Computational Systems Biology: Biology X

Bud Mishra

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

L#6:(Mar-09-2010)
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

Bayesian Interpretation

- Probability $P(e) \mapsto$ our certainty about whether event e is true or false in the real world. (Given whatever information we have available.)
- "**Degree of Belief.**"
- More rigorously, we should write

  Conditional probability $P(e|L) \mapsto$ *Represents a degree of belief with respect to L — The background information upon which our belief is based.*

Probability as a Dynamic Entity

- We update the "degree of belief" as more data arrives: using **Bayes Theorem**:

$$P(e|D) = \frac{P(D|e)P(e)}{P(D)}.$$

  Posterior is proportional to the prior in a manner that depends on the data $P(D|e)/P(D)$.
- **Prior Probability**: $P(e)$ is one's belief in the event e before any data is observed.
- **Posterior Probability**: $P(e|D)$ is one's updated belief in e given the observed data.
- **Likelihood**: $P(D|e) \mapsto$ Probability of the data under the assumption e

Dynamics

- **Note:**

$$\begin{aligned} P(e|D_1, D_2) &= \frac{P(D_2|D_1, e)P(e|D_1)}{P(D_2|D_1)} \\ &= \frac{P(D_2|D_1, e)P(D_1|e)P(e)}{P(D_2D_1)} \end{aligned}$$

- **Further, note:** *The effects of prior diminish as the number of data points increase.*
- **The Law of Large Number:**

  With large number of data points, Bayesian and frequentist viewpoints become indistinguishable.

Parameter Estimation

- Functional form for a model M
  1. Model depends on some parameters Θ
  2. What is the best estimation of Θ?
- Typically the parameters Θ are a set of real-valued numbers
- Both prior $P(\Theta)$ and posterior $P(\Theta|D)$ are defining probability density functions.

MAP Method: Maximum A Posteriori

- Find the set of parameters Θ
  1. Maximizing the posterior $P(\Theta|D)$ or minimizing a score $-\log P(\Theta|D)$

$$\begin{aligned} E'(\Theta) &= -\log P(\Theta|D) \\ &= -\log P(D|\Theta) - \log P(\Theta) + \log P(D) \end{aligned}$$

  2. Same as minimizing

$$E(\Theta) = -\log P(D|\Theta) - \log P(\Theta)$$

  3. If prior $P(\Theta)$ is uniform over the entire parameter space (i.e., uninformative)

$$\min \arg_{\Theta} E_L(\Theta) = -\log P(D|\Theta).$$

Maximum Likelihood Solution

Information theory

- Information theory is based on probability theory (and statistics).
- **Basic concepts**: *Entropy* (the information in a random variable) and *Mutual Information* (the amount of information in common between two random variables).
- The most common unit of information is the **bit** (based log 2). Other units include the **nat**, and the **hartley**.

Entropy

- The entropy H of a discrete random variable X is a measure of the amount uncertainty associated with the value X.
- Suppose one transmits 1000 bits (0s and 1s). If these bits are known ahead of transmission (to be a certain value with absolute probability), logic dictates that no information has been transmitted. If, however, each is equally and independently likely to be 0 or 1, 1000 bits (in the information theoretic sense) have been transmitted.

Entropy

- Between these two extremes, information can be quantified as follows.
- If **X** is the set of all messages x that X could be, and $p(x)$ is the probability of X given x, then the **entropy of** X is defined as

$$H(x) = E_X[I(x)] = -\sum_{x \in X} p(x) \log p(x).$$

Here, $I(x)$ is the self-information, which is the entropy contribution of an individual message, and E_X is the expected value.

- An important property of entropy is that it is maximized when all the messages in the message space are equiprobable $p(x) = 1/n$, i.e., most unpredictable, in which case $H(X) = \log n$.
- The binary entropy function (for a random variable with two outcomes $\in \{0, 1\}$ or $\in \{H, T\}$:

$$H_b(p, q) = -p \log p - q \log q, \quad p + q = 1.$$

Joint entropy

- The joint entropy of two discrete random variables X and Y is merely the entropy of their pairing: $\langle X, Y \rangle$.
- Thus, if X and Y are independent, then their joint entropy is the sum of their individual entropies.

$$H(X, Y) = E_{X,Y}[-\log p(x, y)] = -\sum_{x,y} \log p(x, y).$$

- For example, if (X,Y) represents the position of a chess piece Ñ X the row and Y the column, then the joint entropy of the row of the piece and the column of the piece will be the entropy of the position of the piece.

