Information Theory: Lecture Notes 4

zqy1018

June 5, 2020

Contents

1 Entropy Rate

In AEP, when the random variables are i.i.d., there is some good property. But what if they are dependent?

1.1 Basic Definition

Definition. The **entropy rate** of a stochastic process X_i is defined by

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \cdots, X_n)$$

Define $H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_1, \cdots, X_{n-1})$, where n should be determined by context.

By chain rule, we have

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

Thus we can calculate it by solving the two problems separately:

1. Does $H'(\mathcal{X})$ exist?
2. If $a_n \to a$, does $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} a_i$ exist?

Note. We mainly discuss on the case where $\{X_i\}$ is stationary.

1.2 Limit of Conditional Entropy

Theorem 1. For a stationary stochastic process, $H(X_n|X_1, \cdots, X_{n-1})$ is non-increasing in n and has a limit.

Proof.

$$H(X_{n+1}|X_1, \cdots, X_n) \leq H(X_{n+1}|X_2, \cdots, X_n) = H(X_n|X_1, \cdots, X_{n-1})$$

Since $H(X_n|X_1, \cdots, X_{n-1}) \geq 0$, by the monotonic convergence theorem, $H(X_n|X_1, \cdots, X_{n-1})$ converges. □

1.3 Cesaro Mean

Theorem 2. (Cesaro Mean) If $a_n \to a$, $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} a_i = a$.

Proof. Let $b_n = \frac{1}{n}\sum_{i=1}^{n} a_i$. By some calculus, $\forall \epsilon > 0$, $\exists N(\epsilon)$, $\forall n > N(\epsilon)$,

$$\begin{aligned}
|b_n - a| &= \left| \frac{1}{n} \sum_{i=1}^{n} (a_i - a) \right| \\
&\leq \frac{1}{n} \sum_{i=1}^{n} |(a_i - a)| \\
&\leq \frac{1}{n} \sum_{i=1}^{N(\epsilon)} |a_i - a| + \frac{n - N(\epsilon)}{n} \epsilon \\
&\leq \frac{1}{n} \sum_{i=1}^{N(\epsilon)} |a_i - a| + \epsilon
\end{aligned}$$

Since $\sum_{i=1}^{N(\epsilon)} |a_i - a|$ is finite, $b_n \to a$. □

1.4 Combining Them

Combining the results above, we have the following theorem.

Theorem 3. For a stationary stochastic process,

$$H(\mathcal{X}) = H'(\mathcal{X})$$

Corollary. For a stationary Markov chain,

$$H(\mathcal{X}) = H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_{n-1}) = H(X_2|X_1)$$

2 Random Walk

On a graph $G = (V, E)$ with weighted edges, we can define the random walk. The probability that one goes from i to j is

$$p_{ij} = \frac{w_{ij}}{\sum_k w_{ik}}$$

We can represent the random walk as a Markov chain. Let

$$2W = \sum_i \sum_j w_{ij}$$

Theorem 4. If the graph is undirected, i.e. $\forall i, j, w_{ij} = w_{ji}$, then the stationary distribution is

$$\mathbf{x} = \left(x_1, x_2, \cdots, x_{|V|}\right), x_i = \frac{w_i}{2W}$$

where $w_i = \sum_k w_{ik}$.

Thus we can easily find its entropy rate.

Theorem 5. As a special case, if all the weights are the same and the graph is undirected, then

$$H(\mathcal{X}) = \log(2E) - H\left(\frac{\deg(1)}{2E}, \frac{\deg(2)}{2E}, \cdots, \frac{\deg(|V|)}{2E}\right)$$

Note. It is easy to see that a stationary random walk on a graph is *time-reversible*; that is, the probability of any sequence of states is the same forward or backward:

$$p(X_1 = x_1, \cdots, X_n = x_n) = p(X_1 = x_n, \cdots, X_n = x_1)$$

Rather surprisingly, the converse is also true; that is, any time-reversible Markov chain can be represented as a random walk on an undirected weighted graph.

