PRML 第二次作业报告

22374123 朱帅铭

Abstract

本报告围绕三类经典的监督学习算法——**决策树**、**Adaboost + 决策树**、**支持向量机**（**SVM**）展开，旨在通过可视化手段对比不同方法在合成数据集上的分类性能。实验过程中，利用 sklearn 构建并训练各类模型，记录其在测试集上的准确率，并绘制分类结果图像以辅助直观理解。结果表明，SVM在使用RBF核时表现最优，准确率达到0.9880，而 Adaboost + 决策树也表现出良好的泛化能力，有效提升了分类性能。

Introduction

随着机器学习技术的广泛应用，构建高效的分类模型成为模式识别领域的核心问题之一。分类器的性能直接影响模型在实际应用中的可靠性和泛化能力。因此，比较不同分类器在标准数据集上的表现具有重要意义。本文选用三类具有代表性的分类方法：决策树作为经典的符号学习模型，AdaBoost 作为提升弱分类器性能的集成方法，以及支持向量机（SVM）作为基于几何间隔最大化原理的强分类器。通过构建统一的数据集并在多个模型下进行训练与测试，本报告不仅量化了各模型的分类准确率，还通过图像可视化深入观察其决策边界的差异，为进一步模型选择与优化提供参考。

Methodology

在本部分中，我们将详细描述所使用的数据分类的方法。

M1：Decision Trees

决策树是一种利用树状结构对数据进行分类或回归的监督学习方法，其基本思想是在每个节点上选择最优特征来划分数据，从而形成一系列 $if-then$ 规则。

具体来说，对于一个数据集 D（包含 C 个类别），可以通过计算其**熵** $H(D)=-\sum_{i=1}^{C} p_i \log_2(p_i)$ 来衡量数据的不确定性，其中 p_i 表示第 i 类的样本比例。

当利用某一特征 A 将数据集分为若干子集 $D_1, D_2, \ldots, D_k$ 时，其**条件熵**为 $H(D|A)=\sum_{j=1}^{k} \frac{|D_j|}{|D|} H(D_j)$，从而信息增益定义为 $IG(D,A)=H(D)-H(D|A)$，用于衡量划分前后信息的不确定性减少程度。

另一常用准则是**基尼指数**，其对于数据集 D 定义为 $Gini(D)=1-\sum_{i=1}^{C} p_i^2$，分裂后基尼指数为 $Gini(D,A)=\sum_{j=1}^{k} \frac{|D_j|}{|D|} Gini(D_j)$

决策树构造过程中通过不断选择信息增益最大或基尼指数最小的特征进行划分，直到满足停止条件（如节点纯度达到一定程度或数据量低于阈值），最终在叶节点上根据多数投票（分类）或均值（回归）确定预测结果。

M2：Adaboost + Decision Trees

Adaboost是一种**自适应增强算法**，通过结合多个弱分类器（如决策树）来构建一个强分类器，其核心思想在于依次训练一系列弱分类器，并在每一步调整训练样本的权重以便后续分类器更关注之前被错误分类的样本。

初始时，每个样本赋予相等权重 $w_i^{(1)}=\frac{1}{N}$（其中 N 为样本总数），在第 m 次迭代中，训练一个弱分类器 $h_m(x)$ 并计算其加权错误率 $\epsilon_m=\sum_{i=1}^{N} w_i^{(m)} \mathbf{1}\{y_i \neq h_m(x_i)\}$，然后根据错误率计算**分类器权重** $\alpha_m=\frac{1}{2}\ln\left(\frac{1-\epsilon_m}{\epsilon_m}\right)$。

接下来，更新每个样本的权重为 $w_i^{(m+1)}=w_i^{(m)}\exp\left(-\alpha_m y_i h_m(x_i)\right)$，并归一化使得所有权重之和为1。

最终，通过加权投票的方式构造**最终分类器**，即 $H(x)=\text{sign}\left(\sum_{m=1}^{M}\alpha_m h_m(x)\right)$。这种方法使得每一轮都能“聚焦”于难以分类的样本，从而显著提升整体分类性能。

M3：Support Vector Machine (SVM)

支持向量机（SVM）是一种监督学习算法，其核心思想在于寻找一个最优超平面来区分不同类别，使得样本点到该超平面的**间隔**（**Margin**）最大化。

对于线性可分数据，其优化问题可以表述为：最小化目标函数 $\frac{1}{2}\|w\|^2$，同时满足约束条件 $y_i(w^\top x_i+b)\geq 1$ 对于所有样本 i，其中 w 为超平面法向量，b 为偏置；这一过程确保了间隔最大化，从而提升分类器的泛化能力。

