Programming Assignment 2: Computing AUC and Implementing a Naïve-Bayesian Classifier

This assignment has two problems.

1. Write a program to compute AUC (area under the ROC curve).

    Input: The input consists of two files. In the first file, is like the following, where the first number is the id of the test instance, and the second number is the probability for the positive class.

    1, 0.56
    2, 0.70
    3, 0.33
    …

    The second file contains the ground-truth labels and is of the following form, where the first number is also the id of the test instance, but the second item is the ground-truth class label, P for positive and N for negative.

    1, P
    2, N
    3, P
    …

    Output: the AUC value.

2. Implement a Naïve Bayesian Classifier. Please use a variable to switch on/off smoothing as we want you to print the probabilities in each option. The input data may have any number of attributes. The last attribute is the class attribute. The attribute values are all discrete.

    Input data: The input data will be given in a file in the following format, where the first row gives the attribute names and the rest are the data instances.

    A1, A2, A3, C
    a, b, t, 1
    x, y, f, 0

    Output
    (1). Your program should print all the probabilities in the following format, where the ordering of the conditional probabilities are not important.

    $P(C=1) = 0.4$
    $P(C=0) = 0.6$
    $P(A1= a \mid C=1) = 0.3$
    $P(A1= z \mid C=1) = 0.4$
    $P(A1= a \mid C=0) = 0.2$
    $P(A1= z \mid C=0) = 0.3$
    …

$P(A2 = b \mid C = 1) = 0.2$

…

(2). Given a sequence of test instances in a file in the same format as the input, predict the class for each of them and write the output class in a text file in the same order as the input with one instance per line as follows, where 1 and 0 are class labels,

1

0

1

..

Group project: This is a group project. Each group can have no more than 2 students.

Code submission: The code must be well commented. A readme file must accompany the **code** for each program to describe each function in the program.

Programming language: You may use any language of your choice.

Demo date: April 8, 2022, from 9:00am – 12:00pm.

Deadline for code and readme file submission on Blackboard: April 7, 2022, 11:59pm (CST)