EM Algorithm Tutorial

Expectation Maximization

李修成

Department of Computer Science

Harbin Institute of Technology

Outline

1. EM 算法的提出
2. EM 算法的推导
3. EM Algorithm
4. EM 算法收敛性的证明
5. EM 算法的应用：求解 GMM(高斯混合模型)
6. EM 算法的推广：变分 EM 算法 (Varitional EM)

为什么要有 EM 算法

- EM 算法解决的是含有隐含变量概率模型的参数估计问题（极大似然估计）
- 考虑如下概率模型参数估计

$$\ell(\theta) = \log p(x;\theta) = \log \sum_{z} p(x,z;\theta) \tag{1}$$

问题难以用极大似然估计求解参数！难在 log 与 p 之间有个 Σ！

为什么要有 EM 算法

- EM 算法解决的是含有隐含变量概率模型的参数估计问题（极大似然估计）
- 考虑如下概率模型参数估计

$$\ell(\theta) = \log p(x;\theta) = \log \sum_z p(x,z;\theta) \tag{1}$$

问题难以用极大似然估计求解参数！难在 log 与 p 之间有个 Σ！

- Gaussian Discriminant Analysis model 高斯判别分析
- Gaussian Mixture Model 高斯混合模型

GDA(高斯判别分析),MLE 易解

高斯判别分析

已知有多个正态分布，每个样本数据由一个正态分布产生，生成数据的正态分布已知。给出训练数据，现要估计参数

$p(x^{(i)}, z^{(i)}) = p(x^{(i)}|z^{(i)})p(z^{(i)})$.

$z^{(i)} \sim Multinomial(\phi)(where\ \phi_j \geq 0, \sum_{j=1}^{k} \phi_j = 1,$ and the parameter gives $p(z^{(i)} = j))$, and $(x^{(i)}|z^{(i)} = j) \sim N(\mu_j, \Sigma_j)$.

GDA 似然函数

$$
\begin{aligned}
\ell(\phi, \mu, \Sigma) &= \sum_{i=1}^{m} \log p(x^{(i)}, z^{(i)}; \mu, \Sigma) \\
&= \sum_{i=1}^{m} \log p(x^{(i)} | z^{(i)}; \mu, \Sigma) p(z^{(i)}; \phi) \\
&= \sum_{i=1}^{m} \log \frac{1}{(2\pi)^{n/2} |\Sigma_j|^{1/2}} \exp(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)) \phi_j
\end{aligned} \tag{2}
$$

MLE 求解 GDA

Maximizing this with respect to ϕ, μ, Σ gives the parameters:

- $\phi_j = \frac{1}{m} \sum_{i=1}^{m} 1\{z^{(i)} = j\}$
- $\mu_j = \frac{\sum_{i=1}^{m} 1\{z^{(i)}=j\} x^{(i)}}{\sum_{i=1}^{m} 1\{z^{(i)}=j\}}$
- $\Sigma_j = \frac{\sum_{i=1}^{m} 1\{z^{(i)}=j\} (x^{(i)}-\mu_j)(x^{(i)}-\mu_j)^T}{\sum_{i=1}^{m} 1\{z^{(i)}=j\}}$

GMM(高斯混合模型)，MLE 难解

高斯混合模型

已知有多个正态分布，每个样本数据由一个正态分布产生，但具体是哪一个，我们无法观测（含有隐含变量）。给出训练数据，现要估计参数

$p(x^{(i)}, z^{(i)}) = p(x^{(i)}|z^{(i)})p(z^{(i)})$.

$z^{(i)} \sim Multinomial(\phi)$($where\ \phi_j \geq 0, \sum_{j=1}^{k} \phi_j = 1$, and the parameter gives $p(z^{(i)} = j)$), and $(x^{(i)}|z^{(i)} = j) \sim N(\mu_j, \Sigma_j)$.

