EM算法与高斯混合模型笔记以及公式推导

0.Introduction

原创文章，转载请注明出处，谢谢。本文共有2000多字，阅读需要至少20分钟，主要章节包括：**EM算法简介，预备知识，高斯混合模型 (GMMs)以及EM算法，EM算法，最后回顾GMMs，总结，Reference**。本文先列举了预备知识，然后使用了高斯混合模型作为例子理解EM算法，再将EM算法进行推导。

1. EM算法简介

用于含有隐变量(latent variable/hidden variable)的概率模型参数的极大似然估计或者极大后验概率估计。也就是用于解决无标记样本分类问题（无监督学习）。每次迭代有两步组成：E(Expectation): 求期望；M(Maximisation): 求极大。本质上就是求期望的最大化。类似于K-means算法，但是K-means算法无法给出样本点属于K类的概率。

2. 预备知识

2.1 Jensen不等式：

若：

$$f(x)$$

是凹函数， θ 是一个从0到1的实数

则：

$$f(\theta x+(1-\theta)y) \leq \theta f(x)+(1-\theta)f(y)$$

左边式子表示：函数图形本身

右边式子表示：函数图像的割线，因为是凹函数，割线一定在图像的上方。

再若(离散情况)：

$$\theta_1,...\theta_k \ \geq 0, \ and \, \theta_1+...+\theta_k \ =1$$

则：

$$f\left(\theta_1 x_1+...+\theta_k x_k\right) \leq \theta_1 f\left(x_1\right)+...+\theta_k f\left(x_k\right)$$

左边表示：对

$$\theta_k x_k$$

期望的函数值

右边表示：对

$$f(x_k)$$

求期望

再若(连续情况):

$$p(x) \geq 0 \, f, \int_s p(x) \, dx = 1$$

则:

$$f\left(\int_s p(x) \, x dx\right) \leq \int_s f(x) \, p(x) \, dx$$

则:

$$f(Ex) \leq Ef(x)$$

总结：期望的函数值小于等于函数值的期望

2.2 后验概率:

$$P(A|B)$$

在B的条件下A发生的概率。

$$p(A|B) = \frac{P(AB)}{P(B)}$$

2.3 极大似然估计:

例如抛硬币（伯努利分布），抛10次的结果为(+ + - + + + - - + +)

设得到 + 的概率为p，则能得到这样结果的概率为:

$$P = pp(1-p)ppp(1-p)(1-p)pp$$

$$P = p^7(1-p)^3$$

极大似然估计就是求参数 p 使得 P 最大即

$$\max P = \max p^7(1-p)^3$$

以上式子求导之后设为0，求出最大值

$$\frac{7}{10}$$

似然函数(Likelihood Function)：

$$l(\theta) = \prod_x p(x|\theta)$$

再若：给定的是一组来自高斯分布

$$\mathcal{N}(\mu, \sigma)$$

的样本

$$x_{1...}x_n$$

尝试估计样本参数 μ, σ

同样使用极大似然估计的方法来求解：

先拿出高斯分布的概率密度函数：

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

将样本

$$x_{1...}x_n$$

带入似然函数则得到：

$$L(x) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

因为取对数求导比较方便，所以以上式子变为对数似然函数：

$$l(x) = log\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

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

$$l(x)=\sum_i log\frac{1}{\sigma\sqrt{2\pi}}+\left[\sum_i -(x-\mu)^2/2\sigma^2\right]$$

$$l(x)=\sum_i log\left(\sigma^2 2\pi\right)^{-1/2}+\left[\sum_i -(x-\mu)^2/2\sigma^2\right]$$

因为*i*有*n*个，所以：

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

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

再对以上式子分别对 μ,σ 求偏导，可以得到：

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

$$\mu=\bar{x}=\frac{1}{n}=\Sigma_i x_i$$

$$\nabla\sigma^2=-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_i(x_i-\mu)^2$$

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

（就是样本的伪方差(

$$\frac{1}{n}$$

), 方差是(

$$\frac{1}{n-1}$$

)

也符合直观猜测：样本的均值为总体的均值，样本的伪方差为总体的方差。

以上就是EM算法的基础。下面会用高斯混合模型来解释EM。

3. 高斯混合模型 (GMMs)以及EM算法

问题：假设有10000个人，其中男性身高和女性身高分别服从

$$N(\mu_1, \sigma_1)$$

,以及

$$N(\mu_2, \sigma_2)$$

。

GMMs的直观理解：随机变量X有K个高斯分布混合而成， 取各个高斯分布的概率为

$$\pi_1 \dots \pi_k$$

（也就是每个高斯分布出现的概率，例如有两组样本，第一组由100个男的组成，第二组由900个女的组成，随机选出男性的概率为

$$\pi_1 = 0.1$$

，选出女性的概率为

$$\pi_2 = 0.9$$

），第i个高斯分布的均值为 μ_i, 标差为 Σ_i 。若观测到一组随机变量

$$x_1 \dots x_n$$

,试估计参数

$$\pi, \mu, \Sigma$$

（都是向量）。

建立对数似然函数（目标函数）：

$$l_{\pi,\mu,\Sigma x} = \sum_{i=1}^{N} log(\sum_{k=1}^{K} \pi_k N\left(x_i|\mu_k,\Sigma_k\right))$$

