Midterm II Study Guide and Review

Deepak Bastola

2023-05-21

Midterm II Study Guide

Format: In Class with open-ended questions.

Two-sided Cheat-sheet allowed (A4 paper) and a basic calculator allowed.

- You may use a calculator
- You are not permitted to use a laptop or classroom computer.

Topics

- The exam covers functions, iterations, basic string manipulations, kNN, k-means, logistic regression and random forests machine learning algorithms (through Wed. 05/24)
- Shiny and web scraping tools will not be on the exam!!
- You will be tested on your conceptual understanding of the tools and algorithms, the accuracy metrics, and the associated construction of the workflow in R that we have discussed in the class. I will not make you write extremely complicated code from scratch, but be prepared to write small chunks of code. Additional ways I could assess your understanding of R include (but are not limited to):
  - Identifying the error in written code.
  - Putting lines of code in order to complete a specified task.
  - Describing the output resulting from a code/code-chunk.

Your name:

Questions

Q1

Given below are the monthly deaths from bronchitis, emphysema and asthma in the UK from 1974 to 1979.

```
knitr::kable(mydata)
```

year	jan	feb	mar	apr	may	jun	jul	aug	sep	oct	nov	dec
1974	3035	1721	2933	1607	2787	1489	3102	1498	2815	1529	3084	1461
1975	2552	1524	2889	1545	3891	1300	2294	1361	3137	1366	2605	1354

year	jan	feb	mar	apr	may	jun	jul	aug	sep	oct	nov	dec
1976	2704	1596	2938	1396	3179	1356	2385	1346	2679	1357	2573	1333
1977	2554	2074	2497	1787	2011	1653	2444	1564	1969	1570	2143	1492
1978	2014	2199	1870	2076	1636	2013	1748	1640	1870	1535	1693	1781
1979	1655	2512	1726	2837	1580	2823	1554	2293	1633	2491	1504	1915

a. Write a for loop that will return a vector of the ratio of the mean value to the median value for columns 2-13 in `mydata` (shown above).

b. Describe what is returned by the code below, including the type of R object produced, the length or dimension of the object, and the information contained in the object.

```
map_dbl(mydata %>% select(-1), mean) %>% mean()
```

c. Describe what is returned by the code below, including the type of R object produced, the length or dimension of the object, and the information contained in the object.

```
ratio_fun <- function(x) quantile(x, probs= c(0.25, 0.5, 0.75))
map_dfr(mydata %>% select(-1), ratio_fun, .id = "month")
```

d. Describe what is returned by the code below, including the type of R object produced, the length or dimension of the object, and the information contained in the object.

```
map_dfc(mydata %>% select(-1), ratio_fun)
```

e. Describe what is returned by the code below, including the type of R object produced, the length or dimension of the object, and the information contained in the object.

```
lapply(mydata %>% select(-1), ratio_fun) %>%
  unlist()
```

Q2

What do the following codes do? Provide a thorough and intuitive (2-3 sentences) description of the output from each of the following R chunks.

a

Consider the function below. Give the output produced by `myfun(3)` and explain how you arrived at your answer.

```
myfun <- function(x) {
  if (x < 3 | x > 3){
  rep("hi", x)
  } else{
  rep("bye", x)
  }
}
```

b

Consider the following list called `weekly_sales` and the function called `daily_summary`. Based on these objects, answer what type of data object is weekly_sales_summary and what is stored in it. Be specific in your answer.

```
weekly_sales <- list(
  week1 = c(120, 130, 90, 150, 170, 200, 80),
  week2 = c(100, 200, 150, 180, 170, 160, 100),
  week3 = c(90, 110, 120, 90, 80, 100, 110),
  week4 = c(200, 210, 230, 240, 250, 230, 220)
)


daily_summary <- function(week_sales) {
  return(ifelse(week_sales < 120, "low",
                ifelse(week_sales >= 120 & week_sales <= 200, "average", "high")))
}

weekly_sales_summary <- lapply(weekly_sales, daily_summary)
```

