一个例子

高斯混合模型（Gaussian Mixed Model）指的是多个高斯分布函数的线性组合，理论上GMM可以拟合出任意类型的分布，通常用于解决同一集合下的数据包含多个不同的分布的情况（或者是同一类分布但参数不一样，或者是不同类型的分布，比如正态分布和伯努利分布）。

如图1，图中的点在我们看来明显分成两个聚类。这两个聚类中的点分别通过两个不同的正态分布随机生成而来。但是如果没有GMM，那么只能用一个的二维高斯分布来描述图1中的数据。图1中的椭圆即为二倍标准差的正态分布椭圆。这显然不太合理，毕竟肉眼一看就觉得应该把它们分成两类。

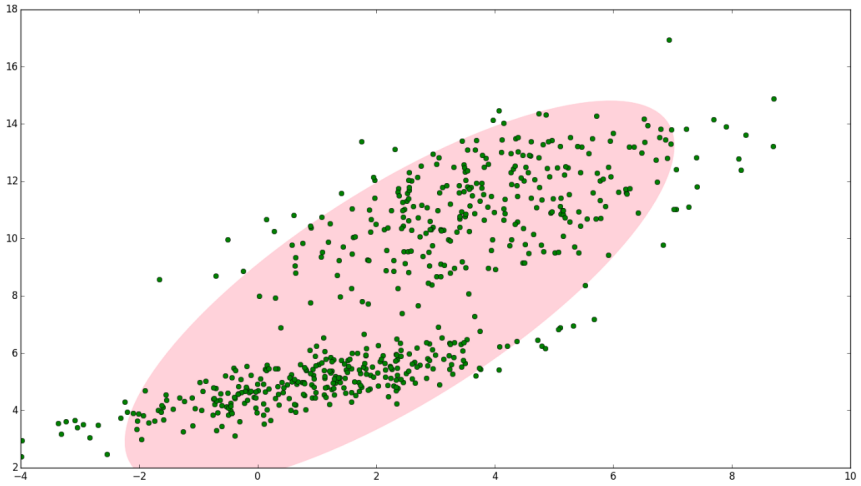

图1

这时候就可以使用GMM了！如图2，数据在平面上的空间分布和图1一样，这时使用两个二维高斯分布来描述图2中的数据，分别记为$\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$和$\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. 图中的两个椭圆分别是这两个高斯分布的二倍标准差椭圆。可以看到使用两个二维高斯分布来描述图中的数据显然更合理。实际上图中的两个聚类的中的点是通过两个不同的正态分布随机生成而来。如果将两个二维高斯分布$\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$和$\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$合成一个二维的分布，那么就可以用合成后的分布来描述图2中的所有点。最直观的方法就是对这两个二维高斯分布做线性组合，用线性组合后的分布来描述整个集合中的数据。这就是高斯混合模型（GMM）。

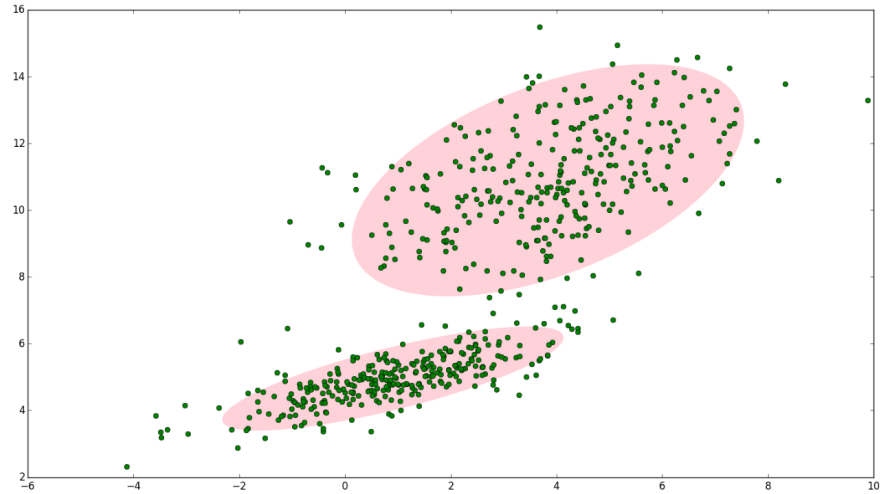

图2

高斯混合模型（GMM）

设有随机变量$\boldsymbol{X}$，则混合高斯模型可以用下式表示：

$$\mathrm{p}(\boldsymbol{x})=\sum_{\mathrm{k}=1}^{\mathrm{K}} \pi_{\mathrm{k}} \mathcal{N}\left(\boldsymbol{x} | \boldsymbol{\mu}_{\mathrm{k}}, \boldsymbol{\Sigma}_{\mathrm{k}}\right)$$

