**2025年春季学期《机器学习》实验报告**

**（第四次实验）**

班级： 230617 学号： 23371007 姓名： 刘嘉明

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*？

K-means是一种基于距离的迭代聚类算法，目标是将数据划分为K个簇，使得每个数据点属于距离最近的簇中心对应的簇。核心步骤有：初始化、分配步骤、更新步骤、迭代。

收敛条件：质心稳定、簇分配稳定

随机初始化：优点是实现简单，计算成本低；缺点是结果不稳定且可能收敛到局部最优

K-means++初始化：优点是质心分散在数据密集区域，减少局部最优风险；缺点是计算成本高  
 确定最优的聚类数量K：根据肘部法则，计算不同K值对应的簇内平方和，绘制K-WCSS曲线。选择WCSS下降速度显著减缓的“肘点”作为K。

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？

PCA是一种通过线性变换将高维数据投影到低维空间的数据降维方法，其核心是最大化投影方差或最小化重构误差。协方差矩阵反映了数据各维度之间的线性相关性。特征值分解的目的是找到一组正交基，使得数据在这些基上的投影方差最大化，其中特征向量 vi表示第 i 个主成分的方向。

选择主成分数量的原则是需平衡信息保留和维度精简，可以采用累计贡献率方法，计算特征值的累计贡献率，选择最小的k，使得累计贡献率超过阈值。

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？

前者的核心思想是通过迭代训练多个弱分类器，每次调整样本权重，使后续分类器更关注之前分类错误的样本，最终加权投票。解决过拟合的思路是限制弱分类器复杂度，通过验证集控制迭代次数；后者的核心思想是通过构建大量决策树每棵树在随机选择的特征和样本子集上训练，最终通过投票或平均预测结果。解决过拟合问题的思路是通过Bagging减少方差。

AdaBoost初始化时，所有样本的权重是一致的，然后训练一个弱分类器，例如决策树，计算分类的错误和分类器的权重α=1/2ln((1-e)/e)（e为错误率），更新样本权重wi=wi\*e的α次幂，并归一化权重，最终得到最终模型；随机森林主要通过Bootstrap随机采样，从m个随即特征中选择最优分裂，构建多棵树并聚合预测等步骤进行。

随机森林的随机性主要体现在两点：每棵树训练时使用不同的Bootstrap样本子集；每个节点分裂时，仅考虑随机选择的 m 个特征。

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？

轮廓系数的定义是衡量样本与同簇和其他簇的相似性，取值范围为 [−1,1]，值越接近1表示聚类质量越好。对样本 i，计算：a(i)：样本 i 到同簇其他样本的平均距离。b(i)：样本 i 到最近其他簇所有样本的平均距离。样本i的轮廓系数为s(i)=(b(i)−a(i))/(max(a(i),b(i)))，整体轮廓系数则为所有样本s(i)的平均值。当s(i)≈1：样本聚类合理，远离其他簇，s(i)≈0：样本位于簇边界。s(i)≈−1：样本可能被误分到其他簇。

DBI 定义是衡量簇内紧密度与簇间分离度的比值，值越小表示聚类质量越好。对簇 Ck，计算：Sk：簇内所有样本到质心的平均距离。对簇 Ck和 Cl，计算簇间距离 Mkl=∥ck−cl∥（质心距离）。簇 Ck的DBI分量：Rk=max(Sk+Sl/Mkl)(其中l!=k)整体DBI：所有簇 Rk的平均值。衡量标准为DBI越小，簇内越紧密，簇间越远离。

若更关注局部精度：以轮廓系数为主。若更关注全局结构：以DBI为主。

1. 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。

选择随机森林和PCA。

随机森林：随着特征维度增加，数据稀疏性增强，特征间相关性可能降低随机森林的决策质量；高维下特征重要性评估可能不稳定，噪声特征可能被误判为重要特征；每棵树需要评估大量特征分裂点，时间复杂度随特征数线性增长。对此，解决方案有利用随机森林输出的特征重要性排序，保留Top-N特征，在宫颈癌数据中，优先选择临床意义明确的特征。

PCA：PCA仅捕捉线性相关性，可能会忽略非线性结构；高维数据中，少数高方差特征可能主导主成分方向，掩盖重要但低方差特征。主要的解决方法有对每个特征进行StandardScaler处理，避免量纲差异扭曲主成分方向，通过核函数映射到高维空间，捕捉非线性结构，利用拐点法，观察特征值下降拐点确定最佳维度。