Lab 9

Contents

Enter the names of the group members here: Austine Do, Graceanne Becker, Catherine Zhong

This assignment is due by the end of the lab. Only one student in the group submits a pdf file on Gradescope.

For all questions, include the R commands/functions that you used to find your answer (show R chunk). Answers without supporting code will not receive credit. Write full sentences to describe your findings.

In this lab, you will continue exploring data originally collected by researchers at the Johns Hopkins Bloomberg School of Public Health. Let's first load the appropriate packages for today:

```
library(tidyverse)
library(plotROC)
```

```
## Warning: package 'plotROC' was built under R version 4.3.1
```

```
library(caret)
library(rpart)
```

Let's re-upload the data from Github and take a quick look again:

```
pollution <- read_csv("https://raw.githubusercontent.com/laylaguyot/datasets/main//pm25.csv") |>
  mutate(violation = ifelse(value > 12, 1, 0))

# Take a quick look!
head(pollution)
```

```
## # A tibble: 6 x 12
##      id state   county  city        value  zcta   lat   lon   pov  CMAQ zcta_pop
##   <dbl> <chr>   <chr>   <chr>       <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>
```

```
## 1 1003. Alabama Baldwin Fairhope     9.60 36532  30.5 -87.9   6.1  8.10    27829
## 2 1027. Alabama Clay    Ashland     10.8  36251  33.3 -85.8  19.5  9.77     5103
## 3 1033. Alabama Colbert Muscle Sho~ 11.2  35660  34.8 -87.7  19    9.40     9042
## 4 1049. Alabama DeKalb  Crossville  11.7  35962  34.3 -86.0  13.8  8.53     8300
## 5 1055. Alabama Etowah  Gadsden     12.4  35901  34.0 -86.0   8.8  9.24    20045
## 6 1069. Alabama Houston Dothan      10.5  36303  31.2 -85.4  15.6  9.12    30217
## # i 1 more variable: violation <dbl>
```

The goal of the lab is to make predictions for the PM2.5 levels with 3 different models and perform cross-validation.

Question 1 (3 pts)

In this report, you will choose to focus on either predicting the PM2.5 values at a given location (`value`) or predicting whether a given location is in `violation` of the national ambient air quality standards (with a `value` greater than 12 $\mu g/m^3$) or not based on `lat`, `lon`, `pov`, and `zcta_pop`.

Which outcome variable will you focus on?

The outcome variable we will focus on is the PM2.5 value at a given location so the outcome variable `value`

Which corresponding measure should be reported to assess the performance of the model?

The RSME and the Adjusted R-squared values are the performance measures we need to report for the model.

To assess the performance of the models, we will perform cross-validation. More specifically, we will perform a 10-fold cross-validation. What's the idea behind the following code?

```
# Make this example reproducible
set.seed(322)

# Choose number of folds
k = 10

# Randomly order rows in the dataset
data <- pollution[sample(nrow(pollution)), ]

# Create k folds from the dataset
folds <- cut(seq(1:nrow(data)), breaks = k, labels = FALSE)
```

The idea of the code is to randomize the order of the rows in the data set and then create 10 folds or 10 subsets of data from the original dataset for cross validation testing of our model later on.

Question 2 (6 pts)

Your first model will either be a linear regression or logistic regression model. Which one is appropriate for the outcome you picked?

Our model will be a linear regression since our outcome variable is numeric.

Complete the following code to perform cross-validation for this regression model:

```
# Initialize a vector to keep track of the performance for each k-fold
perf_k <- NULL

# Use a for-loop to get performance for each k-fold
for(i in 1:k){
  # Split data into train and test data
  train_not_i <- data[folds != i, ] # train data = all observations except in fold i
  test_i <- data[folds == i, ]  # test data = observations in fold i

  # Train model on train data (all but fold i)
  train_model <- lm(value ~ lon + lat + zcta_pop + pov, data = train_not_i)
  # Performance listed for each test data (fold i)
  perf_k[i] <- sqrt(mean((
    test_i$value - predict(train_model, newdata = test_i))^2,
    na.rm = TRUE))

}
```

