Statistical Methods of Language Technology

Exercise 3

Nasrul Huda

May 9, 2025

1 Problem 3.1 Probability

1.1 Problem a

A pattern C {yes, no} to recognize a name event N {name, not_name} has the following properties:

$$P(C = yes|N = name) = 0.9$$
$$P(C = yes|N = not\_name) = 0.2$$

Assume the following:

- In newspaper text, around 5% of the words are names.
- In scientific text, around 1% of the words are names.

1.1.1 What is the probability to really see a name if C says so?

We need to compute $P(N = name|C = yes)$ using Bayes' rule:

$$\begin{aligned} P(N = name|C = yes) &= \frac{P(C = yes|N = name) \times P(N = name)}{P(C = yes)} \\ &= \frac{P(C = yes|N = name) \times P(N = name)}{P(C = yes|N = name) \times P(N = name) + P(C = yes|N = not\_name) \times P(N = not\_name)} \end{aligned}$$

For newspaper text where $P(N = name) = 0.05$:

$$\begin{aligned} P(N = name|C = yes) &= \frac{0.9 \times 0.05}{0.9 \times 0.05 + 0.2 \times 0.95} \\ &= \frac{0.045}{0.045 + 0.19} \\ &= \frac{0.045}{0.235} \\ &\approx 0.1915 \end{aligned}$$

For scientific text where $P(N = name) = 0.01$:

$$\begin{aligned} P(N = name|C = yes) &= \frac{0.9 \times 0.01}{0.9 \times 0.01 + 0.2 \times 0.99} \\ &= \frac{0.009}{0.009 + 0.198} \\ &= \frac{0.009}{0.207} \\ &\approx 0.0435 \end{aligned}$$

So, the probability to really see a name if C says so is approximately 19.15% for newspaper text and 4.35% for scientific text.

1.1.2 How low must the false positive rate $P(C = yes|N = not - name)$ get so that this probability goes up to 50% for both kinds of text?

We need to find the value of $P(C = yes|N = not - name) = x$ such that $P(N = name|C = yes) = 0.5$.

For newspaper text where $P(N = name) = 0.05$:

$$\begin{aligned}
0.5 &= \frac{0.9 \times 0.05}{0.9 \times 0.05 + x \times 0.95} \\
0.5(0.9 \times 0.05 + x \times 0.95) &= 0.9 \times 0.05 \\
0.045 + 0.475x &= 0.045 \\
0.475x &= 0 \\
x &= 0
\end{aligned}$$

This is not realistically possible since it would require a zero false positive rate.

For a more reasonable answer, let's solve for scientific text where $P(N = name) = 0.01$:

$$\begin{aligned}
0.5 &= \frac{0.9 \times 0.01}{0.9 \times 0.01 + x \times 0.99} \\
0.5(0.9 \times 0.01 + x \times 0.99) &= 0.9 \times 0.01 \\
0.0045 + 0.495x &= 0.009 \\
0.495x &= 0.0045 \\
x &\approx 0.0091
\end{aligned}$$

Therefore, the false positive rate needs to be approximately 0.0091 (about 0.91%) for scientific text to achieve a 50% probability. For newspaper text, the false positive rate would need to be even lower (essentially zero) to reach 50%.

1.2 Problem b

Are X and Y as defined in the following table independently distributed?

x	0	0	1	1
y	a	b	a	b
$p(X = x, Y = y)$	0.3	0.1	0.2	0.4

To determine if X and Y are independent, we need to check if $P(X = x, Y = y) = P(X = x) \times P(Y = y)$ for all x and y.

First, let's calculate the marginal probabilities:

$P(X = 0) = P(X = 0, Y = a) + P(X = 0, Y = b) = 0.3 + 0.1 = 0.4$

$P(X = 1) = P(X = 1, Y = a) + P(X = 1, Y = b) = 0.2 + 0.4 = 0.6$

$P(Y = a) = P(X = 0, Y = a) + P(X = 1, Y = a) = 0.3 + 0.2 = 0.5$

$P(Y = b) = P(X = 0, Y = b) + P(X = 1, Y = b) = 0.1 + 0.4 = 0.5$

Now, let's check independence for each combination:

For $(X = 0, Y = a)$: $P(X = 0) \times P(Y = a) = 0.4 \times 0.5 = 0.2$ $P(X = 0, Y = a) = 0.3$ Since $0.2 \neq 0.3$, this doesn't satisfy independence.

For $(X = 0, Y = b)$: $P(X = 0) \times P(Y = b) = 0.4 \times 0.5 = 0.2$ $P(X = 0, Y = b) = 0.1$ Since $0.2 \neq 0.1$, this doesn't satisfy independence.

Therefore, X and Y are not independently distributed.

