

# Starting from Entropy

Ruichen Wang

November 13, 2018

## Abstract

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

## Contents

<b>1</b>	<b>What is information?</b>	<b>2</b>
1.1	Defination of information . . . . .	2
1.2	Property of information . . . . .	2
<b>2</b>	<b>Entorpy (Expectation(sum) of Information)</b>	<b>2</b>
2.1	Shannon's information theory . . . . .	2
2.2	Property of entropy . . . . .	3
<b>3</b>	<b>Families of entropy</b>	<b>3</b>
3.1	Cross-Entropy . . . . .	3
3.1.1	Relation to log-likelihood . . . . .	3
3.1.2	Cross-entropy loss in classification . . . . .	3
3.1.3	The simple and elegant relationship with softmax . . . . .	4
3.2	Kullback-Leibler divergence . . . . .	4
3.2.1	Interpretations . . . . .	5
3.2.2	Property of KL . . . . .	5
3.2.3	Applications . . . . .	5
<b>4</b>	<b>Variational Bayesian Inference</b>	<b>5</b>
4.1	Variational inference . . . . .	5
4.2	KL divergence measure . . . . .	6
4.3	Evidence lower bound . . . . .	6
4.4	KL transformation . . . . .	7
4.5	Mean field variational inference . . . . .	7

# 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

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

## 1.2 Property of information

As a result, the  $-\log P(x)$  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.

$$\begin{aligned} P(x_1, x_2) &= P(x_1) * P(x_2) \\ \log P(x_1, x_2) &= \log P(x_1) + \log P(x_2) \end{aligned}$$

# 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:

$$H(x) = E[I(x)] = \sum_{i=1}^n P(x_i)I(x_i) = -\sum_{i=1}^n P(x_i)\log P(x_i)$$

## 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, multiple events happen, 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 use  $Q(x)$  to approximate  $P(x)$ ? We define cross 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 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  $\frac{\partial p_i}{\partial \alpha_j}$ :

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

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 derivative of softmax we derived earlier,

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

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

## 3.2 Kullback-Leibler divergence

KL divergence is also called relative entropy. It is a measure of how one probability distribution is different from a second.

For discrete probability distributions P and Q defined on the same probability space, the KL divergence from Q to P (Q with respect to P) is defined as :

$$D_{KL}(P \parallel Q) = - \sum_i P(i) \log\left(\frac{Q(i)}{P(i)}\right)$$

Actually, this can also be written as:

$$D_{KL}(P \parallel Q) = H(P, Q) - H(P)$$

Which means the more entropy using Q gets with respect to original distribution P.

### 3.2.1 Interpretations

In machine learning,  $D_{KL}(P \parallel Q)$  is often called the information gain achieved if  $Q$  is used instead of  $P$ .

Expressed in the language of Bayesian inference,  $D_{KL}(P \parallel Q)$  is a measure of the information gained when one revises one's beliefs from the prior probability distribution  $Q$  to the posterior probability distribution  $P$ .

In applications,  $P$  typically represents the true distribution of data.  $Q$  represents the model. Minimize  $D_{KL}(P \parallel Q)$  is to find a  $Q$  that closest to  $P$ .

### 3.2.2 Property of KL

- Non-negative

As a result known as Gibbs's inequality, with  $D_{KL}(P \parallel Q)$  zero if and only if  $P = Q$ .

- Asymmetric

$$D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$$

we can define symmetrised divergence as:

$$\frac{D_{KL}(P \parallel Q) + D_{KL}(Q \parallel P)}{2}$$

### 3.2.3 Applications

Generative models like VAE, we may need a new section to go through this. I will go into details later. Here I just put a bayes equation here :).

$$posterior = \frac{likelihood * prior}{evidence}$$

## 4 Variational Bayesian Inference

### 4.1 Variational inference

**Question description** Suppose we have observations  $x$ , and hidden variables  $z$ , and some fixed parameters  $\alpha$ . What we want is the posterior distribution.

$$p(z|x, \alpha) = \frac{p(z, x|\alpha)}{\int_z p(z, x|\alpha) dz}$$

In many cases, the  $\int_z p(z, x|\alpha) dz$  is intractable. we don't know how to compute it especially in high dimensions.

**Solution** The main idea behind variational methods is to pick a family of distributions over the latent variables with its own **variational parameters**.

$$q(z_{1:m}|v)$$

Then find  $v$  to make  $q$  close to the posterior.

## 4.2 KL divergence measure

As mentioned above, we can use KL for this variational inference:

$$D_{KL}(q(z) \parallel p(z|x)) = E_{q(z)} \left[ \log \frac{q(z)}{p(z|x)} \right]$$

Intuitively, According to this formula, there are three cases:

- If q is low, then we don't care (Because of the expectation)
- If q is high and p is high, good :)
- If q is high and p is low, bad :(

## 4.3 Evidence lower bound

Actually we can not minimize KL divergence. But we can minimize another function which is equal to this. This is evidence lower bound (ELBO).

**Jensen's inequality** Jensen's inequality are widely used in EM algorithm. In convex function, we have :

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

In the context of probability theory, if X is a random variable, and  $\varphi$  is a convex function, then:

$$\varphi(E[x]) \leq E[\varphi(x)]$$

**Back** to the problem, we have observations  $x^1, x^2, \dots, x^n$ , we want  $p(x^i)$  get the max probability. Using MLE on it, which is the sum of the log-likelihood,

$$\log p_{\theta}(x^1, x^2, \dots, x^n) = \sum_{i=1}^N \log p_{\theta}(x^i)$$

and

$$\begin{aligned} \log p(x) &= \log \int_z p(x, z) \\ &= \log \int_z p(x, z) \frac{q(z)}{q(z)} \\ &= \log \left( E_q \left[ \frac{p(x, z)}{q(z)} \right] \right) \\ &\geq E_q \left[ \log \frac{p(x, z)}{q(z)} \right] \\ &\geq E_q[\log p(x, z)] - E_q[\log q(z)] \end{aligned}$$

Note the second term is the entropy

But what does this have to do with the KL divergence?

#### 4.4 KL transformation

As mentioned, we want  $q(z)$  and  $q(z|x)$  are close to each other:

$$\begin{aligned} KL(q(z)||q(z|x)) &= E_q \left[ \log \frac{q(z)}{p(z|x)} \right] \\ &= E_q[\log q(z)] - E_q[\log p(z|x)] \end{aligned}$$

As we know,

$$p(z|x) = \frac{p(z, x)}{p(x)}$$

so

$$\begin{aligned} KL(q(z)||q(z|x)) &= E_q[\log q(z)] - E_q[\log(z, x)] + E_q[\log p(x)] \\ &= -(E_q[\log q(z, x)] - E_q[\log q(z)]) + \log p(x) \end{aligned}$$

The first term is ELBO we just met.

The formula can also be written as :

$$\log p(x) = KL(q(z)||q(z|x)) + (E_q[\log q(z, x)] - E_q[\log q(z)])$$

As I mentioned before, For two different distributions, KL divergence is always larger than zero. and  $p(x)$  is the observation evidence, which is fixed. So minimizing the KL divergence is the same as maximizing the ELBO. This is also called as the variational lower bound.

#### 4.5 Mean field variational inference

Mean field theory is also called **self-consistent field theory**. It studies the behavior of large and complex stochastic models by studying a simpler model. Such models consider a large number of small individual components that interact with each other.

The effect of all the other individuals on any given individual is approximated by a single averaged effect, thus reducing a many-body problem to a one-body problem.

We assume each variable is independent.

$$q(z_1, \dots, z_m) = \prod_{i=1}^m q(z_i)$$