Computational Systems Biology: Biology X

Bud Mishra

Room 1002, 715 Broadway, Courant Institute, NYU, New York, USA

L#4:(Feb-23-2010)
Genome Wide Association Studies

Outline

The law of causality ... is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm ...

–Bertrand Russell, *On the Notion of Cause*. Proceedings of the Aristotelian Society 13: 1-26, 1913.

Outline

Random Variables

- A (discrete) random variable is a numerical quantity that in some experiment (involving randomness) takes a value from some (discrete) set of possible values.
- More formally, these are measurable maps

$$X(\omega), \omega \in \Omega,$$

  from a basic probability space (Ω, F, P) ($\equiv$ outcomes, a sigma field of subsets of Ω and probability measure P on F).
- *Events*

$$\ldots\{\omega \in \Omega | X(\omega) = x_i\}\ldots$$

  same as $\{X = x_i\}$ [X assumes the value x_i].

Few Examples

- Example 1: Rolling of two six-sided dice. Random Variable might be the sum of the two numbers showing on the dice. The possible values of the random variable are 2, 3, . . ., 12.
- Example 2: Occurrence of a specific word *GAATTC* in a genome. Random Variable might be the number of occurrence of this word in a random genome of length 3×10^9. The possible values of the random variable are 0, 1, 2, . . ., 3×10^9.

The Probability Distribution

- The *probability distribution* of a discrete random variable Y is the set of values that this random variable can take, together with the set of associated probabilities.
- Probabilities are numbers in the range between zero and one (inclusive) that always add up to one when summed over all possible values of the random variable.

Bernoulli Trial

- A *Bernoulli trial* is a single trial with two possible outcomes: "success" & "failure."

$$P(\text{success}) = p \text{ and } P(\text{failure}) = 1 - p \equiv q.$$

- Random variable S takes the value -1 if the trial results in failure and $+1$ if it results in success.

$$P_S(s) = p^{(1+s)/2} q^{(1-s)/2}, \quad s = -1, +1.$$

The Binomial Distribution

- A *Binomial random variable* is the number of successes in a fixed number n of independent Bernoulli trials (with success probability = p).
- Random variable Y denotes the total number of successes in the n trials.

$$P_Y(y) = \binom{n}{y} p^y q^{n-y}, \quad y = 0, 1, \ldots, n.$$

The Uniform Distribution

- A random variable Y has the *uniform distribution* if the possible values of Y are a, $a+1$, ..., $a+b-1$ for two integer constants a and b, and the probability that Y takes any specified one of these b possible values is b^{-1}.

$$P_Y(y) = b^{-1}, \quad y = a, a+1, \ldots, a+b-1.$$

The Geometric Distribution

- Suppose that a sequence of independent Bernoulli trials is conducted, each trial having probability p of success. The random variable of interest is the number Y of trials before but not including the first failure. The possible values of Y are 0, 1, 2, $\ldots$.

$$P_Y(y) = p^y q, \quad y = 0, 1, \ldots.$$

The Poisson Distribution

- A random variable Y has a Poisson distribution (with parameter $\lambda > 0$) if

$$P_Y(y) = \frac{e^{-\lambda}\lambda^y}{y!}, \quad y = 0, 1, \ldots.$$

- The Poisson distribution often arises as a limiting form of the binomial distribution.

Continuous Random Variables

- We denote a continuous random variable by X and observed value of the random variable by x.
- Each random variable X with range I has an associated density function $f_X(x)$ which is defined, positive for all x and integrates to one over the range I.

$$\text{Prob}(a < X < b) = \int_a^b f_X(x)dx.$$

The Normal Distribution

- A random variable X has a normal or Gaussian distribution if it has range $(-\infty, \infty)$ and density function

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

where μ and $\sigma > 0$ are parameters of the distribution.

Expectation

- For a random variable Y, and any function $g(Y)$ of Y, the expected value of $g(Y)$ is

$$E(g(Y)) = \sum_y g(y) P_Y(y),$$

when Y is discrete; and

$$E(g(Y)) = \int_y g(y) f_Y(y)\, dy,$$

when Y is continuous.

