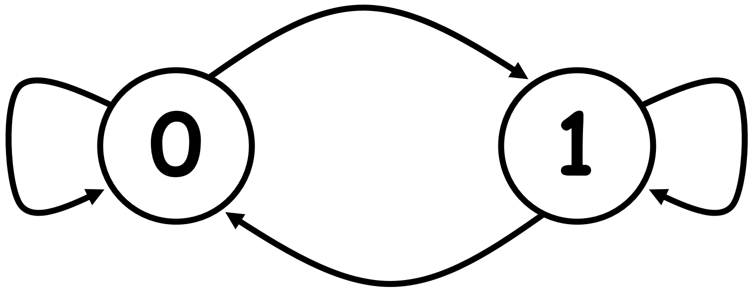
Markov Information Source Model



Hiroki Sayama
sayama@binghamton.edu

Today's topic

Behaviors of most of stochastic systems around us are NOT i.i.d.
(think about letters in a language, weather, baseball winning teams, etc.)

⬇

How can we model & calculate the entropy of a system that involves correlations between events?

Review: Generalization of entropy

- If the stochastic system's behavior is not i.i.d. (i.e., the values are corre

entro

Called "entropy rate"

$$\overline{H}(X) = \lim_{k \to \infty} H(X^k)/k$$

This is $\quad\quad$ $_i$ log p_i

How to calculate it?

Average # of bits needed to describe one event

Markov Information Source

Markov information source

- Information source whose probability distribution at time t depends only on its immediate past value X_{t-1} (or past n values $X_{t-1}, X_{t-2}, \ldots, X_{t-n}$)
  - Cases n>1 can be converted into n=1 form by defining composite events
  - Probabilistic rules are given as a set of conditional probabilities, which can be written in the form of a transition probability matrix (TPM)

Memoryless and Markov information sources

01010010001011011001101000110

Memoryless information source

$p(0) = p(1) = 1/2$

01000000111111001110001111111

Markov information source

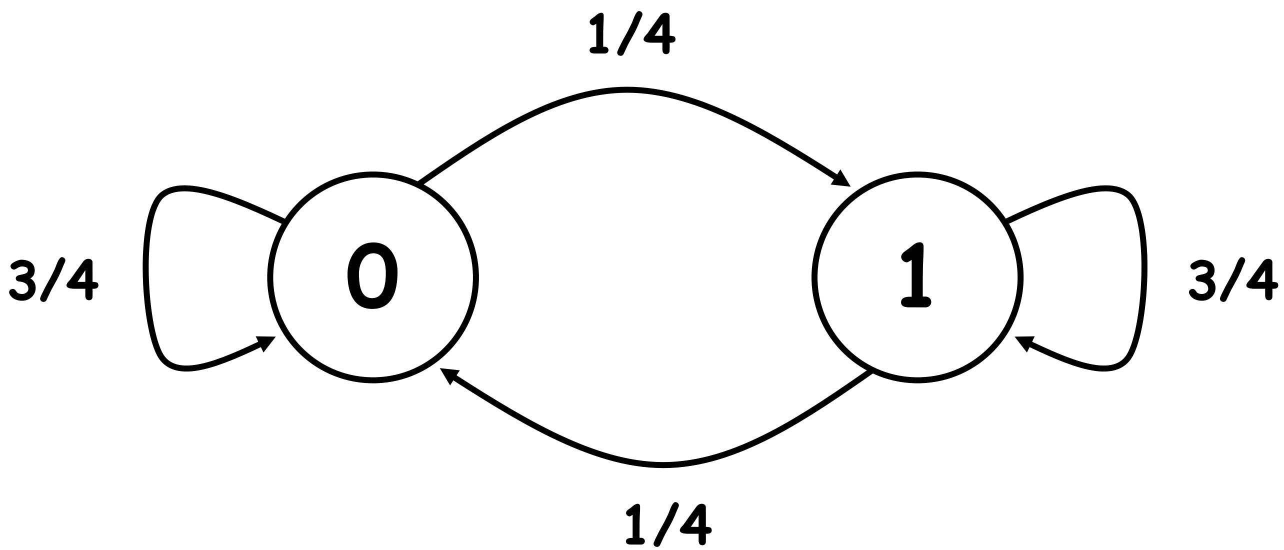
$p(1|0) = p(0|1) = 1/4$

State-transition diagram

010000000111111001110001111111

Markov information source

$p(1|0) = p(0|1) = 1/4$



Exercise

- Draw a state-transition diagram of the following binary state Markov information source
- The next state will be:
  - 1 if the previous three states were "000"
  - 0 if the previous three states were "111"
  - Randomly chosen from {0, 1} otherwise

