高斯混合模型--**GMM**（**Gaussian Mixture Model**）

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。

所谓概率模型，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**）。所谓非概率模型，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**）。

所谓混合高斯模型（**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。然后我们可以选取概率最大的类所为判决结果。

从中心极限定理的角度上看，把混合模型假设为高斯的是比较合理的，当然，也可以根据实际数据定义成任何分布的**Mixture Model,**不过定义为高斯的在计算上有一些方便之处，另外，理论上可以通过增加**Model**的个数，用**GMM**近似任何概率分布。混合高斯模型的定义为：

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

其中**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。在做参数估计的时候，常采用的是最大似然方法。最大似然法就是使样本点在估计的概率密度函数上的概率值最大。由于概率值一般都很小，**N** 很大的时候**,** 连乘的结果非常小，容易造成浮点数下溢。所以我们通常取**log**，将目标改写成：

$$\max\sum_{i=1}^{N}\log\,p(x_i)$$

也就是最大化对数似然函数，完整形式为：

$$\max\sum_{i=1}^{N}\log(\sum_{k=1}^{K}\pi_k N(x_i\mid\mu_k,\sigma_k))$$

一般用来做参数估计的时候，我们都是通过对待求变量进行求导来求极值，在上式中，**log**函数中又有求和，你想用求导的方法算的话方程组将会非常复杂，没有闭合解。可以采用的求解方法是**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）。具体表达如下：

**1、（E step）**

对于第**i**个样本**xi** 来说，它由第**k** 个**model** 生成的概率为：

$$\varpi_i(k)=\frac{\pi_k N(x_i\mid\mu_k,\sigma_k)}{\sum_{j=1}^{K}\pi_j N(x_i\mid\mu_j,\sigma_j)}$$

在这一步，假设高斯模型的参数和是已知的（由上一步迭代而来或由初始值决定）。

**2、（M step）**

得到每个点的$\varpi_i(k)$后，我们可以这样考虑：对样本 xi 来说，它的$\varpi_i(k)x_i$的值是由第 k 个高斯模型产生的。换句话说，第 k 个高斯模型产生了$\varpi_i(k)x_i(i=1\cdots\cdots N)$这些数据。这样在估计第 k 个高斯模型的参数时，我们就用$\varpi_i(k)x_i(i=1\cdots\cdots N)$这些数据去做参数估计。和前面提到的一样采用最大似然的方法去估计：

$$\mu_k=\frac{1}{N}\sum_{i=1}^{N}\varpi_i(k)x_i$$

$$\sigma_k=\frac{1}{N_k}\sum_{i=1}^{N}\varpi_i(k)(x_i-\mu_k)(x_i-\mu_k)^T$$

$$N_k=\sum_{i=1}^{N}\varpi_i(k)$$

**3、**重复上述两步骤直到算法收敛。