一文让你完全入门EM算法

原创 石头 机器学习算法那些事 2019-07-10

EM（Expectation Maximum，期望最大化）是一种迭代算法，用于对含有隐变量概率参数模型的极大似然估计或极大后验估计。模型参数的每一次迭代，含有隐变量概率参数模型的似然函数都会增加，当似然函数不再增加或增加的值小于设置的阈值时，迭代结束。

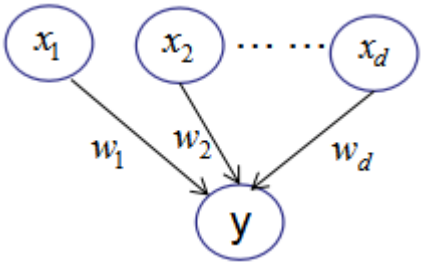
EM算法在机器学习和计算机视觉的数据聚类领域有广泛的应用，只要是涉及到后验概率的应用，我们都可以考虑用EM算法去解决问题。EM算法更像是一种数值分析方法，正确理解了EM算法，会增强你机器学习的自学能力，也能让你对机器学习算法有新的认识，本文详细总结了EM算法原理。

目录

1. 只含有观测变量的模型估计

我们首先考虑比较简单的情况，即模型只含有观测变量不含有隐藏变量，如何估计模型的参数？我们用逻辑斯蒂回归模型（logistic regression model）来解释这一过程。

假设数据集有d维的特征向量X和相应的目标向量Y，其中$X = \{x_1, x_2, ..., x_d\}$，$Y \in \{0,1\}$。下图表示逻辑斯蒂回归模型：



由之前的文章介绍，逻辑斯蒂回归模型的目标预测概率是S型函数计算得到，定义为：

$$P(Y=1|X=x)=\sigma(w\cdot x)=\frac{1}{1+\exp(-w\cdot x)} \quad (1)$$

若$P(Y=1|X=x)\geq P(Y=0|X=x)$，则目标预测变量为1；反之，目标预测变量为0。其中w是待估计的模型参数向量。

机器学习模型的核心问题是如何通过观测变量来构建模型参数w，最大似然方法是使观测数据的概率最大化，下面介绍用最大似然方法（Maximum Likelihood Approach）求解模型参数w。

假设数据集$S=\{x^{(1)},x^{(2)},...,x^{(N)}\}$，样本数据$x^{(i)}=\{x_1^{(i)},x_2^{(i)},...,x_d^{(i)}\}^T$，模型参数$w=\{w_1,w_2,...,w_d\}^T$。

观测数据的对数似然函数可写为：

$$L(w)=\log(P(data))=\log\Pi_{i=1}^{N}P(Y=y^{(i)}|x=x^{(i)})$$

由对数性质可知，上式等价于：

$$L(w)=\log(P(data))=\sum_{i=1}^{N}\log P(Y=y^{(i)}|x=x^{(i)}) \quad (2)$$

式(1)代入式(2)，得：

$$L(w)=\sum_{i=1}^{N}\log[\sigma(w^T\cdot x^{(i)})^{y^{(i)}}\cdot\sigma(-w^T\cdot x^{(i)})^{1-y^{(i)}}]$$

$$L(w)=\sum_{i=1}^{N}[y^{(i)}\cdot\log\sigma(w^T\cdot x^{(i)})+(1-y^{(i)})\cdot\log(\sigma(-w^T\cdot x^{(i)}))] \quad (3)$$

其中：

$$\sigma(w^T\cdot x^{(i)})=1-\sigma(-w^T\cdot x^{(i)}) \quad (4)$$

由于(3)式是各个样本的和且模型参数间并无耦合，因此用类似梯度上升的迭代优化算法去求解模型参数w。因为：

$$\frac{\partial\left[\sigma(w^T\cdot x^{(i)})\right]}{\partial w}=x\cdot\sigma(w^T\cdot x^{(i)})[1-\sigma(w^T\cdot x^{(i)})] \quad (5)$$

$$\frac{\partial\left[\sigma(-w^T\cdot x^{(i)})\right]}{\partial w}=-x\cdot\sigma(-w^T\cdot x^{(i)})[1-\sigma(-w^T\cdot x^{(i)})] \quad (6)$$

