FALL 2022-2023

CS464 HW-1

CS464-2

Necati Furkan Çolak

21803512

Section: 2

Preparation for Question-1:

Before start, I want to describe the probabilities given in the question:

P(H) = 0.64 : The probability that students get high grade.
P(L) = 0.24 : The probability that students get low grade.
P(F) = 0.12 : The probability that students failed.

$\boldsymbol{P}(\boldsymbol{S_M} \mid \text{H}) = 0.87$ (The probability that 87% of the high grades taken by motivated students.)
$\boldsymbol{P}(\boldsymbol{S_M} \mid \text{L}) = 0.21$: (The probability that 21% of the low grades taken by motivated students.)
$\boldsymbol{P}(\boldsymbol{S_M} \mid \text{F}) = 0.04$: (The probability that 0.04% of the grades taken by motivated students.)

As an example, if %87 of the high grades was taken by motivated students then remained high grades would be taken by unmotivated students. So, with the formulas below, we can calculate the same probabilities for the unmotivated students.

$$1 - \boldsymbol{P}(\boldsymbol{S_M} \mid \text{H}) = \boldsymbol{P}(\boldsymbol{S_U} \mid \text{H})$$
$$1 - \boldsymbol{P}(\boldsymbol{S_M} \mid \text{L}) = \boldsymbol{P}(\boldsymbol{S_U} \mid \text{L})$$
$$1 - \boldsymbol{P}(\boldsymbol{S_M} \mid \text{F}) = \boldsymbol{P}(\boldsymbol{S_U} \mid \text{F})$$

$\boldsymbol{P}(\boldsymbol{S_U} \mid \text{H}) = 0.13$: (The probability that 13% of the high grades taken by unmotivated students.)
$\boldsymbol{P}(\boldsymbol{S_U} \mid \text{L}) = 0.79$: (The probability that 79% of the low grades taken by unmotivated students.)
$\boldsymbol{P}(\boldsymbol{S_U} \mid \text{F}) = 0.96$: (The probability that 96% of the high grades taken by unmotivated students.)

Question 1.a:

$$\boldsymbol{P}(\boldsymbol{S_M}) = \boldsymbol{P}(\boldsymbol{S_M} \cap \boldsymbol{H}) + \boldsymbol{P}(\boldsymbol{S_M} \cap \boldsymbol{L}) + \boldsymbol{P}(\boldsymbol{S_M} \cap \boldsymbol{F}) \text{ : Total probability Law.}$$

$$\boldsymbol{P}(\boldsymbol{S_M}) = \boldsymbol{P}(\boldsymbol{S_M} \mid \text{H}) * \text{P(H)} + \boldsymbol{P}(\boldsymbol{S_M} \mid \text{L}) * \text{P(L)} + \boldsymbol{P}(\boldsymbol{S_M} \mid \text{F}) * \text{P(F)}$$

$$\boldsymbol{P}(\boldsymbol{S_M}) = \mathbf{0.87} * \mathbf{0.64} + \mathbf{0.21} * \mathbf{0.24} + \mathbf{0.04} * \mathbf{0.12} = \mathbf{61.2}\,\%$$

Therefore, we can say that 61.2% of the students are motivated for the CS464 lecture.

Question 1.b:

We can calculate the $P(H \mid S_M)$ with the bayes formula. We already know the relevant parts of the question, P (S_M) was calculated in Q1.1 already. After bayes transformation, we can calculate the probability in the description below.

$$\boldsymbol{P(H \mid S_M)} = \frac{\boldsymbol{P(H \cap S_M)}}{\boldsymbol{P(S_M)}} = \frac{\boldsymbol{0.87 * 0.64}}{\boldsymbol{0.612}} = \boldsymbol{0.9098}$$

Question 1.c:

In this question, $\boldsymbol{P(S_U)}$ should be calculated to perform operation below. In the question, a student can be motivated or unmotivated only. In this case,

$$\boldsymbol{P(S_U)} + \boldsymbol{P(S_M)} = 1$$
$$\boldsymbol{P(S_U)} = 1 - \boldsymbol{P(S_M)}$$

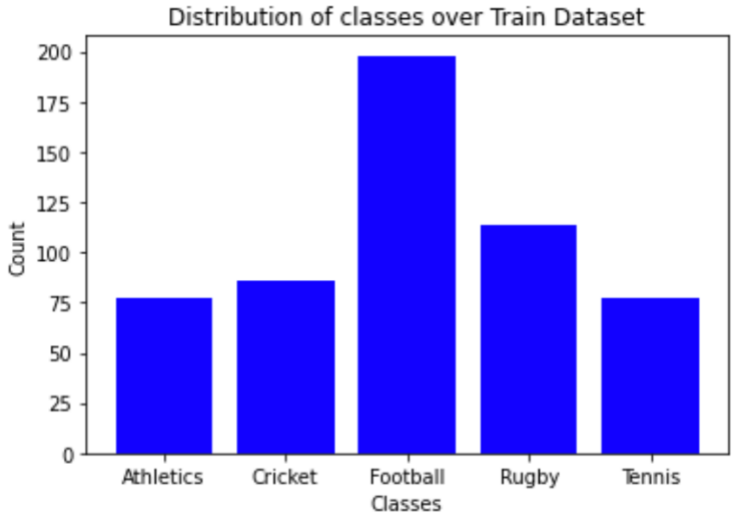
We already calculated $\boldsymbol{P(S_M)}$ in the question 1.1 which is 0.612. Then, $\boldsymbol{P(S_U)}$ = 0.388.

