EM算法原理总结

> 在机器学习问题中，我们一般会通过样本的观测数据来求解模型参数，常用的方法是极大化模型分布的对数似然函数。
>
> 但是在一些情况下，我们得到的观测数据含有未观测到的隐含数据，此时我们未知的有隐含数据和模型参数，就不能直接用极大对数似然函数来求解模型参数了，EM算法就是用来求解这类问题的。
>
> EM算法是一个启发式的迭代方法，既然无法直接求出模型参数，那就先猜测出一版隐含数据(E步)，然后根据观测数据和猜测的隐含数据来求出一版模型参数(M步)；由于隐含数据是猜测得到的，所以这时候求出的模型参数还不是理想的结果，接下来继续用得到的模型参数和观测数据，再猜测出一版隐含数据(E步)，然后再用观测数据和隐含数据求出一般模型参数(M步)，如此循环下去。。。。。知道模型参数收敛为止。
>
> 从上面的描述可以看出，EM算法是迭代求解最大值的算法，同时算法在每一次迭代时分为两步，E步和M步。一轮轮迭代更新隐含数据和模型分布参数，直到收敛，即得到我们需要的模型参数。
>
> 一个最直观了解EM算法思路的是K-Means算法。在K-Means聚类时，每个聚类簇的质心是隐含数据。我们会假设K个初始化质心，即EM算法的E步；然后计算得到每个样本最近的质心，并把样本聚类到最近的这个质心，即EM算法的M步。重复这个E步和M步，直到质心不再变化为止，这样就完成了K-Means聚类。

1. EM算法的数学推导

对n个样本的观测数据为$(x_1, x_2, ..., x_n)$，求解模型参数，通常用极大对数似然估计

$$L(\theta|x) = \sum_{i=1}^{n} logP(x_i;\theta)$$

当存在不可观测数据$(z_1, z_2, ..., z_n)$时，对数似然估计为

$$L(\theta|x) = \sum_{i=1}^{n} log \sum_{z_i} P(x_i, z_i;\theta)$$

由于上式存在θ和z两个未知变量，所以不能直接求解，在上式引入一个分布 $Q(z)$

$$L(\theta|x) = \sum_{i=1}^{n} log \sum_{z_i} Q_i(z_i) * \frac{P(x_i, z_i;\theta)}{Q_i(z_i)}$$

> Jensen不等式：设A为R^k中凸集，$f(\cdot)$为A上的凸函数，即对任意$\lambda \in [0,1]$,$x, y \in A$, 有
>
> $$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y)$$
>
> 如果k维随机变量满足$P(x \in A) = 1$，则有
>
> $$f(E(x)) \le E(f(x))$$
>
> 当且仅当$f(x)$为常数时，等号成立。

由于log是一个凹函数，应用Jensen不等式

$$L(\theta|x) = \sum_{i=1}^{n} log \sum_{z_i} Q_i(z_i) * \frac{P(x_i, z_i; \theta)}{Q_i(z_i)}$$

$$\geq \sum_{i=1}^{n} \sum_{z_i} Q_i(z_i) * log \frac{P(x_i, z_i; \theta)}{Q_i(z_i)}$$

此时，如果要满足Jensen不等式的等号，则有

$$\frac{P(x_i, z_i; \theta)}{Q_i(z_i)} = c$$

c为常数

又由于$Q_i(z_i)$是一个概率密度函数，则有

$$\sum_{z} Q_i(z_i) = 1$$

由上面两式，可以得到

$$Q_i(z_i) = \frac{P(x_i, z_i; \theta)}{\sum_z P(x_i, z_i; \theta)} = \frac{P(x_i, z_i; \theta)}{P(x_i; \theta)} = P(z_i|x_i; \theta) =: q^*(z_i|\theta)$$

所以，新引入的分布$Q_i(x_i)$可以看作x_i条件下z_i的条件概率。 此时

$$L(\theta|x) = \sum_{i=1}^{n} \sum_{z_i} q^*(z_i|\theta) * log \frac{P(x_i, z_i; \theta)}{q^*(z_i|\theta)} = F(q^*(z_i|\theta), \theta)$$

假设第t步的估计值为θ_t,构造如下迭代算法

$$\theta_{t+1} = argmax_\theta \sum_{i=1}^{n} \sum_{z_i} q^*(z_i|\theta_t) * log \frac{P(x_i, z_i; \theta)}{q^*(z_i|\theta_t)}$$

由于$q^*(z_i|\theta_t) * log \frac{P(x_i,z_i;\theta)}{q^*(z_i|\theta_t)} = q^*(z_i|\theta_t) * logP(x_i, z_i; \theta) - q^*(z_i|\theta_t) * log(q^*(z_i|\theta_t))$，最后一项与$\theta$无关，可以省略 所以优化问题等价于

$$\theta_{t+1} = argmax_\theta \sum_{i=1}^{n} \sum_{z_i} q^*(z_i|\theta_t) * logP(x_i, z_i; \theta)$$

注意到，$L(\theta_{t+1}|x) \geq F(q^*(z_i|\theta_t), \theta_{t+1}) \geq F(q^*(z_i|\theta_t), \theta_t) = L(\theta_t|x)$。如果$P(x, z; \theta_{t+1})/P(x, z|\theta_t)$与$z$相关（即不为常数时），有$L(\theta_{t+1}|x) > L(\theta_t|x)$，所以，该迭代算法使得似然函数单调递增，如果$\theta_t$收敛到$\theta^*$，那么在满足一定条件下，$\theta^*$为似然函数的驻点。该性质保证了EM算法的收敛性。

2. EM算法的流程

输入：观测数据$(x_1, x_2, ..., x_m)$，联合概率分布$P(x, z; \theta)$，条件概率分布$P(z|x; \theta)$。

1. E步：计算条件概率期望

$$Q_i(z_j) = P(z_i|x_i;\theta_t)$$

$$L(\theta|x;\theta_t) = \sum_{i=1}^{m}\sum_{z_i} Q_i(x_i) * logP(x_i, z_i;\theta)$$

2. M步：极大化$L(\theta|x;\theta_t)$，得到θ_{t+1}

$$\theta_{t+1} = argmax_\theta L(\theta|x;\theta_t)$$

3. 判断θ_{t+1}是否收敛，否则继续第一步。