INFO 371 AUTUMN 2023
PROBLEM SET 3
Due 11/19/2023 11:59pm

Please turn in your jupyter notebook, its html file, and answers to the following questions in a pdf file. Make sure to name your files LASTNAME-FIRSTNAME-PS3.[filetype].

1. [4 POINTS] Bayes' Theorem shows us how to turn P(E|H) to P(H|E), with E=Evidence and H=Hypothesis. But what does that really mean? Imagine you have to explain this to someone who doesn't understand machine learning or probability at all.

INSTRUCTIONS

a. (2 pts) Explain how to turn P(E|H) to P(H|E), with E=Evidence and H=Hypothesis in layman's terms.

b. (2 pts) Use an example from real life to ground the explanation.

c. The answer should be no more than two paragraphs. (-1 pt if longer than two paragraphs.)

2. [6 POINTS] **Download a YouTube spam collection dataset available at this link.**

This is a public set of comments collected for spam research. It has five datasets composed of 1,956 real messages extracted from five videos. These five videos are popular pop songs that were among the 10 most viewed in the collection period.

All five datasets have the following attributes:

Attribute	Attribute Explained
COMMENT_ID	Unique ID representing the comment
AUTHOR	Author ID
DATE	Date the comment is posted
CONTENT	The comment
TAG	Attribute Explained

INSTRUCTIONS

a. (2 pts) For this exercise use any four of these five datasets to build a spam filter with the Naïve Bayes approach.

b. (2 pts) Use that filter to check the accuracy on the remaining dataset.

c. (2 pts) Make sure to report the details of the training process and the model.

3. **[5 POINTS]** In this exercise, you will use the Portuguese sea battles data that contains outcomes of naval battles between Portuguese and Dutch/British ships between 1583 and 1663.

The dataset has the following features:

Features	Features Explained
Battle	Name of the battle place
Year	Year of the battle
Portuguese ships	Number of Portuguese ships
Dutch ships	Number of Dutch ships
English ships	Number of ships from English side
The ratio of Portuguese to Dutch/British ships	
Spanish Involvement	1=Yes, 0=No
Portuguese outcome	-1=Defeat, 0=Draw, 1=Victory

a. (2 pts) Use an SVM-based model to predict the Portuguese outcome of the battle from the number of ships involved on all sides and Spanish involvement.

b. (2 pts) Try solving the same problem using two other classifiers that you know.

c. (1 pt) Report and compare their results with those from SVM.