Notes about classical information theory in Nielson

Jake

2020/9/26

Shannon entropy:

1. Shannon entropy of X quantifies how much information we gain, on average, after we learn the value of X.

2. Shannon entropy measures the amount of uncertainty about X before we learn its value.

   Shannon entropy can be used to reformulating the uncertainty principle of quantum mechanics.

Shannon entropy associated with the probability distribution:

$$H(X) = H(p_1, p_2 \cdots p_n) = -\sum_x p_x \log p_x$$

Justification for this definition of log:

The information gain when two independent events occur with individual probabilities p and q is the sum of the information gained from each event alone.

binary entropy: the entropy of a two-outcome random variable

$$H_{\text{bin}}(p) \equiv -p \log p - (1-p) \log(1-p)$$

Many of the deepest results in quantum information have their roots in skilful application of concavity properties of classical or quantum entropies.

$$H_{\text{bin}}(px_1 + (1-p)x_2) \geq H_{\text{bin}}(px_1) + H_{\text{bin}}((1-p)x)$$

relative entropy: entropy-like measure of closeness of two probability distributions, $p(x)$ and $q(x)$.

Relative entropy of $p(x)$ to $q(x)$:

$$\begin{aligned} H(p(x)||q(x)) &\equiv \sum_x p(x) \log \frac{p(x)}{q(x)} \equiv \sum_x p(x)[\log p(x) - \log q(x)] \\ &= -H(X) - \sum_x p(x) \log q(x) \end{aligned}$$

relative entropy is useful because other entropic quantities can be regarded as special cases of the relative entropy.

1. Non-negativity of the relative entropy: $H(p(x)||q(x)) \geq 0$, with equality if and only if $p(x) = q(x), \forall x$

   Proof:

$$-\log_a x \geq \frac{1-x}{\ln a} \quad \forall a > 1, \forall x > 0 \text{ with equality } x = 1$$

$$\begin{aligned}
&-\log_2 x \geq \frac{1-x}{\ln 2} \implies \\
&H(p(x)||q(x)) \equiv \sum_x p(x) \log \frac{p(x)}{q(x)} \\
&\geq \frac{1}{\ln 2} \sum_x p(x)(1 - \frac{p(x)}{q(x)}) \\
&= \frac{1}{\ln 2} \cdot 0
\end{aligned}$$

2. Non-negativity of the relative entropy $\implies$ $H(X) \leq \log d$, with equality X is uniformly distributed over d outcomes.

   Proof:

   Let $q(x) \equiv \frac{1}{d}$. Then

$$H(p(x)||q(x)) \equiv -H(X) - \sum_x p(x) \log q(x) \geq 0$$

3. $H(p(x,y)||p(x)p(y)) = H(p(x)) + H(p(y)) - H(p(x,y)) \implies$ Subadditivity of the Shannon entropy $H(X,Y) \leq H(X) + H(Y)$, with equality X and Y are independent random variables.

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) \mathrm{d}y \implies \sum_{x,y} p(x,y) \log p(x) = \sum_x p(x) \log p(x)$$

$$\begin{aligned}
H(p(x,y)||p(x)p(y)) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x,y} p(x,y) \log p(x,y) - \sum_{x,y} p(x,y) \log(p(x)p(y)) \\
&= \sum_{x,y} p(x,y) \log p(x,y) - \sum_{x,y} p(x,y) \log p(x) - \sum_{x,y} p(x,y) \log p(y) \\
&= -H(p(x,y)) + H(p(x)) + H(p(y))
\end{aligned}$$

Conditional entropy and mutual information:

Conditional entropy and mutual information: how is the information content of X related to the information content of Y?

1. joint entropy: $H(X,Y) = -\sum_{x,y} p(x,y) \log p(x,y)$.

$$H(X,Y) = H(p(x)) + H(p(y)) - H(p(x,y)||p(x)p(y))$$

   It is pretty straightforward with the definition of $H(X)$

   measure our total uncertainty about the pair (X,Y)

2. entropy of X conditional on knowing Y: $H(X|Y) = H(X,Y) - H(Y)$.

$$H(X|Y) = H(p(x)) - H(p(x,y)||p(x)p(y)) \quad H(Y|X) = H(p(y)) - H(p(x,y)||p(x)p(y))$$

   defined with joint entropy and entropy

$$\begin{aligned} H(X,Y) &= -\sum_{x,y} p(x,y) \log p(x,y) = -\sum_{x,y} p(x,y) \log p(x)p(y|x) \\ &= -\sum_{x,y} p(x,y) \log p(x) - \sum_{x,y} p(x,y) \log p(y|x) \\ &= -\sum_{x} p(x) \log p(x) - -\sum_{x,y} p(x,y) \log p(y|x) \\ &= H(X) - \sum_{x,y} p(x,y) \log p(y|x) \end{aligned}$$

$$\implies H(Y|X) = -\sum_{x,y} p(x,y) \log p(y|x)$$

   how uncertain we are about X, given that we know Y.

