CS464 Introduction to Machine Learning

Section 2

Fall 2024-25

Homework 1

20.10.2024

Furkan Komaç 22102165

1 Probability Review

Question 1.1

T: landing tail 2 times [P(T)*P(T)]
B : selecting blue coin,
R : selecting red coin
Y : selecting yellow coin
X_1: selecting box 1
X_2: selecting box 2

$$P(T) = P(T|X_1)P(X_1) + P(T|X_2)P(X_2)$$

$$= [P(T|B)P(B) + P(T|Y)P(Y)]P(X_1) + [P(T|B)P(B) + P(T|R)P(R)]P(X_2)$$

$$= [(\frac{1}{2})(\frac{1}{2})(\frac{2}{3}) + (\frac{3}{4})(\frac{3}{4})(\frac{1}{3})](\frac{1}{2}) + [(\frac{1}{2})(\frac{1}{2})(\frac{1}{2}) + (\frac{9}{10})(\frac{9}{10})(\frac{1}{2})](\frac{1}{2})$$

$$= 0.44208$$

$$P(T) = 0.44208 = 44.208\%$$

Question 1.2

P(T) is founded 0.44208 in question 1.1.

$$P(B|T) = \frac{P(B \cap T)}{P(T)} = \frac{P(T|B) \times P(B)}{P(T)} = \frac{P(T|B) \times [P(B|X_1) \times P(X_1) + P(B|X_2) \times P(X_2)]}{P(T)}$$

$$= \frac{(\frac{1}{2})(\frac{1}{2}) \times [(\frac{2}{3})(\frac{1}{2}) + (\frac{1}{2})(\frac{1}{2})]}{0.44208} \approx 0.32948$$

$$P(B|H) \approx 0.32948 = 32.948\%$$

Question 1.3

P(T) is founded 0.44208 in question 1.1.

$$P(R|T) = \frac{P(R \cap T)}{P(T)} = \frac{P(T|R) \times P(R)}{P(T)} = \frac{P(T|R) \times [P(R|X_1) \times P(X_1) + P(R|X_2) \times P(X_2)]}{P(T)}$$

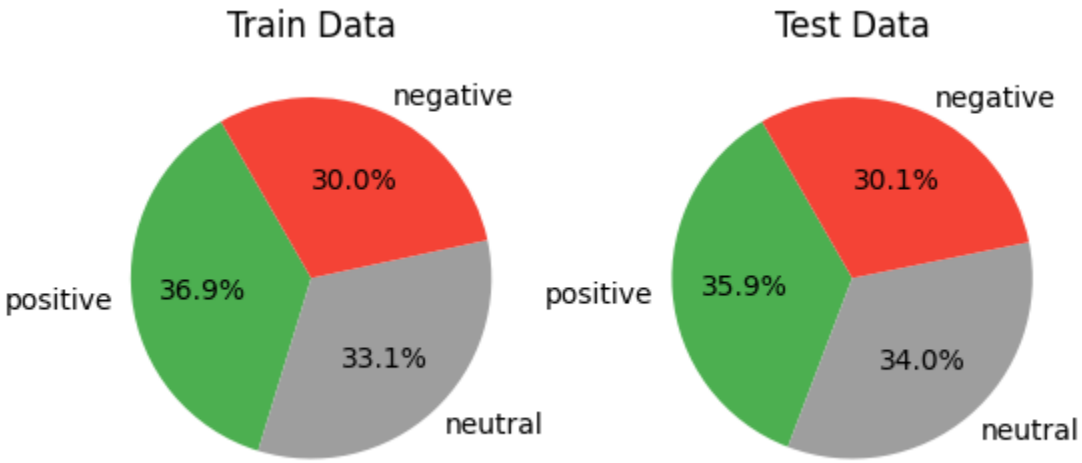
$$= \frac{(\frac{9}{10})(\frac{9}{10}) \times [0 \times (\frac{1}{2}) + (\frac{1}{2})(\frac{1}{2})]}{0.44208} \approx 0.45803$$

$$P(R|T) \approx 0.45803 = 45.803\%$$

2 Amazon Reviews Classification

Question 2.1

1)



2)

Prior probability of each class is $P(Y = y_k) = \frac{Ny_k}{N}$. Hence the prior probability of each label is equal to the proportion in all labels, it is the same with the portion of each label in the pie chart for the train set:

$$P(Y = Positive) = 36.9\%$$

$$P(Y = Neutral) = 33.1\%$$

$$P(Y = Negative) = 30.0\%$$

3)

The training set is a bit skewed towards the positive class, which makes up 36.9% of the data compared to 33.1% neutral and 30% negative. While the difference isn't huge, it could still affect the model because the model might learn to predict positive more often just because there are more positive examples in the training data. An unbalanced data set can lead to biased predictions, where the model favors the majority class, which can reduce performance on less frequent classes.

4)

$$ln(P(good|Y = positive)) = -1.4113403930459782$$

$$ln(P(bad|Y = positive)) = -4.259152536523347$$

Question 2.2

The accuracy of the model is 58.143%. Following table shows the confusion matrix of this model:

Multinomial Naive Bayes				
		Actual Label		
		Negative (0)	Neutral (1)	Positive (2)
Predicted Label	Negative (0)	138	76	32
	Neutral (1)	41	78	28
	Positive (2)	32	84	191

Question 2.3

The accuracy of the model has improved to 64.857% using the Dirichlet prior. This means that the model becomes more generalizable and shows better accuracy for unseen data, assuming that each word appears once more in the training set. Following table shows the confusion matrix of this model:

Multinomial Naive Bayes using a fair Dirichlet prior.				
		Actual Label		
		Negative (0)	Neutral (1)	Positive (2)
Predicted Label	Negative (0)	151	73	13
	Neutral (1)	45	86	21
	Positive (2)	15	79	217

Question 2.4

The accuracy of the model has improved to 65.143% using the Bernoulli Naive Bayes. Unlike Multinomial Naive Bayes, which considers the frequency of each word, Bernoulli Naive Bayes only looks at whether a word is present or absent, making it more effective in handling sparse data. This helped it perform slightly better, especially in distinguishing "Neutral" reviews.

Bernoulli Naive Bayes				
		Actual Label		
		Negative (0)	Neutral (1)	Positive (2)
Predicted Label	Negative (0)	122	33	19
	Neutral (1)	82	174	72
	Positive (2)	7	31	160