Perplexity and Entropy Issue

Boyuan Zhuang

June 2019

1 Some Important Concept

1.1 Stochastic Process

1.1.1 Stochastic Process

Intuitively speaking, a stochastic process can be considered as a group of random variables, where each variable is indexed by an element in some mathematical set, like the set of integers. For example, stochastic process $X(t)$ indexed by the natural numbers N^0, is

$$X(t) = [x_0, x_1, \cdots, x_t, \cdots] \tag{1}$$

where $x_i, i \in \mathrm{N}^0)$ is a random variable.

1.1.2 Stationary

If stochastic process is stationary if the probabilities of the sequence of random variable are invariant **with respect to shifts in the index**, i.e.,

$$P(x_i, x_{i+1}, \cdots, x_{i+n}) = P(x_{i+k}, x_{i+k+1}, \cdots, x_{i+k+n}) \tag{2}$$

1.1.3 Ergodic

A stochastic process is said to be ergodic if its **statistical properties** (like mean, autocorrelation function) can be **deduced** from **a single, sufficiently long, random sample** of the process. In other words, a sufficiently long sample must represent the whole process. Ergordic means “遍

历 (性的)” in Chinese. Non-teachnically speaking, ergordic means the expectation over index set of a sample is the same as the expectation of the stochastic process (In Chinese, we'd like to say ” 随机过程中的平均等于某个轨迹上的时间平均”)

1.2 Entropy Rate

1.2.1 Entropy

Entropy is a measure used to estimate the uncertainty of a random variable. Here we first define Entropy of a random variable X as

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log(p(x)) \tag{3}$$

where $\mathcal{X}$ is sample space (the set of all possible value of X).

1.2.2 Entropy Rate

Entropy Rate is the Entropy of a stochastic process averaged over some index set (time and etc.). Here we define the Entropy Rate of a random sequence $\mathbf{X} = [X_1, X_2, \cdots]$ as the limit of the joint entropy of n random variable from the process X divided by n, as n tends to infinity:

$$HR(\mathbf{X}) = \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \ldots X_n) \tag{4}$$

when the limit exists.

1.2.3 Entropy Rate of Stationary and Ergodic Process

According to **Shannon-McMillan-Breiman theorem**, the Entropy Rate of a **stationary** and **ergodic** process is equal to entropy of a sufficient long sample divided by its length,

$$HR(\mathbf{X}) = \lim_{n \to \infty} -\frac{1}{n} \log(P(X_1, X_2, \ldots X_n)) \tag{5}$$

1.3 Cross Entropy

1.3.1 Cross Entropy

The cross entropy for the distributions p and q over a given set is defined as follows:

$$H(p,q) = \mathrm{E}_p[-\log q] \tag{6}$$

. Also Cross Entropy between p and q can be decomposed into the sum of entropy over p and Kullback–Leibler(KL) divergence between p and q,

$$H(p,q) = H(p) + D_{KL}(p||q) \tag{7}$$

where $D_{KL}(p||q)$ is Kullback–Leibler(KL) divergence (or we can call it relative entropy), which is defined as follows:

$$D_{KL}(p||q) = \sum_{x\in\mathcal{X}} P(x)\log\left(\frac{P(x)}{Q(x)}\right) \tag{8}$$

1.3.2 Estimate the Cross Entropy over Stationary and Ergodic Process

If $\mathbf{X}$ is a **stationary** and **ergodic** process and we want to calculate the the cross entropy between its joint probability p and some other probability q over the same stochastic process,

$$H(p,q) = \lim_{n\to\infty} -\frac{1}{n}\sum_{x\in\mathcal{X}} p\left(x_1,\ldots,x_n\right)\log q\left(x_1,\ldots,x_n\right) \tag{9}$$

again we can apply **Shannon-McMillan-Breiman theorem** and finally obtain:

$$H(p,q) = \lim_{n\to\infty} -\frac{1}{n}\log q\left(x_1 x_2 \ldots x_n\right) \tag{10}$$

2 Related Issues

2.1 Relationship between Cross Entropy and Perplexity

First of all, we assume Language $\mathbf{L}$ as a stationary and ergodic stochastic process and the joint probability of $\mathbf{L}$ is p. We try to use a language model **LM** to approximate the language **L**, in other words, we try to use a probability q, generated by the language model **LM**, to approximate p.

In order to evaluate the effeteness of the proposed **LM** in simulating the **L**, we calculate cross entropy between p and q,

$$H(p,q) = \lim_{n\to\infty} -\frac{1}{n} \sum_{x\in\mathbf{L}} p\left(x_1,\ldots,x_n\right) \log q\left(x_1,\ldots,x_n\right) \tag{11}$$

. Since **L** is stationary and ergodic, we can applying the **Shannon-McMillan-Breiman theorem** to eq.13 and obtain:

$$H(p,q) = \lim_{n\to\infty} -\frac{1}{n} \log q\left(x_1 x_2 \ldots x_n\right) \tag{12}$$

. Therefore we can use a sufficient large corpus **X** from Language **L** to approximate eq.13 to finally obtain the cross entropy between p and q:

$$H(p,q) \approx -\frac{1}{n} \log q\left(x_1 x_2 \ldots x_n\right) \tag{13}$$

.

The perplexity (short for PP) defined on the language model **LM** over a corpus **X** (the Language **L** is represented by the corpus) is shown as follows:

$$PP_{\mathbf{LM}}(\mathbf{X}) = P\left(x_1 x_2 \ldots x_N\right)^{-\frac{1}{N}} \tag{14}$$

We can modify eq.14 and obtain:

$$\begin{aligned} PP_{\mathbf{LM}}(\mathbf{X}) &= P\left(x_1 x_2 \ldots x_N\right)^{-\frac{1}{N}} \\ &= 2^{\log_2 P\left(x_1 x_2 \ldots x_N\right)^{-\frac{1}{N}}} \\ &= 2^{-\frac{1}{N}\log_2 P\left(x_1 x_2 \ldots x_N\right)} \\ &= 2^{H(p,q)} \end{aligned} \tag{15}$$

, that is to say, the **perplexity** of a model on a sequence of words is the same as the **exp of the cross-entropy**. In order words, minimizing the perplexity is the same as minimizing the cross entropy.