Assignment6

John Bute

2024-10-01

First let us set up the dataframe and compute the euclidean distance from test point X1 = X2 = X3 = 0

```r
data <- data.frame(
  Obs = 1:6,
  X1 = c(0, 2, 0, 0, -1, 1),
  X2 = c(3, 0, 1, 1, 0, 1),
  X3 = c(0, 0, 3, 2, 1, 1),
  Y = c("red", "red", "red", "green", "green", "red")
)
euclidian_3 <- function(x1o, x1t, x2o, x2t, x3o, x3t){
  sqrt((x1o - x1t)^2 + (x2o - x2t)^2 + (x3o - x3t)^2)
}

test_point <- c(0, 0, 0)

distances <- numeric(nrow(data))
for (i in 1:nrow(data)) {
  distances[i] <- euclidian_3(data$X1[i], test_point[1],
                              data$X2[i], test_point[2],
                              data$X3[i], test_point[3])
}
data$Distance <- distances
print(data)
```

```
##   Obs X1 X2 X3     Y Distance
## 1   1  0  3  0   red 3.000000
## 2   2  2  0  0   red 2.000000
## 3   3  0  1  3   red 3.162278
## 4   4  0  1  2 green 2.236068
## 5   5 -1  0  1 green 1.414214
## 6   6  1  1  1   red 1.732051
```

The table above calculates the euclidean distance between each observation and the test point X1 = X2 = X3 = 0

2. The prediction with K = 1 would be "Green" as the test point X1 = X2 = X3 = 0 is closest to Observation 5 (-1, 0, 1), who's qualitative variable is green.

3. The prediction with K = 3 neighbors is "red" as the three closest neighbors are Observation 2 (2, 0, 0), observation 5 (-1, 0, 1), and observation 6 (1, 1, 1), which have as response variables red, green, red respectively. Thus, the most common response variable is red, therefore our prediction would be red.

4. If the bayes decision boundary in this problem is highly non-linear, then we would expect to see a line that over fits the data set, moving erratically in order to classify the training set correctly. Thus, we would expect a very small k (k = 1) as it would attempt to fit the training data as closely as possible. Using k=1 means that the model looks at the closest training point to make its prediction, thus causing over fitting.,