数据分类作业报告

姓名：叶子鑫　学号：22373405

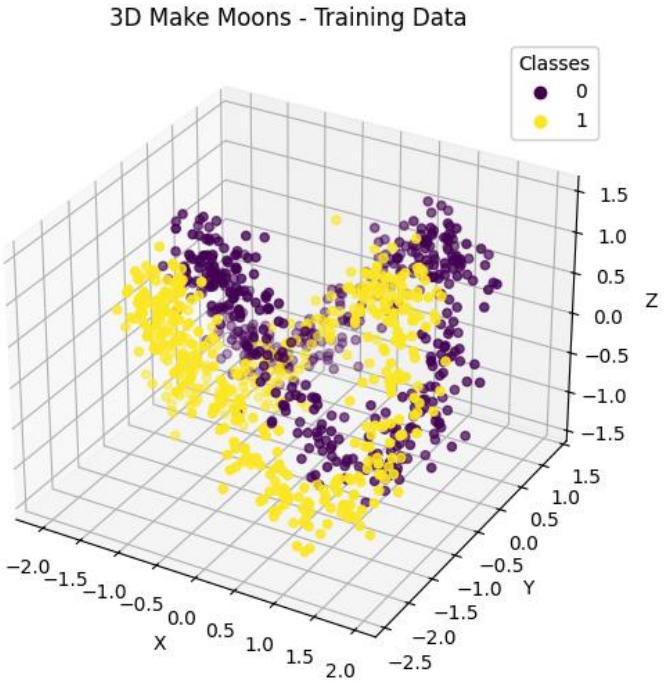
一、摘要

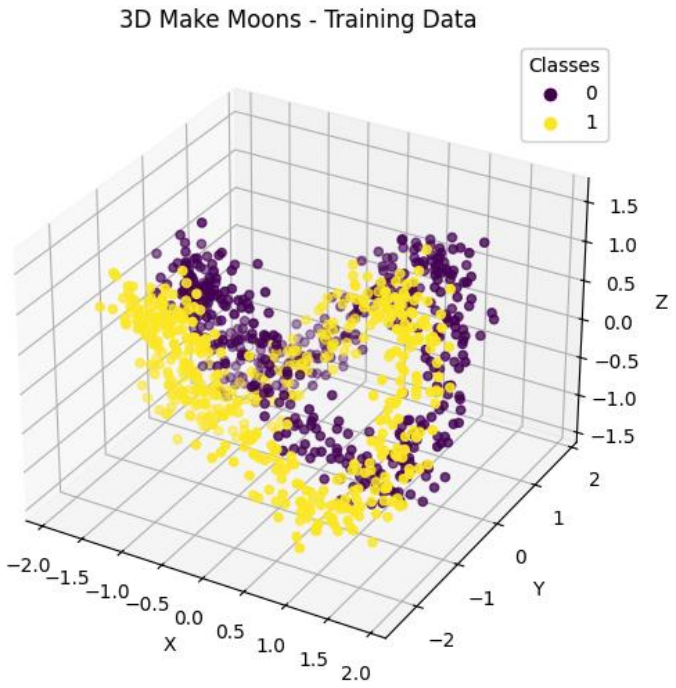
本实验为 PRML 第二次课程作业，任务是对一个非线性三维数据集进行分类建模。我们采用了三种典型的分类模型：决策树（Decision Tree）、AdaBoost 集成学习算法（基分类器为决策树），以及支持向量机（SVM）分别使用三种不同核函数（线性、Polynomial、RBF）进行训练与评估。实验目标是比较不同模型在该数据集上的分类性能，分析其优劣和适用场景。

二、实验数据与可视化

我们使用 make_moons 方法生成了一个 三维双月形数据集，包含两个类别（0 类和 1 类），总共 1000 个训练样本，测试集为新生成的同分布样本，共 500 个（250/250 类别均衡）。

下图是实验数据的三维可视化：





通过图像可见，两类数据呈明显的非线性结构，适合用于测试分类器的非线性建模能力。

三、方法介绍

1. 决策树（Decision Tree）

决策树是一种基于规则的分类模型，通过递归划分特征空间建立一棵分类树。它能够捕捉复杂的特征交互关系，对非线性数据表现良好，训练速度快，易于解释。

2. AdaBoost + 决策树（集成学习）

AdaBoost（Adaptive Boosting）通过迭代方式组合多个弱分类器（通常是浅层决策树），不断强化被错误分类的样本，从而构建强分类器。其对噪声较敏感，模型训练需谨慎调参。

3. 支持向量机（SVM）

SVM 是一种强大的二分类方法，目标是找到最优间隔分隔超平面。其核心在于核函数的选择：

线性核：适合线性可分问题。

多项式核（Poly）：可拟合一定非线性，模型复杂度受多项式次数控制。

RBF 核：通过高维映射处理高度非线性问题，具有良好的鲁棒性。

四、实验结果与评估指标

下表是五个分类模型在测试集上的性能对比：

分类器类型及准确率	Precision 类 0/类 1	Recall 类 0/类 1	F1-Score 类 0/类 1
决策树 0.94	0.93 / 0.95	0.95 / 0.93	0.94 / 0.94
AdaBoost +决策树 0.73	0.76 / 0.71	0.67 / 0.78	0.71 / 0.74
SVM（线性核）0.67	0.66 / 0.67	0.68 / 0.66	0.67 / 0.66
SVM（多项式核）0.84	0.78 / 0.95	0.96 / 0.73	0.86 / 0.82
SVM（RBF 核） 0.97	0.98 / 0.96	0.96 / 0.98	0.97 / 0.97

五、分析与讨论

决策树：在本实验中表现较好（准确率 94%），其能够处理非线性结构，适合初步建模和可视化解释。

AdaBoost：尽管理论上可增强性能，但在本任务中表现不佳（准确率仅 73%），主要原因可能是对训练数据中的噪声敏感，弱分类器能力不足或组合方式不佳。

SVM（线性核）：在该非线性数据上表现最差，说明线性核无法处理复杂边界问题。

SVM（多项式核）：提升了性能（84%），但仍依赖参数（多项式阶数）的优化。

SVM（RBF 核）：表现最佳（97%准确率），RBF 能有效映射复杂结构，具有很强的分类能力和鲁棒性。

六、结论

实验结果表明，在处理三维非线性数据（如本例中的双月形数据）时，SVM + RBF 核函数 是最优选择，能够准确拟合复杂边界并具有出色的泛化能力；决策树作为轻量模型也有良好表现；而 AdaBoost 与线性核的 SVM 因适用性与参数问题表现相对较弱。通过本实验，我们加深了对分类器性能与数据结构之间关系的理解，并体会到核函数和模型选择对分类结果的重大影响。