Midterm II

2024-03-03

Your name:

Questions

Q1 Logistic Regression

We're interested in estimating the probability of developing hypertension based on systolic blood pressure using logistic regression. The logistic regression model has been trained, and the coefficients are $\beta_0 = -3.78$ for the intercept and $\beta_1 = 0.021$ for systolic blood pressure.

a. Calculate the odds for a systolic blood pressure of 135 mmHg. (5 points)

b. Convert the odds in part `a` to the probability of developing hypertension. (5 points)

Table 1: Confusion Matrix

	Predicted Negative	Predicted Positive
Actual Negative	65	25
Actual Positive	15	95

Q2 Classification using k-NN

Consider a binary classification problem where we use a k-Nearest Neighbors (k-NN) algorithm. We have a confusion matrix as above, which shows the classification performance:

a. (5 points)

Given the confusion matrix, calculate the model's precision and recall for the "Positive" class. What does this imply about the model's performance in identifying positive cases?

b. (5 points)

Discuss the influence of the k parameter size on the occurrence of overfitting and underfitting within the k-Nearest Neighbors (k-NN) algorithm. How can adjusting k help in achieving a balance to optimize model performance?

Q3 Loops and functions

Consider the following data frame `df_waste` which represents the waste generation (in tonnes) of five types of waste materials: Plastic, Metal, Glass, Paper, and Organic in a city over 12 months in a city.

```
glimpse(df_waste)
## Rows: 12
## Columns: 6
## $ month   <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
## $ plastic <dbl> 150, 120, 130, 200, 210, 180, 190, 220, 250, 260, 270, 280
## $ metal   <dbl> 100, 110, 120, 130, 140, 150, 130, 120, 110, 115, 125, 135
## $ glass   <dbl> 200, 190, 180, 170, 160, 150, 160, 170, 180, 185, 190, 195
## $ paper   <dbl> 90, 85, 95, 100, 105, 110, 115, 120, 125, 130, 135, 140
## $ organic <dbl> 300, 310, 320, 330, 340, 350, 360, 370, 380, 390, 400, 410
```

a. (10 points)

Given `df_waste` write a loop to calculate the monthly average waste generation for each type of material. Your resulting output should be a data frame with each row representing the average waste generation for a month and each column representing a type of material.

```
#  your r-code
```

b. (10 points)

Given the dataset `df_waste` representing monthly waste generation (in tonnes) for various materials and the vector `seasons` indicating the corresponding season for each month, calculate the average waste generation for each type of material across the four seasons: Winter, Spring, Summer, and Autumn. The seasons are defined as follows:

- Winter: December, January, February
- Spring: March, April, May
- Summer: June, July, August
- Autumn: September, October, November

The `seasons` vector provided is as follows:

```
seasons <- c("Winter", "Winter", "Spring", "Spring", "Spring", "Summer",
             "Summer", "Summer", "Autumn", "Autumn", "Autumn", "Winter")
```

Use `lapply` or `map` functions from R to compute the average waste generation for each material (`Plastic`, `Metal`, `Glass`, `Paper`, `Organic`) for each `season.` The expected output should be a structured data object (like a `list` or a `data frame`) where each entry or row corresponds to a season, showing the average waste generation for each material during that season.

```
#  your r-code
```

Q4 Shiny (10 points)

