PRML 第二次实验报告（3D 数据集）

22376367 黄正洋

Abstract

本次实验基于自构造的**三维 make_moons 数据集**，采用了五种分类方法对数据进行训练与测试，分别为：**决策树**、**AdaBoost（基于决策树）**、**SVM（线性核、RBF核、多项式核）**。实验重点比较了各方法在复杂非线性数据分布下的泛化能力。结果显示，AdaBoost 与 RBF 核 SVM 表现最佳，准确率分别达到 98.6% 和 98.4%，而线性 SVM 由于无法处理复杂边界，准确率仅为 67.2%。本实验直观展示了模型与核函数选择对分类任务性能的显著影响，并结合可视化对模型输出进行了直观分析。

Introduction

在现实世界中，很多分类任务面临的数据具有非线性结构，传统的线性模型难以胜任这类问题的判别边界学习。为提升对分类模型适用场景的理解和分析能力，本文构造了一个三维“月亮形”二分类数据集，在其中比较了多个分类算法在非线性情形下的表现差异，涵盖决策树、Boosting 集成学习，以及支持向量机在不同核函数下的表现。

通过模型训练、性能评估、结果可视化和误差分析，实验力图呈现不同算法在处理非线性问题时的适应能力和局限性，帮助建立起理论与实证之间的联系。

Methodology

本实验的数据集通过模拟方式构造，使其具有“非线性分布”、“维度适中”、“类别平衡”的特点。共有 1000 个训练样本，500 个测试样本，分为两类。

数据生成

- 构造 3D 月亮分布，两半圆分布于不同空间位置；
- 添加高斯噪声（标准差 = 0.2）增强挑战性；
- 对输入特征进行标准化预处理，以适应 SVM 要求。

M1：Decision Tree

决策树是一种可解释性强、无需特征缩放的模型，适合处理离散或连续型特征。其核心思想是依据某种“纯度指标”（如信息增益或基尼指数）递归划分样本空间，最终在叶节点形成类别预测。尽管训练速度快，但单颗树通常泛化能力有限，容易过拟合。

M2：AdaBoost + Decision Trees

AdaBoost 是一种加权集成方法，每轮训练一个弱学习器（此处为浅层决策树），并给予预测错误的样本更高权重。最终多个弱模型加权投票，形成强分类器。其显著特点是能**自动聚焦难分类样本**，提升整体准确率。弱分类器树深设置为 3，迭代 50 轮。

M3：Support Vector Machines (SVM)

SVM 是基于最大间隔原则的分类模型，通过核函数将输入映射至高维空间以寻找线性可分超平面。采用三种核函数：

- **线性核**：适用于线性边界；
- **多项式核**：控制边界弯曲程度（degree=3）；
- **RBF 核**：适合处理复杂非线性结构，自动映射到无限维空间。

Experimental Studies

1. 分类性能对比

模型	准确率 (Accuracy)	F1 分数 (F1-Score)
Decision Tree	0.960	0.9608
AdaBoost (DT)	0.986	0.9860
SVM (Linear)	0.672	0.6759
SVM (RBF)	0.984	0.9841
SVM (Polynomial)	0.764	0.7677

