE AND ANALYSIS HERE
```

Error Rate for Increasing Data Sizes

We will now examine how the error rate stabalizes for larger and larger data sizes. Manipulate the variable `data_size` in the following code to simulate different sizes of the data we created for the last problem.

```
data_size = 10
my_sigma_1 = t(matrix(c(2,0.2,0.2,2)*0.6, ncol=2)) %*% matrix(c(2,0.2,0.2,2)*0.6, ncol=2)
my_sigma_2 = t(matrix(c(2,0.2,0.2,2), ncol=2)) %*% matrix(c(2,0.2,0.2,2), ncol=2)
data_0 = data.frame(rmvnorm(data_size, mean = rep(0, times = 2), sigma = my_sigma_1))
data_1 = data.frame(X1 = NA, X2 = NA)
count = 0
while(count < data_size){
  new_draw = rmvnorm(1, mean = rep(0, times = 2), sigma = my_sigma_2)
  if( sqrt( new_draw[,1]**2 + new_draw[,2]**2) >= 1.5 ) {
    data_1 = rbind(data_1, data.frame(new_draw))
    count = count + 1
  }
}
data_1 = data_1[-1,]
my_data = data.frame(rbind(data_0, data_1))
my_data$class = as.factor(c(rep(0, times = data_size), rep(1, times = data_size)))
```

For `data_size` $\in \{5, 10, 25, 50, 100, 200, 500, 1000, 10000\}$ do the following:

- Generate the "Strange Scatters" data using the given code.
- Split the data into testing and training data sets.
- Build an LDA model and the KNN models for $k \in \{1, 5, 11\}$ on the training data, predict the labels on the training data, and calculate the training error rate.
- Build an LDA model and the KNN models for $k \in \{1, 5, 11\}$ on the training data, predict the labels on the testing data, and calculate the testing error rate.
- Save both error rates

Plot two scatterplots: one for each error type with the data sizes on the x-axis, the error rate on the y-axis, and color by which model type you used.

Do you notice any trends as your data size increases? Do the error rates "stabilize"?

```
YOUR CODE AND ANALYSIS HERE
```

Now compare the testing error rates to the training error rates using boxplots. The modeled used should be on the x-axis, the value of the error rate should on the y-axis, and you should color by whether the error was calculated on the training or the testing data. Comment on any differences between the testing and the training errors.

```
YOUR CODE AND ANALYSIS HERE
```

Introduction to Cross Validation

Use the data generating procedure from the previous section to generate a dataset of size 1000. Perform a 10-fold cross validation based off of the procedure given in the lecture slides. You are **not** permitted to use `R`'s built in functions to perform this cross validation. Report cross validated (CV) error rate $\hat{R}^{10-CV}(\phi)$ for LDA and for the KNN models where $k \in \{1, 5, 11\}$. Compare this CV error rate with the error rates you derived in the previous problem for the same data size. What are these two quantities trying to estimate? Are they trying to estimate the same thing?

```
YOUR CODE AND ANALYSIS HERE
```

What do the CV error rates for LDA and KNN tell us about which of these procedures is better for data generated from our given distribution?

YOUR ANSWER HERE

In the previous problem, we derived the non-CV testing and training error rates for a `data_size` of 1000. What did those error rates tell us about the *classification rules* calculated for that particular data set? Can we use these non-CV results to say anything about the procedures we used?

YOUR ANSWER HERE