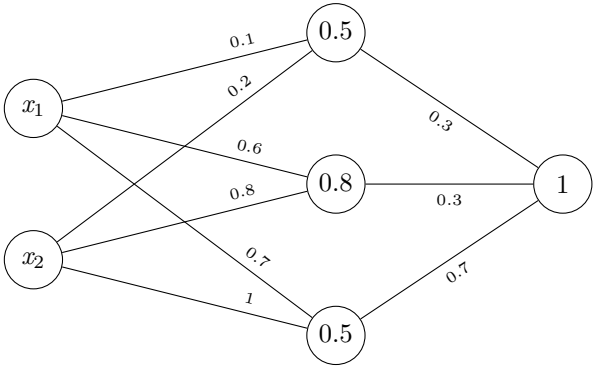
Recap



x_1	x_2	y
0	0	0.86169636
1	0	0.87786007
1	1	0.891605
10	10	0.90814614
⋮	⋮	⋮

Figure: Neural network with randomly initialized weights

Fragerunde vor der Klausur

Sprachverarbeitung (VL + Ü)

Nils Reiter

July 6, 2023

INSTITUT FÜR
DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

Fragen per Mail

- Difference between probabilities und likelihoods?
- Information gain vs. entropy
- What's the »number of numbers«?
  - On the slides in the context of language models
  - In the python script on deep learning
- Will »area under curve« be in the exam? It was skipped in the lecture
- What's B and N in Lidstone's Law?

Lidstone's Law

$$p(\langle w_1, \ldots w_n \rangle) = \frac{c(\langle w_1, \ldots w_n \rangle) + \lambda}{N + B\lambda}$$

- B: Number of different n-grams (i.e., n-gram types)
- λ: Parameter set to control how much mass remains for OOV words
  - Typical setting: $\lambda = \frac{1}{2}$ (for reasons see Manning/Schütze, 1999, 204)

Lidstone's Law

Example

- Corpus with 1000 words
- We look at bigrams ($n = 2$)
- ➔ Corpus has 999 bigrams ($= N$)
- Corpus has 300 different bigrams ($=$ bigram types $= B$)

Lidstone's Law

Example

- Corpus with 1000 words
- We look at bigrams ($n = 2$)
- ➔ Corpus has 999 bigrams ($= N$)
- Corpus has 300 different bigrams ($=$ bigram types $= B$)
- Bigram »morning sun« appears 30 times
  - Base probability (without smooting): $\frac{30}{999} = 0.03003 \simeq 3\%$

Lidstone's Law

Example

- Corpus with 1000 words
- We look at bigrams ($n = 2$)
- ➔ Corpus has 999 bigrams ($= N$)
- Corpus has 300 different bigrams ($=$ bigram types $= B$)
- Bigram »morning sun« appears 30 times
  - Base probability (without smooting): $\frac{30}{999} = 0.03003 \simeq 3\%$
  - Smoothed probability (w/ $\lambda = 1$): $\frac{30+1}{999+300\times1} = 0.02386451 \simeq 2.4\%$

Lidstone's Law

Example

- Corpus with 1000 words
- We look at bigrams ($n = 2$)
- ➔ Corpus has 999 bigrams ($= N$)
- Corpus has 300 different bigrams ($=$ bigram types $= B$)
- Bigram »morning sun« appears 30 times
  - Base probability (without smooting): $\frac{30}{999} = 0.03003 \simeq 3\%$
  - Smoothed probability (w/ $\lambda = 1$): $\frac{30+1}{999+300\times 1} = 0.02386451 \simeq 2.4\%$
  - Smoothed probability (w/ $\lambda = \frac{1}{2}$): $\frac{30+\frac{1}{2}}{999+300\times \frac{1}{2}} = 0.02654482 \simeq 2.7\%$

Lidstone's Law

Example

- Corpus with 1000 words
- We look at bigrams ($n = 2$)
- ➔ Corpus has 999 bigrams ($= N$)
- Corpus has 300 different bigrams ($=$ bigram types $= B$)
- Bigram »morning sun« appears 30 times
  - Base probability (without smooting): $\frac{30}{999} = 0.03003 \simeq 3\%$
  - Smoothed probability (w/ $\lambda = 1$): $\frac{30+1}{999+300\times 1} = 0.02386451 \simeq 2.4\%$
  - Smoothed probability (w/ $\lambda = \frac{1}{2}$): $\frac{30+\frac{1}{2}}{999+300\times \frac{1}{2}} = 0.02654482 \simeq 2.7\%$
  - Smoothed probability (w/ $\lambda = 0.1$): $\frac{30+0.1}{999+300\times 0.1} = 0.0292517 \simeq 2.9\%$