Starts from Entropy

Ruichen Wang

November 12, 2018

Abstract

Starting with the origins of entropy and extend to some brief introduction of its related algorithms like log-likelihood, logistic regression, variational auto encoder(VAE),generative adversarial network(GAN),etc.

Contents

1 What is information?

1.1 Defination of information

How to measure the uncentainty of certain event in a mathematics?

Given x is some event, P(x) is probability which event x happens. Intuitively, the information should have inverse proportion to the probability, which is

$$I(x) = \frac{1}{P(x)}$$

We also want the information more stable,remove the division, so

$$I(x) = log\frac{1}{P(x)} = -logP(x)$$

1.2 Property of information

As a result, the $-logP(x)$ has every properties we want:

- Lower probability, higher information
- Higher probability, lower information
- Multi-event happens, the probability is multiplied. the information is summed

Mathematics As we know, $P(x) \in [0, 1]$, and larger P(x) should have smaller information.

$$P(x_1, x_2) = P(x_1) * P(x_2)$$

$$logP(x_1, x_2) = logP(x_1) + logP(x_2)$$

2 Entorpy (Expectation(sum) of Information)

2.1 Shannon‘s Information Theory

Claude Elwood Shannon(1916-2001).
1937 MIT Master degree.
1940 MIT Ph.D degree from MIT.
1948 Published a landmark paper 'A mathematical Theory of Communication'.
Entropy is defined as:

$$H(x) = \sum_{i=1}^{n} P(x_i)I(x_i) = -\sum_{i=1}^{n} P(x_i)logP(x_i)$$

2.2 Property of entropy

The property of the entropy is quite simple

- Higher probability, the less information, the lower entropy
- Non-negative, every event has some information
- Cumulative, multile events happens, the information is the sum of them.

3 Families of entropy

3.1 Cross-Entropy

Suppose we don't know P(x) yet, so we make an 'artificial' probability distribution Q(x). How can we measure the cost as we using Q(x) to approximate P(x)? We define corss entropy as:

$$H(P, Q) = -\sum_x P(x) log Q(x)$$

Estimation There are many situations where P(x) is unknown. Given a test set N observed, which comes from a Monte Carlo sampling of the true distribution P(x). Cross entropy is calculated using :

$$H(T, Q) = -\sum_{i=1}^{N} \frac{1}{N} log Q(x_i)$$

3.1.1 Relation to log-likelihood

for the maximum likelihood estimation (MAE), we have:

$$\prod_i q_i^{N_{p_i}}$$

So log-likelihood,divided by N is :

$$\frac{1}{N} log \prod_i q_i^{N_{p_i}} = \sum_i p_i log q_i = -H(p, q)$$

So maximum the likelihood is the same as minimizing the cross entropy.

3.1.2 Cross-entropy loss in multi-class classification

In machine learning, cross-entropy loss is widely used now, it often defines as :

$$L = -y log(y^{'}) = H(y, y^{'})$$

It describes the distance between the prediction and truth.

3.1.3 The simple and elegant relationship with softmax

This is worth talking here. As the softmax probability and cross-entropy loss is so so common, and they often work together.
Softmax function:

$$p_i = \frac{e^{a_i}}{\sum_{k=1}^{N} e^{a_k}}$$

Derivative of softmax $\frac{\partial p_i}{\partial \alpha_j}$:

$$\frac{\partial p_i}{\partial \alpha_j} = \begin{cases} p_i(1-p_j) & i=j \\ -p_j * p_i & i \neq j \end{cases} \tag{1}$$

The cross entropy loss:

$$L = -\sum_i y_i log p_i$$

Derivative of cross entropy loss:

$$\frac{\partial L}{\partial o_i} = -\sum y_k \frac{1}{p_k} * \frac{\partial p_i}{\partial \alpha_j}$$

From the dervative of softmax we derived earlier,

$$\frac{\partial L}{\partial o_i} = -y_i(1-p_i) - \sum_{k \neq i} y_k \frac{1}{p_k}(-p_k * p_i)$$

$$= p_i(y_i + \sum_{k \neq i} y_k) - y_i = p_i - y_i$$

This is why we often use softmax and cross entropy together, The gradient is quite simple to calculate.

3.2 Kullback-Leibler divergence