III. Information Measures

- Shannon Entropy
- Differential Entropy
- Mutual Information

Shannon Entropy

Let $\mathcal{X}$ be a countable set $\underline{P} \in \mathcal{P}(\mathcal{X})$ w/ pmf p. The Shannon entropy of $X \sim P$ is:

$$H(X) = H(P) = H(p) := \mathbb{E}_P\left[\log \frac{1}{P(X)}\right]$$

Remark: $H(X) = H(P)$ is a measure for the uncertainty/unpredictability of $X \sim \underline{P}$.

Example: For $X \sim \underline{P}$:

$$H(\underline{P}) = -D_{KL}(P \,\|\, \#)$$

(# : counting measure)

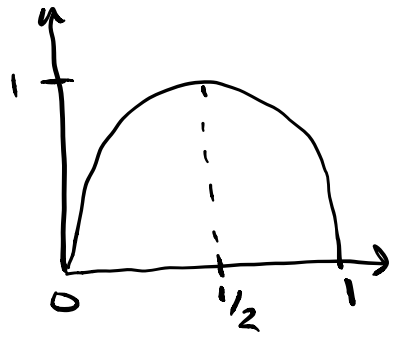
Furthermore, if $|\mathcal{X}| < \infty$, then:

$$H(\underline{P}) = \log|\mathcal{X}| - D_{KL}(P \,\|\, \mathrm{Unif}(\mathcal{X}))$$

Example - Bernoulli:

$X \sim Ber(\alpha)$, $\alpha \in [0,1]$

$$H(X) = H_b(\alpha) = \alpha \log \frac{1}{\alpha} + (1-\alpha) \log \frac{1}{1-\alpha}$$



Example - $X \sim P = Unif(\mathcal{X})$

$$H(P) = \log |\mathcal{X}|$$

Note: $H(P) = +\infty$ is possible! Why?

Proposition: properties of H. Let $X \sim P \in \mathcal{P}(\mathcal{X})$, $\mathcal{X}$ countable

1. Positivity: $H(X) \geq 0$ w/ equality iff X is const.

2. Uniform Upper Bound: If $|\mathcal{X}| < \infty$, then $H(P) \leq \log |\mathcal{X}|$

3. Invariance Under Relabeling: $H(X) = H(f(X))$ for any bijective f

4. Entropy of Functions: For any function f, we have $H(X) \geq H(f(X))$ w/ equality iff f is a bijection

5. Concavity: $P \mapsto H(P)$ is concave

Definition (Joint Entropy)

For $(X_1, \dots, X_n) \sim \mathbb{P}_{X_1,\dots,X_n} \in \mathcal{P}(\mathcal{X}_1 \times \cdots \times \mathcal{X}_n)$ with joint pmf $p_{X_1,\dots,X_n}$, the joint Shannon entropy is

$$H(X_1, \dots, X_n) := \mathbb{E}_{\mathbb{P}_{X_1,\dots,X_n}}\left[\log \frac{1}{p_{X_1,\dots,X_n}(X_1, \dots, X_n)}\right]$$

Definition (Conditional Entropy)

Let $(X, Y) \sim \mathbb{P}_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ with pmf p_{XY}. The conditional entropy of Y given X is

$$H(Y|X) := \mathbb{E}_{\mathbb{P}_{XY}}\left[\log \frac{1}{p_{Y|X}(Y|X)}\right]$$

Remark (Interpretation)

- Note that fixing $x \in \mathcal{X}$, $\mathbb{P}_{Y|X}(\cdot|x) \in \mathcal{P}(\mathcal{Y})$ so $H(P_{Y|X}(\cdot|x))$ is a regular entropy as given in the first definition.
- Now:

$$\begin{aligned} H(Y|X) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{XY}(x,y) \log \frac{1}{p_{Y|X=x}(y|x)} \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_X(x)\, p_{Y|X=x}(y|x) \log \frac{1}{p_{Y|X=x}(y|x)} \\ &= \sum_{x \in \mathcal{X}} p_X(x) \sum_{y \in \mathcal{Y}} p_{Y|X=x}(y|x) \log \frac{1}{p_{Y|X=x}(y|x)} \\ &= \sum_{x \in \mathcal{X}} p_X(x)\, H(Y|X=x) = \mathbb{E}_{x \sim P_X}\left[H(Y|X=x)\right] \end{aligned}$$

Proposition (properties continued)

5. Conditioning cannot increase entropy: $H(X) \geq H(X|Y)$
   w/ equality iff $X \perp\!\!\!\perp Y$

6. Entropy Chain Rule:

(i) Small:

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

(ii) Full:

$$H(X_1,\dots,X_n) = H(X_1) + \sum_{i=2}^{n} H(X_i | X_1,\dots,X_{i-1})$$

$$\leq \sum_{i=1}^{n} H(X_i)$$

w/ equality iff $X_1,\dots,X_n$ mutually independent.

