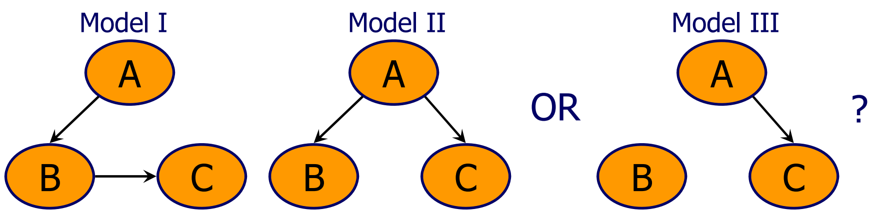
Maximum Likelihood Estimation & Expectation Maximization

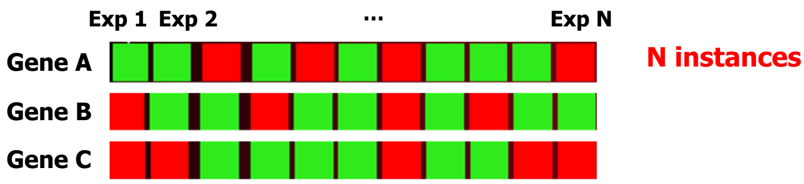
Lectures 3 – Oct 5, 2011
CSE 527 Computational Biology, Fall 2011

Instructor: Su-In Lee
TA: Christopher Miles

Monday & Wednesday 12:00-1:20
Johnson Hall (JHN) 022

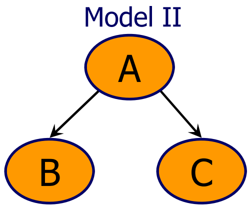
Review



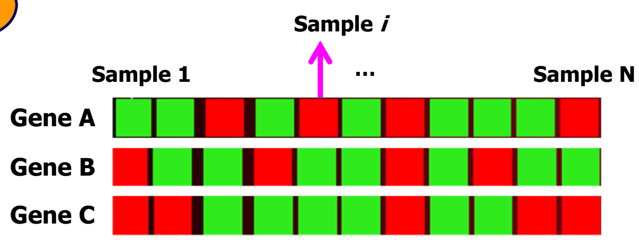


- Model selection: argmax_x P(model x is true | Data)
- How to compute P(model x is true | Data)?
  - Compute P(Data | model x is true)
  - Use Bayes rule

Computing P(Data | model II is true)



- P(A,B,C | model II is true)= ?
  - P(A)P(B|A)P(C|A)
  - P(A=**high**)P(B=**low**|A=**high**) P(C=**low**|A=**high**)



Outline

- Probabilistic models in biology
  - Model selection problem
- Mathematical foundations
- Bayesian networks
- Learning from data
  - Maximum likelihood estimation
  - Maximum a posteriori (MAP)
  - Expectation and maximization

Parameter Estimation

- Assumptions
  - Fixed network structure
  - Fully observed instances of the network variables: D={d[1],…,d[M]}
  - Maximum likelihood estimation (MLE)!

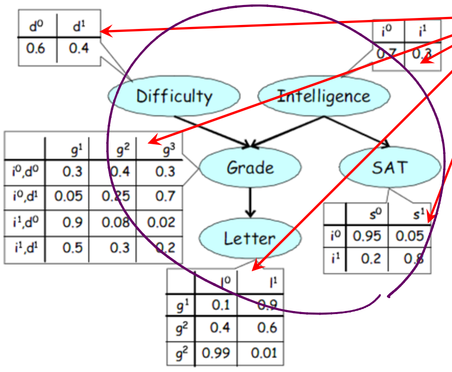
For example, {i0,d1,g1,l0,s0}

from data

M=100

M=1000

"Parameters" of the Bayesian network



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^2	0.99	0.01

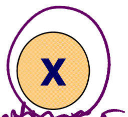
from Koller & Friedman

The *Thumbtack* example

- Parameter learning for a single variable.

