Computational Systems Biology: Biology X

Bud Mishra

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

L#5:(Mar-02-2010)
Genome Wide Association Studies

Outline

1. A Short Introduction to Probability and Causation
   - Probability
   - Causation
   - Association Studies

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

1. A Short Introduction to Probability and Causation
   - Probability
   - Causation
   - Association Studies

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, ...

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

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.

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

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

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

Causation and Correlation

- A fallacy, known as *cum hoc ergo propter hoc* (Latin for "with this, therefore because of this"): Correlations do not imply causation.
- Statements associated with *necessity* and *sufficiency*
- **The INUS condition**: An Insufficient but Non-redundant part of an Unnecessary but Sufficient condition.
- **The Probability Raising condition**
- **Temporal Priority**

Regularity Theories (David Hume)

- **Causes are invariably followed by their effects**: “We may define a cause to be an object, followed by another, and where all the objects similar to the first, are followed by objects similar to the second.”
- Attempts to analyze causation in terms of invariable patterns of succession are referred to as “regularity theories” of causation.
- There are a number of well-known difficulties with regularity theories, and these may be used to motivate probabilistic approaches to causation.

Imperfect Regularities

- The first difficulty is that most causes are not invariably followed by their effects.
- **Penetrance**: The presence of a disease allele does not always lead to a disease phenotype.
- **Probabilistic theories of causation**: simply requires that *causes raise the probability of their effects*; an effect may still occur in the absence of a cause or fail to occur in its presence.
- Thus smoking is a cause of lung cancer, not because all smokers develop lung cancer, but because smokers are more likely to develop lung cancer than non-smokers.

Imperfect Regularities: INUS condition

- **John Stuart Mill and John Mackie** offer more refined accounts of the regularities that underwrite causal relations.
- **An INUS condition**: for some effect is an *insufficient but non-redundant part of an unnecessary but sufficient condition.*
- **Complexity**: raises problems for the epistemology of causation.

INUS condition

- Suppose, for example, that a lit match causes a forest fire. The lighting of the match, by itself, is not sufficient; many matches are lit without ensuing forest fires. The lit match is, however, a part of some constellation of conditions that are jointly sufficient for the fire. Moreover, given that this set of conditions occurred, rather than some other set sufficient for fire, the lighting of the match was necessary: fires do not occur in such circumstances when lit matches are not present.
- Epistasis, and gene-environment interaction.

Asymmetry

- If A causes B, then, typically, B will not also cause A.
- Causation is usually asymmetric.
- This poses a problem for regularity theories, for it seems quite plausible that if smoking is an INUS condition for lung cancer, then lung cancer will be an INUS condition for smoking.
- One way of enforcing the asymmetry of causation is to stipulate that causes precede their effects in time.

Spurious Regularities

- Suppose that a cause is regularly followed by two effects. For instance, a particular allele A is pleiotropic... It causes a disease trait, but also transcription of another gene B. B may be mistakenly thought to be causing the disease.
- B is also an INUS condition for disease state. But it's not a cause.
- Whenever the barometric pressure drops below a certain level, two things happen: First, the height of the column of mercury in a barometer drops . Shortly afterwards, a storm occurs. Then, it may well also be the case that whenever the column of mercury drops, there will be a storm.

Causes raise the probability of their effects.

- This can be expressed formally using the apparatus of conditional probability.
- Let A, B, C, $\ldots$ represent factors that potentially stand in causal relations.
- Let Pr be a probability function... such that $Pr(A)$ represents the empirical probability that factor A occurs or is instantiated.
- Let $Pr(B|A)$ represent the conditional probability of B, given A.

$$Pr(B|A) = \frac{Pr(A \wedge B)}{Pr(A)}.$$

- If $Pr(A)$ is 0, then the ratio in the definition of conditional probability is undefined. (There are other ways of handling this formally).
- "A raises the probability of B" is that

$$Pr(B|A) > Pr(B|\neg A).$$

PR Axiom

PR: A causes B if and only if$Pr(B|A) > Pr(B|\neg A)$.

Problems

- Probability-raising is **symmetric**: if $Pr(B|A) > P(B|\neg A)$, then $Pr(A|B) > P(A|\neg B)$. The causal relation, however, is typically asymmetric.
- Probability-raising has trouble with spurious correlations. If A and B are both caused by some third factor, C, then it may be that $Pr(B|A) > Pr(B|\neg A)$ even though A does not cause B.
- Those with yellow-stained fingers are more likely to suffer from lung cancer ... smoking tends to produce both effects. Because individuals with yellow-stained fingers are more likely to be smokers, they are also more likely to suffer from lung cancer.
- Intuitively, the way to address this problem is to require that causes raise the probabilities of their effects *ceteris paribus*.

Spurious Correlations

- **Screening off**: If $Pr(B|A \wedge C) = P(B|C)$, then C is said to screen A off from B.
- Equivalently $(A \perp B)|C$...
  $[Pr(A \wedge B|C) = Pr(A|C)Pr(B|C))]$ ... Intuitively, C renders A probabilistically irrelevant to B.
- To avoid the problem of spurious correlations, add a '*no screening off*' (NSO)

NSO

Factor A occurring at time t, is a cause of the later factor B if and only if:

$$Pr(B|A) > Pr(B|\neg A)$$

There is no factor C, occurring earlier than or simultaneously with A, that screens A off from B.

