

Entropy

- measure of uncertainty of a random variable.

Notation: For a discrete r.v. X with alphabet (range) \mathcal{X} , denote prob mass function by $p(x) = \Pr\{X=x\}$, for $x \in \mathcal{X}$.

Def: Entropy of X $H(X) \triangleq - \sum_{x \in \mathcal{X}} p(x) \log p(x)$

Convention: \log means logarithm of base 2

\ln means natural logarithm

$0 \log 0 = 0$ (since $\lim_{x \rightarrow 0} x \log x = 0$)

Entropy measured in bits when \log used

Entropy " " nats when \ln used.

$H_b(X)$ denotes entropy with logarithm of base b

Props: (i) $H(X) = -E[\log p(X)] = E[\log \frac{1}{p(X)}]$

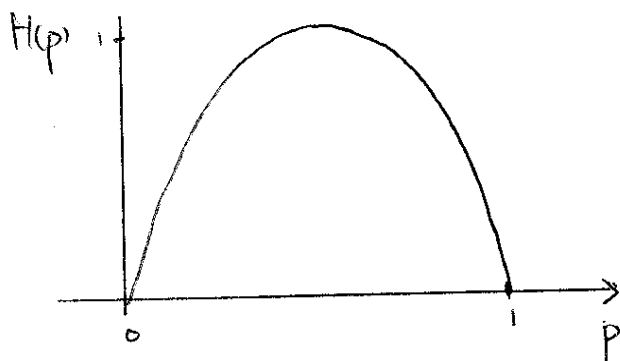
(ii) $H(X) \geq 0$ (Lemma 2.1.1)

(iii) $H_b(X) = (\log_b 2) H(X)$ (Lemma 2.1.2)

Example: $X = \begin{cases} 1 & \text{with prob. } p. \\ 0 & \text{with prob. } 1-p. \end{cases}$

(2)

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



$$\max H(p) = 1 \text{ at } p = \frac{1}{2}$$

Joint Entropy: (X, Y) r.v.'s with joint prob. mass fn. $p(x, y)$

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

* definition extends to more than 2 r.v.'s.

Conditional entropy: $(X, Y) \sim p(x, y)$

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

* def extends to conditioning on more than 1 r.v.'s

Note: $H(Y|X) \neq H(X|Y)$

Chain Rule (Thm 2.2.1)

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

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

Relative Entropy

- measures the difference ("distance") between 2 distributions

Def: Consider 2 prob. mass fns $p(x)$ & $q(x)$

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

Convention: $0 \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$

Remark: Relative entropy is not a true metric

Mutual Information

- measures amount of "information" that one r.v. contains about another r.v.

Def: Two discrete r.v.'s X & Y : $(X, Y) \sim p(x, y)$, $X \sim p(x)$, $Y \sim p(y)$

$$\begin{aligned} I(X; Y) &\triangleq \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\ &= E \log \frac{p(X, Y)}{p(X)p(Y)} \\ &= D(p(x, y) || p(x)p(y)) \end{aligned}$$

- mutual info is the relative entropy between the joint pmf & the product of the marginal pmf's.

* definition extends to more than 1 r.v.'s on both sides of ";

Relationships between entropy & mutual information:

④

(Thm 2.4.1)

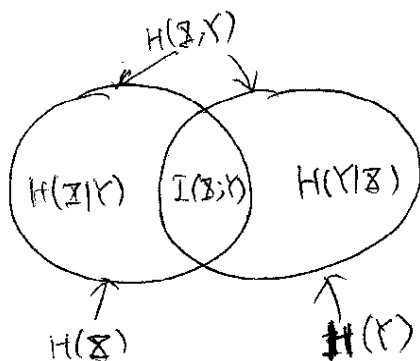
$$(i) \quad I(X; Y) = H(X) - H(X|Y)$$

$$(ii) \quad I(X; Y) = H(Y) - H(Y|X)$$

$$(iii) \quad I(X; Y) = H(X) + H(Y) - H(X, Y)$$

$$(iv) \quad I(X; Y) = I(Y; X)$$

$$(v) \quad I(X; X) = H(X)$$



(i) & (ii): mutual info is reduction of uncertainty of one r.v. due to knowledge of the other.

(v): entropy = self-information.

Ex: To show (v), we need to establish $H(X|X) = 0$.

