Bioinformatics Chapter 3.

Probabilistic Modeling and Inference: Thermodynamic Approach

Outline

- The Single Die Model with Sequence Data
  - Lagrangian Multiplier
  - Dirichlet Prior
  - ML, MAP, MP estimates
- The Single Die Model with Counts Data
  - MaxEnt Distribution
  - The entropic prior (concentration phenomenon)
- Statistical Mechanics
  - Boltzmann-Gibbs Distribution
  - Free Energy
  - Thermodynamics
  - Latent Variables Case

The Single Die Model with Sequence Data (1)

- Data(D)
  - A single sequence of length N from 4 sided die {A,C,G,T}
- Model(M)
  - four parameters with $p_A + p_C + p_G + p_T = 1$
- Likelihood (ML): $P(D \mid M) = \prod_X p_X^{n_X} = p_A^{n_A} p_C^{n_C} p_G^{n_G} p_T^{n_T}$
- Negative log posterior (MAP):

$$-\log P(M \mid D) = -\sum_X n_X \log p_X - \log P(M) + \log P(D)$$

- Cf. Multinomial Data (Counts Data)

The Single Die Model with Sequence Data (2)

- Lagrangian multiplier: $\nabla f(x_0) = \lambda \nabla g(x_0)$
- By optimizing negative log likelihood using Lagrange multiplier constrained with probability condition, parameters can be estimated (ML estimate).

$$L = -\sum_X n_X \log p_X - \lambda(1 - \sum_X p_X);$$

$$\lambda = N, \ : p_X^* = n_X / N \qquad for\ all\ X \in A$$

- The negative log likelihood per letter becomes the entropy of optimal distribution P^* as N becomes larger.

$$-\frac{1}{N}\sum_{X \in A} n_X \log \frac{n_X}{N} \rightarrow -\sum_{X \in A} p_X^* \log p_X^* = H(P^*)$$

The Single Die Model with Sequence Data (3)

- Dirichlet density: $D_{\alpha Q}(P) = \frac{\Gamma(\alpha)}{\prod_i \Gamma(\alpha q_i)} \prod_{i=1}^{K} p_i^{\alpha q_i - 1} = \prod_{i=1}^{K} \frac{p_i^{\alpha q_i - 1}}{Z(i)}$

where

$P = (p_1, \ldots, p_K), Q = (q_1, \ldots, q_K);$

$\sum_i q_i = 1,$ (α *indicates peakness.*)

$$E(p_i) = q_i / \sum q_i = q_i,$$
$$Var(p_i) = q_i(1 - q_i / \sum q_i) / (\sum q_i (\sum q_i + 1))$$
$$= q_i (1 - q_i) / (\alpha + 1)$$
$$Cov(p_i, p_j) = -q_i q_j / (\alpha + 1)$$

- Dirichlet can be seen as a multivariate generalization of the beta distribution.
- When $q_1 = q_2 = \ldots = q_K = 1/\alpha = 1/K,$ Dirichlet becomes a multivariate uniform.
- Under uniform prior: MAP estimation is identical to ML estimation.
  But when n is not observed, the corresponding parameter is estimated as 0 in ML → Apply MAP estimation with Dirichlet prior

