Assignment 1 + Language Properties

(SNLP tutorial 2)

Vilém Zouhar, Awantee Deshpande, Julius Steuer

4th, 6th May 2021

Organisational Issues

- Teammates
- Assignment submissions
  - Naming your assignment folder: `Name1_id1_Name1_id2.zip`
  - Your Notebooks and files should be directly inside the main folder (no unnecessary nesting)
  - Do not submit the following files:
    - `__pycache__`
    - `.ipynb_checkpoints`
    - `data/*`
    - Any other pdf or information file accompanying the assignment
  - Only submit: Notebook + Python files. Otherwise points can be deducted.

Part 1: Discussion of Assignment 1

- Exercise 1: Instructions for setup
- Exercise 2: Mandelbrot distribution + Stick breaking
- Exercise 3: Zipf's Law at word level
- Bonus: Zipf's Law at character level

Part 2: Overview of current topics

- Basics of Probability Theory
- Perplexity
- Maximum Likelihood Estimation
- Smoothing

Probability Theory for Language Models

Predict

$P(w_1, w_2 \ldots w_N)$ which can be decomposed as $\prod P(w_i|h_i)$

Bonus question

Compare for uniform, unigram, bigram, trigram... ngram models.

- Where do we assume statistical independence?
- What is this kind of assumption called?

Probability Theory for Language Models

Entropy as Expectation value

$$E[f(V)] = \sum_{w_i \in V} p(w_i) f(w_i)$$

Entropy is a property of any distribution, e.g. that of a unigram language model.

$$H = E[-\log(p(V))] = -\sum_{w_i \in V} p(w_i) \log(p(w_i))$$

What does this mean? What are we capturing by the entropy of the LM distribution?

Probability Theory for Language Models

Entropy as Expectation value

$$E[f(V)] = \sum_{w_i \in V} p(w_i) f(w_i)$$

Entropy is a property of any distribution, e.g. that of a unigram language model.

$$H = E[-\log(p(V))] = -\sum_{w_i \in V} p(w_i) \log(p(w_i))$$

What does this mean? What are we capturing by the entropy of the LM distribution?

Consider a bigram model where

$E[-\log P(w|'in')] = 10.42$ {e.g. ('in', 'fact'), ('in', 'that'), ('in', 'my')}

$E[-\log P(w|'the')] = 15.11$ {e.g. ('the', 'day'), ('the', 'most'), ('the', 'end')}

What do the expectation values indicate here?

Bonus Questions

1. What is the entropy of a fair die $p = (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$?

Bonus Questions

1. What is the entropy of a fair die $p = (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$?
2. What is the entropy of a loaded die $q = (\frac{1}{12}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{3}{12})$?

Bonus Questions

1. What is the entropy of a fair die $p = (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$?
2. What is the entropy of a loaded die $q = (\frac{1}{12}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{3}{12})$?
3. What is the cross-entropy of the same distribution? $H(p, p)$

Bonus Questions

1. What is the entropy of a fair die $p = (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$?
2. What is the entropy of a loaded die $q = (\frac{1}{12}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{3}{12})$?
3. What is the cross-entropy of the same distribution? $H(p, p)$
4. What is the cross-entropy of the loaded die q if we assume a wrongly loaded die k $(\frac{1}{6}, \frac{1}{12}, \frac{1}{6}, \frac{1}{6}, \frac{3}{12}, \frac{1}{6})$?

Bonus Questions

1. What is the entropy of a fair die $p = (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$?
2. What is the entropy of a loaded die $q = (\frac{1}{12}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{3}{12})$?
3. What is the cross-entropy of the same distribution? $H(p, p)$
4. What is the cross-entropy of the loaded die q if we assume a wrongly loaded die k $(\frac{1}{6}, \frac{1}{12}, \frac{1}{6}, \frac{1}{6}, \frac{3}{12}, \frac{1}{6})$?
5. What would happen if we assumed the correct distribution?

Bonus Questions

1. What is the entropy of a fair die $p = (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$?
2. What is the entropy of a loaded die $q = (\frac{1}{12}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{3}{12})$?
3. What is the cross-entropy of the same distribution? $H(p, p)$
4. What is the cross-entropy of the loaded die q if we assume a wrongly loaded die k $(\frac{1}{6}, \frac{1}{12}, \frac{1}{6}, \frac{1}{6}, \frac{3}{12}, \frac{1}{6})$?
5. What would happen if we assumed the correct distribution?
6. What does the difference tell us?

Perplexity

Formulae

$$PP = 2^{-\frac{1}{n}\sum_1^n \log p(w_i|w_{:i})}$$

$$PP = 2^{-\sum_{w,h} f(w,h)\log_2 P(w|h)}$$

How do these two formulae relate to each other?

Maximum Likelihood Estimation

- A way to estimate language model (distribution) parameters
- Trying to maximize probability of the training data
- NOTE: Separate the text itself from the language model
- LMs exist independent of the text and MLE only maximizes their performance on the text

LM Smoothing

Bonus Questions

1. What happens if an unknown token is encountered and the LM assigns it 0 probability?

LM Smoothing

Bonus Questions

1. What happens if an unknown token is encountered and the LM assigns it 0 probability?
2. What are some quick solutions to this issue?

LM Smoothing

Bonus Questions

1. What happens if an unknown token is encountered and the LM assigns it 0 probability?
2. What are some quick solutions to this issue?
3. How are LMs useful in downstream tasks?

LM Smoothing

Bonus Questions

1. What happens if an unknown token is encountered and the LM assigns it 0 probability?
2. What are some quick solutions to this issue?
3. How are LMs useful in downstream tasks?

LM Smoothing

Bonus Questions

1. What happens if an unknown token is encountered and the LM assigns it 0 probability?
2. What are some quick solutions to this issue?
3. How are LMs useful in downstream tasks?

Different smoothing methods will be covered in the further chapters.

Homework

- Exercise 1: Perplexity calculation by hand
- Exercise 2: Plotting n-gram distributions
- Exercise 3: MLE language models, Perplexity calculation
- Bonus: Custom alternative to perplexity

Resources

1. Why is Perplexity used over Entropy?
   https://stats.stackexchange.com/questions/285798/perplexity-and-cross-entropy-for-n-gram-models
2. On Redundancy in Natural Languages
   http://www-math.ucdenver.edu/~wcherowi/courses/m5410/m5410lc1.html