其中$\mathcal{N}\left(\boldsymbol{x} | \boldsymbol{\mu}_{\mathrm{k}}, \boldsymbol{\Sigma}_{\mathrm{k}}\right)$称为混合模型中的第$\mathrm{k}$个**分量（component）**。如前面图2中的例子，有两个聚类，可以用两个二维高斯分布来表示，那么分量数$\mathrm{K}=2$. π_{k}是**混合系数（mixture coefficient）**，且满足：

$$\sum_{\mathrm{k}=1}^{\mathrm{K}} \pi_{\mathrm{k}}=1$$

$$0 \leq \pi_{\mathrm{k}} \leq 1$$

实际上，可以认为π_{k}就是每个分量$\mathcal{N}\left(\boldsymbol{x} | \boldsymbol{\mu}_{\mathrm{k}}, \boldsymbol{\Sigma}_{\mathrm{k}}\right)$的权重。

GMM的应用

GMM常用于聚类。如果要从 GMM 的分布中随机地取一个点的话，实际上可以分为两步：首先随机地在这 K 个 Component 之中选一个，**每个 Component 被选中的概率实际上就是它的系数**π_{k}，选中 Component 之后，再单独地考虑从这个 Component 的分布中选取一个点就可以了——这里已经回到了普通的 Gaussian 分布，转化为已知的问题。

将GMM用于聚类时，**假设数据服从混合高斯分布（Mixture Gaussian Distribution）**，那么只要根据数据推出 GMM 的概率分布来就可以了；然后 GMM 的 K 个 Component 实际上对应K个 cluster 。根据数据来推算概率密度通常被称作 density estimation 。特别地，当我已知（或假定）概率密度函数的形式，而要估计其中的参数的过程被称作『参数估计』。

例如图2的例子，很明显有两个聚类，可以定义$\mathrm{K}=2$. 那么对应的GMM形式如下：

$$\mathrm{p}(\boldsymbol{x})=\pi_1 \mathcal{N}\left(\boldsymbol{x} | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right)+\pi_2 \mathcal{N}\left(\boldsymbol{x} | \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\right)$$

上式中未知的参数有六个：$\left(\pi_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1 ; \pi_2, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\right)$. 之前提到GMM聚类时分为两步，第一步是**随机地在这K个分量中选一个，每个分量被选中的概率即为混合系数**π_{k}. 可以设定$\pi_1=\pi_2=0.5$，表示每个分量被选中的概率是0.5，即从中抽出一个点，这个点属于第一类的概率和第二类的概率各占一半。但实际应用中事先指定π_{k}的值是很笨的做法，当问题一般化后，会出现一个问题：当从图2中的集合随机选取一个点，怎么知道这个点是来自$\mathrm{N}\left(\boldsymbol{x} | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right)$还是$\mathrm{N}\left(\boldsymbol{x} | \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\right)$呢？换言之怎么根据数据**自动确定**$\pi_1$和$\pi_2$的值？这就是GMM参数估计的问题。要解决这个问题，可以使用EM算法。通过EM算法，我们可以迭代计算出GMM中的参数：$\left(\pi_{\mathrm{k}}, \boldsymbol{x}_{\mathrm{k}}, \boldsymbol{\Sigma}_{\mathrm{k}}\right)$.