X=head	X=tail
θ	$1-\theta$

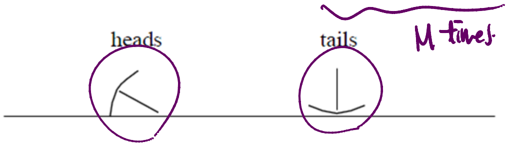
- Variable
  - X: an outcome of a thumbtack toss
  - Val(X) = {head, tail}



X[i] ∈ {head, tail}

- Data
  - A set of thumbtack tosses: x[1],…x[M]

M times



Maximum likelihood estimation

- Say that P(x=head) = Θ, P(x=tail) = 1-Θ
  - P(HHTTHHH…<M_h heads, M_t tails>; Θ) = $\theta\theta(1-\theta)\cdots$

    $= \theta^{M_h}(1-\theta)^{M_t}$
- **Definition:** The likelihood function
  - L(Θ : D) = P(D; Θ) $= \theta^{M_h}(1-\theta)^{M_t}$

    $P(D|\theta)$, θ is a random $P(D:\theta)$ $\frac{P(D|\theta)}{P(\theta)}$
- Maximum likelihood estimation (MLE)
  - Given data D=HHTTHHH…<M_h heads, M_t tails>, find Θ that maximizes the likelihood function L(Θ : D).

    $\theta^{M_h}(1-\theta)^{M_t}$ $P(D:\theta)$

    $[0, 1]$

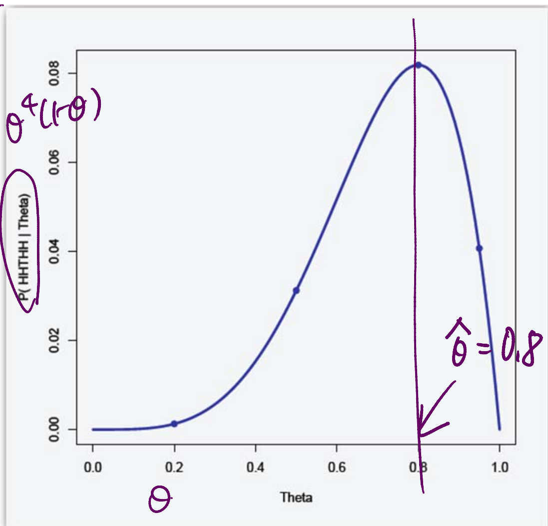
7

Likelihood function

Probability of HHTHH, given P(H) = θ:

$\frac{4}{5}$

θ	θ⁴(1-θ)
0.2	0.0013
0.5	0.0313
0.8	0.0819
0.95	0.0407



8

MLE for the *Thumbtack* problem

- Given data D=HHTTHHH…<M_h heads, M_t tails>,
  - MLE solution $\hat{\Theta}^*$ = $M_h / (M_h+M_t)$.
- Proof:

$$L(\theta:D) = p(D:\theta) = \theta^{M_h}(1-\theta)^{M_t}$$

$L(\theta:D)$ $\quad l(\theta:D)$

$\rightarrow \log(\cdot)$

$$\log L(\theta:D) = l(\theta:D) = \log p(D:\theta) = M_h \log\theta + M_t \log(1-\theta)$$

$$\frac{\partial}{\partial\theta} l(\theta:D)\Big|_{\theta=\hat\theta} = 0$$

$$\frac{M_h}{\hat\theta} + \frac{(-1)M_t}{1-\hat\theta} = 0$$

9

Continuous Space

- Assuming sample $x_1, x_2, \ldots, x_n$ is from a parametric distribution $f(x|\Theta)$, estimate Θ.

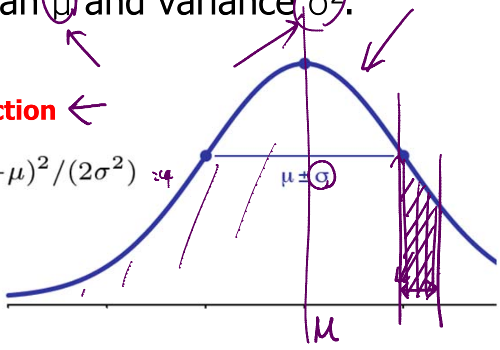
X

- Say that the n samples are from a normal distribution with mean μ and variance σ^2.

Probability density function

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

$$\theta = (\mu, \sigma^2)$$



Continuous Space (cont.)

