Entropy, Relative Entropy, Cross Entropy

Entropy

Entropy, H(x) is a measure of the uncertainty of a discrete random variable.

$$H(x) = -\sum_{x \in X} p(x) log(p(x)) = E_p log \frac{1}{p(X)}$$

Properties:

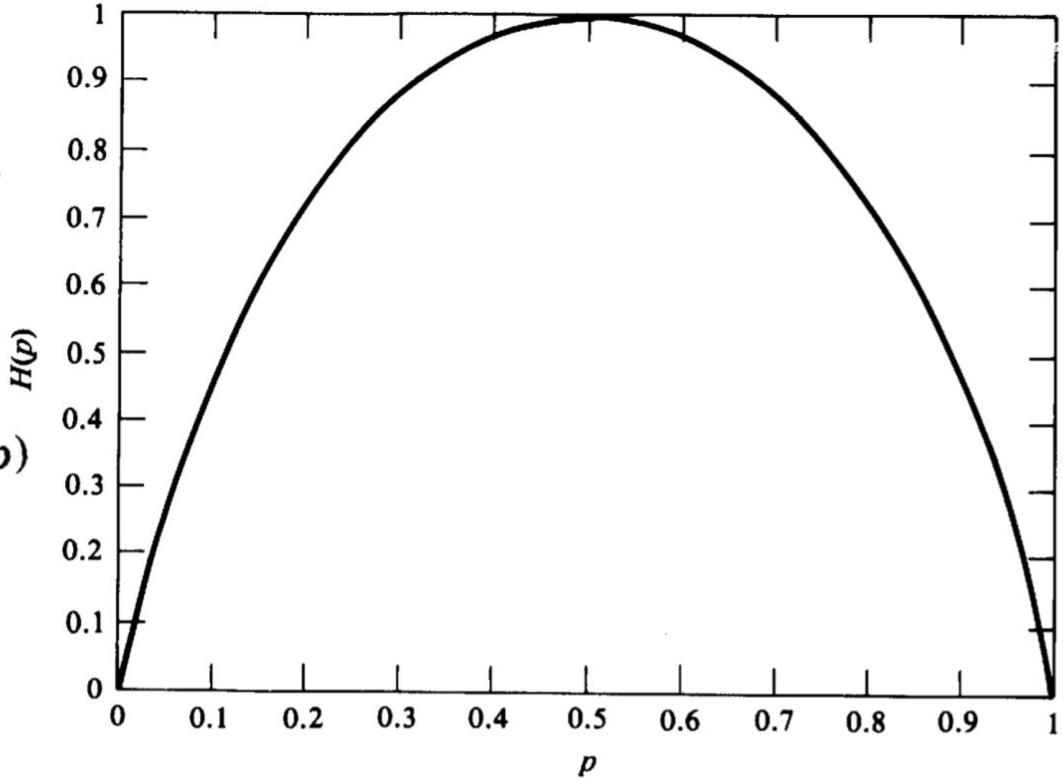
- H(x) >= 0
- $H_b(X) = (\log_b a) H_a(X).$

Entropy

$$X = \begin{cases} 1 & \text{with probability } p\,, \\ 0 & \text{with probability } 1-p\,. \end{cases}$$

$$H(X) = -p \log p - (1-p)\log(1-p)$$

$$\boxed{-\log_2 p + \log_2(1-p) = \log_2 \frac{1-p}{p}}$$



Entropy

- Lesser the probability for an event, larger the entropy.