Conditional Entropy or Equivocation

- The conditional entropy or conditional uncertainty of X given random variable Y (also called the equivocation of X about Y) is the average conditional entropy over Y:

$$\begin{aligned} H(X|Y) &= E_Y[H(X|y)] \\ &= -\sum_{y\in Y} p(y) \sum_{x\in X} p(x|y) \log p(x|y) \\ &= -\sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(y)} \end{aligned}$$

- A basic property of this form of conditional entropy is that:

$$H(X|Y) = H(X,Y) - H(Y).$$

Mutual Information (Transinformation)

- Mutual information measures the amount of information that can be obtained about one random variable by observing another.
- The mutual information of X relative to Y is given by:

$$I(X;Y) = E_{X,Y}[SI(x,y)] = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}.$$

where SI (**Specific mutual Information**) is the pointwise mutual information.

- A basic property of the mutual information is that

$$I(X;Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y) = I(Y;X).$$

That is, knowing Y, we can save an average of $I(X;Y)$ bits in encoding X compared to not knowing Y. Note that mutual information is **symmetric**.

- It is important in communication where it can be used to maximize the amount of information shared between sent and received signals.

Kullback-Leibler Divergence (Information Gain)

- The Kullback-Leibler divergence (or information divergence, information gain, or relative entropy) is a way of comparing two distributions: a "true" probability distribution $p(X)$, and an arbitrary probability distribution $q(X)$.

$$\begin{aligned} D_{KL}(p(X)\|q(X)) &= \sum_{x\in X} p(x)\log\frac{p(x)}{q(x)} \\ &= \sum_{x\in X}[-p(x)\log q(x)] - [-p(x)\log p(x)] \end{aligned}$$

- If we compress data in a manner that assumes $q(X)$ is the distribution underlying some data, when, in reality, $p(X)$ is the correct distribution, the Kullback-Leibler divergence is the number of average additional bits per datum necessary for compression.
- Although it is sometimes used as a 'distance metric,' it is not a true metric since it is not symmetric and does not satisfy the triangle inequality (making it a semi-quasimetric).

- Mutual information can be expressed as the average Kullback-Leibler divergence (information gain) of the posterior probability distribution of X given the value of Y to the prior distribution on X:

$$\begin{aligned} I(X;Y) &= E_{p(Y)}[D_{KL}(p(X|Y=y)\|p(X)] \\ &= D_{KL}(p(X,Y)\|p(X)p(Y)). \end{aligned}$$

  In other words, mutual information $I(X,Y)$ is a measure of how much, on the average, the probability distribution on X will change if we are given the value of Y. This is often recalculated as the divergence from the product of the marginal distributions to the actual joint distribution.
- Mutual information is closely related to the log-likelihood ratio test in the context of contingency tables and the multinomial distribution and to Pearson's χ^2 test.

Source theory

- Any process that generates successive messages can be considered a source of information.
- A memoryless source is one in which each message is an independent identically-distributed random variable, whereas the properties of ergodicity and stationarity impose more general constraints. All such sources are stochastic.

Information Rate

- **Rate** Information rate is the average entropy per symbol. For memoryless sources, this is merely the entropy of each symbol, while, in the case of a stationary stochastic process, it is

$$r = \lim_{n \to \infty} H(X_n | X_{n-1}, X_{n-2} \ldots)$$

- In general (e.g., nonstationary), it is defined as

$$r = \lim_{n \to \infty} \frac{1}{n} H(X_n, X_{n-1}, X_{n-2} \ldots)$$

- In information theory, one may thus speak of the “rate” or “entropy” of a language.

Rate Distortion Theory

- $R(D) =$ Minimum achievable rate under a given constraint on the expected distortion.
- $X =$ random variable; $T =$ alphabet for a compressed representation.
- If $x \in X$ is represented by $t \in T$, there is a distortion $d(x,t)$

$$
\begin{aligned}
R(D) &= \min_{\{p(t|x):\langle d(x,t)\rangle \leq D\}} I(T,X). \\
\langle d(x,t)\rangle &= \sum_{x,t} p(x,t)d(x,t) \\
&= \sum_{x,t} p(x)p(t|x)d(x,t)
\end{aligned}
$$

- Introduce a Lagrange multiplier parameter β and
- Solve the following **variational problem**

$$\mathcal{L}_{min}[p(t|x)] = I(T;X) + \beta\langle d(x,t)\rangle_{p(x)p(t|x)}.$$

- We need

$$\frac{\partial \mathcal{L}}{\partial p(t|x)} = 0.$$

Since

$$\mathcal{L} = \sum_x p(x) \sum_t p(t|x) \log \frac{p(t|x)}{p(t)} + \beta \sum_x p(x) \sum_t p(t|x) d(x,t),$$

we have

$$p(x) \left[\log \frac{p(t|x)}{p(t)} + \beta d(x,t) \right] = 0.$$

$$\Rightarrow \frac{p(t|x)}{p(t)} \propto e^{-\beta d(x,t)}.$$

Summary

- In summary,

$$p(t|x) = \frac{p(t)}{Z(x,\beta)} e^{-\beta d(x,t)} \qquad p(t) = \sum_x p(x)p(t|x).$$

  $Z(x,\beta) = \sum_t p(t) \exp[-\beta d(x,t)]$ is a Partition Function.
- The Lagrange parameter in this case is positive; It is determined by the upper bound on distortion:

$$\frac{\partial R}{\partial D} = -\beta.$$

Outline

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, ... 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.

