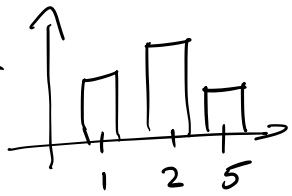
Review of Week 1

Chapter 1 : Information measures

rv X with values in $\{1, 2, 5\}$

$P_X(x) = p_x$



smaller probability
$\Rightarrow$ more surprised

1.1 Surprisal and Entropy

surprisal is a function of p_x, say $s(p)$

Properties we require:

1) Monotonicity: $s(1) = 0$,
$s(x)$ increases monotonically as p decreases

2) Additivity: $s(p_x p_y) = s(p_x) + s(p_y)$
↑ probability of observing $X = x$ and $Y = y$ for independent X and Y

3) Normalisation: $s(\frac{1}{2}) = 1$

$\Rightarrow s(p) = \log \frac{1}{p}$

Define $S(X) = s(P_X) = \log \frac{1}{P_x(X)}$ as the surprisal of the random experiment X governed by the pmf $P_x(x)$

The entropy of X is $H(X) = \mathbb{E}[S(X)]$.

We have

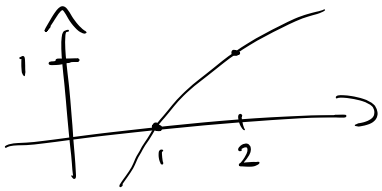
$$H(X) = \mathbb{E}\left[\log \frac{1}{P_X(X)}\right] = \sum_x P_X(x) \log \frac{1}{P_X(x)}$$

use convention that $0 \log 0 = 0$
$0 \log \infty = 0$

$$\lim_{\varepsilon \to 0} \varepsilon \log \frac{1}{\varepsilon} = \lim_{k \to \infty} \frac{\log k}{k} = 0$$

Prop.: $H(X) \geq 0$ with equality if and only if X is deterministic.

Proof: $\log \frac{1}{p} \geq 0$ for any $p \leq 1$



$H(X) = \sum p_x \log \frac{1}{p_x} \geq 0$ can be 0 only if $p_x \in \{0, 1\}$ for all x.

□

$H(X)$ is a concave function of the pmf.

Entropy is ...

$$f(t) = t \log \frac{1}{t} \longrightarrow H(X) = \sum_x f(p_x)$$

$$= -t \log t$$

$f(t)$ is strictly concave since

$$f'(t) = -\log t - \log e$$

$$f''(t) = -\frac{\log e}{t}$$

Prop: $H(X) \leq \log |\mathcal{X}|$ with equality iff X is uniform.

Proof for $\mathcal{X} = \{0, 1\}$:

Take pmf $\{p, 1-p\}$ for $p \in [0, 1]$. We write

$$H(X) = f(p) + f(1-p)$$

$$= \frac{1}{2}\left(f(p) + f(1-p)\right) + \frac{1}{2}\left(f(p) + f(1-p)\right)$$

Jensen's inequality $\leq f\left(\frac{1}{2}p + \frac{1}{2}(1-p)\right) \cdot 2$

$$= f\left(\frac{1}{2}\right) \cdot 2 \quad \leftarrow \text{recall } f(t) = -t \log t$$

$$= -\frac{1}{2} \log \frac{1}{2} \cdot 2 = 1$$

1.2 Conditional entropy and mutual information

Joint entropy

Two rvs X and Y with joint pmf $P_{XY}(x, y) = p_{xy}$

$$H(XY) = \mathbb{E}[S(X,Y)] = \sum_{(x,y)\in \mathcal{X}\times\mathcal{Y}} p_{xy} \log \frac{1}{p_{xy}}$$

$$= \sum_x \sum_y p_{xy} \log \frac{1}{p_{xy}}$$

Note: if X and Y are independent, then

$$H(XY) = \mathbb{E}[S(X,Y)] = \mathbb{E}[S(X) + S(Y)]$$

$$= \mathbb{E}[S(X)] + \mathbb{E}[S(Y)]$$

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

Example where $H(XY) = H(X) = H(Y)$?

$X = Y$, e.g. $p_{xy} = \begin{cases} \frac{1}{2} & x = y \\ 0 & x \neq y \end{cases}$

with $x, y \in \{0, 1\}$

$$p_{y|x} = \begin{cases} 1 & y = x \\ 0 & y \neq x \end{cases}$$

Conditional entropy of Y given X is defined as

$$H(Y|X) = \mathbb{E}\left[\log \frac{1}{p_{Y|X}}\right] = \sum_x \sum_y p_{xy} \log \frac{1}{p_{y|x}}$$

$$= \sum_x p_x \sum_y p_{y|x} \log \frac{1}{p_{y|x}}$$

$=: H(Y_x) = H(Y|X=x)$

$= \mathbb{E}[H(Y_x)]$

$\Rightarrow 0 \le H(Y|X) \le \log |\mathcal{Y}|$

Prop. (Chain rule) $H(XY) = H(X) + H(Y|X)$

Proof. $H(XY) = \sum_x \sum_y p_{xy} \log \frac{1}{p_{xy}}$ ← $p_{xy} = p_x p_{y|x}$

$= \sum_x \underbrace{\sum_y p_{xy}}_{p_x} \log \frac{1}{p_x} + \sum_x \sum_y p_{xy} \log \frac{1}{p_{y|x}}$

$\underbrace{\qquad\qquad}_{H(X)} \qquad\qquad \underbrace{\qquad\qquad}_{H(Y|X)}$

□

Often $H(Y|X) = H(XY) - H(X)$ is used as the definition of $H(Y|X)$!