GMM 似然函数

$$
\begin{aligned}
\ell(\phi, \mu, \Sigma) &= \sum_{i=1}^{m} \log p(x^{(i)}, z^{(i)}; \mu, \Sigma) \\
&= \sum_{i=1}^{m} \log \sum_{z^{(i)}=1}^{k} p(x^{(i)} | z^{(i)}; \mu, \Sigma) p(z^{(i)}; \phi) \\
&= \sum_{i=1}^{m} \log \sum_{j=1}^{k} \frac{1}{(2\pi)^{n/2} |\Sigma_j|^{1/2}} \exp(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)) \phi_j
\end{aligned} \tag{3}
$$

- 怎么办?

- 怎么办? Jensen's Inequality

$$\begin{aligned}\ell(\theta) = \log p(x;\theta) &= \log \sum_z p(x,z;\theta) \\ &= \log \sum_z Q(z) \frac{p(x,z;\theta)}{Q(z)} \\ &\geq \sum_z Q(z) \log \frac{p(x,z;\theta)}{Q(z)}\end{aligned} \tag{4}$$

$Q(z)$ 为关于隐含变量 z 的任意分布

- 令 $\mathcal{L}(x, z; \theta) = \sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)}$ 得到

$$\begin{aligned}
\ell(\theta) - \mathcal{L}(x, z; \theta) &= \log p(x; \theta) - \sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)} \\
&= \sum_z Q(z) \log p(x; \theta) - \sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)} \qquad (5) \\
&= -\sum_z Q(z) \log \frac{p(z|x)}{Q(z)} = D(Q(z) \parallel p(z|x; \theta))
\end{aligned}$$

- 似曾相识！

- 令 $\mathcal{L}(x, z; \theta) = \sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)}$ 得到

$$\begin{aligned} \ell(\theta) - \mathcal{L}(x, z; \theta) &= \log p(x; \theta) - \sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)} \\ &= \sum_z Q(z) \log p(x; \theta) - \sum_z Q(z) \log \frac{p(x, z; \theta)}{Q(z)} \\ &= -\sum_z Q(z) \log \frac{p(z|x)}{Q(z)} = D(Q(z) \parallel p(z|x; \theta)) \end{aligned} \tag{5}$$

- 似曾相识！ LDA!
- $\ell(\theta) - \mathcal{L}(x, z; \theta)$ 得到的是 $Q(z)$ 和 $p(z|x)$ 的 KL 距离，又称相对熵。
-

$$\ell(\theta) = \mathcal{L}(x, z; \theta) + D(Q(z) \parallel p(z|x; \theta)) \tag{6}$$

- 由于 $\mathcal{L}(x, z; \theta)$ 消除了 $\ell(\theta)$ 中 log 后面的 $\sum$，因此它是容易计算的
- 由于 $KL(q \parallel p) \geq 0$，故 $\mathcal{L}(x, z; \theta)$ 为 $\ell(\theta)$ 的下界
- 用 $\mathcal{L}(x, z; \theta)$ 去逼近 $\ell(\theta)$，而 $\ell(\theta) \geq \mathcal{L}(x, z; \theta)$
  故当 $D(Q(z) \parallel p(z|x; \theta)) = 0$ 时，下界函数 $\mathcal{L}(x, z; \theta)$ 可以最大化

- 由于 $\mathcal{L}(x, z; \theta)$ 消除了 $\ell(\theta)$ 中 log 后面的 $\sum$，因此它是容易计算的
- 由于 $KL(q \parallel p) \geq 0$，故 $\mathcal{L}(x, z; \theta)$ 为 $\ell(\theta)$ 的下界
- 用 $\mathcal{L}(x, z; \theta)$ 去逼近 $\ell(\theta)$，而 $\ell(\theta) \geq \mathcal{L}(x, z; \theta)$
  故当 $D(Q(z) \parallel p(z|x; \theta)) = 0$ 时，下界函数 $\mathcal{L}(x, z; \theta)$ 可以最大化
  EM 算法的精髓！
- 显然，当 $Q(z) = p(z|x; \theta)$ 时 $\mathcal{L}(x, z; \theta)$ 极大化

EM Algorithm

1 Repeat until convergence $\{$

2 **E-step:**

3 For each i, set

4 $Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta)$

5 **M-step:**

6 Set

7 $\theta := arg\ max_\theta \sum_i \sum_z Q_i(z^{(i)}) \log \frac{p(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)};\theta)}$

