Entropy, Cross-Entropy and KL-Divergence Brief Review

Lizi Chen

Motivation: In Machine Learning, cross-entropy is commonly used as a cost function when training classifiers.

1 Background

Laws of nature states that things goes from in-order state to dis-order state. The number of micro-states of an in-order state is much greater than the one of a dis-order state. The definition of Entropy in Physics:

$$S = \kappa \ln \Omega$$

κ is Boltzmann constant, Ω is the number of micro-states.

The Mathematical Theory of Communication, Claude Shannon states that: To transmit one bit of information (from sender to recipient), means to reduce the recipient's uncertainty by a factor of 2 (assume that probability of each event is the same).

Example: A event of two equally likely possible options; A is 50% chance and the other B is also 50% chance, given that the sender told the recipient that the next event will be B, the sender just reduce the recipient's uncertainty for the event by 2.

1.1 Compute the actual bit of information:

N is uncertainty reduction factor, the number of possible options.
When $N = 2$, a binary prediction situation, transmitting $\log_2 N = \log_2 2 = 1$ bit is good enough.
When $N = 8$, effective bit of data transmission requires $\log_2 8 = 3$ bits.
(Again, all options have the same possibility.)

1.2 When the possibilities are not equally likely:

Say, A is 75% chance and B is 25% chance. When we are told B is going to occur, the uncertainty reduction factor becomes

$$\frac{1}{25\%} = 4$$

Therefore, we have the actual bit $= \log_2 4 = 2$ bits for transmitting B.
Similarly, for transmitting A, we need $\log_2(1/0.75) = 0.41$ bits of information.

Observation: The event with higher probability; when transmitting, requires less bits. On the opposite, the one with lower probability requires more bits.

2 Entropy:

In the example above; transmitting A with P_A probability and B with P_B probability, each requires b_A bits and b_B bits of information. Thus, the entropy is defined as:

$$E = P_A \times b_A + P_B \times b_B$$

which is the expected amount of bits of information for a transmission from sender to recipient. In the example, we have $75\% \times 0.41 + 25\% \times 2 = 0.81$ bits.

In general, entropy is defined as:

$$H(p) = -\sum_i p_i \times \log_2(p_i)$$

Note: Entropy is always non-negative.

3 Cross-Entropy:

In short, Cross-Entropy is the average weighted length of a message. (i.e., number of bits).

For example: Four different messages; A, B, C, D are sending between a sender and a recipient, we need to know the encoding for these messages. Optimally, we need to know the best encoding so that we use less 0's and 1's for the message representation. Assume all four messages have the same probability; $\frac{1}{4}$, we use the Table 1. Our most optimal length of message encoding is $2 \times \frac{1}{4} \times 4 = 2$ bits.

Message	A	B	C	D
Code	00	01	10	11

Table 1: Same probability messages

Now assume we have different probabilities for the four messages, as shown in Table 2, our expected encoding length is $1_A \times \frac{1}{2} + 2_B \times \frac{1}{4} + 3_C \times \frac{1}{8} + 3_D \times \frac{1}{8} = 1.75$ bits, better than the previous 2 bits for the same-length encoding.

Message	A	B	C	D
Probability	1/2	1/4	1/8	1/8
Code	0	10	110	111

Table 2: Different probability messages

3.1 Encoding Space

In the example that the code for A is 0, and thereafter the rest codes cannot start with 0. This way we can decode unique series of code. For example, if we have A as 0, B as 01, and C as 1 then when the recipient receive 0101, it can be interpreted as ACB, or BB, or BAC. Such coding is wrong. We need follow the **Prefix Codes** method for different code length to avoid ambiguity. To visualize the encoding space, we have Figure 3.1. When we choose 1 for the second bit, $\frac{1}{4}$ encoding space will no longer be use-able.

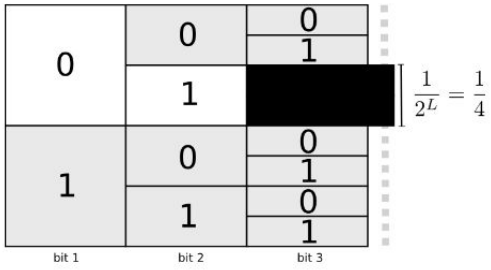


Figure 1: Encoding Space Example

3.2 Optimal Encoding

In general, a code of length $\mathcal{L}$ results in lost of $\frac{1}{2^{\mathcal{L}}}$ encoding space.

Therefore, for limited length of a code, we want to optimize the use of encoding space. That is, (1) use code that is as short as possible, (2) use as many encoding space as possible.

Optimal Encoding can be reached when we assign encoding space loss to a message according to its probability $P(X)$, where X is the message, $P(X)$ means its **prior probability of showing-up among all messages**.

Let $\mathcal{L}(X)$ be the length of X, we have $\frac{1}{2^{\mathcal{L}(X)}} = P(X)$, which is $\mathcal{L}(X) = \log_2 \frac{1}{P(X)}$. Therefore, for a probability distribution P, we have the expected weighted average code length:

$$H(P) = \sum_{x \in X} \underbrace{p(x)}_{prob} \underbrace{\log_2 \frac{1}{p(x)}}_{\mathcal{L}(x)}, \text{ where } H(P) \text{ is the Entropy.}$$

$H(P)$; the entropy, means the average length code using optimal encoding under the P distribution.

