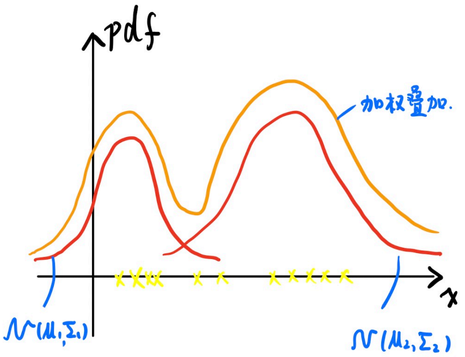
机器学习白板推导系列笔记 (11.1∼11.4)

Dexing Huang
dxhuang@bupt.edu.cn
Beijing University of Posts and Telecommunications

日期：2021 年 11 月 17 日

1 高斯混合模型背景

高斯混合模型 (Gaussian Mixture Model, GMM) 相较于高斯分布模型能够更好的描述复杂数据的分布, 如下所示的一维数据分布



使用一个高斯分布显然不能很好的反映数据的分布，即然一个不行，很自然的就想到采用多个高斯分布加权的方式得到一个混合的高斯分布来描述数据的分布．下面从两种方式来理解 GMM

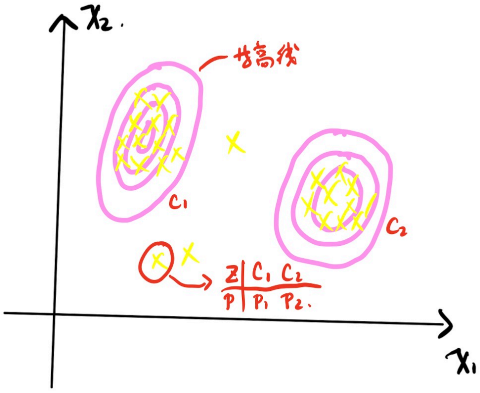
1.1 几何角度

几何角度十分简单明了, 正如上面所说的, 一个高斯分布无法描述数据分布那就采用多个加权的方式, 得到的概率密度函数为

$$p(x)=\sum_{i=1}^{K}\alpha_i\mathcal{N}(\mu_i,\Sigma_i),\quad \sum_{i=1}^{K}\alpha_i=1 \tag{1}$$

1.2 混合模型角度 (生成模型)

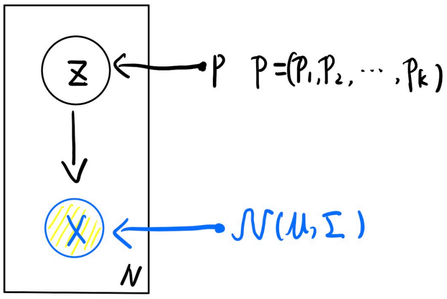
下面从模型的角度来探究加权的另一种理解方式, 使用二维的例子来说明, 首先做一下说明. 我们称观测到的数据点 X 为 Observable Variable, 并且引入一个隐变量 Z 称为 Latent Variable.



上图是一个二维变量 X 的分布, 使用两个高斯分布来描述数据的分布, 两个高斯分布分别记为 c_1 和 c_2. 隐变量Z 描述的是数据点 X 属于哪类高斯分布的概率, 由此可知, Z 是一个随机变量, 且是离散随机变量, 概率分布如下

Z	c_1	c_2	$\cdots$	c_K
p	p_1	p_2	$\cdots$	p_K

如上面红色圈起来的点, 它既属于 c_1 也属于 c_2, 但直观上来说属于 c_1 的概率比较大, 因此 $p(Z=c_1) \geq p(Z=c_2)$, 有点类似于分类问题. GMM 是一个生成模型, 下面来说一下生成的数据的步骤 (在这里假设模型参数已知), 若数据是简单的高斯分布, 那么生成数据很简单, 对分布进行简单的采样即可, 需要几个点采样几个. 对于 GMM 来说, 可以假设有一个 K 面的骰子, 并且概率不均等, 随机转骰子, 当转到 k 面时, 就在第 k 个高斯分布上进行采样, 生成样本点. 概率图模型可以描述成以下形式