The Shiny application below contains three errors. Identify and correct these errors. Additionally, describe the purpose and functionality of this Shiny app in broad terms.

```
library(shiny)
library(gapminder)
dplyr::glimpse(gapminder)
## Rows: 1,704
## Columns: 6
## $ country   <fct> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanistan", ~
## $ continent <fct> Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, ~
## $ year      <int> 1952, 1957, 1962, 1967, 1972, 1977, 1982, 1987, 1992, 1997, ~
## $ lifeExp   <dbl> 28.801, 30.332, 31.997, 34.020, 36.088, 38.438, 39.854, 40.8~
## $ pop       <int> 8425333, 9240934, 10267083, 11537966, 13079460, 14880372, 12~
## $ gdpPercap <dbl> 779.4453, 820.8530, 853.1007, 836.1971, 739.9811, 786.1134, ~
```

```
ui <- fluidPage(
  titlePanel("Gapminder Data for Selected Country"),
  selectInput("countryInput", "Select a Country:", choices = unique(gapminder$country)),
  actionButton("actionButton", "Show Latest Data"),
  htmlOutput("countryInfo")
)

server <- function(input, output) {
  latestData <- eventReactive(input$showButton, {
    selectedCountry <- isolate(input$countryInput)
    gapminder %>%
      filter(country == selectedCountry) %>%
      tail(1)
  })

  output$countryInfo <- renderUI({
    HTML(stringr::str_c(
      "Country:", latestData$country, " ",
      "Life Expectancy:", round(latestData$lifeExp, 2), "years"
    ))
  })
}

shinyApp(ui, server)
```

Q5 String manipulations

a. (5 points)

Given a vector of words, write a function named `find_vowel_5_letter_words` that identifies all five-letter words starting and ending with a vowel. The function should return the indices of these words within the original vector. For this problem, assume vowels are "a", "e", "i", "o", and "u", and consider case insensitivity.

```
```

b. (5 points)

```
comments <- c("Loving the new #season!", "Can't wait for #Friday!")
```

Write a command to remove all instances of # from the `comments` vector.

```
```

c. (5 points)

```
dates <- c("01012023", "15032024")
```

Write a command to insert dashes (-) into the dates vector to achieve the desired format.

d. (10 points)

```
info <- "John Doe:john.doe@example.com, Jane Smith:jane.smith@example.com"
```

Use the `stringr` functions to separate the names and email addresses into two vectors, `names` and `emails`.

Q6 Miscellaneous: Multiple Choice (5 points each)

a. When the K-Means clustering algorithm reaches a point where the assignment of observations to clusters remains constant across successive iterations, it indicates:

A) The algorithm has potentially reached a local minimum, but not necessarily the global optimum.
B) The algorithm has converged, indicating stability in cluster assignment and optimal clustering has been achieved.
C) The algorithm requires re-initialization as it has likely converged to a suboptimal solution.
D) The clustering process is incomplete and requires more iterations for accurate cluster assignment.

b. In the context of the k-Nearest Neighbors (k-NN) algorithm, consider how the choice of k impacts the model's bias and variance. Select the correct statement from the options below:

A) Increasing the value of k decreases both bias and variance, leading to a universally better model performance across different datasets.
B) Decreasing the value of k to 1 minimizes the model's variance while maximizing its bias, as the prediction is entirely based on the nearest neighbor.
C) Increasing the value of k generally increases the model's bias but reduces its variance, as the classification decision is based on a larger, more generalized set of neighbors.
D) A very large value of k (approaching the size of the training set) results in a model with low bias and high variance, as it becomes too sensitive to the noise in the training data.

c. The utilization of cross-validation techniques in machine learning models:

A) Exclusively mitigates overfitting by training the model on multiple subsets of the data.
B) Can effectively eliminate the need for a separate test set by using the training data for validation.
C) Helps in mitigating both overfitting and underfitting by validating the model's performance on different subsets of the dataset.
D) Guarantees improvement in model performance on unseen data by optimizing hyperparameters.

d. In logistic regression, odds are defined as:

A) The ratio of the probability of the event occurring to the probability of it not occurring.
B) A direct measure of probability that an event occurs, identical in interpretation.
C) The logarithmic transformation of probabilities for easier computation.
D) The probability of an event occurring divided by the probability of all other events.

e. In logistic regression, changing the decision threshold (default is 0.5) affects the model's sensitivity and specificity. Which of the following statements is true when increasing the decision threshold above 0.5 ?

A) Sensitivity increases while specificity decreases.
B) Both sensitivity and specificity increase.
C) Sensitivity decreases while specificity increases.
D) Both sensitivity and specificity decrease.