Chapter 2: Entropy, Relative Entropy, and Mutual Information

Definition

The *entropy* $H(X)$ of a discrete random variable X is defined by

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

Properties of H

1. $H(X) \geq 0$.
2. $H_b(X) = (\log_b a) H_a(X)$.
3. (Conditioning Reduces Entropy) For any two random variables, X and Y, we have

$$H(X|Y) \leq H(X) \tag{2}$$

   with equality if and only if X and Y are independent.
4. $H(X_1, ..., X_n) \leq \sum_{i=1}^{n} H(X_i)$, with equality if and only if the X_i are independent.
5. $H(X) \leq \log(|\mathcal{X}|)$, with equality if X is distributed uniformly over $\mathcal{X}$.
6. $H(p)$ is concave in p.

Definition

The *relative entropy* $D(p||q)$ of the probabiltiy mass function p with respect to the probability mass function q is defined by

$$D(p||q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)} \tag{3}$$

Definition

The *mutual information* between two random variable X and Y is defined as

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{4}$$

Alternative Expressions

$$H(X) = E_p \log \frac{q}{p(X)}, \tag{5}$$

$$H(X,Y) = E_p \log \frac{1}{p(X,Y)}, \tag{6}$$

$$H(X|Y) = E_p \log \frac{1}{p(X|Y)} \tag{7}$$

$$I(X;Y) = E_p \log \frac{p(X,Y)}{p(X)p(Y)} \tag{8}$$

$$D(p||q) = E_p \log \frac{p(X)}{q(X)} \tag{9}$$

Properties of D and I

1. $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$.
2. $D(q||q) \geq 0$ with equality if and only if $p(x) = q(x)$, for all $x \in \mathcal{X}$.
3. $I(X;Y) = D(p(X,Y)||p(x)p(y)) \geq 0$, with equality if and only if $p(x,y) = p(x)p(y)$ (i.e., X and Y are independent)
4. If $|\mathcal{X}| = m$, and u is the uniform distribution over $\mathcal{X}$, then $D(p||u) = \log m - H(p)$.
5. $D(p||q)$ is convex in the pair (p,q).

Chain Rules

Entropy: $H(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} I(X_i; Y|X_1, X_2, ..., X_{i-1}$.
Mutual information: $I(X_1, X_2, ..., X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_1, X_2, ..., X_{i-1})$.
Relative entropy: $D(p(x, y)||q(x, y)) = D(p(x)||q(x)) + D(p(x)||q(x)) + D(p(y|x)||q(y|x))$.

Jensen's Inequality

If f is a convex function, then $Ef(X) \geq f(EX)$.

Log Sum Inequality

For n positive numbers, $a_1, a_2, ..., a_n$ and $b_1, b_2, ..., b_n$,

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq (\sum_{i=1}^{n} a_i) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}. \tag{10}$$

with equality if and only if $\frac{a_i}{b_i}$ = constant.

Data-processing Inequality

If $X \to Y \to Z$ forms a Markov chain, $I(X;Y) \geq I(X;Z)$.

Sufficient Statistic

$T(X)$ is sufficient relative to $\{f_\theta(x)\}$ if and only if $I(\theta; X) = I(\theta; T(X))$ for all distributions on θ.

Fano's Inequality

Let $P_e = Pr\{\hat{X}(Y) \neq X\}$. Then

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|Y). \tag{11}$$

Inequality

If X and X' are independent and identically distributed, then

$$Pr(X = X') \geq 2^{-H(x)}. \tag{12}$$