Tutorial on Expectation-Maximization Algorithm

Long Qiu

`qiul@comp.nus.edu.sg`

School of Computing

National University of Singapore

Outline

- Maximum Likelihood Estimation

Outline

- Maximum Likelihood Estimation
- Example #1 (from [1]): MLE with missing data

Outline

- Maximum Likelihood Estimation
- Example #1 (from [1]): MLE with missing data
- EM Algorithm

Outline

- Maximum Likelihood Estimation
- Example #1 (from [1]): MLE with missing data
- EM Algorithm
- Example #2 (from [1]): MLE with a hidden variable

Outline

- Maximum Likelihood Estimation
- Example #1 (from [1]): MLE with missing data
- EM Algorithm
- Example #2 (from [1]): MLE with a hidden variable
- Discussion

Maximum Likelihood Estimation

- Parameter Estimation:
  Given
  - a set of observed data $\{x_1, x_2, ...\}$, and
  - a proposed model $p(x \mid \theta)$

  to find
  - the value for one or more parameters θ that maximize the likelihood:

$$L(\theta; x) = p(x_1, x_2, ... \mid \theta).$$

Example #1

A	B
0	0
0	0
0	0
0	?
0	1
1	0
1	1
1	1

Estimate parameters in joint distribution: θ

	$\overline{A}$	A
$\overline{B}$	$p(\overline{AB})$	$p(A\overline{B})$
B	$p(\overline{A}B)$	$p(AB)$

Example #1

A	B
0	0
0	0
0	0
0	1
1	0
1	1
1	1

Estimate parameters in joint distribution: θ

	$\overline{A}$	A
$\overline{B}$	3/7	1/7
B	1/7	2/7

- Opt 1: Ignore the incomplete data

Example #1

A	B
0	0
0	0
0	0
0	0
0	1
1	0
1	1
1	1

Estimate parameters in joint distribution: θ

	$\overline{A}$	A
$\overline{B}$	4/8	1/8
B	1/8	2/8

- Opt 1: Ignore the incomplete data
- Opt 2: Fill in a best guessed value

Example #1

A	B
0	0
0	0
0	0
0	?
0	1
1	0
1	1
1	1

Estimate parameters in joint distribution: θ

	$\overline{A}$	A
$\overline{B}$	0.25	0.25
B	0.25	0.25

- Opt 1: Ignore the incomplete data
- Opt 2: Fill in a best guessed value
- Opt 3: Fill in with distribution

Example #1

A	B
0	0
0	0
0	0
0	0.5, 0 0.5, 1
0	1
1	0
1	1
1	1

Estimate parameters in joint distribution: θ

	$\overline{A}$	A
$\overline{B}$	0.25	0.25
B	0.25	0.25

- Opt 1: Ignore the incomplete data
- Opt 2: Fill in a best guessed value
- Opt 3: Fill in with distribution

Example #1

A	B
0	0
0	0
0	0
0	0.5, 0 0.5, 1
0	1
1	0
1	1
1	1

Estimate parameters in joint distribution: θ

	$\overline{A}$	A
$\overline{B}$	3.5/8	1/8
B	1.5/8	2/8

- Opt 1: Ignore the incomplete data
- Opt 2: Fill in a best guessed value
- Opt 3: Fill in with distribution

Example #1

A	B
0	0
0	0
0	0
0	0.7, 0 0.3, 1
0	1
1	0
1	1
1	1

Estimate parameters in joint distribution: θ

	$\overline{A}$	A
$\overline{B}$	3.5/8	1/8
B	1.5/8	2/8

- Opt 1: Ignore the incomplete data
- Opt 2: Fill in a best guessed value
- Opt 3: Fill in with distribution

Problem Setting

- observation: $x_i = (x_{i1}, x_{i2}, ..., x_{iN}), i = 1...k$
- modeling:
  - observable variables: $X = (X_1, X_2, ..., X_N)$
  - hidden/latent variables: $Z = (Z_1, Z_2, ..., Z_M)$
  - probability model: $p(x, z \mid \theta)$
- parameter estimation:
  - find the values for θ that maximize the *log likelihood* $l(\theta; x, z) \triangleq \log p(x, z \mid \theta)$

Why observability matters?

