Outline

- On average, $nH(X)+1$ bits suffices to describe n i.i.d. random variables. But what if the random variables are dependent?
- Markov Chain: a simplest way to model the correlations among random variables in a stochastic process.
- Entropy Rate: average number of bits suffices to describe one random variable in a stochastic process.

How to model Dependence: Markov chains

- A stochastic process $\{X_i\}$ is an indexed sequence of random variables $(X_1, X_2, \dots)$ characterized by the joint PMF $p(x_1, x_2, \dots, x_n)$, where $(x_1, x_2, \dots, x_n) \in \mathcal{X}^n$ for $n = 0, 1, \dots$.

Definition

A stochastic process is said to be stationary if the joint distribution of any subset of the sequence of random variables is invariant with respect to shifts in the time index, i.e.,

$$\Pr[X_1 = x_1, X_2 = x_2, \dots, X_n = x_n] = \Pr[X_{1+\ell} = x_1, X_{2+\ell} = x_2, \dots, X_{n+\ell} = x_n]$$

for every n and every shift ℓ and for all $x_1, x_2, \dots, x_n \in \mathcal{X}$.

即每个变量同分布.(但不一定独立)

Markov Chains

Definition

A discrete stochastic process $X_1, X_2, \dots$ is said to be a Markov chain or a Markov process if for $n = 1, 2, \dots$,

$$\Pr[X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \dots, X_1 = x_1] = \Pr[X_{n+1} = x_{n+1} | X_n = x_n]$$

for all $x_1, x_2, \dots, x_n, x_{n+1} \in \mathcal{X}$.

给定系统当前状态，系统的未来状态与过去状态无关.
也即过去状态均包含在了当前状态中.

In this case, the joint PMF can be written as

$$p(x_1, x_2, \dots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_2)\cdots p(x_n|x_{n-1}) := p(x_1)\prod_{i=1}^{n-1} p(x_{i+1}|x_i)$$

Hence, a Markov chain is completely characterized by initial distribution $p(x_1)$ and transition probabilities $p(x_n|x_{n-1})$, $n = 2, 3, 4, \dots$

Definition

The Markov chain is called time invariant if the transition probability $p(x_{n+1}|x_n)$ does NOT depend on n, i.e., for $n = 1, 2, \dots$,

$$\Pr[X_{n+1} = b | X_n = a] = \Pr[X_2 = b | X_1 = a], \quad \forall a, b \in \mathcal{X}.$$

无论哪个时刻，即$n=1,2,\cdots$，$a\to b$的状态转移概率不变.

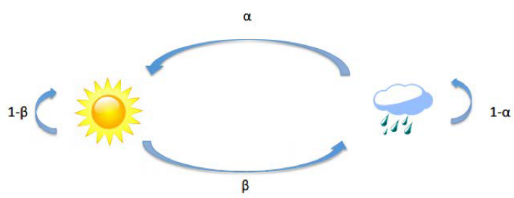
We deal with time invariant Markov chains. If $\{X_i\}$ is a Markov chain, X_n is called the state at time n. A time invariant Markov chain is characterized by its initial state and a probability transition matrix $P = [P_{ij}], i, j \in \{1, 2, \dots, m\}$, where $P_{ij} = \Pr[X_{n+1} = j | X_n = i]$.

Example: Simple Weather Model

- $\mathcal{X} = \{\text{Sunny: S, Rainy: R}\}$

$$p(S|S) = 1-\beta, p(R|R) = 1-\alpha, p(R|S) = \beta, p(S|R) = \alpha$$

$$P = \begin{bmatrix} 1-\beta & \beta \\ \alpha & 1-\alpha \end{bmatrix}$$



- Probability of seeing a sequence SSRR:

$$p(SSRR) = p(S)p(S|S)p(R|S)p(R|R) = p(S)(1-\beta)\beta(1-\alpha)$$

Suppose the first day is "Sunny" with probability γ, what is the weather distribution of the second day, third day, ...?

