CS464 Introduction to Machine Learning

Homework 1 Report

İrem Ecem Yelkanat
21702624
Section 2

1 The White Library

Question 1.1 Define the sample space of Donald's experiment.

Ω = {1-word title novel, 2-word title novel, 1-word title poetry, 2-word title poetry, 3-word title poetry, 1-word title story, 2-word title story, 3-word title story}

Question 1.2 Based on your sample space, write the elements of the event
A = {"a novel with a single-word title or a poem book with a 2- or 3-word title but not a story book"}.

A = {1-word title novel, 2-word title poetry, 3-word title poetry}

Question 1.3 Write down the axioms of probability.

Let Ω be the sample space, A be an event in Ω, and P (A) be the probability of event A.

- For any event $A \in \Omega$, $0 \leq P(A) \leq 1$, that is, all probabilities are real values between 0 and 1
- $P(\Omega) = 1$, that is, probability of the sample space is 1
- For any sequence of mutually exclusive events $\{A_1, A_2, A_3, \ldots, A_n\}$
  $P(\cup_i A_i) = \sum_i P(A_i)$

Question 1.4 After collecting some statistics, Donald estimates that, in a random book selection experiment,
P ({a 3-word title story book or a 2-word title novel}) = 0.045
P ({a 3-word tile story book or a 2-word title novel or 2-word title poetry book}) = 0.11
P ({a 2-word title poetry book or a 3-word title story book}) = 0.06
Would you agree Donald on his estimation? Why?

Let represent the probability values of the sample space in a table

Type / **Title**	Novel	Poetry	Story
1-word	n_1	p_1	s_1
2-word	n_2	p_2	s_2
3-word	$n_3 = 0$	p_3	s_3

Table 1: The probability values of the sample space

Write down the probabilities given in the question using the values in *Table 1*.

$s_3 + n_2 = 0.045$ (1)

$s_3 + n_2 + p_2 = 0.11$ (2)

$p_2 + s_3 = 0.06$ (3)

Solving the equations (1) and (2), the value of p_2 is 0.065. Putting the value of p_2 into the equation (3) and solving it for s_3, s_3 is -0.005. But this is impossible according to the axioms of probability. Since first axiom of probability in *Question 1.3*, the probability of an event in sample space Ω must be a real value between 0 and 1.

Similarly, solving the equations (2) and (3), the value of n2 is 0.05. Putting the value of n2 into the equation (1) and solving it for s_3, s_3 is -0.005. But this is impossible according to the axioms of probability. Since first axiom of probability in *Question 1.3*, the probability of an event in sample space Ω must be a real value between 0 and 1.

Hence, I don't agree Donald on his estimation.

2 Cafe Customers

Question 2.1 If 10 people arrive to the cafe in a break, what is the probability that less than 3 people will buy coffee?

We need to calculate P (Y < 3 | X = 10). Since X and Y are independent,
P (Y < 3 | X = 10) = P (Y < 3)

P (Y < 3) = P (0) + P (1) + P (2)

$$= \binom{10}{0} * \left(\frac{3}{10}\right)^0 * \left(\frac{7}{10}\right)^{10} + \binom{10}{1} * \left(\frac{3}{10}\right)^1 * \left(\frac{7}{10}\right)^9 + \binom{10}{2} * \left(\frac{3}{10}\right)^2 * \left(\frac{7}{10}\right)^8$$

$$= 0.3827827864$$

Question 2.2 What is the probability that a total of 2 people arrive to the cafe in a break and none of them buys coffee?

We need to calculate P (X = 2, Y = 0). Since X and Y are independent,
P (X = 2, Y = 0) = P (X = 2) * P (Y = 0)

$$P(X = 2) * P(Y = 0) = \frac{e^{-20} * 20^2}{2!} * \binom{2}{0} * \left(\frac{3}{10}\right)^0 * \left(\frac{7}{10}\right)^2$$

$$= 2.01993055 * 10^{-7}$$

Question 2.3 Compute the expected value of Y , i.e. E[Y] without computing the probability mass function of Y.