Indeed,

$$H(X|X) = \sum_{x \in \mathcal{X}} p(x) H(X|X=x)$$

But the conditional pmf of X given $\{X=x\}$ is

$$p(x'|X=x) = \begin{cases} 1 & \text{if } x' = x \\ 0 & \text{if } x' \neq x \end{cases}$$

$$\begin{aligned} \text{Hence } H(X|X=x) &= - \sum_{x' \in \mathcal{X}} p(x'|X=x) \log p(x'|X=x) \\ &= 1 \cdot \log 1 = 0. \end{aligned}$$

Chain rules:

X_1, X_2, \dots, X_n discrete r.v.s $\sim p(x_1, x_2, \dots, x_n)$

Entropy (Thm 2.5.1)

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

Conditional mutual information b/w X and Y given Z

$$\begin{aligned} \text{Def: } I(X; Y | Z) &\triangleq H(X|Z) - H(X|Y, Z) \\ &= E \log \frac{p(X, Y | Z)}{p(X|Z)p(Y|Z)} \end{aligned}$$

(* definition extends to conditioning on more than 1 r.v.'s)

Mutual Information (Thm 2.5.2)

$$I(X_1, X_2, \dots, X_n; Y) = \sum_{i=1}^n I(X_i; Y | X_{i-1}, X_{i-2}, \dots, X_1)$$

Conditional relative entropy between $p(y|x)$ & $q(y|x)$

$$\begin{aligned} \text{Def: } D(p(y|x) || q(y|x)) &\triangleq \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log \frac{p(y|x)}{q(y|x)} \\ &= E \log \frac{p(Y|X)}{q(Y|X)} \end{aligned}$$

Relative entropy (Thm 2.3.3)

$$D(p(x, y) || q(x, y)) = D(p(x) || q(x)) + D(p(y|x) || q(y|x))$$

Properties of entropy, conditional entropy, & mutual information

(6)

Main math result employed here:

Jensen's inequality (Thm. 2.6.2)

If f is a convex function and X is a r.v. (cont./dis) then

$$E f(X) \geq f(E X).$$

Moreover if f is strictly convex, then equality above implies that $X = EX$ with prob. 1, i.e., X is a constant.

Remark: If f is concave, then $-f$ is convex.

So we can flip the inequality sign for concave fn.

Mutual Information ≥ 0 (Thm 2.6.3)

(i) Let $p(x)$ & $q(x)$, $x \in \mathcal{X}$ be two pmf's,

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

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

(ii) For 2 r.v.'s X & Y ,

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

with equality iff X & Y are independent.

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

with equality iff $p(y|x) = q(y|x)$ for all y & x

with $p(x) > 0$.

(7)

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

with equality iff X & Y are conditionally independent given Z .

Proof: Direct application of the Jensen's inequality, realizing that \log is a concave function!

Uniform distribution has maximum entropy: (Thm 2.6.4)

$$H(X) \leq \log |\mathcal{X}|$$

with equality iff X has a unif. distribution over \mathcal{X} .

Proof: Let $X \sim p(x)$ and $u(x) = \frac{1}{|\mathcal{X}|} \sim$ unif. over \mathcal{X}

$$0 \leq D(p \| u) = \sum_x p(x) \log \frac{p(x)}{u(x)} = \log |\mathcal{X}| - H(X)$$

Conditioning Reduces entropy: (Thm 2.6.5)

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

with equality iff X & Y are independent.

Proof: $0 \leq I(X; Y) = H(X) - H(X|Y)$.

Remark: $H(X|Y=y)$ can be greater than $H(X)$ (see example in text)

Independence bound on entropy. (Thm 2.6.6.)

(8)

Let $(X_1, X_2, \dots, X_n) \sim p(x_1, x_2, \dots, x_n)$, then

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

with equality iff X_i are independent.

Proof: Use chain rule of entropy (Thm 2.5.1) and conditioning reduces entropy.

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

Concavity of entropy & mutual information

Main math result: log sum inequality (Thm 2.7.1)

For non-negative numbers, a_1, a_2, \dots, a_n & b_1, b_2, \dots, b_n ,

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

Proof: Direct consequence (rather tricky) of Jensen inequality.

Convexity of relative entropy (Thm 2.7.2)

Let (p, q) be 2 pmfs.

Then $D(p||q)$ is convex in the pair (p, q) ,

i.e. if (p_1, q_1) & (p_2, q_2) are two pairs of pmfs,

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

for all $0 \leq \lambda \leq 1$.