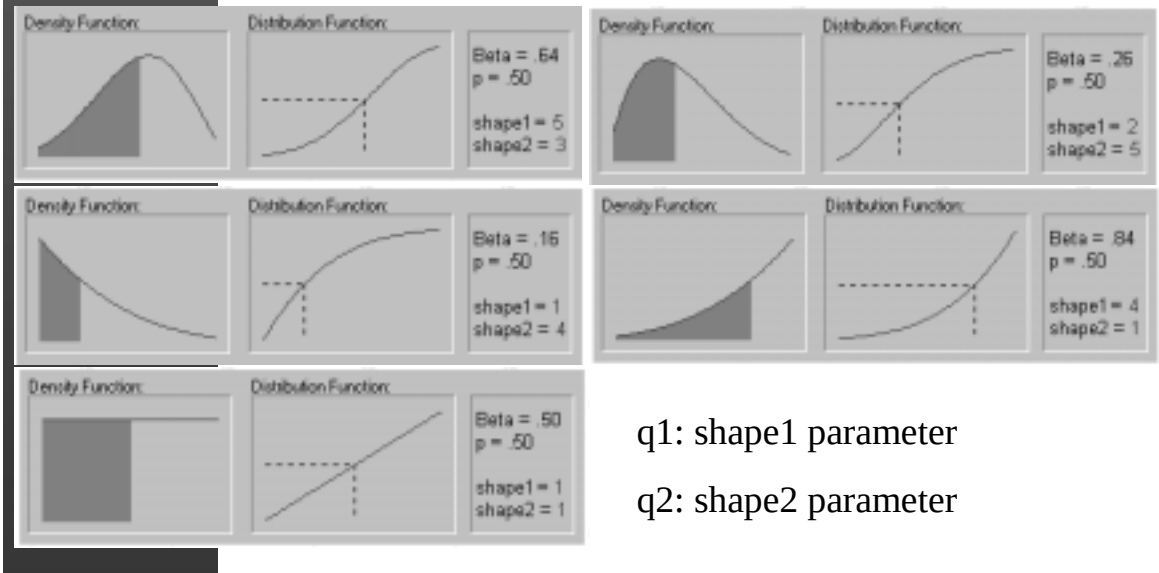
The Single Die Model with Sequence Data (4)

- Dirichlet density: a generalization of beta density.

(x = (x , 1-x) ~ beta (q_1 , q_2) = Dirichlet (q_1 , q_2))

$0 \le X \le 1$

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$



q1: shape1 parameter

q2: shape2 parameter

The Single Die Model with Sequence Data (5)

- MAP under a Dirichlet prior w.r.t. sequence data:

$$-\log P(M \mid D) = -\log\{P(D \mid M)P(M)/P(D)\}$$
$$= -\sum_X (n_X + \alpha q_X - 1)\log p_X + \log Z + \log P(D)$$

- MAP estimate: $$p_X^* = \frac{n_X + \alpha q_X - 1}{N + \alpha - |K|},$$
- The effect is same as adding pseudo-counts to the observed counts.
- The posterior $P(M \mid D)$ is also a Dirichlet $D_{\beta R}$ with

$$\beta = N + \alpha \quad , \quad r_X = (n_X + \alpha q_X)/(N + \alpha).$$

- MP(mean posterior) estimate:

$$p_X^{**} = \frac{n_X + \alpha q_X}{N + \alpha}$$

The Single Die Model with Sequence Data (6)

- MP estimate minimizes the expected relative entropy distance (check!) ($P|D$: posterior) $f(P^*) \equiv E(H(P, P^*))$ (Expectation w.r.t. posterior)
- Alternative ways of estimation:
  - ♦ Use evidence (below) to fq_Xl (mean posterior).
  - ♦ Define a prior (q_X (hyper parameter).
  - ♦ Study MAP, MP estimates when the prior is a mixture of Dirichlet. (In general when the prior is a mixture of conjugate, the posterior is also a mixture of conjugate.)
- Calculation of evidence: ratio of normalizing constants (prior/posterior)

$$P(D) = \int P(D \mid w)P(w)dw = \int_{\sum p_X = 1} \prod_X p_X^{n_X} \prod_X p_X^{\alpha q_X - 1} \frac{\Gamma(\alpha)}{\Gamma(q_X)} dp_X$$

$$= \frac{\Gamma(\alpha)}{\prod_X \Gamma(\alpha q_X)} \frac{\prod_X \Gamma(\beta r_X)}{\Gamma(\beta)}, \quad (\beta = N + \alpha, \beta r_X = n_X + \alpha q_X, \sum_i r_i = 1)$$

The Single Die Model with Counts Data (1)

- Maximum entropy distributions
  - ♦ On $E=[a,b]$: Uniform [a,b]
  - ♦ On real line with constrain $mean=\mu$, $\mathrm{var}=\sigma^2$: Normal

- Concentration phenomenon: Almost all outcomes that satisfy a given constraint have frequencies very close to the corresponding maximum entropy distribution as the sample size is increasing.