Entropy of a six-headed fair dice is $\log_2 6$.

$$-\sum_{i=1}^{n} \frac{1}{n} \log_2 \frac{1}{n} = \log_2 n.$$

Entropy : Properties

Primer on Probability Fundamentals

- Random Variable $X : \Omega \mapsto \mathbb{R}$
- Probability $p(X = a) = \sum_{s \in \Omega, X(s)=a} p(s)$
- Expectation $\mathbb{E}(\mathbb{X}) = \sum_a a.\, p(X = a)$
- Linearity of Expectation $\mathbb{E}(\sum_{\mathrm{i=1:n}} \mathbb{X}_{\mathrm{i}}) = \sum_{i=1:n} \mathbb{E}(\mathbb{X}_{\mathrm{i}})$

Entropy : Properties

Primer on Probability Fundamentals

- Jensen's Inequality $$\mathbb{E}[\mathbf{f}(\mathbf{x})] \geq f(\mathbb{E}[\mathbf{x}])$$

  Ex:- $$\mathbb{E}[\mathbb{X}^2] \geq \mathbb{E}[\mathbb{X}]^2$$

  Subject to the constraint that, f is a convex function.

Entropy : Properties

- H(U) >= 0,

$$H(U) = \mathbb{E}\left[\log \frac{1}{p(U)}\right] \geq 0 \text{ because } \log \frac{1}{p(U)} \geq 0$$

Where, U = {u_1, u_2, …, u_M}

- H(U) <= log(M)

$$\begin{aligned} H(U) &= \mathbb{E}\left[\log \frac{1}{p(U)}\right] \\ &\leq \log \mathbb{E}\left[\frac{1}{p(U)}\right] \\ &= \log \sum_u p(u) \cdot \frac{1}{p(u)} \\ &= \log M. \end{aligned}$$

Entropy between pair of R.Vs

- Joint Entropy

$$H(X,Y) = -\sum_{x,y \in X,Y} p(x,y) log(p(x,y))$$

- Conditional Entropy

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

Relative Entropy aka Kullback Leibler Distance

D(p||q) is a measure of the inefficiency of assuming that the distribution is q, when the true distribution is p.

- H(p) : avg description length when true distribution.
- H(p) + D(p||q) : avg description length when approximated distribution.

If X is a random variable and p(x), q(x) are probability mass functions,

$$D(p||q) = \sum_{x \in X} p(x) log \frac{p(x)}{q(x)}$$

Relative Entropy/ K-L Divergence : Properties

$D(p||q)$ is a measure of the inefficiency of assuming that the distribution is q, when the true distribution is p.

Properties:

- Non-negative.
- $D(p||q) = 0$ if $p=q$.
- Non-symmetric and does not satisfy triangular inequality - it is rather divergence than distance.

Relative Entropy/ K-L Divergence : Properties

<u>Asymmetricity:</u>

Let, X = {0, 1} be a random variable. Consider two distributions p, q on X. Assume, p(0) = 1-r, p(1) = r ; q(0) = 1-s, q(1) = s;

$$D(p||q) = (1-r)log\frac{1-r}{1-s} + rlog\frac{r}{s}$$

$$D(q||p) = (1-s)log\frac{1-s}{1-r} + slog\frac{s}{r}$$

If, r=s, then D(p||q) = D(q||p) = 0, else for r!=s, D(p||q) != D(q||p)

Relative Entropy/ K-L Divergence : Properties

Non-negativity:

$$D(p||q) = \sum_{x \in X} p(x) log \frac{p(x)}{q(x)}$$

$$= - \sum_{x \in X} p(x) log \frac{q(x)}{p(x)}$$

$$= -\mathbb{E}[\mathrm{log(q|p)}]$$

$$>= -log(\mathbb{E}[\mathrm{q|p}])$$

$$= -log(\sum_{x} p(x). \frac{q(x)}{p(x)})$$

$$= -log(1) = 0$$

Relative Entropy/ K-L Divergence : Properties

For a PMF q define

$$H_q(U) \triangleq \mathbb{E}\left[\log \frac{1}{q(U)}\right] = \sum_{u \in \mathcal{U}} p(u) \log \frac{1}{q(u)}.$$

Then:

$$H(U) \leq H_q(U),$$

with equality iff $q = p$.

