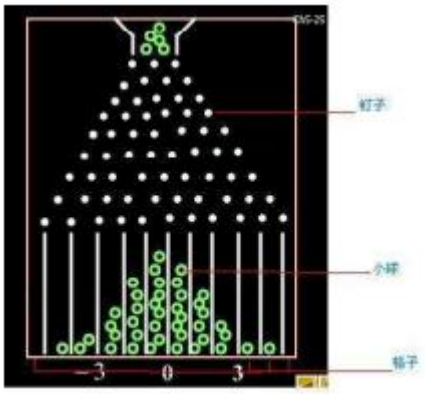
EM 算法与 GMM 模型

单个高斯分布 GM 的估计参数

关于高斯分布

自然界中，很多事物的取值服从高斯分布
高尔顿钉板



每次弹珠往下走的时候，碰到钉子会随机往左还是往右走，可以观测到多次随机过程结合的结果趋近于正太分布

给定一组样本 X1,X2,...Xn，已知它们来自于高斯分布 N(μ,σ)。如何估计μ，σ？

高斯分布的概率密度函数：

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

将 X 的取值带入 f(x)得到每个 x 取值的概率表达形式(带着μ，σ)

高斯分布的似然函数

X1-XN 全部发生的总概率为：

$$L(x)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{(x_i-\mu)^2}{2\sigma^2}}$$

根据 MLE 思想，让 L 取得最大值的μ，σ 就是我们估计出来最合理的μ，σ

求似然公式最大，取对数

$$\begin{aligned}l(x)&=\log\prod_{i}\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{(x_i-\mu)^2}{2\sigma^2}}\\&=\sum_{i}\log\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{(x_i-\mu)^2}{2\sigma^2}}\\&=\left(\sum_{i}\log\frac{1}{\sqrt{2\pi}\sigma}\right)+\left(\sum_{i}-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\\&=-\frac{n}{2}\log(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i}(x_i-\mu)^2\end{aligned}$$

分别对μ和σ求偏导，令偏导数=0，得到

$$\mu=\frac{1}{n}\sum_{i}x_i$$

$$\sigma^2=\frac{1}{n}\sum_{i}(x_i-\mu)^2$$

混合高斯分布 GMM 的参数估计

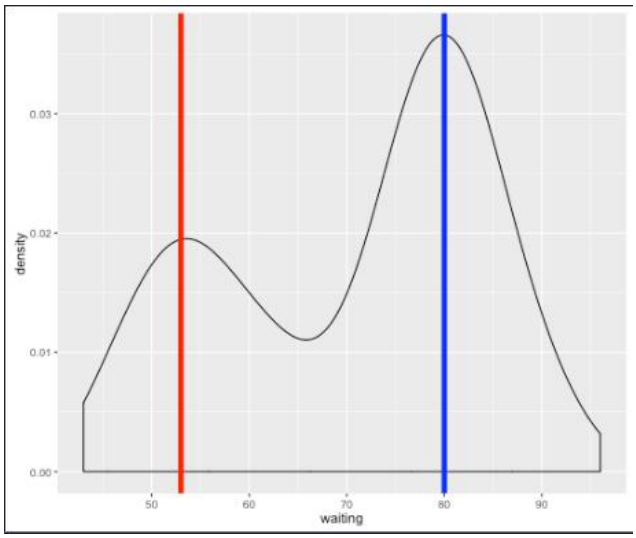
说个关于身高的例子

人群中随机选出 10000 个人来，测量他们的身高

男女身高虽然都服从高斯分布，但是方差均值不同

假设男 N(μ1，σ1) 女 N(μ2，σ2)

能否估计 μ1 σ1？ μ2 σ2？



GMM 混合高斯分布

假设随机变量 X 是由 K 个高斯分布混合而来，取到各个高斯分布的概率为Π1，Π2 直到Πk，第 i 个高斯分布的均值为μi，标准差为σi，若只观测到一系列样本 X1,X2,X3...Xn，试估计Π，μ，σ

这里需要提一下的是Π1 和Π2 等可以理解为先验概率，比如男人女人的例子，10000 个人里面根据经验知道男人大概 8000 个，女人大概 2000 个，这样男人的Π1 就是 0.8，女人的Π2 就是 0.2，它俩加一起就是 1

