Hidden Markov model and its application in sequence analysis

Xiaohui Xie

University of California, Irvine

CpG islands

- Example: CpG islands
- CpG site: CpG is a pair of nucleotides C and G, appearing successively in this order along one DNA strand. "CpG" stands for cytosine and guanine separated by a phosphate, which links the two nucleosides together in DNA.
- CpG is relative rare in vertebrate genomes due to certain biochemical properties.
- Frequency of CpG dinucleotides in the human genome:
  Based on 42% GC content (like the human genome).
  1. Expected frequency: 0.21*0.21 = 4.41%
  2. Actual frequnency: 1%

CpG islands detection

- CpG islands

  Regions of the DNA that have a higher concentration of CpG sites, typically several hundreds of nucleotides long.

  What are CpG islands interesting? They are typically associated with the promoter regions of genes. About half of the genes in human have a CpG island present in their promoters.

- CpG island detection problem

  Input: A long DNA sequences $X = (x_1, \cdots, x_L) \in \Sigma^*$

  Question: Locate all CpG lands within X.

An occasionally dishonest casino dealer

Suppose a dealer in a casino rolls a die. The dealer use a fair die most of the time, but occasionally he switches to a loaded die. The loaded die has probability 0.5 of a six and probability 0.1 for the numbers 1 to 5.

Hidden Markov Model

Definition: A hidden Markov model (HMM) is a triplet $M = (\Sigma, Q, \Theta)$, where

- Σ is an alphabet of symbols, e.g. $\Sigma = \{1, 2, 3, 4, 5, 6\}$.
- Q is a finite set of states, capable of emiting symbols from Σ, e.g. $Q = \{F, L\}$.
- Θ is a set of probabilities, consisting of
  1) State transition probabilities, a_{kl} for all $k, l \in Q$. and 2) Emission probabilities.

Hidden Markov Model

- State transition probabilities,a_{kl} for all $k, l \in Q$.

$$a = \begin{bmatrix} a_{FF} & a_{FL} \\ a_{LF} & a_{LL} \end{bmatrix} = \begin{bmatrix} 0.95 & 0.05 \\ 0.1 & 0.9 \end{bmatrix}$$

- Emission probabilities: $e_k(b)$ for all $k \in Q$ and $b \in \Sigma$.

$$e_F = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6) \quad (1)$$

$$e_L = (0.1, 0.1, 0.1, 0.1, 0.1, 0.5) \quad (2)$$

Denote $\Theta = (a, e_F, e_L)$.

Path and joint prob

- A path $Z = (z_1, \cdots, z_L)$ in the model M is a sequence of states, modeled by a Markov chain.

$$P(z_i = l | z_{i-1} = k) = a_{kl}$$

- Joint probability:

$$P(X, Z) = P(z_1) \prod_{i=2}^{L} P(z_i | z_{i-1}) \prod_{i=1}^{L} P(x_i | z_i) \quad (3)$$

$$= \prod_{i=1}^{L} e_{z_i}(x_i) a_{z_{i-1}, z_i} \quad (4)$$

where we define $a_{z_0, z_1} = P(z_1)$.

- A sequence $X = (x_1, \cdots, x_L) \in \Sigma^L$:

$$P(x_i = b | z_i = k) = e_k(b)$$

Decoding problem

- The decoding problem
  Input: A HMM $M = (\Sigma, Q, \Theta)$ and a sequence $X = \Sigma^*$, for which the generating path $Z = (z_1, \cdots, z_L)$ is unknown.
  Question: Find the most probable path $\hat{Z}$ for X, i.e.

$$\hat{Z} = \arg\max_{Z} P(X, Z)$$

Decoding problem: Viterbi algorithm

Consider a path ending at $k \in Q$, and the probability of Z generating the prefix $(x_1, \cdots, x_i)$ of X.

$$v_k(i) = \max_{z_1, \cdots, z_{i-1}} P(x_1, \cdots, x_i, z_1, \cdots, z_{i-1}, z_i = k)$$

- Initialize

$$v_k(1) = P(z_1 = k) e_k(x_1) \quad \text{forall } k \in Q$$

- For $i = 1, \cdots, L-1$ and $l \in Q$,

$$v_l(i+1) = e_l(x_{i+1}) \max_{k \in Q} \{v_k(i) a_{kl}\}$$

- Finally

$$P(X, \hat{Z}) = \max_{k \in Q} v_k(L)$$

The posterior decoding problem

- The posterior decoding problem

  Input: A HMM $M = (\Sigma, Q, \Theta)$, and a sequence $X \in \Sigma^*$, for which the generating path $Z = (z_1, \cdots, z_L)$ is unknown.

