Machine Translation: Summer Term 2021

Language Model Basics

Sangeet Sagar (7009050)
sasa00001@stud.uni-saarland.de

July 8, 2021

1. **Given a string of words $w_1w_2w_3\cdots w_n$ what do language models (LMs) tell us?**
   The Language model
   - tells us the probability of the given string $P(w)$
   - helps predict which words are likely continuations of the given string.

2. **How can you model $P(w_1w_2w_3\cdots w_n)$ faithfully using the chain rule from Probability Theory? Is this modeling "faithful", i.e. do you get exact equality?**
   Using chain rule of probability

   $$P(w_1w_2w_3\cdots w_n) = P(w_1) \times P(w_2|w_1) \times P(w_3|w_1w_2) \times \cdots \times P(w_n|w_1w_2\cdots w_{n-1})$$

   This is an exact equality.

3. **How can you estimate the component probabilities of the chain rule in (ii) above using MLE and counts from data?**

   $$P(w_1) = \frac{C(w_1)}{C(words)}$$
   $$P(w_2|w_1) = \frac{C(w_1w_2)}{C(w_2)}$$
   $$P(w_3|w_1w_2) = \frac{C(w_1w_2w_3)}{C(w_1w_2)}$$

4. **What's "wrong" (in the sense of making this not very useful) with this application of the chain rule for practical language modeling?**
   When computing $P(w_n|w_1w_2\cdots w_{n-1})$ we are relying heavily on the entire history except the word w_n. Theoretically, this is correct, but in practical condition, the history sequence $w_1w_2\cdots w_{n-1}$ is very rare to be found in the training data. If this one term becomes zero, the entire product vanishes, and the probability of the sentence becomes 0. Hence we can not afford the exact equality and we must limit the maximum context on which the probability of a word is conditioned upon.

5. **What is the Markov assumption and how can you use the Markov assumption to overcome the practical modeling problem above? In what sense is the Markov assumption an approximation?**
   Markov assumption states that for an n-gram the relevant history for a current word is only $n-1$ n-gram. It limits the context on which the probability of a word is conditioned upon.

   Let $P(w_i|w_{i-1})$ be the probability that word w_i follows word w_{i-1}. Using MLE

   $$P(w_i|w_{i-1}) =_{MLE} \frac{C(w_{i-1}w_i)}{C(w_{i-1})}$$

   It is an approximation because we can not afford to compute $P(w_1w_2w_3\cdots w_n)$ in exact. Instead we now approximate this quantity by limiting the relevant history of a word.

6. **Please turn $P(w_1 w_2 w_3 \cdots w_n)$ into bigram model using Markov assumption and beginning and end of sentence markers $< s >$ and $< /s >$.**

$$P(w_n|w_1 w_2 \cdots w_{n-1}) \approx P(w_1) \prod_{i=2}^{n} P(w_i|w_{i-1})$$

We start from 2, because we don't have any context for the first word.

7. **Please define *surprisal* and *surprisal in context* using probabilities and explain the intuitions.**
For a very likely event and if it does not occur, the surprisal is low and for a less likely event, if it happens, the surprisal is high. Therefore we have inverse relationship between probability $P(x)$ and surprisal $S(x)$

$$S(x) \sim \frac{1}{P(x)}$$
$$S(x) = \log_2(\frac{1}{P(x)}) = -\log_2 P(x)$$

If $P(x|context)$ is large, then $S(x|context)$ is small and vice-versa.

$$S(x|context) \sim \frac{1}{P(x|context)}$$
$$S(x) = \log_2(\frac{1}{P(x|context)}) = -\log_2 P(x|context)$$

If the surprisal is high then the information content transmitted is high.

8. **In your own words, what is (are) the intuition(s) behind perplexity? What is perplexity used for?**

$$PPL(w_1 w_2 \ldots w_n) = 2^{-\frac{\log_2 P(w_1 w_2 \ldots w_n)}{n}} = \left(P(w_1 w_2 \ldots w_n)\right)^{-\frac{1}{n}} = \sqrt[n]{\frac{1}{P(w_1 w_2 \ldots w_n)}}$$

In NLP, perplexity is a measure of how well a model predicts a word. Better language models have lower perplexity. Perplexity is the average de-facto size of the vocabulary. It may be used to compare probability models.

9. **Given a unigram LM:**

$$P(x) = \frac{1}{4}, P(y) = \frac{1}{2}, P(z) = \frac{1}{4}$$

What is the perplexity of the string x, y, z?

$$\begin{aligned}
\log_2(pp) &= -\frac{1}{3}(\log_2(x) + \log_2(y) + \log_2(z)) \\
&= -\frac{1}{3}(-\log_2(4) - \log_2(2) - \log_2(4)) \\
&= -\frac{1}{3}(-5) \\
&= \frac{5}{3} \\
pp &= 3.17
\end{aligned}$$

10. **Given another unigram LM:**

$$P(x) = \frac{1}{3}, P(y) = \frac{1}{3}, P(z) = \frac{1}{3}$$

What is the perplexity of the string x, y, z?

$$\begin{aligned}\log_2(pp) &= -\frac{1}{3}(\log_2(x) + \log_2(y) + \log_2(z)) \\ &= -\frac{1}{3}(-\log_2(3) - \log_2(3) - \log_2(3)) \\ &= \log_2(3) \\ pp &= 3\end{aligned}$$

11. **Given the ten numerals (the digits) $0, 1, 2, 3, 4, 5, \cdots, 9$, and assuming that all are equally probable, what is the perplexity of any 3 digit string?**

$$\begin{aligned}\log_2(pp) &= -\frac{1}{10}(\log_2(0) + \log_2(1) + \cdots + \log_2(9)) \\ &= -\frac{1}{10}(-\log_2(10) - \log_2(10) - \cdots - \log_2(10)) \\ &= \log_2(10) \\ pp &= 10\end{aligned}$$

References