2022-2023 Spring Semester

Homework 1

Course: CS464

Section: 1

Name: Alper

Surname: Mumcular

Student ID: 21902740

Date: 22 March 2023

1) The Online Shopping Case

Question 1.1)

We are given that:

$P(F_P|P) = 0.95$ $P(F_P|M) = 0.6$ $P(F_P|U) = 0.1$

$P(P) = 0.45$ $P(M) = 0.3$ $P(U) = 0.25$

For the answer, we need to calculate $P(F_P)$.

$$P(F_P) = \sum_i P(F_P|X_i) * P(X_i)$$

Law of Probability

$P(F_P) = 0.95 * 0.45 + 0.6 * 0.3 + 0.1 * 0.25$

$P(F_P) =$ **0.6325**

Question 1.2)

We need to calculate $P(P|F_P)$

$$P(P|F_P) = \frac{P(F_P|P) * P(P)}{\sum_i P(F_P|X_i) * P(X_i)}$$

$$P(P|F_P) = \frac{0.95 * 0.45}{0.6325} = \frac{0.4275}{0.6325} \approx \mathbf{0.6759}$$

Question 1.3)

$P(F_N|P) = 1 - P(F_P|P) = 1 - 0.95 = 0.05$

$P(F_N|M) = 1 - P(F_P|M) = 1 - 0.6 = 0.4$

$P(F_N|U) = 1 - P(F_P|U) = 1 - 0.1 = 0.9$

$$P(P|F_N) = \frac{P(F_N|P) * P(P)}{\sum_i P(F_N|X_i) * P(X_i)}$$

$$P(P|F_P) = \frac{0.05 * 0.45}{0.3675} = \frac{0.0225}{0.3675} \approx \mathbf{0.0612}$$

2) Spam Mail Detection

Libraries that are used:

NumPy: This library was used because some parts needed logarithmic and max functions.

Pandas: This library is used to open .csv files and concatenate training sets. Also, the data is stored in a Pandas dataframe structure.

Question 2.1)

1- 28.6% ($\frac{1183}{4137}$)

2- The training set is skewed towards the non-spam (negative) class. Yes, I think having an imbalanced training set affects our Multinomial Naive-Bayes model because this difference (in this homework) creates a bias towards non-spam emails. Because the prior probability will be much higher for non-spam emails.

3- If my dataset is skewed towards one of the classes, then the model will classify most of the test emails as the majority class. This can decrease our accuracy if the test dataset is balanced because our training dataset is imbalanced. On the other hand, if the test dataset consists of a majority class with a high percentage, then the accuracy will be very high because our model always classifies as a majority class. In this case, the reported accuracy will be misleading for our model.

Question 2.2)

After training our Multinomial Naive Bayes model, the following results were obtained:

Accuracy: 0.9584541062801932

(TP = True Positives - FP = False Positives - TN = True Negatives - TP = True Positives)

TP: 289 **FP**: 15
FN: 28 **TN**: 703
Precision: 0.9506578947368421
Recall: 0.9116719242902208
Specificity: 0.9791086350974930
F-Measure: 0.9307568438003221

Wrong Prediction Count: 43

Question 2.3)

After training our Multinomial Naive Bayes model using a fair Dirichlet prior, the following results were obtained:

Accuracy: 0.9478260869565217

TP: 300 **FP**: 37
FN: 17 **TN**: 681
Precision: 0.8902077151335311
Recall: 0.9463722397476341
Specificity: 0.9484679665738162
F-Measure: 0.9174311926605505
Wrong Prediction Count: 54

The count of true positives is increased and false negatives is decreased. On the other hand, the count of false positives is increased and true negatives is decreased. There is a trade-off between them. Also, the accuracy rate has decreased.

The Effect of the Dirichlet prior α:

In order to observe the effect of the Dirichlet prior, the value of α has been changed many times. Tested α values and their accuracy rates are below:

For α = 1 -> Accuracy: 0.9507246376811594
For α = 2 -> Accuracy: 0.9487922705314009
For α = 4 -> Accuracy: 0.9478260869565217
For α = 10 -> Accuracy: 0.9458937198067633
For α = 50 -> Accuracy: 0.9391304347826087
For α = 100 -> Accuracy: 0.8753623188405797

As you can see from the sample tests, for this dataset, when we increase the value of α, the accuracy rate decreases. The reason for that can be while we are increasing the value of α, the count of false negatives increases. This causes a decrease in the accuracy rate.

Question 2.4)

After training our Bernoulli Naive Bayes model, the following results were obtained:

Accuracy: 0.9082125603864735

TP: 231 **FP**: 9
FN: 86 **TN**: 709
Precision: 0.9625
Recall: 0.7287066246056783
Specificity: 0.9874651810584958
F-Measure: 0.8294434470377020
Wrong Prediction Count: 95

Comparing these results with the Multinomial model, we can observe that the accuracy rate is noticeably less than the accuracy rate obtained in previous models. The count of false positives in this model is the lowest one among these three models. However, the count of false negatives is the highest among these three models.

Bernoulli Naive Bayes Model differ from Multinomial Naive Bayes Model in these ways:

1- In Bernoulli Naive Bayes Model, we are using binary data. In this dataset, for example, if the word j is absent we represent it as 0. If the email contains the word j, we represent it as 1. On the other hand, in Multinomial Naive Bayes, the number of occurrences of the word j in email including the multiple occurrences is important. For example, for the word j, we can represent it as 0,1,2,3...

2- Their probability distributions are different. Bernoulli Naive Bayes Model assumes each feature has a Bernoulli distribution. On the other hand, Multinomial Naive Bayes assumes each feature has a discrete probability distribution.

3- Their estimators' formulas are different. The formula for predicting label is also different from each other because, in Bernoulli Naive Bayes Model, multiple occurrences of a word in an email is not important.

Question 2.5)

Multinomial Naive Bayes Model (MNB)

Multinomial Naive Bayes Model with Dirichlet prior (MNBD)

Bernoulli Naive Bayes Model (BNB)

Comparison of Accuracy:

MNB > MNBD > BNB

Comparison of Precision:

BNB > MNB > MNBD

Comparison of Recall:

MNBD > MNB > BNB

Comparison of Specificity:

BNB > MNB > MNBD

Comparison of F-Measure:

MNB > MNBD > BNB

The best model can change according to the type of our dataset. For this example, the best model is Multinomial Naive Bayes Model (without Dirichlet prior) because this model has a higher accuracy rate than the others. Accuracy can be deceiving as a performance metric because our dataset can be considered imbalanced. As a result of an imbalanced dataset, the accuracy rate can be misleading for us.