第5-2章 EM算法

- Maximum likelihood estimation (MLE)
- EM算法
- EM for Multinomial distribution

部分Slides来源于
faculty.washington.edu/fxia/courses/LING572/EM_part2.ppt

What is MLE?

- Given
  - A sample $X=\{X_1, \ldots, X_n\}$
  - A vector of parameters θ

- We define
  - Likelihood of the data: $L(\theta)=P(X \mid \theta)$
  - Log-likelihood of the data: $l(\theta)=\log P(X|\theta)$

- Given X, find

$$\theta_{ML} = \arg\max_{\theta\in\Omega} \; l(\Theta)$$

MLE (cont)

- Often we assume that X_is are independently identically distributed (i.i.d.)

$$
\begin{aligned}
\theta_{ML} &= \arg\max_{\theta \in \Omega} \ l(\Theta) \\
&= \arg\max_{\theta \in \Omega} \ \log P(X \mid \Theta) \\
&= \arg\max_{\theta \in \Omega} \ \log \prod_i \ P(X_i \mid \Theta) \\
&= \arg\max_{\theta \in \Omega} \ \sum_i \log P(X_i \mid \Theta)
\end{aligned}
$$

- Depending on the form of p(x|θ), solving optimization problem can be easy or hard.

An Easy Case

- Assuming
  - A coin has a probability p of being heads, 1-p of being tails.
  - Observation: We toss a coin N times, and the result is a set of Hs and Ts, and there are m Hs.

- What is the value of p based on MLE, given the observation?

An Easy Case (cont)

$$l(\Theta) = \log P(X \mid \Theta) = \log p^{m}(1-p)^{N-m}$$

$$= m \log p + (N-m)\log(1-p)$$

$$\frac{d\mathrm{l}(\Theta)}{dp} = \frac{d(m\log p + (N-m)\log(1-p))}{dp} = \frac{m}{p} - \frac{N-m}{1-p} = 0$$

$$\Rightarrow \quad \hat{p} = \frac{m}{N}$$

以频率来估计概率

Basic Setting in EM

- X is a set of data points: **observed** data
- Θ is a parameter vector.
- EM is a method to find θ_{ML} where

$$\theta_{ML} = \arg\max_{\theta \in \Omega} \; l(\Theta)$$
$$= \arg\max_{\theta \in \Omega} \; \log P(X \mid \Theta)$$

- Calculating $P(X \mid \theta)$ directly is hard.
- Calculating $P(X, Z|\theta)$ is much simpler, where Z is “hidden” data (or “missing” data).

The Basic Setting in EM

- Y = (X, Z)
  - Y: complete data (“augmented data”)
  - X: observed data (“incomplete” data)
  - Z: hidden data (“missing” data)

- Given a fixed x, there could be many possible z’s.
  - Ex: given a sentence x, there could be many state sequences in an HMM that generates x.

The Iterative Approach for MLE

- When missing data is available, it's hard to find the MLE directly

$$\theta_{ML} = \underset{\theta}{\text{Argmax}} \log \left(\sum_Z P(X, Z|\theta) \right)$$

- An alternative is to find a sequence

$$\theta^{(0)}, \theta^{(1)}, \cdots, \theta^{(t)}, \cdots,$$

$$\text{s.t.} \quad l(\theta^{(0)}) < l(\theta^{(1)}) < \cdots < l(\theta^{(t)}) < \cdots$$

$$
\begin{aligned}
l(\theta) - l(\theta^{(t)}) &= \log P(X|\theta) - \log P(X|\theta^{(t)}) \\
&= \log \left(\frac{\sum_Z P(X, Z|\theta)}{\sum_Z P(X, Z|\theta^{(t)})} \right) \\
&= \log \left(\sum_Z \frac{P(X, Z|\theta)}{\sum_{Z'} P(X, Z'|\theta^{(t)})} \right) \\
&= \log \left(\sum_Z \frac{P(X, Z|\theta)}{\sum_{Z'} P(X, Z'|\theta^{(t)})} \times \frac{P(X, Z|\theta^{(t)})}{P(X, Z|\theta^{(t)})} \right) \\
&= \log \left(\sum_Z \frac{P(X, Z|\theta^{(t)})}{\sum_{Z'} P(X, Z'|\theta^{(t)})} \times \frac{P(X, Z|\theta)}{P(X, Z|\theta^{(t)})} \right)
\end{aligned}
$$