其中

$$N\left(x_i|\mu_k,\Sigma_k\right)$$

是在给定第k个高斯分布的 μ 和 Σ 下第i个样本点的概率,等于

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

;

π_k 是第k个高斯分布的概率;

所以

$$\pi_k N\left(x_i|\mu_k,\Sigma_k\right)$$

就是第i个样本点属于第k个高斯分布的概率;

最后

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

其实就是似然函数取对数。

接下来，求这个似然函数的偏导为0求出极大值。但是在对数函数里面又有加和，没办法直接用求导的方程的办法直接求出最大值。所以分成2步求解:

第一步（E步）：估计数据来自K样本的概率

对于每个样本xi, 它来自第k个类别（高斯分布）的概率为:

$$\gamma\left(i,k\right) = \frac{\pi_k N\left(x_i|\mu_k,\Sigma_k\right)}{\left(\Sigma_{k=1}^{K} \pi_k N\left(x_i|\mu_k,\Sigma_k\right)\right)}$$

以上式子其实不是全概率公式，只是做归一化，将和变为1；

分母是样本分别来自k个样本概率之和，其实等于1；

其中分子是第i个样本来自k个类别的概率，也是EM的关键。

在上面式子

$$\gamma(i,k)$$

，μ 和Σ也是待估计值，因此采用迭代算法：先初始化所有k的参数 μ 和 Σ (先验给定)。

第二步（M步）：估计每个样本组中的参数

对于每个k（组/类别）而言， 可以看作是生成了

$$\gamma(i,k) \ * \ x_i(i=1,\ \ldots n)$$

(相当于贡献：比如说身高为1.9米的人来自A组的概率为0.9， 那么1.9米中1.71是由A组贡献的，剩下的0.19是由其他组贡献的) 组分k是一个标准的高斯分布，利用以上结论可得到：

