Assignment 1 + Language Properties

(SNLP tutorial 2)

Vilém Zouhar, Awantee Deshpande, Julius Steuer

May 3, 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 subtracted..

Part 1: Discussion of Assignment 1

- Exercise 1: Instructions for setup
- Exercise 2: 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

$P(w_1, w_2 ... w_N)$ 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?
- How is this 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.

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

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

Bonus Questions

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

Perplexity

Formulas

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

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

How do these two formulas relate to each other?

Other metrics

- Q: What is the advantage of mean rank over perplexity?
- Q: What is the advantage of perplexity over mean rank?

Maximum Likelihood Estimation

- A way to estimate language model (distribution) parameters
- Trying to maximize probability of the training data
- ALERT: 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

- Q: What happens if an unknown token is encountered and LM assigns it 0 probability?
- Q: What are some quick solutions to this issue?

Different smoothing methods will be covered in the further chapters.

- Q: What are LMs useful in downstream tasks?

Homework

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

Resources

1. TODO