- Let $\Theta_1=\mu$, $\Theta_2=\sigma^2$

$\mu \quad \sigma^2$

$$L(\theta_1,\theta_2 : x_1, x_2,\ldots,x_n) = p(x_1 \cdots x_n : \theta_1, \theta_2) = \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right\}$$

$$\log L(\theta_1,\theta_2 : x_1, x_2,\ldots,x_n) = \sum_{i=1}^{n} \left[\log \frac{1}{\sqrt{2\pi\sigma^2}} + \log e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right]$$

$$- \frac{(x_i-\mu)^2}{2\sigma^2}$$

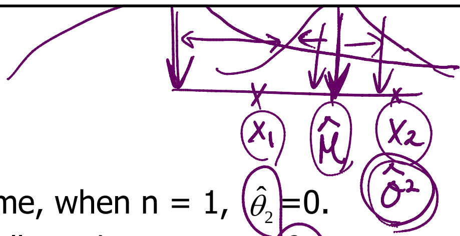
$$\frac{\partial}{\partial \theta_1} \log L(\theta_1,\theta_2 : x_1, x_2,\ldots,x_n) = \frac{\partial}{\partial \mu} \left[\sum_{i=1}^{n} -\frac{(x_i-\mu)^2}{2\sigma^2} \right] = 0$$

$$\boxed{\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i}$$

$$\frac{\partial}{\partial \theta_2} \log L(\theta_1,\theta_2 : x_1, x_2,\ldots,x_n) = \frac{\partial}{\partial \sigma^2} \left[\quad \right] = 0$$

$$\boxed{\hat{\sigma}^2 = \frac{1}{n} \sum_i (x_i - \hat{\mu})^2}$$

Any Drawback?



- Is it biased?
  - Is it? Yes. As an extreme, when n = 1, $\hat{\theta}_2$=0.
  - The MLE $\hat{\theta}_2$ systematically underestimates θ_2.

Why? A bit harder to see, but think about $n = 2$. Then θ_1 is exactly between the two sample points, the position that exactly minimizes the expression for $\hat{\theta}_2$. Any other choices for (θ_1, θ_2) make the likelihood of the observed data slightly lower. But it's actually pretty unlikely that two sample points would be chosen exactly equidistant, and on opposite sides of the mean, so the MLE $\hat{\theta}_2$ systematically underestimates θ_2.

Maximum A Posteriori

$\theta \to 0.5$

HHHHH $\boxed{\hat{\theta} = 1}$

- Incorporating priors. How?

  MLE: $p(D \mid \theta)$ $\boxed{\log P(D \mid \theta)}$

  MAP: $p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{p(D)}$ $\boxed{\log P(D \mid \theta)} + \log P(\theta)$

- MLE vs MAP estimation

MLE for General Problems

- Learning problem setting $X_1 \cdots X_n$
  - A set of random variables X from unknown distribution P*
  - Training data D = M instances of X: {d[1],…,d[M]}

    H or T $\to (0, 0, \cdots)$

- A *parametric model* P(X; Θ) (a 'legal' distribution)

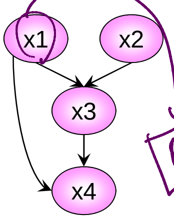
- Define the likelihood function:
  - L(Θ : D) = $p(D : \theta)$

  $\hat{\theta} = \arg\max_{\theta} P(D : \theta)$

- Maximum likelihood estimation
  - Choose parameters ~~Θ*~~ $\hat{\theta}$ that satisfy: $\frac{\partial}{\partial \theta} \log P(D : \theta) \Big|_{\theta = \hat{\theta}} = 0$

MLE for Bayesian Networks

Structure G



$P_G = P(x1,x2,x3,x4) = \prod_i P(X_i \mid Pa X_i)$

$= P(x1)\ P(x2)\ P(x3|x1,x2)\ P(x4|x1,x3)$

More generally? $\quad P_G = \prod_i P(x_i \mid \mathbf{pa}_i)$

$\boxed{\theta_{x_1} \mid 1-\theta_{x_1}}$

Parameters $\boldsymbol{\theta}$ — Θ_{x1}, Θ_{x2}, $\Theta_{x3|x1,x2}$, $\Theta_{x4|x1,x3}$ (more generally $\Theta_{xi|pai}$)

$(X1[1], X2[1], X3[1], X4[1])$

Given D: x[1],…x[m]…,x[M], estimate θ. (x1[m],x2[m],x3[m],x4[m])

$L(\theta : D) = P(X[1] \ldots X[M] \mid \theta_{x_1} \ldots \theta_{x_4|x_1,x_3}) = \prod_{m=1}^{M} P(X[m] \mid \theta_{x_1} \ldots \theta_{x_4})$

- Likelihood decomposition:

$P(X[m]) = \prod_{m=1}^{M} \{ P(X_1[m])\, P(X_2[m])\, P(X_3[m] \mid X_1[m] X_2[m])\, P(X_4[m] \mid \ldots) \}$