Prop : $H(Y|X) \le H(Y) \iff H(XY) \le H(X) + H(Y)$

↑ sub-additivity of entropy

Proof: Only need to show

$H(Y|X) = \mathbb{E}\left[\sum_y p_{y|X} \log \frac{1}{p_{y|X}}\right]$

$H(Y|X)$

$f(t) = -t \log t$

$$= \sum_Y \mathbb{E}[f(p_{Y|X})]$$

Jensen's inequality $\leq \sum_Y f(\mathbb{E}[p_{Y|X}])$

$$= \sum_x p_x p_{Y|x} = p_Y$$

$$= \sum_Y f(p_Y) = H(Y)$$

Strong sub-additivity: $H(X|YZ) \leq H(X|Z)$

Mutual information

Define $I(X:Y) = H(X) + H(Y) - H(XY)$

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

from chain rule $\rightarrow$ $= H(X) - H(X|Y)$

Prop. $0 \leq I(X:Y) \leq \min\{\log|\mathcal{X}|, \log|\mathcal{Y}|\}$

Conditional mutual information

$$I(X:Y|Z) = \sum_z p_z \, I(X:Y|Z=z)$$

mutual information evaluated for p_{XYZ}

$$= H(Y|Z) - H(Y|XZ)$$
$$= H(X|Z) - H(X|YZ)$$

Chain rule: $I(X:YZ) = I(X:Y) + I(X:Z|Y)$

If $X-Z-Y$ form a Markov chain, then
$I(X:Y|Z) = 0$.

---

data-processing inequality

Prop.: Let $X-Y-Z$ form a Markov chain. Then $I(X:Y) \geq I(X:Z)$.

Proof: $I(X:Y) = I(X:YZ)$
due to chain rule and Markov condition

It remains to show $I(X:YZ) \geq I(X:Z)$

$$\Leftrightarrow H(X) - H(X|YZ) \geq H(X) - H(X|Z)$$

$$\Leftrightarrow H(X|Z) \geq H(X|YZ)$$

↑ strong sub-additivity

□

Relative entropy

Def. Let P and Q be two pmf on an alphabet $\mathcal{X}$.
The relative entropy of P with regards to Q is defined as

$$D(P\|Q) = \sum_{\substack{x\in\mathcal{X} \\ P(x)>0}} P(x) \log \frac{P(x)}{Q(x)}$$

if $P(x)>0 \Rightarrow Q(x)>0$ for all $x\in\mathcal{X}$, otherwise $D(P\|Q) = +\infty$.

We can see relative entropy as an expectation of the log-likelihood ratio

$$Z(x) = \log \frac{P(x)}{Q(x)}$$

$$D(P\|Q) = \mathbb{E}_{x\leftarrow P}[Z(x)].$$

Prop: $I(X:Y) = D(P_{XY} \| P_X \times P_Y)$

$$H(X) = \log|\mathcal{X}| - D(P_X \| U_X)$$

U_X: uniform distribution, $U_X(x) = \frac{1}{|\mathcal{X}|}$ for all x

$$H(X|Y) = \log|\mathcal{X}| - D(P_{XY} \| U_X \times P_Y)$$

Proof: Homework

Prop: $D(P\|Q) \geq 0$ with equality only if $P=Q$.

Proof: $g(x) = -\log x$ is strictly convex

$$D(P\|Q) = \sum_{\substack{x \\ P(x)>0}} P(x) \log \frac{P(x)}{Q(x)}$$

$$= \sum_{\substack{x \\ P(x)>0}} P(x)\, g\left(\frac{Q(x)}{P(x)}\right)$$

$$\geq g\left(\sum_{\substack{x \\ P(x)>0}} P(x) \frac{Q(x)}{P(x)}\right)$$

$$= g\left(\sum_{\substack{x \\ P(x)>0}} Q(x)\right)$$

$$\geq g\left(\sum_{x} Q(x)\right) = g(1) = 0 \qquad \square$$

Prop (Data-processing inequality)

Let P_X and Q_X be two pmfs on $\mathcal{X}$ (the input distributions). Let $P_{Y|X}$ be a conditional pmf (the channel / noise channel).

Define $P_Y(y) = \sum_{x \in \mathcal{X}} \underbrace{P_{Y|X}(y|x) P_X(x)}_{P_{XY}}$

$Q_Y(y) = \sum_{x \in \mathcal{X}} \underbrace{P_{Y|X}(y|x) Q_X(x)}_{Q_{XY}}$

(output distributions)

Thus, the data-processing inequality (DPI) states

$$D(P_X \| Q_X) \geq D(P_Y \| Q_Y)$$

Proof: $D(P_{XY} \| Q_{XY}) - D(P_X \| Q_X)$ ②

$$= \left(\sum_{x,y} P_{xy} \log \frac{P_{xy}}{q_{xy}} \right) - \left(\sum_x P_x \log \frac{P_x}{q_x} \right)$$

$$= \sum_{x,y} P_{xy} \left(\log \frac{P_{xy}}{q_{xy}} - \log \frac{P_x}{q_x} \right)$$

$$= \sum_x P_x \sum_y P_{y|x} \log \frac{P_{y|x}}{q_{y|x}}$$

$$= \sum_x P_x D(P_{Y|X=x} \| Q_{Y|X=x}) = 0 !$$

$D(P_{XY} \| Q_{XY}) - D(P_Y \| Q_Y)$ ①

$$= \sum_y P_Y D(P_{X|Y=y} \| Q_{X|Y=y}) \geq 0$$

① − ② ≥ 0

$\|$

$D(P_X \| Q_X) - D(P_Y \| Q_Y)$ □