- If Z is observable, we can maximize the *complete log likelihood*:

$$l_c(\theta; x, z) \triangleq \log p(x, z \mid \theta);$$

- If Z is hidden, we need to maximize the *incomplete log likelihood*:

$$l(\theta; x) \triangleq \log p(x \mid \theta) = log \sum_z p(x, z \mid \theta).$$

Auxiliary Function $\mathcal{L}(q, \theta)$

an arbitrary distribution: $q(z \mid x)$

$$
\begin{aligned}
l(\theta; x) &= \log p(x \mid \theta) \\
&= \log \sum_z p(x, z \mid \theta) \\
&= \log \sum_z q(z \mid x) \frac{p(x, z \mid \theta)}{q(z \mid x)}
\end{aligned}
$$

Auxiliary Function $\mathcal{L}(q, \theta)$

an arbitrary distribution: $q(z \mid x)$

$$\begin{aligned} l(\theta; x) &= \log p(x \mid \theta) \\ &= \log \sum_z p(x, z \mid \theta) \\ &= \log \sum_z q(z \mid x) \frac{p(x, z \mid \theta)}{q(z \mid x)} \end{aligned}$$

Jensen's Inequality: For a real continuous concave function $f(x)$:

$$\sum_x p(x) f(x) \leqq f\left(\sum_x p(x) x\right)$$

Auxiliary Function $\mathcal{L}(q, \theta)$

an arbitrary distribution: $q(z \mid x)$

$$
\begin{aligned}
l(\theta; x) &= \log p(x \mid \theta) \\
&= \log \sum_z p(x, z \mid \theta) \\
&= \log \sum_z q(z \mid x) \frac{p(x, z \mid \theta)}{q(z \mid x)} \\
&\geqq \sum_z q(z \mid x) \log \frac{p(x, z \mid \theta)}{q(z \mid x)} \\
&\triangleq \mathcal{L}(q, \theta)
\end{aligned}
$$

Auxiliary Function $\mathcal{L}(q,\theta)$

an arbitrary distribution: $q(z \mid x)$

$$
\begin{aligned}
l(\theta; x) &= \log p(x \mid \theta) \\
&= \log \sum_z p(x, z \mid \theta) \\
&= \log \sum_z q(z \mid x) \frac{p(x, z \mid \theta)}{q(z \mid x)} \\
&\geqq \sum_z q(z \mid x) \log \frac{p(x, z \mid \theta)}{q(z \mid x)} \\
&\triangleq \mathcal{L}(q, \theta)
\end{aligned}
$$

$\mathcal{L}(q,\theta)$ is a lower bound of $l(\theta; x)$

Expectation Step

With θ fixed, find q that maximizes $\mathcal{L}(q, \theta)$.

$$\mathcal{L}(q, \theta) = \sum_z q(z \mid x) \log \frac{p(x, z \mid \theta)}{q(z \mid x)}$$

Expectation Step

With θ fixed, find q that maximizes $\mathcal{L}(q, \theta)$.

$$\mathcal{L}(q, \theta) = \sum_z q(z \mid x) \log \frac{p(x, z \mid \theta)}{q(z \mid x)}$$

$$q^{(t+1)}(z \mid x) \Downarrow$$

$$\Downarrow p(z \mid x, \theta^{(t)})$$

Expectation Step

With θ fixed, find q that maximizes $\mathcal{L}(q, \theta)$.

$$\mathcal{L}(q, \theta) = \sum_z q(z \mid x) \log \frac{p(x, z \mid \theta)}{q(z \mid x)}$$

$$q^{(t+1)}(z \mid x)$$

$$\Downarrow p(z \mid x, \theta^{(t)})$$

$$\begin{aligned}
\mathcal{L}(p(z \mid x, \theta^{(t)}), \theta^{(t)}) &= \sum_z p(z \mid x, \theta^{(t)}) \log \frac{p(x, z \mid \theta^{(t)})}{p(z \mid x, \theta^{(t)})} \\
&= \sum_z p(z \mid x, \theta^{(t)}) \log p(x \mid \theta^{(t)}) \\
&= \log p(x \mid \theta^{(t)}) \\
&= l(\theta^{(t)}; x)
\end{aligned}$$