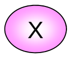
$= P(X_1 = X_1[m]) = \theta \ldots$

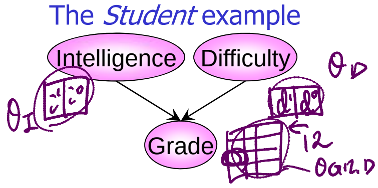
$\theta_{x_1} \quad \theta_{x_2} \quad \theta_{x_3|x_1,x_2} \quad \theta_{x_4|x_1,x_3}$

- The local likelihood function for X_i is: $X_i : \prod_{m=1}^{M} P(X_i[m])$

Bayesian Network with Table CPDs

The *Thumbtack* example vs The *Student* example





	The *Thumbtack* example	The *Student* example
Joint distribution	P(X)	P(I,D,G) = $P(I) \cdot P(D) P(G \mid I, D)$
Parameters	θ	θ_I, θ_D, $\theta_{G\|I,D}$
Data	D: {H…x[m]…T}	D: $\{(i^1,d^1,g^1)\ldots(i[m],d[m],g[m])\ldots\}$
Likelihood function L(θ:D) = P(D;θ)	$\theta^{Mh}(1-\theta)^{Mt}$	$\theta_{I=i^1}^{M_{I=i^1}}\ \theta_{I=i^0}^{M_{I=i^0}} \cdot \theta_{D=d^1}^{M_{D=d^1}} \ldots$
MLE solution	$\hat{\theta} = \frac{Mh}{Mh + Mt}$	$\theta_{G=A \mid I=i^1, D=d^1} = \frac{M_{G=A, I=i^1, D=d^1}}{M_{I=i^1, D=d^1}}$

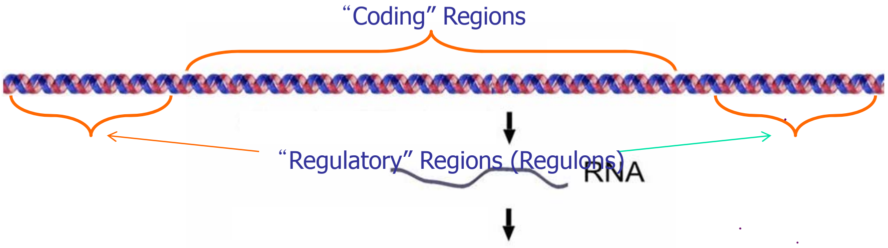
Maximum Likelihood Estimation Review

- Find parameter estimates which make observed data most likely $\hat{\theta} = \arg\max_{\theta} P(D|\theta)$
- General approach, as long as tractable likelihood function exists $\frac{\partial}{\partial\theta} \log P(D|\theta)$
- Can use all available information $\leftarrow D$

17

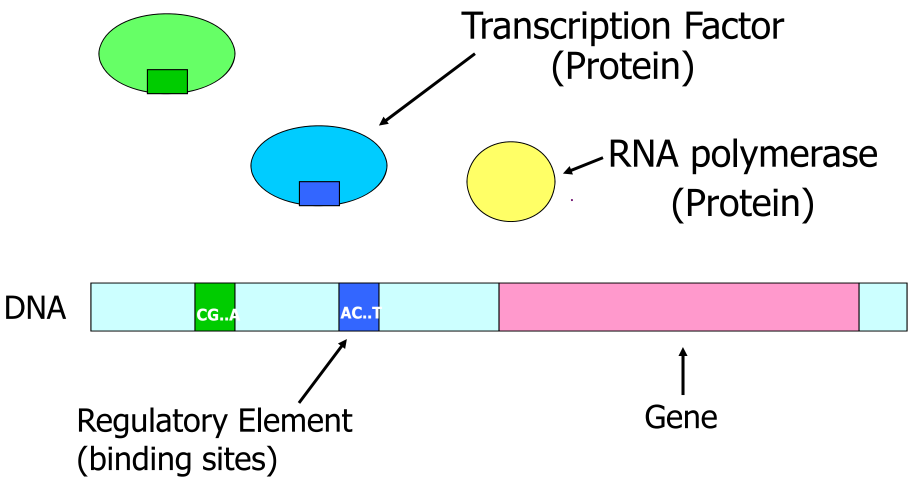
Example – Gene Expression

- Instruction for making the proteins
- Instruction for when and where to make them