$$
\begin{aligned}
l(\theta) - l(\theta^{(t)}) &= \log\left(\sum_Z \frac{P(X,Z|\theta^{(t)})}{\sum_{Z'} P(X,Z'|\theta^{(t)})} \times \frac{P(X,Z|\theta)}{P(X,Z|\theta^{(t)})}\right) \\
&= \log\left(\sum_Z P(Z|X,\theta^{(t)}) \times \frac{P(X,Z|\theta)}{P(X,Z|\theta^{(t)})}\right) \\
&\geq \sum_Z P(Z|X,\theta^{(t)}) \times \log\left(\frac{P(X,Z|\theta)}{P(X,Z|\theta^{(t)})}\right) \\
&= E_{P(Z|X,\theta^{(t)})}\left[\log\left(\frac{P(X,Z|\theta)}{P(X,Z|\theta^{(t)})}\right)\right] \\
&= E_{P(Z|X,\theta^{(t)})}\left[\log P(X,Z|\theta)\right] \\
&\qquad - E_{P(Z|X,\theta^{(t)})}\left[\log P(X,Z|\theta^{(t)})\right]
\end{aligned}
$$

Jensen’s inequality

Maximizing the Lower Bound

- The Jensen's inequality gives a lower bound to maximize,

$$\theta^{(t+1)} = \underset{\theta}{\mathrm{Argmax}}\, E_{P(Z|X,\theta^{(t)})}\left[\log P(X, Z|\theta)\right]$$

- Q-function

$$Q(\theta|\theta^{(t)}) = E_{P(Z|X,\theta^{(t)})}\left[\log P(X, Z|\theta)\right]$$

Increasing the Likelihood

- Increasing the likelihood by maximizing the lower bound

$$l(\theta) - l(\theta^{(t)}) \geq Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)})$$

$$Q(\theta^{(t+1)}|\theta^{(t)}) > Q(\theta^{(t)}|\theta^{(t)}) \Rightarrow l(\theta^{(t+1)}) > l(\theta^{(t)})$$

- Which means that a better estimation of the parameter.