Since a variable X that follows a binomial distribution is sum of discrete random variables Y_i that follow binomial distribution,

$$X = \sum_{i=1}^{n} Y_i$$

$$E(X) = E\left(\sum_{i=1}^{n} Y_i \right)$$

$$E(X) = \sum_{i=1}^{n} E(Y_i)$$

$$E(X) = \sum_{i=1}^{n} p$$

$$E(X) = np$$

3 Spam Email Detection

Question 3.1 If the the ratio of the classes in a dataset is close to each other, it is a called "balanced" class distribution; i.e it is not skewed. What is the percentage of spam e-mails in the y_train.csv?.
Is the training set balanced or skewed towards one of the classes? Do you think having an imbalanced training set affects your model? If yes, please explain how it affects and propose a possible solution if needed.

The percentage of spam emails in the y_train.csv file is approximately %71.26071 as the number of spam emails is 2911 and the total number of emails is 4085.

The training set is not balanced as the percentage of spam emails in the y_train.csv is %71.26071 and the percentage of normal emails in the y_train.csv is %28.73929. Hence the training set is skewed towards spam class.

Having an imbalanced training set may affect my model. The model may be overfitted to the class that the training set is skewed towards to, which is spam class, and not be sensible to the other class, which is normal class. The model may have a good accuracy rate but it may mostly predict spam class when it was supposed to predict normal class. A possible solution to the skewness problem is resampling, which can be divided into two types which are over-sampling and under-sampling. Over-sampling is a technique that involves adding more samples to the minority class, in our case normal class. Under-sampling is a technique that involves removing samples from the skewed class, in our case spam class. However, both can have also negative effects on the model as over-sampling can cause overfitting and under-sampling can cause information loss.

Question 3.2 Train a Multinomial Naive Bayes model on the training set and evaluate your model on the test set given. Find and report the accuracy and the confusion matrix for the test set as well as how many wrong predictions were made.

Total number of predictions: 1086
Number of true predictions: 928
Number of wrong predictions: 158
Accuracy is: 0.85451197053407

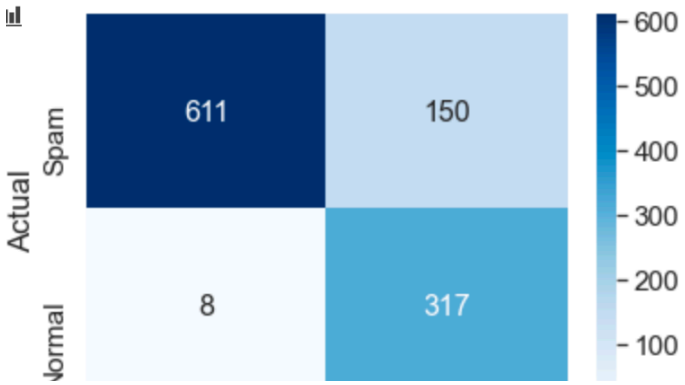

Figure 1: Confusion Matrix for Multinomial Naive Bayes Model without smoothing

Question 3.3 Extend your classifier so that it can compute an MAP estimate of θ parameters using a fair Dirichlet prior. This corresponds to additive smoothing. The prior is fair in the sense that it "hallucinates" that each word appears additionally α times in the train set.
For this question set α = 1. Train your classifier using all of the training set and have it classify all of the test set and report test-set classification accuracy and the confusion matrix. Explicitly discuss your results.