3. mutual information content of X and Y: $H(X:Y) = H(X) + H(Y) - H(X,Y)$.

$$H(X:Y) = H(p(x,y)||p(x)p(y))$$

   defined with joint entropy and entropy

   measure how much information X and Y have in common.

basic properties of Shannon entropy apart from the definition

1. symmetry of joint entropy: $H(X,Y) = H(Y,X)$
2. symmetry of mutual information: $H(X:Y) = H(Y:X)$

3. Non-negativity of conditional entropy: $H(Y|X) \geq 0$, with equality Y is a deterministic function of X.

$$\begin{cases} H(Y|X) = -\sum\limits_{x,y} p(x,y) \log p(y|x) \\ \log p(y|x) < 0 \end{cases} \implies H(Y|X) \geq 0$$

4. mutual information relationship with conditional entropy and entropy without condition: $H(X:Y) = H(X) - H(X|Y)$.

$$\begin{cases} H(X:Y) = H(X) + H(Y) - H(X,Y) \\ H(X|Y) = H(X,Y) - H(Y) \end{cases} \implies H(X:Y) = H(X) - H(X|Y)$$

5. mutual information less than entropy: $H(X:Y) \leq H(Y)$, $H(X:Y) \leq H(X)$, with equality Y is a deterministic function of X. ($H(X:Y) \leq H(X)$, with equality ...)

$$\begin{cases} H(X|Y) \geq 0 \\ H(X:Y) = H(X) - H(X|Y) \end{cases} \implies H(X:Y) \leq H(Y)$$

$$\begin{cases} H(Y|X) \geq 0 \\ H(Y:X) = H(Y) - H(Y|X) \end{cases} \implies H(X:Y) = H(Y:X) \leq H(Y)$$

6. entropy less than joint entropy: $H(X) \leq H(X,Y)$, $H(Y) \leq H(X,Y)$, with equality Y is a deterministic function of X. ($H(Y) \leq H(X,Y)$, with equality...)

$$\begin{cases} H(X:Y) \leq H(Y) \\ H(X,Y) = H(X) + H(Y) - H(X:Y) \end{cases} \implies H(X) \leq H(X,Y)$$

$$\begin{cases} H(X:Y) \leq H(X) \\ H(X,Y) = H(X) + H(Y) - H(X:Y) \end{cases} \implies H(Y) \leq H(X,Y)$$

$$\begin{cases} H(Y|X) \geq 0 \\ H(X,Y) = H(X) + H(Y|X) \end{cases} \implies H(X) \leq H(X,Y)$$

$$\begin{cases} H(X|Y) \geq 0 \\ H(X,Y) = H(Y) + H(X|Y) \end{cases} \implies H(Y) \leq H(X,Y)$$

7. subadditivity of entropy (joint entropy less than sum of entropy): $H(X,Y) \leq H(X) + H(Y)$, with equality X and Y are independent.

   We have proven this before. The following paragragh is just a more straightforward way to prove it. But actually, we can just need to rearrange the proof-needing expression and make it a relative entropy and then use the non-negativity of relative entropy, which is what we did at the first proof.

$$\iint_{-\infty}^{\infty} f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y = \iint_{-\infty}^{\infty} f_X(x) \cdot f_Y(y)\mathrm{d}x\mathrm{d}y$$

$$\begin{aligned} H(p(x,y)||p(x)p(y)) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &\leq \frac{1}{\ln 2} \sum_{x,y} p(x,y) \left(\frac{p(x,y)}{p(x)p(y)} - 1 \right) \\ &= \frac{1}{\ln 2} \sum_{x,y} (p(x)p(y) - p(x,y)) = 0 \end{aligned}$$

8. conditional entropy less than entropy without condition: $H(Y|X) \leq H(Y)$, with equality X and Y are independent.

$$\begin{cases} H(X,Y) \leq H(X) + H(Y) \\ H(X,Y) = H(X) + H(X|Y) \end{cases} \implies H(Y|X) \leq H(Y)$$

9. Non-negativity of mutual information: $H(X:Y) \geq 0$, with equality X and Y are independent.

$$\begin{cases} H(Y|X) \leq H(Y) \\ H(X:Y) = H(X) - H(X|Y) \end{cases} \implies H(X:Y) \geq 0$$

10. Strong subadditivity: $H(X,Y,Z)+H(Y) \leq H(X,Y)+H(Y,Z)$, $H(X,Y,Z)+H(Z) \leq H(X,Z)+H(Z,Y)$, $H(X,Y,Z)+H(X) \leq H(Y,X)+H(Z,X)$with equality $Z \to Y \to X$ forms a Markov chain. ($H(X,Y,Z) + H(Z) \leq H(X,Z) + H(Z,Y)$, $H(X,Y,Z) + H(X) \leq H(Y,X) + H(Z,X)$. with equality...)

