作业二报告

钟经佑 22373168

摘要

本次作业通过比较决策树、AdaBoost 和支持向量机（SVM）在非线性分类任务中的表现，探讨了不同模型和核函数的效果。首先，决策树被用于基本分类任务，随后通过引入 AdaBoost 集成方法提高了分类性能。对于 SVM，重点分析了不同核函数（线性核、Polynomial 核、Sigmoid 核和 RBF 核）的优缺点，发现 RBF 核在处理复杂数据时表现优越

方法简介

一、Decision Tree

决策树是一种基本的分类和回归方法，基于树状结构来决策。在每个节点上，决策树通过选择特征和阈值将数据划分为子集，直到达到叶节点。其核心思想是递归地根据特征的最佳分割点来划分数据，使得每个分割后的数据集尽可能纯净。常见的决策树算法包括 ID3、CART（Classification and Regression Trees）等，决策树在处理有噪声或复杂数据时可能出现过拟合问题，通常通过剪枝等技术来减轻这一问题。

二、AdaBoost

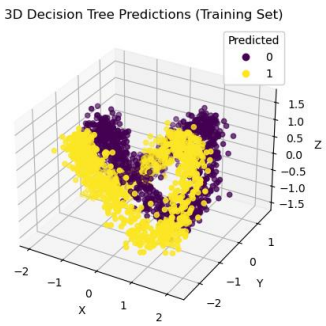
AdaBoost 是一种集成学习方法，旨在通过组合多个弱分类器来提升分类性能。其核心思想是反复训练一系列弱分类器，每个分类器都关注前一轮分类器错误分类的数据点，从而不断调整训练数据的权重，使得分类器能够更好地处理难以分类的样本。最终的分类结果是多个弱分类器加权投票的结果。AdaBoost 特别适用于提高简单模型（如决策树）的性能，并且能够显著减少过拟合风险。

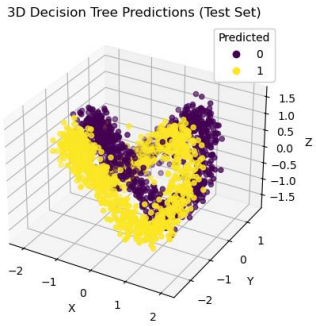
三、SVM

支持向量机（SVM）是一种强大的监督学习算法，用于分类和回归任务。SVM 的目标是找到一个最优的超平面，使得不同类别的数据点尽可能远离超平面。为了应对非线性问题，SVM 采用核函数（如线性核、RBF 核）将数据映射到高维空间，使得在高维空间中数据可以通过线性分割超平面分离。SVM 的优点在于其强大的理论基础和良好的泛化能力，适用于各种复杂的分类任务。通过调节正则化参数（C）和核函数的参数（如 gamma），SVM 可以在多种数据集上表现出色

实验结果

1. Decision Trees



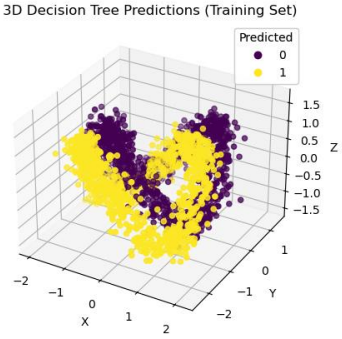


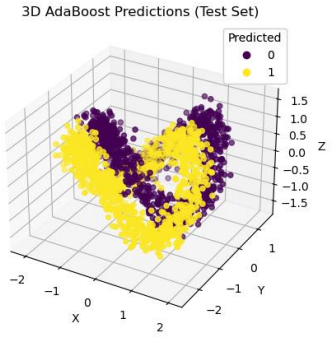
Classification Report (Test Set):

	precision	recall	f1-score	support
0.0	0.95	0.97	0.96	1000
1.0	0.97	0.95	0.96	1000
accuracy			0.96	2000
macro avg	0.96	0.96	0.96	2000
weighted avg	0.96	0.96	0.96	2000

2. AdaBoost+Decision Tree

决策树的最大深度为 3，训练 1000 轮





Training Accuracy: 1.0000

Testing Accuracy: 0.9845

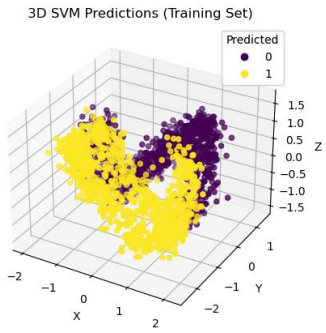
Classification Report (Test Set):

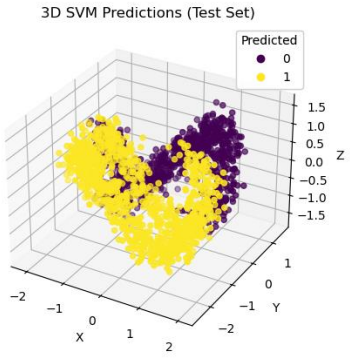
	precision	recall	f1-score	support
0.0	0.98	0.99	0.98	1000

1.0	0.99	0.98	0.98	1000
accuracy			0.98	2000
macro avg	0.98	0.98	0.98	2000
weighted avg	0.98	0.98	0.98	2000

3. SVM

A.使用 Linear kernel





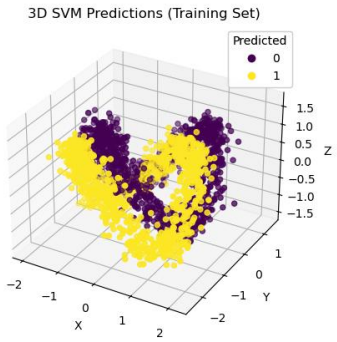
Training Accuracy: 0.6800
Testing Accuracy: 0.6685