Proof: Use log-sum inequality

Let $a_1 = \lambda p_1(x)$, $a_2 = (1-\lambda) p_2(x)$
 $b_1 = \lambda q_1(x)$, $b_2 = (1-\lambda) q_2(x)$

Then $[\lambda p_1(x) + (1-\lambda) p_2(x)] \log \frac{\lambda p_1(x) + (1-\lambda) p_2(x)}{\lambda q_1(x) + (1-\lambda) q_2(x)}$
 $\leq \lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda) p_2(x) \log \frac{(1-\lambda) p_2(x)}{(1-\lambda) q_2(x)}$

Summing over $x \in \mathcal{X}$ on both sides, we get the result.

Concavity of entropy (Thm 2.7.3)

$H(p)$ is a concave fn of p .

Proof: Recall $H(p) = \log |\mathcal{X}| - D(p||u)$, where u is the uniform distribution over \mathcal{X} .

Consider the pairs (p_1, u) & (p_2, u) , convexity of $D(\cdot||\cdot)$ implies that

$$D(\lambda p_1 + (1-\lambda) p_2 || u) \leq \lambda D(p_1 || u) + (1-\lambda) D(p_2 || u) \quad \text{for } 0 \leq \lambda \leq 1$$

$$\Rightarrow H(\lambda p_1 + (1-\lambda) p_2) \geq \lambda H(p_1) + (1-\lambda) H(p_2)$$

Concavity of mutual information (Thm 2.7.4)

Let $(\mathcal{X}, \mathcal{Y}) \sim p(x, y) = p(x) p(y|x)$.

- (i) $I(\mathcal{X}; \mathcal{Y})$ is a concave fn of $p(x)$ for fixed $p(y|x)$.
- (ii) $I(\mathcal{X}; \mathcal{Y})$ is a convex fn of $p(y|x)$ for fixed $p(x)$.

Proof: (i) $I(\mathcal{X}; \mathcal{Y}) = H(\mathcal{Y}) - H(\mathcal{Y}|\mathcal{X}) = H(\mathcal{Y}) - \sum_x p(x) H(\mathcal{Y}|\mathcal{X}=x)$

Fix $p(y|x)$. Then $p(y) = \sum_x p(x) p(y|x)$ is a linear fn of $p(x)$.

Since $H(\mathcal{Y})$ is concave of $p(y)$, $H(\mathcal{Y})$ is concave of $p(x)$
 $\sum_x p(x) H(\mathcal{Y}|\mathcal{X}=x)$ is a linear fn of $p(x) \Rightarrow$ it is concave of $p(x)$.

(10)

(ii) Fix $p(x)$ and consider $p_1(y|x)$ & $p_2(y|x)$.For $0 \leq \lambda \leq 1$, let $p_\lambda(y|x) = \lambda p_1(y|x) + (1-\lambda) p_2(y|x)$.

$$\begin{aligned} \text{Since } I(X; Y) &= D(p(x, y) \| p(x) p(y)) \\ &= D(p(x) p(y|x) \| p(x) \cdot \sum_x p(x) p(y|x)), \end{aligned}$$

suffices to show

$$\begin{aligned} &D(p(x) p_\lambda(y|x) \| p(x) \cdot \sum_x p(x) p_\lambda(y|x)) \\ &\leq \lambda D(p(x) p_1(y|x) \| p(x) \cdot \sum_x p(x) p_1(y|x)) + (1-\lambda) D(p(x) p_2(y|x) \| p(x) \cdot \sum_x p(x) p_2(y|x)) \end{aligned}$$

$$\text{Let } p_\lambda(x, y) \triangleq p(x) p_\lambda(y|x) = \lambda \underbrace{p(x) p_1(y|x)}_{p_1(x, y)} + (1-\lambda) \underbrace{p(x) p_2(y|x)}_{p_2(x, y)}$$

$$p_\lambda(y) \triangleq \sum_x p(x) p_\lambda(y|x) = \lambda \underbrace{\sum_x p(x) p_1(y|x)}_{p_1(y)} + (1-\lambda) \underbrace{\sum_x p(x) p_2(y|x)}_{p_2(y)}$$

Rewriting the inequality above,

$$D(p_\lambda(x, y) \| p(x) p_\lambda(y)) \leq \lambda D(p_1(x, y) \| p(x) p_1(y)) + (1-\lambda) D(p_2(x, y) \| p(x) p_2(y))$$

But this inequality is evident from the convexity of $D(\cdot \| \cdot)$.