综合上面分析, GMM 模型的概率密度函数可以得出

$$
\begin{aligned}
p(x) &= \sum_z p(x,z) \\
&= \sum_{i=1}^K p(x, z=c_i) \\
&= \sum_{i=1}^K p(z=c_i)p(x|z=c_i) \\
&= \sum_{i=1}^K p_i \mathcal{N}(x|\mu_i,\Sigma_i)
\end{aligned}
$$

2 极大似然 (模型参数估计)

下面对 GMM 的参数进行估计, 模型的参数包括

$$\theta=\{p_1,p_2,\cdots,p_K,\mu_1,\mu_2,\cdots,\mu_K,\Sigma_1,\Sigma_2,\cdots,\Sigma_K\} \tag{2}$$

按照极大似然估计的做法是

$$\begin{aligned}\hat{\theta}&=\arg\max_{\theta}\log P(X|\theta)\\&=\arg\max_{\theta}\sum_{i=1}^{N}\log p(x_i|\theta)\\&=\arg\max_{\theta}\sum_{i=1}^{N}\log\sum_{j=1}^{K}p_j\mathcal{N}(x_i|\mu_j,\Sigma_j)\end{aligned}$$

然后对似然函数求导并令导数为 0 得到参数的估计值, 但是这在 GMM 中几乎是不可能的, 即无法求出解析解, 如对 p_i 进行求导

$$\frac{\partial\mathcal{L}}{\partial p_j}=\sum_{i=1}^{N}\frac{\mathcal{N}(x_i|\mu_j,\Sigma_j)}{\sum_{k=1}^{K}p_k\mathcal{N}(x_i|\mu_k,\Sigma_k)} \tag{3}$$

可以看到若令 $\frac{\partial\mathcal{L}}{\partial p_j}=0$, 方程中还有其他未知数, 求解起来十分困难, 或者说根本无法解, 因此简单的使用极大似然估计 GMM 的模型参数是不可行的, 因此需要使用一些方法来得到近似解, 可以使用梯度的方法, 但在 GMM 中, 使用最广泛的是EM 算法.

3 EM 算法 (模型参数估计)

首先为了避免混淆, 先对符号做一下规定. 观测数据 $X=\{x_i\}_{i=1}^{N}$, 并且 x_i 是个 p 维向量. 隐变量 $Z=\{z_i\}_{i=1}^{N}$, 在本问题中 z_i 是某个类别, 可以视为标量. p_z 是指当给定 z 后就知道 p_z 的值, 比如 $z=c_2$ 时, 那么 $p_z=p_2$. 大写随机变量 $X=x_1,x_2,\cdots,x_N$ 均表示数据的联合分布, 小写表示每个数据点的分布.

回顾一下 EM 算法迭代的关键步骤

$$\begin{aligned}E-step:Q(\theta,\theta^{(t)})&=\int_{Z}P(Z|X,\theta^{(t)})\log P(X,Z|\theta)dZ\\M-step:\theta^{(t+1)}&=\arg\max_{\theta}Q(\theta,\theta^{(t)})\end{aligned}$$

下面使用 EM 算法来求解 GMM 参数

3.1 E-Step

$$
\begin{aligned}
Q(\theta,\theta^{(t)}) &= \int_Z P(Z|X,\theta^{(t)}) \log P(X,Z|\theta) dZ \\
&= \sum_Z P(Z|X,\theta^{(t)}) \log P(X,Z|\theta) \\
&= \sum_{z_1}\sum_{z_2}\cdots\sum_{z_N}\prod_{i=1}^{N} p(z_i|x_i,\theta^{(t)}) \log \prod_{i=1}^{N} p(x_i,z_i|\theta) \\
&= \sum_{z_1}\sum_{z_2}\cdots\sum_{z_N}\sum_{i=1}^{N} \log p(x_i,z_i|\theta) \prod_{i=1}^{N} p(z_i|x_i,\theta^{(t)}) \\
&= \sum_{z_1}\sum_{z_2}\cdots\sum_{z_N}(\log p(x_1,z_1|\theta) + \log p(x_2,z_2|\theta) + \cdots + \log p(x_N,z_N|\theta)) \prod_{i=1}^{N} p(z_i|x_i,\theta^{(t)})
\end{aligned}
\tag{4}
$$

