Lecture Notes: N-gram Language Models

CS375: NLP / Williams College / Spring 2023

Recall basic definitions of probability

- We denote the *joint probability* of random variables X and Y as $P(X, Y)$. In other classes, you may have used different notation such as $P(X \cap Y)$.
- We define the *conditional probability* of random variable X given random variable Y as $P(X|Y) = P(X, Y)/P(Y)$
- We define the *marginal probability* of random variable X as $P(X) = \sum_{y \in \text{domain}(Y)} P(X, Y = y)$
- Random variables X and Y are *independent* if and only if $P(X, Y) = P(X)P(Y)$
- The *chain rule of probability* follows from the definitions above

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} P(X_i|X_1, X_2, \dots, X_{k-1})$$

As example, suppose $n = 3$, then by the chain rule of probability

$$P(X_1, X_2, X_3) = P(X_1)P(X_2|X_1)P(X_3|X_2, X_1) \tag{1}$$

We formally define a **language model** as one that computes the probability of a sequence of words

$$P(W) = P(w_1, w_2, \dots, w_n) \tag{2}$$

or computes the probability of an upcoming word

$$P(w_n|w_1, w_2, \dots, w_{n-1}) \tag{3}$$

which we also sometimes rewrite as

$$P(w_n|w_{1:n-1}) \tag{4}$$

Using the definition of conditional probabilities (chain rule), we can rewrite the previous equation as

$$P(w_n|w_1, w_2, \dots, w_{n-1}) = \frac{P(w_1, w_2, \dots, w_n)}{P(w_1, w_2, \dots, w_{n-1})} \tag{5}$$

How do we estimate the probability above from data? We can use the **maxium likelihood estimate (MLE)** which is the relative frequency based on the empirical counts in a **training set**

$$P(w_n|w_1, w_2, \dots, w_{n-1}) = \frac{\text{Count}(w_1, w_2, \dots, w_n)}{\text{Count}(w_1, w_2, \dots, w_{n-1})} \tag{6}$$

The issue is that if n is sufficiently large, we'll never see enough data to estimate the counts. So we need to make a simplifying assumption, called the **Markov assumption** that the probability of word n only depends on the previous $N - 1$ words

$$P(w_n|w_{1:n-1}) \approx P(w_n|w_{n-(N-1):n-1}) \tag{7}$$
$$= P(w_n|w_{n-N+1:n-1}) \tag{8}$$

If $N = 2$ this is called a **bigram** assumption and the equation above simplifies to

$$\begin{aligned} P(w_n|w_{1:n-1})) &\approx P(w_n|w_{n-2+1:n-1}) && (9) \\ &= P(w_n|w_{n-1:n-1}) && (10) \\ &= P(w_n|w_{n-1}) && (11) \end{aligned}$$

Combining this bigram assumption with the maximum likelihood estimate we get

$$P(w_n|w_{n-1}) = \frac{\text{Count}(w_{n-1}, w_n)}{\text{Count}(w_{n-1})} \tag{12}$$