Maximization Step

With q fixed, find θ that maximizes $\mathcal{L}(q, \theta)$.

$$\begin{aligned}
\mathcal{L}(q, \theta) &= \sum_z q(z \mid x) \log \frac{p(x, z \mid \theta)}{q(z \mid x)} \\
&= \sum_z q(z \mid x) \log p(x, z \mid \theta) - \sum_z q(z \mid x) \log q(z \mid x) \\
&= \langle l_c(\theta; x, z) \rangle_q - \sum_z q(z \mid x) \log q(z \mid x)
\end{aligned}$$

Maximization Step

With q fixed, find θ that maximizes $\mathcal{L}(q, \theta)$.

$$\begin{aligned}
\mathcal{L}(q, \theta) &= \sum_z q(z \mid x) \log \frac{p(x, z \mid \theta)}{q(z \mid x)} \\
&= \sum_z q(z \mid x) \log p(x, z \mid \theta) - \sum_z q(z \mid x) \log q(z \mid x) \\
&= \langle l_c(\theta; x, z) \rangle_q - \sum_z q(z \mid x) \log q(z \mid x)
\end{aligned}$$

Just need to maximize $\langle l_c(\theta; x, z) \rangle_q$, which is no more difficult than maximizing $\log p(x, z \mid \theta)$, the complete log likelihood.

Maximization Step

With q fixed, find θ that maximizes $\mathcal{L}(q, \theta)$.

$$\begin{aligned}
\mathcal{L}(q, \theta) &= \sum_z q(z \mid x) \log \frac{p(x, z \mid \theta)}{q(z \mid x)} \\
&= \sum_z q(z \mid x) \log p(x, z \mid \theta) - \sum_z q(z \mid x) \log q(z \mid x) \\
&= \langle l_c(\theta; x, z) \rangle_q - \sum_z q(z \mid x) \log q(z \mid x)
\end{aligned}$$

Just need to maximize $\langle l_c(\theta; x, z) \rangle_q$, which is no more difficult than maximizing $\log p(x, z \mid \theta)$, the complete log likelihood.
Now we have improved θ: 1)it leads to higher likelihood; and 2) q is ready to be further, better, estimated.

EM: an Iterative Process

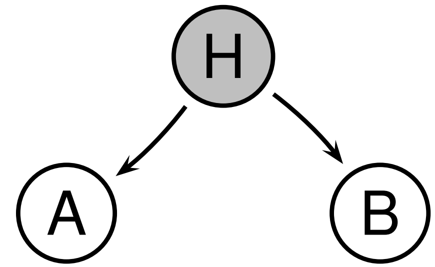
- Start with initial θ^0
- Repeat until convergence

E-Step: $p(z \mid x, \theta^{(t)}) \to q^{(t+1)}(= \mathrm{argmax}_q\, \mathcal{L}(q, \theta^{(t)}))$

M-Step: $\theta^{(t+1)} = \mathrm{argmax}_\theta\, \mathcal{L}(q^{(t+1)}, \theta)$

Example #2: MLE with hidden variables

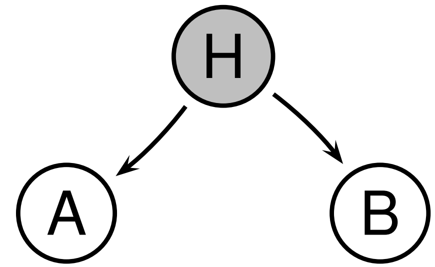
A	B	#	$Pr(H^m \mid D^m, \theta)$
0	0	6	
0	1	1	
1	0	1	
1	1	4	



- Parameters to estimate:

$$Pr(H) = ?$$
$$Pr(A \mid H) = ?$$
$$Pr(A \mid \overline{H}) = ?$$
$$Pr(B \mid H) = ?$$
$$Pr(B \mid \overline{H}) = ?$$

