Lecture 3 More properties of entropy and mutual information

September 6th, 2022

Outline

Definition (Rényi entropy)

Given the parameter $\alpha > 0$ with $\alpha \neq 1$, and given a discrete random variable X with alphabet $\mathcal{X}$ and distribution P_X, its Rényi entropy of order α is given by

$$H_\alpha = \frac{1}{1-\alpha} \log(\sum_{x \in \mathcal{X}} P_X(x)^\alpha).$$

Definition (Rényi divergence)

Given a parameter $0 < \alpha < 1$, and two discrete random variables X and $\hat{X}$ with common alphabet $\mathcal{X}$ and distribution P_X and $P_{\hat{X}}$, respectively, then the Rényi divergence of order α between X and $\hat{X}$ is given by

$$D_\alpha(X\|\hat{X}) = \frac{1}{\alpha - 1} \log(\sum_{x\in\mathcal{X}} [P_X^\alpha(x) P_{\hat{X}}^{1-\alpha}(x)]).$$

This definition can be extended to $\alpha > 1$ if $P_{\hat{X}}(x) > 0$ for all $x \in \mathcal{X}$.

Lemma

When $\alpha \to 1$, we have the following:

$$\lim_{\alpha \to 1} H_\alpha(X) = H(X)$$

and

$$\lim_{\alpha \to 1} D_\alpha(X \| \hat{X}) = D(X \| \hat{X}).$$

Fundamental inequality

Lemma (Fundamental inequality (FI))

For any $x > 0$ and $D > 1$, we have that

$$\log_D(x) \leq \log_D(e) \cdot (x - 1),$$

with equality if and only if $x = 1$.

Setting $y = 1/x$ and using FI above directly that for any $y > 0$, we also have that

$$\log_D(y) \geq \log_D(e)(1 - \frac{1}{y}),$$

also with equality iff $y = 1$. In the above the base-D logarithm was used. Specifically, for a logarithm with base-2, the above inequalities become

$$\log_2(e)(1 - \frac{1}{x}) \leq \log_2(x) \leq \log_2(e) \cdot (x - 1),$$

with equality iff $x = 1$.

Information inequality

Theorem

Let X and $\hat{X}$ be two random variables, with probability mass functions P_X and $P_{\hat{X}}$. Then

$$D(X\|\hat{X}) \geq 0,$$

with equality if and only if $P_X(x) = P_{\hat{X}}(x)$ for all $x \in \mathcal{X}$, i.e., X and $\hat{X}$ have the same distribution.

Proof.

$$
\begin{aligned}
D(X\|\hat{X}) &= \sum_{x\in\mathcal{X}} P_X(x)\log_2\frac{P_X(x)}{P_{\hat{X}}(x)} \\
&\geq (\log_2 e)\sum_{x\in\mathcal{X}} P_X(x)(1-\frac{P_{\hat{X}(x)}}{P_X(x)}) \\
&= (\log_2 e)\sum_{x\in\mathcal{X}} P_X(x) - \sum_{x\in\mathcal{X}} P_{\hat{X}}(x) \\
&= 0,
\end{aligned}
$$

where the second step follows from FI, and the equality holds if and only if for every $x\in\mathcal{X}$,

$$\frac{P_X(x)}{P_{\hat{X}}(x)} = 1$$

for all $x\in\mathcal{X}$. □

Corollary

For any two random variables X, Y,

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

with equality if and only if X *and* Y *are independent.*

Corollary

$$D(p(y|x)\|q(y|x)) \geq 0,$$

with equality if and only if $p(y|x) = q(y|x)$ *for all* y *and* x *such that* $p(x) > 0$*.*

Corollary

$$I(X;Y|Z) \geq 0,$$

with equality if and only if X *and* Y *are conditionally independent given* Z*.*

Upper bound on entropy

Theorem

If a random variable X takes values from a finite set $\mathcal{X}$, then

$$H(X) \leq \log_2 |\mathcal{X}|,$$

where $|\mathcal{X}|$ is the size of the set $\mathcal{X}$. Equality holds if and only if X is equiprobable or uniformly distributed over $\mathcal{X}$ (i.e. $P_X(x) = \frac{1}{|\mathcal{X}|}$ for all $x \in \mathcal{X}$).

Proof.

$$
\begin{aligned}
& \log_2|\mathcal{X}| - H(X) \\
= & \sum_{x\in\mathcal{X}} P_X(x)\cdot\log_2|\mathcal{X}| + \sum_{x\in\mathcal{X}} P_X(x)\log_2 P_X(x) \\
= & \sum_{x\in\mathcal{X}} P_X(x)\cdot\log_2[|\mathcal{X}|\cdot P_X(x)] \\
\geq & \sum_{x\in\mathcal{X}} P_X(x)\cdot\log_2(e)(1-\frac{1}{|\mathcal{X}|\cdot P_X(x)}) \\
= & \log_2(e)\sum_{x\in\mathcal{X}}(P_X(x)-\frac{1}{|\mathcal{X}|}) \\
= & \log_2(e)(1-1)=0.
\end{aligned}
$$

with equality if and only if $|\mathcal{X}|\cdot P_X(x)=1$. □

- Intuitively, entropy tells us how random X is.