3.3 Not Using Optimal Encoding

Now we have messages in probability distribution Q; however, we choose to use another distribution P. Thus, the expected average length of code is:

$$H(Q||P) = H_P(Q) = \sum_{x \in X} q(x) \log_2 \frac{1}{p(x)}$$

$H(Q||P)$ or $H_P(Q)$ is the **Cross Entropy** for Q on P. It measures the average length of code for using P distribution on Q message distribution.

Example: We have Table 3. The third row; Probability (Q), is the true probability for the message to show up. The fifth row; P(X), is the distribution we use according to the code. It is obvious to notice that we use longer code for message that appears more frequently; D is 111 but has 40% to show up. Such encoding is not a good choice, since we will be more likely to send more bits.

$$H(Q||P) = 0.1 \times \log_2 1/(1/2) + 0.2 \times \log_2 1/(1/4) \cdots = 2.6$$

Message	A	B	C	D
Code	0	10	110	111
Probability (Q)	10%	20%	30%	40%
Code Length	1	2	3	3
P(X)	1/2	1/4	1/8	1/8

Table 3: Different probability messages

This means on average a sender has to send 2.6 bits. However, if we set Q = P and follow Table 2, we can have $H = 1.75$ bits, which is much better than 2.6 bits for such set-up.

When P is 'perfect'; meaning when $P = Q$, Cross Entropy equals to Entropy. Otherwise, Cross Entropy is greater than Entropy. Such difference between Cross Entropy and Entropy is called **Relative Entropy**, or more commonly called **KullbackLeibler (KL) Divergence**, as reviewed in the next section.

4 KullbackLeibler (KL) Divergence (Relative Entropy)

Two probability distributions; Q and P, the KL-divergence measures the similarity of Q and P. If $KL\_Divergence(Q||P) = 0$, the two distributions are equal.

For discrete probability distributions:

$$D_{KL}(Q||P) = \sum_i Q(i) \log \frac{Q(i)}{P(i)}$$

For continuous probability distributions:

$$D_{KL}(Q||P) = \int_{-\infty}^{\infty} q(x) \log \frac{q(x)}{p(x)} dx$$

Also, as mentioned previously, the difference between Cross Entropy and Entropy is KL Divergence. Therefore, we can have:

$$D_{KL}(Q||P) = H(Q||P) - H(Q)$$

Figure 4 illustrates such relationship.

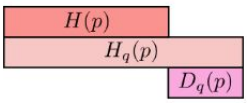


Figure 2: KL Divergence, Entropy and Cross Entropy

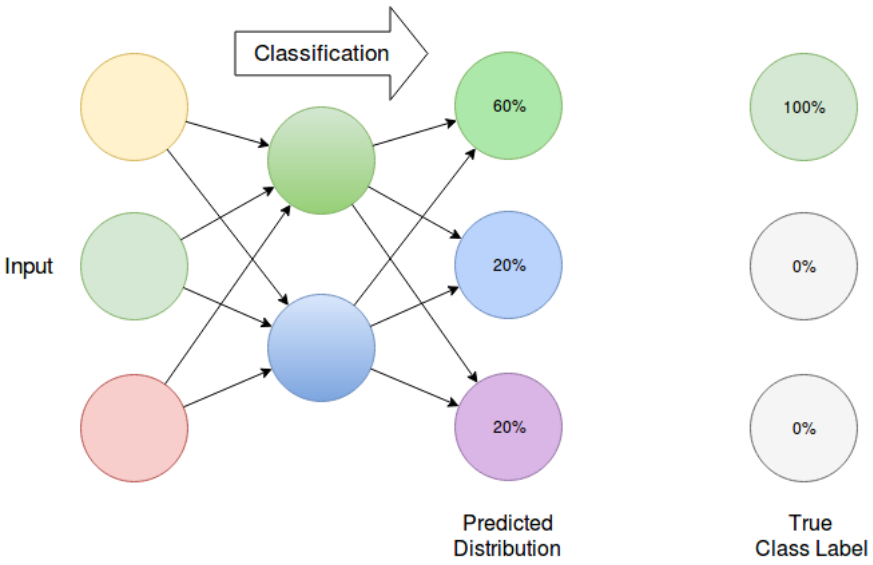
5 Cross-Entropy as Cost Function in Machine Learning

Cross-Entropy is widely used in machine learning as a cost function for classifier. In a supervised learning set-up, after a classifier, the true class distribution is Q; commonly

represented as y_i, and the predicted class distribution is P. The Cross-Entropy Loss is defined as:

$$\text{Loss} = H(Q, P) = -\sum_{i}^{N} q_{y_i} \log p_{x_i}$$

, and the objective is to optimize $\arg\min_{P(X)} Loss$.



6 Jensen-Shannon Divergence

7 Joint Entropy

8 Conditional Entropy

Given r.v. X, the uncertainty of r.v. Y is defined as:

$$\begin{aligned} H(Y|X) &= \sum_{x} p(x) H(Y|X = x) \\ &= -\sum_{x} p(x) \sum_{y} p(y|x) \log p(y|x) \\ &= -\sum_{x} \sum_{y} p(x, y) \log p(y|x) \\ &= -\sum_{x,y} p(x, y) \log p(y|x) \end{aligned} \tag{1}$$

References

[1] "Journey into information theory, Khan Academy," available at https://www.khanacademy.org/computing/computer-science/informationtheory.

[2] A. Geron, "Short introduction to entropy, cross-entropy, and kl-divergence," available at https://www.youtube.com/watch?v=ErfnhcEV1O8&t=393s.

[3] S. L. Paul Penfield, "MIT 6.050j information and entropy, spring 2008," available at https://www.youtube.com/playlist?list=PLDDE03B3BDCA1D9B1.