下面取出一项进行分析, 为简单起见就取第一项

$$
\begin{aligned}
\sum_{z_1,z_2,\cdots,z_N} \log p(x_1,z_1|\theta) \prod_{i=1}^{N} p(z_i|x_i,\theta^{(t)}) &= \sum_{z_1}\sum_{z_2}\cdots\sum_{z_N} \log p(x_1,z_1|\theta) p(z_1|x_1,\theta^{(t)}) \prod_{i=2}^{N} p(z_i|x_i,\theta^{(t)}) \\
&= \sum_{z_1} p(z_1|x_1,\theta^{(t)}) \log p(x_1,z_1|\theta) \sum_{z_2}\cdots\sum_{z_N}\prod_{i=2}^{N} p(z_i|x_i,\theta^{(t)}) \\
&= \sum_{z_1} p(z_1|x_1,\theta^{(t)}) \log p(x_1,z_1|\theta) \underbrace{\sum_{z_2} p(z_2|x_2,\theta^{(t)}) \cdots \sum_{z_N} p(z_N|x_N,\theta^{(t)})}_{1\cdot 1\cdots 1} \\
&= \sum_{z_1} p(z_1|x_1,\theta^{(t)}) \log p(x_1,z_1|\theta)
\end{aligned}
\tag{5}
$$

将式 (5) 带入式 (4) 得

$$
Q(\theta,\theta^{(t)}) = \sum_{i=1}^{N}\sum_{z_i} p(z_i|x_i,\theta^{(t)}) \log p(x_i,z_i|\theta) \tag{6}
$$

又有

$$
\begin{aligned}
p(x,z) &= p(z)p(x|z) \\
&= p_z\mathcal{N}(x|\mu_z,\Sigma_z)
\end{aligned}
\tag{7}
$$

$$
\begin{aligned}
p(z|x) &= \frac{p(x,z)}{p(x)} \\
&= \frac{p_z\mathcal{N}(x|\mu_z,\Sigma_z)}{\sum_{i=1}^{K} p_i\mathcal{N}(x|\mu_i,\Sigma_i)}
\end{aligned}
\tag{8}
$$

将式 (7), (8) 带入式 (6) 得

$$
Q(\theta,\theta^{(t)}) = \sum_{i=1}^{N}\sum_{z_i} \frac{p_{z_i}^{(t)}\mathcal{N}(x_i|\mu_{z_i}^{(t)},\Sigma_{z_i}^{(t)})}{\sum_{k=1}^{K} p_k^{(t)}\mathcal{N}(x_i|\mu_k^{(t)},\Sigma_k^{(t)})} \log(p_{z_i}\mathcal{N}(x_i|\mu_{z_i},\Sigma_{z_i})) \tag{9}
$$

3.2 M-Step

在 M-step 中, 使用求得的 $Q(\theta,\theta^{(t)})$ 估计新的参数, 首先对式 (9) 做一点变换

$$\begin{aligned}
Q(\theta,\theta^{(t)}) &= \sum_{i=1}^{N}\sum_{z_i} p(z_i|x_i,\theta^{(t)})\log(p_{z_i}\mathcal{N}(x_i|\mu_{z_i},\Sigma_{z_i})) \\
&= \sum_{z_i}\sum_{i=1}^{N} p(z_i|x_i,\theta^{(t)})\log(p_{z_i}\mathcal{N}(x_i|\mu_{z_i},\Sigma_{z_i})) \\
&= \sum_{k=1}^{K}\sum_{i=1}^{N} p(z_i=c_k|x_i,\theta^{(t)})\log(p_k\mathcal{N}(x_i|\mu_k,\Sigma_k)) \\
&= \sum_{k=1}^{K}\sum_{i=1}^{N} p(z_i=c_k|x_i,\theta^{(t)})(\log p_k + \log\mathcal{N}(x_i|\mu_k,\Sigma_k))
\end{aligned}$$