3 Relations With Second Law of Thermodynamics

One of the basic laws of physics, the second law of thermodynamics, states that the entropy of an isolated system is non-decreasing. We now model the isolated system as a *Markov chain* with transitions obeying the physical laws governing the system.

Remark. Implicit in this assumption is the notion of an overall state of the system and the fact that knowing the present state, the future of the system is independent of the past.

Some results can be derived from this model:

- Relative entropy $D(\mu_n \| \mu'_n)$ decreases with n, where μ_n, μ'_n are two different distributions at time n.
- The conditional entropy $H(X_n|X_1)$ increases with n for a stationary Markov process.
- Shuffles increase entropy. That is, for a shuffle operator T, $H(TX) \geq H(X)$.

4 Functions of Markov Chains

Given a stationary Markov chain $\{X_i\}$, let $Y_i = \phi(X_i)$ be a new stochastic process. What is $H(\mathcal{Y})$?

Note. $\{Y_i\}$ is a very special case of *hidden Markov model (HMM).* And it is not a Markov chain in general.

Since $\{X_i\}$ is stationary, $\{Y_i\}$ is stationary. So

$$H(\mathcal{Y}) = \lim_{n\to\infty} H(Y_n|Y_1, \cdots, Y_{n-1})$$

One approach to find this limit is to find a lower bound for $H(\mathcal{Y})$, which is closer to $H(Y_n|Y_1, \cdots, Y_{n-1})$ with larger n. And our choice of this lower bound is $H(Y_n|Y_1, \cdots, Y_{n-1}, X_1)$, because X_1 contains a little more information than Y_1.

Theorem 6.

$$H(Y_n|Y_1, \cdots, Y_{n-1}, X_1) \leq H(\mathcal{Y}) \leq H(Y_n|Y_1, \cdots, Y_{n-1})$$

where all of them converges to $H(\mathcal{Y})$ when $n \to \infty$.

Proof. For the first inequality, we have

$$\begin{aligned} H\left(Y_n|Y_{n-1}, \ldots, Y_2, Y_1, X_1\right) &\overset{\text{(a)}}{=} H\left(Y_n|Y_{n-1}, \ldots, Y_2, X_1\right) \\ &\overset{\text{(b)}}{=} H\left(Y_n|Y_{n-1}, \ldots, Y_1, X_1, X_0, X_{-1}, \ldots, X_{-k}\right) \\ &\overset{\text{(c)}}{=} H\left(Y_n|Y_{n-1}, \ldots, Y_1, X_1, X_0, X_{-1}, \ldots \right. \\ &\left. X_{-k}, Y_0, \ldots, Y_{-k}\right) \\ &\overset{\text{(d)}}{\leq} H\left(Y_n|Y_{n-1}, \ldots, Y_1, Y_0, \ldots, Y_{-k}\right) \\ &\overset{\text{(e)}}{=} H\left(Y_{n+k+1}|Y_{n+k}, \ldots, Y_1\right) \end{aligned}$$

where

- (a) for $Y_1 = \phi(X_1)$.
- (b) for the property of Markov chains.
- (c) for $Y_i = \phi(X_i)$. So it does not matter to add them.
- (d) for conditioning reducing entropy.
- (e) for stationarity.

Thus taking $k \to \infty$, we have $H\left(Y_n|Y_{n-1}, \ldots, Y_2, Y_1, X_1\right) \leq H(\mathcal{Y})$.

For the first equality, we prove that $I(X_1; Y_n|Y_{n-1}, \cdots, Y_1) \to 0$. We have

$$I(X_1; Y_n, Y_{n-1}, \cdots, Y_1) = \sum_{i=1}^{n} I(X_1; Y_i|Y_{i-1}, \cdots, Y_1) \leq H(X_1)$$

Letting $n \to \infty$, by the property of positive series, we know that $I(X_1; Y_n|Y_{n-1}, \cdots, Y_1) \to 0$. □

Acknowledgment

The contents are mainly based on the course materials of CS258, 2020 Spring, Shanghai Jiao Tong University and *Elements of Information Theory* by Thomas M. Cover and Joy A. Thomas.