Q3

Consider the following string:

```
my_string <- "A 5.1 magnitude earthquake strikes near San Jose, US Geological Survey 10 reports.
There are myriad implications of this Survey according to Channel 10 news"
```

a

Carefully explain what `str_view_all` will highlight in `x` for the given pattern.

```
pattern <- "[\\d?.\\d]"
str_view_all(my_string, pattern)
```

b

Explain how you would use regular expression to count the number of words in x. Include the actual commands you would need to accomplish this task.

```
str_replace_all(my_string, pattern = "[\\d\\.]+", "") %>%
  str_extract_all(pattern = "\\w+") %>%
  unlist() %>%
  length()
```

c

What does the following command do?

```
pattern <- "\\d+"
str_replace_all(my_string, pattern, "X")
```

d

Suppose you want to place a period at the end of the my_string only if it doesn't already have one. The replacement attempt below contains one mistake. Write down the correct command that can place a period at the end of `my_string`.

```
pattern <- "$?<!\\."
str_replace_all(my_string, pattern, ".")
```

e

Use the negative lookbehind operator to remove all the occurrences of word 'Survey' only when it is not preceded by 'Geological'. Provide the commands to accomplish this.

```
pattern <- "(?<!Geological )Survey"
str_replace_all(my_string, pattern, "study")
```

Q4 : K-nearest neighbor

Consider the `Smarket` dataset, which provides daily percentage returns for the S&P 500 stock index from 2001 to 2005. Your task is to fit a K-nearest neighbor model to predict the direction of stock returns.

```r
set.seed(1234)

data_Smarket <- as_tibble(Smarket)
split <- initial_split(data_Smarket, strata = Direction, prop = 4/5)
Smarket_train <- training(split)
Smarket_test <- testing(split)

# glimpses of data
glimpse(Smarket_train)
## Rows: 999
## Columns: 9
## $ Year      <dbl> 2001, 2001, 2001, 2001, 2001, 2001, 2001, 2001, 2001, 2001, ~
## $ Lag1      <dbl> 1.032, 1.392, -0.498, 0.546, 0.359, -0.623, 1.183, -0.865, 0~
## $ Lag2      <dbl> 0.959, 0.213, 0.287, -0.562, -1.747, -0.841, -1.334, 1.183, ~
## $ Lag3      <dbl> 0.381, 0.614, 1.303, 0.701, 0.546, -0.151, -0.623, -1.334, -~
## $ Lag4      <dbl> -0.192, -0.623, 0.027, 0.680, -0.562, 0.359, -0.841, -0.623,~
## $ Lag5      <dbl> -2.624, 1.032, -0.403, -0.189, 0.701, -1.747, -0.151, -0.841~
## $ Volume    <dbl> 1.4112, 1.4450, 1.2580, 1.1188, 1.0130, 1.1072, 1.0391, 1.07~
## $ Today     <dbl> -0.623, -0.403, -0.189, -1.747, -0.151, -1.334, -0.865, -0.2~
## $ Direction <fct> Down, Down, Down, Down, Down, Down, Down, Down, Down, Down, ~
glimpse(Smarket_test)
## Rows: 251
## Columns: 9
## $ Year      <dbl> 2001, 2001, 2001, 2001, 2001, 2001, 2001, 2001, 2001, 2001, ~
## $ Lag1      <dbl> 0.381, 0.614, 0.213, 0.287, 0.701, -0.562, -0.151, -0.841, -~
## $ Lag2      <dbl> -0.192, -0.623, 0.614, 1.303, 0.680, 0.701, 0.359, -0.151, -~
## $ Lag3      <dbl> -2.624, 1.032, -0.623, 0.027, -0.189, 0.680, -1.747, 0.359, ~
## $ Lag4      <dbl> -1.055, 0.959, 1.032, -0.403, -0.498, -0.189, 0.546, -1.747,~
## $ Lag5      <dbl> 5.010, 0.381, 0.959, 1.392, 0.287, -0.498, -0.562, 0.546, 0.~
## $ Volume    <dbl> 1.19130, 1.20570, 1.34910, 1.30900, 1.14980, 1.29530, 1.0596~
## $ Today     <dbl> 0.959, 0.213, 1.392, -0.498, -0.562, 0.546, -0.841, -0.623, ~
## $ Direction <fct> Up, Up, Up, Down, Down, Up, Down, Down, Up, Up, Up, Up, Up, ~
```