GMM参数估计过程

GMM的贝叶斯理解

在介绍GMM参数估计之前，先改写GMM的形式，改写之后的GMM模型可以方便地使用EM估计参数。GMM的原始形式如下：

$$\mathrm{p}(\boldsymbol{x})=\sum_{\mathrm{k}=1}^{\mathrm{K}} \pi_{\mathrm{k}} \mathcal{N}\left(\boldsymbol{x} | \boldsymbol{\mu}_{\mathrm{k}}, \boldsymbol{\Sigma}_{\mathrm{k}}\right) \tag{1}$$

前面提到π_{k}可以看成是第k类被选中的概率。我们引入一个新的K维随机变量$\boldsymbol{z}$. $\mathrm{z}_{\mathrm{k}}(1 \leq \mathrm{k} \leq \mathrm{K})$只能取0或1两个值；$\mathrm{z}_{\mathrm{k}}=1$表示第$\mathrm{k}$类被选中的概率，即：$\mathrm{p}\left(\mathrm{z}_{\mathrm{k}}=1\right)=\pi_{\mathrm{k}}$；如果$\mathrm{z}_{\mathrm{k}}=0$表示第$\mathrm{k}$类没有被选中的概率。更数学化一点，$\mathrm{z}_{\mathrm{k}}$要满足以下两个条件：

$$\mathrm{z}_{\mathrm{k}} \in\{0,1\}$$

$$\sum_{\mathrm{K}} \mathrm{z}_{\mathrm{k}}=1$$

例如图2中的例子，有两类，则$\boldsymbol{z}$的维数是2. 如果从第一类中取出一个点，则$\boldsymbol{z}=(1,0)$；，如果从第二类中取出一个点，则$\boldsymbol{z}=(0,1)$.

$\mathrm{z}_{\mathrm{k}}=1$的概率就是$\pi_{\mathrm{k}}$，假设$\mathrm{z}_{\mathrm{k}}$之间是独立同分布的（iid），我们可以写出$\boldsymbol{z}$的联合概率分布形式，就是连乘：

$$\mathrm{p}(\boldsymbol{z})=\mathrm{p}\left(\mathrm{z}_1\right) \mathrm{p}\left(\mathrm{z}_2\right) \ldots \mathrm{p}\left(\mathrm{z}_{\mathrm{K}}\right)=\prod_{\mathrm{k}=1}^{\mathrm{K}} \pi_{\mathrm{k}}^{\mathrm{z}_{\mathrm{k}}} \tag{2}$$

因为z_{k}只能取0或1，且$\boldsymbol{z}$中只能有一个z_{k}为1而其它$\mathrm{z}_{\mathrm{j}}(\mathrm{j} \neq \mathrm{k})$全为0，所以上式是成立的。

图2中的数据可以分为两类，显然，每一類中的数据都是服从正态分布的。这个叙述可以用条件概率来表示：

$$\mathrm{p}\left(\boldsymbol{x} | \mathrm{z}_{\mathrm{k}}=1\right)=\mathcal{N}\left(\boldsymbol{x} | \boldsymbol{\mu}_{\mathrm{k}}, \boldsymbol{\Sigma}_{\mathrm{k}}\right)$$

即第k类中的数据服从正态分布。进而上式有可以写成如下形式：

$$\mathrm{p}(\boldsymbol{x} | \boldsymbol{z})=\prod_{\mathrm{k}=1}^{\mathrm{K}} \mathcal{N}\left(\boldsymbol{x} | \boldsymbol{\mu}_{\mathrm{k}}, \boldsymbol{\Sigma}_{\mathrm{k}}\right)^{\mathrm{z}_{\mathrm{k}}} \tag{3}$$

上面分别给出了$\mathrm{p}(\boldsymbol{z})$和$\mathrm{p}(\boldsymbol{x}|\boldsymbol{z})$的形式，根据条件概率公式，可以求出$\mathrm{p}(\boldsymbol{x})$的形式：

$$\begin{aligned}\mathrm{p}(\boldsymbol{x}) &= \sum_{\boldsymbol{z}} \mathrm{p}(\boldsymbol{z})\mathrm{p}(\boldsymbol{x}|\boldsymbol{z}) \\ &= \sum_{\boldsymbol{z}} \left(\prod_{\mathrm{k}=1}^{\mathrm{K}} \pi_{\mathrm{k}}^{\mathrm{z_k}} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{\mathrm{k}}, \boldsymbol{\Sigma}_{\mathrm{k}})^{\mathrm{z_k}} \right) \\ &= \sum_{\mathrm{k}=1}^{\mathrm{K}} \pi_{\mathrm{k}} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{\mathrm{k}}, \boldsymbol{\Sigma}_{\mathrm{k}})\end{aligned} \tag{4}$$