Total number of predictions: 1086
Number of true predictions: 1066
Number of wrong predictions: 20
Accuracy is: 0.9815837937384899

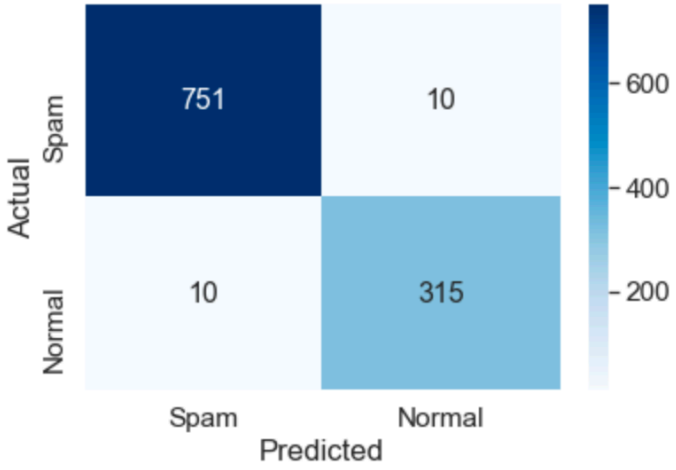

Figure 2: Confusion Matrix for Multinomial Naive Bayes Model with smoothing

Applying additive smoothing to the estimates of θ parameters had significant effect on the accuracy of the model compared to the Multinomial Naive Bayes model that didn't have smoothing in its estimators. The accuracy increased from %85 to %98. That is because, when additive smoothing is applied to θ parameters, estimates that are equal to 0 do not effect the whole probability as estimates that are equal to 0 might cause the prediction to be mislead since they cause the probability of a test sample being belong to a class to be equal to zero. Therefore, in smoothing, α times appearance added to each word and it eliminates the situation of estimators to be equal to zero, rather it increases the probabilities to a small number. Therefore, the model makes more accurate predictions.

Question 3.4 Train a Bernoulli Naive Bayes classifier using all of the data in the training set, and report the testing accuracy and the confusion matrix as well as how many wrong predictions were made. What did your classifier end up predicting? Compare your results with the Multinomial Model. Discuss your findings explicitly.

Total number of predictions: 1086
Number of true predictions: 913
Number of wrong predictions: 173
Accuracy is: 0.8406998158379374

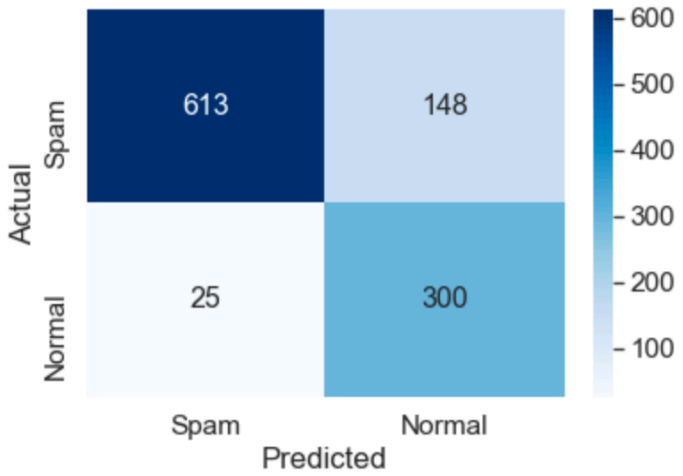

Figure 3: Confusion Matrix for Bernoulli Naive Bayes Model

Bernoulli Naive Bayes model, like the first model that was using Multinomial approach, didn't apply additive smoothing to estimates of θ parameters. Similarly, estimates that are equal to 0 affected the prediction negatively as they caused the overall probability of a test sample to belong to a class to be zero. Therefore Bernoulli Naive Bayes Model performed worse than the Multinomial Naive Bayes Model with smoothing. Additionally, the number of false normal predictions decreased and the number of true spam predictions increased slightly, whereas, the number of true normal predictions decreased and the number of false spam predictions increased compared to the first Multinomial Naive Bayes Model. The Bernoulli Naive Bayes model performed better on spam test data but performed worse on normal test data more, overall, leading to less accuracy compared to Multinomial Naive Bayes Model without smoothing. Also, although the training data is skewed through spam class, the model performed better on normal class, similar to Multinomial Naive Bayes Model without smoothing.