由式(4)(5)(6)可得：

$$\frac{\partial L(w)}{\partial w} = x^{(i)} y^{(i)} - x^{(i)} \sigma(w \cdot x^{(i)})$$

$$\frac{\partial L(w)}{\partial w} = x^{(i)} \left[y^{(i)} - \sigma(w \cdot x^{(i)}) \right]$$

因此，模型参数w的更新方程为：

$$w = w + \eta \frac{\partial L(w)}{\partial w}$$

$$w = w + \eta \left[\sum_{i=1}^{N} [1 - \sigma(w \cdot x^{(i)})] \cdot x^{(i)} \right] \quad (7)$$

其中η是学习率。

根据梯度更新方程（7）迭代参数w，似然函数L(w)逐渐增加，当似然函数收敛时，模型参数w不再更新，这种参数估计方法称为最大似然估计。

2.含有观测变量和因变量的模型参数估计

上节介绍当模型只含有观测变量时，我们用极大似然估计方法计算模型参数w。但是当模型含有隐变量或潜在变量（latent）时，是否可以用极大似然估计方法去估计模型参数，下面我们讨论这一问题：

假设V是观测变量，Z是隐变量，θ是模型参数，我们考虑用极大似然估计方法去计算模型参数：

$$L(w) = \sum_{i=1}^{N} \log P(data) = \log \Pi_{i=1}^{N} P(Y_i \mid V_i)$$

$$L(w) = \sum_{i=1}^{N} \log P(Y_i \mid V_i) = \sum_{i=1}^{N} \log \sum_{h \in Z_i} P(Y_i \mid V_i, h) \quad (8)$$

由于隐变量在log内部求和，造成不同参数间相互耦合，因此用极大似然方法估计模型参数非常难。(8)式不能估计模型参数的主要原因是隐变量，若隐变量Z已知，完全数据的似然函数为 $P(V, Y, Z \mid \theta)$，为了书写方便，观测变量V，Y统一用V表示，即$P(V, Z \mid \theta)$。

那么问题来了，如何通过已观测变量估计隐变量Z的值？这个时候我们想到了后验概率：$P(Z \mid V, \theta)$

EM算法最大化完全数据在隐变量分布的对数似然函数期望，得到模型参数θ，即：

$$\theta = \arg\max_{w} \sum_{z} P(Z \mid V, \theta^{old}) \cdot \log P(V, Z \mid \theta)$$

现在我们总结EM算法的流程：

1）初始化模型参数θ^{old}；

2）E步估计隐变量的后验概率分布：

$$P(Z|V,\theta^{old})$$

3）M步估计模型参数θ^{new}：

$$\theta^{new} = \arg\max_{\theta} \sum_{z} P(Z|V,\theta^{old}) \cdot \log P(V,Z|\theta)$$

4）当模型参数θ^{new}或对数似然函数收敛时，迭代结束；反之$\theta^{old} = \theta^{new}$，返回第（2）步，继续迭代。

3.EM算法的更深层分析

上节我们介绍了EM算法的模型参数估计过程，相信大家会有个疑问：为什么最大化下式来构建模型参数。

$$\sum_{z} P(Z|V,\theta^{old}) \cdot \log P(V,Z|\theta)$$

下面我给大家解释这一算法的推导过程以及其中蕴含的含义：

假设隐藏变量的理论分布为$q(z)$，观测数据的对数似然函数可以分解为下式：

$$\log p(V|\theta) = \sum_{z} q(z) \log p(V|\theta) \quad (9)$$

由贝叶斯理论可知：

$$p(V|\theta) = \frac{p(V,Z|\theta)}{p(Z|V,\theta)}$$

（9）式得：

$$\log p(V|\theta) = \sum_{z} q(Z) \log \frac{p(V,Z|\theta)}{p(Z|V,\theta)}$$

分子分母除q(Z)，得：

$$\log p(V|\theta) = \sum_{z} q(Z) \log \frac{p(V,Z|\theta)}{q(Z)} - \sum_{z} q(Z) \log \frac{p(Z|V,\theta)}{q(Z)}$$