（注：上式第二个等号，对$\boldsymbol{z}$求和，实际上就是$\sum_{\mathrm{k}=1}^{\mathrm{K}}$。又因为对某个$\mathrm{k}$，只要$\mathrm{i} \neq \mathrm{k}$，则有$\mathrm{z_i} = 0$，所以$\mathrm{z_k} = 0$的项为1，可省略，最终得到第三个等号）

可以看到GMM模型的(1)式与(4)式有一样的形式，且(4)式中引入了一个新的变量$\boldsymbol{z}$，通常称为**隐含变量（latent variable）**。对于图2中的数据，『隐含』的意义是：我们知道数据可以分成两类，但是随机抽取一个数据点，我们不知道这个数据点属于第一类还是第二类，它的归属我们观察不到，因此引入一个隐含变量$\boldsymbol{z}$来描述这个现象。

注意到在贝叶斯的思想下，$\mathrm{p}(\boldsymbol{z})$是先验概率，$\mathrm{p}(\boldsymbol{x}|\boldsymbol{z})$是似然概率，很自然我们会想到求出后验概率$\mathrm{p}(\boldsymbol{z}|\boldsymbol{x})$：

$$\begin{aligned}\gamma(\mathrm{z_k}) &= \mathrm{p}(\mathrm{z_k} = 1|\boldsymbol{x}) \\ &= \frac{\mathrm{p}(\mathrm{z_k} = 1)\mathrm{p}(\boldsymbol{x}|\mathrm{z_k} = 1)}{\mathrm{p}(\boldsymbol{x}, \mathrm{z_k} = 1)} \\ &= \frac{\mathrm{p}(\mathrm{z_k} = 1)\mathrm{p}(\boldsymbol{x}|\mathrm{z_k} = 1)}{\sum_{\mathrm{j}=1}^{\mathrm{K}} \mathrm{p}(\mathrm{z_j} = 1)\mathrm{p}(\boldsymbol{x}|\mathrm{z_j} = 1)} \text{（全概率公式）} \\ &= \frac{\pi_{\mathrm{k}} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})}{\sum_{\mathrm{j}=1}^{\mathrm{K}} \pi_{\mathrm{j}} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu_j}, \boldsymbol{\Sigma_j})} \text{（结合(3)、(4)）}\end{aligned} \tag{5}$$

（第2行，贝叶斯定理。关于这一行的分母，很多人有疑问，应该是$\mathrm{p}(\boldsymbol{x}, \mathrm{z_k} = 1)$还是$\mathrm{p}(\boldsymbol{x})$，按照正常写法，应该是$\mathrm{p}(\boldsymbol{x})$。但是为了强调$\mathrm{z_k}$的取值，有的书会写成$\mathrm{p}(\boldsymbol{x}, \mathrm{z_k} = 1)$，比如李航的《统计学习方法》，这里就约定$\mathrm{p}(\boldsymbol{x})$与$\mathrm{p}(\boldsymbol{x}, \mathrm{z_k} = 1)$是等同的）

上式中我们定义符号$\gamma(\mathrm{z_k})$来表示来表示第k个分量的后验概率。在贝叶斯的观点下，π_{k}可视为$\mathrm{z_k} = 1$的先验概率。

上述内容改写了GMM的形式，并引入了隐含变量$\boldsymbol{z}$和已知$\boldsymbol{x}$后的的后验概率$\gamma(\mathrm{z_k})$，这样做是为了方便使用EM算法来估计GMM的参数。

EM算法估计GMM参数

EM算法（Expectation-Maximization algorithm）分两步，第一步先求出要估计参数的粗略值，第二步使用第一步的值最大化似然函数。因此要先求出GMM的似然函数。

假设$\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$，对于图2，$\boldsymbol{x}$是图中所有点（每个点有在二维平面上有两个坐标，是二维向量，因此$\boldsymbol{x}_1, \boldsymbol{x}_2$等都用粗体表示）。GMM的概率模型如(1)式所示。GMM模型中有三个参数需要估计，分别是$\boldsymbol{\pi}$，$\boldsymbol{\mu}$和$\boldsymbol{\Sigma}$. 将(1)式稍微改写一下：

$$\mathrm{p}(\boldsymbol{x}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{\mathrm{k}=1}^{\mathrm{K}} \pi_{\mathrm{k}} \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{\mathrm{k}}, \boldsymbol{\Sigma}_{\mathrm{k}}) \tag{6}$$