Example #2: MLE with hidden variables

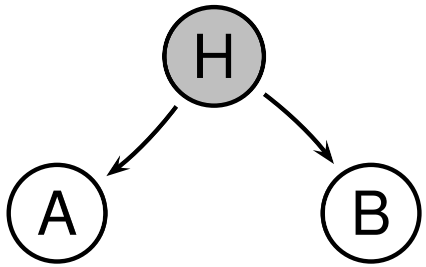


A	B	#	$Pr(H^m \mid D^m, \theta)$
0	0	6	
0	1	1	
1	0	1	
1	1	4	

- Parameters initialized:

$$Pr(H) = 0.4$$
$$Pr(A \mid H) = 0.55$$
$$Pr(A \mid \overline{H}) = 0.61$$
$$Pr(B \mid H) = 0.43$$
$$Pr(B \mid \overline{H}) = 0.52$$

Example #2: MLE with hidden variables



A	B	#	$Pr(H^m \mid D^m, \theta)$
0	0	6	.48
0	1	1	.39
1	0	1	.42
1	1	4	.33

- Parameters initialized:

$$\begin{aligned} Pr(H) &= 0.4 \\ Pr(A \mid H) &= 0.55 \\ Pr(A \mid \overline{H}) &= 0.61 \\ Pr(B \mid H) &= 0.43 \\ Pr(B \mid \overline{H}) &= 0.52 \end{aligned}$$

- Iteration 1: E-Step

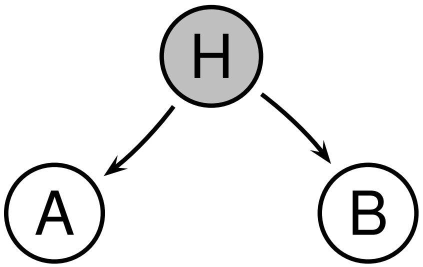
$$\begin{aligned} Pr(H \mid A, B) &= \frac{Pr(A, B, H)}{Pr(A, B)} \\ &= \frac{Pr(A, B \mid H) Pr(H)}{Pr(A, B)} \\ &= \frac{Pr(A \mid H) Pr(B \mid H) Pr(H)}{Pr(A, B \mid H) Pr(H) + Pr(A, B \mid \overline{H})(1 - Pr(H))} \end{aligned}$$

Example #2: MLE with hidden variables

A	B	#	$Pr(H^m \mid D^m, \theta)$
0	0	6	.48
0	1	1	.39
1	0	1	.42
1	1	4	.33

- Iteration 1: E-Step

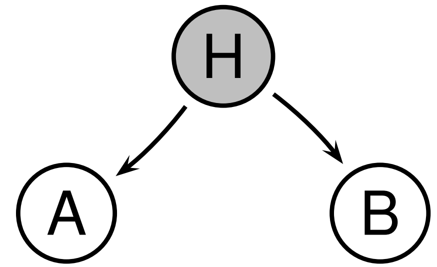
$$Pr(H \mid A, B) = \frac{Pr(A,B,H)}{Pr(A,B)} = \frac{Pr(A,B \mid H)Pr(H)}{Pr(A,B)} = \frac{Pr(A \mid H)Pr(B \mid H)Pr(H)}{Pr(A,B \mid H)Pr(H) + Pr(A,B \mid \overline{H})(1 - Pr(H))}$$



- Iteration 1: M-step (parameters re-estimated):

$$Pr(H) = 0.42$$
$$Pr(A \mid H) = 0.35$$
$$Pr(A \mid \overline{H}) = 0.46$$
$$Pr(B \mid H) = 0.34$$
$$Pr(B \mid \overline{H}) = 0.47$$

Example #2: MLE with hidden variables



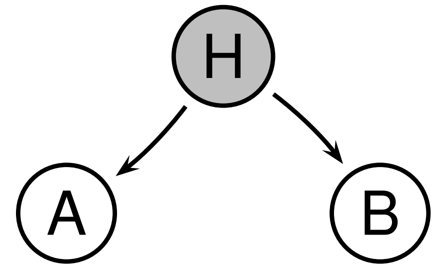
A	B	#	$Pr(H^m \mid D^m, \theta)$
0	0	6	.52
0	1	1	.39
1	0	1	.39
1	1	4	.28

- Iteration 2: E-Step

- Iteration 1:
  M-step
  (parameters re-estimated):

$$
\begin{aligned}
Pr(H) &= 0.42 \\
Pr(A \mid H) &= 0.35 \\
Pr(A \mid \overline{H}) &= 0.46 \\
Pr(B \mid H) &= 0.34 \\
Pr(B \mid \overline{H}) &= 0.47
\end{aligned}
$$