$$\log p(V|\theta) = L(q,\theta) + \sum_{z} q(Z) \log \frac{q(Z)}{p(Z|V,\theta)}$$

$$\log p(V|\theta) = L(q,\theta) + KL(q \| p) \quad (10)$$

（10）式第二项表示相对熵，含义为隐变量后验概率分布与理论概率分布的差异，相对熵的一个性质是：

$$KL(q \| p) \geq 0$$

根据（10）式我们推断：

$$\log p(V|\theta) \geq L(q,\theta) \quad (11)$$

因此观测数据的对数似然函数的下界为$L(q,\theta)$，如果我们能够极大化这个下界，那么同时也极大化了可观测数据的对数似然函数。

当相对熵等于0时，即：

$$KL(q \parallel p) = \sum_Z q(Z) \log \frac{q(Z)}{p(Z|V,\theta)} = 0$$

由上式得到隐藏变量的后验概率分布与理论分布相等，即：

$$q(Z) = p(Z|V,\theta) \quad (12)$$

进而（11）式等号成立，即：

$$\log p(V|\theta) = L(q,\theta) \quad (13)$$

$L(q,\theta)$取得上界，现在我们需要最大化$L(q,\theta)$的上界，即：

$$L(q,\theta) = \sum_Z q(Z) \log \frac{p(V,Z|\theta)}{q(Z)}$$

$$L(q,\theta) = \sum_Z q(Z) \log p(V,Z|\theta) - \sum_Z q(Z) \log(q(Z)) \quad (14)$$

当相对熵等于0时，式(12)代入式(13)得到$L(q,\theta)$的上界为：

$$L(q,\theta) = \sum_Z p(Z|V,\theta) \log p(V,Z|\theta) - \sum_Z q(Z) \log(q(Z)) \quad (15)$$

式（15）的第二项对应隐变量的熵，可看成是常数，因此最大化（15）式等价于最大化$Q(\theta,\theta^{old})$，其中：

$$Q(\theta,\theta^{old}) = \sum_Z p(Z|V,\theta^{old}) \log p(V,Z|\theta) \quad (16)$$

最大化（16）式对应上节介绍EM算法的M步。

是不是对EM算法有了新的认识，本节重新整理算法EM的流程：

1）初始化模型参数为θ^{old}；

2）当等式（12）成立时，$L(q,\theta)$取得上界，最大化$L(q,\theta)$等价于最大化下式：

$$Q(\theta,\theta^{old})=\sum_{Z}p(Z\mid V,\theta^{old})\log p(V,Z\mid\theta)$$

3）最大化$Q(\theta,\theta^{old})$，返回参数θ^{new}；

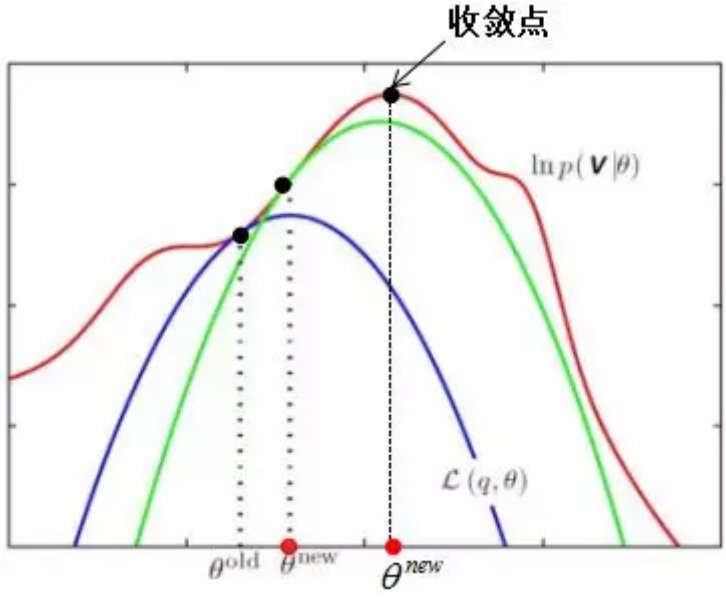
4）当$Q(\theta,\theta^{old})$收敛时，迭代结束；否则$\theta^{old}=\theta^{new}$，算法返回到第（2）步继续迭代；

