LUND UNIVERSITY

Information Theory

Lecture 2

Properties of information measures

STEFAN HÖST

Entropy

Definition (Entropy)

The Entropy is the average self information for a random variable,

$$H(X) = E\big[-\log p(X=x)\big] = -\sum_{x} p(x)\log p(x)$$

Theorem (Limit on the Entropy)

The entropy is a measure of the uncertainty of the outcome of the random variable X. It is bounded by

$$0 \le H(X) \le \log k$$

LUND UNIVERSITY

Entropy

Definition (Joint Entropy)

The joint entropy is the entropy for a pair of random variables with the joint distribution $p(x, y)$,

$$H(X, Y) = E_{XY}\big[-\log p(X, Y)\big] = -\sum_{x,y} p(x, y) \log p(x, y)$$

Definition (Conditional Entropy)

The conditional entropy is

$$H(X|Y) = E_{XY}\big[-\log p(X|Y)\big] = -\sum_{x,y} p(x, y) \log p(x|y)$$

Entropy

Theorem

A natural way to derive the conditional entropy is by

$$H(X|Y) = \sum_y H(X|Y = y)p(y)$$

where

$$H(X|Y = y) = -\sum_x p(x|y) \log p(x|y)$$

is the entropy of X, conditioned on the event $\{Y = y\}$.

Entropy

Theorem (Chain Rule)

The (first order) chain rule for entropies can be expressed as

$$H(X, Y) = H(X|Y) + H(Y) = H(Y|X) + H(X)$$

Theorem (Chain Rule)

Let $X_1, X_2, \ldots, X_n$ be an n-dimensional random variable drawn according to $p(x_1, x_2, \ldots, x_n)$. Then

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_1, \ldots, X_{i-1})$$

Mutual information

Definition

The mutual information between the random variables X and Y is defined as

$$I(X;Y) = E_{X,Y}\big[I(X=x;Y=y)\big]$$

$$= E_{X,Y}\Big[\log\frac{p(X|Y)}{p(X)}\Big] = \sum_{x,y} p(x,y)\log\frac{p(x|y)}{p(x)}$$

Mutual information

Alternative definition

Alternatively, the mutual information can be defined as the relation between the joint and the marginal probabilities,

$$I(X;Y) = E_{X,Y}\left[\log \frac{p(X,Y)}{p(X)p(Y)}\right] = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Mutual information

Properties of the mutual information

The mutual information between the random variables X and Y is symmetric,

$$I(X;Y) = I(Y;X)$$

It can be derived as

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned}$$

Conditioned mutual information

Conditioned mutual information

The mutual information between the random variables X and Y, conditioned on Z, is

$$I(X;Y|Z) = E_{X,Y,Z}\left[\log \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}\right] = \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}$$

It can be derived as

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

Relative Entropy

Definition

Given two probability distributions $p(x)$ and $q(x)$ for the same sample set $\mathcal{X}$. The relative entropy, or Kullback-Leibler divergence, is defined as

$$D(p||q) = E_p\left[\log \frac{p(X)}{q(X)}\right] = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

The relative entropy is not symmetric, i.e. in general $D(p||q) \neq D(q||p)$.

Information and Relative Entropy

Mutual information

The mutual information between X and Y can be seen as the relative entropy from the independent marginal distributions to the joint distribution

$$I(X;Y) = E_{p(x,y)}\Big[\log \frac{p(X,Y)}{p(X)p(Y)}\Big] = D\Big(p(x,y)\big|\big|p(x)p(y)\Big)$$

Entropy

Let $X \in \{x_1,\dots,x_k\}$ and $u(x) = 1/k$ be the uniform distribution. Then

$$H(X) = \log k - D(p||u)$$

Relative entropy

Theorem

Given two probability distributions $p(x)$ and $q(x)$ for the same sample set $\mathcal{X}$. Then the relative entropy is positive

$$D(p||q) \geq 0$$

with equality if and only if $p(x) = q(x)$ for all x.

Corollary

For any two random variables X and Y

$$I(X;Y) \geq 0$$

with equality if and only if they are independent.

Entropy

Theorem

For any two random variables X and Y

$$H(X|Y) \leq H(X)$$

with equality if and only if they are independent.

Theorem

Let $X_1, X_2, \ldots, X_n$ be an n-dimensional random variable. Then

$$H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$$

with equality if and only if all X_i are independent.

Jensen's inequality

Definition (Convex function)

A function $f(x)$ is convex in the interval $[a, b]$ if, for any $x_1, x_2 \in [a, b]$ and any λ, $0 \leq \lambda \leq 1$,

$$f\big(\lambda x_1 + (1-\lambda)x_2\big) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

A function $f(x)$ is concave in the interval $[a, b]$ if $-f(x)$ is convex in the same interval.

Example

The functions x^2 and e^x are convex.

The functions $-x^2$ and $\log x$ are concave.

Jensen's inequality

Theorem (Jensen's inequality)

If $f(x)$ is a convex function and X a random variable we have

$$E[f(X)] \geq f(E[X])$$

If $f(x)$ is a concave function and X a random variable we have

$$E[f(X)] \leq f(E[X])$$

Log-sum inequality

Theorem

Let $a_1, \dots, a_n$ and $b_1, \dots, b_n$ be non-negative numbers. Then

$$\sum_i a_i \log \frac{a_i}{b_i} \geq \Big(\sum_i a_i\Big) \log \frac{\sum_i a_i}{\sum_i b_i}$$

Covexity of information measures

Theorem (Relative entropy)

The relative entropy $D(p||q)$ is a convex function in (p, q), i.e.

$$D\Big(\lambda p_1 + (1-\lambda)p_2\Big|\Big|\lambda q_1 + (1-\lambda)q_2\Big) \leq \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2)$$

Covexity of information measures

Let $p_\lambda(x) = \lambda p_1(x) + (1-\lambda)p_2(x)$, $0 \leq \lambda \leq 1$.

Theorem (Entropy)

The entropy $H(X)$ is a concave function, i.e.

$$H_{p_\lambda}(X) \geq \lambda H_{p_1}(X) + (1-\lambda)H_{p_2}(X)$$

Covexity of information measures

Theorem (Mutual information)

The mutual information $I(X;Y)$ is

- *concave in $p(x)$ if $p(y|x)$ is fixed, i.e.*
$$I_{p_\lambda(x)}(X;Y) \geq \lambda I_{p_1(x)}(X;Y) + (1-\lambda) I_{p_2(x)}(X;Y)$$
- *convex in $p(y|x)$ if $p(x)$ is fixed, i.e.*
$$I_{p_\lambda(y|x)}(X;Y) \leq \lambda I_{p_1(y|x)}(X;Y) + (1-\lambda) I_{p_2(y|x)}(X;Y)$$