Assignment 3

Maximum points: 20 Due date: 6th Dec. 2024 at 23:00

Guidelines

- Individual Assignment: This is an individual assignment. Please do not seek help from others or collaborate with classmates.
- Problem Understanding: Carefully read the problem description and each question before starting your analysis.
- Choice of Software: You are free to use either R or Python to perform your analysis.
- Code Requirements: Your code should be well-structured and free of errors. Please use clear and descriptive variable names.
- Deliverables: Submit the following two files on Canvas:
    1. Code file (.R/.Rmd/.py/.ipynb): This file should contain the complete executable code for the analysis.
    2. Report file (.pdf): This document should contain your answers to all the questions asked below. Please use the solution template provided on Canvas (Assignment_3_Template.docx) to answer the questions. On the cover page, please write your name, student number, and date. After writing your answers, submit the document as a pdf file.
- Plagiarism Warning: Ensure that all the code and analysis are your original work. Plagiarism will lead to disqualification and academic consequences.
- Late Submissions: Late submissions will be accepted but penalized.
- Points Distribution: The distribution of points and grading rubric for the questions is provided at the end of this document.

Use the "insurance.csv" dataset from Assignment 2 to answer questions 1 and 2:

1. Create a column binary_charges (as you did in Question 9 of Assignment 2). Build a logistic regression model (as you did in Question 10 of Assignment 2) using age, gender, bmi, children, smoker, and region as predictor variables and binary_charges as the target variable. For this logistic regression model:
    a. Set a threshold (cut-off probability) of **0.2** and use the model to classify all the training observations. Report the confusion matrix.
    b. Set a threshold of **0.8** and use the model to classify all the training observations. Report the confusion matrix.
    c. Create an ROC plot for the logistic regression model and report the area under the curve (AUC). Explain the meaning of AUC in 1-2 lines.

2. Suppose we decide to use a threshold of **0.5** for the logistic regression model. Do the following:
    a. Report the confusion matrix.
    b. Report the accuracy of the model and explain it (1-2 lines).
    c. Report the precision of the model and explain it (1-2 lines).
    d. Report the sensitivity of the model and explain it (1-2 lines).
    e. Report the specificity of the model and explain it (1-2 lines).
    f. Report the true positive rate of the model and explain it (1-2 lines).
    g. Report the false positive rate of the model and explain it (1-2 lines).

To answer the questions below, access the "books.csv" file from Canvas. The dataset contains 500 transactions of customers across 5 categories of books. An entry of 1 indicates purchase and an entry of 0 indicates no purchase.

3. Compute the following (based on the 5 categories):
    a. Euclidean distance between customers 245 and 431
    b. Manhattan distance between customers 82 and 197
    c. Centroid of the first 50 customers
4. Which two genres of books have:
    a. the highest co-occurrence?
    b. the lowest co-occurrence?
5. Suppose we cluster the customers based on the total number of books purchased. What is the size of each cluster?
6. Compute the **support** of the following itemsets:
    a. {fiction}
    b. {non_fiction}
    c. {fiction, self_help}
7. Compute the **confidence** of the following association rules:
    a. {fiction} → {mystery}
    b. {non_fiction} → {self_help}

c. {fiction, self_help} → {childrens_books}

8. Compute the **lift** of the following association rules:
   a. {fiction, self_help} → {childrens_books}
   b. {fiction} → {non_fiction}
   c. {non_fiction} → {self_help}
9. Explain the meaning of the following:
   a. Support of {fiction, self_help}
   b. Confidence of {fiction, self_help} → {childrens_books}
   c. Lift of {fiction, self_help} → {childrens_books}

Points distribution and grading rubric

Question	Point(s)	Grading criteria		
		Correct answer	Incorrect answer, but coding logic partially correct	Incorrect answer and coding logic
1a	1	1	0.5	0
1b	1	1	0.5	0
1c	1	1	0.5	0
2a	1	1	0.5	0
2b	0.5	0.5	0.25	0
2c	0.5	0.5	0.25	0
2d	0.5	0.5	0.25	0
2e	0.5	0.5	0.25	0
2f	0.5	0.5	0.25	0
2g	0.5	0.5	0.25	0
3a	0.5	0.5	0.25	0
3b	0.5	0.5	0.25	0
3c	1	1	0.5	0
4a	1	1	0.5	0
4b	1	1	0.5	0
5	1	1	0.5	0
6a	0.5	0.5	0.25	0
6b	0.5	0.5	0.25	0
6c	0.5	0.5	0.25	0
7a	0.5	0.5	0.25	0
7b	0.5	0.5	0.25	0
7c	1	1	0.5	0
8a	1	1	0.5	0
8b	1	1	0.5	0
8c	1	1	0.5	0
9a	0.5	0.5	0.25	0
9b	0.5	0.5	0.25	0
9c	0.5	0.5	0.25	0