GI12/4C59: Information Theory

Lecture 10

Massimiliano Pontil

Outline

Theme of this lecture: We introduce the notion of of stochastic process, provide some examples of it and discuss entropy and coding theory in this context.

- Stationary process
- Markov process
- Entropy rate
- Random walk

Stochastic processes

A stochastic process is an indexed sequence of r.v. $X_n, n \in \mathbb{N}$. We say that the process is

- *Stationary* if $P(\{X_{\ell+1} = x_1, \ldots, X_{n+\ell} = x_n\}) = p(x_1, \ldots, x_n)$
- *A Markov chain* if

  $P(\{X_{n+1} = x_{n+1}\}|\{X_n = x_n, \ldots, X_1 = x_1\}) = P(\{X_{n+1} = x_{n+1}\}|\{X_n = x_n\})$

  X_n is called the state of the Markov process at time n.
- *Invariant Markov chain* if the above probability does not depend on n.

In the last case we define $p(x_{n+1}|x_n) := P(\{X_{n+1} = x_{n+1}\}|\{X_n = x_n\})$

Invariant Markov chain

If the process is an invariant Markov chain, we have

$$p(x_1, \ldots, x_n) = p(x_n|x_{n-1})p(x_{n-1}|x_{n-2}) \cdots p(x_2|x_1)p(x_1)$$

We also introduce the transition matrix $P_{ij} = P(X_{n+1} = j|X_n = i)$.

We have $P(X_{n+1} = j) = \sum_i P(X_n = i)P_{ij}$ (another notation is $p_{n+1}(x_{n+1}) = \sum_{x_n} p_n(x_n)P_{x_n x_{n+1}}$.

p_n is called a *stationary distribution* if $p_{n+1} = p_n$. If the initial distribution is stationary it follows that the process is stationary.

Example

Let $\mathcal{X} = \{1, 2\}$ and $P_{11} = 1 - \alpha, P_{12} = \alpha, P_{21} = \beta, P_{22} = 1 - \beta$, with $\alpha, \beta \in [0, 1]$.

The stationary distribution solves the eigenvalue equation

$$\mu P = \mu, \qquad \text{or } P^\top \mu = \mu.$$

A direct computation gives

$$\mu_1 = \frac{\beta}{\alpha + \beta}, \qquad \mu_2 = \frac{\alpha}{\alpha + \beta}$$

where P is the 2×2 matrix whose elements are the P_{ij} above.

Alternatively, this distribution can be obtained by balancing the probability flow across any cut-set in the state transition graph of the process (use $\mu_1 \alpha = \mu_2 \beta$ and $\mu_1 + \mu_2 = 1$)

Entropy rate of a stochastic process

It is defined by

$$h(\mathcal{X}) = \lim_{n \to \infty} \frac{H(X_1, \ldots, X_n)}{n}$$

when the limit exists.

Example 1: If X_i are *identically independent distributed* (i.i.d),

$$H(X_1, \ldots, X_n) = nH(X)$$

which implies that $h(\mathcal{X}) = H(X)$.

Example 2: A typewriter has m equally likely output letters with which can produce m^n equiprobable sequences of length n. In this case we have $H(X_1, \ldots, X_n) = \log m^n$ and, so, $h(\mathcal{X}) = \log m$

Example 3: If X_i are independent but not identical one can have cases where $H(X_i)$ oscillates in a way that $h(\mathcal{X})$ does not exist.

Entropy rate of a stochastic process (cont.)

We also define $\bar{h}(\mathcal{X}) := \lim_{n\to\infty} H(X_n|X_{n-1},\ldots,X_1)$, when the limit exists.

$\bar{h}(\mathcal{X})$ measures the conditional entropy of the last symbol given the past (as opposed to $h(\mathcal{X})$ which measures the per symbol entropy rate).

Theorem: If $X_n, n \in \mathbb{N}$ is a stationary process, the entropy rate exists and $h(\mathcal{X}) = \bar{h}(\mathcal{X})$.

