Maximum Likelihood Estimation 
& Expectation Maximization

Review

- Model selection: \( \text{argmax}_x \ P(\text{model } x \text{ is true } | \text{ Data}) \)
- How to compute \( P(\text{model } x \text{ is true } | \text{ Data}) \)?
  - Compute \( P(\text{ Data } | \text{ model } x \text{ is true }) \)
  - Use Bayes rule
Computing $P(\text{Data} \mid \text{model II is true})$

- $P(A,B,C \mid \text{model II is true}) = ?$
  - $P(A)P(B \mid A)P(C \mid A)$
  - $P(A=\text{high})P(B=\text{low} \mid A=\text{high})$
  - $P(C=\text{low} \mid A=\text{high})$

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: \(D=\{d[1], \ldots, d[M]\}\)
  - Maximum likelihood estimation (MLE)

For example, \((i0, d1, g1, l0, s0)\)

“Parameters” of the Bayesian network from Koller & Friedman

The **Thumbtack** example

- Parameter learning for a single variable.

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

- Data
  - A set of thumbtack tosses: \(x[1], \ldots, x[M]\)
Maximum likelihood estimation

- Say that $P(x=\text{head}) = \Theta$, $P(x=\text{tail}) = 1-\Theta$
  - $P(\text{HHTTHHH...<M h heads, M t tails>} ; \Theta) = \Theta^M (1-\Theta)^M$

- **Definition:** The likelihood function
  - $L(\Theta : D) = P(D; \Theta) = \Theta^M (1-\Theta)^M$

- Maximum likelihood estimation (MLE)
  - Given data $D=\text{HHTTHHH...<M h heads, M t tails>}$, find $\Theta$ that maximizes the likelihood function $L(\Theta : D)$. 

Likelihood function

<table>
<thead>
<tr>
<th>Probability of HHTTHH given $P(H) = \Theta$</th>
<th>0.2</th>
<th>0.5</th>
<th>0.8</th>
<th>0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Theta$</td>
<td>0.0013</td>
<td>0.0313</td>
<td>0.0819</td>
<td>0.0407</td>
</tr>
</tbody>
</table>

![Graph showing the likelihood function with a peak at $\hat{\Theta} = 0.8$]
MLE for the *Thumbtack* problem

- Given data \( D = \text{HHTTHHH} \ldots <M_h \text{ heads}, M_t \text{ tails}> \),
- MLE solution \( \hat{\Theta} = \frac{M_h}{M_h + M_t} \).

**Proof:**

\[
\log(L(\Theta; D)) = \log(p(D; \Theta)) = \Theta^M_h (1-\Theta)^M_t
\]

\[
\frac{\partial}{\partial \Theta} \log(L(\Theta; D)) = \frac{M_h \log \Theta + M_t \log (1-\Theta)}{\Theta^M_h (1-\Theta)^M_t} = 0
\]

Continous Space

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

- Say that the \( n \) samples are from a normal distribution with mean \( \mu \) and variance \( \sigma^2 \).

**Probability density function**

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

\[
\Theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}
\]
Continuous Space (cont.)

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

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

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

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

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

Any Drawback?

- Is it biased? Yes. As an extreme, when $n = 1$, $\hat{\mu} = 0$.
- The MLE $\hat{\theta}_2$ systematically underestimates $\theta_2$.

Why? A bit harder to see, but think about $n = 2$. Then $\theta_1$ is exactly between the two sample points, the position that exactly minimizes the expression for $\hat{\theta}_1$. Any other choices for $(\theta', \theta_2)$ 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 $\theta_2$. 
Maximum A Posteriori

- Incorporating priors. How?
  - MLE: \( P(D|\theta) \frac{\log P(D|\theta)}{P(\theta)} \)
  - MAP: \( P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \)

- MLE vs MAP estimation

MLE for General Problems

- Learning problem setting
  - A set of random variables \( X \) from unknown distribution \( P^* \)
  - Training data \( D = M \) instances of \( X: \{d[1],...,d[M]\} \)

- A parametric model \( P(X; \Theta) \) (a 'legal' distribution)

- Define the likelihood function:
  - \( L(\theta : D) = P(D:\theta) \)
  - Choose parameters \( \hat{\theta} \) that satisfy:
    \( \hat{\theta} = \arg\max_{\theta} P(D:\theta) \)
MLE for Bayesian Networks

\[ P_G = P(x_1, x_2, x_3, x_4) = \prod P(x_i | pa_i) \]

Parameters \( \theta \)

- \( \theta_{x_1}, \theta_{x_2}, \theta_{x_3|x_1, x_2}, \theta_{x_4|x_3} \)
- (more generally \( \theta_{x_i|pa_i} \))

Given \( D: x[1], \ldots, x[m], \ldots, x[M] \), estimate \( \theta \).

Likelihood decomposition:

\[ L(\theta; D) = \prod_{i=1}^{m} P(x_i | pa_i) \]

The local likelihood function for \( x_i \) is:

\[ P(x_{1:m}, x_{2:m}, x_{3:m}, x_{4:m}) \]

Bayesian Network with Table CPDs

The Thumbtack example

- Joint distribution \( P(X) \)
- Parameters \( \theta \)
- Data \( D: \{H...x[m]...T\} \)
- Likelihood function \( L(\theta; D) = P(D; \theta) \)
- MLE solution \( \hat{\theta} = \frac{M_h}{M_h + M_t} \)

The Student example

- Joint distribution \( P(I, D, G) = P(I) P(D; I) P(G; I, D) \)
- Parameters \( \theta_I, \theta_D, \theta_{G|I,D} \)
- Data \( D: \{(i_1, d_1, g_1) \ldots (i[m], d[m], g[m]) \ldots\} \)
- Likelihood function \( L(\theta; D) = P(D; \theta) \)
- MLE solution

\[ \hat{\theta}_{G=A|I=i, D=d} = \frac{M_{G=A, I=i, D=d}}{M_{I=i, D=d}} \]
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

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

Transcription Factor (Protein)

RNA polymerase (Protein)

DNA

Regulatory Element (binding sites)

Gene

source: M. Tompa, U. of Washington
Regulation of Genes

RNA polymerase

Transcription Factor (Protein)

Regulatory Element

DNA

Gene

source: M. Tompa, U. of Washington

Regulation of Genes

RNA polymerase

Transcription Factor

Regulatory Element

DNA

New protein

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 \( \Rightarrow \) want to infer!

![Diagram showing gene regulation and expression levels](image)

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,

This?

Or this?
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