2252654 NLP Lab 4 Math Exercise

Phuong Huynh

February 2025

1 Problem 1

The Naive Bayes formula for classification is:

$$P(\text{class}|\text{words}) \propto P(\text{sentence}|\text{class})P(\text{class})$$

Since the prior probabilities are equal, we only need to compute:

$$P(\text{sentence}|\text{pos}) = P(I|\text{pos}) \times P(always|\text{pos}) \times P(like|\text{pos}) \times P(foreign|\text{pos}) \times P(films|\text{pos})$$

$$P(\text{sentence}|\text{neg}) = P(I|\text{neg}) \times P(always|\text{neg}) \times P(like|\text{neg}) \times P(foreign|\text{neg}) \times P(films|\text{neg})$$

Using the provided likelihoods:

$$P(\text{sentence}|\text{pos}) = (0.09)(0.07)(0.29)(0.04)(0.08)$$

$$= 0.0000583$$

$$P(\text{sentence}|\text{neg}) = (0.16)(0.06)(0.06)(0.15)(0.11)$$

$$= 0.0000950$$

Since $P(\text{sentence}|\text{neg}) > P(\text{sentence}|\text{pos})$, the Naïve Bayes classifier assigns the sentence to the **negative** class.

2 Problem 2

$$Vocabulary = \{fun, couple, love, fast, furious, shoot, fly\}$$

$$|V| = 7$$

The prior probability of each class is $P(comedy) = \frac{2}{5}$ and $P(action) = \frac{3}{5}$

Word	Comedy	Action
fun	3	1
couple	2	0
love	2	1
fast	1	2
furious	0	2
shoot	0	4
fly	1	1
Total	9	11

Table 1: Count occurences of word

Laplace smoothing:

$$P(w_i|c) = \frac{count(w_i, c) + 1}{\Sigma(count(w_i, c)) + |V|}$$

Document D consists of: $D = \{fast, couple, shoot, fly\}$

$$\begin{aligned} P(D|\text{comedy}) &= P(\text{fast}|\text{comedy}) \times P(\text{couple}|\text{comedy}) \times P(\text{shoot}|\text{comedy}) \times P(\text{fly}|\text{comedy}) \\ &= \frac{1+1}{9+7} \times \frac{2+1}{9+7} \times \frac{0+1}{9+7} \times \frac{1+1}{9+7} \\ &= \frac{2 \times 3 \times 1 \times 2}{16^4} \approx 1.831 \times 10^{-4} \end{aligned}$$

$$\begin{aligned} P(D|\text{action}) &= P(\text{fast}|\text{action}) \times P(\text{couple}|\text{action}) \times P(\text{shoot}|\text{action}) \times P(\text{fly}|\text{action}) \\ &= \frac{2+1}{11+7} \times \frac{0+1}{11+7} \times \frac{4+1}{11+7} \times \frac{1+1}{11+7} \\ &= \frac{3 \times 1 \times 5 \times 2}{18^4} \approx 2.858 \times 10^{-4} \end{aligned}$$

$$P(comedy|D) \propto P(D|comedy) \times P(comedy) = \frac{2}{5} \times 1.831 \times 10^{-4} \approx 7.324 \times 10^{-5}$$

$$P(action|D) \propto P(D|action) \times P(action) = \frac{3}{5} \times 2.858 \times 10^{-4} \approx 1.715 \times 10^{-4}$$

Since $P(action|D) > P(comedy|D)$, **action** will be the most likely class for D.

3 Problem 3

$$Vocabulary = \{good, poor, great\} => |V| = 3$$

Sentence after ignoring the word that never occurred: $\boldsymbol{sentence = \{good, good, great, poor\}}$

3.1 Multinomial Naive Bayes

Word	class: pos	class: neg
good	3	2
poor	1	10
great	5	2
Total	9	14

Table 2: Table for multinomial Naive Bayes (count occurrences)

3.1.1 Calculate positive:

There are 2 positive documents out of 5 documents: $P(pos) = \frac{2}{5}$

$$P(good|pos) = \frac{3+1}{9+3} = \frac{1}{3}$$
$$P(poor|pos) = \frac{1+1}{9+3} = \frac{1}{6}$$
$$P(great|pos) = \frac{5+1}{9+3} = \frac{1}{2}$$

$$\begin{aligned} P(pos|sentence) &\propto P(sentence|pos) \times P(pos) \\ &= P(good|pos) \times P(good|pos) \times P(poor|pos) \times P(great|pos) \times P(pos) \\ &= \frac{1}{3} \times \frac{1}{3} \times \frac{1}{6} \times \frac{1}{2} \times \frac{2}{5} \\ &= \frac{1}{270} \approx 3.7 \times 10^{-3} \end{aligned}$$