- If $\mu = [\mu_S, \mu_R] = \begin{bmatrix} \frac{\alpha}{\alpha+\beta} & \frac{\beta}{\alpha+\beta} \end{bmatrix}$

$$P = \begin{bmatrix} 1-\beta & \beta \\ \alpha & 1-\alpha \end{bmatrix}$$

$$p(X_{n+1} = S) = p(S|S)\mu_S + p(S|R)\mu_R$$
$$= (1-\beta)\frac{\alpha}{\alpha+\beta} + \alpha\frac{\beta}{\alpha+\beta} = \frac{\alpha}{\alpha+\beta} = \mu_S.$$

该系统是time invariant Markov Chain

设$P(X_1=S)=a$, $P(X_1=R)=b$, 则

$$P(X_2=S) = a(1-\beta)+b\alpha = (a\ b)\begin{pmatrix} 1-\beta \\ \alpha \end{pmatrix}$$

$$P(X_2=R) = a\beta + b(1-\alpha) = (a\ b)\begin{pmatrix} \beta \\ 1-\alpha \end{pmatrix}$$

$$\begin{array}{ccc} & S & R \\ P = & \Bigg(\begin{matrix} 1-\beta & \beta \\ \alpha & 1-\alpha \end{matrix} & \Bigg) \end{array}$$

$$\Rightarrow (P(X_2=S)\ \ P(X_2=R)) = (a\ b)\, P$$

$$\Rightarrow (P(X_{n+1}=S)\ \ P(X_{n+1}=R)) = (a\ b)\, P^n$$

因为该系统是time invariant 的，故

$$(a\ b)P = (a\ b)$$

$$\Rightarrow \begin{cases} a(1-\beta)+b\alpha = a \\ a\beta + b(1-\alpha) = b \end{cases} \Rightarrow \begin{cases} a = \frac{\alpha}{\alpha+\beta} \\ b = \frac{\beta}{\alpha+\beta} \end{cases}$$

又 $a+b=1$

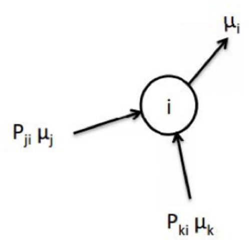
Stationary Distribution

- If the PMF of the random variable at time n is $\mu_i^n = \Pr[X_n = i]$, the PMF at time $n+1$, say $\mu_j^{n+1} = \Pr[X_{n+1} = j]$, can be written as

$$\mu_j^{n+1} = \sum_i \mu_i^n \Pr[X_{n+1} = j | X_n = i] = \sum_i \mu_i^n P_{ij}.$$

- $\{\mu_i^n | \forall i\}$ is called a stationary distribution if $\mu_i^n = \mu_i^{n+1}, \forall i$.
- For notation convenience, let $\mu_i = \mu_i^n = \mu_i^{n+1}, \forall i$.
- How to calculate stationary distribution?
  - Stationary distribution $\mu_i, i = 1, 2, \ldots, |\mathcal{X}|$ satisfies

$$\mu_j = \sum_{i=1}^{|\mathcal{X}|} \mu_i P_{ij} \quad \text{and} \quad \sum_{i=1}^{|\mathcal{X}|} \mu_i = 1.$$



推广：$\mathcal{X} = \{x_1, \ldots, x_m\}$

time invariant 需满足 $(\mu_1, \mu_2, \ldots \mu_m) = (\mu_1, \mu_2, \ldots \mu_m)\underset{}{\overset{m\times m}{P}}$

即 $(\mu_1, \mu_2, \ldots \mu_m)(P-I) = 0$

又有归一化 $(\mu_1, \mu_2, \ldots, \mu_m)e = 1$

则有 $(\mu_1, \mu_2, \ldots, \mu_m)\underbrace{[P-I, e]}_{m\times(m+1)} = (\underbrace{0 \ldots 0}_{m}\ 1)$

令 $\tilde{P} = [P-I, e]$，并左右同乘伪逆 $\tilde{P}^T(\tilde{P}\tilde{P}^T)^{-1}$

$$(\mu_1\ \mu_2\ \ldots\ \mu_m) = (0 \ldots 0\ 1)\tilde{P}^T(\tilde{P}\tilde{P}^T)^{-1}$$

Entropy Rate

- When X_i's are i.i.d., the entropy

$$H(X^n) = H(X_1, X_2, \dots, X_n) = \sum_{i=1}^{n} H(X_i) = nH(X).$$

- With dependent sequences X_i's, how does $H(X^n)$ grow with n?