对于非线性可分问题，可以引入核函数 $K(x,x')$ 将输入数据映射到更高维空间，在该空间中构造线性超平面进行分类，同时也可以通过引入松弛变量 ξ_i 处理不可分问题，构造软间隔SVM。最终的决策函数形式为 $f(x)=\text{sign}\left(\sum_{i=1}^{N}\alpha_i y_i K(x_i,x)+b\right)$，其中 α_i 是通过拉格朗日对偶问题求解得到的权重系数。这种方法不仅能够处理线性问题，也能通过合适的**核函数**（如多项式核、RBF核等）有效解决非线性分类问题。

Experimental Studies

1. 不同分类器性能对比

Classifier	Accuracy
Decision Trees	0.9360
AdaBoost + Decision Trees	0.9520
SVM (Linear Kernel)	0.6720
SVM (Poly Kernel)	0.8620
SVM (RBF Kernel)	0.9880

2. 可视化结果

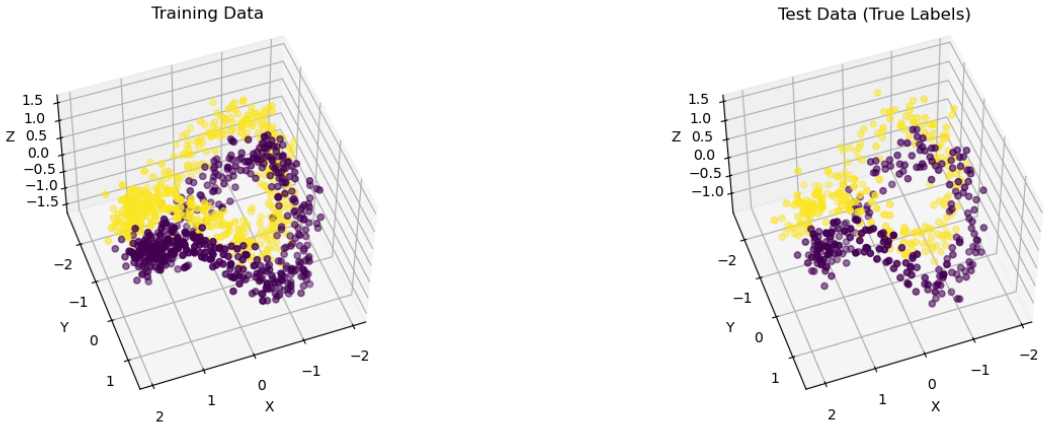

图1. 训练数据以及测试数据

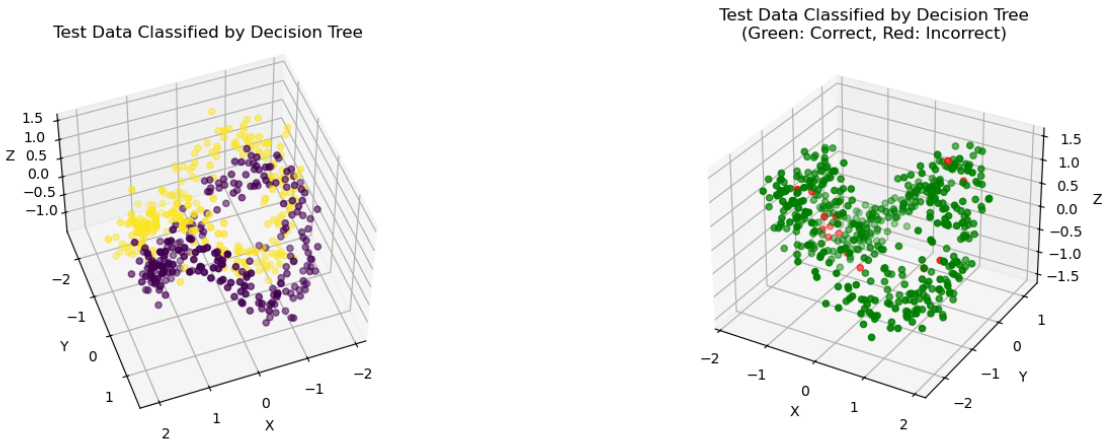

图2. Decision Trees 分类结果

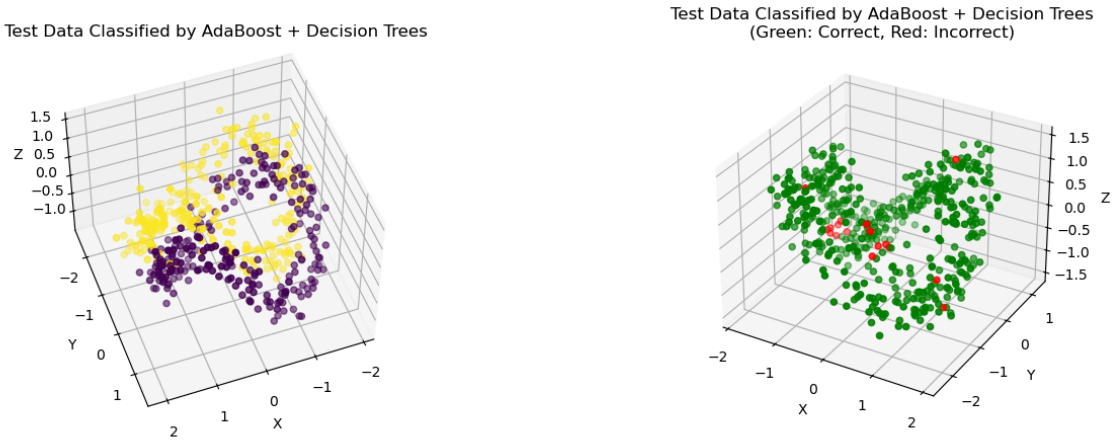

图3. Adaboost + Decision Trees 分类结果

