高斯混合模型 (Gaussian Mixture Model, GMM)

(1) Likelihood & Log-Likelihood

$X = \{x_1, x_2, x_3, \cdots, x_N\}$ 共有N条数据，并且 $x_i \overset{iid}{\sim} P(x|\theta)$ → 即：当确定θ后，所有的数据都是独立同分布所得的

Log-Likelihood

$$L(\theta|X) = \sum_{i=1}^{N} \log P(x_i|\theta)$$

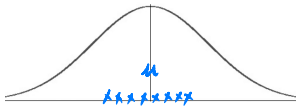
→ 为什么要log化？
① 变Π为Σ，简化计算
② $P(x_1) > P(x_2) \Rightarrow \log P(x_1) > \log P(x_2)$，具有严格单调性
③ 因为概率值都是 $0<P<1$ 的，若干P连乘会造成计算时的 under float 问题，用log将相乘转化为相加，则可解决这个问题

求它的最大值，即MLE

只对斯公式中的 Likelihood

(2) 单个高斯模型

当数据来自单个高斯模型时，样本数据的分布将呈现为如下形状



(若这些数据来自单一的高斯分布，那么越靠近μ，数据越集中)

⇓

$$L(\theta|X) = L(\mu, \Sigma|X)$$
$$= \arg\max_{\theta} \sum_{i=1}^{N} \log N(x_i|\theta)$$
$$= \arg\max_{\theta} \sum_{i=1}^{N} \log N(x_i|\mu, \Sigma)$$

然后，分别对 μ、Σ 求导（并令其导数为0即可）

→ 即：μ_{MLE} = 所有样本的均值

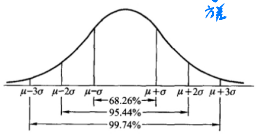
$\boxed{\mu_{MLE}}$ $\frac{\partial}{\partial \mu} L(\mu, \Sigma|x) = 0 \Rightarrow \mu_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$

$\boxed{\Sigma_{MLE}}$ $\frac{\partial}{\partial \Sigma} L(\mu, \Sigma|x) = 0 \Rightarrow \Sigma_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

一元高斯分布

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

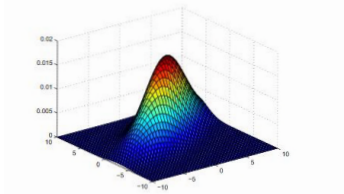
→ 均值　↓ 方差



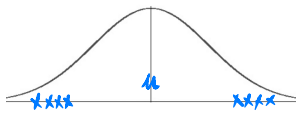
多元高斯分布

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

→ 协方差矩阵



(2) 混合高斯模型



Q：这样的数据会是来自单一高斯模型吗？

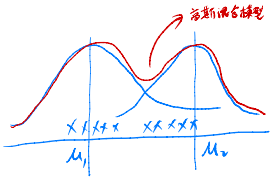
A：不可能！

因为根据MLE，我们求得 μ = 样本的平均值

而根据高斯分布的特点，越靠近中心(μ)采样所得数应该更密集才对。

⇓

那么，这样的数据分布可能是多个高斯模型的混合分布所得。



混合高斯模型的参数：

$$\theta = \{\mu_1, \mu_2, \cdots, \mu_k, \Sigma_1, \Sigma_2, \Sigma_3, \cdots, \Sigma_k, \alpha_1, \alpha_2, \cdots \alpha_{k-1}\}$$

↳ 每个高斯模型在"混合高斯模型"中所占的比重（必须要加上符合概率要求）

因为当 $\alpha_1 + \alpha_2 + \cdots + \alpha_{k-1}$ 确定后，α_k 的值也就确定了。
所以，参数中只需要 k-1 个自由度。

并且，$\alpha_1 + \alpha_2 + \cdots \alpha_{k-1} + \alpha_k = 1$

$$P(x_i|\theta)=\sum_{l=1}^{k}\mathcal{N}(\mu_l,\Sigma_l) \quad ✗$$

$$\Rightarrow \int_{x_i} P(x_i|\theta)=\int_{x_i}\sum_{l=1}^{k}\mathcal{N}(\mu_l,\Sigma_l)\,dx \quad ✗$$

积分结果不确定1

每个高斯在积分后都是1
那么k个高斯累加后的积是k，不符合概率定义

解决的办法就是为高斯混合模型中的不同高斯分布分配不同的权重 α_i，并使得 $\sum_{l=1}^{k}\alpha_l=1$

即：

$$P(x_i|\theta)=\sum_{l=1}^{k}\alpha_l\,\mathcal{N}(\mu_l,\Sigma_l) \quad ✓$$

$$\Downarrow$$

当数据来自多个高斯模型时，

$$\mathcal{L}(\theta|X)=\mathcal{L}(\mu,\Sigma|X)$$

$$=\arg\max_{\theta}\sum_{i=1}^{N}\log\left[\sum_{l=1}^{k}\alpha_l\,\mathcal{N}(x_i|\mu_l,\Sigma_l)\right]$$

数据量

法l高斯模型中，高斯模型的个数

服从 $\mathcal{N}(\mu_l,\Sigma_l)$ 分布的数据为 x_i

然后，分别对 μ、Σ 求导（并令其导数为0即可）

$$\boxed{\mu_{MLE_1}}\quad \frac{\partial}{\partial\mu}\mathcal{L}(\mu_1,\mu_2,\cdots,\mu_k,\Sigma_1,\Sigma_2,\cdots\Sigma_k,\alpha_1,\alpha_2,\cdots,\alpha_{k-1}|X)=0$$

$$\boxed{\mu_{MLE_2}}\quad \frac{\partial}{\partial\mu}\mathcal{L}(\mu_1,\mu_2,\cdots,\mu_k,\Sigma_1,\Sigma_2,\cdots\Sigma_k,\alpha_1,\alpha_2,\cdots,\alpha_{k-1}|X)=0$$

$$\boxed{\mu_{MLE_3}}\quad \frac{\partial}{\partial\mu}\mathcal{L}(\mu_1,\mu_2,\cdots,\mu_k,\Sigma_1,\Sigma_2,\cdots\Sigma_k,\alpha_1,\alpha_2,\cdots,\alpha_{k-1}|X)=0$$

$$\vdots$$

$$\boxed{\Sigma_{MLE_1}}\quad \frac{\partial}{\partial\Sigma}\mathcal{L}(\mu_1,\mu_2,\cdots,\mu_k,\Sigma_1,\Sigma_2,\cdots\Sigma_k,\alpha_1,\alpha_2,\cdots,\alpha_{k-1}|X)=0$$

$$\vdots$$

$$\boxed{\alpha_{MLE_1}}\quad \frac{\partial}{\partial\alpha}\mathcal{L}(\mu_1,\mu_2,\cdots,\mu_k,\Sigma_1,\Sigma_2,\cdots\Sigma_k,\alpha_1,\alpha_2,\cdots,\alpha_{k-1}|X)=0$$

$$\vdots$$

$$\boxed{\alpha_{MLE_{k-1}}}\quad \frac{\partial}{\partial\alpha}\mathcal{L}(\mu_1,\mu_2,\cdots,\mu_k,\Sigma_1,\Sigma_2,\cdots\Sigma_k,\alpha_1,\alpha_2,\cdots,\alpha_{k-1}|X)=0$$

$\xrightarrow{\text{计算复杂度高}}$

因为，$\log(A\cdot B)=\log A+\log B$ → 容易计算

$\log(A+B)=$ 憋逼 → 计算复杂

怎么办？

$$\Downarrow$$

EM算法