Differentiable Entropy

Let $P \in \mathcal{P}(\mathbb{R}^d)$. The differentiable entropy of $X \sim P$ is:

$$h(X) = h(P) := -D_{KL}\left(P \,\|\, \text{Leb}(\mathbb{R}^d)\right).$$

In particular, if $P \ll \text{Leb}(\mathbb{R}^d)$ with pdf p, then:

$$h(X) = h(P) = h(p) = \mathbb{E}_P\left[\log \frac{1}{p(X)}\right] = \int_{\text{supp}(P)} p(x) \log \frac{1}{p(x)}\, dx$$

Examples

① $X \sim \mathbb{P} = \text{Unif}([0,a])$, $a > 0$

$$h(X) = \log(a)$$

② $X \sim \mathbb{P} = \mathcal{N}(\mu, \sigma^2)$

$$h(X) = \frac{1}{2}\log_2(2\pi e \sigma^2)$$

$$X \sim \mathbb{P} = \mathcal{N}(\vec{\mu}, \Sigma)$$

$$h(X) = \frac{1}{2}\log\left((2\pi e)^d \det(\Sigma)\right)$$

The definition of $h(X)$ extends to $h(X_1, \ldots, X_n)$ and $h(X|Y)$ similarly to what we saw for Shannon entropy.

Aside

could have

$$H(X|Y) = \mathbb{E}_{y \sim \mathbb{P}_Y}\left[\underbrace{H(X|Y=y)}_{\text{discrete}}\right] = \int \mathbb{P}_Y(y)\, H(X|Y=y)\, dy$$

also have

$$h(X|Y) = \mathbb{E}_{y \sim \mathbb{P}_Y}\left[\underbrace{h(X|Y=y)}_{\text{continuous}}\right] = \sum_{y \in \mathcal{Y}} \mathbb{P}_Y(y)\, h(X|Y=y)\, dy$$

Proposition (properties of h)

1. Uniform distribution maximizes h over a bounded domain

If $\text{supp}(\mathbb{P}) \subseteq \mathbb{R}^d$ is bounded, then

$$h(\mathbb{P}) \leq h(\text{Unif}(\text{supp}(\mathbb{P})))$$

2. Gaussian maximizes h under variance constraint

If $\text{supp}(P) = \mathbb{R}^d$ and $X \sim P$ with

$$\Sigma := \mathbb{E}\left[(X - \mathbb{E}[X])^T (X - \mathbb{E}[X])\right]$$

then

$$h(P) \leq h(\mathcal{N}(0, \Sigma))$$

3. Scaling and Shifting $X \sim P \in \mathcal{P}(\mathbb{R}^d)$

(i) $h(X + \mu) = h(X)$

(ii) $h(aX) = h(X) + d \log |a|$

(iii) $h(AX) = h(X) + \log(\det(A))$ for invertible A

4. Conditioning Cannot Increase Entropy

$$h(X) \geq h(X|Y)$$

w/ equality iff $X \perp\!\!\!\perp Y$

5. Chain Rule

$$h(X_1, \ldots, X_n) = h(X_1) + \sum_{i=2}^{n} h(X_i | X_1, \ldots, X_{i-1})$$

$$\leq \sum_{i=1}^{n} h(X_i)$$

w/ equality iff $(X_1, \ldots, X_n)$ mutually independent.

Remark (Warning): Note that h and H, while seemingly quantifying similar ideas, sometimes have vastly different behavior.

(i) While $H(P) \geq 0$, $h(P)$ can be posite/negative/unbounded

(ii) $H(X) \geq H(f(X))$ → not true for h

(iii) For H, we have

$$H(X+Y) \leq H(X+Y, Y)$$
$$= H(X,Y) \leq H(X) + H(Y)$$

BUT this doesn't hold for h