Yule-Simpson Effect

- *NSO does not suffice to resolve the problem of spurious correlations*
- Suppose, for example, that smoking is highly correlated with exercise: those who smoke are much more likely to exercise as well. Smoking is a cause of heart disease, but suppose that exercise is an even stronger preventative of heart disease. Then it may be that smokers are, over all, less likely to suffer from heart disease than non-smokers.
- $A \mapsto$ smoking, $C \mapsto$ exercise, and $B \mapsto$ heart disease, $Pr(B|A) < Pr(B|\neg A)$. Note, however, that if we conditionalize on whether one exercises or not, this inequality is reversed:

$$Pr(B|A \wedge C) > Pr(B|\neg A \wedge C)$$

$$Pr(B|A \wedge \neg C) > Pr(B|\neg A \wedge \neg C).$$

Test Situations

- Causes must raise the probability of their effects in test situations:

TS

TS: A causes B if $Pr(B|A \wedge T) > Pr(B|\neg A \wedge T)$ $\forall$ test situation T.

- A test situation is a conjunction of factors, which are “held fixed.” This suggests that in evaluating the causal relevance of A for B, we need to hold fixed other causes of B, either positively or negatively.

Notations

- We will use y to represent the trait under study; x to represent the genotype data; and z to represent covariates.
- **Example:** $y_i =$ the trait value for the ith individual in a sample, where $i = 1, \ldots, n$; and n is the total sample size.
- Similarly, x_{ij} is the genotype at the jth SNP for individual i, where $j = 1, \ldots, p$ is the total number of SNPs under study.
- Finally, z_{ik} is the value of the kth covariate for individual i, where $k = 1, \ldots, m$ and m is the total number of covariates.

Notations

- Thus, we will write $\mathbf{x} = (x_1, \ldots, x_n)^T$ to represent an $n \times 1$ vector of genotypes at a single site on the genome across all individuals in our sample.
- Thus, $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T$ will represent the genotypes at the jth site.
- Additionally, we will write $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ to denote the genotype data for the ith individual.

- Similarly, $\mathbf{y} = (y_1, \ldots, y_n)^T$ is a vector with its ith element corresponding to the trait for individual i. $\mathbf{y}$ can be quantitative; e.g., CD4 count or total cholesterol level.
- Finally, an $n \times p$ matrix of genotype variables is given by $\mathbf{X}$, with the (i, j)th element corresponding to the jth genotype for individual i.
- Similarly, $n \times m$ matrix $\mathbf{Z}$ denotes the entire set of covariates. (Multiple clinical, demographic and environmental variables, such as age, sex, weight and second hand smoke exposures.)

Explanatory Variables

- The combined matrix [**XZ**] represents the combined explanatory variables.
- Greek letters α, β, μ and θ are used to represent the model parameters. The parameters are unobservable quantities and are estimated from the data.

- The genotype for individual i at site j (denoted x_{ij}) is a categorical variable taking two or more levels.
- For instance, x_{ij} may be a three level factor variable taking three possible genotypes at a biallelic site: AA, Aa and aa, where A is the major haplotype and a is the minor haplotype.
- As another example, we may assign $x_{ij} = 0$ if the observed genotype is homozygous in major alleles, i.e., AA and $x_{ij} = 1$ otherwise.
- Sometimes, we will think of x_{ij} as an indicator for the presence of any variant alleles across multilocus genotype. Thus $x_{ij} = 0$ if the multilocus genotype is (AA, BB) and $x_{ij} = 1$ otherwise.

Difficulties

- Effects leading to spurious causal explanations:
- **Confounding and effect mediation**
- A *confounder* is a variable that is: (1) associated with the exposure (cause) variable; (2) independently associated with the outcome (effect) variable; and (3) not in the causal pathway between exposure and disease.
- *Example: Heavy alcohol consumption (the exposure) is associated with the total cholesterol level (the outcome). However smoking tends to be associated with heavy alcohol consumption. Smoking is also associated with high cholesterol levels among the individuals who are not heavy alcohol users.*
- A confounder is defined as a clinical or demographic variable that is associated with the genotype and the trait under investigation.

Difficulties

- A variable lying on the causal pathway between the predictor and the outcome is called an *effect mediator* or causal pathway variable.
- Genotype affects the trait through alteration of the mediator variable.
- A particular SNP variant may make an individual more likely to smoke and smoking would then cause cancer. Here smoking is an effect mediator.

Difficulties

- **Effect modification**: Effect of a predictor variable on the outcome depends on the level of another variable, called a *modifier*. Thus, the predictor variable and the modifier *interact* (in a statistical sense) in their association with the outcome.
- **Conditional Association**: The causal pathways between a predictor variable and the outcome depends on the values taken by the third modifying variable.

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:

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

Then by Bayes rule:

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

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

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

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 #5]