Proof:

$$H(U) - H_q(U) = \mathbb{E}\left[\log \frac{1}{p(u)}\right] - \mathbb{E}\left[\log \frac{1}{q(u)}\right]$$

$$\begin{aligned} H(U) - H_q(U) &= \mathbb{E}\left[\log \frac{q(u)}{p(u)}\right] \\ &\leq \log \mathrm{E}\left[\frac{q(u)}{p(u)}\right] \\ &= \log \sum_{u \in \mathcal{U}} p(u) \frac{q(u)}{p(u)} \\ &= \log \sum_{u \in \mathcal{U}} q(u) \\ &= \log 1 \\ &= 0 \end{aligned}$$

Thus,

$$H(U) - H_q(U) \leq 0.$$

Relative Entropy of joint distributions as Mutual Information

Mutual Information, which is a measure of the amount of information that one random variable contains about another random variable. It is the reduction in the uncertainty of one random variable due to the knowledge of the other.

$$I(X;Y) = D(p(x,y)||p(x)p(y)) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}$$

- Unlike Relative Entropy, Mutual Information is symmetric. And, it is non-negative.

Relationship between Entropy and Mutual Information

$$
\begin{aligned}
I(X;Y) &= \sum_{x,y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x,y} p(x,y) log \frac{p(x|y)}{p(x)} \\
&= -\sum_{x,y} p(x,y) log(p(x)) + \sum_{x,y} p(x,y) log(p(x|y)) \\
&= -\sum_{x} p(x) log(p(x)) - \{-\sum_{x,y} p(x,y) log(p(x|y))\} \\
&= H(X) - H(X|Y) = H(Y) - H(Y|X)
\end{aligned}
$$

Relationship between Entropy and Mutual Information

- I(X;X) = H(X) + H(X|X) = H(X)
  Mutual Information of a random variable with itself is the entropy of the random variable. This is the reason that entropy is sometimes referred to as **self-information**.

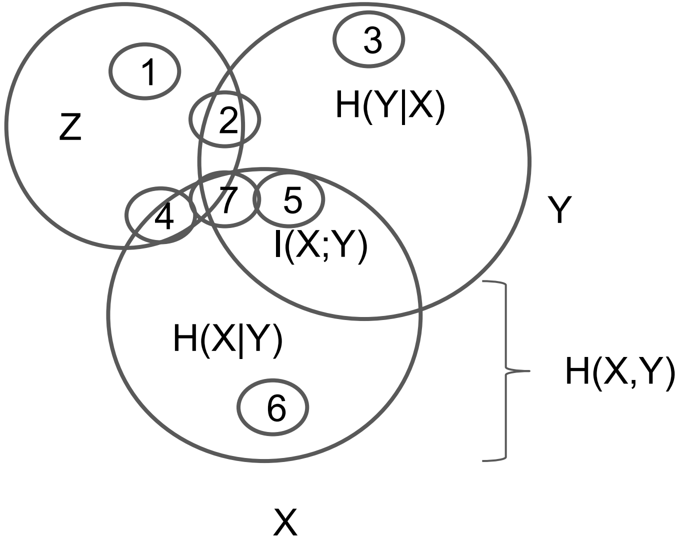
- Intuitively, the entropy of a random variable X with a probability distribution p(x) is related to how much p(x) diverges from the uniform distribution on the support of X. The more p(x) diverges the lesser its entropy and vice versa.

$$
\begin{aligned}
H(X) &= \sum_{x\in\mathcal{X}} p(x)\log\frac{1}{p(x)} \\
&= \log|\mathcal{X}| - \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{\frac{1}{|\mathcal{X}|}} \\
&= \log|\mathcal{X}| - D(p||uniform)
\end{aligned}
$$

Relationship between Entropy and Mutual Information

Conditioning reduces Entropy: H(X|Y) <= H(X)
as 0 <= I(X; Y) = H(X) - H(XIY).

