Bilkent University

CS 464-2 - Homework 1 Report

Mehmet Akif Şahin - 22203673

October 29, 2024

1 Probability Review

1.1 What is the probability that you get two tails in a row?

Let's define following events:

TT : we got two tails in a row	B : the selected coin is blue
F : the coin is selected from first box	Y : the selected coin is yellow
S : the coin is selected from second box	R : the selected coin is red

assuming,

$$P(F) = P(S) = \frac{1}{2}$$

then:

$$P(B) = P(B|F) \cdot P(F) + P(B|S) \cdot P(S) = \frac{2}{3} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{7}{12}$$

$$P(Y) = P(Y|F) \cdot P(F) + P(Y|S) \cdot P(S) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$$

$$P(R) = P(R|F) \cdot P(F) + P(R|S) \cdot P(S) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

It's obvious that the events $B,\ Y,\ R$ are disjoint and $P(B \cup Y \cup R) = 1$. Then by the total probability theorem:

$$P(TT) = P(TT|B) \cdot P(B) + P(TT|Y) \cdot P(Y) + P(TT|R) \cdot P(R)$$

$$P(TT) = (\frac{1}{2})^2 \cdot \frac{7}{12} + (\frac{3}{4})^2 \cdot \frac{1}{6} + (\frac{9}{10})^2 \cdot \frac{1}{4} = 0.44208$$

1.2 You toss the coin two times and got two tails in a row. What is the probability that the selected coin was fair?

Let's define the same events and assume same probabilities in previous question, using bayes's rule:

$$P(B|TT) = \frac{P(TT|B) \cdot P(B)}{P(TT)} = \frac{(\frac{1}{2})^2 \cdot \frac{7}{12}}{0.44208} = 0.32988$$

1.3 You toss the coin two times and got two tails in a row. What is the probability that the selected coin was the red one?

Let's define the same events and assume same probabilities in first question, using bayes's rule:

$$P(R|TT) = \frac{P(TT|R) \cdot P(R)}{P(TT)} = \frac{(\frac{9}{10})^2 \cdot \frac{1}{4}}{0.44208} = 0.45806$$

3.1

3.1.1 What are the percentages of each category in the y_train.csv y_test.csv? Draw a pie chart showing percentages.

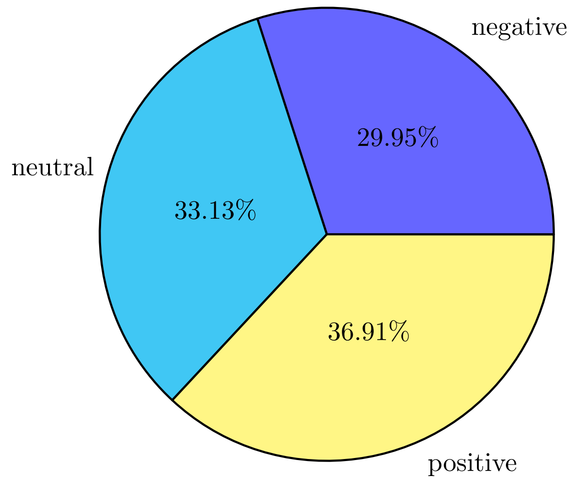

Figure 1: y_train category percentages

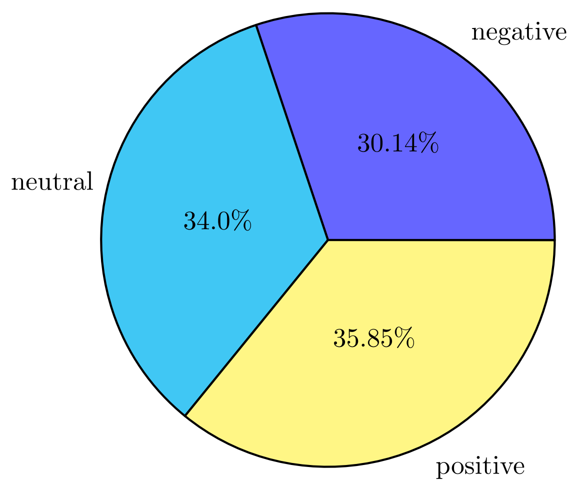

Figure 2: y_test category percentages

3.1.2 What is the prior probability of each class?

Classes are negative(0), neutral(1), positive(2).

$$P(Y = 0) = 29.95\%$$

$$P(Y = 1) = 33.13\%$$

$$P(Y = 2) = 36.91\%$$

3.1.3 Is the training set balanced or skewed towards one of the classes? Do you think having an imbalanced training set affects your model?

The training set is fairly balanced, with approximately 30% negative, 33% neutral, and 37% positive reviews, showing no significant skew toward any class.

3.1.4 How many times do the words "good" and "bad" appear in the training documents with the label "positive", including multiple occurrences, and what is the log ratio of their occurrences within those documents, i.e, $ln(P(good|Y = positive))$ and $ln(P(bad|Y = positive))$?

good	bad
207	12

Figure 3: Number of occurences of words 'good' and 'bad'

$$ratio = \frac{ln(P(good|Y = positive)}{ln(P(good|Y = positive)} = \frac{-4.287609}{-7.135421} = 0.600890$$

3.2 Train a Multinomial Naive Bayes Model

Accuracy: 0.583

Confusion Matrix:

		True Labels		
		0	1	2
Predicted Labels	0	138	41	32
	1	76	78	84
	2	31	28	192

Figure 4: Confusion Matrix for Multinomial Naive Bayes Model

3.3 Train a Multinomial Naive Bayes Model with Drichlet Prior

Accuracy: 0.649

Confusion Matrix:

		True Labels		
		0	1	2
Predicted Labels	0	151	45	15
	1	73	86	79
	2	13	21	217

Figure 5: Confusion Matrix for Multinomial Naive Bayes Model with Drichlet Prior $\alpha = 1$

The model's accuracy increased by 0.066 with the introduction of additive smoothing using the Dirichlet prior. However, the confusion matrix shows that some previously correct classifications were misclassified. Overall, though, smoothing enhanced the model's performance by assigning small probabilities to words that were unseen in certain classes, thus preventing zero-frequency words from having a probability of zero.

3.4 Train a Bernoulli Naive Bayes Model

Accuracy: 0.641

Confusion Matrix:

		True Labels		
		0	1	2
Predicted Labels	0	113	90	8
	1	29	180	29
	2	19	76	156

Figure 6: Confusion Matrix for Bernoulli Naive Bayes Model with Drichlet Prior $\alpha = 1$

This model considers the presence or absence of each word, rather than the frequency. Thus focuses on whether a term is present rather than how many times it occurs. The Bernoulli Naive Bayes model may be less sensitive to the high-frequency words but can capture the pattern in the documents where the presence of certain words is more indicative rather than their counts. It's expected for Bernoulli Naive Bayes model to perform better on sparse datasets.