(Janes, Probability theory: the logic of science(1996)

$D=\{n_X\}=(n_1,\ \mathrm{L}\ ,n_K)$ ∨ustl.edu/etj/prob.html)

- Likelihood:

$$P(D\mid M)=P(\{n_X\}\mid\{p_X\})=\binom{N}{n_{X_1}\ \mathrm{L}\ n_{X_A}}\prod_{X\in A}p_X^{n_X},\ (\textstyle\sum_X n_X=N)$$

The Single Die Model with Counts Data (2)

- Stirling's approximation: $\ln n!\ =n\ln n-n+\ln\sqrt{2\pi n}+O(\frac{1}{n})$
- By the concentration phenomenon, uniform P (named as entropic prior) induces an entropic distribution to the data.
- Cf. For a Dirichlet prior - Dirichlet posterior.
- The entropic distribution is not the conjugate of a multinomial: the posterior is not entropic nor Dirichlet. (?)
- Derivation of an entropic distribution:

$$\ln P(D\mid P)=\ln\frac{N!}{n_1!n_2!\ \mathrm{L}\ n_X!}(\frac{1}{K})^N=N\ln N-N+\ln\sqrt{2\pi N}$$

$$+O(\frac{1}{N})-\sum_i\ \{n_i\ln n_i-n_i+\ln\sqrt{2\pi n_i}+O(\frac{1}{n_i})\}-N\ln K$$

$$\cong N\cdot H(n_i/N)\ \ (\textit{Thus Maximizing the Entropy wrt. } n_i)$$

Boltzmann-Gibbs Distribution (1)

- Question: Given the observation D , that is, the average of f (a function of states), what can we say about the state distribution P ?
- Answer: Maximum entropy principle (least assumptions, most spread out). Choose P that has the highest entropy with constraint D (average of f)

$$L = -\sum_s p_s \log p_s - \lambda(\sum_s p_s f(s) - D) - \mu(\sum_s p_s - 1) \quad D = \sum_s p_s f(s)$$

$$p_s^*(\lambda) = \frac{e^{-\lambda f(s)}}{Z(\lambda)} \ , \ Z(\lambda) = \sum_s e^{-\lambda f(s)} \ , \ \lambda = 1/kT$$

  - In statistical mechanics, the normalizing factor Z is called the partition function and P becomes the Maxwell-Boltzmann distribution.
  - The Lagrange multiplier λ is related to the temperature T of the system and are entirely determined by $\sum_s \frac{e^{-\lambda f(s)}}{Z(\lambda)} f(s) = D$ observation D.

Boltzmann-Gibbs Distribution (2)

- Notes
  1. Any distribution can be represented as Boltzmann-Gibbs Distribution if the energy function f is proportional to negative log P, and λ is fixed (e.g. λ= 1) .
  2. The Lagrange multiplier λ (or Temperature) is determined by D.
- Weaknesses of the Boltzmann-Gibbs distribution
  1. The prior is not explicit.(How to incorporate additional information of parameters)
  2. The probabilistic model is not explicit. (e.g. How to calculate the likelihood)
  3. The justification for the use of MaxEnt is weak (Connections between MaxEnt, ML, MAP is not clear.)

Various Distributions

- Maxwell-Boltzmann distribution (classical): identical, distinguishable particles. (ex) ideal gas

$$g(E) = \frac{1}{Ae^{E/kT}}$$

- Bose-Einstein distribution (quantum): identical, indistinguishable particles with integer spin (bosons).