之前式子(高斯分布极大似然方法估计参数)：

$$\mu = \bar{x} = \frac{1}{n} = \Sigma_i x_i$$

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

可以得到：

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

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

其中 N_k 是所有数据点属于k类的概率之和， μk 则是

$$\gamma(i,k) \ * x_i$$

之和除以

$$\gamma(i,k)$$

之和则为均值。 设为

$$\frac{1}{N_k}$$

而不是

$$\frac{1}{n}$$

是因为贡献（比如，某个样本组中，被选中的人的身高为1.9米，1.9米中1.71是由第k组贡献的，并不是完整的1.9米）

还可以得到：

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

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

其中

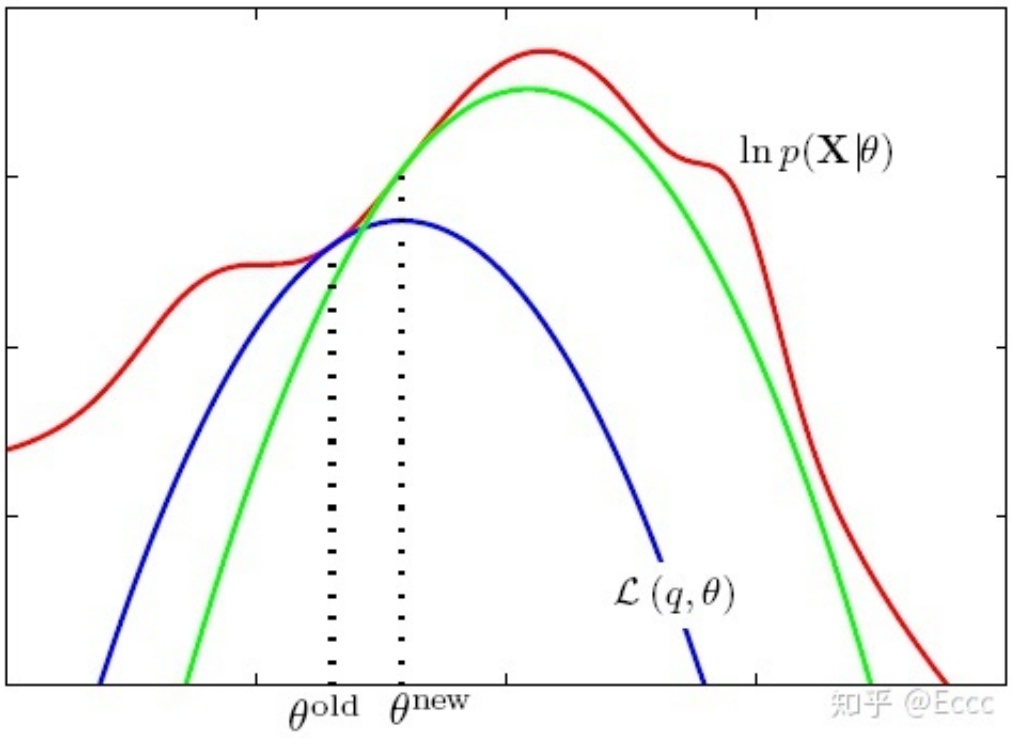
$$\frac{N_k}{N}$$

可以理解为属于k组的有 N_k 个，总共有 N 个人，则得到k组的概率。

第三步：得到这些参数之和带入第一步

$$\gamma(i,k)$$

，再通过第二步得到新的参数，如此迭代直到converge。



迭代如所示，不断更新 θ 。

小结：上述方法(第一步，第二步，第三步)就是使用了EM算法思想，EM算法对初始值很敏感（与K-means一样），如果初始值很极端，得到的结果会很差。所以EM算法仅仅只能得到局部最优。似然函数只有在指数族分布（泊松分布，指数分布）的时候会有一个全局最优点。但是GMMs是有多个极值点的。如果GMMs里面不全是高斯分布则不能使用这个方法。看到这里如果有疑惑，请继续往下看。

4. EM算法

假设有训练集{

$$x_1 \ldots . x_m$$

}, 包含m个独立样本，目标是找出该数据模型

$$p(x,z)$$

的参数。

然后使用极大似然估计，

取对数似然函数(目标函数)：

$$l(\theta) = \sum_{i=1}^{m} log\, p(x;\theta)$$

使隐变量z暴露出来：

$$l(\theta) = \sum_{i=1}^{m} log \sum_{z} p(x,z;\theta)$$

这里z是隐变量，直接找参数估计是很难的。所以策略是找个某个函数 $g(\theta)$ 小于等于 $l(\theta)$, 再用 $g(\theta)$ 的极大值代替 $l(\theta)$ 的极大值。

$$l(\theta) = \sum_{i=1}^{m} log \sum_{z^i} p\left(x^i, z^i;\theta\right)$$

令

$$Q_i\left(z^i\right)$$

是 z^i (隐变量)的分布，

$$Q_i\left(z^i\right) \geq 0$$

则有：

这里需要用到Jensen不等式：

$$l(\theta) = \sum_{i=1}^{m} log \sum_{z^i} Q_i\left(z^i\right) \frac{p(x^i, z^i;\theta)}{Q_i\left(z^i\right)}$$

$$\geq \sum_{i} \sum_{z^i} Q_i\left(z^i\right) log \frac{p(x^i, z^i;\theta)}{Q_i\left(z^i\right)}$$

本来应该是 ≤ ，但是log是凸函数如图，所以是大于等于，解释如下：

左边式子是两点沿着蓝线相连，右边式子是任意两点直线相连（黄线）， 则两条线上任取一点，蓝色一定大于黄色。

左边式子和右边式子相等的情况(最接近Jensen不等式左边的值)：

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

则：

$$Q_i\left(z^i\right) 与 p(xi, zi; \theta)$$

成正比例 (正比例是指两种相关联的量，一种量变化，另一种量也随着变化。)；

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

因为：

$$\sum_z Q_i\left(z^i\right) = 1$$

，

$$p\left(A|B\right) = \frac{P\left(AB\right)}{P\left(B\right)}$$

所以：

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

EM算法框架：

Repeat until convergence {

(E-step) For each i, set

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

(M-step) set,

$$\theta:=\mathbf{argmax}_\theta\sum_i\sum_{z^i}Q_i\left(z^i\right)log\frac{p(x^i,z^i;\theta)}{Q_i\left(z^i\right)}$$

}

5. 最后回顾GMMs

设随机变量X有K个高斯分布混合而成， 取各个高斯分布的概率为

$$\phi_1\ldots\phi_k$$

（第i个高斯分布的均值为 μ_i 标差为 Σ_i 。若观测到一组随机变量

$$x_1\ldots x_n$$

,试估计参数

$$\theta(\phi,\mu,\Sigma)$$

（都是向量）。

E-step:

M-step: 将

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

带入

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

得到：

对 μ 求偏导：

令上式为0，得：

对 Σ 求偏导，令其为0得：

对 ϕ 求偏导，得到：

因为

$$s.t. \sum_{j=1}^{k} \phi_j = 1$$

所以用拉格朗日乘子法得到：

令其偏导为0，得到：

6.总结

其实最后得到得式子与第三章是一模一样的。

$$\gamma(i,k) = w_j^i$$

$$\pi_k = \phi_j$$

最后得：

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

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

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

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

此时M-step中的

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

为最大值，所以每次迭代只需要跟新这几个值。

7.Reference

1. 李航《统计学习方法》p155-p163。

2. 最后图片来自于网络，无作者标明，侵权则删。

写在最后：

1. 这是我的第一篇博客，感谢阅读，希望你有所收获。

2. 看到知乎更新了上传文本的功能，然后用word很开心得写完博客，上传后发现公式无法显示。然后又花了1个多小时一个一个把公式输入到文章里~~~~