EM、GMM

七月算法 邹博

2015年4月19日

主要内容

- 通过实例直观求解高斯混合模型GMM
  - 适合快速掌握GMM，及编程实现
- 通过极大似然估计详细推导EM算法
  - 适合理论层面的深入理解
  - 用坐标上升理解EM的过程
- 推导GMM的参数φ、μ、σ
  - 复习多元高斯模型
  - 复习拉格朗日乘子法

极大似然估计

- 找出与样本的分布最接近的概率分布模型。
- 简单的例子
  - 10次抛硬币的结果是：正正反正正正反反正正
- 假设p是每次抛硬币结果为正的概率。则：
- 得到这样的实验结果的概率是：

$$
\begin{aligned}
P &= pp(1-p)ppp(1-p)(1-p)pp \\
&= p^7(1-p)^3
\end{aligned}
$$

极大似然估计MLE

- 目标函数：$\max P = \max\limits_{0 \le p \le 1} p^7 (1-p)^3$
- 最优解是：p=0.7
  - 思考：如何求解？

- 一般形式：$L_{\overline{p}} = \prod\limits_x p(x)^{\overline{p}(x)}$

$p(x)$模型是估计的概率分布

$\overline{p}(x)$是实验结果的分布

进一步考察

- 若给定一组样本$x_1,x_2\ldots x_n$，已知它们来自于高斯分布$N(\mu,\sigma)$，试估计参数μ,σ。

按照MLE的过程分析

- 高斯分布的概率密度函数：

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

- 将x1,x2...xn带入，得到：

$$L(x)=\prod_i \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

化简对数似然函数

$$l(x)=\log\prod_i \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$=\sum_i \log\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$=\left(\sum_i \log\frac{1}{\sqrt{2\pi}\sigma}\right)+\left(\sum_i -\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$=-\frac{n}{2}\log\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2$$

参数估计的结论

- 目标函数 $$l(x) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2$$
- 将目标函数对参数μ，σ分别求偏导，很容易得到μ，σ的式子：

$$\mu = \frac{1}{n}\sum_i x_i$$

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

符合直观想象

$$\mu = \frac{1}{n}\sum_i x_i$$

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

- 上述结论和矩估计的结果是一致的，并且意义非常直观：样本的均值即高斯分布的均值，样本的方差即高斯分布的方差。
  - 注：经典意义下的方差，分母是n-1；在似然估计的方法中，求的方差是n
- 该结论将作为下面分析的基础。

思考：若随机变量无法直接(完全)观察到

- 在西单商场随机挑选100位顾客，测量这100位顾客的身高：
- 若这100个样本服从正态分布N(μ, σ)，试估计参数μ和σ。
- 若样本中存在男性和女性顾客，它们服从N(μ_1, σ_1)和N(μ_2, σ_2)的分布，试估计μ_1, σ_1, μ_2, σ_2。

从直观理解猜测GMM的参数估计

- 随机变量X是有K个高斯分布混合而成，取各个高斯分布的概率为φ1，φ2...φK，第i个高斯分布的均值为μi，方差为Σi。若观测到随机变量X的一系列样本x1,x2...xn，试估计参数φ，μ，Σ。

建立目标函数

- 对数似然函数

$$l(x)=\sum_{i=1}^{N}\log\left(\sum_{k=1}^{K}\pi_k N(x_i \mid \mu_k,\Sigma_k)\right)$$

目标函数

- 由于在对数函数里面又有加和，我们没法直接用求导解方程的办法直接求得极大值。为了解决这个问题，我们分成两步。

第一步：估算数据来自哪个组份

- 估计数据由每个组份生成的概率：对于每个数据xi来说，它由第k个组份生成的概率为

$$\gamma(i,k)=\frac{\pi_k N(x_i\mid\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j N(x_i\mid\mu_j,\Sigma_j)}$$

- 由于式子里的μ和Σ也是需要我们估计的值，我们采用迭代法，在计算γ(i,k)的时候我们假定μ和Σ均已知；
  - 第一次计算时，需要先验知识给定μ和Σ。

第二步：估计每个组份的参数

- 假设上一步中得到的$\gamma(i,k)$就是正确的“数据xi由组份k生成的概率”，亦可以当做该组份在生成这个数据上所做的贡献；
- 或者，我们可以看作xi其中有$\gamma(i,k)$*xi部分是由组份k所生成的。