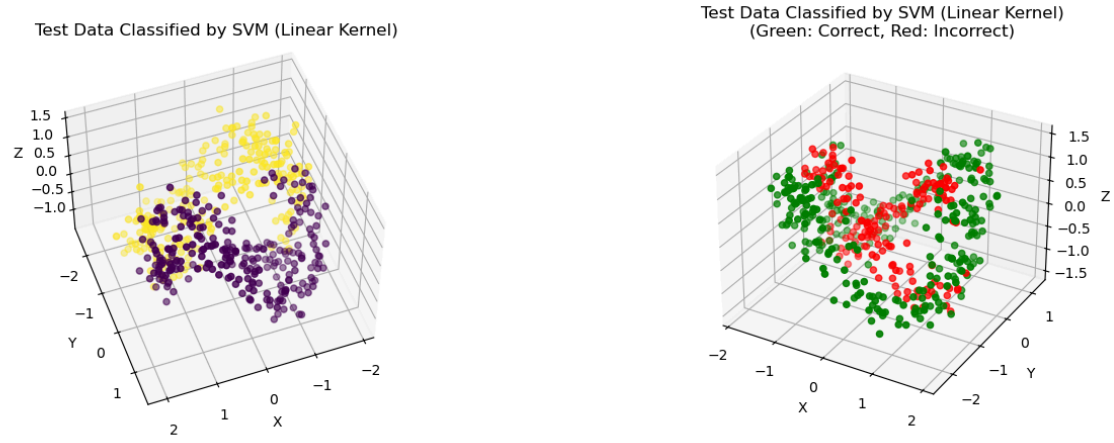

图4. SVM with Linear Kernel 分类结果

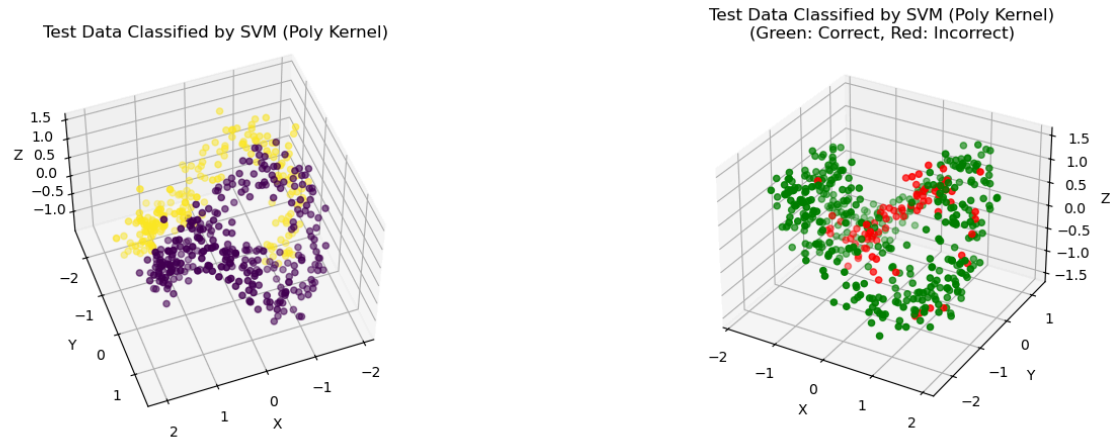

图5. SVM with Poly Kernel 分类结果

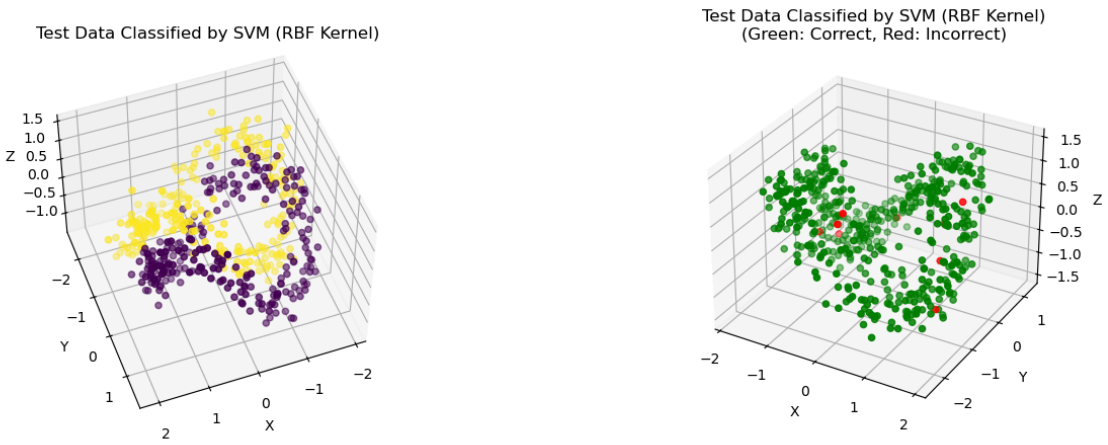

图6. SVM with RBF Kernel 分类结果

Conclusions

1. 实验性能分析

通过对比三类分类器在合成数据集上的表现，可以得出以下结论：

- **SVM（RBF核）**：以0.9880的准确率表现最优，表明其在处理复杂非线性边界时具有显著优势。高斯核通过隐式映射到高维空间，能够灵活适应数据的非线性分布，从而构建更精确的决策边界。
- **Adaboost + 决策树**：准确率为0.9520，通过集成弱分类器有效提升了模型泛化能力。Adaboost 的权重调整机制聚焦于错误分类样本，减少了单棵决策树的过拟合风险。
- **决策树**：准确率为0.9360，虽然简单直观，但在复杂边界场景下表现略逊于集成方法和非线性SVM，可能因单棵树对噪声敏感或划分深度不足。
- **SVM（线性核）**：准确率为0.6720，性能最差，说明数据集具有明显的非线性可分特性，线性超平面难以有效分割类别。

2. 可视化结果的分析

