United International University
Department of Computer Science and Engineering
CSI 416 Pattern Recognition Laboratory,
Assignment 1, Fall 2021

Problem Description

- First learn about ***Naive Bayes*** from this link:

  Naive Bayes Classifiers (GeeksforGeeks)

- Your task is to implement the ***Naive Bayes*** Model.
- You have to implement two functions, a ***fit*** function and a ***predict*** function.
- To implement the ***fit*** function -
  - Take two parameters- ***features*** and ***labels***
  - Count how many labels are **0** and how many labels are **1**
  - Calculate the probabilities ***P*(*y* = 0)** and ***P*(*y* = 1)** and store them
  - Now traverse **each column** and for each column-
    - If the column is **categorical**
      - Identify the **unique** values for that column
      - for **each unique value**, count **how many** values have the label **0** and how many values have the label **1**
      - Calculate the necessary **probabilities** and store them using a data structure of your choice (e.g. Dictionary)
    - If the column is **numeric** (***bonus***)
      - Assume that the numeric feature follows ***normal*** distribution.
      - Identify which **rows** have label **0** and which ones have label **1**
      - For the rows with label **0**, identify the **mean** and **standard deviation** for that column. Do the same for the rows with label **1**
      - Store the **means** and **standard deviations** using a data structure of your choice (e.g. Dictionary)

- To implement the ***predict*** function -
  - – Take one parameter- ***features***
  - – For **each row** of the dataset, identify the **features**, calculate the **probabilities** and **classify**
  - – Store all the **predictions** in a list
  - – Return the list
- Now, download the **dataset** from this link:

  Telco Customer Churn

- The **label** column of this dataset is named ***Churn***
- In the ***Churn*** column, replace the labels according to the following:
  - **No** : ***0***
  - **Yes** : ***1***
- Drop the **unnecessary** columns. (There is at least one such column in the given dataset.)
- Drop the **rows** that have **missing** values in any column (if there are any such)
- Identify which columns are **categorical** and which are **numeric**
- Drop the **numeric** columns if you are **not attempting** the **bonus** part.
- **Split** the dataset (***80% training***, ***20% testing***) both with and without **stratification** (use **random_state** = **911**)
- **Do not scale** the dataset here. Naive Bayes do not need scaling of data.
- Now **train** and **test** the dataset using the functions you have written.
- Determine ***accuracy***, ***precision***, ***recall*** and ***F1 score*** (You **can not** use library functions for this task.)
- **Print** the scores for both **with** and **without** stratification

Marks Breakdown

Task	Marks
Train function	4
Test function	3
Split dataset	1
Report performance	2
Bonus	2
Total Marks	10 + 2

Assignment Rules

- Assignment must be submitted in eLMS. Submission via email won't be accepted. **Submit your code as a pdf file. To do this press *ctrl* + *p* in colab and save the pdf. Any other file type will not be accepted.**

- ***Rename*** your file to your ***student_id***.

- Deadline for the assignment is ***16/11/2021*** at ***01:30 PM***.

- **DO NOT COPY ANY CODE**. Penalty for ***plagiarism*** is **-100%**. Also, a powerful **plagiarism checker** is now included in eLMS. So, your submitted assignment will be automatically checked for plagiarism against your classmates and against the internet by eLMS.

- No request for extending the assignment deadline will be entertained.