为了大家清晰理解这一算法流程，下面用图形表示EM算法的含义。

E步：模型参数是θ^{old}时，由（13）式可知$\ln p(V\mid\theta)=L(q,\theta)$，用黑色实心点标记；

M步：最大化$L(q,\theta)$，返回参数θ^{new}，用红色实心点标记；

令$\theta^{old}=\theta^{new}$，重复E步和M步，当$L(q,\theta)$收敛时，迭代结束。



4.抛硬币问题举例

我们有两种硬币A和B，选择硬币A和硬币B的概率分别为π和（1-π），硬币A和硬币B正面向上的概率分别为p和q，假设观测变量为$x_i\in\{0,1\}$，1，0表示正面和反面，i表示硬币抛掷次数；隐变量$z_i\in\{0,1\}$，1，0表示选择硬币A和硬币B进行抛掷。

问题：硬币共抛掷n次，观测变量已知的情况下求模型参数$\theta=(\pi,p,q)$的更新表达式。

根据EM算法，完全数据的对数似然函数的期望：

$$Q(\theta,\theta^{(t)})=E\left[\log\prod_{i=1}^{n}[\pi p^{x_i}(1-p)^{1-x_i}]^{z_i}[(1-\pi)q^{x_i}(1-q)^{1-x_i}]^{1-z_i}\right]$$

$$=\sum_{i=1}^{n}E[z_i\mid x_i,\theta^{(t)}][\log\pi+x_i\log p+(1-x_i)\log(1-p)]$$

$$+\sum_{i=1}^{n}(1-E[z_i\mid x_i,\theta^{(t)}])[\log(1-\pi)+x_i\log q+(1-x_i)\log(1-q)] \quad (17)$$

其中$E[z_i\mid x_i,\theta^t]$表示观测数据x_i来自掷硬币A的概率，用$u_i^{(t)}$表示：

$$u_i^{(t)}=E(z_i\mid x_i,\theta^{(t)})=p(z_i=1\mid x_i,\theta^t)$$

$$=\frac{p(x_i\mid z_i,\theta^{(t)})p(z_i=1\mid\theta^{(t)})}{p(x_i\mid\theta^{(t)})}$$

$$=\frac{\pi^{(t)}[p^{(t)}]^{x_i}[1-p^{(t)}]^{1-x_i}}{\pi^{(t)}[p^{(t)}]^{x_i}[1-p^{(t)}]^{1-x_i}+(1-\pi^{(t)})[q^{(t)}]^{x_i}[1-q^{(t)}]^{1-x_i}} \quad (18)$$

最大化$Q(\theta,\theta^{(t)})$，得到如下更新表达式：

$$\frac{\partial Q(\theta,\theta^{(t)})}{\partial\pi}=0 \quad\Rightarrow\quad \pi^{(t+1)}=\frac{1}{n}\sum_i u_i^{(t)}$$

$$\frac{\partial Q(\theta,\theta^{(t)})}{\partial p}=0 \quad\Rightarrow\quad p^{(t+1)}=\frac{\sum_i u_i^{(t)}x_i}{\sum_i u_i^{(t)}}$$

$$\frac{\partial Q(\theta,\theta^{(t)})}{\partial q}=0 \quad\Rightarrow\quad q^{(t+1)}=\frac{\sum_i (1-u_i^{(t)})x_i}{\sum_i (1-u_i^{(t)})}$$