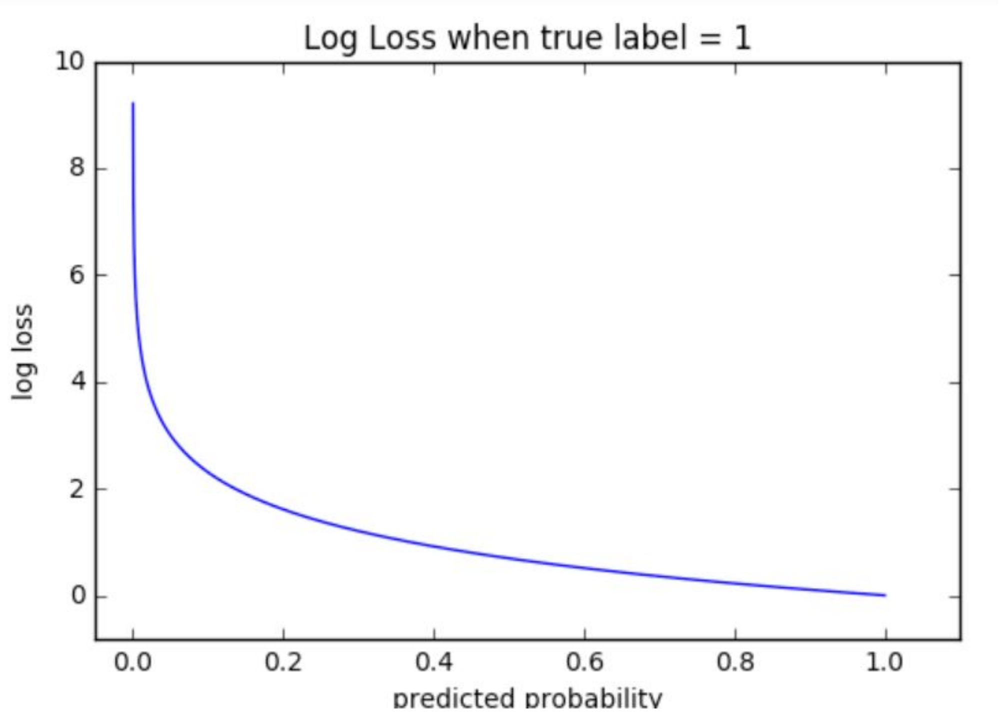
$$I(X;X) = H(X)$$
$$H(X,Y) = H(X) + H(Y|X)$$
$$H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$$
$$I(X;Y) = H(X) - H(X|Y)$$
$$I(X;Y) = H(Y) - H(Y|X)$$
$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$



Cross Entropy vs K-L Divergence

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

$$KL(P|Q) = \sum_x P(x)\log\frac{P(x)}{Q(x)}$$



Cross Entropy vs K-L Divergence

$$S(v) = -\sum_i p(v_i) \log p(v_i),$$

$$D_{KL}(A \parallel B) = \sum_i p_A(v_i) \log p_A(v_i) - p_A(v_i) \log p_B(v_i),$$

$$H(A, B) = -\sum_i p_A(v_i) \log p_B(v_i).$$

$$\boxed{H(A, B) = D_{KL}(A \parallel B) + S_A.}$$

Cross Entropy vs K-L Divergence

$$S(v) = -\sum_i p(v_i) \log p(v_i),$$

Entropy: A random variable has information about itself - self-informativeness.

True distribution

How B differs from A

$$D_{KL}(A \parallel B) = \sum_i p_A(v_i) \log p_A(v_i) - p_A(v_i) \log p_B(v_i) = \sum_i p_A(v_i)[log(p_A(v_i) - p_B(v_i)]$$

$$H(A, B) = -\sum_i p_A(v_i) \log p_B(v_i).$$

Cross-Entropy: A random variable compares true distribution A with approximated distribution B.

Cross-entropy = divergence + entropy
[A random variable knows about itself (**entropy**) and **from its perspective** compares its true distribution with approximated distribution **through divergence**]
Minimizing divergence and cross-entropy are said to have the same effects.

Relative-Entropy: A random variable compares true distribution A with how the approximated distribution B differs from A at each sample point (divergence or difference).

$$\boxed{H(A, B) = D_{KL}(A \parallel B) + S_A.}$$

Questions?

Thank You