Lecture 2: Rényi Entropy

Anatomy of the paper:

1. Introduce $H_\alpha(P)$
2. "Motivate" $H_\alpha(P)$ via axioms/postulates
3. Rényi divergence
4. Applications to Markov Chains

Def$^{\underline{n}}$: (Rényi) Entropy of order α, where $\alpha > 0$ and $\alpha \neq 1$, is given by

$$H_\alpha(p_1, p_2, \ldots, p_n) = \frac{1}{1-\alpha} \log_2\left(\sum_{k=1}^{n} p_k^\alpha\right)$$
$$= H_\alpha(\mathcal{P})$$

← probability distribution

e.g. $p_k > 0$ & $\sum_{k=1}^{n} p_k = 1$

Special Cases (or Limiting cases):

$\alpha = 1$: Shannon Entropy

$$\lim_{\alpha \to 1} H_\alpha(p_1, p_2, \dots, p_n) = \sum_{k=1}^{n} p_k \log_2 \frac{1}{p_k}$$

$$= H(p_1, p_2, \dots, p_n) = H_1(\mathbb{P})$$

$\alpha = 0$: max-entropy or Hartley Entropy

$$H_0(p_1, p_2, .. p_n) = \log n$$

size of support set

$\alpha = \infty$: min-entropy

$$\lim_{\alpha \to \infty} H_\alpha(p_1, p_2, \dots, p_n) = \min_{k \in \{1, \dots, n\}} \log \frac{1}{p_k}$$

$$= H_\infty(p_1, p_2, \dots, p_n)$$

$\alpha = 2$: "collision" entropy

$$H_2(p_1, p_2, \dots, p_n) = -\log_2\left(\sum_{k=1}^{n} p_k^2\right)$$

$$= \log_2 \frac{1}{\mathbb{P}[X_1 = X_2]}$$

where X_1, X_2 are ind & $X_1, X_2 \sim P$

Cases:

- $\mathcal{P}$ is equiprobable on $\{1, \dots, n\}$

$$H_\alpha(\mathcal{P}) = \log n$$

- $\mathcal{P}$ is deterministic

$$\mathbb{P}[X = 1] = 1$$

$$H_\alpha(\mathcal{P}) = 0$$

Axiomatic Perspective

Postulates for Shannon Entropy (due to Fadeev)

(a) $H(p_1, p_2, \dots, p_n)$ is a symmetric func. of variables for $n = 2, 3, \dots$

(b) $H(p, 1-p)$ is continuous function of p for $0 \le p \le 1$.

(c) $H(1/2, 1/2) = 1$

(d) $H(tp_1, (1-t)p_1, p_2, \dots, p_n)$

$= H(p_1, p_2, \dots, p_n) + p_1 H(t, 1-t)$

for any distribution $\mathcal{P}$ and $0 \le t \le 1$

Chain rule

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

- Consider changing (d) to

(d') $H(\mathcal{P} * Q) = H(P) + H(Q)$

direct product (product distribution)

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

if X and Y are independent

Claim: $H_\alpha(\cdot)$ satisfies (a), (b), (c), (d'). (check).

$$H_\alpha(X, Y) = H_\alpha(X) + H_\alpha(Y)$$

if X, Y are independent.

2. Motivating $H_\alpha(\mathcal{P})$

- "Generalized distributions"

$\mathcal{P} = (p_1, p_2, \dots, p_n)$ s.t. $0 \le \sum_{k=1}^{n} p_k \le 1$ and $p_k \ge 0$.

Consider this setting for technical reasons.

- Mean values: "information" in observation $k \in \{1, \dots, n\}$ it is given by $i_p(k) = \log \frac{1}{p_k}$ where $\mathbb{P}[X=k] = p_k$

- Shannon Entropy: arithmetic mean of information

$$H_1(\mathcal{P}) = \sum_{k=1}^{n} p_k \log \frac{1}{p_k} = \sum_{k=1}^{n} p_k \, i_p(k)$$

- Kolmogorov-Nagumo mean

$$g^{-1}\left(\sum_{k=1}^{n} w_k \, g(x_k) \right)$$

g ← strictly monotonic and continuous

x_k ← data points

Examples:

- $g(x) = ax + b$ arithmetic mean

- $g(x) = \frac{1}{x}$ gives harmonic mean

- $g(x) = x^2$ quadratic mean

• Rényi Entropy: g-mean with

$$g_\alpha(x) = 2^{(\alpha-1)x}$$

Thus, it is a different way to find average information in an information source.

Postulates

P1: $\mathcal{H}(\mathcal{P})$ is a symmetric function of all the elements

same as in sec 1

P2: $H(\{p\})$ is a continuous function of p on the interval $0<p\leq 1$

P3: $H(\{1/2\}) = 1$

P4: For $P, Q \in \Delta$ we have

$$H(P * Q) = H(P) + H(Q)$$

Entropy for independent events is additive

(d') in Sec 1.

P5: For $P, Q \in \Delta$ and $w(P) + w(Q) \leq 1$

weight

$$H(P \cup Q) = \frac{w(P)H(\mathcal{P}) + w(Q)H(Q)}{w(P) + w(Q)}$$

- Thm 1: Shannon Entropy is characterized by $P_1 - P_5$.

$$P5': \; H(P \cup Q) = g^{-1}\left[\frac{w(P)\, g(H(P)) + w(Q)\, g(H(Q))}{w(P) + w(Q)}\right]$$

- Thm 2: $H_\alpha(P)$ satisfies $P_1 - P_4$, P_5'

- Main takeaway:

this is a motivation for $H_\alpha(P)$ being a "good" measure of information

Sec. 3 – Rényi divergence of order α.

Defn: Rényi divergence of order α, $\alpha > 0$ & $\alpha \neq 1$ is given by

$$D_\alpha(Q \| P) = I_\alpha(Q | P)$$
$$= \frac{1}{\alpha - 1} \log_2 \left(\sum_{k=1}^{n} q_k^\alpha \, p_k^{1-\alpha} \right)$$

– As before

KL-divergence

$$\lim_{\alpha \to 1} D_\alpha(Q \| P) = D(Q \| P)$$

– KL Divergence / relative entropy is an arithmetic mean of $i_{Q\|P}(k) = \log \frac{q_k}{p_k}$

– Rényi divergence is the g-mean of $i_{Q\|P}(k)$

– thm 3 characterize $D(Q\|P)$ an $D_\alpha(Q\|P)$