  Question: For all $1 \leq i \leq L$ and $k \in Q$, find $P(z_i = k|X)$.

Posterior decoding: forward algorithm

Denote by $f_k(i)$ the prob of emitting the prefix $(x_1, \cdots, x_i)$ and eventually reaching state $z_i = k$:

$$f_k(i) = P(x_1, \cdots, x_i, z_i = k)$$

- Initialize

$$f_k(1) = P(z_1 = k)e_k(x_1) \quad \text{for all } k \in Q$$

- For $i = 1, \cdots, L-1$ and $l \in Q$,

$$f_l(i+1) = e_l(x_{i+1}) \sum_{k \in Q} f_k(i) a_{kl}$$

- Finally

$$P(X) = \sum_{k \in Q} f_k(L)$$

Posterior decoding: backward algorithm

Denote by $b_k(i)$ the prob of emitting the suffix $(x_{i=1}, \cdots, x_L)$ given $z_i = k$:

$$b_k(i) = P(x_{i+1}, \cdots, x_L | z_i = k)$$

- Initialize

$$b_k(L) = 1 \quad \forall k \in Q$$

- For $i = L-1, \cdots, 1$ and $l \in Q$,

$$b_l(i) = \sum_{l \in Q} a_{kl} e_l(x_{i+1}) b_k(i+1)$$

- Finally

$$P(X) = \sum_{l \in Q} P(z_1 = l) e_l(x_1) b_l(1)$$

The posterior decoding problem

- The posterior decoding problem

$$P(X, z_i = k) = P(x_1, \cdots, x_i, z_i = k) P(x_{i+1}, \cdots, x_L | x_1, \cdots, x_i, z_i = k) \tag{5}$$

$$= P(x_1, \cdots, x_i, z_i = k) P(x_{i+1}, \cdots, x_L | z_i = k) \tag{6}$$

$$= f_k(i) b_k(i) \tag{7}$$

Hence,

$$P(z_i = k | X) = \frac{f_k(i) b_k(i)}{P(X)}$$

Parameter estimation

- Parameter estimation for HMM

  Input: Given n example sequences $X^{(1)}, \cdots, X^{(n)} \in \Sigma^*$ of length $L^{(1)}, \cdots, L^{(n)}$, respectively, which were generated from a HMM $M = (\Sigma, Q, \Theta)$ with unknown Θ.

  Question: Find the most probable $\hat{\Theta}$, that is

$$\hat{\Theta} = \arg\max_{\Theta} P(X^{(1)}, \cdots, X^{(n)} | \Theta)$$

Parameter estimation I: state is known

- When the state sequence is known
  Suppose we know the state sequences $Z^{(1)}, \cdots, Z^{(n)} \in \Sigma^*$ corresponding to $X^{(1)}, \cdots, X^{(n)} \in \Sigma^*$, respectively. We can then scan the sequences and compute
  1. A_{kl} - the number of transitions from state k to l.
  2. $E_k(b)$ - the number of times that an emission of the symbol b occurred in state k.
  Then the ML estimators will be

$$a_{kl} = \frac{A_{kl}}{\sum_{q \in Q} A_{kq}} \tag{8}$$

$$e_k(b) = \frac{E_k(b)}{\sum_{\sigma \in \Sigma} E_k(\sigma)} \tag{9}$$

Parameter estimation I: state is unknown

- When the state sequence is unknown: Baum-Welch algorithm
  1. Initialization: Assign random values to Θ.
  2. Expectation

EM-algorithm: Expectation

- Calculate the expected number of state transtions from statet k to l. Note that

$$P(z_i = k, z_{i+1} = l | X, \Theta) = \frac{f_k(i) a_{kl} e_l(x_{i+1}) b_l(i+1)}{P(X)}$$

Let $\{f_k^{(j)}(i), b_k^{(j)}\}$ denote the forward and backward probabilities of the sequence $X^{(j)}$. Then the expectation is

$$A_{kl} = \sum_{j=1}^{n} \frac{1}{P(X^{(j)})} \sum_{i=1}^{L^{(j)}} f_k^{(j)}(i)\ a_{kl}\ e_l(x_{i+1}^{(j)})\ b_l^{(j)}(i+1)$$

EM-algorithm: Expectation

- Calculate the expected number of emissions of the symbol b that occurred at the state k,

$$E_k(b) = \sum_{j=1}^{n} \frac{1}{P(X^{(j)})} \sum_{\{i:x_i^{(j)}=b\}} f_k^{(j)}(i)\ b_k^{(j)}(i)$$

EM-algorithm: Expectation

- Initialize
- Repeat until convergence
  - Expectation
  - Maximization: Update Θ using A_{kl} and $E_k(b)$.