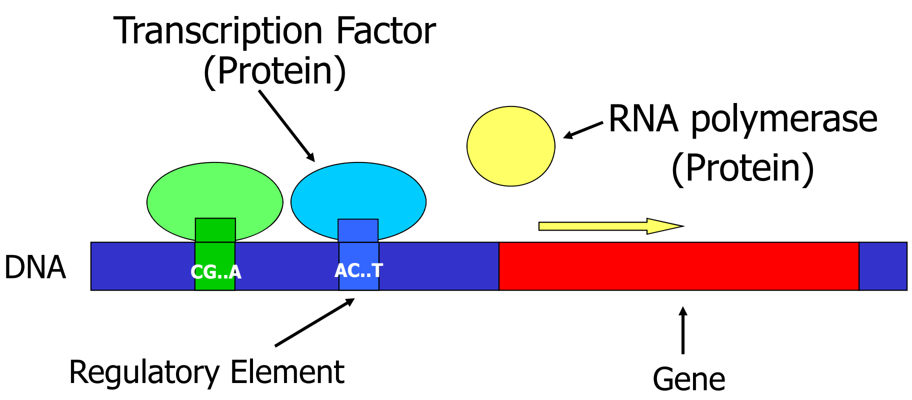
- Regulatory regions contain "binding sites" (6-20 bp).
- "Binding sites" attract a special class of proteins, known as "transcription factors".
- Bound transcription factors can initiate transcription (making RNA).
- Proteins that inhibit transcription can also be bound to their binding sites.