8 $\}$

其中 **M-step** 亦可为，$\theta := arg\ max_\theta \sum_i \sum_z Q_i(z^{(i)}) \log p(x^{(i)}, z^{(i)};\theta)$

因为，$\sum_i \sum_z Q_i(z^{(i)}) \log Q_i(z^{(i)})$ 是一个与 θ 独立的常量

EM 算法收敛性的证明

由于

$$\ell(\theta) \geq \sum_i \sum_z Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \tag{7}$$

$\Rightarrow$

$$\begin{aligned} \ell(\theta^{(t+1)}) &\geq \sum_i \sum_z Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{(t+1)})}{Q_i^{(t)}(z^{(i)})} \\ &\geq \sum_i \sum_z Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{(t)})}{Q_i^{(t)}(z^{(i)})} \qquad (8) \\ &= \ell(\theta^{(t)}) \end{aligned}$$

其中第 2 步成立是因为，θ^{t+1} 是由第 t 次迭代经极大似然估计得到。

EM 算法应用：求解 GMM(高斯混合模型)

E-step:令

$$
\begin{aligned}
w_j^{(i)} &= Q_i(z^{(i)} = j) \\
&= p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma) \\
&= \frac{p(x^{(i)} | z^{(i)} = j; \mu, \Sigma) p(z^{(i)} = j; \phi)}{\sum_{l=1}^{k} p(x^{(i)} | z^{(i)} = l; \mu, \Sigma) p(z^{(i)} = l; \phi)}
\end{aligned} \tag{9}
$$

M-step:似然函数

$$
\begin{aligned}
\ell(\theta) &= \sum_{i=1}^{m} \sum_{z} Q_i(z^{(i)}) \log p(x^{(i)}, z^{(i)}; \mu, \Sigma) \\
&= \sum_{i=1}^{m} \sum_{j=1}^{k} Q_i(z^{(i)} = j) \log p(x^{(i)} | z^{(i)} = j; \mu, \Sigma) p(z^{(i)} = j; \phi) \\
&= \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(i)} \log \frac{1}{(2\pi)^{n/2} |\Sigma_j|^{1/2}} \exp(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)) \phi_j
\end{aligned}
\tag{10}
$$

下面进行参数极大似然估计：

- μ_j
- ϕ_j
- Σ_j

一些矩阵论中的知识

兵马未动，粮草先行

- $\nabla_x x^T A x = 2Ax$, A is symmetric
- $\nabla_x^2 x^T A x = 2A$, A is symmetric
- $\nabla_x b^T x = b$
- $\nabla_A \log |A| = A^{-1}$, A is invertible
- $\nabla_A x^T A x = x x^T$, A is $n \times n$ matrix

μ_j 的估计

$$\nabla_{\mu_j}\ell(\theta) = \nabla_{\mu_j}\sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)}\frac{1}{2}(x^{(i)}-\mu_j)^T\Sigma_j^{-1}(x^{(i)}-\mu_j)$$

$$= \sum_{i=1}^{m} w_j^{(i)}\Sigma_j^{-1}(x^{(i)}-\mu_j) = 0 \tag{11}$$

$$\mu_j = \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}}$$

ϕ_j 的估计

$$\begin{aligned}\nabla_{\phi_j}\ell(\theta) &= \nabla_{\phi_j}\sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)}\log\phi_j \\ &= \sum_{i=1}^{m}\frac{w_j^{(i)}}{\phi_j}\end{aligned} \tag{12}$$

- 解不下去了，梯度 $\nabla_{\phi_j}\ell(\theta) > 0$，似乎只要使 ϕ_j 取无穷大，似然函数就可以无限大！

ϕ_j 的估计

$$\begin{aligned}\nabla_{\phi_j}\ell(\theta) &= \nabla_{\phi_j}\sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)}\log\phi_j \\ &= \sum_{i=1}^{m}\frac{w_j^{(i)}}{\phi_j}\end{aligned} \tag{12}$$

- 解不下去了，梯度 $\nabla_{\phi_j}\ell(\theta) > 0$，似乎只要使 ϕ_j 取无穷大，似然函数就可以无限大！
- 哪有那么好的事！别忘了还有约束条件 $\sum_{j=1}^{k}\phi_j = 1$！

