Math 254B Lecture 3 Notes

Daniel Raban

April 5, 2019

1 Properties of Shannon Entropy

1.1 Motivation and intuition

Recall that the Shannon entropy of p is

$$H(p) = -\sum_{x \in \mathcal{X}} p(x) \log(p(x)).$$

Here is a slogan to help you understand $H(p)$. $H(\alpha)$ is the canonical way to measure how "uncertain" α is.

Imagine an experiment modeled using $(\Omega, \mathcal{F}, \mathbb{P})$.

- Let $\Omega = E \cup E^c$, where $\mathbb{P}(E) = \mathbb{P}(E^c) = 1/2$. Here we have "1 bit of information" about the outcome.
- Let $\Omega = E_1 \cup \cdots \cup E_{2^m}$, where $\mathbb{P}(E_i) = 2^{-m}$. So E_i tells you "m bits of information."
- Localize: if $\mathbb{P}(E) = 2^{-m}$, then E conveys m bits of information.
- For all $E \subseteq \Omega$, if $\mathbb{P}(E) = 0$, we can interpolate to say the "information content" is $-\log_2(\mathbb{P}(E))$.

  We want to use natural logarithms, so instead of measuring information in "bits," we measure it in "nats."[1]
- $\alpha : \Omega \to \mathcal{X}$ partitions $\Omega = \{\alpha = x\}$, $x \in \mathcal{X}$. The expected information conveyed by α is $\sum_{x \in \mathcal{X}} \mathbb{P}(\alpha = x)[-\log \mathbb{P}(\alpha = x)]$

Lemma 1.1. *$H(\alpha) \geq 0$ with equality if and only if $p = \delta_x$ for some $x \in \mathcal{X}$.*

Proof. $-x \log(x)$ (with the convention that $-0 \log 0 = \lim_{x \to 0}[-x \log x] = 0$) is greater than 0 for $x \in (0, 1)$. □

Remark 1.1. $H(\alpha)$ is really a property of the partition $\Omega = \bigcup_x \{\alpha = x\}$.

[1] This word actually shows up in engineering textbooks.

1.2 Chain rule and conditional entropy

Suppose α, β are $\mathcal{X}, \mathcal{Y}$-valued respectively. Then we can regard (α, β) as a single $\mathcal{X} \times \mathcal{Y}$-valued random variable (also denoted sometimes by $\alpha \wedge \beta$).

Lemma 1.2 (chain rule)**.** *$H(\alpha, \beta) = H(\alpha) + H(\beta \mid \alpha)$, where*

$$H(\beta \mid \alpha) = \sum_{x \in \mathcal{X}} \mathbb{P}(\alpha = x) H_{\mathbb{P}(\cdot \mid \{\alpha = x\})}(\beta) = \sum_{x \in \mathcal{X}} p(x) H(q_x)$$

*is the **conditional entropy of** β **given** α. Here, p is the distribution of α, and $q_x(y) = \mathbb{P}(\beta = y \mid \alpha = x)$.*

Proof. Let $r(x, y) = \mathbb{P}(\alpha = x, \beta = y)$. Then

$$\begin{aligned} H(\alpha, \beta) &= -\sum_{(x,y)} r(x,y) \log(r(x,y)) \\ &= -\sum_x \sum_y p(x) q_x(y) \log(p(x) q_x(y)) \\ &= -\sum_x p(x) \log(p(x)) \underbrace{\sum_y q_x(y)}_{=1} - \sum_x p(x) \left(\sum_y q_x(y) \log(q_x(y)) \right). \end{aligned}$$

□

Lemma 1.3 (relative chain rule)**.** *Let α, β, ξ be random variables.*

$$H(\alpha, \beta \mid \xi) = H(\alpha \mid \xi) + H(\beta \mid \alpha, \xi).$$

Proof. Apply the chain rule with $\mathbb{P}(\cdot \mid \{\xi = z\})$. Take a weighted average. □

Lemma 1.4 (full chain rule)**.** *Let $\alpha_1, \ldots, \alpha_n, \xi$ be random variables.*

$$H(\alpha_1, \ldots, \alpha_n \mid \xi) = H(\alpha_1 \mid \xi) + H(\alpha_2 \mid \alpha_1, \xi) + \cdots + H(\alpha_n \mid \alpha_1, \ldots, \alpha_{n-1}, \xi)$$

Proof. Induction on n. □

1.3 Inequalities and mutual information

Lemma 1.5. *$H(\alpha, \beta) \geq H(\alpha)$, with equality if and only if there exists $\varphi : \mathcal{X} \to \mathcal{Y}$ such that $\beta = \varphi(\alpha)$ a.s.*

Proof. By the chain rule, $H(\alpha, \beta) - H(\alpha) = \sum_x p(x) H(q_x)$. This is 0 if and only if whenever $p(x) > 0$, we have $q_x = \delta_{\varphi(x)}$ for some $\varphi(x) \in \mathcal{Y}$. □

Corollary 1.1. *IF $\beta = \varphi(\alpha)$ a.s., then $H(\beta) \leq H(\alpha)$, and $H(\beta \mid \xi) \leq H(\alpha \mid \xi)$.*

Proof. $H(\alpha) = H(\alpha, \beta) \geq H(\beta)$. □

Proposition 1.1. *$H(\alpha) \geq H(\alpha \mid \beta)$, with equality if and only if α, β are independent.*

Remark 1.2. This says that conditioning reduces uncertainty. Similarly, $H(\alpha \mid \xi) \geq H(\alpha \mid \beta, \xi)$.

Lemma 1.6. *Fix $\mathcal{X}$ and consider $H : P(\mathcal{X}) \to [0, \infty)$. H is continuous and strictly concave.*

Proof. $H(p) = -\sum_x p(x) \log(p(x))$. Now apply known facts for each x. □

Proof. The law of total probability says that

$$q(y) = \mathbb{P}(\beta = y) = \sum_{x \in \mathcal{X}} \mathbb{P}(\{\beta = y\} \cap \{\alpha = x\}) = \sum_x p(x) q_x(y).$$

So by Jensen's inequality,

$$H(\beta) = H(q) = H(\sum_x p(x) q_x) \geq \sum_x p(x) H(q_x) = H(\beta \mid \alpha),$$

with equality if and only if $q_x = q$ whenever $p(x) > 0$; i.e. α, β are independent. □

Lemma 1.7. *Given $\mathcal{X}$, $H(p) \leq \log |\mathcal{X}|$ with equality if and only if $p(x) = 1/|\mathcal{X}|$ for all $x \in \mathcal{X}$.*

Definition 1.1. The **mutual information between α and β** is $I(\alpha; \beta) = H(\alpha) - H(\alpha \mid \beta) = H(\alpha + H(\beta) - H(\alpha, \beta)$.

The second formula follows from the first by the chain rule. Rearranging this, we get

$$H(\alpha) = H(\beta \mid \alpha) + U(\alpha, \beta).$$

Let α, β both be $\mathcal{X}$-valued.

Lemma 1.8 (Fano's inequality)**.** *Let $p = \mathbb{P}(\alpha \neq \beta)$ be the "probability of error." Then*

$$H(\alpha) - H(\beta) \leq H(\alpha \mid \beta) \leq H(p, 1-p) + p \log(|\mathcal{X}| - 1),$$

where $p, 1-p$ is a distribution on $\{0, 1\}$ with probability $p, 1-p$.