**2025年春季学期《机器学习》实验报告**

**（第四次实验）**

班级： 230617 学号： 23373179 姓名： 周子强

一、在K-means聚类实验中，初始聚类中心的选择对算法收敛性和聚类结果有重要影响。请说明K-means算法的基本原理和收敛条件，并分析随机初始化、K-means++初始化方法的优缺点。如何确定最优的聚类数量*K*？

**（1）K-means算法的基本原理**

目标：将数据集划分为K个簇，使得簇内样本的相似度最大，簇间样本的相似度最小。

算法流程：

1. 初始化：选择K个初始聚类中心。
2. 分配样本：将每个样本分配给最近的中心点。
3. 更新中心：对每个簇，计算所有成员的均值作为新的中心。
4. 重复：重复步骤2和3，直到聚类结果不再变化。

**（2）K-means算法的收敛条件**

K-means总是单调收敛于一个局部最优解，收敛条件如下：

1. 聚类标签不再发生变化。
2. 所有中心点的位置不再更新（或更新幅度小于某个阈值）。

从数学角度看，K-means等价于最小化目标函数：

其中表示第个簇，是该簇的中心。

**（3）初始化方法分析**

1. 随机初始化：随机从数据中选取K个点作为初始中心。

优点：简单，速度快。

缺点：对初始点敏感，容易陷入局部最优；不稳定，多次运行结构差异可能较大。

2. K-means++初始化：从数据集中随机选一个点作为第一个中心；对于每个剩余点，计算其与现有中心最小距离的平方，作为概率加权；根据这些概率选择下一个中心；重复知道选出K个初始中心。

优点：初始中心分布更合理，减少局部最优问题，提高收敛速度和最终聚类质量。

缺点：相比随机初始化稍慢。

**（4）确定最优聚类数K**

1. 肘部法：绘制K值与总聚类误差（SSE）曲线；选择SSE开始减缓处的K。直观、常用。

2. 轮廓系数：取值范围，越接近1，聚类越合理；对比多个K选择轮廓系数最大的那个。兼顾簇内紧凑型和簇间分离性。

3. 间隙统计量：与随机数据集的聚类结果进行比较；当Gap值最大时，对应的K被认为是最优。理论支持更强，但计算复杂。

二、PCA（主成分分析）在数据降维中起到重要作用。请结合协方差矩阵和特征值分解，说明PCA的数学原理。在图像数据处理中，PCA降维后如何选择合适的主成分数量？

**（1）PCA的数学原理**

目标是通过线性变换，找出数据中最有代表性的方向（主成分），并在保留尽量多信息的同时降低维度。

1. 数据中心化  
   设原始数据矩阵为：  
   其中是样本数，d是特征维度。首先对每一维特征做中心化处理（减去均值）：
2. 构造协方差矩阵  
   协方差矩阵描述了各特征间的线性相关性：  
   它是一个对称正定矩阵。
3. 特征值分解  
   对协方差矩阵进行特征值分解：  
   其中是由特征向量组成的正交矩阵，表示主成分方向；是特征值，对应每个主成分的方差大小（即信息量）
4. 投影变换  
   选取前k个最大特征值对应的特征向量，组成矩阵，则降维后的数据为：  
   此时原始d维数据变为k维，同时最大限度保留了原始数据的方差信息。

**（2）图像数据中主成分数量的选择**

1. 累积方差贡献率  
每个主成分解释的数据方差比例为：

累计贡献率：

选择使累计贡献率超过某个阈值的最小k，即可达到较好压缩效果并保留足够信息。

三、请比较AdaBoost和随机森林两种集成学习方法的核心思想和实现机制。说明AdaBoost中样本权重更新的数学公式，以及随机森林中"随机"体现在哪些方面？两种方法分别如何解决过拟合问题？

**（1）AdaBoost与随机森林核心思想对比**

|  |  |  |
| --- | --- | --- |
| **特点** | **AdaBoost** | **随机森林** |
| **集成方式** | 顺序加法模型 | 并行训练多个决策树 |
| **核心思想** | 关注难以分类的样本，不断调整样本权重 | 通过特征和样本的随机性，构建多样性模型 |
| **弱学习器** | 通常是浅层决策树 | 完整的决策树 |
| **投票机制** | 加权投票 | 多数投票/平均 |