Theorem: If $X_n, n \in \mathbb{N}$ is a time invariant Markov chain, then

$$h(\mathcal{X}) = \bar{h}(\mathcal{X}) = H(X_2|X_1)$$

Proof: Note that $H(X_n|X_{n-1},\ldots,X_1) = X(X_n|X_{n-1}) = H(X_2|X_1)$. The result follows form the previous theorem

Example (cont.)

For a stationary Markov chain we have

$$H(\mathcal{X}) = H'(\mathcal{X}) = \lim H(X_n|X_{n-1},\ldots,X_1) = \lim H(X_n|X_{n-1}) = H(X_2|X_1).$$

Here the conditional entropy is computed using a given stationary distribution μ, and we have

$$H(\mathcal{X}) = -\sum_{ij} \mu_i P_{ij} \log P_{ij} \qquad (\text{from } H(X_2|X_1) = -\sum_{x_1,x_2} p(x_1)p(x_2|x_1)\log p(x_2|x_1)$$

If we go back to the above example we see that:

$$H(X_n) = H(\frac{\alpha}{\alpha+\beta}, \frac{\beta}{\alpha+\beta})$$

and

$$H(X_2|X_1) = \frac{\beta}{\alpha+\beta}H(\alpha) + \frac{\alpha}{\alpha+\beta}H(\beta)$$

Thus, the rate at which the entropy of the process grows is different from the entropy of the state X_n (n is arbitrary).

Random walk

Let G be a connected weighted graph with vertex set $V = \{1, \dots, n\}$ and $n \times n$ symmetric weight matrix W: W_{ij} is the weight of the edge (i, j) (if $W_{ij} = 0$ there is no edge between i and j). We also require that $W_{ii} = 0$ for every i.

A random walk on this graph is the process $X_i, i \in \mathbb{N}$ with $range(X_i) = V$ and given that $X_n = i$ the next vertex j is chosen with probability

$$P_{ij} = \frac{W_{ij}}{\sum_k W_{ik}}$$

(so the next vertex can only be one among those connected to i)

Show that the stationary distribution of this process is $\mu_i = \frac{\sum_j W_{ij}}{\sum_{i,j} W_{ij}}$.

Shannon code

Recall than the average description length L of an optimal code for a r.v. X satisfies:

$$H(X) \leq L < H(X) + 1$$

If X has a $D-$adic distribution, that is, $P(X = x_k) = p_k = D^{-\ell_k}$ for some $\ell_k \in \mathbb{N}$, there exist an optimal code whose average L^* equal $H(X)$. Otherwise we may pay up to an extra bit more than the entropy to describe X.

If we use the (sub-optimal) Shannon code the average description length is still in the above bound. According to this code x_k has codelength equal to $\lceil \log \frac{1}{p_k} \rceil$.

Coding a stochastic process

If we wish to encode a sequence of r.v., $X^n = (X_1, \ldots, X_n)$, we can use the same idea above and have a code for sequence $x^n = (x_1, \ldots, x_n)$ with length

$$\ell(x^n) = \left\lceil \log \frac{1}{p(x^n)} \right\rceil < \log \frac{1}{p(x^n)} + 1$$

and, as before,

$$\frac{H(X^n)}{n} \leq \frac{E[\ell^*(X^n)]}{n} \leq \frac{E[\ell(X^n)]}{n} < \frac{H(X^n)}{n} + \frac{1}{n}.$$

The expected code length per unit symbol is defined by $L_n = E[\ell(X^n)]/n$. Our discussion above tells us that if the process is stationary L_n and L_n^* converge to the entropy rate of the process.

Coding a stochastic process (cont.)

If $X_i = X,\ i = 1, \ldots, n$ (i.i.d. random variables), we have

$$p(x^n) = \prod_{i=1}^{n} p(x_i)$$

and, so,

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

Note that, even in this simple case, unless $p(x)$ is $D-$adic, the codeword lengths for X^n are different from the codeword lengths obtained by concatenating the Shannon code for X, as

$$\ell(x^n) = \log \left\lceil \frac{1}{p(x^n)} \right\rceil \leq \sum_{i=1}^{n} \log \left\lceil \frac{1}{p(x_i)} \right\rceil$$

Bibliography

See Chapter 4 of
T.M. Cover and J.A. Thomas, *The elements of information theory*, Wiley, 1991.