EM 算法估计 GMM 参数

ZY2203402 冯士轩

1、 摘要

从高斯分布中抽取身高样本数据，基于这些数据使用 EM 算法对高斯混合分布进行参数估算并进行预算。

2、 介绍

（1）高斯混合模型

高斯混合模型（Gaussian Mixture Model，GMM）是一种对数据进行建模的方法。它假设数据来自于多个高斯分布，这些高斯分布之间又存在一定的关系或者权重。因此，GMM 是一种概率模型，可以用来描述离散或连续变量的数据分布。

$$P(y)=\sum_{k=1}^{K}\alpha_k\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(y-\mu_k)^T\Sigma^{-1}(y-\mu_k)\right)$$

其中 d 为数据的维度，μ_k为均值，Σ为协方差矩阵。

对于二维高斯混合模型，d=2，y 和μ_k都是二维的数据，用矩阵表示就是一行两列，Σ则是两行两列的协方差矩阵。

在 GMM 中，我们首先假设目标数据集中存在多个不同的高斯分布。然后，我们使用每个高斯分布的均值、方差和权重参数来描述这些分布，从而得到一个总体的分布模型。我们还需要通过某些手段，如 EM 算法等，来确定各个高斯分布的均值、方差和权重参数。

在实际应用中，GMM 可以用来进行聚类、密度估计和异常检测等任务。与传统的 k-Means 算法相比，GMM 能够更有效地处理数据分布复杂、离群点较多的情况。同时，GMM 也被广泛应用于图像处理、语音识别及模式识别等领域。

（2）EM 算法

EM 算法（Expectation-Maximization algorithm）是一种经典的迭代优化算法，主要用于寻找具有隐变量（latent variable）的概率模型中的最大似然估计或最大后验估计。EM 算法的基本思路是在每次迭代中，通过先验分布和当前参数下的观测数据来计算隐变量的后验分布，并求出对数似然函数关于隐变量后验分布的期望值（E 步）。然后，在保持隐变量的后验分布不变的情况下，通过最大化隐变量的后验分布相关的对数似然函数，来获取新的参数估计值（M 步）。迭代这个过程直到收敛为止。

EM 算法主要分为 E 步和 M 步

E 步（Expectation）：计算隐变量的后验概率分布，即给定当前参数下的观测数据，计算每个可能状态的后验概率分布。

$$\hat{\gamma}_{jk}=\frac{\alpha_k\phi(y_j\mid\theta_k)}{\sum_{k=1}^{K}\alpha_k\phi(y_j\mid\theta_k)},j=1,2,\dots,N;k=1,2,\dots,K$$

M 步（Maximization）：使用 E 步得到的隐变量的后验概率分布更新模型参数，即对隐变量的后验概率分布加权平均得到模型对数似然函数的期望，从而得到新的参数估计值。

$$\hat{\mu}_k = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk} y_j}{\sum_{j=1}^{N} \hat{\gamma}_{jk}}, k = 1,2,\dots,K$$

$$\hat{\Sigma}_k = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk} (y_j - \mu_k)^T (y_j - \mu_k)}{\sum_{j=1}^{N} \hat{\gamma}_{jk}}, k = 1,2,\dots,K$$

$$\hat{\alpha}_k = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk}}{N}, k = 1,2,\dots,K$$

EM 算法的优点在于它可以在存在隐变量的问题中进行参数估计，而且在许多情况下可以保证收敛到全局最优解。EM 算法常常被用于聚类、密度估计和图像处理等领域中，比如高斯混合模型的参数估计就是通过 EM 算法来实现的。