- Entropy rate characterized the growth rate.

熵率描述了$H(X^n)$随n增长的比率

- **Definition 1:** average entropy per symbol

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

- **Definition 2:** conditional entropy of the last r.v. given the past

$$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n | X_{n-1}, X_{n-2}, \dots, X_1)$$

Theorem 4.2.2

For a stationary stochastic process, $H(X_n|X_{n-1}, \dots, X_1)$ is nonincreasing in n and has a limit $H'(\mathcal{X})$.

Proof.

$$H\left(X_{n+1}|X_1, X_2, \dots, X_n\right) \overset{\text{conditional reduces entropy}}{\leq} H\left(X_{n+1}|X_n, \dots, X_2\right) \overset{\text{stationary}}{=} H(X_n|X_{n-1}, \dots, X_1),$$

- $H(X_n|X_{n-1}, \dots, X_1)$ decreases as n increases
- $H(X) \geq 0$
- The limit must exist.

□

Theorem 4.2.1

For a stationary stochastic process, $H(\mathcal{X}) = H'(\mathcal{X})$.

对平稳随机过程，上述两个定义相等

Proof.

By the chain rule,

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

- $H(X_n|X_{n-1}, \dots, X_1) \to H'(\mathcal{X})$
- *Cesaro mean*: If $a_n \to a$, $b_n = \frac{1}{n}\sum_{i=1}^{n} a_i$, then $b_n \to a$.
- So

$$\frac{1}{n} H(X_1, \dots, X_n) \to H'(\mathcal{X})$$

□

AEP for Stationary Ergodic Process

$$-\frac{1}{n} \log p(X_1, \dots, X_n) \to H(\mathcal{X})$$

- $p(X_1, \dots, X_n) \approx 2^{-nH(\mathcal{X})}$
- Typical sequences in typical set of size $2^{-nH(\mathcal{X})}$
- We can use $nH(\mathcal{X})$ bits to reprensent typical sequences

Entropy Rate for Markov Chain

- For a stationary Markov chain, the entropy rate is

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

- Let $P_{ij} = \Pr[X_2 = j|X_1 = i]$. By definition, entropy rate of stationary Markov chain

$$H(\mathcal{X}) = H(X_2|X_1) = \sum_i \mu_i (\sum_j -P_{ij} \log P_{ij}) = -\sum_{ij} \mu_i P_{ij} \log P_{ij}$$

Calculate Entropy Rate

1. Find *stationary distribution* μ_i

$$\mu_i = \sum_j \mu_j p_{ji} \text{ and } \sum_{i=1}^{|\mathcal{X}|} \mu_i = 1$$

2. User *transition probability* P_{ij}

$$H(\mathcal{X}) = -\sum_{ij} \mu_i P_{ij} \log P_{ij}$$

Entropy Rate of Weather Model

- Stationary distribution $\mu(S) = \frac{\alpha}{\alpha+\beta}$, $\mu(R) = \frac{\beta}{\alpha+\beta}$

$$P = \begin{bmatrix} 1-\beta & \beta \\ \alpha & 1-\alpha \end{bmatrix}$$

$$\begin{aligned} H(\mathcal{X}) &= \mu(S)H(\beta) + \mu(R)H(\alpha) \\ &= \frac{\alpha}{\alpha+\beta}H(\beta) + \frac{\beta}{\alpha+\beta}H(\alpha) \\ &\overset{\text{Jensen's inequality}}{\leq} H(2\frac{\alpha\beta}{\alpha+\beta}) \end{aligned}$$

Maximum when $\alpha = \beta = 1/2$: degenerate to independent process