Assignment 1

Junbin Wu

2024-03-10

Content

1. Model functions (in chronological order)

1.1 Simple Linear Model

```
# input a training data set
# output predictions

sim_liner_mod <- function(train) {
  # use `52` as the independent varaible, and `281` as the dependent variable.
  fit <- lm(`281` ~ `7`+`52`+`61`+`62`, data = train)
  return(fit)
}
```

1.2 All zeros

By predicting all outcomes as zeros, the MSE is even better than 2.1 Linear Model, which indicate that a simple linear regression would not work.

1.3 Decision tree

```
# input a training data set
# output a fit model
deci_tree <- function(train) {
  tree <- rpart(`281` ~ ., data = train, method = "anova")
  return(tree)
}
```

2. Model Evaluation functions

2.1 MSE calculation

```
#input a vector of prediction value, the test data set
#output a mse
MSE <- function(prediction, test) {
  mse <- mean((prediction - test$`281`)^2)
  return(mse)
}
```

3. Main functions

The Main function calls all other functions above intermittently for the purpose of high efficiency and maintenance, due to high cohesion and low coupling.

3.1 Library loading

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.2.1
```

```
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.2.3
```

```
## Loading required package: Matrix
```

```
## Warning: package 'Matrix' was built under R version 4.2.1
```

```
## Loaded glmnet 4.1-8
```

```
library(rpart)
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 4.2.3
```

3.2 Data loading

The data used here have been pre-processed. I added headers 1-281 for every data set. For the training set, I removed all duplicates, so the total observations come down from 52,397 to 49,203.

```
train_set <- read_csv("data/Processed Data Set/blogData_train duplicate removed.csv")
```

```
## Rows: 49203 Columns: 281
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (281): 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
final_test_set <- read_csv("data/Processed Data Set/blogData_test.csv")
```

```
## Rows: 214 Columns: 281
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (281): 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
test_set_201 <- read_csv("data/Processed Data Set/blogData_test-2012.02.01.00_00.csv")
```

```
## Rows: 115 Columns: 281
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (281): 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

3.3 Models running

```
predictions <- predict(deci_tree(train_set), test_set_201 , type = "matrix")
```

3.4 Model evaluation

```
MSE(predictions,test_set_201)
```

```
## [1] 730.9276
```

3.5 Output-as-a-CSV

```
outputcsv <- data.frame(ID = c(0:213))

outputcsv$num_comments <- as.vector(predict(deci_tree(train_set), final_test_set , type = "matrix"))

write.csv(outputcsv, "csv_for_submission/031401.csv", row.names = FALSE)
```