Summary: EM Algorithm

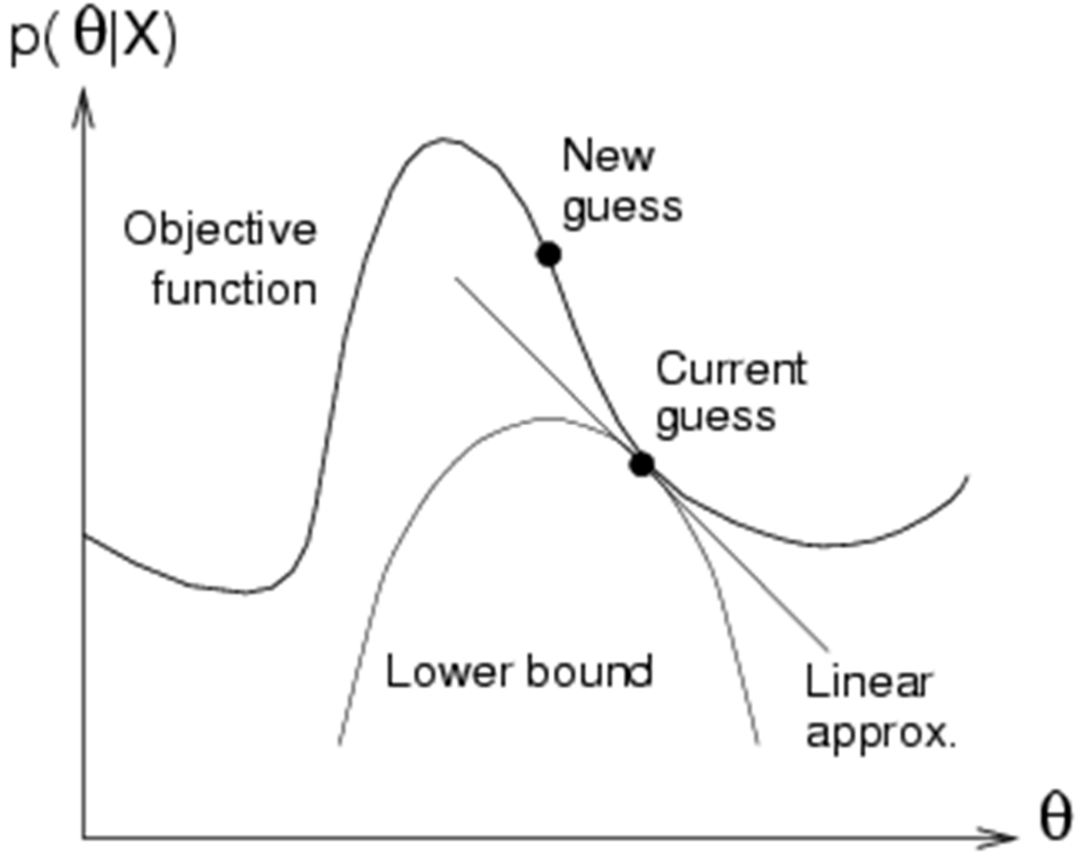
- Define a auxiliary function

$$
\begin{aligned}
Q(\theta|\theta') &= \sum_Z P(Z|X,\theta') \log P(X,Z|\theta) \\
&= E_{P(Z|X,\theta')}\left[\log P(X,Z|\theta)\right]
\end{aligned}
$$

- EM algorithm iterates with two step
  - E-Step, compute $Q(\theta|\theta^{(t)})$
  - M-Step:

$$\theta^{(t+1)} = \operatorname*{Argmax}_\theta Q(\theta|\theta^{(t)})$$

Illustration of EM Algorithm



Jensen’s Inequality

- Convex function

$$\forall x_1, x_2 \in (a,b), \lambda \in [0,1]$$
$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

Jensen's Inequality

- For convex function f(x)

$$E[f(X)] \geq f(E[X])$$

- For discrete random variable with two mass points

$$E[X] = p_1 x_1 + p_2 x_2$$
$$E[f(X)] = p_1 f(x_1) + p_2 f(x_2)$$
$$\geq f(p_1 x_1 + p_2 x_2) = f(E[x])$$

- It's easy to induce to random variable with more points

Jensen's Inequality Corollary

- Log(x) is a concave function, for any positive function g(x)

$$\log(E[g]) \geq E[\log(g)]$$

$$\log\left(\sum_j q_j g(j)\right) \geq \sum_j q_j \log(g(j))$$

where

$$q_j \in [0,1], \quad \sum_j q_j = 1$$

Example

- Rao (1965, pp.368-369), *Genetic Linkage Model*
- Suppose 197 animals are distributed multinomially into four categories,

$$X = (125, 18, 20, 34) = (x_1, x_2, x_3, x_4)$$

- A genetic model for the population specifies cell probabilities

$$\left(\frac{1}{2} + \frac{\theta}{4}, \frac{1}{4} - \frac{\theta}{4}, \frac{1}{4} - \frac{\theta}{4}, \frac{\theta}{4}\right)$$

Multinomial Distribution

- Likelihood function

$$L(\theta) = \frac{197!}{x_1!x_2!x_3!x_4!}(\frac{1}{2}+\frac{\theta}{4})^{x_1}(\frac{1}{4}-\frac{\theta}{4})^{x_2+x_3}(\frac{\theta}{4})^{x_4}$$

- log-likelihood function

$$l(\theta) = \log\frac{197!}{x_1!x_2!x_3!x_4!} + x_1\log(\frac{1}{2}+\frac{\theta}{4}) + (x_2+x_3)\log(\frac{1}{4}-\frac{\theta}{4}) + x_4\log(\frac{\theta}{4})$$

MLE

- Take derivative, solve equation

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{1}{4} \times \frac{x_1}{\frac{1}{2} + \frac{\theta}{4}} - \frac{1}{4} \times \frac{x_2 + x_3}{\frac{1}{4} - \frac{\theta}{4}} + \frac{1}{4} \times \frac{x_4}{\frac{\theta}{4}} = 0$$

- It's not easy to solve this equation!

$$\frac{x_1}{2+\theta} - \frac{x_2 + x_3}{1-\theta} + \frac{x_4}{\theta} = 0$$

Missing Data Problem

- Split the first category into two group

$$x_1 = z_1 + z_2, \quad z_1, z_2 \text{ missing}$$

With Probability

$$p(z_1) = \frac{1}{2}, p(z_2) = \frac{\theta}{4}$$

- Log-likelihood function of complete data

$$l(\theta) = \log \frac{197!}{z_1! z_2! x_2! x_3! x_4!}$$
$$+ z_1 \log(\frac{1}{2}) + (z_2 + x_4) \log(\frac{\theta}{4}) + (x_2 + x_3) \log(\frac{1}{4} - \frac{\theta}{4})$$

