EE6435 Homework 3

Points: 80.
Out: Oct. 15, 2020 (Thursday)
Due: **11:59PM**, Oct. 28, 2020 (Wed.). **No late homework will be accepted**.
Handin method and requirement: name your notebook file (.ipynb) as **yourlastname-firstname-studentID-hw3.ipynb**. For example, if your name is Amy Zhang, the file should be named as zhang-amy-5678910-hw3.ipynb. Also, attach an html file (generated by the notebook file) with your notebook using the naming rule: **yourlastname-firstname-studentID-hw2.html**. **(-10 points if missing these files)**

You are allowed to form a group of size <=2 for this homework. In that case, you two will get the same grade for this homework. If you choose to do this by yourself, +10 points (that means, you could get 90)
==========

Homework overview: Implement the Naïve Bayes Classifier (NBC) for the given training data, apply it to the given testing data, and report the accuracy on both the training and testing data.
You need to use two methods to learn the conditional probabilities for continuous attributes. One is based on a parametric distribution such as normal distribution. The other is discretization (you need to figure out how to do this and describe it in the report).
You need to implement NBC yourself. Calling any API or existing functions of NBC will lead to 0 for this homework.

Data:
Canvas→files/homework/hw3 → training sample.csv, testing sample.csv

The data format and meanings are self-explanatory.

Requirement and grading:

1. (20 pts) Submit two programs with the two different methods of computing conditional probabilities for continuous attributes. Each python program must take two files as inputs. One is the training data and the other is the testing data. In practice, you can separate the training and testing. But in this homework, we will look at them altogether. Don't hardcode the input files because we can change the contents of the test data (while keeping the same format).
    a. If you did not follow the naming and submission requirement (ipynb+html files), -10 pts
    b. If you hardcode your input files, -10 pts.
    c. Please use the given file format directly. If you feel you must convert the format, you should wrap the conversion part in a function and generalize it to other files. If any manual intervention is needed, it is called "hard code" and we will deduct 10 points (refer to b).

2. (15 pts) Required outputs:

a. Part 1: output the **accuracy** on both the training data and the testing data using the following format:
    i. The accuracy on training data is ____. The accuracy on testing data is _____.
b. Part 2: for the first five samples in the test files, output the results of P(class | attributes) for all classes using the following format.
    i. P(class 1| sample 1)=_____. P(class 2 |sample 1)=____ ..
    ii. P(class 1| sample 2)=_____. P(class 2 |sample 2)=_____...
    iii. …
c. Part 3: output the **learned conditional probabilities** P(attribute | class) for each class and attribute. For continuous attributes, you should output the **normal distribution parameters and the discretization results**.
d. The output should only include the required information. Don't include any debug information in the final program
e. For any new test files, your code should be able to generate the outputs for the new input. Unless you hardcode your input files, you should not have this issue.

3. (10 pts) Test your program using the given data. 10 pts for correct outputs.
4. (15 pts) Test your program using a different data file (same format, but different number of samples and values). 20 pts for correct outputs.

5. (20 pts) Submit one report in pdf format containing the following information:
    a. Describe your method of discretization of the continuous attributes
    b. Compare the two methods of computing conditional probabilities of continuous attributes. Which one is better and why?
        i. **You should not use Chinese in your report**. English writing is part of the training of this course.
        ii. **Be concise and accurate**. **Just include the required information** (e.g. don't print the data you read, debugging information in the report).
        iii. **5 pts are reserved for clarity**. If we need to read multiple times to understand your report, -5.
6. Note that items 4 and 5 depend on whether your programs can run correctly for our test files. If the codes have bugs and we cannot generate the output, you cannot get credit for items 4 and 5.