Midterm II

2024-05-21

Your name:

Questions

Q1

Consider the following data frame `df_waste` which represents the waste generation (in tonnes) of three types of waste materials: Plastic, Metal, and Glass in a city.

```r
df_waste <- data.frame(
  day = seq(ymd("2023-01-01"), by="day", length.out=50),
  plastic = runif(50, 80, 250),
  metal = rnorm(50, 120, 40),
  glass = rnorm(50, 100, 60)
)
```

```r
head(df_waste)
##          day   plastic     metal      glass
## 1 2023-01-01 150.49954 215.54986 194.948021
## 2 2023-01-02 161.36857  79.01086 104.627406
## 3 2023-01-03 155.58448 135.22436 -27.092387
## 4 2023-01-04 152.32341 106.34783  24.511645
## 5 2023-01-05  96.80489 150.87599   2.441849
## 6 2023-01-06 119.03242 121.87534  49.211641
```

(a). (5 points)

Consider the function `calc_moving_avg` which calculates the moving average of a numeric vector. Given the vector from 1 to 8, what is the output of `calc_moving_avg(seq(1:8))`? Recall that the function uses a window width of 7 for the moving average.

```r
library(dplyr)
library(purrr)

calc_moving_avg <- function(vec, width = 7) {
  if (length(vec) < width) return(numeric(0))
  map_dbl(width:length(vec), ~ mean(vec[.x:(.x-width+1)]))
}

calc_moving_avg(seq(1:8))
## [1] 4 5
```

(b). (5 points)

Bootstrap is a statistical technique for estimating population parameters by resampling with replacement from an observed dataset using the `sample()` function in R. This method involves generating multiple small samples (bootstrap samples), each of which is drawn randomly, allowing elements to be picked multiple times.

Given the third column of `df_waste` which represents `metal` waste data, use the function `bootstrap_ci` to calculate the 95% confidence intervals for the average of the moving averages (with a window width of 7). What would typically be the output of this function call: `bootstrap_ci(df_waste[[2]], n_boot = 500)`? Describe conceptually what the resulting object would convey without needing to calculate the values. Report the length or dimension of the output.

```
bootstrap_ci <- function(vec, n_boot=500, width = 7) {
  map_dbl(1:n_boot, ~ {
    sample_vec <- sample(vec, size = length(vec), replace = TRUE)
    ma_values <- calc_moving_avg(sample_vec, width)
    if (length(ma_values) > 0) mean(ma_values) else NA
  }) %>%
  quantile(probs = c(0.025, 0.975), na.rm = TRUE)
}

bootstrap_ci(df_waste[[3]], n_boot = 500)
##     2.5%    97.5%
## 114.4058 137.6844
```

The function `bootstrap_ci` calculates the 95% confidence intervals for the average of moving averages of a numeric vector vec with a window width of 7, using 500 bootstrap samples. Conceptually, the output object would convey the lower and upper bounds of the confidence interval for the mean of the moving averages of the sampled metal waste data. Typically, the output would be a numeric vector with two elements, representing the 2.5th percentile and the 97.5th percentile of the bootstrapped mean moving averages, thus providing an interval estimate that is likely to contain the true mean moving average of the metal waste data 95% of the time.

(c). (5 points)

Using the `bootstrap_ci` function and the `df_waste` dataframe, write a code snippet to calculate the 95% confidence intervals for the average moving averages of the `plastic`, `metal`, and `glass` waste columns, excluding the `day` column. Use a bootstrap sample size of 500 for each.

```
results <- df_waste %>%
  select(-day) %>%
  map_df(~ bootstrap_ci(.x, n_boot = 500), .id = "material")

results
## # A tibble: 3 x 3
##   material `2.5%` `97.5%`
##   <chr>     <dbl>   <dbl>
## 1 plastic   148.     173.
## 2 metal     113.     138.
## 3 glass      85.0    123.
```

Q2 (5 points each)

The Shiny app below is designed to visualize data distributions. It allows users to select a dataset, choose a variable from that dataset, and display a histogram of that variable.

```r
library(shiny)
library(ggplot2)
library(babynames)
library(dplyr)

ui <- fluidPage(
  titlePanel("Baby Names Trends"),
  sidebarLayout(
    sidebarPanel(
      selectInput("year", "Choose a year:",
                  choices = unique(babynames$year)),
      selectInput("gender", "Select gender:",
                  choices = c("Male" = "M", "Female" = "F")),
      actionButton("update", "Update Chart")
    ),
    mainPanel(
      plotOutput("namePlot")
    )
  )
)

server <- function(input, output) {
  filtered_data <- eventReactive(input$update, {
    babynames %>%
      filter(year == input$year, sex == input$gender) %>%
      arrange(desc(n)) %>%
      head(10)
  })

  output$namePlot <- renderPlot({
    req(filtered_data())
    ggplot(filtered_data(), aes(x = reorder(name, n), y = n)) +
      geom_bar(stat = "identity", fill = "steelblue") +
      labs(title = "Top 10 Popular Baby Names",
           x = "Names", y = "Number of Babies") +
      theme_minimal() +
      coord_flip()
  })
}

shinyApp(ui = ui, server = server)
```

(a). Considering the need to control computational load in a data-intensive Shiny application, what is the best use of actionButton in conjunction with reactive expressions?