$$\theta^{(t+1)} = \arg\max_{\theta}\sum_{k=1}^{K}\sum_{i=1}^{N} p(z_i=c_k|x_i,\theta^{(t)})(\log p_k + \log\mathcal{N}(x_i|\mu_k,\Sigma_k)) \tag{10}$$

对上式求导即可得到下一时刻迭代参数, 下面以 p 为例来说明一下求解过程

$$p_k^{(t+1)} = \arg\max_{p_k}\sum_{k=1}^{K}\sum_{i=1}^{N} p(z_i=c_k|x_i,\theta^{(t)})\log p_k \tag{11}$$

因为 p_k 的归一性, 得到下面的约束优化问题

$$\begin{cases}
\arg\min\limits_{p_k} -\sum_{k=1}^{K}\sum_{i=1}^{N} p(z_i=c_k|x_i,\theta^{(t)})\log p_k \\
s.t. \sum_{k=1}^{K} p_k = 1
\end{cases} \tag{12}$$

构造 Lagrange 函数

$$\mathcal{L}(p,\lambda) = -\sum_{k=1}^{K}\sum_{i=1}^{N} p(z_i=c_k|x_i,\theta^{(t)})\log p_k + \lambda(1-\sum_{k=1}^{K} p_k) \tag{13}$$

$$\begin{aligned}
&\frac{\partial}{\partial p_k}\mathcal{L}(p,\lambda) = -\sum_{i=1}^{N} p(z_i=c_k|x_i,\theta^{(t)})\frac{1}{p_k} - \lambda = 0 \\
&\sum_{i=1}^{N} p(z_i=c_k|x_i,\theta^{(t)})\frac{1}{p_k} + \lambda = 0 \\
&\sum_{i=1}^{N} p(z_i=c_k|x_i,\theta^{(t)}) + \lambda p_k = 0 \\
&\sum_{k=1}^{K}\sum_{i=1}^{N} p(z_i=c_k|x_i,\theta^{(t)}) + \sum_{k=1}^{K}\lambda p_k = 0 \\
&N + \lambda = 0 \\
&\lambda = -N
\end{aligned}$$

因此

$$p_k^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} p(z_i=c_k|x_i,\theta^{(t)}) \tag{14}$$

那么

$$p^{(t+1)} = (p_1^{(t+1)}, p_2^{(t+1)}, \cdots, p_K^{(t+1)}) \tag{15}$$

①EM Algorithm

Observed variable $X=\{x_i\}_{i=1}^{N}$ $x_i\in\mathbb{R}^p$

latent variable $Z=\{z_j\}_j$

complete data (X,Z) param: θ

$$\hat{\theta}=\arg\max_{\theta}\log P(X|\theta)$$

$$\text{model}\xrightarrow{MLE}\hat{\theta}\ (x)\quad \text{model}\xrightarrow{EM}\approx\hat{\theta}(V)$$

$$\log P(X|\theta)=\log P(X,Z|\theta)-\log P(Z|X,\theta)$$

$$=\log\frac{P(X,Z|\theta)}{q(Z)}-\log\frac{P(Z|X,\theta)}{q(Z)}$$

$$\text{left}=\int_Z q(Z)\log P(X|\theta)dZ=\log P(X|\theta)$$

$$\text{right}=\underbrace{\int_Z q(Z)\log\frac{P(X,Z|\theta)}{q(Z)}dZ}_{ELBO}\ \underbrace{-\int_Z q(Z)\log\frac{P(Z|X,\theta)}{q(Z)}dZ}_{KL\{q(Z)\|P(Z|X,\theta)\}\geq 0}$$