现在我们知道了模型参数θ的更新方程，假设共抛掷硬币10次，观测结果如下：1,1,0,1,0,0,1,0,1,1。

初始化模型参数为：

$$\pi^{(1)}=0.5,\ p^{(1)}=0.5,\ q^{(1)}=0.5$$

由式（18）得：

$$u_i^{(1)}=0.5$$

利用模型参数更新得：

$$\pi^{(2)}=0.5,\ p^{(2)}=0.6,\ q^{(2)}=0.6$$

由式（18），得：

$$u_i^{(2)} = 0.5 \quad i = 1,2,...,10$$

模型参数继续更新：

$$\pi^{(3)} = 0.5, p^{(3)} = 0.6, q^{(3)} = 0.6$$

因此，$Q(\theta, \theta^{(t)})$收敛时，最终的模型参数为：

$$\hat{\pi} = 0.5, \hat{p} = 0.6, \hat{q} = 0.6$$

$\hat{\pi} = 0.5$表示选择硬币A和硬币B的概率是一样的，如果模型参数的初始值不同，得到的最终模型参数也可能不同，模型参数的初始化和先验经验有关。

5.高斯混合模型的参数估计

一维变量的高斯分布：

$$N(x \mid u, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-u)^2}{2\sigma^2}}$$

其中u和σ分别表示均值和标准差。

n维变量的高斯分布：

$$N(x \mid u, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}(x-u)^T \Sigma^{-1} (x-u)}$$

其中u是n维均值向量，Σ是n×n的协方差矩阵。

n维变量的混合高斯分布：

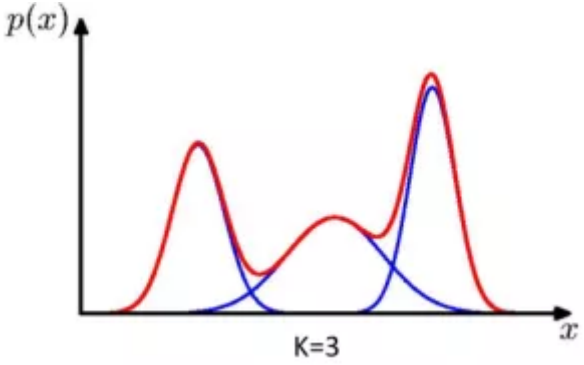
$$p(x) = \sum_{k=1}^{K} \pi_k N(x \mid u_k, \Sigma_k)$$

该分布共由k个混合成分组成，每个混合成分对应一个高斯分布，其中u_k与Σ_k是第k个高斯混合成分的均值和协方差。

π_k是归一化混合系数，含义为选择第k个高斯混合成分的概率，满足以下条件：

$$0 \le \pi_k \le 1, \sum_{k=1}^{K} \pi_k = 1$$

下图为k=3的高斯混合成分的概率分布图（红色）：



假设由高斯混合分布生成的观测数据$X=\{x_1,x_2,...,x_N\}$，其对数似然函数：

$$\ln p(X\mid\pi,u,\Sigma)=\sum_{i=1}^{N}\ln p(x_n)=\sum_{i=1}^{N}\ln\{\sum_{k=1}^{K}\pi_k N(x_n\mid u_k,\Sigma_k)\} \quad (19)$$

我们用EM算法估计模型参数，其中隐变量对应模型的高斯混合成分，即对于给定的数据x，计算该数据属于第k个高斯混合分布生成的后验概率，记为$\gamma_k(x)$。

根据贝叶斯定律：

$$\gamma_k(x)=p(k\mid x)=\frac{p(k)p(x\mid k)}{p(x)}$$
$$=\frac{\pi_k N(x\mid u_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j N(x\mid u_j,\Sigma_j)} \quad (20)$$