a. . The following code splits the Smarket dataset into a training set and a test set. Explain the purpose of splitting the data and what stratification accomplishes in this context.

```
Smarket_recipe <- recipe(Direction ~ Lag1 + Lag2 + Lag3 + Year + Volume,
                         data = Smarket_train) %>%
  step_center(all_predictors()) %>%
  step_scale(all_predictors()) %>%
  prep()

Smarket_knn_spec <- nearest_neighbor(mode = "classification",
                             engine = "kknn",
                             weight_func = "rectangular",
                             neighbors = 5)

Smarket_workflow <- workflow() %>%
  add_recipe(Smarket_recipe) %>%
  add_model(Smarket_knn_spec)

Smarket_fit <- fit(Smarket_workflow, data = Smarket_train)

test_features <- Smarket_test %>% select(Direction, Lag1, Lag2, Lag3, Year, Volume)
nn1_pred3 <- predict(Smarket_fit, test_features, type = "raw")
Smarket_results <- Smarket_test %>%
  select(Direction) %>%
  bind_cols(predicted = nn1_pred3) %>% mutate(Direction = as.factor(Direction))
```

b. The following trained model is used to produce a data-frame of the actual and predicted `Direction` in the test dataset. Call this data-frame `Smarket_results`. What information does `Smarket_results` contain? What is the dimension of this dataset? Explain.

```
conf_mat(Smarket_results, truth = Direction, estimate = predicted)
##           Truth
## Prediction Down Up
##       Down   50 58
##       Up     71 72
```

c. Calculate the sensitivity, specificity, accuracy, and positive predictive value of the classifier based on the confusion matrix below. Discuss the interpretation of these metrics in the context of this problem.

Q5 Miscellaneous

(a) Differentiate between supervised and unsupervised learning by providing examples of situations where each would be most appropriate.

(b) Discuss the significance of feature scaling in the K-NN algorithm and illustrate its impact with a hypothetical example.

(c) Is it guaranteed to obtain identical clustering results across different runs of the K-Means algorithm? Explain your answer considering the initialization process.

(d) What does it imply if the assignment of observations to clusters does not change between successive iterations in K-Means?

(e) (True/False) Recall is a more relevant metric than precision in scenarios where False Negatives are more detrimental than False Positives. Explain your answer.

(f) How does the concept of the 'elbow method' assist in selecting the optimal number of clusters in K-means clustering?

(g) Explain the rationale behind data preprocessing in the KNN algorithm, discussing how the algorithm's performance might be influenced by unprocessed data.

(h) (Multiple Choice) Among the following values for K in K-NN, which is most likely to cause underfitting and why?

1. 30
2. 5
3. 1

(i) Discuss how centroid initialization affects the K-means algorithm, providing examples to illustrate the potential consequences of different initialization strategies.

(j) Logistic regression is a machine learning algorithm that is typically used to predict the probability of what kind of variable? Explain the reasoning behind your choice.

(A) categorical independent variable
(B) categorical dependent variable.
(C) numerical dependent variable.
(D) numerical independent variable.

(k) Explain the role of the 'Gini index' in the Random Forest algorithm and illustrate how it is used to construct the decision trees within the forest.

(l) For a K-NN classification model, how would you expect the model's performance to change as K increases? Discuss considering both bias-variance trade-off and the model's complexity.

(m) Explain the rationale behind using 'ensemble methods' like Random Forest. How do they generally improve on the performance of individual learners?