

**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

1. **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：

其中是指示函数，若样本被分到簇则为1，否则为0；为第个簇的均值（质心）；为第个样本的簇标号。

常使用硬分配规则：

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：

其中为混合系数，是多元高斯密度函数。

对第个样本，属于第个分量的后验概率（责任度）为：

令，可解出参数

**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：

采用EM算法：

E步：计算每个点属于各高斯分布的后验概率；

M步：采用的加权统计量更新模型参数

**（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

1. **实验环境**

google colab云平台

jupyter notebook

1. **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%。

1. **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。

1. **参考文献**

[1]（美）SHELDON AXLER著；杜现坤，刘大艳，[马晶](https://zhida.zhihu.com/search?content_id=183206416&content_type=Article&match_order=1&q=%E9%A9%AC%E6%99%B6&zhida_source=entity)译. [线性代数应该这样学](https://zhida.zhihu.com/search?content_id=183206416&content_type=Article&match_order=1&q=%E7%BA%BF%E6%80%A7%E4%BB%A3%E6%95%B0%E5%BA%94%E8%AF%A5%E8%BF%99%E6%A0%B7%E5%AD%A6&zhida_source=entity) 第3版[M]. 北京：人民邮电出版社, 2016.10.

[2][周志华](https://zhida.zhihu.com/search?content_id=183206416&content_type=Article&match_order=1&q=%E5%91%A8%E5%BF%97%E5%8D%8E&zhida_source=entity)著. 机器学习[M]. 北京：清华大学出版社, 2016.01.

[3]谢文睿，[秦州](https://zhida.zhihu.com/search?content_id=183206416&content_type=Article&match_order=1&q=%E7%A7%A6%E5%B7%9E&zhida_source=entity)编著. 机器学习公式详解[M]. 北京：[人民邮电出版社](https://zhida.zhihu.com/search?content_id=183206416&content_type=Article&match_order=2&q=%E4%BA%BA%E6%B0%91%E9%82%AE%E7%94%B5%E5%87%BA%E7%89%88%E7%A4%BE&zhida_source=entity), 2021.03.

[4][李航](https://zhida.zhihu.com/search?content_id=183206416&content_type=Article&match_order=1&q=%E6%9D%8E%E8%88%AA&zhida_source=entity)著. 统计学习方法 第2版[M]. 北京：清华大学出版社, 2019.05.