经典聚类：混合高斯知多少

---整理者：潘振福

最近学习，想搞搞无监督的聚类问题，所以挑选一个经典的混合高斯聚类重点学习学习，今天给大家分享一下，目前，研究还比较浅，该笔记适合菜鸟交流学习。

1.高斯分布

高斯分布又称正态分布，广泛应用于物理数据，信号数据的拟合与聚类中。混合高斯分布由单变量的高斯分布组成，单变量 $\boldsymbol{x}$ 的高斯分布为

$$N(x\mid\mu,\sigma^2)=\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\} \tag{1}$$

其中 μ 是均值，σ^2 是方差。

对于一个 D 维的向量 $\vec{x}$，多维的高斯分布为

$$N(\vec{x}\mid\vec{\mu},\Sigma)=\frac{1}{((2\pi)^D\mid\Sigma\mid)^{1/2}}\exp\{-\frac{1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})\} \tag{2}$$

$$\mu=\frac{1}{m}\sum_{i=1}^{m}x^{(i)} \tag{3}$$

$$\Sigma=\frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu)(x^{(i)}-\mu)^T \tag{4}$$

其中 $\vec{\mu}$ 是一个 D 维的均值向量，Σ 是 $D\times D$ 的协方差矩阵，$|\Sigma|$ 是 Σ 的行列式。

高斯分布有很多很重要的性质，建议前往《统计与模式识别》教材拜读。

公式(3)(4) μ 归纳了样本数据的中心位置，协方差矩阵 Σ 是个对称矩阵，$x^{(i)}$ 表示样例，共有 m 个，每个样例 D 个特征，因此 μ 是 n 维向量，Σ 是 n*n 协方差矩阵。

当 m<<D 时，我们会发现 Σ 是奇异阵（$|\Sigma|=0$），也就是说 Σ^{-1} 不存在，没办法拟合出多元高斯分布了，确切的说是我们估计不出来 Σ。

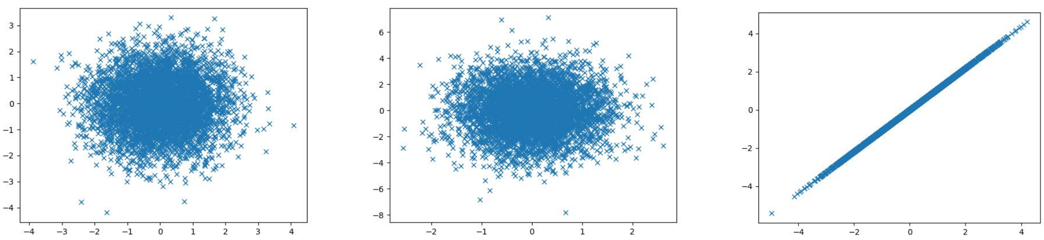
以下代码展示，高斯分布与 Σ 的关系，分布的聚散程度，跟协方差矩阵的秩的大小有关：

```
#coding=utf-8
import numpy as np
import matplotlib.pyplot as plt
#  标准圆形,如图(a)
mean = [0,0]
```

```
cov = [[1,0],
       [0,1]]
#  椭圆形，其中椭圆的横向与纵向都与坐标轴平行,如图(b)
# cov = [[0.5,0],
#        [0,3]]
#  椭圆，其轴向任意,如图(c)

# cov = [[1,2.3],
#        [2.3,1.4]]

x,y = np.random.multivariate_normal(mean,cov,5000).T
plt.plot(x,y,'x')
plt.show()
```



(a)(b)(c)

图 1 二维高斯分布平面示意图

2.混合高斯分布模型

高斯分布有一些重要的分析性质，但在利用它为实际数据建模时有很大的局限性。考虑如图 2 的例子，数据集形成两个占主导地位的团，简单高斯分布并不能捕捉这里的结构，而两个高斯分布的线性叠加可以更好地给出数据集中的特征。

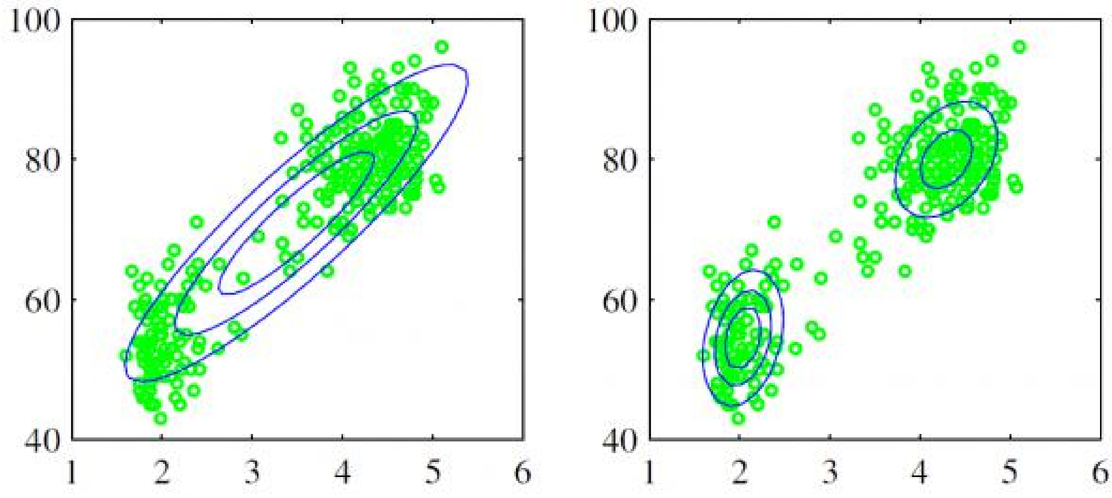


图 2 基本高斯分布拟合与混合高斯分布拟合示意图

由基本分布(如高斯分布)的线性组合得到的概率模型被称为混合分布(Mixture Distribution),高斯分布的线性组合可以给出非常复杂的密度。如图 3 所示，三个高斯分布（蓝线）混合为一个更复杂的分布（红线）。如果使用足够多的高斯分布，通过调整均值、方差和线性组合中的参数，基本能得到任何连续密度的任意精度的近似。