$$\Rightarrow\log P(X|\theta)=ELBO+KL\{q(Z)\|P(Z|X,\theta)\}\geq ELBO$$

$$\theta^{(t+1)}=\arg\max_{\theta}ELBO=\arg\max_{\theta}\int_Z q(Z)\log\frac{P(X,Z|\theta)}{q(Z)}dZ$$

$$=\arg\max_{\theta}\int_Z P(Z|X,\theta^{(t)})\log\frac{P(X,Z|\theta)}{P(Z|X,\theta^{(t)})}dZ$$

$$=\arg\max_{\theta}\int_Z P(Z|X,\theta^{(t)})\log P(X,Z|\theta)dZ$$

$$Q(\theta,\theta^{(t)})=\int_Z P(Z|X,\theta^{(t)})\log P(X,Z|\theta)dZ$$

$$=E_{Z\sim P(Z|X,\theta^{(t)})}[\log P(X,Z|\theta)]$$

E-step: 求出 $E_{Z\sim P(Z,X,\theta^{(t)})}[\log P(X,Z|\theta)]$

M-step: $\arg\max_{\theta}Q(\theta,\theta^{(t)})$

②广义EM

$$ELBO=\int_Z q(Z)\log\frac{P(X,Z|\theta)}{q(Z)}dZ=\mathcal{L}(q,\theta)$$

E-step: 固定θ $q^{(t+1)}=\arg\max_{q}\mathcal{L}(q,\theta^{(t)})$

M-step: $\theta^{(t+1)}=\arg\max_{\theta}\mathcal{L}(q^{(t+1)},\theta)$

③GMM

latent variable $Z=\{z_j\}_{j=1}^{N}$ $z_j\in\mathbb{R}$ $x_i\in\mathbb{R}^p$

z_j	C_1	C_2	$\cdots$	C_k
P	P_1	P_2	$\cdots$	P_k.

$$P(x)=\int_z P(x,z)dz=\sum_z P(x,z)=\sum_z P(z)P(x|z)$$

$$=\sum_{k=1}^{K}P_k\mathcal{N}(x|\mu_k,\Sigma_k)$$

$$\theta=\{P_1,P_2,\cdots,P_k,\mu_1,\cdots,\mu_k,\Sigma_1,\cdots,\Sigma_k\}$$

$$Q(\theta,\theta^{(t)})=\int_Z P(Z|X,\theta^{(t)})\log P(X,Z|\theta)dZ$$

$$=\sum_Z\prod_{i=1}^{N}P(z_i|x_i,\theta^{(t)})\sum_{i=1}^{N}\log P(x_i,z_i|\theta)\quad -①$$

$$=\sum_Z(\log P(x_1,z_1|\theta)+\log P(x_2,z_2|\theta)+\cdots)\prod_{i=1}^{N}P(z_i|x_i,\theta^{(t)})$$

$$\sum_Z\log P(x_1,z_1|\theta)\prod_{i=1}^{N}P(z_i|x_i,\theta^{(t)})=$$

$$\sum_{z_1}\sum_{z_2}\cdots\sum_{z_N}\log P(x_1,z_1|\theta)P(z_1|x_1,\theta^{(t)})\prod_{i=1}^{N}P(z_i|x_i,\theta^{(t)})$$

$$=\sum_{z_1}\log P(x_1,z_1|\theta)P(z_1|x_1,\theta^{(t)})\underbrace{\sum_{z_2}\cdots\sum_{z_N}\prod_{i=1}^{N}P(z_i|x_i,\theta^{(t)})}_{1\times1\times\cdots\times1}$$

$$=\sum_{z_1}\log P(x_1,z_1|\theta)P(z_1|x_1,\theta^{(t)})\quad -②$$

②代入①

$$Q(\theta,\theta^{(t)})=\sum_{i=1}^{N}\sum_{z_i}\log[P(x_i,z_i|\theta)]P(z_i|x_i,\theta^{(t)})$$