    Rearrange this expression to make it a relative entropy, which is

$$H(p(x,y,z)||\frac{p(x,y)p(y,z)}{p(y)}) \geq 0$$

    So relative entropy is really a useful tool.

11. Chaining rule for conditional entropies: $H(X_1, X_2, \cdots X_n|Y) = \sum_{i=1}^{n} H(X_i|Y, X_1, \cdots, X_{i-1})$.

$$
\begin{aligned}
H(X_1, X_2, \cdots X_n|Y) &= \sum_{i=1}^{n} H(X_i|Y, X_1, \cdots, X_{i-1}) \\
&= H(X_1|Y) + H(X_2|Y, X_1) + H(X_3|Y, X_1, X_2) + \cdots + H(X_n|Y, X_1, \cdots, X_{n-1})
\end{aligned}
$$

Proof: prove the result for $n = 2$, and then induct on n.

$$
\begin{aligned}
H(X_1, X_2|Y) &= H(X_1, X_2, Y) - H(Y) \quad \text{construct } H(X_2|Y, X_1) \\
&= H(X_1, X_2, Y) - H(X_1, Y) + H(X_1, Y) - H(Y) \\
&= H(X_2|Y, X_1) + H(X_1|Y)
\end{aligned}
$$

$$
\begin{cases}
H(X_1, \cdots, X_{n+1}|Y) = H(X_2, \cdots, X_{n+1}|Y, X_1) + H(X_1|Y) \\
H(X_2, \cdots, X_{n+1}|Y, X_1) = \sum_{i=2}^{n} H(X_i|Y, X_1, \cdots, X_{i-1})
\end{cases}
$$

$$
\implies H(X_1, X_2, \cdots X_n|Y) = \sum_{i=1}^{n} H(X_i|Y, X_1, \cdots, X_{i-1})
$$

12. non-subadditivity of mutual information: $H(X, Y : Z) \nleq H(X : Z) + H(Y : Z)$.

    $Z = X \oplus Y$. X, Y independent. $p_0 = 0.5, p_1 = 0.5$

13. non-superadditivity of mutual information: $H(X_1 : Y_1) + H(X_2 : Y_2) \nleq H(X_1, X_2 : Y_1, Y_2)$.

    $X_2 = X_1 = Y_1 = Y_2$, $p_0 = 0.5, p_1 = 0.5$

Markov process and data processing ineqaulity

Markov process: $p(X_{n+1} = x_{n+1}|X_n = x_n, \cdots, X_1 = x_1) = p(X_{n+1} = x_{n+1}|X_n = x_n)$.

Data processing inequality:

$$
X \to Y \to Z \text{ is a Markov process} \implies H(X) \geq H(X : Y) \geq H(X : Z)
$$

$H(X) \geq H(X : Y)$ is the 'mutual information less than entropy'.

$$
\begin{aligned}
H(X : Y) \geq H(X : Z) &\Leftarrow H(X) - H(X|Y) \geq H(X) - H(X|Z) \\
&\Leftarrow \begin{cases} H(X|Z) \geq H(X|Y) \\ H(X|Y) = H(X|Y, Z) \end{cases} \quad \Leftarrow H(X|Z) \geq H(X|Y, Z) \\
&\begin{cases} H(X|Z) \geq H(X|Y, Z) \\ H(X|Z) = H(X, Z) - H(Z) \\ H(X|Y, Z) = H(X, Y, Z) - H(Y, Z) \end{cases} \quad \Leftarrow H(X, Y, Z) - H(Y, Z) \leq H(X, Z) - H(Z) \\
&\Leftarrow H(X, Y, Z) + H(Z) \leq H(X, Z) + H(Y, Z)
\end{aligned}
$$

But it seems the second one can't make it to the equality.
Data pipelining inequality:

$$X \to Y \to Z \text{ is a Markov process } \implies H(Z:Y) \geq H(Z:X)$$

$$X \to Y \to Z \text{ is a Markov process } \implies Z \to Y \to X \text{ is a Markov process}$$

$$\begin{aligned} p(X_{n+1} = x_{n+1} | X_n = x_n, \cdots, X_1 = x_1) &= p(X_{n+1} = x_{n+1} | X_n = x_n) \\ &= p(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \cdots, X_2 = x_2) \implies \end{aligned}$$

$$\frac{p(x_{n+1}, \cdots, x_1)}{p(x_n, \cdots, x_1)} = \frac{p(x_{n+1}, \cdots, x_2)}{p(x_n, \cdots, x_2)} \implies$$

$$\frac{p(x_{n+1}, \cdots, x_1)}{p(x_{n+1}, \cdots, x_2)} = \frac{p(x_n, \cdots, x_1)}{p(x_n, \cdots, x_2)} = \cdots = \frac{p(x_2, x_1)}{p(x_2)} \implies$$

$$p(X_1 = x_1 | X_{n+1} = x_{n+1}, X_n = x_n, \cdots, X_2 = x_2) = p(X_1 = x_1 | X_2 = x_2)$$