**（2）AdaBoost的样本权重更新公式**

在第t轮训练中：

1. 给定样本权重
2. 训练弱分类器，计算加权错误率：
3. 计算该分类器的权重（可信度）：
4. 更新样本权重：  
   其中，是归一化因子，保证是概率分布；正确分类的样本权重减小，错误分类的样本权重增大。

**（3）随机森林的随机性**

1. 样本随机性：对原始训练集采用有放回抽样，每棵树训练集略有不同。

2. 特征随机性：每个节点分裂时，不是考虑所有特征，而是从特征集中随机选取m个特征寻找最佳分裂。

**（4）如何解决过拟合问题**

1. AdaBoost：逐步聚焦难分类样本，可能对噪声敏感，但通过浅树（低方差）、较强的正则性可缓解。剪枝或控制迭代次数也有效。

2. 随机森林：通过“投票/平均”减少单棵树的高方差。树的独立性由Bagging和特征子集增强，有效避免过拟合。

四、在聚类任务评估中，由于缺乏真实标签，需要使用无监督评估指标。请说明轮廓系数（Silhouette Coefficient）和Davies-Bouldin指数的定义和计算方法，并解释它们如何衡量聚类质量。当这两个指标出现冲突时（一个高一个低），应该如何判断聚类效果？

**（1）轮廓系数**

1. 定义：衡量单个样本的聚类合理性，综合考虑簇内紧密度和簇间分离度。对于每个样本i：

- ：样本i到本簇内其他样本的平均距离（簇内距离）。

- ：样本i到最近的其他簇的平均距离（最近邻簇间距）。

轮廓系数公式：

* ：聚类合理，样本距离本簇近，离其他簇远。
* ：样本位于边界附近。
* ：样本可能被错误分簇。

2.全体轮廓系数

全体数据集的轮廓系数（均值）。越大越好，接近1表示较好聚类质量。

**（2）DB指数**

1. 定义：衡量聚类结果中每个簇的“紧密度”与“分离度”，即簇内方差与簇间距离的比值。

设有k个簇，每个簇的：

* ：簇内样本到簇中心的平均距离。
* ：簇i和簇j之间的距离（两个中心点的欧氏距离）。

对任意两个不同簇ij，定义：

每个簇i对应一个最大值：

最终DB指数为：

越小越好，表示簇之间分离更清晰、簇内更紧凑。

**（3）冲突情况分析**

可能原因：

* 数据分布不均匀/簇形状不规则：轮廓系数更关注样本本地结构，DB更全局
* 某些小簇非常紧密但靠得很近：轮廓系数好，DB由于小而偏高。
* 部分簇内部较松但远离其他簇：DB好，轮廓系数可能较低。

五、在本次实验中涉及了监督学习（AdaBoost、随机森林）和无监督学习（K-means、PCA）方法。请选择其中一种监督学习和一种无监督学习方法，分析它们在处理高维数据时面临的挑战（如维度灾难），并说明如何通过特征选择、特征工程或算法改进来解决这些问题。

**（1）监督学习法：AdaBoost**

1. 面临的挑战：维度灾难

- 噪声特征干扰强：高维数据中可能包含大量无关或冗余特征，容易使弱学习器学到错误模式。

- 计算复杂度增加：随着特征数增加，训练时间和模型复杂度迅速上升。

- 过拟合风险上升：反复拟合残差，高维下容易对噪声点过度学习。

2. 应对策略：

- 特征选择：过滤法剔除无关特征。

- 特征提取：使用PCA或LDA将高维空间压缩为低维特征子空间。

**（2）无监督学习法：PCA**

1. 面临的挑战：

- 计算复杂度高：需要对协方差矩阵进行特征值分解。

- 样本数量不足：高维小样本情况下，协方差矩阵不稳定，主成分方向可能不准确，导致降维效果差。

- 信息损失风险：若特征空间中重要信息在多维上分散，PCA线性降维可能丢失非线性关系中的有用信息。

2. 应对策略：

- 样本预处理：归一化、标准化。

- 采用稀疏PCA或核PCA。

- 选择合理的主成分数量。