哈爾濱工業大學

模式识别与机器学习实验报告

实验 三

题　　　目	K-Means 与 GMM 聚类
学　　　院	未来技术学院
专　　　业	人工智能
学　　　号	2023112419
学　　　生	陈铠
任 课 教 师	刘扬

哈尔滨工业大学计算机科学与技术学院

2025 年秋季

一、 实验内容

（一）Kmeans 和 GMM 简要介绍

K-Means 聚类算法和高斯混合模型（GMM）聚类算法均为针对无监督学习的分类模型，前者属于硬聚类，后者属于软聚类。GMM 是基于生成式的

1.数学模型

K-means 聚类：将数据集$\{x_i\}_{i=1}^{n}$划分为 K 个簇（cluster），每个簇由一个质心（mean）表示。K-Means 的目标是最小化每个点到其簇中心的距离和：

$$J=\sum_{i=1}^{n} \sum_{k=1}^{K} \mathbf{1}\,(c_i=k)||x_i-\mu_k||^2$$

其中$\mathbf{1}(\cdot)$是指示函数，若样本i被分到簇k则为 1，否则为 0；μ_k为第k个簇的均值（质心）；c_i为第i个样本的簇标号。
常使用硬分配规则：

$$c_i=\arg\min_k ||x_i-\mu_k||^2$$

GMM 聚类：假设数据由权重不同的多簇高斯分布（高斯混合分布）结合而成：

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

其中π_k为混合系数($\pi_k\geq 0,\sum_k\pi_k=1$)，$\mathcal{N}(x\mid\mu,\Sigma)$是多元高斯密度函数。

对第j个样本x_j，属于第i个分量的后验概率（责任度）为：

$$\gamma_{j,i}\equiv p\left(z_j=i\mid x_j\right)=\frac{\pi_i,\mathcal{N}\left(x_j\mid\mu_i,\Sigma_i\right)}{\sum_{l=1}^{K}\pi_l,\mathcal{N}\left(x_j\mid\mu_l,\Sigma_l\right)}$$

$$\hat{c}_j=\arg\max_i \gamma_{j,i}$$

令$N_i=\sum_{j=1}^{n}\gamma_{j,i}$，可解出参数$\pi_k,\mu_k,\Sigma_k$

$$\pi_i=\frac{N_i}{n}$$

$$\mu_i=\frac{1}{N_i}\sum_{j=1}^{n}\gamma_{j,i},x_j$$

$$\Sigma_i=\frac{1}{N_i}\sum_{j=1}^{n}\gamma_{j,i}(x_j-\mu_i)(x_j-\mu_i)^\top$$

2.优化方法

K-means 的最优解求解问题是组合问题，属于 NP 困难问题，现实中常采用迭代方法求解：

初始化 k 个聚类中心，指派数据到最近的中心的类中；

将每个类的样本均值作为新的聚类中心。重复上述步骤，直到收敛为止。

GMM 的优化常基于最大似然：

$$\mathcal{L}(\Theta) = \sum_{j=1}^{n} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(x_j \mid \mu_k, \Sigma_k)$$

