# 混合高斯模型（GMM）

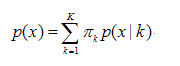
1、统计学习的模型有两种，一种是概率模型，一种是非概率模型。

（1）概率模型，是指训练模型的形式是P(Y|X)。输入是X，输出是Y，训练后模型得到的输出不是一个具体的值，而是一系列的概率值（对应于分类问题来说，就是输入X对应于各个不同Y（类）的概率），然后我们选取概率最大的那个类作为判决对象（软分类--soft assignment）。

（2）非概率模型，是指训练模型是一个决策函数Y=f(X)，输入数据X是多少就可以投影得到唯一的Y，即判决结果（硬分类--hard assignment）。

2、混合高斯模型（GMM）就是指对样本的概率密度分布进行估计，而估计采用的模型（训练模型）是几个高斯模型的加权和（具体是几个要在模型训练前建立好）。每个高斯模型就代表了一个类（一个Cluster）。对样本中的数据分别在几个高斯模型上投影，就会分别得到在各个类上的概率（数据点服从未知参数的高斯分布）。然后我们可以选取概率最大的类所为判决结果。

混合高斯模型的定义公式如下：



K 为模型的个数；

πk为第k个高斯的权重；

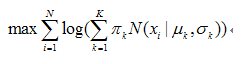
p（x / k） 则为第k个高斯概率密度，其均值为μk，方差为σk。

对此概率密度的估计就是要求出πk、μk 和σk 各个变量。当求出p(x)的表达式后，求和式的各项的结果就分别代表样本x 属于各个类的概率。

=====>>>>

对混合高斯模型的定义公式做参数估计，采用最大对数似然法，再对最大对数似然法公式采用EM算法：

1. 将混合高斯模型的定义公式改写成：



（2）EM算法——

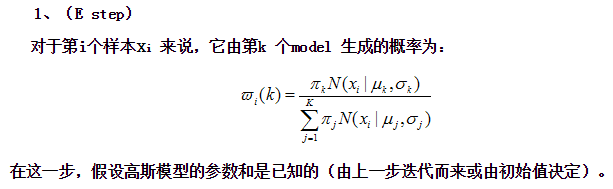
将求解分为两步：

第一步,假设知道各个高斯模型的参数（可以初始化一个，或者基于上一步迭代结果），去估计每个高斯模型的权值；

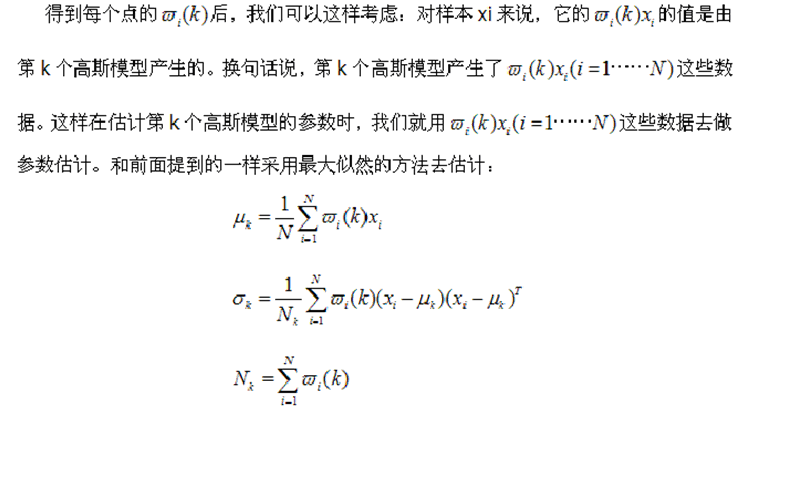
第二步,基于估计的权值，回过头再去确定高斯模型的参数。重复这两个步骤，直到波动很小，近似达到极值（注意这里是极值不是最值，EM算法会陷入局部最优）。

确定出参数，就确定了对数似然函数，再带入数据点就可以得到概率。

具体描述如下：



2、



3、重复上述两步骤直到算法收敛。

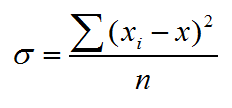
1. GMM 协方差（Covariance）

如果协方差为正，说明X，Y同向变化，协方差越大说明同向程度越高；如果协方差为负，说明X，Y反向运动，协方差越小说明反向程度越高。

（https://www.zhihu.com/question/20852004）

尽管GMM通常被用来聚类，但是我们还可以比较 得到的簇和数据集中真实的类别。我们需要初始化the means of the Gaussians with the means of the classes from the training set to make this comparison valid.

概率论中方差用来度量[随机变量](https://baike.baidu.com/item/%E9%9A%8F%E6%9C%BA%E5%8F%98%E9%87%8F" \t "https://baike.baidu.com/item/_blank)和其[数学期望](https://baike.baidu.com/item/%E6%95%B0%E5%AD%A6%E6%9C%9F%E6%9C%9B" \t "https://baike.baidu.com/item/_blank)（即[均值](https://baike.baidu.com/item/%E5%9D%87%E5%80%BC" \t "https://baike.baidu.com/item/_blank)）之间的偏离程度。



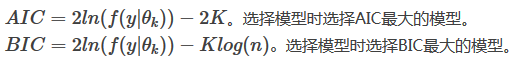
当数据分布比较分散（即数据在平均数附近波动较大）时，各个数据与平均数的差的平方和较大，方差就较大；当数据分布比较集中时，各个数据与平均数的差的平方和较小。因此方差越大，数据的波动越大；方差越小，数据的波动就越小。[方差](https://baike.baidu.com/item/%E6%96%B9%E5%B7%AE" \t "https://baike.baidu.com/item/%E5%8D%8F%E6%96%B9%E5%B7%AE/_blank)是协方差的一种特殊情况，即当两个变量X,Y是相同的情况。

1. 选择高斯混合模型的分量的数量

模型选择同时考虑协方差类型和分量数。

模型选择主要有两个思路：   
 1.解释性框架   
在已有数据下，重点关注哪些变量是模型的重要变量，模型的形式应该怎样。好的模型应该是最能解释现有数据的模型。   
 2.预测性框架   
重点关注哪些变量是模型的潜在变量以及模型的可能形式。好的模型应该是最能预测结果的模型。

给定数据下存在一个最佳的模型，且该模型可以通过已有数据估计出来，根据某个选择标准选择出来的模型，用它所做的推理应该是最合理的。这个选择标准就可以是AIC和BIC。没有模型的选择的绝对标准，好的选择标准应该根据数据分布不同而不同，并且要能融入到统计推理的框架中去。

（？？？？）