最大化式（19），令

$$\frac{\partial\ln(X)}{\partial u_j}=0 \quad (21)$$

$$\frac{\partial\ln(X)}{\partial\Sigma_j}=0 \quad (22)$$

$$\frac{\partial\ln(X)}{\partial\pi_j}=0 \quad (23)$$

由式（20）（21）（22）（23）可得模型参数：

$$u_j=\frac{\sum_{n=1}^{N}\gamma_j(x_n)x_n}{\sum_{n=1}^{N}\gamma_j(x_n)} \quad (24)$$

$$\Sigma_j = \frac{\sum_{n=1}^{N} \gamma_j(x_n)(x_n - u_j)(x_n - u_j)^T}{\sum_{n=1}^{N} \gamma_j(x_n)} \quad (25)$$

$$\pi_j = \frac{1}{N} \sum_{n=1}^{N} \gamma_j(x_n) \quad (26)$$

下面小结EM算法构建高斯混合模型的流程：

1）初始化高斯混合模型的均值u_j，协方差Σ_j和混合系数π_j，计算完全数据的对数似然值（式（19））；

2）E步：使用当前的参数值，通过下式计算均值：

$$\gamma_k(x) = \frac{\pi_k N(x \mid u_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x \mid u_j, \Sigma_j)}$$

$\gamma_k(x)$表示观测数据x属于第k个高斯混合成分的后验概率；

3）M步：最大化对数似然函数，得到式（24）（25）（26）的模型更新参数；

4）根据更新的参数值，重新计算完全数据的对数似然函数：

$$\ln p(X \mid \pi, u, \Sigma) = \sum_{i=1}^{N} \ln p(x_n) = \sum_{i=1}^{N} \ln \{ \sum_{k=1}^{K} \pi_k N(x_n \mid u_k, \Sigma_k) \}$$

若收敛，则得到最终的模型参数值；反之，回到算法第（2）步继续迭代。

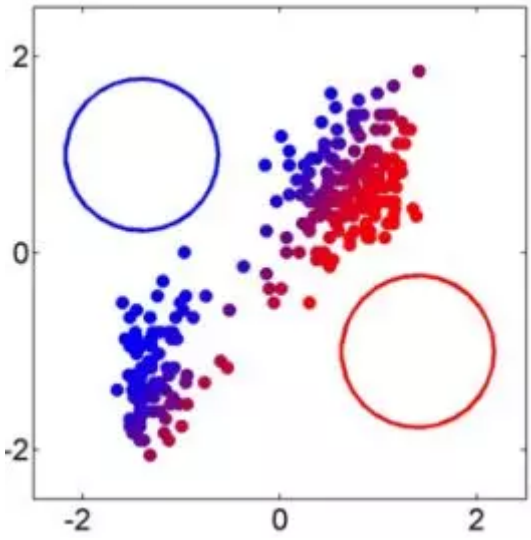
6.聚类蕴含的EM算法思想

我们可以把聚类理解为：计算观测数据x属于不同簇类的后验概率，记为$\gamma_j(x)$，其中j是簇类个数（j=1,2,...,K），观测数据x所属的簇标记λ_j由如下确定：

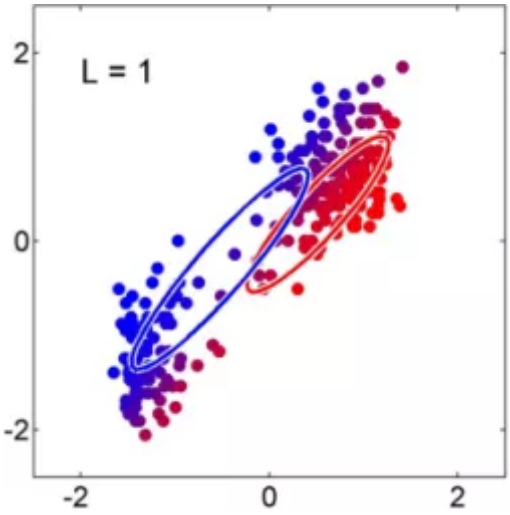
$$\lambda_j = \arg\max_{j \in \{1,2,\ldots,k\}} \gamma_j(x) \quad (27)$$