Regulation of Genes



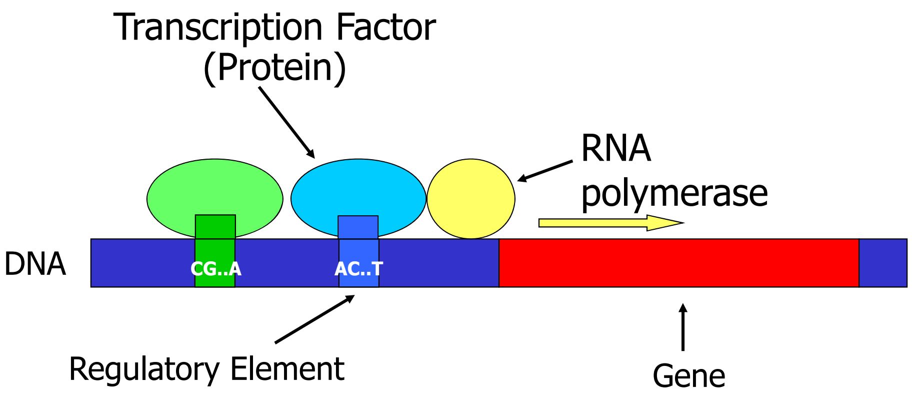
source: M. Tompa, U. of Washington

Regulation of Genes



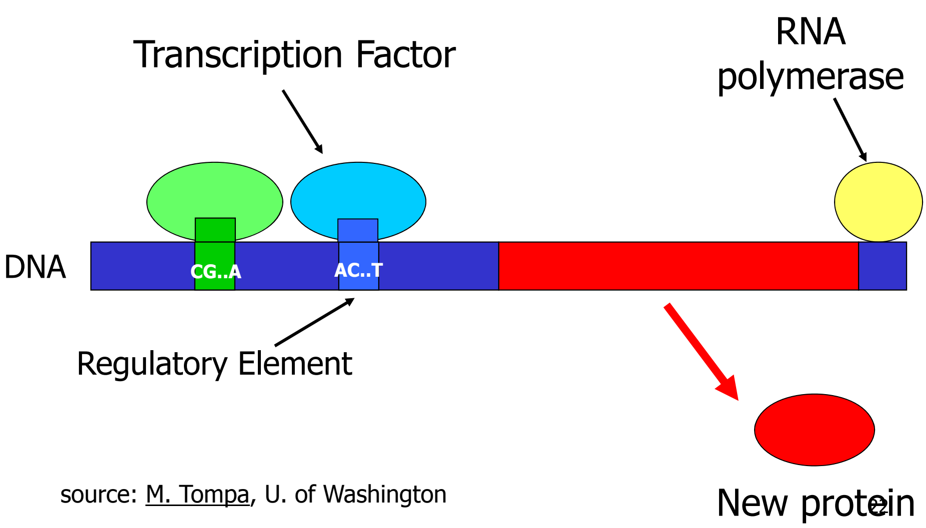
source: M. Tompa, U. of Washington

Regulation of Genes



source: M. Tompa, U. of Washington

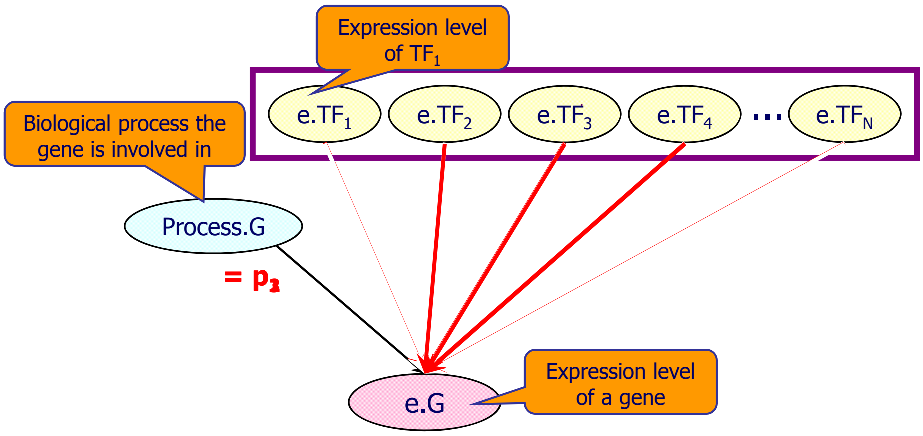
Regulation of Genes



source: M. Tompa, U. of Washington

The *Gene regulation* example

- What determines the expression level of a gene?
- What are observed and hidden variables?
  - e.G, e.TF's: observed; Process.G: hidden variables ⇒ want to infer!



Not All Data Are Perfect

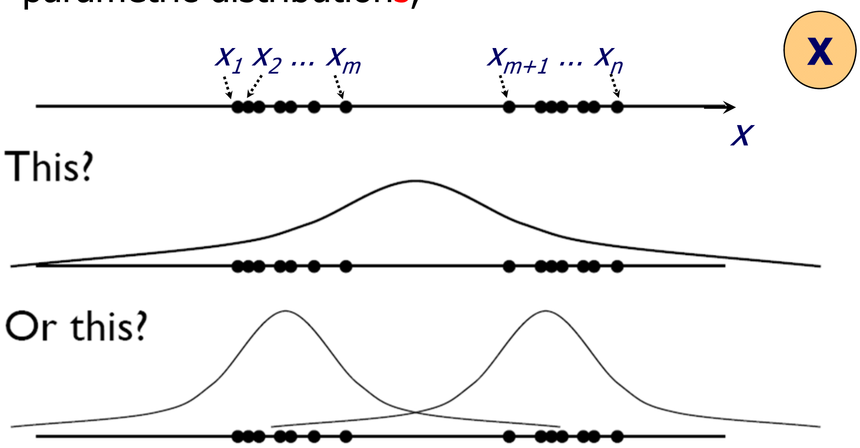
- Most MLE problems are simple to solve with complete data.
- Available data are "incomplete" in some way.

Outline

- Learning from data
  - Maximum likelihood estimation (MLE)
  - Maximum a posteriori (MAP)
  - Expectation-maximization (EM) algorithm

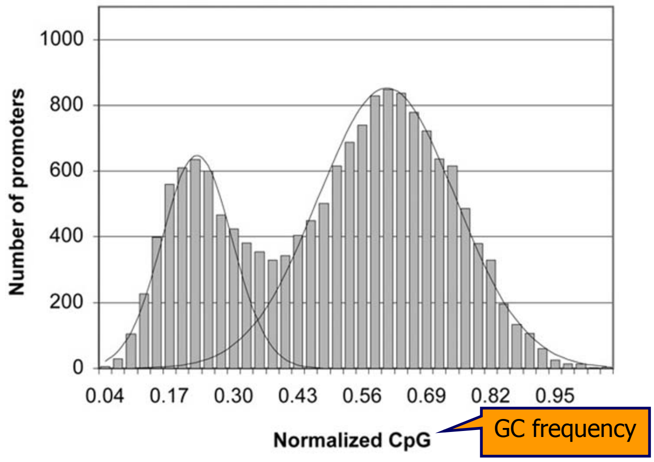
Continuous Space Revisited...

- Assuming sample $x_1, x_2, \ldots, x_n$ is from a mixture of parametric distributions,



A Real Example

- CpG content of human gene promoters



"A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters" Saxonov, Berg, and Brutlag, PNAS 2006;103:1412-1417

Acknowledgement

- Profs Daphne Koller & Nir Friedman, "Probabilistic Graphical Models"

- Prof Larry Ruzo, CSE 527, Autumn 2009