理解 GMM 模型的似然函数

$$l_{\pi,\mu,\Sigma}(x)=\sum_{i=1}^{N}\log\left(\sum_{k=1}^{K}\pi_k N(x_i\mid\mu_k,\Sigma_k)\right)$$

这里的 notation 我们先来理解一下

Πk=P(X∈Zk)，Zk 是某个类别

Σk=σk

N 表示是正太分布的概率密度函数，当知道是哪个分布的情况下，公式里带入 Xi，μk，σk 就可以得到概率 P(X=Xi|X∈Zk)，Xi 属于类别 k 的条件下这条样本出现的概率

条件概率：

$$P(A|B)=\frac{P(AB)}{P(B)}$$

两者一乘等于 P(X=Xi,Xi∈Zk)联合分布概率

全概率公式：

$$P(A)=\sum_i P(A|B_i)P(B_i)$$

这里有Σ加和从 k=1 到 k，把每种情况全考虑到了，那么就是 P(Xi)
所以前面有个Σ加和，log，说明这个公式小 l 代表的就是 MLE 的对数似然形式嘛！！！

作业

1，理解 GM 和 GMM 的区别在于什么？
2，理解 GM 高斯分布参数估计和多元线性回归假设误差服从高斯分布的参数估计区别？
3，单个高斯分布的参数估计μ和σ如何推导出来？
4，正向的把 GMM 的对数似然函数式子推出来？

分两步求解 GMM

解决方法：
EM 算法！！！

第一步：估计数据来源于哪个分布

$$\gamma(i,k)=\frac{\pi_k N(x_i\mid\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j N(x_i\mid\mu_j,\Sigma_j)}$$

这里γ(i，k)=P(X∈Zk|X=Xi)，这和分子里面的 N(Xi|μk，σk)是 P(X=Xi|Xi∈Zk)可不一样！

贝叶斯(Bayes)公式：

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

也就是说γ(i，k)其实是根据贝叶斯公式来的，Πk 是 P(X∈Zk)，就是先验概率
分母就是 P(X=Xi)

$$P(x\in z_k \mid x=x_i)=\frac{P(x\in z_k)\cdot P(x=x_i \mid x\in z_k)}{P(x=x_i)}$$

$$=\frac{P(x=x_i, x\in z_k)}{\sum_{j}^{K} P(x\in z_j)\cdot P(x=x_i \mid x\in z_j)}$$

这里我们是知道 X1..Xi..Xn，然后去看 P(X∈Zk|X=Xi)属于哪个类别的概率最大，这样我们就可以去 soft assignment ,换一个角度也可以看成第 k 个类别在生成这个样本时做出的贡献比

如果我们可以估计出来三组Π，μ，σ，我们就可去求，或者说我们如果已经随机出来三组Π，μ，σ，我们是不是就可以去求概率γ，如果我们有一个人的身高是 1 米 8，我们是不是就要去求在男和女中的概率值

比如：

γ(1.8,男)=0.7

γ(1,8,女)=0.3

第 0 步：确定Π，μ，σ，哪怕是随机出来的

然后第一步，带入公式里面去求出γ

举例讲解 GMM 流程

GMM 是分维度去估计参数的，所以当维度很多的时候 GMM 的计算量可想而知是很大的

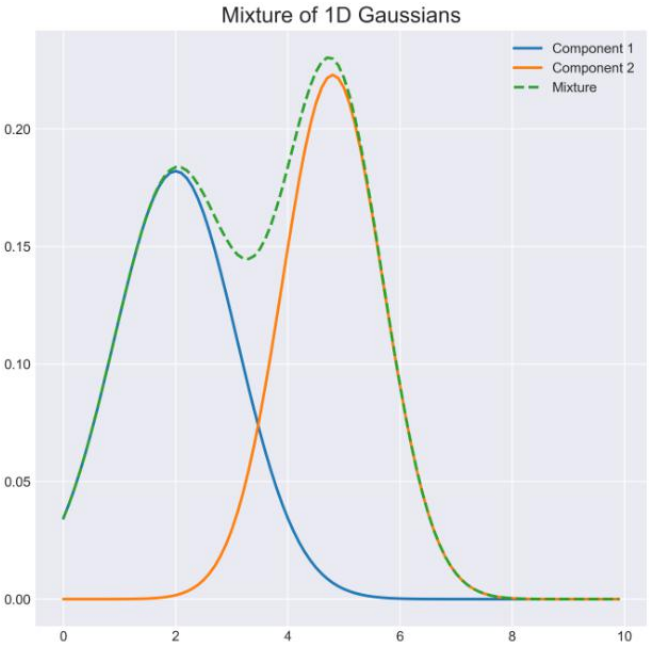
			男	女			
Π	μ	σ	γ1	γ2	X	X1	X2
Π1	μ1	σ1	0.3	0.7	100	30	70
Π2	μ2	σ2	0.65	0.35	90	58.5	31.5
			0.27	0.73	80	21.6	58.4
					70		