Example #2: MLE with hidden variables

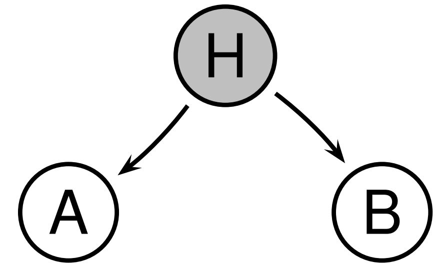


A	B	#	$Pr(H^m \mid D^m, \theta)$
0	0	6	.52
0	1	1	.39
1	0	1	.39
1	1	4	.28

- Iteration 2: E-Step
- Iteration 2: M-step (parameters re-estimated):

$$Pr(H) = 0.42$$
$$Pr(A \mid H) = 0.31$$
$$Pr(A \mid \overline{H}) = 0.50$$
$$Pr(B \mid H) = 0.30$$
$$Pr(B \mid \overline{H}) = 0.50$$

Example #2: MLE with hidden variables



A	B	#	$Pr(H^m \mid D^m, \theta)$
0	0	6	.79
0	1	1	.31
1	0	1	.31
1	1	4	.05

- Iteration 5: E-Step

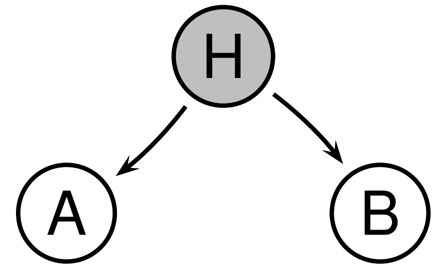
- Iteration 5: M-step:

$$
\begin{aligned}
Pr(H) &= 0.46 \\
Pr(A \mid H) &= 0.09 \\
Pr(A \mid \overline{H}) &= 0.69 \\
Pr(B \mid H) &= 0.09 \\
Pr(B \mid \overline{H}) &= 0.69
\end{aligned}
$$

Example #2: MLE with hidden variables

A	B	#	$Pr(H^m \mid D^m, \theta)$
0	0	6	.971
0	1	1	.183
1	0	1	.183
1	1	4	.001

- Iteration 10: E-Step



- Iteration 10: M-step:

$$Pr(H) = 0.52$$
$$Pr(A \mid H) = 0.03$$
$$Pr(A \mid \overline{H}) = 0.83$$
$$Pr(B \mid H) = 0.03$$
$$Pr(B \mid \overline{H}) = 0.83$$

Discussion

- Properties of EM:

Discussion

- Properties of EM:
  - Each iteration increases the log likelihood;

Discussion

- Properties of EM:
  - Each iteration increases the log likelihood;
  - Susceptible to local maximum.
    Solution: Initialize $\theta^{(0)}$ randomly, or for multiple times.

Discussion

- Properties of EM:
  - Each iteration increases the log likelihood;
  - Susceptible to local maximum.
    Solution: Initialize $\theta^{(0)}$ randomly, or for multiple times.
- Commonly used for:

Discussion

- Properties of EM:
  - Each iteration increases the log likelihood;
  - Susceptible to local maximum.
    Solution: Initialize $\theta^{(0)}$ randomly, or for multiple times.
- Commonly used for:
  - mixture models;

Discussion

- Properties of EM:
  - Each iteration increases the log likelihood;
  - Susceptible to local maximum.
    Solution: Initialize $\theta^{(0)}$ randomly, or for multiple times.
- Commonly used for:
  - mixture models;
  - HMM;

Discussion

- Properties of EM:
  - Each iteration increases the log likelihood;
  - Susceptible to local maximum.
    Solution: Initialize $\theta^{(0)}$ randomly, or for multiple times.
- Commonly used for:
  - mixture models;
  - HMM;
  - models with hidden variables (why do we need hidden variables?), *etc*.

References

- [1] Leslie Pack Kaelbling, sma5504/MIT6.825 lecture notes (available on MIT OpenCourseWare: ocw.mit.edu/OcwWeb/Electrical-Engineering-and-Computer-Science)
- [2] Michael I. Jordan, *An Introduction to Probabilistic Graphical Models* (Book Draft)