E Step: Multinomial

$$E\left(\log f(x,\theta)|\theta^{(k)}\right) = E\left(\log \frac{197!}{z_1!z_2!x_2!x_3!x_4!}\right)$$

$$+ z_1^{(k)} \log(\frac{1}{2}) + (z_2^{(k)} + x_4) \log(\frac{\theta}{4}) + (x_2 + x_3) \log(\frac{1}{4} - \frac{\theta}{4})$$

- Where

$$\begin{cases} E(z_1) = 125\dfrac{\frac{1}{2}}{\frac{1}{2} + \frac{\theta^{(k)}}{4}} = z_1^{(k)} \\ E(z_2) = 125\dfrac{\frac{\theta^{(k)}}{4}}{\frac{1}{2} + \frac{\theta^{(k)}}{4}} = z_2^{(k)} \end{cases}$$

M Step: Multinomial

- Take derivative

$$E\left(\log f(x,\theta)|\theta^{(k)}\right) = E\left(\log \frac{197!}{z_1!z_2!x_2!x_3!x_4!}\right)$$

$$+ z_1^{(k)} \log(\frac{1}{2}) + (z_2^{(k)} + x_4) \log(\frac{\theta}{4}) + (x_2 + x_3) \log(\frac{1}{4} - \frac{\theta}{4})$$

- One can obtain

$$\theta^{(k+1)} = \frac{z_2^{(k)} + x_4}{z_2^{(k)} + x_4 + x_2 + x_3} = \frac{z_2^{(k)} + 34}{z_2^{(k)} + 18 + 20 + 34}$$

Back to Motif Finding

- Given the missing data, it's a multinomial distribution

$$Pr(X_i, Z_{ij} = 1|P) = \underbrace{\prod_{k=1}^{j-1} p_{x_{ik},0}}_{\text{before motif}} \underbrace{\prod_{k=j}^{j+w-1} p_{x_{ik},k-j+1}}_{\text{motif}} \underbrace{\prod_{k=j+w}^{L} p_{x_{ik},0}}_{\text{after motif}}$$

X_i is the *i*th sequence

Z_{ij} is 1 if motif starts at position *j* in sequence *i*

Log-likelihood

$$l(p) = \sum_{k=1}^{j-1} \log p_{x_{ik},0} + \sum_{k=j}^{j+W-1} \log p_{x_{ik},k-j+1} + \sum_{k=j+W}^{L} \log p_{x_{ik},0}$$

- Q function

$$\begin{aligned} Q(p|p^{(t)}) &= E_{P(Z|X,p^{(t)})}\left[\log P(X,Z|p)\right] \\ &= \sum_{Z} P(Z|X,p^{(t)}) \log P(X,Z|p) \end{aligned}$$

Q-function

$$
\begin{aligned}
Q(p|p^{(t)}) &= \sum_{Z} P(Z|X, p^{(t)}) \log P(X, Z|p) \\
&= \sum_{Z} P(Z|X, p^{(t)}) \sum_{k=1}^{j-1} \log p_{x_{ik},0} \\
&+ \sum_{Z} P(Z|X, p^{(t)}) \sum_{k=j}^{j+W-1} \log p_{x_{ik},k-j+1} \\
&+ \sum_{Z} P(Z|X, p^{(t)}) \sum_{k=j+W}^{L} \log p_{x_{ik},0}
\end{aligned}
$$

Q-function

- For each sequence i, the missing value Z_{ij} can take value

$$Z_{i1} = 1, Z_{i2} = 1, \cdots, Z_{i,L-W+1} = 1$$

- So the coefficient of $\log P_{c,k}$ is

$$\sum_{i} \sum_{m=1}^{L-W+1} P(Z_{im} = 1 | X_i, p^t) \delta(X_{i,m+k}, c)$$

Q-function

- The coefficient of $\log P_{c,0}$ is

$$\sum_i \sum_{m=1}^{L-W+1} P(Z_{im}=1|X_i,p^t)\left(\sum_{k=1}^{m-1}\delta(X_{i,k},c)+\sum_{k=m+W}^{L}\delta(X_{i,k},c)\right)$$