3 实验操作

（1） 对数据进行处理

对均值分别为 164、176，方差为 3 和 5 的高斯分布进行取样，生成 500 个分布 1 的数据和 1500 个分布 2 的数据。

```
mean1, std1 = 164, 3
mean2, std2 = 176, 5

# 从两个高斯分布中生成各50个样本
data1 = np.random.normal(mean1, std1, 500)
data2 = np.random.normal(mean2, std2, 1500)
data = np.concatenate((data1, data2), axis=0)

# 将数据写入 CSV 文件
df = pd.DataFrame(data, columns=['height'])
df.to_csv('height_data.csv', index=False)

# 绘制数据的直方图
plt.hist(data, bins=20)
plt.xlabel('Height (cm)')
plt.ylabel('Count')
plt.title('Distribution of Heights')
plt.show()
```

（2） 然后对高斯混合分布模型进行初始化，根据经验将均值、方差、权重分别设为如下图所示的数据

```
mu1 = 170; sigma1 = 6; w1 = 0.6
mu2 = 160; sigma2 = 2; w2 = 0.4
```

（3） 定义 EM 算法函数，得到更新后的高斯混合分布参数

```python
def em(h, mu1, sigma1, w1, mu2, sigma2, w2):
    d = 1
    n = len(h)  # 样本长度
    p1 = w1 * stats.norm(mu1, sigma1).pdf(h)
    p2 = w2 * stats.norm(mu2, sigma2).pdf(h)
    # p1, p2权重 * 男女生的后验概率
    R1i = p1 / (p1 + p2)
    R2i = p2 / (p1 + p2)
    # M-step
    # mu的更新
    mu1 = np.sum(R1i * h) / np.sum(R1i)
    mu2 = np.sum(R2i * h) / np.sum(R2i)
    # sigma1的更新
    sigma1 = np.sqrt(np.sum(R1i * np.square(h - mu1)) / (d * np.sum(R1i)))
    sigma2 = np.sqrt(np.sum(R2i * np.square(h - mu2)) / (d * np.sum(R2i)))
    # w的更新
    w1 = np.sum(R1i) / n
    w2 = np.sum(R2i) / n

    return mu1, sigma1, w1, mu2, sigma2, w2
```

（4） 通过 500 次迭代可以得到高斯混合分布参数

```python
for iteration in range(1000):
    mu1,sigma1,w1,mu2,sigma2,w2=em(h,mu1,sigma1,w1,mu2,sigma2,w2)
```

（5） 通过计算输入身高的概率密度的对比，判断所处类别。

```python
#预测
while(1):
 height =int(input("输入身高："))
 prob1 = [w1 * norm.pdf(height, mu1, sigma1)]
 prob2 = [w2 * norm.pdf(height, mu2, sigma2)]
 if prob1 < prob2:
    print("可能是女性")
 else:
    print("可能是男性")
```

```
输入身高：160
可能是女性
输入身高：170
可能是男性
输入身高：180
可能是男性
```

4　结果分析

（1）　当初始化参数设为：

mu1 = 170; sigma1 = 6; w1 = 0.6

mu2 = 160; sigma2 = 2; w2 = 0.4，

可得出结果

```
mu1:  175.9467262250708
sigma1:  5.065749732805323
w1:  0.7459849184834813
mu2:  164.1983279357066
sigma2:  2.9377119101520455
w2:  0.2540150815165188
```

可以看出预估男性均值为 175.9467262250708 方差为 5.065749732805323 所占比例为 0.7459849184834813。

女性均值为 164.1983279357066 方差为 2.9377119101520455 所占比例 0.2540150815165188。

（2）当初始参数设为

mu1 = 180; sigma1 = 5; w1 = 0.9

mu2 = 170; sigma2 = 4; w2 = 0.1

结果为

```
mu1:  175.63961871324972
sigma1:  5.153750796261923
w1:  0.7676216302896826
mu2:  163.7914508273397
sigma2:  2.7340756070345735
w2:  0.2323783697103174
```

可以看出预估男性均值为 175.63961871324972 方差为 5.153750796261923 所占比例为 0.7676216302896826。

女性均值为 163.7914508273397 方差为 2.7340756070345735 所占比例 0.2323783697103174。

从上面两次实验对比可知，第一次拟合效果比第二次要好，可以推断出初始值的设定可以决定着参数评估效果的好坏。