110

30

这个例子中如果你有 10000 个样本，那么就会有 20000 个γ值

还有 GMM 为什么叫这个名，首先就是假设高斯分布，然后就是混合，其实这里的混合会认为每条样本都是由这两个分布贡献而生成的，最后叫模型，因为有参数需要去求解嘛。



每个样本的值都是有多个分布混合融合出来的！

总结流程：

0，随机初始化Π，μ，σ

1，计算求得γ，这样每条脚本可以根据γ进行划分类别

（注意，这里是用γ，不是用Π，除非未来有一个人数据拿不到生猜就用Π1，Π2）

2，计算多个分布的多个分量，比如例子中的 X1，X2，然后更新Π，μ，σ

（μ如何来更新？用μ1 等于 X1 加和除以γ1 的求和）

$$\mu_k = \frac{\sum_{i}^{N} \gamma_{ik} \cdot x_i}{N_k}$$

3，判断收敛停止迭代，直到Π，μ，σ不变了

第二步：估计每个分布的参数

对于所有的样本点 1 到 N，对于分布 k 而言，可以看作生成了

$\{\gamma(i,k)x_i \mid i=1,2,\cdots N\}$ 这些值，由于每个分布都是高斯分布，利用上面的结论可以得到如下，当然后面会推导出来为什么是如下结论

$$\begin{cases} N_k = \sum_{i=1}^{N} \gamma(i,k) \\ \mu_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma(i,k) x_i \\ \Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma(i,k)(x_i - \mu_k)(x_i - \mu_k)^T \\ \pi_k = \frac{N_k}{N} = \frac{1}{N} \sum_{i=1}^{N} \gamma(i,k) \end{cases}$$

作业

1，总结 GMM 迭代计算流程
2，使用 Excel 将例子自行复现一遍

EM 算法

隐变量

通过极大似然估计建立目标函数，EM 不单单用在 GMM，其实它是更泛化的一种求解思想！
Z（Z1...Zk）是隐变量，很难直接找到参数的估计
我们需要先建立似然函数的下界，

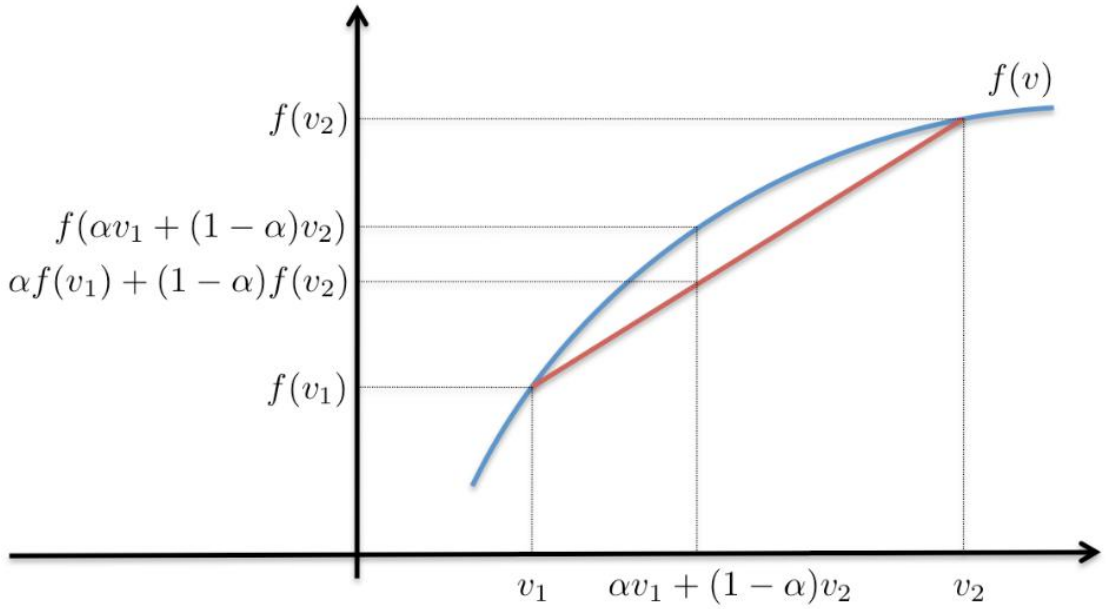
$$l(\theta) = \sum_{i=1}^{m} \log p(x;\theta)$$

