模式识别 第三次作业 报告

杨永祎 17300240038

一 任务简述和数据

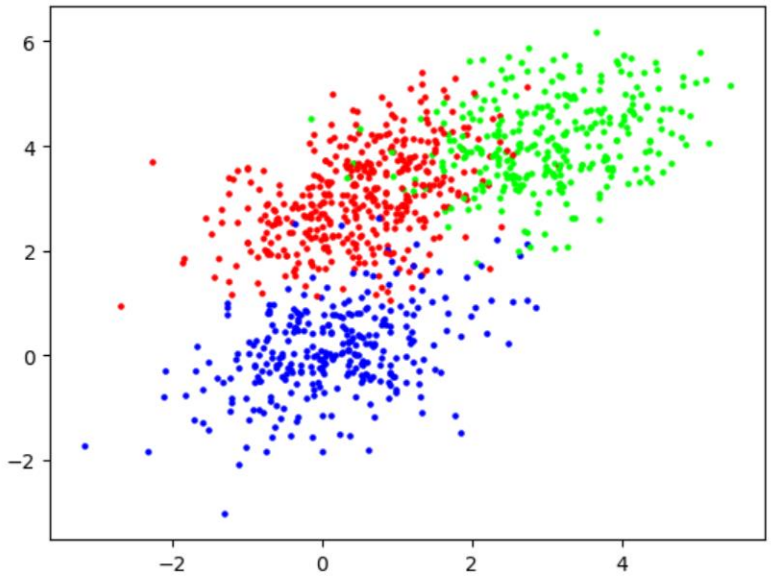
本次任务要求生成一组集群数据并用混合高斯模型进行分类。

我简单地造了一个由三个二维高斯分布生成的数据。

三个分布的均值分别为 (0.1,0.1)，(3,4)，(0.5,3)，协方差分别为 $\begin{bmatrix}0.9 & 0.3\\0.3 & 0.7\end{bmatrix}$，$\begin{bmatrix}0.8 & 0.2\\0.2 & 0.7\end{bmatrix}$，$\begin{bmatrix}0.8 & 0.3\\0.3 & 0.8\end{bmatrix}$。

一共 1000 个点，并且三种分布以 0.3/0.3/0.4 的概率采样。

可视化如下



二 模型

使用高斯混合模型进行分类。

具体来说，需要拟合三个二维高斯分布，并且模型需要拟合三种分布的先验概率 p。
使用 EM 算法来迭代地求模型参数。
每次迭代分为两个阶段，E-step 和 M-step。

E-step：求出每个样本点属于每一类的概率密度，可以由先验概率乘以正态分布的概率密度公式算出：

$$\gamma_{ik} = p(x_i|\mu_k, \Sigma_k) = p_k \times \frac{1}{\sqrt{(2\pi)^2|\Sigma_k|}} e^{(-\frac{1}{2}(x_i-\mu_k)^T \Sigma_k^{-1}(x_i-\mu_k))}$$

M-step：根据γ更新参数

$$\mu_k = \frac{\sum_i \gamma_{ik} x_j}{\sum_i \gamma_{ik}}$$

$$\Sigma_k = \frac{\sum_i \gamma_{ik}(x_j - \mu_k)(x_j - \mu_k)^T}{\sum_i \gamma_{ik}}$$

$$p_k = \frac{\sum_i \gamma_{ik}}{N}$$

如此迭代 1000 步之后结果基本收敛。

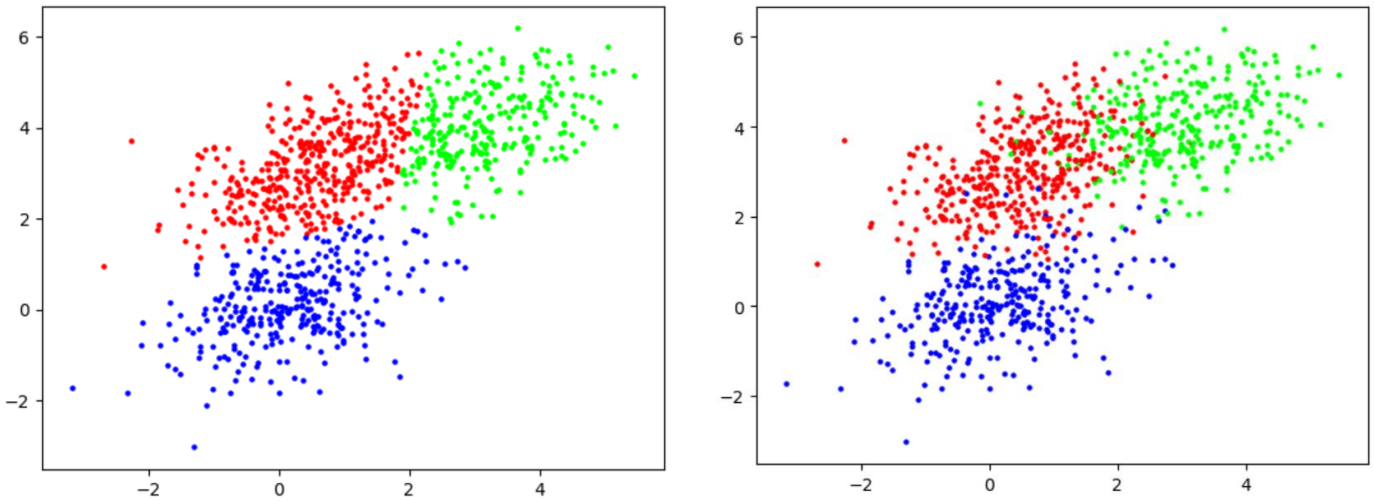
三 评测

因为这是一个无监督分类任务，所以不能直接算 accurancy。不过既然这是个自己造的数据集，可以用这样一种简单的方式来给出一个评测：
首先用某种将模型拟合出的分布与生成数据集所用的分布（我称后者为神谕分布）一对一地对齐。然后检查预测为此分布的样本中有多少确实是其对应的神谕分布生成的（也就是对齐后检查 accurancy）。

我用一种非常简单的贪心策略来将预测分布与神谕分布对齐：每次取一个神谕分布，然后从还未找到对应的预测分布中找到与其均值（均值是二维向量）的欧氏距离最近的预测分布，将这两者对应。

四 结果

迭代 1000 步之后模型基本收敛。用上一节阐述的算 accurancy 的方法求出的 accurancy 是 0.9130。
预测数据可视化如下左图，作为对比，我把真实标签放在右图



左：模型预测聚类 右：真实聚类

可以看到在生成的时候这几个聚类本身有重叠部分，模型用非常明显的边界把这几个聚类分隔开了，这是比较符合预期的行为。