$$\boldsymbol{P(H \mid S_U)} = \frac{\boldsymbol{P(H \cap S_U)}}{\boldsymbol{P(S_U)}} = \frac{\boldsymbol{0.13 * 0.64}}{\boldsymbol{0.388}} = \boldsymbol{0.2144}$$

Question 2.1.1:

In my submission, the following instances were calculated with relevant code. In plot 1, you can find the distribution of classes over the train dataset. In the plot below, we can say that classes are in non-uniform distribution.

PLOT-1: (For train set)



Figures for classes in train dataset:

- **Athletics (class 0):** There are 77 documents that are labeled under athletics.
  In this case, class 0 can be found in 0.1395 probability in the given dataset.
- **Cricket (class 1):** There are 86 documents that are labeled under cricket.
  In this case, class 1 can be found in 0.1558 probability in the given dataset.
- **Football (class 2):** There are 198 documents that are labeled under football. In this case, class 2 can be found in 0.3587 probability in the given dataset.
- **Rugby (class 3):** There are 33 documents that are labeled under rugby.
  In this case, class 3 can be found in 0.0598 probability in the given dataset.
- **Tennis (class 4):** There are 23 documents that are labeled under tennis.
  In this case, class 4 can be found in 0.0416 probability in the given dataset.

Question 2.1.2:

As observed in the graph, the training set is skewed toward class 2. In this case, the ML model would be influenced negatively.

Imbalanced datasets can affect the performance of the model since the classes that are dominant over the dataset, would certainly affect the training process and makes it harder to estimate other classes. In this case, the model tends to predict the class that is over the dataset. Therefore, an ML model that has skewed data, builds a bias toward the dominant class.

To make the process clearer, an example can be useful. Let us think of two classes, class 1, and class 2, with the sizes 400 and 1600 respectively. In this case, prior probabilities can be seen both class 1 and class 2.

$$P(class\ 1) = 0.2\ and\ P(class\ 2) = 0.8$$

P(class 1 | class 3) ∝ P(class 1)* P(class 3 |class 1) and P(class 2 |class 3) ∝ P(class 2)* P(class 3| class 2).

Note that P(class 1 |class 3) > P(class 2 |class 3) since P(class 1) > P(class 2).

We want to estimate the class of class 3.

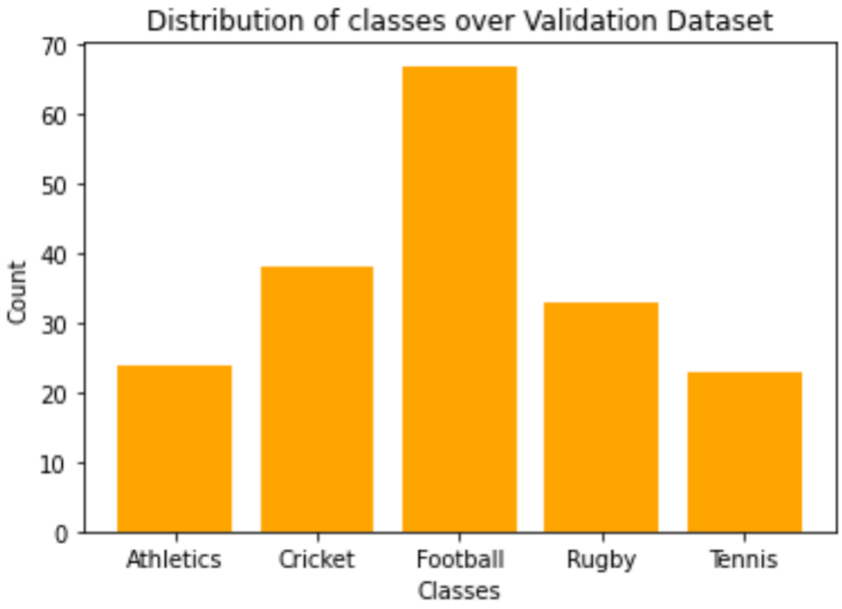
P(class 3 |class 1) ∝ P(class 3)* P(class 1|class 3) and P(class 3 | class 2) ∝ P(class 3)* P(class 2| class 3)

In this example, Naïve Bayes can be used as a classification algorithm. By Naïve Bayes, we compute the probabilities and try to decide which class "class_3" belongs to. In the operations above, we observed that P (class 1 | class 3) > P (class 2 |class 3). So, class_ 3 tends to be class_1 according to the Naïve Bayes algorithm.

To solve the problem of skewness, log transformation can be used. In this case, the distribution will be close to normal distribution.
Normalization is also another way to solve the problem of skewness.