第二步：估计每个组份的参数

- 对于所有的数据点，现在实际上可以看作组份k生成了{γ(i,k)*xi|i=1,2,…N}这些点。组份k是一个标准的高斯分布，利用上面的结论：

$$\mu_k = \frac{1}{N_k}\sum_{i=1}^{N}\gamma(i,k)x_i$$

$$\Sigma_k = \frac{1}{N_k}\sum_{i=1}^{N}\gamma(i,k)(x_i-\mu_k)(x_i-\mu_k)^T$$

EM算法的提出

- 假定有训练集

$$\{x^{(1)},\ldots,x^{(m)}\}$$

- 包含m个独立样本，希望从中找到该组数据的模型p(x,z)的参数。

通过极大似然估计建立目标函数

- 取对数似然函数

$$
\begin{aligned}
\ell(\theta) &= \sum_{i=1}^{m} \log p(x;\theta) \\
&= \sum_{i=1}^{m} \log \sum_{z} p(x,z;\theta)
\end{aligned}
$$

问题的提出

- 这里，z是隐随机变量，直接找到参数的估计是很困难的。我们的策略是建立$l(\theta)$的下界，并且求该下界的最大值；重复这个过程，直到收敛到局部最大值。

Jensen不等式

- 令Qi是z的某一个分布，Qi≥0，有：

$$
\begin{aligned}
\sum_i \log p(x^{(i)};\theta) &= \sum_i \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)};\theta) \\
&= \sum_i \log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} \\
&\geq \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}
\end{aligned}
$$

寻找尽量紧的下界

- 为了使等号成立

$$\frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} = c$$

进一步分析

$$Q_i(z^{(i)}) \propto p(x^{(i)}, z^{(i)}; \theta) \qquad \sum_z Q_i(z^{(i)}) = 1$$

$$\begin{aligned} Q_i(z^{(i)}) &= \frac{p(x^{(i)}, z^{(i)}; \theta)}{\sum_z p(x^{(i)}, z; \theta)} \\ &= \frac{p(x^{(i)}, z^{(i)}; \theta)}{p(x^{(i)}; \theta)} \\ &= p(z^{(i)} | x^{(i)}; \theta) \end{aligned}$$

EM算法整体框架

Repeat until convergence {

(E-step) For each i, set

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

(M-step) Set

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

}

坐标上升

Remark. If we define

$$J(Q,\theta)=\sum_i\sum_{z^{(i)}}Q_i(z^{(i)})\log\frac{p(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})},$$

the we know $\ell(\theta)\geq J(Q,\theta)$ from our previous derivation. The EM can also be viewed a coordinate ascent on J, in which the E-step maximizes it with respect to Q, and the M-step maximizes it with respect to θ.

从理论公式推导GMM

- 随机变量X是有K个高斯分布混合而成，取各个高斯分布的概率为φ1，φ2...φK，第i个高斯分布的均值为μi，方差为Σi。若观测到随机变量X的一系列样本x1,x2...xn，试估计参数φ，μ，Σ。

E-step

$$w_j^{(i)} = Q_i(z^{(i)} = j) = P(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$$

M-step

- 将多项分布和高斯分布的参数带入：

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

对均值求偏导

$$
\begin{aligned}
&\nabla_{\mu_l} \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)}-\mu_j)^T \Sigma_j^{-1} (x^{(i)}-\mu_j)\right) \cdot \phi_j}{w_j^{(i)}} \\
&= -\nabla_{\mu_l} \sum_{i=1}^{m} \sum_{j=1}^{k} w_j^{(i)} \frac{1}{2} (x^{(i)}-\mu_j)^T \Sigma_j^{-1} (x^{(i)}-\mu_j) \\
&= \frac{1}{2} \sum_{i=1}^{m} w_l^{(i)} \nabla_{\mu_l} 2\mu_l^T \Sigma_l^{-1} x^{(i)} - \mu_l^T \Sigma_l^{-1} \mu_l \\
&= \sum_{i=1}^{m} w_l^{(i)} \left(\Sigma_l^{-1} x^{(i)} - \Sigma_l^{-1} \mu_l\right)
\end{aligned}
$$

高斯分布的均值

- 令上式等于0，解的均值：

$$\mu_l := \frac{\sum_{i=1}^m w_l^{(i)} x^{(i)}}{\sum_{i=1}^m w_l^{(i)}}$$

高斯分布的方差：求偏导，等于0

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

多项分布的参数

- 考察M-step的目标函数，对于ϕ，删除常数项

$$\sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)}-\mu_j)^T \Sigma_j^{-1}(x^{(i)}-\mu_j)\right) \cdot \phi_j}{w_j^{(i)}}$$

- 得到

$$\sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} \log \phi_j$$

拉格朗日乘子法

- 由于多项分布的概率和为1，建立拉格朗日方程

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

- 注：这样求解的φi一定非负，所以，不用考虑φi≥0这个条件

求偏导，等于0

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

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

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

参考文献

- Prof. Andrew Ng, Machine Learning, Stanford University

我们在这里

- 更多机器学习问题在 7 七月算法
  - http://www.julyedu.com/
    - 免费视频
    - 直播课程
    - 问答社区
- contact us：微博
  - @研究者July
  - @七月问答
  - @邹博_机器学习

感谢大家！

恳请大家批评指正！