3.1.2 Calculate negative:

There are 3 negative documents out of 5 documents: $P(pos) = \frac{3}{5}$

$$P(good|neg) = \frac{2+1}{14+3} = \frac{3}{17}$$
$$P(poor|neg) = \frac{10+1}{14+3} = \frac{11}{17}$$
$$P(great|neg) = \frac{2+1}{14+3} = \frac{3}{17}$$

$$\begin{aligned} P(neg|sentence) &\propto P(sentence|neg) \times P(neg) \\ &= P(good|neg) \times P(good|neg) \times P(poor|neg) \times P(great|neg) \times P(neg) \\ &= \frac{3}{17} \times \frac{3}{17} \times \frac{11}{17} \times \frac{3}{17} \times \frac{3}{5} \\ &= \frac{1}{270} \approx 2.13 \times 10^{-3} \end{aligned}$$

3.1.3 Multinomial NB conclusion:

With multinomial Naive Bayes, the classification is most likely to be **positive** because $P(pos|sentence) > P(neg|sentence)$

3.2 Binarized Naive Bayes

For binarized naive Bayes, we consider only the presence or absence of words.

$$P(word|class) = \frac{\text{number of documents in class where word is present} + 1}{\text{total number of documents in class} + 2}$$

3.2.1 Calculate positive:

Total documents: 2.

$$P(good|pos) = \frac{1+1}{2+2} = \frac{1}{2}$$
$$P(poor|pos) = \frac{1+1}{2+2} = \frac{1}{2}$$
$$P(great|pos) = \frac{2+1}{2+2} = \frac{3}{4}$$

$$\begin{aligned} P(pos|sentence) &\propto P(sentence|pos) \times P(pos) \\ &= P(good|pos) \times P(good|pos) \times P(poor|pos) \times P(great|pos) \times P(pos) \\ &= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{3}{4} \times \frac{2}{5} \\ &= \frac{3}{80} \approx 0.0375 \end{aligned}$$

3.2.2 Calculate negative:

Total documents: 3.

$$P(good|neg) = \frac{2+1}{3+2} = \frac{3}{5}$$
$$P(poor|neg) = \frac{3+1}{3+2} = \frac{4}{5}$$
$$P(great|neg) = \frac{1+1}{3+2} = \frac{2}{5}$$

$$\begin{aligned} P(neg|sentence) &\propto P(sentence|neg) \times P(neg) \\ &= P(good|neg) \times P(good|neg) \times P(poor|neg) \times P(great|neg) \times P(neg) \\ &= \frac{3}{5} \times \frac{3}{5} \times \frac{4}{5} \times \frac{2}{5} \times \frac{3}{5} \\ &= \frac{216}{3125} \approx 0.06912 \end{aligned}$$

3.2.3 Binarized NB conclusion:

With binarized Naive Bayes, the classification is most likely to be **negative** because $P(neg|sentence) > P(pos|sentence)$

3.3 Conclusion:

The 2 models **disagree** because Multinomial NB classifies postive while Binarized NB classifies negative.

4 Problem 4

4.1 Term Frequency (TF)

The term frequency $tf(w, D)$ is the count of a word w in a document D. The term frequencies for each document are given in the table below:

Word	D_1	D_2	D_3	D_4	D_5
Data	1	0	0	2	0
System	1	0	0	0	0
Algorithm	0	1	1	1	0
Computer	0	0	1	0	1
Geometry	0	0	1	0	0
Structure	0	0	0	1	0
Analysis	0	0	0	1	0
Organization	0	0	0	0	1