ϕ_j 的估计

$$\begin{aligned}\nabla_{\phi_j}\ell(\theta) &= \nabla_{\phi_j}\sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)}\log\phi_j \\ &= \sum_{i=1}^{m}\frac{w_j^{(i)}}{\phi_j}\end{aligned} \tag{12}$$

- 解不下去了，梯度 $\nabla_{\phi_j}\ell(\theta) > 0$，似乎只要使 ϕ_j 取无穷大，似然函数就可以无限大！
- 哪有那么好的事！别忘了还有约束条件 $\sum_{j=1}^{k}\phi_j = 1$！
- 什么？约束条件！求极大值！无限遐想中 ...

ϕ_j 的估计

$$\begin{aligned}\nabla_{\phi_j}\ell(\theta) &= \nabla_{\phi_j}\sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)}\log\phi_j \\ &= \sum_{i=1}^{m}\frac{w_j^{(i)}}{\phi_j}\end{aligned} \tag{12}$$

- 解不下去了，梯度 $\nabla_{\phi_j}\ell(\theta) > 0$，似乎只要使 ϕ_j 取无穷大，似然函数就可以无限大！
- 哪有那么好的事！别忘了还有约束条件 $\sum_{j=1}^{k}\phi_j = 1$！
- 什么？约束条件！求极大值！无限遐想中 ...
- 想到了 Lagrangian，因为目标函数是凸的嘛！

ϕ_j 的估计

$$\begin{aligned}\nabla_{\phi_j}\ell(\theta) &= \nabla_{\phi_j}\sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)}\log\phi_j \\ &= \sum_{i=1}^{m}\frac{w_j^{(i)}}{\phi_j}\end{aligned} \tag{12}$$

- 解不下去了，梯度 $\nabla_{\phi_j}\ell(\theta) > 0$，似乎只要使 ϕ_j 取无穷大，似然函数就可以无限大！
- 哪有那么好的事！别忘了还有约束条件 $\sum_{j=1}^{k}\phi_j = 1$！
- 什么？约束条件！求极大值！无限遐想中 ...
- 想到了 Lagrangian，因为目标函数是凸的嘛！（其实是凹的，正好可以求极大值）

ϕ_j 的估计

构造 Lagrangian 函数

$$\mathcal{L}(\phi) = \sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} \log \phi_j + \beta(\sum_{j=1}^{k} \phi_j - 1) \tag{13}$$

求偏导

$$\frac{\partial \mathcal{L}}{\partial \phi_j} = \sum_{i=1}^{m} \frac{w_j^{(i)}}{\phi_j} + \beta \Rightarrow -\beta\phi_j = \sum_{i=1}^{m} w_j^{(i)}$$

$$-\beta \sum_{j=1}^{k} \phi_j = \sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} = m \Rightarrow -\beta = m \tag{14}$$

$$\phi_j = \sum_{i=1}^{m} \frac{w_j^{(i)}}{m}$$

Σ_j 的估计

令 $S_j = \Sigma_j^{-1}, b = (x^{(i)} - \mu_j)$

$$
\begin{aligned}
\nabla_{s_j} \ell(\theta) &= \nabla_{s_j} \sum_{i=1}^{m} w_j^{(i)} (\frac{1}{2} \log |S_j| - \frac{1}{2} b^T S_j b) \\
&= \frac{1}{2} \sum_{i=1}^{m} w_j^{(i)} (S_j^{-1} - bb^T) = 0 \\
S_j^{-1} &= \frac{\sum_{i=1}^{m} w_j^{(i)} bb^T}{\sum_{i=1}^{m} w_j^{(i)}} \\
\Sigma_j &= \frac{\sum_{i=1}^{m} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} w_j^{(i)}}
\end{aligned}
\tag{15}
$$

EM 算法的推广：变分 EM 算法 (Varitional EM)

EM 算法的推广：变分 EM 算法 (Varitional EM)

- 由于昨晚去喝酒了，而且喝得挺多，所以这部分内容没来的及做
- 回忆 $\mathcal{LDA}$(Latent Dirichlet Allocation)
- ...