采用 EM 算法：

E 步：计算每个点属于各高斯分布的后验概率$\gamma_{j,i} = p(z_j = i \mid x_j; \Theta^{(t)})$；

M 步：采用$\gamma_{j,i}$的加权统计量更新模型参数π_k, μ_k, Σ_k

（二）实验研究内容

分别使用两种算法对生成的高斯数据分类；

对 GMM 使用真实数据集进行检验

二、 实验环境

google colab 云平台

jupyter notebook

三、 实验结果及分析

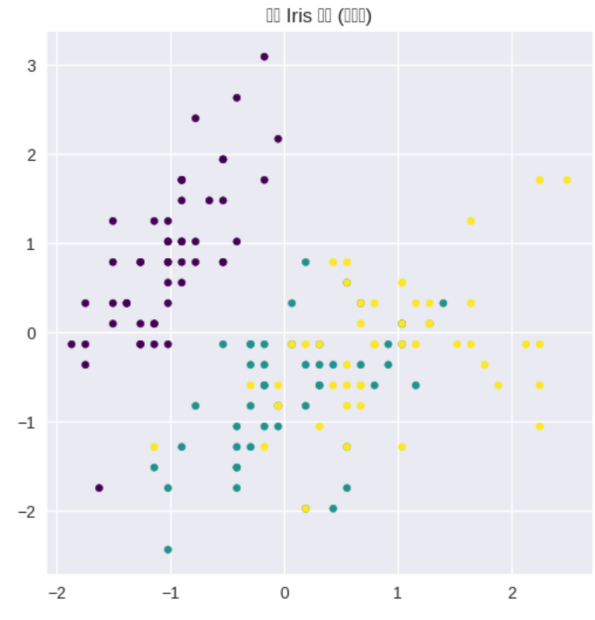
（一）两种算法对生成的高斯数据分类

第一行为 3 次随机生成的不同数据集

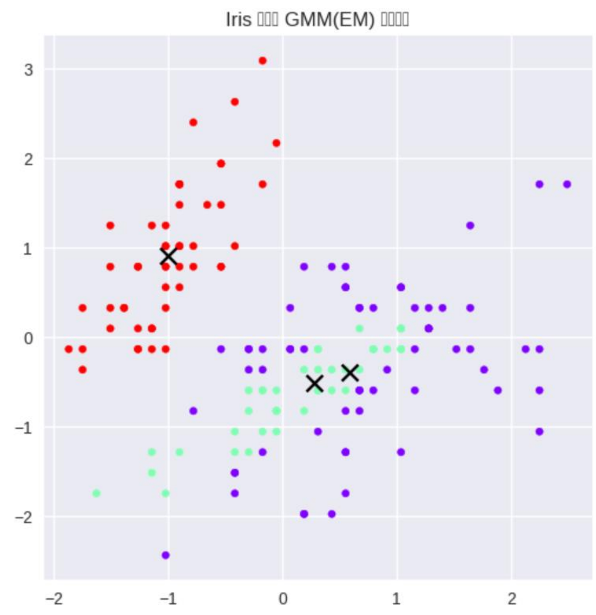
第二、三行分别为 Kmeans 和 GMM 的聚类结果

结果表明，在对生成的高斯数据进行分类时，数据集的分布状态对聚类算法的效果影响显著，且 K-means 算法对初始聚类中心的选择，对最终聚类结果的影响尤为突出。

(二)GMM 对真实数据集进行验证



iris 数据集前两维



GMM 训练结果

最终准确率为 72.67%。

四、 结论

K-means 与 GMM 均为 EM 算法的具体实现形式，二者均包含隐变量，且遵循“E 步（期望步）-M 步（最大化步）”的迭代优化逻辑，能够有效解决简单数据分类问题。但两者存在共同固有局限：仅能收敛至局部最优解，且最终聚类效果对初始值（初值）高度敏感，初值选取不当会显著降低分类精度。

K-means 的模型假设更强：认为各聚类对总模型的贡献相等，且样本对聚类的归属为“硬分配”（即样本属于某一聚类的概率为 1，属于其他聚类的概率为 0）；同时假设数据呈球状分布，以欧氏距离衡量样本与聚类中心的相似度，其本质是 GMM 的特殊形式（聚类贡献固定为 1/k、变量间协方差矩阵为对角阵）。

相比之下，GMM 的假设更宽松：允许各高斯模型对总模型的贡献存在权重差异，样本对聚类的归属为“软分配”（即样本属于某一聚类的概率为连续值），无需依赖数据球状分布假设，可适配更复杂的数据分布。但 GMM 存在特有风险：若初始高斯模型的均值、方差选取不佳，易出现“极大似然值为 0”（样本几乎无法由初始模型生成）或“协方差矩阵不可逆”的问题，而 K-means 无此类额外计算风险。

在高斯数据分类场景中，相同聚类数下，K-means 的分类效果优于基于 EM 算法的 GMM，且 K-means 的迭代次数更少，运算效率更高。考虑 EM 算法对初值的敏感程度高于 K-means，实验中可利用 K-means 的聚类结果作为 GMM（或 EM 算法）的初始值，以降低初值选取不当对 GMM 分类效果的影响。此外，K-means 的性能依赖欧氏距离计算，其优缺点会随距离衡量方式的变化而体现。

五、 参考文献

[1]（美）SHELDON AXLER 著；杜现坤，刘大艳，马晶译. 线性代数应该这样学 第 3 版[M]. 北京：人民邮电出版社, 2016.10.
[2]周志华著. 机器学习[M]. 北京：清华大学出版社, 2016.01.
[3]谢文睿，秦州编著. 机器学习公式详解[M]. 北京：人民邮电出版社, 2021.03.
[4]李航著. 统计学习方法 第 2 版[M]. 北京：清华大学出版社, 2019.05.