Contingency Table

- Three genotypes for a given SNP: *homozygous wildtype aa*, *heterozygous Aa* and *homozygous rare/ AA*.
- The data can be represented by the 2×3 contingency table. See below.
- **Odds Ratio**: *Ratio of the odds of disease among the exposed to the odds of disease among the unexposed.*
- Genotype $\equiv$ *exposure*

	Gen: *aa*	Gen: *Aa*	Gen: *AA*	
Dis: $+$	n_{11}	n_{12}	n_{13}	$n_{1\cdot}$
Dis: $-$	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	n

Odds Ratio

- Odds Ratio:

$$OR = \frac{Pr(D^+|E^+)/[1 - Pr(D^+|E^+)]}{Pr(D^+|E^-)/[1 - Pr(D^+|E^-)]}$$

- In genetics, we calculate the OR for each genotype with relation to the homozygous wildtype genotype, AA.

$$OR_{aa,AA} = \frac{(n_{11}/n_{\cdot 1})/(n_{21}/n_{\cdot 1})}{(n_{13}/n_{\cdot 3})/(n_{23}/n_{\cdot 3})} = \frac{n_{11}n_{23}}{n_{21}n_{13}}$$

Dichotomized Contingency Table

- Dichotomizing genotype priors
- $E^+ = \{Aa, aa\}$ and $E^- = \{AA\}$
- The data can be represented by the 2×2 contingency table. See below.

	Gen: $\{aa, Aa\}$	Gen: AA	
Dis: $+$	n_{11}	n_{12}	$n_{1.}$
Dis: $-$	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	n

Odds Ratio

- Odds Ratio:

$$\widehat{OR} = \frac{(n_{11}/n_{.1})/(n_{21}/n_{.1})}{(n_{12}/n_{.2})/(n_{22}/n_{.2})} = \frac{n_{11}n_{22}}{n_{21}n_{12}}$$

Fisher's Exact Test

- What is the probability of getting the 2×2 table by *chance*

$$p = \binom{n_{1\cdot}}{n_{11}}\binom{n_{2\cdot}}{n_{21}}/\binom{n}{n_{\cdot 1}} = \frac{n_{1\cdot}!n_{2\cdot}!n_{\cdot 1}!n_{\cdot 2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}$$

- This formula gives the exact probability of observing this particular arrangement of the data, assuming the given marginal totals, on the null hypothesis that te two categories of genotypes are equally likely to have the disease.
- In other words, the probability p indicates how well the data fit the hypothesis: "the single or double mutation ($A \mapsto a$) cause the disease."
- If $p \ll \theta$ (i.e., the probability is very very small), we can reject the null hypothesis, and conclude that "the mutation ($A \mapsto a$) has a necessary causal role in the disease."

Fisher's exact test

- Fisher's exact test is a statistical test used to determine if there are nonrandom associations between two categorical variables. — E.g., Genotypes and a Categorical Trait.
- Let there exist two such variables X and Y, with m and n observed states, respectively.
- Now form an $m \times n$ matrix in which the entries a_{ij} represent the number of observations in which $x = i$ and $y = j$. Calculate the row and column sums R_i and C_j, respectively, and the total sum

$$N \sum_i R_i = \sum_j C_j.$$

of the matrix.

- Then calculate the conditional probability of getting the actual matrix given the particular row and column sums, given by

$$P_{cutoff} = \frac{(R_1!R_2!\cdots R_m!)(C_1!C_2!\cdots C_n!)}{N!\prod_{ij} a_{ij}!}$$

  which is a multivariate generalization of the **hypergeometric probability function**.
- Now find all possible matrices of nonnegative integers consistent with the row and column sums R_i and C_j. For each one, calculate the associated conditional probability using this formula, where the sum of these probabilities must be 1.

- To compute the P-value of the test, the tables must then be ordered by some criterion that measures dependence, and those tables that represent equal or greater deviation from independence than the observed table are the ones whose probabilities are added together.
- There are a variety of criteria that can be used to measure dependence. In the 2×2 case, which is the one Fisher looked at when he developed the exact test, either the Pearson chi-square or the difference in proportions (which are equivalent) is typically used.
- Other measures of association, such as the likelihood-ratio-test, -squared, or any of the other measures typically used for association in contingency tables, can also be used.

[End of Lecture #6]