Write a sentence to report the average performance and how the performance varies from fold to fold. Round both measures to 0.01.

```
# Performance of the model using cross validation for 10 folds in the data
perf_k
```

```
##  [1] 2.028749 2.310223 2.180019 2.416585 2.706735 2.469577 2.172594 2.333821
##  [9] 3.059730 2.914597
```

```
mean(perf_k)
```

```
## [1] 2.459263
```

```
sd(perf_k)
```

```
## [1] 0.3357426
```

The average RSME for the linear regression model is 2.46 and the variance of the performance of the model is 0.34

Question 3 (6 pts)

Your second model will use the k-Nearest Neighbors algorithm. Which kNN function is appropriate for the outcome you picked: `knnreg` or `knn3`?

The appropriate function to use is `knnreg` since our outcome variable is numeric

Complete the following code to perform cross-validation with 5 nearest neighbors:

```
# Initialize a vector to keep track of the performance for each k-fold
perf_k <- NULL

# Use a for-loop to get performance for each k-fold
for(i in 1:k){
  # Split data into train and test data
  train_not_i <- data[folds != i, ] # train data = all observations except in fold i
  test_i <- data[folds == i, ]  # test data = observations in fold i

  # Train model on train data (all but fold i)
  train_model <- knnreg(value ~ lon + lat + zcta_pop + pov, data = train_not_i, k = 5)
  # Performance listed for each test data (fold i)
  perf_k[i] <- sqrt(mean((
    test_i$value - predict(train_model, newdata = test_i))^2,
    na.rm = TRUE))
}
```

Write a sentence to report the average performance and how the performance varies from fold to fold. Round both measures to 0.01.

```
# Performance of the model using cross validation for 10 folds in the data
perf_k
```

```
##  [1] 2.490394 2.706403 2.245503 2.857880 2.782978 2.621346 2.647528 2.606170
##  [9] 3.101083 3.013553
```

```
mean(perf_k)
```

```
## [1] 2.707284
```

```
sd(perf_k)
```

```
## [1] 0.2491656
```

The average RSME for the kNN model is 2.71 and the variance of the performance of the model is 0.25

Question 4 (6 pts)

Your third model will use the decision tree algorithm. Which function is used to build a decision tree?

The function used to build a decision tree is `rpart()`

Complete the following code to perform cross-validation for this decision tree:

```
# Initialize a vector to keep track of the performance for each k-fold
perf_k <- NULL

# Use a for-loop to get performance for each k-fold
```

```
for(i in 1:k){
  # Split data into train and test data
  train_not_i <- data[folds != i, ] # train data = all observations except in fold i
  test_i <- data[folds == i, ]  # test data = observations in fold i

  # Train model on train data (all but fold i)
  train_model <- rpart(value ~ lon + lat + zcta_pop + pov, data = train_not_i)
  # Performance listed for each test data (fold i)
  perf_k[i] <- sqrt(mean((
    test_i$value - predict(train_model, newdata = test_i))^2,
    na.rm = TRUE))
}
```

Write a sentence to report the average performance and how the performance varies from fold to fold. Round both measures to 0.01.

```
# Performance of the model using cross validation for 10 folds in the data
perf_k
```

```
##  [1] 1.616849 1.595855 1.722734 1.690954 2.146426 2.279502 1.607617 2.074972
##  [9] 2.731677 1.875213
```

```
mean(perf_k)
```

```
## [1] 1.93418
```

```
sd(perf_k)
```

```
## [1] 0.3725693
```

The average RSME of this decision tree model is 1.93 and the variance of the performance of the model is 0.37.

Question 5 (3 pts)

Comparing the cross-validation for each of the three models, which model appears to perform better? Why?

The model that appears to perform the best out of the three is the decision tree model because it has the lowest RSME and a comparable variance for model performance across the three models.

Formatting: (1 pt)

Make sure the names of all group members are included at the beginning of the document.

Knit your file! You can knit into pdf directly or into html. Once it knits in html, click on `Open in Browser` at the top left of the window pops out. Print your html file into pdf from your browser.

Any issue? Ask other classmates or TA!

Finally, remember to select pages for each question when submitting your pdf to Gradescope and to identify your group members.