Lesson 4
Entropy = Self information

4-5: Mutual Information

Jan Reimann

Math 574, Topics in Logic
Penn State, Spring 2014

Mutual Information

Question: How can we describe the mutual information between two random variables / strings?

Idea:

$$I(\sigma;\tau) = K(\sigma) + K(\tau) - K(\sigma,\tau)$$

Transfer to probabilistic setting:

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

Kullback-Leibler Divergence

Another way to gauge mutual information in the probabilistic setting is to ask how "far away" the joint distribution of two random variables is from a product distribution.

Independent random variables should share no mutual information.

Let P and Q be two probability distributions on a finite set A.

DEF: The Kullback-Leibler (KL-) divergence between P and Q is given as

$$D(P \parallel Q) = \sum_{a \in A} P(a) \log \frac{P(a)}{Q(a)} = \mathbb{E}_P \log \frac{P}{Q}$$

(We put $0 \log(0/0) = 0$ and $p \log(p/0) = \infty$ for $p > 0$.)

- This is not a metric (not even symmetric).
- Do we have at least $D(P \| Q) \geqslant 0$?

Mutual Information via KL-Divergence

Let X, Y be two A-valued random variables with distributions $P(x)$, $P(y)$, respectively, and joint distribution $P(x, y)$.

$$\boxed{I(X;Y) = D(\underline{P(x,y)} \,\|\, \underline{P(x)P(y)})}$$

$$= \sum_{a,b \in A} P(a,b) \log \frac{P(a,b)}{P(a)P(b)}$$

$$= \mathbb{E}_{P(x,y)} \log \frac{P(X,Y)}{P(X)P(Y)}.$$

Examples

- X, Y independent: $P(x,y) = P(x) \cdot P(y)$

$$D\big(P(x)\cdot P(y) \,\|\, P(x)\cdot P(y)\big) = 0$$

- $X = Y$: $$I(X;X) = \sum_{a \in A} P(a) \cdot \log \frac{P(a)}{P(a)^2}$$

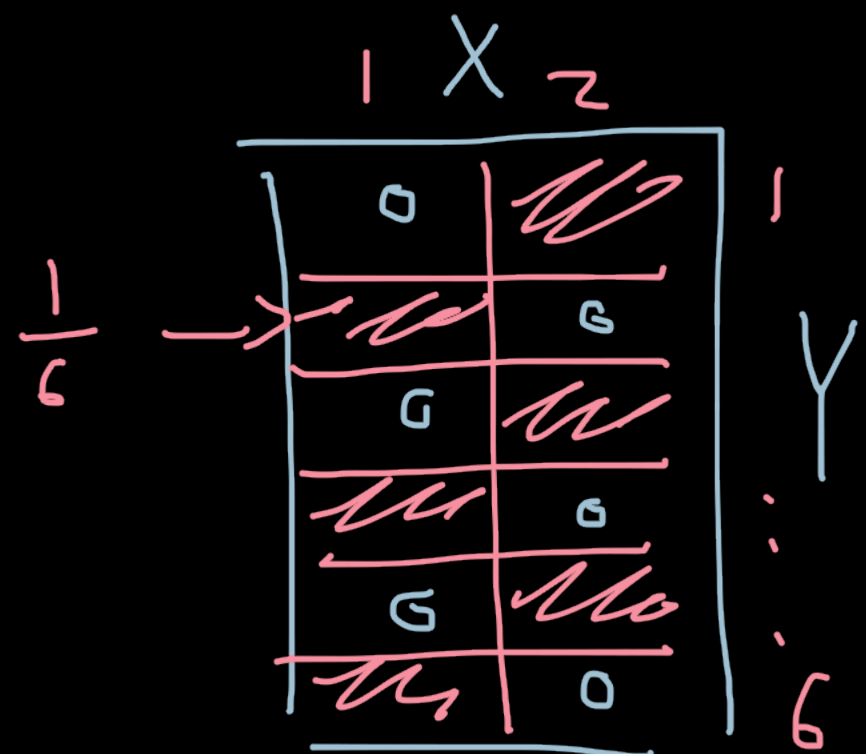
$$= \sum P(a) \cdot \log \frac{1}{P(a)}$$

$$= H(X)$$

Examples

- Y fair dice $\{1, 2, 3, 4, 5, 6\}$

X
$\hookrightarrow$ 1 if even
$\hookrightarrow$ 2 if odd



$$I(X;Y)$$

$$= \sum_{\substack{a=1,2\\ b=1,\dots,6}} P(a,b) \log \frac{P(a,b)}{P(a)\cdot P(b)}$$