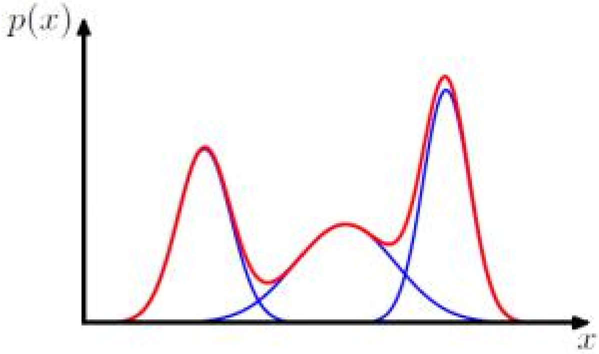


图 3 混合高斯拟合

K 个高斯密度的线性组合形式为

$$p(\vec{x})=\sum_{k=1}^{K}\pi_k N(\vec{x}\mid\vec{\mu},\Sigma_k) \tag{5}$$

它被称为混合高斯。每个高斯密度 $N(\vec{x}\mid\vec{\mu},\Sigma_k)$ 是混合的一个组成部分。混合模型可以包含任意的分布模型，当然我们只是拿高斯分布作为一个例子，高斯模型有众多的优良的性质，公式规范，方便参数估计。

在公式(5)中，参数 π_k 被称为混合系数，其中

$$\sum_{k=1}^{K}\pi_k=1 \tag{6}$$

根据概率论的和法则与积法则，边缘概率密度为

$$p(\vec{x})=\sum_{k=1}^{K}p(k)p(\vec{x}\mid k) \tag{7}$$

如果把 $\pi_k=p(k)$ 看做选中第 k 个部分的概率，则上式(7)等价于式(5)，其中密度 $N(\vec{x}\mid\vec{\mu}_k,\Sigma_k)=p(\vec{x}\mid k)$ 是以 k 为条件 $\vec{x}$ 的概率密度。

根据贝叶斯定理得到

$$\begin{aligned}\gamma_k(\vec{x})&=p(k\mid\vec{x})\\&=\frac{p(k)p(\vec{x}\mid k)}{\sum_{l=1}^{K}p(l)p(\vec{x}\mid l)}\\&=\frac{\pi_k N(\vec{x}\mid\vec{\mu}_k,\Sigma_k)}{\sum_{l=1}^{K}\pi_l N(\vec{x}\mid\vec{\mu}_l,\Sigma_l)}\end{aligned} \tag{8}$$

这里是个矩阵值，混合高斯分布被参数 $\vec{\pi}=\{\pi_1,\pi_2,\ldots,\pi_{\mathrm{k}}\}$、$\vec{\mu}=\{\vec{\mu}_1,\ldots,\vec{\mu}_{\mathrm{k}}\}$ 和 $\vec{\Sigma}=\{\Sigma_1,\ldots,\Sigma_{\mathrm{k}}\}$ 所控制。确定这些参数的一种方法是利用最大似然，的对数似然函数为

$$\ln p(X|\vec{\pi},\vec{\mu},\vec{\Sigma})=\sum_{n=1}^{N}\ln\{\sum_{k=1}^{K}\pi_k N(\vec{x}_n|\vec{\mu}_k,\Sigma_k)\} \tag{9}$$

其中 $X=\{\vec{x}_1,...,\vec{x}_N\}$。我们注意到情况比单高斯时更加复杂，因为对数中出现了和。这种情况下，我们不能利用封闭的解析解来表示最大似然解，一种方法是利用迭代数值优化技术，另外一种方法是利用期望最大。

通过 EM 算法可以求出：

1.E：求期望值：

$$\begin{aligned}\gamma_k(\vec{x}_i)&=p(k|\vec{x}_i)\\&=\frac{p(k)p(\vec{x}_i|k)}{\sum_{l=1}^{K}p(l)p(\vec{x}_i|l)}\\&=\frac{\pi_k N(\vec{x}_i|\vec{\mu}_k,\Sigma_k)}{\sum_{l=1}^{K}\pi_l N(\vec{x}_i|\vec{\mu}_l,\Sigma_l)}\end{aligned} \tag{10}$$

2.M:最大似然函数值

EM 算法是的鸡生蛋，蛋生鸡的问题，根据 $\gamma_k(\vec{x}_i)$ 计算出新的 π_k^{new}， μ_k^{new}， $\sum_k^{new}$，有

$$\mu_k^{new}=\frac{1}{N_k}\sum_{n=1}^{N}\gamma_k(\vec{x}_n)x_n$$

$$\Sigma_k^{new}=\frac{1}{N_k}\sum_{n=1}^{N}\gamma_k(\vec{x}_n)(\vec{x}_n-\mu_k^{new})^T(\vec{x}_n-\mu_k^{new})$$

$$\pi_k^{new}=\frac{N_k}{N}$$

其中 $N_k=\sum_{n=1}^{N}\gamma_k(x_n)$

然后计算似然函数公式(9),判断是否达到最大值(只能是局部最大值)，或者达到最大迭代数，终止迭代。（具体公式推导，详见：http://www.cs.cmu.edu/~awm/doc/gmm-algebra.pdf）

经典代码实现：https://github.com/panzhenfu/GMM_py

最终代码实现的结果：如图所示(**喜欢我的代码记得打星哦**)