Assignment Instructions: Assignment 3

Purpose

The purpose of this assignment is to use Naive Bayes for classification.

Directions

The file UniversalBank.csv contains data on 5000 customers of Universal Bank. The data include customer demographic information (age, income, etc.), the customer's relationship with the bank (mortgage, securities account, etc.), and the customer response to the last personal loan campaign (Personal Loan). Among these 5000 customers, only 480 (= 9.6%) accepted the personal loan that was offered to them in the earlier campaign. In this exercise, we focus on two predictors: Online (whether or not the customer is an active user of online banking services) and Credit Card (abbreviated CC below) (does the customer hold a credit card issued by the bank), and the outcome Personal Loan (abbreviated Loan below).

Partition the data into training (60%) and validation (40%) sets.

A. Create a pivot table for the training data with Online as a column variable, CC as a row variable, and Loan as a secondary row variable. The values inside the table should convey the count. In R use functions *melt()* and *cast()*, or function table(). In Python, use panda dataframe methods *melt()* and *pivot()*.
B. Consider the task of classifying a customer who owns a bank credit card and is actively using online banking services. Looking at the pivot table, what is the probability that this customer will accept the loan offer? [This is the probability of loan acceptance (Loan = 1) conditional on having a bank credit card (CC = 1) and being an active user of online banking services (Online = 1)].
C. Create two separate pivot tables for the training data. One will have Loan (rows) as a function of Online (columns) and the other will have Loan (rows) as a function of CC.
D. Compute the following quantities [P(A | B) means "the probability ofA given B"]:
    i. P(CC = 1 | Loan = 1) (the proportion of credit card holders among the loan acceptors)
    ii. P(Online = 1 | Loan = 1)
    iii. P(Loan = 1) (the proportion of loan acceptors)
    iv. P(CC = 1 | Loan = 0)
    v. P(Online = 1 | Loan = 0)
    vi. P(Loan = 0)
E. Use the quantities computed above to compute the naive Bayes probability P(Loan = 1 | CC = 1, Online = 1).
F. Compare this value with the one obtained from the pivot table in (B). Which is a more accurate estimate?
G. Which of the entries in this table are needed for computing P(Loan = 1 | CC = 1, Online = 1)? Run naive Bayes on the data. Examine the model output on training data, and find the entry that corresponds to P(Loan = 1 | CC = 1, Online = 1). Compare this to the number you obtained in (E).

File Attached: UniversalBank.csv

Learning Outcomes

The assignment will help you with the following course outcomes:

1. Think critically about how to use machine learning algorithms to solve a given business problem.
2. Know how to formulate business problems and identify relevant data to use in modeling frameworks.
3. Know how to evaluate the appropriateness and estimate the performance of using Naive Bayes for a given task.
4. Know how to use software tools (such as R) effectively to implement Naive Bayes.
5. Foster the communication and presentation of statistical results and inferences.

Requirements

All due dates are included in the Assignment Schedule.

General Submission Instructions

All work must be your own. Copying other people's work or from the Internet is a form of plagiarism and will be prosecuted as such.

1. Create a new folder called **Assignment_3** in your previously created GitHub repository.
2. If you are using R, then upload the R Markdown file, the knitted pdf/html file, and any other data file you might have used for the assignment.
3. If you using Python, then share the Jupyter/Google Colab notebook in our Assignment_3 folder on GitHub

Provide the link to your git repository in Canvas for the assignment.