推理 E-Step 和 M-Step 的解析表达式

目标

利用 EM算法，对二元高斯混合模型（Gaussian Mixture Model）来推算出来 E-Step 和 M-Step 的解析表达式。

分析

首先了解下专有名词的解释和背后的意义，以及对这个目标提出的一些问题。

EM 算法 (Expectation Maximization Algorithm) 即最大期望算法。作用主要是求参数极大似然估计的一种方法，它可以从非完整数据集中对参数进行估计。

Q1：这里我们引申出一个问题，极大似然估计是干什么的？他的意义是什么？为啥要求它。

极大似然估计方法是求参数估计的一种方法。在实际生活中我们往往知道一些事物的结果从而反推导致这些样本结果的模型参数，最终建模成功来进行新样本预估。参数估计就是通过若干次试验，观察其结果，利用结果推出参数的大概值。

Q2：那为什么又叫极大似然估计，似然是什么意思，和概率什么区别

对于函数P(x|θ)输入有两个输入：x表示某一个具体的数据；θ表示模型的参数。

如果θ是已知确定的，x是变量，这个函数叫做概率函数(probability function)，它描述对于不同的样本点x，其出现概率是多少。

如果x是已知确定的，θ是变量，这个函数叫做似然函数(likelihood function), 它描述对于不同的模型参数，出现x这个样本点的概率是多少。

求最大似然函数估计值的一般步骤：
（1）写出似然函数；
（2）对似然函数取对数，并整理；
（3）求导数，令导数为 0，得到似然方程；
（4）解似然方程，得到的参数即为所求；

极大似然估计，通俗理解来说，就是利用已知的样本结果信息，反推最具有可能（最大概率）导致这些样本结果出现的模型参数值！

而EM 算法就是来求解参数极大似然估计。

Q3： EM 算法怎么能做到求解最合适的参数，怎么做到？核心思想是什么？

由于我们一开始在预估的时候，并不能给定最好的值是多少，所以我们先可以拍脑袋定一个，不然就永远陷入先有鸡还是现有蛋的逻辑。假设我们想估计知道A和B两个参数，在开始状态下二者都是未知的，但如果知道了A的信息就可以得到B的信息，反过来知道了B也就得到了A。可以考虑首先赋予A某种初值，以此得到B的估计值，然后从B的当前值出发，重新估计A的取值，这个过程一直持续到收敛为止。

EM 算法的 E步骤：根据参数初始值或上一次迭代的模型参数来计算出隐性变量的后验概率，其实就是隐性变量的期望。作为隐藏变量的现估计值。M步骤 将似然函数最大化以获得新的参数值，这个不断的迭代，就可以得到使似然函数L(θ)最大化的参数θ了。 所以其实其核心思想就是不断的重复 E step 和 M step， 持续收敛到最佳状态。稍后我们会进一步证明 E step 和 M step

Q4：二元高斯混合模型是什么？

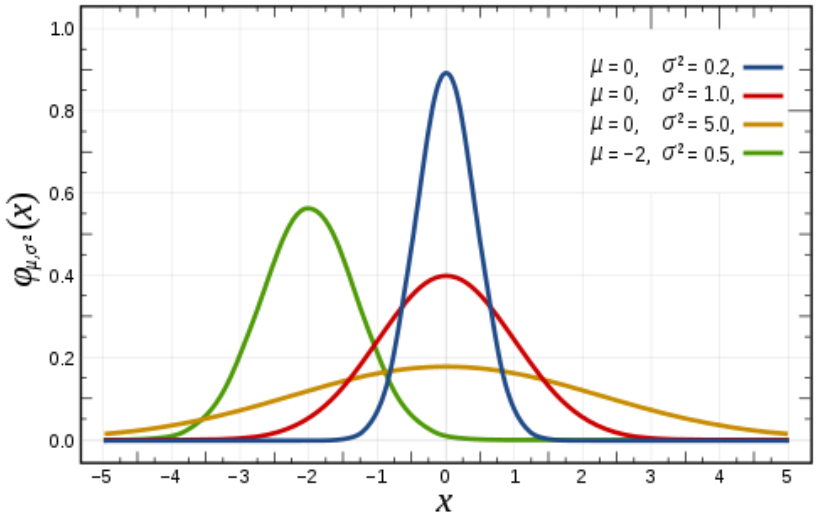
高斯分布（Gaussian Distribution)，又称为正态分布。若随机变量 X 服从一个位置参数为μ 、尺度参数为σ的概率分布，记为：

$$X \sim N(\mu, \sigma^2)$$

则其概率密度函数为:

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

正态分布的数学期望值 μ 等于位置参数，决定了分布的位置；其方差 σ2 的开平方或标准差 σ 等于尺度参数，决定了分布的幅度。



高斯混合模型，就是多个高斯模型的线性组合，二元就是两个线性组合。

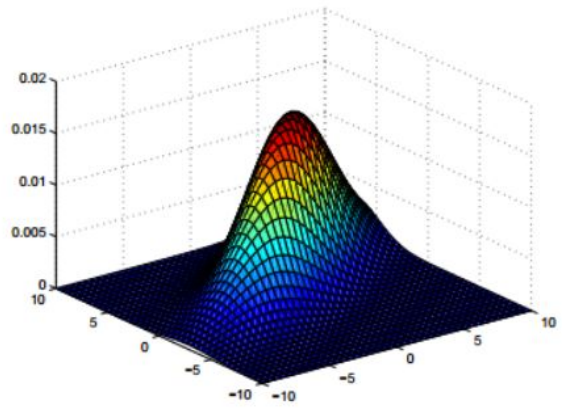
如果是多元高斯

多维变量 X 服从高斯分布时，它的概率密度函数为：

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

μ为模型期望，∑为模型协方差矩阵。

二元高斯如下：



Q5 ：为什么预估要用高斯分布？不用别的函数分布

高斯混合模型（Gaussian Mixture Model，GMM），顾名思义，就是多个高斯模型的线性组合。由于高斯分布的普遍性（可由中心极限定理得到），在人类世界中，几乎所有的数据都可近似满足高斯分布，问题只不过是用多少个高斯分布能更精确的描述数据的分布。当然，应用其他分布建立混合模型，理论上也能构成复杂的描述模型的，但是高斯函数还具有良好的数学性质，所以能被广泛的应用。

另外高斯混合分布首先将待解决的问题转换为包含隐变量（即每条样本属于不同类别的概率）和模型参数的极大似然估计问题。由于该极大似然估计问题中包含隐变量和模型参数，所以无法用传统的求偏导的方法求得。这时，需要利用EM算法，即期望最大化算法求解参数。

理解了这么多概念和问题之后，我们来推理E-Step 和M-Step 的解析表达式。

推理

高斯混合模型主要分为两步，Expectation（期望）和 Maximization（最大化），其实也就是 E-Step 和 M-Step

开始推理：

假设观察到的样本都是独立的，那么其似然的概率是各自概率的积：

为了方便计算，我们对概率积加上log，得到log似然函数，变积为和

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

其中θ是三个模型的参数，每个聚类对应高斯分布的均值为u 方差Σ

由于样本分类也是未知的，所以这个问题中包含一个隐含变量，每个样本属于不同类别的概率

将该隐变量求和：

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

目标求解上式最大值，得到参数，由于存在隐变量，无法直接求偏导，进行上式转化

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

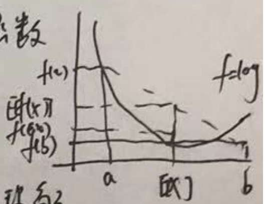
上下分子分母乘以一个Q函数

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

Jensen不等式 由于log为凹函数

所以 $E[f(x)]\geq f(E[x])$

右图的E[f(x)]为 f(E[x]) 可以理解



E步骤：为了求最下界，为了让上面不等式成立，则

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

由贝叶斯公式：$P(B|A)=\frac{P(B)P(A|B)}{P(A)}$ 得：

$$P(z^{(i)}=j|x^{(i)};\phi,u,\Sigma)=\frac{p(x^{(i)}|z^{(i)}=j;u,\Sigma)p(z^{(i)}=j;\phi)}{\sum_{l=1}^{k}p(x^{(i)}|z^{(i)}=l;u,\Sigma)p(z^{(i)}=l;\phi)}$$

M步骤：最大化其最小下界，由于最小下界的隐含变量在E中求得，直接利用极大似然估计求解：

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

对每个参数求偏导 令其为0 求参数

得

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

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

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

推理结束，整个推理过程中比较重要的是引入了Q 函数，利用了Jensen 不等式，使用Bayes 公式，通过似然函数对各个参数求偏导解出结果。

思考

EM 用来解决聚类的问题，而不像前几节课学的SVM，LR，决策树是做分类相关的。聚类和分类的区别在于，待聚类的样本标签是未知的，也就是模型包含隐变量的情况下，需要根据样本分布情况，将样本聚成不同的簇，每一簇代表相似的群体。而且现实生活中样本混合在一起，很多时候我们往往无法得知采样样本是来自哪个类。

EM 算法的核心，其实就是不断的重复 E step 和 M step，持续收敛到最佳状态，E-Step 就相当于猜参数，反思（M-step）,重复上述步骤来选出最好的参数。

由于存在隐含变量，不能直接最大化l(θ)l(θ)，所以只能不断地建立ll的下界（E-step），再优化下界（M-step），依次迭代，直至算法收敛到局部最优解。这种思想应该可以推广举一反三用到更多的生活应用上。