EM 算法

推导(证明收敛性，以及推导)

说明一些优点(不需要label，非监督学习，EM是非监督学习里面非常强的算法之一)

引入GMM,应用EM. 假设我们有数据服从2个高斯分布（画个图），对于每个x他会属于其中一个高斯，那么如果我们有个隐变量Z来告诉我们他会属于哪个高斯，那么我们就训练好了这个GMM模型。
当然一开始，我们是不可能知道这个latten variable Z，不然的话我们就已经知道了每个x属于哪个子高斯。这有点像chinken and egg的problem。首先我们来看下引入隐藏变量后x的分布。首先我们现在的x分布由p(z)来决定，我们可以通过联合概率公式推导出$p(x^i, z^i) = p(x^i|z^i)p(z^i)$, where $p(x^i|z^i = k) \sim \mathcal{N}(\mu^k, \sigma^k)$ here k referes the k^{th} sub-gaussian.

Since we dont know what is the value of z, 我们不能直接用MLE来求解（列出MLE带Z的式子来说明）。
所以，EM算法就发挥他的作用了，

首先我们要了解的是the E of EM，which is call Expectation step or E-step. 这里我们用它来求 Z的取值。

对于每个z，通过贝叶斯公式$p(z^i = k|x^i) = \frac{p(x^i|z^i=k)\cdot p(z^i=k)}{p(x^i)} = \frac{p(x^i|z^i=k)\cdot p(z^i=k)}{\sum_{k=1}^{K} p(x^i|z^i=k)\cdot p(z^i=k)}$
为了后面方便说明，这里把式子记作w^i
补充说明：$\sum_{k=1}^{K} p(z = k) = 1$

然后又因为 $p(x^i|z^i = k)$是高斯分布，我们可以带入式子
$\frac{1}{2\pi^{n/2}|\Sigma_k|^{\frac{1}{2}}} \exp -\frac{1}{2}(x^i - \mu_k)\Sigma_k^{-1}(x^i - \mu_k)$ 进行计算
$p(z^i = k)$是z属于第k个高斯的概率，通常我们对每个z初始化为 $\frac{1}{K}$

以上这就是详细的E-step

M-step 的过程

在这里我们需要更新 $(\sum_{i=1}^{m} w^i)^{new}, p(z^i = k)^{new}, \mu_k^{new}, \Sigma_k^{new}$
Here m referes to m traning data.

所谓M-step 就是通过对上面所得到的值来对我们需要得variable求最大化，最常见的做法就是求他们分别的导数

for example
since the time constrain I will directly show the result

$N_k^{\text{new}} = \sum_{i=1}^m w^i$
$p(z^i = k)^{\text{new}} = \frac{N_k^{\text{new}}}{N}$
$\mu_k^{\text{new}} = \frac{1}{N_k^{\text{new}}} \sum_{i=1}^m w^i)\mathbf{x}_n$
$\Sigma_k^{\text{new}} = \frac{1}{N_k^{\text{new}}} \sum_{i=1}^m w^i(\mathbf{x}_n - \mu_k^{\text{new}})(\mathbf{x}_n - \mu_k^{\text{new}})^T$

收敛性证明

如果我们引入Jensen不等式，我们可以知道 对于任何convex function, $f(E[x]) \leq \mathbf{E}[f(x)]$
更进一步， 如果 f(x)的二次导>0, 我们可以令 上式等号成立：$f(E[x]) = \mathbf{E}[f(x)]$. 这也等价于说x是一个常数。

在EM算法收敛性证明中我们需要用到Jesen不等式的另一种形式，那就是我们取concave function。事实上concave function可以被等价的认为是convex function 取相反值。 所以我们同样会有这些结论(but note the inqeuality sign takes take the opposite, which is $f(E[x]) \geq \mathbf{E}[f(x)],如果f(x)的二次导 < 0, f(E[x]) = \mathbf{E}[f(x)]$)

为什么要介绍这个不等式公式，以及为什么他能够证明EM的收敛性我很快会提到，现在让我先暂时记住他，然后进入下一个部分.

假设我们有了一个高斯混合模型，这里我们用变量θ来capture 其他变量(Σ, μ, etc).然后我们想通过MLE训练他， Then we will have ℓ = $\sum_{i=1}^m \log p(x^i;\theta)$ (simicolum means θ parameterize x^i) 然后这里的 $p(x^i)$ 就跟我们之前说过的一样他其实是与latten varble z的joint distribution, 所以我们rewrite这个公式就成为了ℓ = $\sum_{i=1}^m \log \sum_Z p(x^i, z^i;\theta)$

||有点不知道怎么解释$\sum Z$，因为Z是1 hot vector，所以其实是算他本身有值的地方 但是这个地方因为之前没有提过vector 我应该怎么解释呢||

对于当前这个式子ℓ = $\sum_{i=1}^m \log \sum_Z p(x^i, z^i;\theta)$我们rewrite成
ℓ = $\sum_{i=1}^m \log \sum_Z Q^i(z^i)\frac{p(x^i,z^i;\theta)}{Q^i(z^i)}$ where $Q^i(z^i)$ is probility of distriubution. And acoording to Expcetion function properties, we can rewrite it to
$\sum_{i=1}^m \log \mathbf{E}_{z^i \sim Q^i} \frac{p(x^i,z^i;\theta)}{Q^i(z^i)}$

首先我们知道log function 是一个strictlly concave function 因为他的形状是buound down（画图），然后我们这个时候就可以用之前提到的Jensen不等式，因为$\log''(x) < 0$

然后我们这个时候就可以用之前提到的Jensen不等式 得到
$\sum_{i=1}^m \log[\mathbf{E}_{z^i \sim Q^i} \frac{p(x^i,z^i;\theta)}{Q^i(z^i)}] \geq \sum_{i=1}^m \mathbf{E}_{z^i \sim Q^i}[\log \frac{p(x^i,z^i;\theta)}{Q^i(z^i)}]$

and moreover we unpack back $\sum_{i=1}^m \mathbf{E}_{z^i \sim Q^i}[\log \frac{p(x^i,z^i;\theta)}{Q^i(z^i)}]$ we can get a function $\sum_{i=1}^m \sum_Z Q^i(z^i)[\log \frac{p(x^i,z^i;\theta)}{Q^i(z^i)}]$ let denote this function as $G(\theta)$

why we try hard to get this equation? The answer is we can combiand this euqation and Jesen ineuqllity to prove EM converge.

我们首先对这个函数ℓ作图, 我们可以知道log函数是一个concave函数，然后大概函数是这个样子（画出来） 然后用刚证明的$G(\theta)$ 一定是小于等于$\ell(\theta)$，那么我们可以在$\ell(\theta)$的范围内去画出这个$G(\theta)$，初始点我们用他们取等号的情况。然后这个$G(\theta)$函数其实就是$\ell(\theta)$的lower bound，每次我们在这个$G(\theta)$上去取一个\theta最大值的点，然后下次我们用这个最大值作为初始点再次构造这个lower bound，迭代这个过程。直到他的lower bound的θ^{new}不再大于上一次的θ,这也就意味着我们找到了一个局部最优解。为什么说是局部最优解（延长log的图，发现还有更大的情况，但是取不到下一个峰值）。

总结(缺点，以及如何改进)

EM 有可能陷入局部极值，这和初始值的选取十分相关。
迭代速度慢，次数多，容易陷入局部最优；对初始值敏感

拓展(GMM +PCA +Newton method)