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

这里如果是前面的 GMM，θ就是Π，μ，σ，除了高斯分布，其它的分布也是有一些参数的，而且其它的分布 p 概率密度函数是不一样的，EM 不管那么多
这里为了引入 z 隐变量，全概率公式用上，如果对应上，隐变量的分布函数就是之前说的γ k=P(xi∈Zk)，比如某条样本属于男的 0.7，属于女的 0.3，如果有三个类别，那比如属于类别 A 是 1/3，属于类别 B 是 1/2，属于类别 C 是 1/6，总之加起来是 1

Jenson 不等式

举个例子：



- $y = f(x) = \exp(-(x-2)^2)$
- $v_1 = 1; v_2 = 2.5; \alpha = .3$

$$f(v_1) \approx .3679$$
$$f(v_2) \approx .7788$$
$$\alpha f(v_1) + (1-\alpha) f(v_2) \approx \mathbf{.6555}$$

$$\alpha v_1 + (1-\alpha)v_2 = 2.05$$
$$f(\alpha v_1 + (1-\alpha)v_2) \approx \mathbf{.9975}$$

we see that $\alpha f(v_1) + (1-\alpha) f(v_2) \leq f(\alpha v_1 + (1-\alpha)v_2)$.

Jensen 不等式表述如下：

如果 f 是凸函数，X 是随机变量，那么：

$$f(\mathbf{E}\, x) \leq \mathbf{E}\, f(x)$$

特别地，如果 f 是严格凸函数，当且仅当 X 是常量时，上式取等号。

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$$

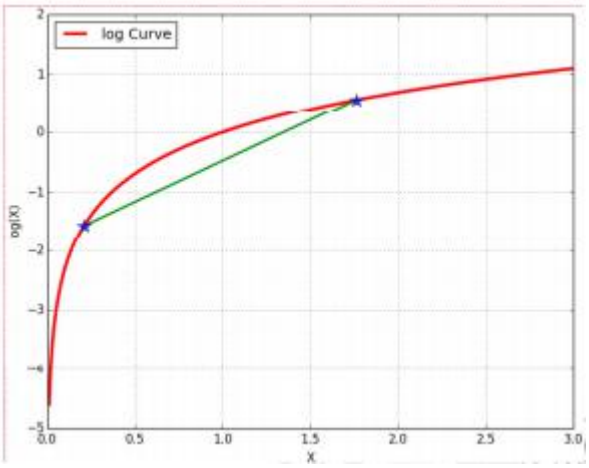
若 $\theta_1, \ldots, \theta_k \geq 0, \theta_1 + \cdots + \theta_k = 1$

则 $f(\theta_1 x_1 + \cdots + \theta_k x_k) \leq \theta_1 f(x_1) + \cdots + \theta_k f(x_k)$

如果θk 是隐变量的分布函数，是概率γk 呢？

求下界

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



这里有 log，就意味着它是单调递增的函数，根据上面 Jensen 不等式，应该是大于等于的符号，这里的 log 就是 Jensen 不等式中的 f()函数

令 Qi(zi)是 z 的分布函数，在上面公式中，直接就是上下乘以除以了一个 Qi(zi)，Q(zi)就是 Jensen 公式中的α或θ，然后除以的那部分就是 v 或 x

