ENTROPY

1 Definition of ENTROPY

ENTROPY is used to evaluate text generation tasks.
Entropy is also called self-information.
In information theory and probability statistics, entropy is a measure of uncertainty in a random variable. Suppose X is a discrete random variable taking a finite number of values, and its probability distribution is

$$P(X=x_i)=p_i, \quad i=1,2,\ldots,n$$

Then the entropy of random variable X is defined as

$$H(X)=-\sum_{i=1}^{n} p_i log p_i$$

If $p_i=0$, then define $0 log 0=0$. Usually the logarithm in the above formula is based on 2 or e as the base. In this case, the unit of entropy is called bit or nat respectively. It can be seen from the definition that entropy is only Depends on the distribution of X and has nothing to do with the value of X, so the entropy of X can also be recorded as $H(p)$, that is

$$H(p)=-\sum_{i=1}^{n} p_i log p_i$$

The greater the entropy, the greater the uncertainty of the random variable. From the definition of entropy it can be verified that

$$0 \leq H(p) \leq logn$$

2 Advantages and Disadvantages of ENTROPY

2.1 Advantages of ENTROPY

- By comparing the entropy values of the models on the same data set, it can be judged which model can better capture the regularity of the data.

2.2 Disadvantages of ENTROPY

- Entropy can only provide information about the uncertainty of the model, but not about the structure or complexity of the model. Therefore, it cannot fully represent the performance of the model.
- In some cases, calculating the entropy of a model may require a large number of data samples, especially in high-dimensional spaces, which can lead to high computational costs.

References

Li Hang. Statistical Learning Methods. Tsinghua University Press. 2016.