  (ex) thermal radiation

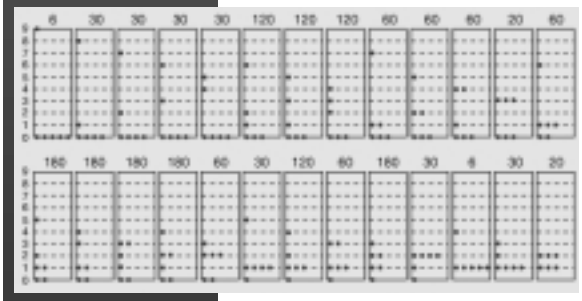
$$g(E) = \frac{1}{Ae^{E/kT} - 1}$$

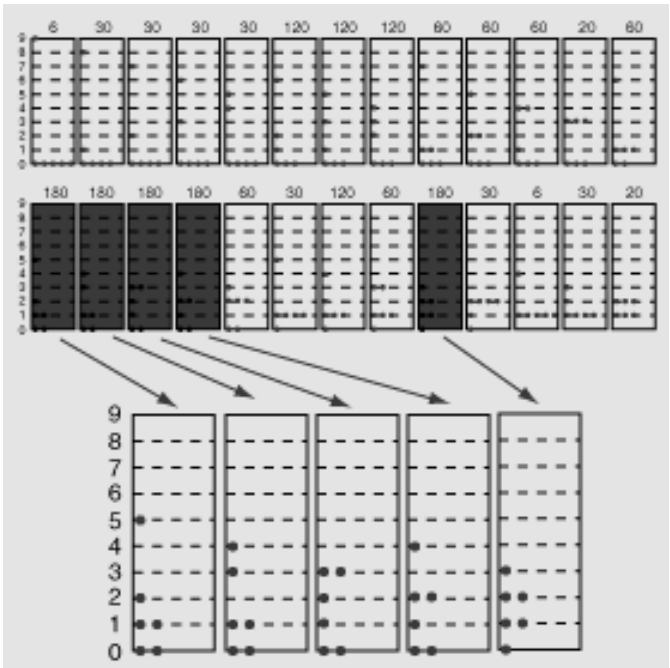
- Fermi-Dirac distribution function (quantum): identical indistinguishable particles with half-integer spin (fermions).

  (ex) electrons in a metal

$$g(E) = \frac{1}{Ae^{E/kT} + 1}$$

- Distributing 9 energy levels for 6 particles
  - ♦ Maxwell-Boltzmann distribution: 2002 different states
  - ♦ Einstein-Bose distribution: 26 different states
  - ♦ Fermi-Dirac disribution: 5 states (Max 2 spin states for each energy label - Pauli exclusion principle)





*** The Pauli exclusion principle is an assertion that no two electrons in an atom can be at the same time in the same state or configuration**

Bayesian Derivation of the Boltzmann-Gibbs Distribution

- Assume that the system is observed over a certain period N and parameterize the model using the count n_s . And estimate the parameters n_s by MAP with data *D* and entropic distribution.

$$D = (\sum_s n_s f(s)) / N \qquad (\neq \sum_s p_s f(s)).$$

$$L_{MAP} = -\sum_s \frac{n_s}{N} \log \frac{n_s}{N} - \lambda(\sum_s \frac{n_s}{N} f(s) - D) - \mu(\sum_s n_s - 1)$$

- The Boltzmann-Gibbs distribution corresponds to a first step of Bayesian inference by MAP ($P(D|n_s)$), with an entropic distribution.

Thermodynamic Limit and Phase Transitions

- Partition function: contains most of the information of the system.
  - Mean and variance of *f*: $E(f) = D = -\frac{\partial}{\partial \lambda} \log Z(\lambda), \quad Var(f) = -\frac{\partial^2}{\partial \lambda^2} \log Z(\lambda))$
  - Entropy of the Boltzmann-Gibbs distribution

$$H(P^*) = -E \log P^* = \log Z(\lambda) + \lambda E(f)$$

- Phase transition: abrupt change of the system as T or λ varies.
  - First order phase transition at λ_C : *E(f)* is discontinuous at λ_C.
  - Second order phase transition at λ_C : *E(f)* is continuous but *Var(f)* is discontinuous at λ_C.

The Free Energy (1)

- Gibb's Free Energy: $F = F(f, \lambda) = F(\lambda) \equiv -\frac{1}{\lambda} \log Z(\lambda),$

$$H(P^*) = -\lambda F(\lambda) + \lambda E(f),$$

$$F(\lambda) = E(f) - \frac{1}{\lambda} H(P^*) = F(f, P^*, \lambda) \text{ (alternative definition of free energy)}$$

$$F(f, Q, \lambda) = E_Q(f) - \frac{1}{\lambda} H(Q) \text{ (extension to any distribution } Q)$$

  - f can be chosen as a negative log probability which might have an important statistical application.
  - Boltzmann-Gibbs distribution minimizes the free energy (see next).