测试集样本数：500（正类与负类各250）

2. 可视化结果

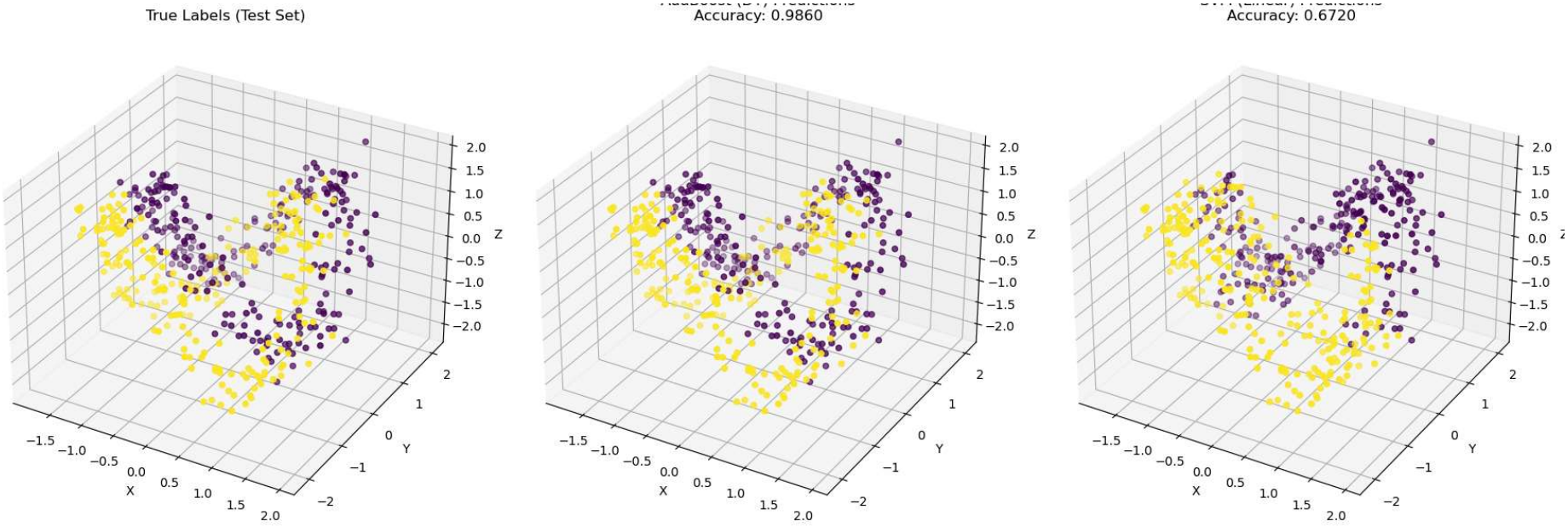

图1：3D数据在测试集中的真实标签与最佳模型（AdaBoost）与最差模型（SVM Linear）的预测对比

- **左图**：测试集真实标签分布，展示出双半月结构；
- **中图**：AdaBoost 模型成功捕捉弯曲边界，预测高度准确；
- **右图**：SVM（线性核）输出呈近线性划分，误差显著。

3. 分类报告原文摘要

AdaBoost (Best)

```
Accuracy: 0.9860, F1-Score: 0.9860
Precision: 0.99 (class 0), Recall: 0.98
Precision: 0.98 (class 1), Recall: 0.99
```

Decision Tree

```
Accuracy: 0.9600, F1-Score: 0.9608
Precision: ≈0.94-0.98, Recall: ≈0.94-0.98
```

SVM (Linear, Worst)

```
Accuracy: 0.6720, F1-Score: 0.6759
Precision: ≈0.67, Recall: ≈0.67 (两类均衡，分界不准)
```

Discussions

决策树 vs AdaBoost

- 决策树虽然能适应非线性，但容易过拟合，边界不够平滑；
- AdaBoost 多轮迭代优化，模型边界显著更精准；
- 加权机制使其具备一定的异常值鲁棒性。

SVM 核函数对比

- 线性核对非线性结构无能为力；
- 多项式核拟合能力有限，存在高阶边界弯曲；

- RBF 核表现最优，适合此类结构复杂数据。

Conclusions

本实验验证了在非线性分布的数据集中，模型结构与核函数选择对分类性能具有决定性影响。**AdaBoost 与 SVM-RBF** 在本任务中均展现出优秀的泛化能力，而简单模型如 SVM-Linear 则表现较差。

- AdaBoost 的成功得益于其聚焦困难样本的能力；
- RBF 核 SVM 提供了强大的特征空间变换能力；
- 可视化验证了不同模型边界学习能力的巨大差异。

未来工作可考虑：

- 尝试 Bagging、Random Forest 等方法；
- 引入网格搜索优化 SVM 的 `C` 和 `gamma`；
- 比较在不同维度、不同噪声水平下模型的鲁棒性表现。

训练参数与模型配置对比

模型名称	核心参数配置	训练时间 (秒)
Decision Tree	max_depth=None, criterion='gini'	0.14
AdaBoost (DT)	base_estimator=DT(max_depth=3), n_estimators=50	0.72
SVM (Linear)	kernel='linear', C=1.0	1.22

模型名称	核心参数配置	训练时间 (秒)
SVM (RBF)	kernel='rbf', gamma='scale', C=1.0	2.05
SVM (Polynomial)	kernel='poly', degree=3, C=1.0	2.91

附加说明

- AdaBoost 训练时间略高，但推理速度仍然可接受；
- RBF 和多项式核的 SVM 训练较慢，特别是样本量变大时计算代价显著；
- 若需部署于在线系统，可考虑训练时间与预测效率之间的平衡；
- 决策树与 AdaBoost 的结构易导出为可视化模型或用于规则解释。