为了估计这三个参数，需要分别求解出这三个参数的最大似然函数。先求解μ_{k}的最大似然函数。样本符合iid，(6)式所有样本连乘得到最大似然函数，对(6)式取对数得到对数似然函数，然后再对$\boldsymbol{\mu}_k$求导并令导数为0即得到最大似然函数。

$$0 = -\sum_{\mathrm{n}=1}^{\mathrm{N}} \frac{\pi_{\mathrm{k}} \mathcal{N}(\boldsymbol{x}_{\mathrm{n}}|\boldsymbol{\mu}_{\mathrm{k}}, \boldsymbol{\Sigma}_{\mathrm{k}})}{\sum_{\mathrm{j}} \pi_{\mathrm{j}} \mathcal{N}(\boldsymbol{x}_{\mathrm{n}}|\boldsymbol{\mu}_{\mathrm{j}}, \boldsymbol{\Sigma}_{\mathrm{j}})} \boldsymbol{\Sigma}_{\mathrm{k}}^{-1}(\boldsymbol{x}_{\mathrm{n}} - \boldsymbol{\mu}_{\mathrm{k}}) \tag{7}$$

注意到上式中分数的一项的形式正好是(5)式后验概率的形式。两边同乘$\boldsymbol{\Sigma}_{\mathrm{k}}$，重新整理可以得到：

$$\boldsymbol{\mu}_{\mathrm{k}} = \frac{1}{\mathrm{N}_{\mathrm{k}}} \sum_{\mathrm{n}=1}^{\mathrm{N}} \gamma(\mathrm{z}_{\mathrm{nk}}) \boldsymbol{x}_{\mathrm{n}} \tag{8}$$

其中：

$$\mathrm{N}_{\mathrm{k}} = \sum_{\mathrm{n}=1}^{\mathrm{N}} \gamma(\mathrm{z}_{\mathrm{nk}}) \tag{9}$$

(8)式和(9)式中，N表示点的数量。$\gamma(\mathrm{z}_{\mathrm{nk}})$表示点$\mathrm{n}$（$\boldsymbol{x}_{\mathrm{n}}$）属于聚类$\mathrm{k}$的后验概率。**则$\mathrm{N}_{\mathrm{k}}$可以表示属于第$\mathrm{k}$个聚类的点的数量**。那么$\boldsymbol{\mu}_{\mathrm{k}}$表示所有点的加权平均，每个点的权值是$\sum_{\mathrm{n}=1}^{\mathrm{N}} \gamma(\mathrm{z}_{\mathrm{nk}})$，跟第$\mathrm{k}$个聚类有关。

同理求$\boldsymbol{\Sigma}_{\mathrm{k}}$的最大似然函数，可以得到：

$$\boldsymbol{\Sigma}_{\mathrm{k}} = \frac{1}{\mathrm{N}_{\mathrm{k}}} \sum_{\mathrm{n}=1}^{\mathrm{N}} \gamma(\mathrm{z}_{\mathrm{nk}})(\mathrm{x}_{\mathrm{n}} - \boldsymbol{\mu}_{\mathrm{k}})(\mathrm{x}_{\mathrm{n}} - \boldsymbol{\mu}_{\mathrm{k}})^{\mathrm{T}} \tag{10}$$

最后剩下π_k的最大似然函数。注意到π_k有限制条件$\sum_{k=1}^{K}\pi_k=1$，因此我们需要加入拉格朗日算子：

$$\ln p(\boldsymbol{x}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma})+\lambda(\sum_{k=1}^{K}\pi_k-1)$$

求上式关于π_k的最大似然函数，得到：

$$0=\sum_{n=1}^{N}\frac{\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_j\pi_j\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}+\lambda \tag{11}$$

上式两边同乘π_k，我们可以做如下推导：

$$0=\sum_{n=1}^{N}\frac{\pi_k\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_j\pi_j\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}+\lambda\pi_k \tag{11.1}$$