Question 2.1.3:

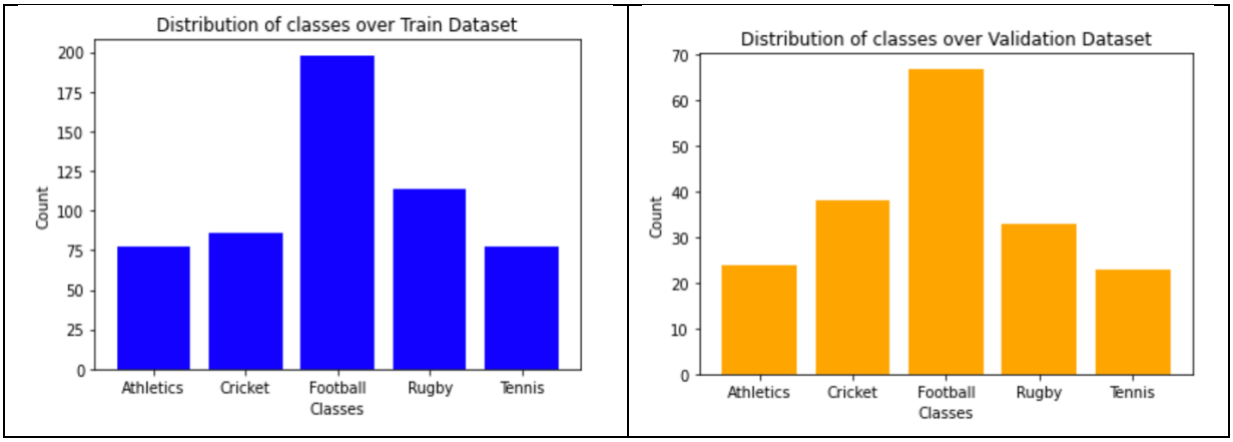
PLOT-2: (For validation set)



Figures for classes in validation dataset:

- **Athletics (class 0):** There are 24 documents that are labeled under athletics. In this case, class 0 can be found in 0.1297 probability in the validation dataset.
- **Cricket (class 1):** There are 38 documents that are labeled under cricket. In this case, class 1 can be found in 0.2054 probability in the validation dataset.
- **Football (class 2):** There are 67 documents that are labeled under football. In this case, class 2 can be found in 0.3621 probability in the validation dataset.
- **Rugby (class 3):** There are 33 documents that are labeled under rugby. In this case, class 3 can be found in 0.1784 probability in the validation dataset.
- **Tennis (class 4):** There are 23 documents that are labeled under tennis. In this case, class 4 can be found in 0.1243 probability in the validation dataset.

Then, we should compare training and validation set as question asks. Let we check the bar charts that are belong to training and validation set respectively.



If we look at the graphs, we can observe the similar type of distributions for every class. As in the training set, the dominant element of the validation set is also class-2 which is football class. If training and validation sets have different characteristics, accuracy can be affected negatively since the amount of FP and FN would increase.

Question 2.1.4:

Unbalanced datasets can make models more unreliable because if the dataset is skewed, then the model tends to predict the dominant element of the dataset. In this situation, accuracy can be affected negatively. Let us check accuracy formula, $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$. If model predicts model wrongly then the number of FP and FN would increase. This situation has led to decrease accuracy in skewed datasets.

Question 2.2:

Confusion Matrix	Athletics	Cricket	Football	Rugby	Tennis
Athletics	24	34	54	31	21
Cricket	0	4	0	0	0
Football	0	0	13	0	0
Rugby	0	0	0	2	0
Tennis	0	0	0	0	2

Accuracy of MLE Estimation: 24.3243 %

In the confusion matrix above, it's obvious that athletics is the most predictab le class for MLE estimation. If we look at the values of our probability calculations, th en we can see that they're close to minus infinity with the np. inf statement. Although each probability is different, since we used np.inf, values are assigned to the same val ue which is minus infinity. Then, the argmax function that I used in my prediction var iable, gives the index corresponding to the first occurrence of max value which is clas s 0 or Athletics in my dataset. Because of this situation, Athletics is the most predicted class in my multinomial naïve Bayes implementation.

Question 2.3:

Confusion Matrix	Athletics	Cricket	Football	Rugby	Tennis
Athletics	24	0	0	0	1
Cricket	0	35	0	0	0
Football	0	2	66	0	0
Rugby	0	1	1	33	0
Tennis	0	0	0	0	22

Accuracy of MAP Estimation with drichlet prior : 97.2973%.

Question 3.4:

If we use Dirichlet prior for accuracy, the performance of the model has increased significantly. In the model that we used in the MLE estimator, values approached minus infinity. Therefore, multinomial naïve Bayes is incapable to detect which y_hat has larger than the other. On the other hand, Dirichlet prior has solved this problem by smoothing the dataset with alpha = 1. So, the low probability values which are close to minus infinity were eliminated from the dataset. Thus, the algorithm works properly and detects which y_hat is larger than the other