Megan Tan, A20527707

**CS 585 Spring 2024 Programming Assignment #02**

Due: **Sunday, March 24, 2023 at 11:59 PM CST**

Points: **100**

**Instructions:**

1. Place **all your deliverables (as described below) into a single ZIP** file named:

LastName\_FirstName\_CS585\_Programming02.zip

1. Submit it to Blackboard Assignments section before the due date. **No late submissions will be accepted**.

**Objectives:**

1. (100 points) Implement and evaluate a Naïve Bayes classifier algorithm.

**Task:**

Your task is to implement, train, and test a Naïve Bayes classifier using a publicly available data set. **You can work in groups of two or by yourself. Two individual students / groups can use the same data set**.

**Data set:**

Pick a publicly available data (**follow the guidelines provided in Blackboard**) set first and do an initial exploratory data analysis.

**Deliverables:**

Your submission (if you are working as a group of two, both partners should submit the same work) should include:

* Python code file(s). Your py file should be named:

CS585\_P02\_AXXXXXXXX.py

where AXXXXXXXX is your IIT A number (this is REQUIRED!). If your solution uses multiple files, makes sure that the main (the one that will be run to solve the problem) is named that way and others include your IIT A number in their names as well.

* Presentation slides in PPTX or PDF format. Name it:

LastName\_FirstName\_CS585\_P02\_Slides.pptx or pdf

* This document with your observations and conclusions. You should rename it to:

LastName\_FirstName\_CS585\_P02.pdf

**Implementation:**

Your task is to implement (**from scratch – you can’t use out-of-the-box Python package classifier**), train, and test a Naïve Bayes classifier (as outlined in class) and apply it to classify sentences entered using keyboard.

Your program should:

* Accept one (1) command line argument, i.e. so your code could be executed with

python CS585\_P02\_AXXXXXXXX.py TRAIN\_SIZE

where:

* + CS585\_P02\_AXXXXXXXX.py is your python code file name,
  + TRAIN\_SIZE is a number between 20 and 80 defining the size (in percentages) of the training set. For example: 60 would mean **FIRST** (as ordered in the dataset file) 60% of samples. **Note that your test set is always going to be the LAST (as ordered in the dataset file) 20% of samples.**

Example:

python CS585\_P01\_A11111111.py YES

If the number of arguments provided is NOT one (none, two or more) or the TRAIN\_SIZE argument is out of the specified range, assume that the value for TRAIN\_SIZE is 80.

* Load and process input data set:
  + Apply any data clean-up / wrangling you consider necessary first (mention and discuss your choices in the Conclusions section below).
  + Text pre-processing:
    - treat every document in the data set as a single sentence, even if it is made of many (no segmentation needed),
* Train your classifier on your data set:
  + assume that vocabulary V is the set of ALL words in the data set,
  + divide your data set into:
    - training set: FIRST (as they appear in the data set) TRAIN\_SIZE % of samples / documents,
    - test set: LAST 20 % of samples / documents,
  + use **binary** BAG OF WORDS with **“add-1” smoothing** representation for documents,
  + train your classifier (find its parameters. HINT: use Python dictionary to store them),
* Test your classifier:
  + use the test set to test your classifier,
  + calculate (and display on screen) following metrics:
    - number of true positives,
    - number of true negatives,
    - number of false positives,
    - number of false negatives,
    - sensitivity (recall),
    - specificity,
    - precision,
    - negative predictive value,
    - accuracy,
    - F-score,
* Ask the user for keyboard input (a single sentence S):
  + use your Naïve Bayes classifier to decide (HINT: use log-space calculations to avoid underflow – but bring it back to linear space after!) which class S belongs to,
  + display classifier decision along with P(CLASS\_A |S) and P(CLASS\_B | S) values on screen

Your program output should look like this (if pre-processing step is NOT ignored, output NONE):

Last Name, First Name, AXXXXXXXX solution:

Training set size: 80 %

Training classifier…

Testing classifier…

Test results / metrics:

Number of true positives: xxxx

Number of true negatives: xxxx

Number of false positives: xxxx

Number of false negatives: xxxx

Sensitivity (recall): xxxx

Specificity: xxxx

Precision: xxxx

Negative predictive value: xxxx

Accuracy: xxxx

F-score: xxxx

Enter your sentence:

Sentence S:

<entered sentence here>

was classified as <CLASS\_LABEL here>.

P(<CLASS\_A> | S) = xxxx

P(<CLASS\_B> | S) = xxxx

Do you want to enter another sentence [Y/N]?

If user responds Y, classify new sentence (you should not be re-training your classifier).

where:

* 80 would be replaced by the value specified by TRAIN\_SIZE,
* xxxx is an actual numerical result,
* <entered sentence here> is actual sentence entered y the user,
* <CLASS\_LABEL here> is the class label decided by your classifier,
* <CLASS\_A>, <CLASS\_B> are available labels (SPAM/HAM, POSITIVE/NEGATIVE, etc.).

**Classifier testing results:**

Enter your classifier performance metrics below:

|  |  |
| --- | --- |
| With TRAIN\_SIZE set to 80: | With TRAIN\_SIZE set to 20 (not 80) |
| Number of true positives: 1788  Number of true negatives: 5421  Number of false positives: 791 Number of false negatives: 529  Sensitivity (recall): 0.77169  Specificity: 0.87267  Precision: 0.69329  Negative predictive value: 0.91109  Accuracy: 0.84523  F-score: 0.73039 | Number of true positives: 355  Number of true negatives: 5907  Number of false positives: 2224  Number of false negatives: 43  Sensitivity (recall): 0.89196  Specificity: 0.72648  Precision: 0.13765  Negative predictive value: 0.99277  Accuracy: 0.7342  F-score: 0.23849 |

What are your observations and conclusions? When did the algorithm perform better? a summary below

|  |
| --- |
| **Summary / observations / conclusions** |
| When train size = 80    When TRAIN\_SIZE = 20    A larger training set leads to much better results as per the reported metrics. When inputting user sentences, the model with the larger training set also more accurately classifies the user sentences (shown above).  However, the model is still highly inaccurate as seen in the example below.    This could be due to a much larger number of positive sentiments in the data with nearly 80% of the review being positive. Downsampling the training set can help make the model more accurate. K fold validation can also make the model more accurate. |