结合公式(5)、(9)，可以将上式改写成：

$$0=N_k+\lambda\pi_k \tag{11.2}$$

注意到$\sum_{k=1}^{K}\pi_k=1$，上式两边同时对k求和。此外N_k表示属于第个聚类的点的数量（公式(9)）。对N_k，从$k=1$到$k=K$求和后，就是所有点的数量N：

$$0=\sum_{k=1}^{K}N_k+\lambda\sum_{k=1}^{K}\pi_k \tag{11.3}$$

$$0=N+\lambda \tag{11.4}$$

从而可得到$\lambda=-N$，带入(11.2)，进而可以得到π_k更简洁的表达式：

$$\pi_k=\frac{N_k}{N} \tag{12}$$

EM算法估计GMM参数即最大化(8)，(10)和(12)。需要用到(5)，(8)，(10)和(12)四个公式。我们先指定$\boldsymbol{\pi}$，$\boldsymbol{\mu}$和$\boldsymbol{\Sigma}$的初始值，带入(5)中计算出$\gamma(z_{nk})$，然后再将$\gamma(z_{nk})$带入(8)，(10)和(12)，求得π_k，$\boldsymbol{\mu}_k$和$\boldsymbol{\Sigma}_k$；接着用求得的π_k，$\boldsymbol{\mu}_k$和$\boldsymbol{\Sigma}_k$再带入(5)得到新的$\gamma(z_{nk})$，再将更新后的$\gamma(z_{nk})$带入(8)，(10)和(12)，如此往复，直到算法收敛。

EM算法

1. 定义分量数目K，对每个分量k设置π_{k}，$\boldsymbol{\mu}_{\mathrm{k}}$和$\boldsymbol{\Sigma}_{\mathrm{k}}$的初始值，然后计算(6)式的对数似然函数。
2. **E step**
   根据当前的π_{k}、$\boldsymbol{\mu}_{\mathrm{k}}$、$\boldsymbol{\Sigma}_{\mathrm{k}}$计算后验概率$\gamma(\mathrm{z}_{\mathrm{nk}})$

$$\gamma(\mathrm{z}_{\mathrm{nk}}) = \frac{\pi_k N(x_n|\mu_{\mathrm{k}}, \sum_k)}{\sum_{j=1}^{K} N(x_n|\mu_j, \sum_j)}$$

   $\gamma(\mathrm{z}_{\mathrm{nk}})$:当前第$\mathrm{n}$个样本属于第$\mathrm{k}$类的概率,维度为$400$，$\mathrm{N}(x_n|\mu_{\mathrm{k}}, \sum_k)$: 多元高斯正态分布
3. **M step**
   根据E step中计算的$\gamma(\mathrm{z}_{\mathrm{nk}})$再计算新的$\pi_{\mathrm{k}}$、$\boldsymbol{\mu}_{\mathrm{k}}$、$\boldsymbol{\Sigma}_{\mathrm{k}}$

$$\boldsymbol{\mu}_{\mathrm{k}}^{\mathrm{new}} = \frac{1}{\mathrm{N}_{\mathrm{k}}} \sum_{\mathrm{n}=1}^{\mathrm{N}} \gamma(\mathrm{z}_{\mathrm{nk}}) \boldsymbol{x}_{\mathrm{n}}$$

$$\boldsymbol{\Sigma}_{\mathrm{k}}^{\mathrm{new}} = \frac{1}{\mathrm{N}_{\mathrm{k}}} \sum_{\mathrm{n}=1}^{\mathrm{N}} \gamma(\mathrm{z}_{\mathrm{nk}}) (\boldsymbol{x}_{\mathrm{n}} - \boldsymbol{\mu}_{\mathrm{k}}^{\mathrm{new}})^{\mathrm{T}} (\boldsymbol{x}_{\mathrm{n}} - \boldsymbol{\mu}_{\mathrm{k}}^{\mathrm{new}})$$

$$\pi_{\mathrm{k}}^{\mathrm{new}} = \frac{\mathrm{N}_{\mathrm{k}}}{\mathrm{N}}$$

   其中:

$$\mathrm{N}_{\mathrm{k}} = \sum_{\mathrm{n}=1}^{\mathrm{N}} \gamma(\mathrm{z}_{\mathrm{nk}})$$