从分类边界可视化图像（图2-6）可进一步观察到：

- 决策树的边界呈现阶梯状，符合其基于轴对齐划分的特点，但对复杂曲面的拟合能力有限；
- Adaboost + 决策树的边界更平滑，说明集成方法通过组合多棵树的预测结果降低了局部噪声的影响；
- SVM（RBF核）的边界高度贴合数据分布，验证了其在非线性问题中的优越性；
- SVM（多项式核）的边界虽有一定复杂度，但可能因阶数选择不当或过拟合导致性能低于RBF核。

3. 模型选择建议

根据实验结果，针对类似合成数据集：

- **优先选择非线性核SVM**（如RBF核），尤其是数据分布复杂时；
- **Adaboost + 决策树**可作为替代方案，平衡性能与计算效率；
- **避免单独使用线性核SVM或浅层决策树**，除非数据明确线性可分。

4. 总结

本实验中，SVM（RBF核）的显著优势源于其通过核技巧隐式构建高维非线性边界的能力，完美契合合成数据复杂的分布模式；而线性核SVM因无法捕捉非线性关系导致性能受限。Adaboost + 决策树的提升效果则来自集成学习机制——通过迭代调整样本权重，强化对难分类样本的关注，弥补了单棵决策树划分粗糙、易受噪声干扰的缺陷。决策树虽直观简单，但其轴对齐的划分方式难以拟合复杂曲面，最终表现略逊。由此可见，**算法的优越性高度依赖于数据内在结构**，非线性方法（如SVM结合RBF核）在处理复杂模式时具有天然优势，而集成策略（如Adaboost）则通过组合弱模型有效提升了鲁棒性与泛化能力。