11.Two Simple Approaches to Prediction: Least Squares and Nearest Neighbors

Avery Holloman

2024-08-21

```
# Loading the ggplot2 package for better visualization and the caret package in R is a powerful tool for building predictive models. It stands for Classification And Regression Training and pr ovides a unified interface for training, tuning, and evaluating various machine learning model. From my perspective, the class package in R is a great tool for applying k-nearest neighbors (k-NN) classification and k-means clustering, both of which are key methods in machine learning and statistical analysis. I find this package particularly useful because it's lightweight and speci fically designed for k-NN algorithms, making it straightforward to implement and experiment with these popular classification techniques in my work.
library(ggplot2)
library(caret)
```

```
## Loading required package: lattice
```

```
library(class)
library(readxl) # Required for reading Excel files

# Next I need to load the data from my folder locally on my computer
healthcare <- read_excel("C:/Users/jacob/OneDrive/Desktop/R Studio Projects 2024/Datasets/health care.xlsx")

# Next, lets take a look at the structure of my data and a quick summary
str(healthcare)
```

```
## tibble [5,834 × 27] (S3: tbl_df/tbl/data.frame)
##  $ Patient_ID                    : num [1:5834] 1001 1002 1003 1004 1005 ...
##  $ Age                           : num [1:5834] 69 32 89 78 38 41 20 39 70 19 ...
##  $ Gender                        : chr [1:5834] "Male" "Female" "Female" "Male" ...
##  $ Diagnosis                     : chr [1:5834] "Arthritis" "Arthritis" "Diabetes" "Asthma"
...
##  $ Medication_1                  : chr [1:5834] "None" "Drug_B" "Drug_C" "Drug_C" ...
##  $ Medication_2                  : chr [1:5834] "Drug_Y" "Drug_Y" "Drug_Z" "Drug_Y" ...
##  $ Dosage_Med1_mg                : num [1:5834] 246 255 397 426 384 275 258 301 214 237 ...
##  $ Dosage_Med2_mg                : num [1:5834] 393 69 486 329 187 61 154 325 86 128 ...
##  $ Days_On_Treatment             : num [1:5834] 363 349 260 138 319 183 57 240 105 65 ...
##  $ Outcome                       : chr [1:5834] "Improved" "Improved" "Improved" "No Change"
...
##  $ Hospital_Visits_Last_Year     : num [1:5834] 1 17 17 11 2 1 5 5 16 15 ...
##  $ Smoker                        : chr [1:5834] "Yes" "No" "No" "No" ...
##  $ BMI                           : num [1:5834] 15.4 19.6 32.7 25.7 21.2 31.8 24.1 25.1 37 2
9.6 ...
##  $ Blood_Pressure_mmHg           : num [1:5834] 129 158 103 99 171 170 138 106 175 178 ...
##  $ Cholesterol_mg_dL             : num [1:5834] 175 217 185 121 265 111 263 149 116 259 ...
##  $ Genetic_Risk_Factor           : chr [1:5834] "Low" "Low" "Low" "Medium" ...
##  $ Retinal_Scan_Image            : chr [1:5834] "None" "Image_B" "Image_A" "None" ...
##  $ Clinical_Notes                : chr [1:5834] "Patient exhibits symptoms of hypertension."
"Heart disease symptoms, patient reports chest pain." "Heart disease symptoms, patient reports c
hest pain." "Heart disease symptoms, patient reports chest pain." ...