- Intuitively, entropy tells us how random X is.
- X is deterministic if and only if $H(X) = 0$.

- Intuitively, entropy tells us how random X is.
- X is deterministic if and only if $H(X) = 0$.
- If X is uniform (equiprobable), $H(X)$ is maximized and equal to $\log_2 |\mathcal{X}|$.

Theorem

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

with equality if and only if X *and* Y *are independent.*

Theorem

Let $X_1, X_2, \cdots, X_n$ be drawn according to $p(x_1, x_2, \cdots, x_n)$. Then

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

with equality if and only if the X_i are independent.

Theorem (Log-sum inequality)

For non-negative numbers $a_1, a_2, \cdots, a_n$ *and* $b_1, b_2, \cdots, b_n$*,*

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

with equality if and only if $\frac{a_i}{b_i} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$*, which is a constant that does not depend on* i*.*

Outline

Convex and concave function

Definition

A function $f(x)$ is said to be convex over an interval (a, b) if for every $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$,

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

A function f is said to be strictly convex if equality holds only if $\lambda = 0$ or $\lambda = 1$.

Convex and concave function

Definition

A function $f(x)$ is said to be convex over an interval (a, b) if for every $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$,

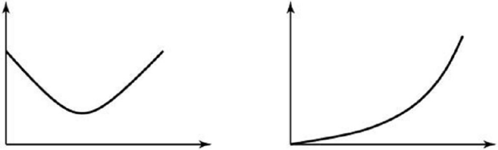
$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2).$$

A function f is said to be strictly convex if equality holds only if $\lambda = 0$ or $\lambda = 1$.

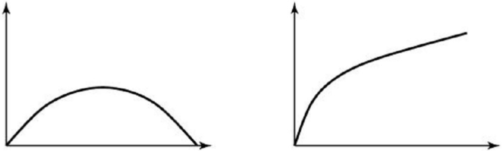
Definition

A function f is concave if $-f$ is convex.

A function is convex if it always lies below any chord. A function is concave if it always lies above chord.



(a)



(b)

Theorem

If the function f has a second derivative that is non-negative (positive) over an interval, the function is convex (strictly convex) over that interval.

Jensen's inequality

Theorem

If f is a convex function and X is a random variable,

$$Ef(x) \geq f(EX).$$

Moreover, if f is strictly convex, the above inequality implies that $X = EX$ with probability 1.

- All the inequalities in last section can be also proved using Jensen's inequality.
- Let f be a strictly convex function, $\alpha_i \geq 0$, and $\sum_{i=1}^{n} \alpha_i = 1$. Jensen's inequality states that

$$\sum_{i=1}^{n} \alpha_i f(t_i) \geq f(\sum_{i=1}^{n} \alpha_i t_i).$$

- Equality holds if and only if t_i is a constant for all i.
- To prove the log-sum inequality, set $\alpha_i = b_i / \sum_{j=1}^{n} b_j$, $t_i = a_i/b_i$, and $f(t) = t \cdot \log_D(t)$, we obtain the desired result.

Outline

Theorem

$H(P_X)$ is a concave function of P_X, namely

$$H(\lambda P_X + (1-\lambda)P_{\tilde{X}}) \geq \lambda H(P_X) + (1-\lambda)H(P_{\tilde{X}})$$

for all $\lambda \in [0,1]$.

Theorem

Noting that $I(X;Y)$ can be written as $I(P_X, P_{Y|X})$, where

$$I(P_X, P_{Y|X}) := \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_{Y|X}(y|x) P_X(x) \log_2 \frac{P_{Y|X}(y|x)}{\sum_{a\in\mathcal{X}} P_{Y|X}(y|a) P_X(a)},$$

then $I(X;Y)$ is a concave function of P_X (for fixed $P_{Y|X}$, and a convex function of $P_{Y|X}$ (for fixed P_X).

Theorem

$D(P_X \| P_{\hat{X}})$ is convex in pair $(P_X, P_{\hat{X}})$, i.e., if $(P_X, P_{\hat{X}})$ and $(Q_X, Q_{\hat{X}})$ are two pairs of probability mass functions, then

$$\begin{aligned} D(\lambda P_X + (1-\lambda) Q_X \| \lambda P_{\hat{X}} + (1-\lambda) Q_{\hat{X}}) \\ \leq \lambda \cdot D(P_X \| P_{\hat{X}}) + (1-\lambda) \cdot D(Q_X \| Q_{\hat{X}}), \end{aligned}$$

for all $\lambda \in [0, 1]$.