Exercise Sheet 9

Neural Computation
University of Birmingham

Week 9 2025

Exercise 1.

The following table shows word counts from a small training corpus: The <EOS> token is used to mark the end of the sequence and should be the last token in any sequence.

Word	Count
the	15
cat	4
sat	3
on	6
mat	3
dog	4
chased	2
and	4
ran	2
quickly	2
<EOS>	15

1. How can you compute the probability table for a unigram model from the word counts?
2. Compute the probability of the sentence:

   "the cat sat on the mat"

   including the end-of-sequence token.
3. What is the most probable sequence according to the model? What is it's proability?
4. What is the most probable sequence of length 4 (including the <EOS> token)? What is it's probability?

Solution:

Step 1: Compute the unigram probabilities

The probability of a word in a unigram model is:

$$P(w) = \frac{\text{count}(w)}{\text{total count of all words}}$$

Compute the total count:

$$\text{Total} = 15 + 4 + 3 + 6 + 3 + 4 + 2 + 4 + 2 + 2 + 15 = 60$$

So the probabilities are:

$$P(\text{the}) = \frac{15}{60} = 0.25,$$
$$P(\text{cat}) = \frac{4}{60} \approx 0.0667,$$
$$P(\text{sat}) = \frac{3}{60} = 0.05,$$
$$P(\text{on}) = \frac{6}{60} = 0.10,$$
$$P(\text{mat}) = \frac{3}{60} = 0.05,$$
$$P(\text{dog}) = \frac{4}{60} \approx 0.0667,$$
$$P(\text{chased}) = \frac{2}{60} \approx 0.0333,$$
$$P(\text{and}) = \frac{4}{60} \approx 0.0667,$$
$$P(\text{ran}) = \frac{2}{60} \approx 0.0333,$$
$$P(\text{quickly}) = \frac{2}{60} \approx 0.0333,$$
$$P(\texttt{<EOS>}) = \frac{15}{60} = 0.25.$$

Step 2: Probability of "the cat sat on the mat <EOS>"

The sentence has 6 words plus the end token:

$$P = P(\text{the}) \cdot P(\text{cat}) \cdot P(\text{sat}) \cdot P(\text{on}) \cdot P(\text{the}) \cdot P(\text{mat}) \cdot P(\texttt{<EOS>})$$

$$P = (0.25)^2 \cdot 0.0667 \cdot 0.05 \cdot 0.10 \cdot 0.05 \cdot 0.25$$

Compute:

$$P \approx (0.0625) \cdot 0.0667 \cdot 0.05 \cdot 0.10 \cdot 0.05 \cdot 0.25$$

$$P \approx 2.6 \times 10^{-6}$$

Step 3: Most probable sequence

The most probable sequence is obtained by repeating the most frequent word and ending with `<EOS>`. Here, "the" and `<EOS>` both have the highest probability (0.25). So the most probable sequence is:

"the the the ... the <EOS>"

If we consider the shortest valid sequence (just `<EOS>`), its probability is:

$$P = 0.25$$

If we require at least one word before `<EOS>`, then:

$$P = 0.25 \cdot 0.25 = 0.0625$$

Step 4: Most probable sequence of length 4 (including `<EOS>`)

Length 4 means 3 words plus `<EOS>`. To maximise probability, choose the most frequent word for all positions:

"the the the <EOS>"

Probability:

$$P = (0.25)^4 = 0.0039$$

Exercise 2. Now consider the following bigram counts. Rows represent the **previous word** w_{i-1}, columns represent the **next word** w_i. Here the <SOS> token comes at the beginning of all sequences to indicate the start of the sequence.

	the	cat	sat	on	mat	dog	chased	and	ran	quickly	<EOS>
<SOS>	5	0	0	0	0	0	0	0	0	0	0
the	0	4	0	3	2	3	0	0	0	0	0
cat	0	0	3	0	0	0	0	2	0	0	1
sat	0	0	0	2	0	0	0	0	0	0	1
on	4	0	0	0	1	0	0	0	0	0	1
mat	0	0	0	0	0	0	0	0	0	0	1
dog	0	0	1	0	0	0	2	3	1	0	1
chased	2	0	0	0	0	0	0	0	0	0	1
and	2	0	0	0	0	1	0	0	0	1	0
ran	0	0	0	0	0	0	0	0	0	2	1
quickly	0	0	0	0	0	0	2	1	2	0	1

- Compute the probability of the sentence:

  "the cat sat on the mat"

  including the <SOS> and <EOS> tokens.

- Use the greedy search algorithm for finding high probability sequences. Start with the <SOS> token and continue until you reached a sequence length of 8.

Solution:

Step 1: Compute the probability of "the cat sat on the mat" including <SOS> and <EOS>

The bigram probability is:

$$P(\text{sentence}) = \prod_{i=1}^{n} P(w_i \mid w_{i-1})$$

where $w_0 =$ <SOS> and $w_{n+1} =$ <EOS>. For the sentence:

<SOS> the cat sat on the mat <EOS>

Compute each conditional probability:

$$
\begin{aligned}
P(\text{the} \mid \texttt{<SOS>}) &= \frac{5}{\sum \text{row}(\texttt{<SOS>})} = \frac{5}{5} = 1.0, \\
P(\text{cat} \mid \text{the}) &= \frac{4}{\sum \text{row}(\text{the})} = \frac{4}{12} \approx 0.3333, \\
P(\text{sat} \mid \text{cat}) &= \frac{3}{\sum \text{row}(\text{cat})} = \frac{3}{6} = 0.5, \\
P(\text{on} \mid \text{sat}) &= \frac{2}{\sum \text{row}(\text{sat})} = \frac{2}{3} \approx 0.6667, \\
P(\text{the} \mid \text{on}) &= \frac{4}{\sum \text{row}(\text{on})} = \frac{4}{6} \approx 0.6667, \\
P(\text{mat} \mid \text{the}) &= \frac{2}{12} \approx 0.1667, \\
P(\texttt{<EOS>} \mid \text{mat}) &= \frac{1}{\sum \text{row}(\text{mat})} = \frac{1}{1} = 1.0.
\end{aligned}
$$

So:

$$P = 1.0 \times 0.3333 \times 0.5 \times 0.6667 \times 0.6667 \times 0.1667 \times 1.0 \approx 0.0123.$$

—

Step 2: Greedy search for sequence of length 11 starting from <SOS>
Greedy search picks the most probable next word at each step:

$$\texttt{<SOS>} \rightarrow \text{the (highest from <SOS> row)}$$

From **the**, the highest probability next word is:

$$\text{cat } (\frac{4}{12} = 0.3333)$$

From **cat**, highest is:

$$\text{sat } (\frac{3}{6} = 0.5)$$

From **sat**, highest is:

$$\text{on } (\frac{2}{3} = 0.6667)$$

From **on**, highest is:

$$\text{the } (\frac{4}{6} = 0.6667)$$

From **the**, again pick cat (0.3333), then sat, then on, then the, and so on until length 11. Final greedy sequence:

$$\texttt{<SOS>} \text{ the cat sat on the cat sat on the}$$

—