##  $ Speech_Analysis_Result        : chr [1:5834] "Normal" "Normal" "Normal" "Anxious" ...
##  $ Texting_Behavior              : num [1:5834] 915 532 989 781 785 719 951 622 293 338 ...
##  $ Social_Interaction_Score      : num [1:5834] 44 7 5 50 93 22 78 59 40 13 ...
##  $ Daily_Vital_Signs             : chr [1:5834] "Stable" "Stable" "Stable" "Critical" ...
##  $ Geographic_Location           : chr [1:5834] "Suburban" "Suburban" "Urban" "Urban" ...
##  $ Socioeconomic_Status          : chr [1:5834] "Middle" "Middle" "Middle" "Middle" ...
##  $ Access_to_Healthcare_Resources: chr [1:5834] "Limited" "Limited" "Good" "Good" ...
##  $ Data_Privacy_Permissions      : chr [1:5834] "Partial" "Full" "Full" "Restricted" ...
##  $ Synthetic_Bias_Flag           : chr [1:5834] "None" "Mild" "None" "None" ...
```

```
summary(healthcare)
```

```
##    Patient_ID        Age           Gender           Diagnosis
##  Min.   :1001   Min.   :18.00   Length:5834        Length:5834
##  1st Qu.:2459   1st Qu.:35.00   Class :character   Class :character
##  Median :3918   Median :53.00   Mode  :character   Mode  :character
##  Mean   :3918   Mean   :53.32
##  3rd Qu.:5376   3rd Qu.:71.00
##  Max.   :6834   Max.   :89.00
##  Medication_1       Medication_2       Dosage_Med1_mg  Dosage_Med2_mg
##  Length:5834        Length:5834        Min.   : 10.0   Min.   : 10.0
##  Class :character   Class :character   1st Qu.:132.0   1st Qu.:129.2
##  Mode  :character   Mode  :character   Median :253.0   Median :252.5
##                                        Mean   :254.1   Mean   :252.4
##                                        3rd Qu.:378.0   3rd Qu.:377.0
##                                        Max.   :499.0   Max.   :499.0
##  Days_On_Treatment   Outcome          Hospital_Visits_Last_Year
##  Min.   :  1.0     Length:5834        Min.   : 0.000
##  1st Qu.: 93.0     Class :character   1st Qu.: 4.000
##  Median :185.0     Mode  :character   Median : 9.000
##  Mean   :184.2                        Mean   : 9.466
##  3rd Qu.:275.0                        3rd Qu.:14.000
##  Max.   :364.0                        Max.   :19.000
##     Smoker               BMI        Blood_Pressure_mmHg Cholesterol_mg_dL
##  Length:5834        Min.   : 6.30   Min.   : 90.0       Min.   :100.0
##  Class :character   1st Qu.:21.52   1st Qu.:112.0       1st Qu.:149.0
##  Mode  :character   Median :24.90   Median :134.0       Median :201.0
##                     Mean   :24.95   Mean   :134.2       Mean   :199.7
##                     3rd Qu.:28.30   3rd Qu.:156.0       3rd Qu.:250.0
##                     Max.   :41.50   Max.   :179.0       Max.   :299.0
##  Genetic_Risk_Factor Retinal_Scan_Image Clinical_Notes
##  Length:5834         Length:5834        Length:5834
##  Class :character    Class :character   Class :character
##  Mode  :character    Mode  :character   Mode  :character
##
##
##
##  Speech_Analysis_Result Texting_Behavior Social_Interaction_Score
##  Length:5834            Min.   :100.0    Min.   : 0.00
##  Class :character       1st Qu.:333.2    1st Qu.:25.00
##  Mode  :character       Median :555.0    Median :50.00
##                         Mean   :553.8    Mean   :50.07
##                         3rd Qu.:780.8    3rd Qu.:76.00
##                         Max.   :999.0    Max.   :99.00
##  Daily_Vital_Signs  Geographic_Location Socioeconomic_Status
##  Length:5834        Length:5834         Length:5834
##  Class :character   Class :character    Class :character
##  Mode  :character   Mode  :character    Mode  :character
##
##
##
##  Access_to_Healthcare_Resources Data_Privacy_Permissions Synthetic_Bias_Flag
##  Length:5834                    Length:5834              Length:5834
##  Class :character               Class :character         Class :character
```

```
##  Mode  :character              Mode  :character         Mode  :character
##
##
##
```

```
# I will utilize the 'Outcome' as the target variable and 'Age', 'BMI', 'Blood_Pressure_mmHg', a
nd 'Cholesterol_mg_dL' will be my predictors
predictors <- healthcare[, c('Age', 'BMI', 'Blood_Pressure_mmHg', 'Cholesterol_mg_dL')]
outcome <- healthcare$Outcome

# Next, I ran into a few problems plotting my model so after looking at the str or the structure
of my data I realized that I needed to convert 'outcome' variable from a categorical variables t
o factor instead.
healthcare$Outcome <- as.factor(healthcare$Outcome)

# Next, I need to Split the data into training and testing sets to evaluate how well my predicti
ve model performs on new, unseen data in the future
set.seed(123)
trainIndex <- createDataPartition(outcome, p = .8, list = FALSE)
healthcareTrain <- healthcare[trainIndex, ]
healthcareTest <- healthcare[-trainIndex, ]

# Lets get started by training and testing my healthcare data
trainX <- healthcareTrain[, c('Age', 'BMI', 'Blood_Pressure_mmHg', 'Cholesterol_mg_dL')]
trainY <- healthcareTrain$Outcome
testX <- healthcareTest[, c('Age', 'BMI', 'Blood_Pressure_mmHg', 'Cholesterol_mg_dL')]
testY <- healthcareTest$Outcome

# First step is the Least Squares Regression model
lm_model <- lm(as.numeric(as.factor(trainY)) ~ Age + BMI + Blood_Pressure_mmHg + Cholesterol_mg_
dL, data = healthcareTrain)
lm_pred <- predict(lm_model, testX)

# Ok great, now lets evaluate the performance of my Least Square model
lm_rmse <- sqrt(mean((lm_pred - as.numeric(as.factor(testY)))^2))
cat("Least Squares RMSE:", lm_rmse, "\n")
```

```
## Least Squares RMSE: 0.7901535
```

```
# Next, I will do the Nearest Neighbors (k-NN)
k <- 5  # Set the number of neighbors
knn_model <- knn(train = trainX, test = testX, cl = trainY, k = k)

# Ok great, now lets take a closer look by evaluating the k-NN model
knn_rmse <- sqrt(mean((as.numeric(knn_model) - as.numeric(testY))^2))
cat("k-NN RMSE:", knn_rmse, "\n")
```

```
## k-NN RMSE: 1.081975
```

```r
# Well now I want to compare the performance of the models
comparison <- data.frame(
  Model = c("Least Squares", "k-NN"),
  RMSE = c(lm_rmse, knn_rmse)
)

# Just out of curirosity lets see a bar graph for the comparison
print(comparison)
```

```
##           Model      RMSE
## 1 Least Squares 0.7901535
## 2          k-NN 1.0819746
```

```r
# This is were I will use ggplot2 as I like the color structure compared to simple black and whi
te
ggplot(comparison, aes(x = Model, y = RMSE)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_minimal() +
  labs(title = "Model Comparison: Least Squares vs k-NN", y = "RMSE", x = "Model")
```