Matrix representation

01000000111111001110001111111

Markov information source

$p(1|0) = p(0|1) = 1/4$

Probability vector at time t — TPM — Probability vector at time t-1

$$\begin{pmatrix} p_0 \\ p_1 \end{pmatrix}_t = \begin{pmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{pmatrix} \begin{pmatrix} p_0 \\ p_1 \end{pmatrix}_{t-1}$$

Exercise

abcaccaabccccaaabc
aaccacaccaaaaabcc

- Consider the above sequence as a Markov information source and create its state-transition diagram and matrix representation

Convenient properties of transition probability matrix

- The product of two TPMs is also a TPM
- All TPMs have eigenvalue 1
- $|\lambda| \leq 1$ for all eigenvalues of any TPM
- If the transition network is strongly connected, the TPM has one and only one eigenvalue 1 (no degeneration)

Exercise: Prove the following

- The product of two TPMs is also a TPM
- All TPMs have eigenvalue 1
- $|\lambda| \leq 1$ for all eigenvalues of any TPM
- If the transition network is strongly connected, the TPM has one and only one eigenvalue 1 (no degeneration)

Solution (1)

- All TPMs have eigenvalue 1
  - You can show that there exists a non-zero vector q that satisfies $A\ q = q$, i.e. $(A-I)\ q = 0$

    $\rightarrow |A-I| = 0$

    This holds when column vectors of $A-I$ are linearly dependent with each other (i.e., $A-I$ maps vectors to a subspace of fewer dimensions)

Solution (1)

- $A-I$ actually looks like this:

$$\begin{pmatrix} a_{11}-1 & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22}-1 & \dots & a_{2n} \\ : & : & \dots & : \\ a_{n1} & a_{n2} & \dots & a_{nn}-1 \end{pmatrix}$$

- Note that each column vector is in a subspace $s_1+s_2+...+s_n = 0$

$\rightarrow |A-I| = 0$

Solution (2)

- $|\lambda| \leq 1$ for all eigenvalues of any TPM
  - For any λ, $A^m q = \lambda^m q$ (q: eigenvector)
  - A^m is a product of TPM, therefore it must be a TPM as well whose elements are all <= 1
  - $A^m q = \lambda^m q$ can't diverge

    $\rightarrow |\lambda| \leq 1$

TPM and asymptotic probability distribution

- $|\lambda| \leq 1$ for all eigenvalues of any TPM
- If the transition network is strongly connected, the TPM has one and only one eigenvalue 1 (no degeneration)

$\rightarrow$ This eigenvalue is a unique dominant eigenvalue and the probability vector will eventually converge to its corresponding eigenvector

Exercise

- Calculate the asymptotic probability distribution of the following:

01000000111111001110001111111

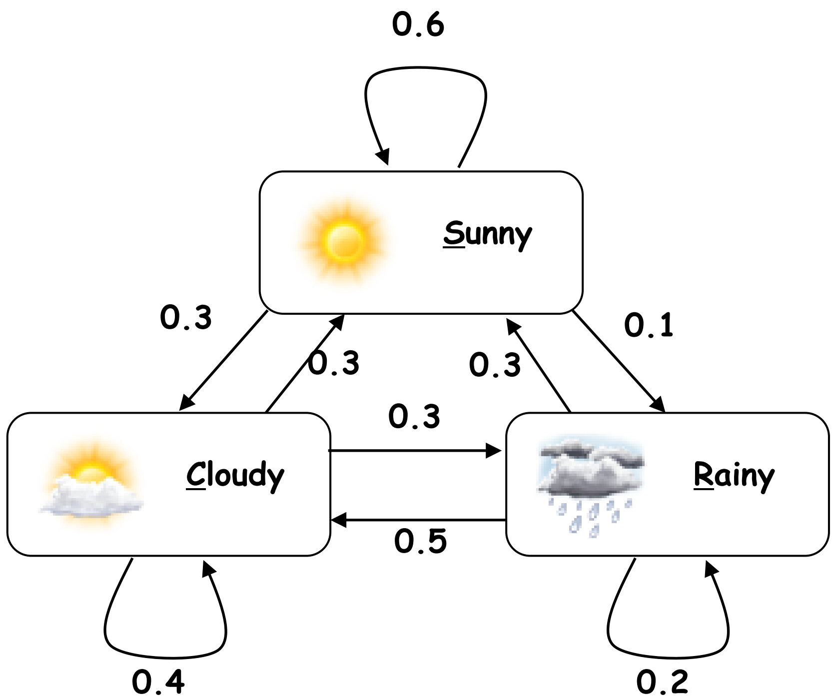
Markov information source

$p(1|0) = p(0|1) = 1/4$

$$\begin{pmatrix} p_0 \\ p_1 \end{pmatrix}_t = \begin{pmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{pmatrix} \begin{pmatrix} p_0 \\ p_1 \end{pmatrix}_{t-1}$$

