DATA-570 Lab 1

Dhun Sheth

2023-10-19

Question 4

Classification is primarily used when you need to categorize data into specific buckets. As an example, you could use classification to classify emails as spam or not. Here, the response variable is whether the email is spam or not. The predictor variables would be the fields used to determine if the email should be classified as spam or not, examples could include the sender email, subject line or attachments.

Either prediction or inference could be the primary purpose. If you are exploring more efficient models in being able to classify emails as spam or not - then the primary purpose would be inference. However, if you are developing this model to implement in an application, then the primary goal would be prediction, as you want a low error rate in classification.

Question 5

Regression is typically used for numeric response variables. As an example, if we wanted to predict the price of a stock based on market cap, earnings, cash flow, and revenue, regression could be used were the response variable would be the stock price, and the predictor variables are market cap, earnings, cash flow, and revenue.

Because stock price is very difficult to model because it is affected by numerous factors not included in the above example, this would primarily be done for inference, however if there is a particular stock for which this method is accurate using test data, then prediction can be argued to be of more interest than inference.

Question 6

I would expect a smaller value for K because as K gets larger, the boundary becomes less flexible. For K = n, where n is the total number of observations, the boundary is a straight line and not very flexible, therefore, if the bounder is very non-linear, I expect the boundary to be more flexible indicating a small K.

Question 7

Part A

```
x <- rbind(c(0, 2, 0),c(-2, -1, 0), c(-1, 0, 1), c(0, 1, 3), c(1, 1, 1), c(0, 3, 0))
origin <- matrix(0, nrow=6, ncol=3)

eu_distance <- matrix(sqrt(rowSums((x-origin)**2)), ncol=1)

print(eu_distance)
```

```
##          [,1]
## [1,] 2.000000
## [2,] 2.236068
## [3,] 1.414214
## [4,] 3.162278
## [5,] 1.732051
## [6,] 3.000000
```

Part B

```
smallest_eu_distance <- min(eu_distance)

print(smallest_eu_distance)
```

```
## [1] 1.414214
```

Because K=1, it will classify the new observation the same class as its nearest neighbor. The smallest Euclidean distance is 1.4142136 which occurs for the 3rd data point which is classified as "Yellow" - so the new observation would be classified as "Yellow."

Part C

```
print(eu_distance)
```

```
##          [,1]
## [1,] 2.000000
## [2,] 2.236068
## [3,] 1.414214
## [4,] 3.162278
## [5,] 1.732051
## [6,] 3.000000
```

Based on the Euclidean distances calculated in Part A, for K=3, we can see the 3 smallest distances are:

"1.414214" -> Class: "Yellow"
"1.732051" -> Class: "Yellow"
"2.000000" -> Class: "Blue"

Thus, the origin can be of class "Blue" which has a probability of 1/3 or class "Yellow" which has a probability 2/3.
So the KNN classification would classify the origin as "Yellow".