A) To trigger data processing only after all user inputs have been finalized, using `actionButton` to activate `eventReactive()` computations.
B) Use `actionButton` to reset user inputs to their default states without affecting any reactive expressions or outputs.
C) Implement `actionButton` as a method to bypass reactivity and directly manipulate output displays without data reprocessing.
D) `actionButton` should only be used to trigger UI updates, such as showing or hiding elements, not for controlling reactive data processes.

E) Combine `actionButton` with `isolate()` to ensure that reactive expressions are updated without user intervention.

Correct Answer: A) To trigger data processing only after all user inputs have been finalized, using actionButton to activate eventReactive() computations.

(b). Which statement best describes the roles and differences between eventReactive, `observe`, and `renderPlot` in managing the app's reactivity and user interactions?

A) `eventReactive` is used to update the data when the 'Update Chart' button is clicked, ensuring data processing is limited to this action, whereas observe would continuously monitor changes in inputs without requiring an action button.

B) `renderPlot` is an observer that automatically updates the plot whenever the `filtered_data()` reactive expression changes, regardless of any user interaction.

C) The `actionButton` is ineffective without an observe function because eventReactive does not properly capture button clicks.

D) Using `eventReactive` rather than observe with `actionButton` unnecessarily complicates the app since both serve the same purpose of updating outputs based on input changes.

E) `renderPlot` should be used with observe instead of eventReactive to improve the efficiency and responsiveness of the plot updates.

Correct Answer: A) eventReactive is used to update the data when the 'Update Chart' button is clicked, ensuring data processing is limited to this action, whereas observe would continuously monitor changes in inputs without requiring an action button.

(c) In the Shiny application provided, how would you modify the app to use `observeEvent()` for a specific scenario where you only want to update the data and re-render the plot when both the year and gender have been changed by the user?

A) Replace `eventReactive(input$update, ...)` with `observeEvent(c(input$year, input$gender), {...})` to trigger updates only when both inputs are altered.

B) Use `observeEvent(input$year & input$gender, {...})` to monitor changes in both inputs simultaneously.

C) Maintain `eventReactive` and add `observeEvent(input$update, {...})` to separately handle other UI updates.

D) `observeEvent()` cannot be used with multiple inputs and should only be used with single input actions.

E) Incorporate `observeEvent(list(input$year, input$gender), {...})` to ensure joint reactivity without using `actionButton.`

Correct Answer: A) Replace `eventReactive(input$update, ...)` with `observeEvent(c(input$year, input$gender), {...})` to trigger updates only when both inputs are altered.

Q3

This HTML code is part of a webpage containing product information, including prices and descriptions.

```
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <title>Product Page</title>
</head>
<body>
    <div id="products">
```

```
        <h1>Our Products</h1>
        <div class="product-item" data-id="1">
            <span class="product-name">Gadget</span>
            <span class="product-price">$19.99</span>
        </div>
        <div class="product-item" data-id="2">
            <span class="product-name">Widget</span>
            <span class="product-price">$25.99</span>
        </div>
    </div>
</body>
</html>
```

(a). Using R and the `rvest` package, write the code to extract and print the names of the products listed on the page.

```r
library(rvest)

html_content <- read_html('path_to_your_html_file.html')

product_names <- html_content %>%
  html_elements(".product-name") %>%
  html_text()

product_names
```

(b). Write R code using `rvest` to extract the data-id attributes of each product item.

```r
library(rvest)

html_content <- read_html('path_to_your_html_file.html')

product_ids <- html_content %>%
  html_elements(".product-item") %>%
  html_attr("data-id")

product_ids
```

Q4 (5 points each)

Consider a binary classification problem where we use a k-Nearest Neighbors (k-NN) algorithm. A confusion matrix shows the classification performance:

	Predicted Negative	Predicted Positive
Actual Negative	55	15
Actual Positive	19	91

(a). What is the accuracy of the model according to the provided confusion matrix?

A) 0.755
B) 0.811
C) 0.852
D) 0.903
E) 0.954

Correct Answer: B) 0.811

(b). In the context of k-NN, how does the use of Euclidean distance affect the classification in high-dimensional spaces?

A) Euclidean distance becomes more meaningful as the number of dimensions increases, improving classification accuracy.

