Entropy and Information*

Zhesheng Zheng
15220162202530
Xiamen University

March 17, 2019

1 Information

- Information can be described as 'degree of surprise'.
- Information can be thought as the reduction in uncertainty from learning an outcome.
  That is an event that is more unlikely happening contains more information. But we need a measurement of uncertainty to measure information. Three conditions the measurement of uncertainty should satisfies:
  1. The measurement must be continuous so that there are no jumps.
  2. The measurement must be additive across events or states.
  3. The measurement must increase as the number of events/states increase.

2 Entropy and Uncertainty

A function that satisfies these three conditions is the information entropy:(The expectation of log probability)

$$H(p) = -E_p[log(p)] = -\int p(x)log(p(x))dx$$

*Homework for Foundations of Statistical Learning.

OR

$$H(p) = -E_p[log(p)] = -\sum_i p_i log p_i$$

1. The expectation of log probability is continuous: the first derivative of log function exists in (0,1].

2. The expectation of log probability is additive across events/states: for two independent events A and B, $p(AB) = p(A)p(B)$, then the entropy of these two events is $H(AB) = H(A) + H(B)$

3. The expectation of log probability is increase as the number of events/states increase: the logic is the same as above, the entropy is always positive.

The expectation of log probability, that is the entropy is a good measurement of uncertainty and a good measurement of information.

3 Maximum Entropy (maxent)

3.1 Cross Entropy

We denote cross entropy to measure the level of uncertainty contained in the distribution q we use to describe x, where the true distribution of x is p.

$$H_c(p, q) = -E_p[log(q(x)] = -\int p(x) log(q(x)) dx$$

OR

$$H_c(p, q) = -E_p[log(q(x)] = -\sum_i p_i log(q_i)$$

3.2 KL Divergence

Kullback-Leibler Divergence measures the additional information required to find the true distribution of x conditional on we think the distribution of x is q. (Notice that the information is the decrease of the uncertainty.)

$$D_{KL}(p||q) = H_c(p, q) - H(p) = \int p(x) log(\frac{p(x)}{q(x)}) dx$$

OR

$$D_{KL}(p||q) = H_c(p,q) - H(p) = \sum_i p_i log(\frac{p_i}{q_i})$$

3.3 Maxent

Maximum entropy is the notion of finding distributions consistent with constraints and the current state of our knowledge. In other words, it is the distribution has highest uncertainty and lowest information, the distribution is the least surprising distribution. For a simple example, we gonna maximize

$$H(p) = -\sum_i p_i log(p_i)$$

with the constrain

$$\sum_i p_i = 1$$

in order to find the maximum value, we solve this problem using langrangian method

$$L = -\sum_i p_i log(p_i) + \lambda(\sum_i p_i - 1)$$

$$\frac{\partial L}{\partial p_j} = 0 \Rightarrow -(1 + log(p_j)) + \lambda = 0$$

The result is that all p_j are equal, thus the distribution maximizes entropy.

4 Application: Exponential Family is maxent distributions

Exponential family is the most common distribution, whose pdf or pmf can be defined as follow:

$$p(x|\theta) = \frac{1}{Z(\theta)} h(x) e^{\theta^T \phi(x)}$$

which contains Gamma distribution, Normal distribution, Exponential distribution,

Poisson distribution and Binomial distribution. All of them are maximum entropy under different constraints. So, the exponential family is one of the least surprising distribution family!

5 Reference

[1]Harvard AM207, http://am207.info/wiki/Entropy.html

[2]Jiaming Mao, 2019, Foundations of statistical learning, https://github.com/jiamingmao/data-analysis/blob/master/Lectures/Foundations_of_Statistical_Learning.pdf