M Step: Optimization

- For multinomial distribution, the optimization is of form

$$\begin{cases} \text{Max: } \sum_k c_k \log x_k \\ \text{subject to: } \sum_k x_k = 1 \end{cases}$$

$$\text{Estimation: } x_i = \frac{c_i}{\sum_k c_k}, i = 1, \cdots, N.$$

M Step: Optimization

- So the estimation of $p_{c,k}$ is

$$\frac{\sum_i \sum_{m=1}^{L-W+1} P(Z_{im}=1|X_i,p^t)\delta(X_{i,m+k},c)}{\sum_b \sum_i \sum_{m=1}^{L-W+1} P(Z_{im}=1|X_i,p^t)\delta(X_{i,m+k},b)}$$

- So the estimation of $p_{c,0}$ is

$$\frac{\sum_i \sum_{m=1}^{L-W+1} P(Z_{im}=1|X_i,p^t)\left(\sum_{k=1}^{m-1}\delta(X_{i,k},c)+\sum_{k=m+W}^{L}\delta(X_{i,k},c)\right)}{\sum_b \sum_i \sum_{m=1}^{L-W+1} P(Z_{im}=1|X_i,p^t)\left(\sum_{k=1}^{m-1}\delta(X_{i,k},b)+\sum_{k=m+W}^{L}\delta(X_{i,k},b)\right)}$$

Example

- Finding motif (length 3) in following sequences

A C A G C A

A G G C A G

T C A G T C

EM Updating

- Let

$$z_{ij}(c) = Pr(Z_{ij} = 1 | X_i, p^{(t)}) \delta(x_{i,m+k}, c)$$

1	2	3	1	2	3
z11(A)	z11(C)	z11(A)	z21(A)	z21(G)	z21(G)
z12(C)	z12(A)	z12(G)	z22(G)	z22(G)	z22(C)
z13(A)	z13(G)	z13(C)	z23(G)	z23(C)	z23(A)
z14(G)	z14(C)	z14(A)	z24(C)	z24(A)	z24(G)

1	2	3
z31(T)	z31(C)	z31(A)
z32(C)	z32(A)	z32(G)
z33(A)	z33(G)	z33(T)
z34(G)	z34(T)	z34(C)

EM Updating

$$p_{A,1} = \frac{z_{11} + z_{13} + z_{21} + z_{33}}{z_{11} + z_{12} + z_{13} + z_{14} + \cdots + z_{31} + z_{32} + z_{33} + z_{34}}$$

$$p_{C,1} = \frac{z_{12} + z_{24} + z_{32}}{z_{11} + z_{12} + z_{13} + z_{14} + \cdots + z_{31} + z_{32} + z_{33} + z_{34}}$$

$$p_{G,1} = \frac{z_{14} + z_{22} + z_{23} + z_{32}}{z_{11} + z_{12} + z_{13} + z_{14} + \cdots + z_{31} + z_{32} + z_{33} + z_{34}}$$

$$p_{T,1} = \frac{z_{31}}{z_{11} + z_{12} + z_{13} + z_{14} + \cdots + z_{31} + z_{32} + z_{33} + z_{34}}$$

Background

- z11: A,C,G
- z12: 2A,C
- z13:2A,C
- z14: 2A, C
- z21:A,C,G
- z22:2A,G
- z23:A,2G
- z24:A,2G
- z31:C,G,T
- z32:C,2T
- z33:2C,T
- z34:A,C,G

Background Updating

- A $z_{11} + 2z_{12} + 2z_{13} + 2z_{14} + z_{21} + 2z_{22} + z_{23} + z_{24} + z_{34}$
- C $z_{11} + z_{12} + z_{13} + z_{14} + z_{21} + z_{31} + z_{32} + 2z_{33} + z_{34}$
- G $z_{11} + z_{21} + z_{22} + 2z_{23} + 2z_{24} + z_{31} + z_{34}$
- T $z_{31} + 2z_{32} + z_{33}$

- Normalization factor

$$3(z_{11} + z_{12} + z_{13} + z_{14} + z_{21} + z_{22} + z_{23} + z_{24} + z_{31} + z_{32} + z_{33} + z_{34})$$

References

- Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological),* Vol. 39, No. 1, , pp. 1-38