- Thus,

$$\text{mean}(Y) = E(Y) = \mu(Y),$$

$$\text{variance}(Y) = E(Y^2) - E(Y)^2 = \sigma^2(Y).$$

Conditional Probabilities

- Suppose that A_1 and A_2 are two events such that $P(A_2) \neq 0$. Then the conditional probability that the event A_1 occurs, given that event A_2 occurs, denoted by $P(A_1|A_2)$ is given by the formula

$$P(A_1|A_2) = \frac{P(A_1 \& A_2)}{P(A_2)}.$$

Bayes Rule

- Suppose that A_1 and A_2 are two events such that $P(A_1) \neq 0$ and $P(A_2) \neq 0$. Then

$$P(A_2|A_1) = \frac{P(A_2)P(A_1|A_2)}{P(A_1)}.$$

Markov Models

- Suppose there are n states S_1, S_2, ..., S_n. And the probability of moving to a state S_j from a state S_i depends only on S_i, but not the previous history. That is:

$$\begin{aligned} P(s(t+1) = S_j | s(t) = S_i, s(t-1) = S_{i_1}, \ldots) \\ = \quad P(s(t+1) = S_j | s(t) = S_i). \end{aligned}$$

  Then by Bayes rule:

$$\begin{aligned} P(s(0) = S_{i_0}, s(1) = S_{i_1}, \ldots, s(t-1) = S_{i_{t-1}}, s(t) = S_{i_t}) \\ = \quad P(s(0) = S_{i_0}) P(S_{i_1} | S_{i_0}) \cdots P(S_{i_t} | S_{i_{t-1}}). \end{aligned}$$

HMM: Hidden Markov Models

Defined with respect to an **alphabet** Σ

- A set of (hidden) **states** Q,
- A $|Q| \times |Q|$ matrix of **state transition probabilities** $A = (a_{kl})$, and
- A $|Q| \times |\Sigma|$ matrix of **emission probabilities** $E = (e_k(\sigma))$.

States

Q is a set of states that emit symbols from the alphabet Σ. Dynamics is determined by a state-space trajectory determined by the state-transition probabilities.

A Path in the HMM

- Path $\Pi = \pi_1 \pi_2 \cdots \pi_n$ = a sequence of states $\in Q^*$ in the hidden markov model, M.
- $x \in \Sigma^* =$ sequence generated by the path Π determined by the model M:

$$P(x|\Pi) = P(\pi_1) \left[\prod_{i=1}^{n} P(x_i|\pi_i) \cdot P(\pi_i|\pi_{i+1}) \right]$$

A Path in the HMM

- Note that

$$\begin{aligned} P(x|\Pi) &= P(\pi_1)\left[\prod_{i=1}^{n} P(x_i|\pi_i)\cdot P(\pi_i|\pi_{i+1})\right] \\ P(x_i|\pi_i) &= e_{\pi_i}(x_i) \\ P(\pi_i|\pi_{i+1}) &= a_{\pi_i,\pi_{i+1}} \end{aligned}$$

- Let π_0 and π_{n+1} be the initial (“begin”) and final (“end”) states, respectively

$$P(x|\Pi) = a_{\pi_0,\pi_1} e_{\pi_1}(x_1) a_{\pi_1,\pi_2} e_{\pi_2}(x_2) \cdots e_{\pi_n}(x_n) a_{\pi_n,\pi_{n+1}}$$

i.e.

$$P(x|\Pi) = a_{\pi_0,\pi_1} \prod_{i=1}^{n} e_{\pi_i}(x_i) a_{\pi_i,\pi_{i+1}}.$$

Decoding Problem

- For a given sequence x, and a given path π, the model (Markovian) defines the probability $P(x|\Pi)$
- In a casino scenario: the dealer knows Π and x, the player knows x but not Π.
- "The path of x is hidden."
- **Decoding Problem**: Find an optimal path π^* for x such that $P(x|\pi)$ is maximized.

$$\pi^* = \arg\max_{\pi} P(x|\pi).$$

Dynamic Programming Approach

Principle of Optimality

Optimal path for the $(i+1)$-prefix of x

$$x_1 x_2 \cdots x_{i+1}$$

uses a path for an i-prefix of x that is optimal among the paths ending in an unknown state $\pi_i = k \in Q$.

Dynamic Programming Approach

Recurrence: $s_k(i) =$ the probability of the most probable path for the i-prefix ending in state k

$$\forall_{k \in Q} \forall_{1 \le i \le n} \qquad s_k(i) = e_k(x_i) \cdot \max_{l \in Q} s_l(i-1) a_{lk}.$$

Dynamic Programming

- $i = 0$, Base case

$$s_{begin}(0) = 1, s_k(0) = 0, \forall_{k \neq begin}.$$

- $0 < i \leq n$, Inductive case

$$s_l(i+1) = e_l(x_{i+1}) \cdot \max_{k \in Q}[s_k(i) \cdot a_{kl}]$$

- $i = n+1$

$$P(x|\pi^*) = \max_{k \in Q} s_k(n) a_{k,end}.$$

Viterbi Algorithm

- Dynamic Programing with "**log-score**" function

$$S_l(i) = \log s_l(i).$$

- Space Complexity = $O(n|Q|)$.
- Time Complexity = $O(n|Q|)$.
- Additive formula:

$$S_l(i+1) = \log e_l(x_{i+1}) + \max_{k \in Q}[S_k(i) + \log a_{kl}].$$

[End of Lecture #4]