$$P(x,z)=P(z)P(x|z)=P_z\mathcal{N}(x|\mu_z,\Sigma_z)$$

$$P(z|x)=\frac{P(x,z)}{P(x)}=\frac{P_z\mathcal{N}(x|\mu_z,\Sigma_z)}{\sum_{k=1}^{K}P_k\mathcal{N}(x|\mu_k,\Sigma_k)}$$

$$Q(\theta,\theta^{(t)})=\sum_{i=1}^{N}\sum_{z_i}\left(\frac{P_{z_i}^{(t)}\mathcal{N}(x_i|\mu_{z_i}^{(t)},\Sigma_{z_i}^{(t)})}{\sum_{k=1}^{K}P_k^{(t)}\mathcal{N}(x_i|\mu_k^{(t)},\Sigma_k^{(t)})}\cdot\log P_{z_i}\mathcal{N}(x_i|\mu_{z_i},\Sigma_{z_i})\right)$$

E-step

M-step

$$Q(\theta,\theta^{(t)})=\sum_{i=1}^{N}\sum_{z_i}P(z_i|x_i,\theta^{(t)})\cdot\log P_{z_i}\mathcal{N}(x_i|\mu_{z_i},\Sigma_{z_i})$$

$$=\sum_{z_i}\sum_{i=1}^{N}P(z_i|x_i,\theta^{(t)})\cdot\log P_{z_i}\mathcal{N}(x_i|\mu_{z_i},\Sigma_{z_i})$$

$$=\sum_{k=1}^{K}\sum_{i=1}^{N}P(z_i=C_k|x_i,\theta^{(t)})\cdot\log P_k\cdot\mathcal{N}(x_i|\mu_k,\Sigma_k)$$

$$=\sum_{k=1}^{K}\sum_{i=1}^{N}(\log P_k+\log\mathcal{N}(x_i|\mu_k,\Sigma_k))\cdot P(z_i=C_k|x_i,\theta^{(t)})$$

$$\theta^{(t+1)}=\arg\max_{\theta}Q(\theta,\theta^{(t)})$$

$$\begin{cases} \arg\max\limits_{\theta} \sum\limits_{k=1}^{K}\sum\limits_{i=1}^{N} \left(\log P_k + \log \mathcal{N}(x_i \mid \mu_k, \Sigma_k)\right) \cdot P(z_i = c_k \mid x_i, \theta^{(t)}) \\ \sum\limits_{k=1}^{K} P_k = 1 \end{cases}$$

估计 $P_k^{(t+1)}$

$$\Rightarrow \begin{cases} \arg\min\limits_{P} - \sum\limits_{k=1}^{K}\sum\limits_{i=1}^{N} \log P_k \cdot P(z_i = c_k \mid x_i, \theta^{(t)}) \\ s.t. \sum\limits_{i=1}^{K} P_k = 1 \end{cases}$$

$$\mathcal{L}(P, \lambda) = - \sum_{k=1}^{N}\sum_{i=1}^{N} \log P_k \cdot P(z_i = c_k \mid x_i, \theta^{(t)}) + \lambda\left(1 - \sum_{i=1}^{K} P_k\right)$$

$$\frac{\partial \mathcal{L}}{\partial P_k} = - \sum_{i=1}^{N} \frac{1}{P_k} P(z_i = c_k \mid x_i, \theta^{(t)}) - \lambda = 0$$

$$\sum_{i=1}^{N} P(z_i = c_k \mid x_i, \theta^{(t)}) + \lambda P_k = 0$$

$$\sum_{k=1}^{K}\sum_{i=1}^{N} P(z_i = c_k \mid x_i, \theta^{(t)}) + \sum_{k=1}^{K} \lambda P_k = 0$$

$$N + \lambda = 0 \Rightarrow \lambda = -N$$

$$\Rightarrow P_k^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} P(z_i = c_k \mid x_i, \theta^{(t)}) \quad \square$$