Exercise

- Obtain the asymptotic probability distribution of the following system



Calculating Entropy of Markov Information Source

Calculating entropy (1)

$$\overline{H}(X) = \lim_{k\to\infty} H(X^k)/k$$

- Because $H(XY) = H(Y|X) + H(X)$, $H(X^k)$ can be rewritten as

$$H(X^k) = H(X_k|X_{k-1}X_{k-2}\ldots X_1)$$
$$+ H(X_{k-1}X_{k-2}\ldots X_1)$$
$$= H(X_k|X_{k-1}X_{k-2}\ldots X_1)$$
$$+ H(X_{k-1}|X_{k-2}\ldots X_1) + H(X_{k-2}\ldots X_1)$$
$$\ldots \quad = \Sigma_{i=1\sim k} H(X_i|X_{k-i}\ldots X_1)$$

Calculating entropy (2)

$\overline{H}(X) = \lim_{k\to\infty} \Sigma_{i=1\sim k} H(X_i|X_{k-i}\ldots X_1)/k$

- $\lim_{k\to\infty} H(X_k|X_{k-1}\ldots X_1)$ converges to a finite value if the probability of a sequence of events doesn't depend on time (called "stationary"), because

$$H(X_k|X_{k-1}\ldots X_1) \leq H(X_k|X_{k-1}\ldots X_2)$$
$$= H(X_{k-1}|X_{k-2}\ldots X_1)$$

Because left hand side has a more constraining condition

Because of stationarity

Calculating entropy (3)

$$\overline{H}(X) = \lim_{k\to\infty} \Sigma_{i=1\sim k} H(X_i|X_{k-i}\ldots X_1)/k$$

- $\lim_{k\to\infty} H(X_k|X_{k-1}\ldots X_1)$ converges to a finite value for stationary systems
- Therefore,

$$\overline{H}(X) = \lim_{k\to\infty} H(X_k|X_{k-1}\ldots X_1)$$

(because the sum is averaging H over $k\to\infty$, which will be dominated by the asymptotic value)

Calculating entropy (4)

$$\overline{H}(X) = \lim_{k\to\infty} H(X_k|X_{k-1}\ldots X_1)$$

- If it is a Markov information source, the next state depends only on X_{k-1}:

$$\overline{H}(X) = \lim_{k\to\infty} H(X_k|X_{k-1})$$

$$= \lim_{k\to\infty} - \Sigma_{x_{k-1}} p(x_{k-1}) \Sigma_{x_k} p(x_k|x_{k-1}) \log p(x_k|x_{k-1})$$

$$= - \Sigma_j q_j \Sigma_i A_{ij} \log A_{ij}$$

Calculating entropy (5)

$$\overline{H}(X) = - \sum_j q_j \sum_i A_{ij} \log A_{ij}$$

- q is the asymptotic probability distribution (A's dominant eigenvector)
- $h_j = - \sum_i A_{ij} \log A_{ij}$ is the entropy of A's j-th column
- With $h = (h_1, h_2, \ldots)^T$:

$$\overline{H}(X) = h \cdot q$$

Bottom line

- Information entropy of a Markov information source is given by **the average of entropies of its TPM's column vectors weighted by its asymptotic probability distribution**

  (The source needs to have only one asymptotic probability distribution though)

Exercise

- Calculate information entropy of the following Markov information source we discussed earlier:

01000000111111001110001111111

abcaccaabccccaaabc
aaccacaccaaaaabcc

Summary

- Even if the stochastic system involves correlations between events, you can:
  - Construct a Markov information source model
  - Represent it in a diagram or a TPM
  - Calculate its asymptotic probability distribution
  - Calculate its entropy (entropy rate)

Exercise

- Choose some real-world data and model it as a Markov information source
- Calculate its asymptotic probability distribution
- Calculate its entropy