B) The distance metric tends to not differentiate well between the nearest and farthest neighbors as dimensionality increases (curse of dimensionality).
C) Euclidean distance calculations become computationally cheaper as more dimensions are added.
D) Increasing dimensionality reduces the overall distance between points, leading to less meaningful neighbor calculations.
E) None of the above; Euclidean distance is independent of the dimensionality of the space.

Correct Answer: B) The distance metric tends to not differentiate well between the nearest and farthest neighbors as dimensionality increases (curse of dimensionality).

(c). When tuning k for a k-NN model, how can ROC curves be utilized to select an optimal k value?

A) Choose k that maximizes the area under the ROC curve, balancing sensitivity and (1 - specificity).
B) Select k that minimizes the area under the ROC curve to ensure maximum model complexity.
C) Use the ROC curve to find a k where the true positive rate equals the false positive rate.
D) The ROC curve should be used to select a k that maximizes both sensitivity and specificity simultaneously.
E) ROC curves are irrelevant in the context of tuning k in k-NN models.

Correct Answer: A) Choose k that maximizes the area under the ROC curve, balancing sensitivity and (1 - specificity).

(d). How is class assignment determined in k-NN?

A) The class with the most representatives among the nearest neighbors determines the class of the query point.
B) The class is assigned based on the average distance to all points within the k nearest neighbors.
C) The class is assigned randomly among the classes of the k nearest neighbors.
D) The class of the nearest single neighbor is always chosen, regardless of k.
E) The class is determined by a weighted vote based on the distance of the k nearest neighbors.

Correct Answer: A) The class with the most representatives among the nearest neighbors determines the class of the query point.

(e). Which of the following best describes the relationship between sensitivity, specificity, and the choice of k in k-NN?

A) A smaller k typically increases sensitivity by capturing finer details, but may decrease specificity due to overfitting to noise.
B) A larger k enhances specificity at the cost of decreasing sensitivity, as the model generalizes by considering more neighbors.
C) Both sensitivity and specificity are maximized at an intermediate value of k, which perfectly balances the local and global data structures.
D) Sensitivity and specificity are independent of k; they are more influenced by the dataset's intrinsic characteristics.
E) A larger k always increases both sensitivity and specificity by utilizing more data for decision-making.

Correct Answer: A) A smaller k typically increases sensitivity by capturing finer details, but may decrease specificity due to overfitting to noise.

(f). How does averaging the results from k-fold cross-validation affect the selection of k in k-NN?

A) It identifies a k that performs consistently well across different subsets of the data, enhancing generalization.
B) Cross-validation tends to select smaller values of k, as they usually yield higher accuracy on smaller test sets.

C) Averaging results tends to favor larger k values, as variability in the validation results is minimized.
D) The method biases the selection towards k values that overfit the training data, as averaged results mask the variance.
E) Cross-validation is not useful in k-NN because the algorithm is non-parametric and does not generalize beyond its training set.

Correct Answer: A) It identifies a k that performs consistently well across different subsets of the data, enhancing generalization.

Q5 Logistic Regression (5 points each)

Consider a logistic regression model used for a binary classification problem that predicts whether a student will pass (1) or fail (0) an exam. The model uses one predictor: the student's hours of study per week. The logistic regression equation is given by:

$$\log\left(\frac{p}{1-p}\right) = -3 + 0.5 \times \text{ Hours of Study}$$

where p is the probability that the student passes the exam.

(a). Given a student studies 10 hours per week, calculate the probability that the student will pass the exam.

A) 0.8808
B) 0.9526
C) 0.5
D) 0.6225
E) 0.7311

(b). If we want the probability that a student passes the exam to be at least 70% before we classify them as likely to pass, what should the study hours threshold be?

A) 16 hours
B) 14 hours
C) 12 hours
D) 10 hours
E) 8 hours

(c). The coefficient of hours of study in the logistic regression model is -0.5. What is the interpretation of this coefficient?

A) The odds of passing increase by a factor of `exp(0.5)` for each additional hour of study.
B) The probability of passing increases as study hours increase.
C) The odds of passing are doubled for each additional hour of study.
D) The probability of passing increases by 50% for each additional hour of study.
E) The odds of passing increase by 50% for each additional hour of study.

(d). In logistic regression, if the outcome variable represents "pass" (1) or "fail" (0) in an exam, and a significant predictor has a negative coefficient, what does this imply about the relationship between the predictor and the likelihood of passing?

A) As study hours increase, the likelihood of passing the exam increases.
B) As study hours increase, the likelihood of passing the exam decreases.
C) There is no relationship between study hours and the likelihood of passing.
D) The predictor inversely predicts the likelihood of passing but does not affect the odds.
E) The effect of study hours on the likelihood of passing is ambiguous and requires further testing.