Fabinger's Data Science Class Notes, July 15th-22nd 2019

Jose Zero

July 30, 2019

Abstract

Entropy, different distribution comparison methods, different types of divergence and their properties. I have compiled the end of last week's discussions into this week's. Also added links that I found clear and useful in the Resources section. There will be a separate write up with examples.

1 Introduction

Given an event ν with probability p we define surprise of ν as $-log(\nu)$. This can also be interpreted as the minimum amount of bits needed to half uncertainty.

Note: In many formulas entropy is preceded by a constant "k" which comes from using different logarithm base.

If ν is a "sure thing" with 100 percent probability its "surprise" is 0, on the other hand the more unlikely is an event the higher its surprise. We define entropy as the probability weighted "surprise" in a set:

$$Entropy(V) = -\sum_{\nu} p(V = \nu) logp(V = \nu)$$

Let's look at Entropy in a slight different way:

$$Entropy(V) = \sum_{\nu} p(V = \nu) log(1/p(V = \nu)) \tag{1}$$

If we are transmitting a message from a station to an outpost. There are certain number of messages which can be transmitted as bits. The number of bits needed depends on the number of outcomes. The logarithm in base 2 of the number of outcomes is the number of bits needed.

As an example for an eight event outcome three bits are needed: $log_2(2^3) = 3$

If all the message outcomes are equally distributed the possibility of each one is .125.

We can then think that number of outcomes as $1/p = 1/.125 = 8$. We can say that receiving a 3 bit message multiplied our knowledge by eight, we fully used the information in the 3 bits and hence Entropy in this case is 3.

On the other hand if there is only one outcome possible we have E=0; we are not using any of the received bits.

In a general distribution we can define Entropy as a general case usign the formula (1).

We can think of the two extreme cases above (outcome of message only one value ,E=0 ; or equally distributed as the input, E=3) as particular cases. We define the Kullback-Leibler cross entropy as:

$$D_{KL}(P||Q) = \sum_{\nu} P(V = \nu) log(Q(V = \nu)/P(V = \nu))$$

Let's look at a case where $D_{KL}(P||Q)$ but $D_{KL}(Q||P)$ is large:

Take P to be bimodal with cusps at say $x = A$ and $x = B$ and Q to be single mode with cusp at just $x = A$. Because of this $D_K L$ is not suitable to define a metric (not symmetric).

Expanding the log yields that the D_{KL} is the difference between the Q and P entropies. Note the $D_{KL}(P||P) = 0$

We want to prove that optimum situation is when the target distribution is identical to P. For that we can simply prove that $D_K L \geq 0$.

The proof can be based on the Gibb's and Jensen's inqualities (if $\rho(x) concave then E(x) \geq E(\rho(x))$ as log is a concave function (take $X = Q/P$ and $\rho = log$) Also using the follwoing Lemma:

Lemma: $log(x) \leq x - 1$ for $x > 0$. Consider $g(x) = x - log(x)$ $g(x)^{'} = 1 - 1/x > 0, \forall x > 1, g(x) increasing$ so the lemma holds. For $0 < x < 1, log(x) < 0$ and $x > 0$; the lemma is true.

1.1 Mutual Information

Consider the distributio P of two random varibale X and Y, P(X,Y). Knowing the happening of an outcome of X means nothing about why if the variables are independent. On the other hand we want to measure how indeppendent thiese two vairable are.

Define $P_X(Y) = \sum_X P(X, Y)$, the sum is replaced with an integral in the continuous case. In the same manner $P_Y(X) = \sum_Y P(X, Y)$. These are called the marginal distributions of $P(X, Y)$. Then we define the Mutual Information of X and Y as:

$$I(X, Y) = D_{KL}(P(X, Y)||P_x \times P_Y)$$

Clearly if X and Y are independent $P(X, Y) = P_X \times P_Y$ and then $I(X, Y) = 0$ Note: In the case of multidimensional variables the product of $P_X \times P_Y$ is replaced with a tensor product $P_X \bigotimes P_Y$

Some properties:

- $I(X, Y) \geq 0$
- $I(X, Y) = I(Y, X)$

- $I(X,Y) = H(y) - H(Y|X)$

To see this consider:

$$I(X,Y) = \sum_{X,Y} P(X,Y) \times log(P(X,Y)/P_X P_Y)$$

$$I(X,Y) = \sum_{X,Y} P(X,Y) \times log(P(X,Y)/P_X) - \sum_{X,Y} P(X,Y) \times log(P_Y)$$

Aplaying Bayes and comparing with the defitinion of H get, $P_{Y|X}(X,Y) = P(X,Y)/P_X(Y)$, and the result follows.

d) $I(X,Y) = E_Y(D_{KL}(P_{X|Y}))$

$$I(X,Y) = \sum_{X,Y} P(X,Y) \times log(P(X,Y)/P_X P_Y)$$

$$= \sum_{X,Y} P_X(x) P_{Y|X=x} \times log(P_{Y|X=x}(X,Y)/P_Y(y))$$

$$= \sum X P_X(x) \sum_{Y} P_{Y|X=x}(x,y) \times log(P_{Y|X=x}(x,y)/P_Y(y))$$

Hence,

$$I(X,Y) = E_X D_{KL}(P_{Y|X}||P_Y)$$

Let's define Cross Entropy of two distributions P,Q as

$$H_{Cross}(P,Q) = E_P(-log(Q)) = \sum_x P(x)(-log(Q(x))$$

Another divergence is the Jensen-Shannon Divergence, $JSD(P||Q)$

$$JSD(P||Q) = [D_{KL}(P||M) + D_{KL}(Q||M)]/2 \; where \; M = [P+Q]/2$$

Here M is the distribution of the mixture random variable X of P and Q using a binary random variable Z which is 0 with probability 1/2 and 1 with probability 1/2. Claim:

$$I(X,Z) = H(X) - H(X|Z)$$

$$I(X,Z) = -\sum_X M log(M) - H(X|Z)$$

$$I(X,Z) = -(1/2)\sum_X P(X) log(M(X)) - (1/2)\sum_X Q(X) log(M) +$$

$$+(1/2)\sum_X P log(P + (1/2)\sum_X Q log(Q)$$

combining

$$I(X,Z) = 1/2 \sum_X P(X)(log(P(X)) - log(M(X)) +$$

$$1/2 \sum_X Q(X)(log(Q(X)) - log(M(X))$$

which is

$$I(X,Z) = JSD(P||Q)$$

Also

$$I(X,Z) = H(X) - H(Y|Z)$$

and since $I(Z,X) = I(X,Z) \Rightarrow I(X,Z) = H(Z) - H(Z|X)$. We get

$$0 \leq JSD(P||Q) \leq H(Z)$$

This divergence has the property of being symmetric (unlike D_{KL})

1.2 Recommended materials

A repeat but very good worked out example. Worth the time watching and redoing the calculations. `https://youtu.be/IPkRVpXtbdY`

Entropy is not disorder! Nice video with simple graphics: `https://youtu.be/vX_WLrcgikc`

This video by Tsallis himself is informative in the sense of the background motivations of the entropy generalizations:
`https://youtu.be/-pnT_MKc9S4`

Very nice short and meaty video on Cross Entropy and D_{kl}
`https://youtu.be/ErfnhcEV1O8`

The wikipedia page for the Jensen-Shannon divergence.