1.3 Problem c

1.3.1 Compute the entropies for:

First, let's compute $H(X)$ and $H(Y)$:

$$\begin{aligned}
H(X) &= -\sum_x P(X = x) \log_2 P(X = x) \\
&= -[P(X = 0) \log_2 P(X = 0) + P(X = 1) \log_2 P(X = 1)] \\
&= -[0.4 \log_2 0.4 + 0.6 \log_2 0.6] \\
&= -[0.4 \times (-1.32) + 0.6 \times (-0.74)] \\
&= 0.528 + 0.444 \\
&= 0.972 \text{ bits}
\end{aligned}$$

$$
\begin{aligned}
H(Y) &= -\sum_{y} P(Y=y)\log_2 P(Y=y) \\
&= -[P(Y=a)\log_2 P(Y=a) + P(Y=b)\log_2 P(Y=b)] \\
&= -[0.5\log_2 0.5 + 0.5\log_2 0.5] \\
&= -[0.5\times(-1) + 0.5\times(-1)] \\
&= 0.5 + 0.5 \\
&= 1 \text{ bit}
\end{aligned}
$$

Now, compute $H(X,Y)$:

$$
\begin{aligned}
H(X,Y) &= -\sum_{x,y} P(X=x,Y=y)\log_2 P(X=x,Y=y) \\
&= -[0.3\log_2 0.3 + 0.1\log_2 0.1 + 0.2\log_2 0.2 + 0.4\log_2 0.4] \\
&= -[0.3\times(-1.74) + 0.1\times(-3.32) + 0.2\times(-2.32) + 0.4\times(-1.32)] \\
&= 0.522 + 0.332 + 0.464 + 0.528 \\
&= 1.846 \text{ bits}
\end{aligned}
$$

Next, compute $H(X|Y)$ and $H(Y|X)$:
Using the fact that $H(X|Y) = H(X,Y) - H(Y)$:

$$
\begin{aligned}
H(X|Y) &= H(X,Y) - H(Y) \\
&= 1.846 - 1 \\
&= 0.846 \text{ bits}
\end{aligned}
$$

Similarly, for $H(Y|X)$:

$$
\begin{aligned}
H(Y|X) &= H(X,Y) - H(X) \\
&= 1.846 - 0.972 \\
&= 0.874 \text{ bits}
\end{aligned}
$$

Finally, compute $D(X||Y)$, which is the Kullback-Leibler divergence:

$$
D(X||Y) = \sum_{x} P(X=x)\log_2 \frac{P(X=x)}{P(Y=x)}
$$

However, since X and Y have different domains ($X \in \{0,1\}$ and $Y \in \{a,b\}$), the Kullback-Leibler divergence is not directly applicable in this form.

Alternatively, if we're asked to compute $I(X;Y)$ (mutual information):

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) \\
&= 0.972 - 0.846 \\
&= 0.126 \text{ bits}
\end{aligned}
$$

2 Problem 3.2 Language Models

2.1 Most Frequent Words

The 20 most frequent words from the training set are:

- die
- der
- und
- in
- den
- von
- zu
- das

- mit
- sich
- des
- auf
- für
- ist
- im
- dem
- nicht
- ein
- Die
- eine

2.2 Unseen Tokens in Test Data

The percentage of tokens in the test data that have not been seen in the training data is approximately 12.71%.

2.3 Most Frequent Bigrams

The 20 most frequent bigrams from the training set are:

1. (in, der)
2. (sich, die)
3. (von, der)
4. (für, die)
5. (in, den)
6. (auf, die)
7. (die, der)
8. (in, die)
9. (an, der)
10. (mit, der)
11. (an, die)
12. (aus, der)
13. (von, den)
14. (zu, den)
15. (mit, dem)
16. (über, die)
17. (die, Welt)
18. (ist, die)
19. (bei, der)
20. (und, der)

2.4 Unseen Bigrams in Test Data

The percentage of bigrams in the test data that have not been seen in the training data is approximately 56.33%.

2.5 Unseen Trigrams in Test Data

The percentage of trigrams in the test data that have not been seen in the training data is approximately 83.73%.

2.6 Zero Probability Sentences

Under an MLE bigram model from the training data, approximately 99.64% of sentences in the test data have zero probability (36,357 out of 36,486 sentences).

2.7 Linear Interpolation Model

Using a linear combination of 0-gram, unigram, bigram, and trigram model with $\lambda_0 = 1.0 \times 10^{-10}$, $\lambda_1 = 0.01$, $\lambda_2 = 0.2$, $\lambda_3 = 1 - (\lambda_0 + \lambda_1 + \lambda_2) \approx 0.79$, the probabilities of the first 3 sentences from the test data are:

1. Sentence 1: "Aufnahme von DDR-Flüchtlingen : Lob für Ungarn in ganz Europa"
   Log probability: -41.247362
   Probability: 1.2221647772e-18

2. Sentence 2: "Bei der Aufnahme der DDR-Flüchtlinge handelt Ungarn im Einklang mit dem Völkerrecht und den internationalen Vereinbarungen"
   Log probability: -103.529861
   Probability: 1.1564869639e-45

3. Sentence 3: "So findet Außenminister Genscher"
   Log probability: -26.831436
   Probability: 4.5026105964e-12