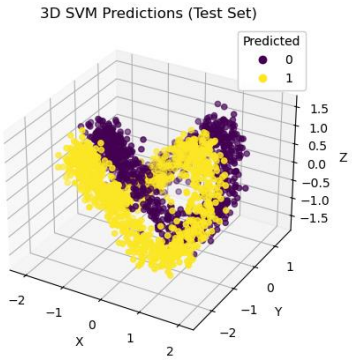
Classification Report (Test Set):

	precision	recall	f1-score	support
0.0	0.67	0.66	0.67	1000
1.0	0.67	0.68	0.67	1000
accuracy			0.67	2000
macro avg	0.67	0.67	0.67	2000
weighted avg	0.67	0.67	0.67	2000

可以看到线性模型对于一个非线性数据表现较差

B. 使用 RBF kernel





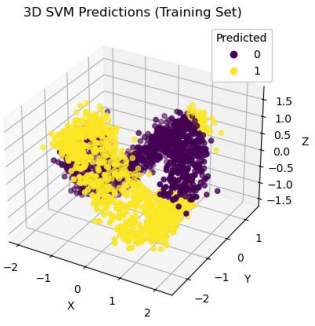
Training Accuracy: 0.9825
Testing Accuracy: 0.9840

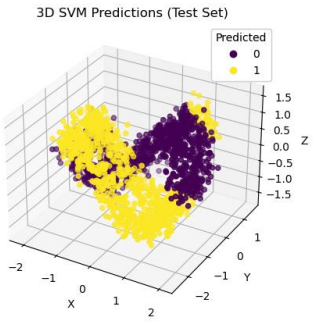
Classification Report (Test Set):

	precision	recall	f1-score	support
0.0	0.98	0.99	0.98	1000
1.0	0.99	0.98	0.98	1000
accuracy			0.98	2000
macro avg	0.98	0.98	0.98	2000
weighted avg	0.98	0.98	0.98	2000

可以看出 RBF kernel 表现较好

C. Sigmoid Kernel





Training Accuracy: 0.4595
Testing Accuracy: 0.4625

Classification Report (Test Set):

	precision	recall	f1-score	support
0.0	0.46	0.47	0.47	1000
1.0	0.46	0.46	0.46	1000
accuracy			0.46	2000
macro avg	0.46	0.46	0.46	2000
weighted avg	0.46	0.46	0.46	2000

Sigmoid Kernel 在此数据集上表现较差

实验结果分析

1. RBF 核的数学表达：

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2\right)$$

这是一个 **高斯核**，也可以看作是：

- 对于每个样本点，RBF 创建了一个**局部感知“影响力”区域**；
- 样本越接近，核值越接近 1；距离越远，核值迅速趋近于 0。

2. 几何和直觉解释

特性	RBF 表现好的原因
局部性	它对距离敏感，只考虑附近的样本，因此能处理复杂的非线性边界
无限维映射	RBF 实际上传递到一个无限维的特征空间，意味着它有极强的表达能力
自动建构复杂边界	无需手动特征工程，RBF 可以自动构造复杂、弯曲的边界
能拟合又能正则化	`gamma` 控制边界复杂度，`C` 控制容错能力，可平衡欠拟合/过拟合

3. 相比其他核的优势

核函数	表现差异解释
Linear	只能处理线性分界，非线性数据下准确率低
Polynomial	对高维不稳定，参数多，容易过拟合或欠拟合
Sigmoid	对参数非常敏感，而且容易陷入梯度消失区域（tanh 饱和）
RBF	稳定、高表达能力、参数少、适用于大多数问题

4. 可控性强（调参好用）

`gamma` 控制 RBF 的“视野范围”：

小 `gamma`：看得很远，边界更平滑；

大 `gamma`：看得很近，边界更精细。

`C` 控制误分类惩罚，和 `gamma` 联合调参几乎能适应各种情况。

结论

首先，我们使用了 决策树 作为基础分类器，并对其在训练集和测试集上的表现进行了评估。决策树算法通过学习数据的特征划分规则，构建了分割超平面。然而，单一的决策树容易出现过拟合问题，因此我们引入了 AdaBoost，一种集成学习方法，来提高模型的准确性和泛化能力。通过对比 AdaBoost 和决策树的单独表现，我们展示了集成方法在复杂任务中的显著优势。

接下来，我们引入了 支持向量机（SVM） 作为强有力的分类工具，重点探讨了不同的 核函数（Kernel Functions） 对分类性能的影响。我们详细分析了 线性核、Polynomial 核、Sigmoid 核和 RBF 核 的优缺点，特别是 RBF 核在处理非线性数据时的突出表现。RBF 核因其强大的非线性映射能力，在大多数情况下表现出了较高的分类准确性和较好的泛化性能。相比之下，Sigmoid 核虽然可以模拟神经网络的激活函数，但因对参数高度敏感，往往导致较差的分类效果。

为了进一步优化 SVM 的性能，我们对 gamma 和 C 等超参数进行了调优。通过网格搜索（Grid Search），我们找到了适合数据集的最佳参数组合，显著提高了分类精度。调参结果表明，RBF 核函数与适当的超参数设置能够在大多数应用中提供优越的分类效果。

实验结果表明，无论是决策树与 AdaBoost 的组合，还是 SVM 的核函数选择与调参，都能在复杂的分类任务中发挥重要作用，尤其是在数据呈现非线性关系时。RBF 核的强大表现使其成为许多机器学习任务中的首选模型。