CS1571
Fall 2019
11/20 In-Class Worksheet

Name: Simran Gidwani
Where were you sitting in class today: Back Right

A. Pre-Reflection

On a scale of 1-5, with 5 being most confident, how well do you think you could execute these learning objectives:

- 21.1 Define machine learning _____
- 21.2 Explain how a Naïve Bayes classifier works _____
- 21.3 Execute a Naïve Bayes classification _____

B. Naïve Bayes Classification

Your goal is to classify text into whether it is talking about sports or not. For the sake of this example, the following is your training data, consisting of sentences and their categories.

Text	Tag
"A great game"	Sports
"The election was over"	Not sports
"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Not sports

You want to determine whether "A very close game" is a sports sentence or non-sports sentence.

1. Your goal is to determine both P(Sports | A very close game) and P(~Sports | A very close game). Use Bayes Rule to write an expression for P(Sports | A very close game) and P(~ Sports | A very close game).

P(Sport|A very close game) = P(a very close game|sports) * P(sports)/P(a very close game)

P(~Sport|A very close game) = P(A very close game| ~Sport) * P(~Sports)/P(A very close game)

2. You can compare these two probabilities derived in #1, and assume that "A very close game" belongs to the category that has the larger probability. What quantities do you need to know based on the probabilities and your answer to question #1?

How many times a very close game occurs in talk regarding sports and not sports
Probability of sport
Probability of not sports

3. Given the training data, what are the probabilities of P(Sports) and P(Not Sports)?

   P(Sports) = 3/5
   P(Not Sports) = 2/5

4. Given the training data, how do you calculate P(A very close game|Sports)? *Hint: You need to apply the Naïve Bayes assumption here.*

5. Fill out the following table, based on the training data. *Note, if you are counting frequencies of words, you need to add 1 to every count, so that no probability is 0, and add the number of total possible words to each divisor. This is called Laplace smoothing.*

Word	**P(word | Sports)**	**P(word | Not Sports)**
a		
very		
close		
game		

6. Now, compute the two probabilities to compare based on the strategy in #2. What category does "A very close game" belong to.

7. Explain how you would apply the same process to determine whether an email is spam or not. Once you've composed your explanation, feel free to check out this link: https://pythonmachinelearning.pro/text-classification-tutorial-with-naive-bayes/, which also walks you through the process with code.

C. Post-Reflection

On a scale of 1-5, with 5 being most confident, how well do you think you could execute these learning objectives:

- 21.1 Define machine learning _____
- 21.2 Explain how a Naïve Bayes classifier works _____
- 21.3 Execute a Naïve Bayes classification _____