Table 3: Term Frequency Table

From the table:

$$tf_{D_1} = \left[\frac{1}{2}, \frac{1}{2}, 0, 0, 0, 0, 0, 0\right]$$

$$tf_{D_2} = [0, 0, 1, 0, 0, 0, 0, 0]$$

$$tf_{D_3} = \left[0, 0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0, 0\right]$$

$$tf_{D_4} = \left[\frac{2}{5}, 0, \frac{1}{5}, 0, 0, \frac{1}{5}, \frac{1}{5}, 0\right]$$

$$tf_{D_5} = \left[0, 0, 0, \frac{1}{2}, 0, 0, 0, \frac{1}{2}\right]$$

4.2 Inverse Document Frequency (IDF)

The inverse document frequency is computed as:

$$idf(w) = \log \frac{N}{df(w)}$$

where $N = 5$ (total documents) and $df(w)$ is the number of documents containing w. The IDF values are:

$$\begin{aligned}
idf(\text{Data}) &= \log \frac{5}{2} \approx 0.916 \\
idf(\text{System}) &= \log \frac{5}{1} \approx 1.609 \\
idf(\text{Algorithm}) &= \log \frac{5}{3} \approx 0.511 \\
idf(\text{Computer}) &= \log \frac{5}{2} \approx 1.609 \\
idf(\text{Geometry}) &= \log \frac{5}{1} \approx 1.609 \\
idf(\text{Structure}) &= \log \frac{5}{1} \approx 1.609 \\
idf(\text{Analysis}) &= \log \frac{5}{2} \approx 0.916 \\
idf(\text{Organization}) &= \log \frac{5}{1} \approx 1.609
\end{aligned}$$

4.3 TF-IDF

$$\begin{aligned}
tfidf_{D_1} &= \left[\frac{1}{2}\log\frac{5}{2}, \frac{1}{2}\log 5, 0, 0, 0, 0, 0, 0\right] \\
tfidf_{D_2} &= \left[0, 0, \log\frac{5}{3}, 0, 0, 0, 0, 0\right] \\
tfidf_{D_3} &= \left[0, 0, \frac{1}{3}\log\frac{5}{3}, \frac{1}{3}\log\frac{5}{2}, \frac{1}{3}\log 5, 0, 0, 0\right] \\
tfidf_{D_4} &= \left[\frac{2}{5}\log\frac{5}{2}, 0, \frac{1}{5}\log\frac{5}{3}, 0, 0, \frac{1}{5}\log 5, \frac{1}{5}\log 5, 0\right] \\
tfidf_{D_5} &= \left[0, 0, 0, \frac{1}{2}\log\frac{5}{2}, 0, 0, 0, \frac{1}{2}\log 5\right]
\end{aligned}$$

4.4 Query: "Geometry Algorithm Concepts"

With $Geometry = w_5$ and $Algorithm = w_3$:

$$tfidf_{Query} = \left[0, 0, \frac{1}{2}\log\frac{5}{3}, 0, \frac{1}{2}\log 5, 0, 0, 0\right]$$

4.5 Cosine similarity

$$||Q|| = \sqrt{(\frac{1}{2}\log\frac{5}{3})^2 + (\frac{1}{2}\log 5)^2} = 0.844$$

Document 1:

$$cos(D_1, Q) = \frac{D_1 \cdot Q}{||D_1|| \times ||Q||} = \frac{\frac{1}{2}\log\frac{5}{3} \times 0 + \frac{1}{2}\log 5 \times 0}{||D_1|| \times ||Q||} = 0$$

Document 2:

$$||D_3|| = 0.511$$

$$cos(D_2, Q) = \frac{D_2 \cdot Q}{||D_2|| \times ||Q||} = \frac{\frac{1}{2}\log\frac{5}{3} \times \log\frac{5}{3} + \frac{1}{2}\log 5 \times 0}{0.511 \times 0.844} = 0.303$$

Document 3:

$$||D_3|| = 0.64$$

$$cos(D_3, Q) = \frac{D_3 \cdot Q}{||D_3|| \times ||Q||} = \frac{\frac{1}{2}\log\frac{5}{3} \times \frac{1}{3}\log\frac{5}{3} + \frac{1}{2}\log 5 \times \frac{1}{3}\log 5}{0.64 \times 0.844} = 0.879$$

Document 4:

$$||D_4|| = 0.593$$

$$cos(D_2, Q) = \frac{D_2 \cdot Q}{||D_1|| \times ||Q||} = \frac{\frac{1}{2}\log\frac{5}{3} \times \frac{1}{5}\log\frac{5}{3} + \frac{1}{2}\log 5 \times 0}{0.593 \times 0.844} = 0.0521$$

Document 5:

$$cos(D_5, Q) = \frac{D_5 \cdot Q}{||D_5|| \times ||Q||} = \frac{\frac{1}{2}\log\frac{5}{3} \times 0 + \frac{1}{2}\log 5 \times 0}{||D_5|| \times ||Q||} =$$

The ranking of cosine similarity is:

$$D_3 > D_2 > D_4 > D_1 \approx D5$$