**2025年春季学期《机器学习》实验报告**

**（第四次实验）**

班级： 230611 学号： 23373193 姓名： 王天一

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算法基本原理：首先会随机选择K个点作为初始的聚类中心 。然后，它会不断重复两个步骤：1）将每个数据点分配给离它最近的中心；2）更新每个中心的位置，使其成为该组所有点的平均值。当中心点不再移动时，算法就完成了 。

随机选初始中心点虽然简单，但缺点在于可能导致算法收敛到局部最优解，使得最终的聚类效果不佳，且对初始点的选择非常敏感；K-means++则是一种更聪明的初始化方法，的核心思想是让初始聚类中心之间尽可能地分散，具体做法是随机选择第一个中心后，后续每个中心的选择概率与其到现有中心的最短距离的平方成正比。这种方法能够有效改善聚类质量和算法的收敛稳定性，但其计算过程比随机初始化稍显复杂。

要确定最佳的聚类数量K，最常用的方法是“肘部法则”：尝试不同的K值并画出图表，图表拐点（像手肘的地方）对应的K值就是最佳选择 。

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的原理是通过数学计算（协方差矩阵和特征值分解），找到数据中能最大程度解释信息差异的新坐标轴，也就是“主成分” 。这样，它就可以用较少的新特征来代表大部分原始信息，从而达到降维的目的 。

在处理图像时，我们可以通过计算“累计解释方差比”来选择主成分的数量 。比如可以选择保留95%原始信息所需的最少主成分数量，这样既能大幅压缩数据，又能保证图像质量不会损失太多 。

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？

AdaBoost是串行工作的，它会一步步建立模型，每个新模型都会更关注上一个模型分错的样本。随机森林则是并行工作的，它同时建立很多棵独立的决策树 。

AdaBoost中样本权重更新的数学工具为Dt+1​(i)=Dt​(i)exp(−αt​yi​ht​(xi​))​/Zt，其中Dt​(i)是第t轮第i个样本的权重，αt​是第t个学习器的权重， ht​(xi​)是其预测结果，yi​是真实标签，Zt​是归一化因子。

随机森林的“随机”体现在两方面：一是每棵树都从原始数据中随机抽样（Bagging），二是在树的每个节点分裂时，只考虑部分随机选择的特征。

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？

轮廓系数衡量的是一个样本与其所属簇的紧密程度（内聚性）以及与其他簇的分离程度（分离性）。其计算方法为，对每个样本计算其轮廓值 s=(b−a)/max(a,b)，其中 a 是该样本与同簇内其他所有点的平均距离， b 是该样本与最近的非所属簇中所有点的平均距离。

DBI则通过计算每个簇与其最相似簇之间的相似度来评估聚类质量。对于每个簇，它会找到另一个与其最相似（距离最近）的簇，并计算一个比率，该比率的分子是两个簇内数据点的平均离散度之和，分母是两个簇中心点之间的距离。

当两个指标给出的结论矛盾时（比如轮廓系数很高，但DB指数也很高），意味着簇的形状可能不规则或密度不均。最好的办法是把数据降维后画出来看一看，结合视觉和业务需求来判断哪个聚类结果更合理。

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。

对于K-means而言，高维空间中的数据点会变得异常稀疏，导致点与点之间的距离概念变得不那么有意义。随着维度增加，任意两点之间的欧氏距离趋于相等，这使得基于距离的聚类变得困难且不可靠，同时算法的计算复杂度也会显著增加。

在算法改进方面，对于K-means来说，可以在其应用前先通过PCA等降维技术对数据进行预处理。PCA可以将原始高度相关的特征转化为一组线性无关的主成分，从而在保留大部分数据方差的同时减少维度。