The Free Energy (2)

- Comparison of two distributions Q and R:

$$F(Q, \lambda) - F(R, \lambda) = \sum_s [Q(s) - R(s)][f(s) + \frac{1}{\lambda} \log R(s)] + \frac{1}{\lambda} H(Q, R)$$

  - If the form of energy function f is taken to be a negative log likelihood
    $f(s) = -\log R(s)$ for a distribution R over states and $\lambda = 1$
    1. $F(Q,1) - F(R,1) = H(Q, R)$
    2. Boltzmann-Gibbs distribution becomes R with 0 free energy.
  - Since relative entropy is nonnegative in (1), Boltzmann-Gibbs distribution minimizes the free energy.
  - There is nothing special for λ=1

$$f(s) = -\log R(s)/\lambda, \quad \rightarrow \; F(Q, \lambda) - F(R, \lambda) = H(Q, R)/\lambda$$

The Free Energy (3)

- Gibbs free energy $Free\,Energy = Enthalpy - T \cdot Entropy$, $(G = H - TS)$
  - Entropy: 1. Degree of disorder, Goes from orderly state to disorderly state (Increasing). / 2. The amount of energy not available for doing a work.
  - $G < 0$: Spontaneous reaction, $G > 0$: not spontaneous reaction.
  - Conservation of energy principle : Energy can't be created nor destroyed

    (The total energy of an isolated system remains constant.)

The Free Energy (4)

- First law of thermodynamics: $\Delta U = Q + W$ (U: internal (microscopic) energy of the system , Q; heat added to the system, W: work done on the system, W = PΔV). -- You can't get more energy than you put in.
- Second law of thermodynamics: The entropy of the universe increases. (Time's arrow, tendency, not prediction)
- Third law of thermodynamics: The entropy of a perfect crystal at 0 temperature is zero.

The Free Energy (5)

- Entropy measures the number of ways that the energy of a system can be distributed among the motions of its microparticles (its atoms, molecules, and/or ions).
- Gibbs Free energy:

  $\Delta G = \Delta H - T\Delta S,$

  $(H : \textit{Heat needed (absorbed) for the process},\ S : \textit{Entropy of the system})$

  $\Delta G < 0$: spontaneous process. $\Delta G > 0$: not spontaneous

- Enthalpy: Energy needed for the process to go.
- Free Energy: Net energy to run the process.

The Hidden Variables Case (1)

- Let *D*: Observable variable, *H*: Latent variable (state), *w*: parameter.

(Q) What can be said about the distribution on the state after observing *D* ?

- The Boltzmann-Gibbs distribution

  Suppose the states of the system are assumed by *H* and define

  $f(H) = -\log P(D, H \mid w)$ then at λ=1

  - Partition function , $Z(1) = p(D \mid w)$
  - Boltzmann-Gibbs distribution is given by the posterior,

    $$P^* = P^*(H,1) = P(H \mid D, w)$$

  - Free energy is given by the negative log likelihood,

    $$\mathrm{F}(P^*,1) = -\log P(D \mid w)$$

The Hidden Variables Case (2)

- Notes: 1. The difference in free energies between P^* and any Q is given by the KL divergence: $\mathrm{F}(Q,1) - \mathrm{F}(P^*,1) = \mathrm{H}(Q, P^*)$

  2. The corresponding data likelihood (for free energy) becomes

$$-\log p(D \mid w) = E(f) - H(P^*)$$

- Variational method:

  When P^* (posterior) and Ef (expectations) are difficult to calculate, one can maximize the log likelihood based on a different family of distribution Q for which the calculations are more tractable instead of P^*.

$$\log P(D \mid w) = -\mathrm{F}(Q,1) + \mathrm{H}(Q, P^*)$$

Discussions

- Data likelihood maximization : ML, MAP
- Entropy maximization : MaxEnt
- MDL
- Free energy minimization: maximize data and entropy at the same time. (relations with Boltzmann-Gibbs / Einstein-Bose / Fermi-Dirac Solutions)
- The role of temperature for identifying phase changes.