我们可以用EM算法计算每个样本由不同高斯混合成分生成的后验概率，步骤可参考上一节。

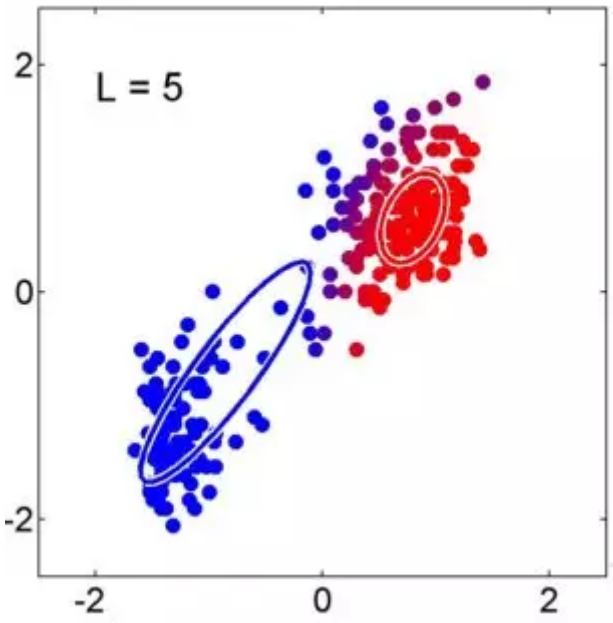
【例】 如下的观测数据，假设簇类个数K=2，初始化每个高斯混合参数得到$\gamma_j(x)$，根据式（27）得到聚类结果：



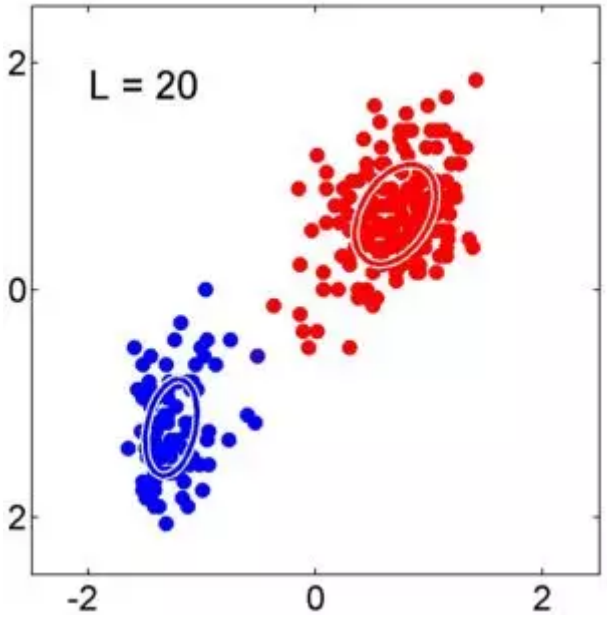
根据上一节介绍的EM算法步骤，迭代1次后得到$\gamma_j(x)$，根据式（27）得到聚类结果：



迭代5次后得到$\gamma_j(x)$，根据式（27）得到聚类结果：



迭代20次后的$\gamma_j(x)$，根据式（27）得到聚类结果：



k均值聚类是高斯混合聚类的特例，k均值假设各个维是相互独立的，其算法过程也可用EM思想去理解：

1）初始化簇类中心；

2）E步：通过簇类中心计算每个样本所属簇类的后验概率$\gamma_j(x)$；

3）M步：最大化当前观测数据的对数似然函数，更新簇类中心

4）当观测数据的对数似然函数不再增加时，迭代结束；反之，返回（2）步继续迭代；

7.小结

EM算法在各领域应用极广，运用了后验概率，极大似然方法和迭代思想构建最优模型参数，后续文章会介绍EM算法在马尔科夫模型的应用，希望通过这篇文章能让读者对EM算法不再陌生。

参考

https://towardsdatascience.com

推荐阅读

k-means聚类算法原理总结