什么时候 LHS 和 RHS 可以取等号？

X1=X2=...=C

E-step 寻找紧的下界

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

$$Q_i(z^{(i)}) \propto p(x^{(i)}, z^{(i)}; \theta)$$

$$\sum_z Q_i(z^{(i)}) = 1$$

同时根据约束条件，0<=Qi<=1

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

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

$$= p(z^{(i)} \mid x^{(i)}; \theta)$$

上面的 Qi(Zi)实际上就是我们的γ，以上就是 EM 算法中的 E-step

M-step 估计一组 theta

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

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

这时上一步已经求出来了γ，就是 Q，这时我们就是最大化总似然，找到最优解θ可以使得上式总似然最大

到这里 EM 就讲完了

EM 框架运用到 GMM 问题上

$$w_j^{(i)} = Q_i(z^{(i)} = j) = P(z^{(i)} = j|x^{(i)}; \phi, \mu, \Sigma)$$

这里的 W 就是γ，ϕ就是Π，Σ就是σ

带入到多维高斯分布的的似然函数中

$$\sum_{i=1}^{m}\sum_{z^{(i)}} Q_i(z^{(i)})\log\frac{p(x^{(i)},z^{(i)};\phi,\mu,\Sigma)}{Q_i(z^{(i)})}$$
$$=\sum_{i=1}^{m}\sum_{j=1}^{k} Q_i(z^{(i)}=j)\log\frac{p(x^{(i)}|z^{(i)}=j;\mu,\Sigma)p(z^{(i)}=j;\phi)}{Q_i(z^{(i)}=j)}$$
$$=\sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)}\log\frac{\frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}}\exp\left(-\frac{1}{2}(x^{(i)}-\mu_j)^T\Sigma_j^{-1}(x^{(i)}-\mu_j)\right)\cdot\phi_j}{w_j^{(i)}}$$

ϕj 就是Πj，高斯分布的概率密度函数就是关于某个高斯分布下的μ和σ有关

这里的 n 是多少没关系，后面会把这项整体约掉，然后Σj 是σj 的平方

对均值求偏导

$$\nabla_{\mu_l}\sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)}\log\frac{\frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}}\exp\left(-\frac{1}{2}(x^{(i)}-\mu_j)^T\Sigma_j^{-1}(x^{(i)}-\mu_j)\right)\cdot\phi_j}{w_j^{(i)}}$$
$$=-\nabla_{\mu_l}\sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)}\frac{1}{2}(x^{(i)}-\mu_j)^T\Sigma_j^{-1}(x^{(i)}-\mu_j)$$
$$=\frac{1}{2}\sum_{i=1}^{m} w_l^{(i)}\nabla_{\mu_l}2\mu_l^T\Sigma_l^{-1}x^{(i)}-\mu_l^T\Sigma_l^{-1}\mu_l$$
$$=\sum_{i=1}^{m} w_l^{(i)}\left(\Sigma_l^{-1}x^{(i)}-\Sigma_l^{-1}\mu_l\right)$$
$$\mu_l:=\frac{\sum_{i=1}^{m} w_l^{(i)}x^{(i)}}{\sum_{i=1}^{m} w_l^{(i)}}$$

这里的 Wl 是γ，m 是之前 excel 算的时候的大 N

方差的解

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

会用到的导函数公式

$y = x^n$	$y' = nx^{n-1}$

等等！还差个参数呢

$$\sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1}(x^{(i)} - \mu_j)\right) \cdot \phi_j}{w_j^{(i)}}$$

就是ϕj 就是Πj 还没有求解呢，那么一样对ϕj 求导，可以删除和它无关的项

$$\sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} \log \phi_j$$

但是这个ϕj 有约束，ϕj 的加和是 1，前面讲过是先验概率

拉格朗日函数！

$$\mathcal{L}(\phi) = \sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} \log \phi_j + \beta\left(\sum_{j=1}^{k} \phi_j - 1\right)$$

ci(x)<=0　hj(x)=0

分别求偏导，得到了ϕj 的取值

$$\frac{\partial}{\partial \phi_j}\mathcal{L}(\phi) = \sum_{i=1}^{m} \frac{w_j^{(i)}}{\phi_j} + \beta$$

$$-\beta = \sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} = \sum_{i=1}^{m} 1 = m$$

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

上面的技巧是对ϕj 求导的式子等于 0，然后把β放到右边，等式两边乘以一个ϕj，然后对 k 求和可得结果

联想 EM 框架运用在 KMeans 算法上

EM 对应 KMeans 第一步就是归堆，根据中心点进行划分

hard assignment

one-hot 编码，这样无法得到某个样本属于某个类别的后验概率，但是是有γ的，就是 1 或 0

然后第二步求μ，可不就是均值嘛，方差可不就是会和单个高斯分布的方差一样了！