$$= 6 \cdot \left(\frac{1}{6} \log \frac{\frac{1}{6}}{\frac{1}{12}} \right)$$

$$= \log 2 = H(X)$$

Entropy and Mutual Information

$$I(X;Y) = \sum_{a,b} P(a,b) \log \frac{P(a,b)}{P(a)P(b)} = \sum_{a,b} P(a,b) \log \frac{P(a|b)}{P(a)}$$

$$= -\sum_{a,b} P(a,b) \log P(a) + \sum_{a,b} P(a,b) \log P(a|b)$$

$$= -\sum_{a,b} P(a) \log P(a) - \left(-\sum_{a,b} P(a,b) \log P(a|b) \right)$$

$$= H(X) - H(X|Y).$$

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

COR: Symmetry of Information

$$\underline{I(X;Y)} = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

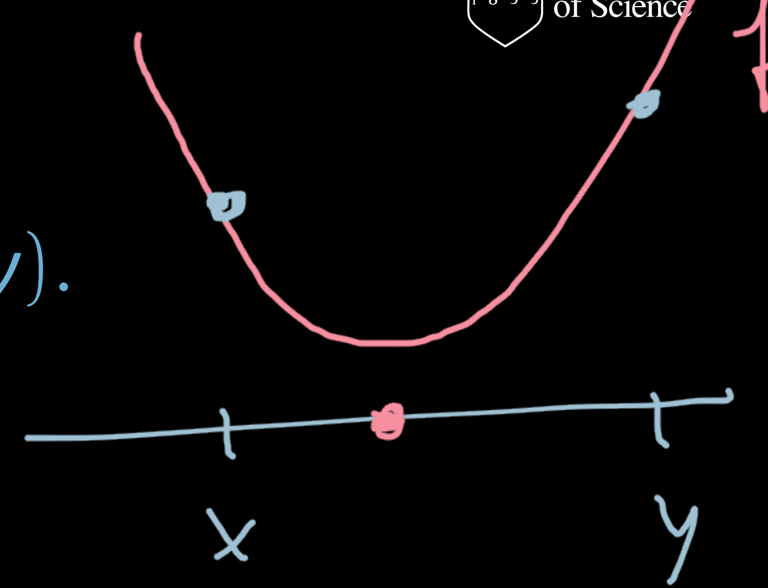
$$\geq 0 \, ?$$

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

entropy = self-information

In particular, $I(X;X) = H(X)$.

Jensen's Inequality



$f : \mathbb{R} \to \mathbb{R}$ is convex if for every $0 \leqslant \lambda \leqslant 1$,

$$f(\lambda x + (1 - \lambda)y) \leqslant \lambda f(x) + (1 - \lambda)f(y).$$

Concave: $-f$ convex.

Jensen's inequality: If f is a convex function and X is a random variable, then

$$\mathbb{E}f(X) \geqslant f(\mathbb{E}X).$$

Proof: For binary random variables (say $\{x_1, x_2\}$-valued), the inequality follows directly from the assumption of convexity:

$$\begin{aligned}\mathbb{E}f(X) &= P(x_1)f(x_1) + P(x_2)f(x_2) = P(x_1)f(x_1) + (1 - P(x_1))f(x_2) \\ &\geqslant f(P(x_1)x_1 + (1 - P(x_1))x_2) = f(\mathbb{E}X).\end{aligned}$$

Use induction to extend this to arbitrary finite-valued random variables.

(For continuous random variables, use continuity arguments.)

Information Inequality

$-\log$

We can use the fact that log is (strictly) convex to infer that

$$D(P \parallel Q) \geqslant 0,$$

with equality iff $P(a) = Q(a)$ for all $a \in A$.

Proof: Let $S_A = \{a \in A \colon P(a) > 0\}$ be the support of P.

$$-D(P \parallel Q) = -\sum_{a \in S_A} P(a) \log \frac{P(a)}{Q(a)} = \sum_{a \in S_A} P(a) \log \frac{Q(a)}{P(a)}$$

$$\leqslant \log \sum_{a \in S_A} P(a) \frac{Q(a)}{P(a)} = \log \sum_{a \in S_A} Q(a) \leqslant \log \sum_{a \in A} Q(a) = 0$$

COR: $I(X;Y) \geqslant 0$ with equality iff X and Y are independent.

Conditioning Reduces Entropy

Since $I(X;Y) = H(X) - H(X|Y) \geqslant 0$, we have

$$H(X|Y) \leqslant H(X).$$

On average, knowing another random variable Y reduces uncertainty in X.

$Y \backslash X$	1	2
1	0	3/4
2	1/8	1/8

$H(X) = H\left(1/8, 7/8\right) \approx 0.54$

$H(X|Y) = 0.25$

Question: Do we have analogues regarding Kolmogorov complexity, i.e. regarding the mutual information between two strings?

$\longrightarrow$ Coding