N-Gram Language Model
Log of Probabilities, Laplace Smoothing, Perplexity

Dr. Varsha Patil

Link of N-gram video
https://youtu.be/zz1CFBS4NaY

N-gram Model

An **n-gram** is a contiguous sequence of **n** items from a given sample of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. The **n-grams** typically are collected from a text or speech corpus.

Chain Rule:

$$P(x_1, x_2, \ldots. x_n) = P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_1, x_2) \ldots\ldots P(x_n \mid x_1, \ldots\ldots, x_{n-1})$$

P("about five minutes from")= P(about) × P(five | about) × P(minutes | about five) × P(from | about five minutes)

Probability of words in sentences:

$$P(w_1, w_2, \ldots w_n) = \prod_i P(w_i \mid w_1, w_2, w_3 \ldots\ldots, w_{i-1})$$

Unigram(1-gram)**: No history is used.**

Bi-gram(2-gram): **One word history**

Tri-gram(3-gram): **Two words history**

Four-gram(4-gram): **Three words history**

Five-gram(5-gram):**Four words history**

As no. of previous state (history) increases, it is very difficult to match that set of words in corpus.

Generally in practical applications, Bi-gram(previous one word), Tri-gram(previous two word, Four-gram (previous three word) are used.

Advantages:

- Easy to understands, implement
- Can be easily convert to any gram

Disadvantages:

- Underflow due to multiplication of probabilities
- **Solution:** Use log. Add probabilities.
- Zero probability problem
- **Solution:** Use Laplace smoothing

Given Corpus

<S> I am Henry </S>

<S> I like college </S>

<S> Do Henry like college </S>

<S> Henry I am </S>

<S> Do I like Henry </S>

<S> Do I like college </S>

<S> I do like Henry </S>

Word	Frequency
<S>	7
</S>	7
I	6
am	2
Henry	5
like	5
college	3
do	4

Bi-gram(2-gram): **One word history**

$$P(w_1, w_2) = \prod_{i=2} P(w_2 \mid w_1)$$

$$P(w_i \mid w_{i-1}) = \frac{\text{count}(w_{i-1}, w_i)}{\text{count}(w_{i-1})}$$

Which of the following sentence is better. i.e. Gets a higher probability with Bi-gram model.

<S> I am Henry </S>

<S> I like college </S>

<S> Do Henry like college </S>

<S> Henry I am </S>

<S> Do I like Henry </S>

<S> Do I like college </S>

<S> I do like Henry </S>

Word	Frequency
<S>	7
</S>	7
I	6
am	2
Henry	5
like	5
college	3
do	4

First statement is more probable

1. <S> I like college </S>

$$=P(I \mid \langle S\rangle) \times P(\text{like} \mid I) \times P(\text{college} \mid \text{like}) \times P(\langle /S\rangle \mid \text{college})$$

$$=3/7 \times 3/6 \times 3/5 \times 3/3 = 9/70 = \mathbf{0.13}$$

$$= \log(3/7) + \log(3/6) + \log(3/5) + \log(3/3) = \mathbf{-2.0513}$$

2. <S> Do I like Henry </S>

$$=P(\text{do} \mid \langle S\rangle) \times P(I \mid \text{do}) \times P(\text{like} \mid I) \times P(\text{Henry} \mid \text{like}) \times P(\langle /S\rangle \mid \text{Henry})$$

$$=3/7 \times 2/4 \times 3/6 \times 2/5 \times 3/5 = 9/350 = \mathbf{0.0257}$$

$$=\log(3/7) + \log(2/4) + \log(3/6) + \log(2/5) + \log(3/5) = \mathbf{-3.6607}$$

<S> I am Henry </S>

<S> I like college </S>

<S> Do Henry like college </S>

<S> Henry I am </S>

<S> Do I like Henry </S>

<S> Do I like college </S>

<S> I do like Henry </S>

Word	Frequency
<S>	7
</S>	7
I	6
am	2
Henry	5
like	5
college	3
do	4

Second statement is more probable

1. <S> like college </S>

=P(like | <S>) × P(college| like) × P(</S> | college)

$=0/7 \times 3/5 \times 3/3 = \mathbf{0}$

2. <S> Do I like Henry </S>

=P(do | <S>) × P(I | do) × P(like | I) × P(Henry | like) × P(</S> | Henry)

$=3/7 \times 2/4 \times 3/6 \times 2/5 \times 3/5 = 9/350 = \mathbf{0.0257}$

Laplace Smoothing

<S> I am Henry </S>

<S> I like college </S>

<S> Do Henry like college </S>

<S> Henry I am </S>

<S> Do I like Henry </S>

<S> Do I like college </S>

<S> I do like Henry </S>

Word	Frequency
<S>	7
</S>	7
I	6
am	2
Henry	5
like	5
college	3
do	4

Unique words are : <S>, </S>, I, Henry do, like, am, college

Total unique words: 8

But we exclude <S> as it never comes in bi-gram calculations

Total unique words: 7

Give the following bi-gram probabilities estimated by Laplace model.

1. <S> like college </S>

=P(like | <S>) × P(college| like) × P(</S> | college)

=(0+1)/(7+7) × (3+1)/(5+7) ×(3+1)/(3+7)

=1/14 × 4/12 × 4/10

=**0.0095**

2. <S> Do I like Henry </S>

=P(do | <S>) × P(I | do) × P(like | I) × P(Henry | like) × P(</S> | Henry)

=(3+1)/(7+7) × (2+1)/(4+7) × (3+1)/(6+7) ×(2+1)/(5+7) ×(3+1)/(5+7)

= 4/14 × 3/11 × 4/13 × 3/12 × 4/12

=**0.0020**

First statement is more probable

Perplexity

The language model is best when it predicts an unseen test set.

Definition of Perplexity:

It is the inverse probability of the test data which is normalized by the number of words.

$$PP(w) = P(w_1, w_2, w_3, \ldots. w_N)^{-\frac{1}{N}}$$

$$PP(w) = \left(\prod_i \frac{1}{P(w_i \mid w_1, w_2, \ldots\ldots, w_{i-1})} \right)^{\frac{1}{N}} \qquad PP(w) = \left(\prod_i \frac{1}{P(w_i \mid w_{i-1})} \right)^{\frac{1}{N}}$$

Lower the value of perplexity: **Better Model**

More value of perplexity: **Confused for prediction**

WSJ Corpus **Training:** 38 million words **Test:** 1.5 million words

N-gram Order	Unigram	Bigram	Trigram
Perplexity	962	170	109

Perplexity for Bigram <S> I like college </S>

$=P(I \mid \langle S\rangle) \times P(\text{like} \mid I) \times P(\text{college} \mid \text{like}) \times P(\langle /S\rangle \mid \text{college})$

$=3/7 \times 3/6 \times 3/5 \times 3/3 = 9/70 = \mathbf{0.13}$

$\mathbf{PP(w)=(1/0.13)^{1/4}=1.67}$

Perplexity for Trigram <S> I like college </S>

$P(w)=P(\text{like} \mid \langle S\rangle\, I) \times P(\text{college} \mid I\ \text{like}) \times P(\langle /S\rangle \mid \text{like college})$

$P(w) = 1/3 \times 2/3 \times 3/3 = 2/9 = \mathbf{0.22}$

$\mathbf{PP(w)=(1/0.22)^{1/3}=1.66}$

References:

Daniel Jurafsky, James H. Martin —Speech and Language Processing, Second Edition, Prentice Hall, 2008.

Christopher D.Manning and Hinrich Schutze, — Foundations of Statistical Natural Language Processing , MIT Press, 1999.