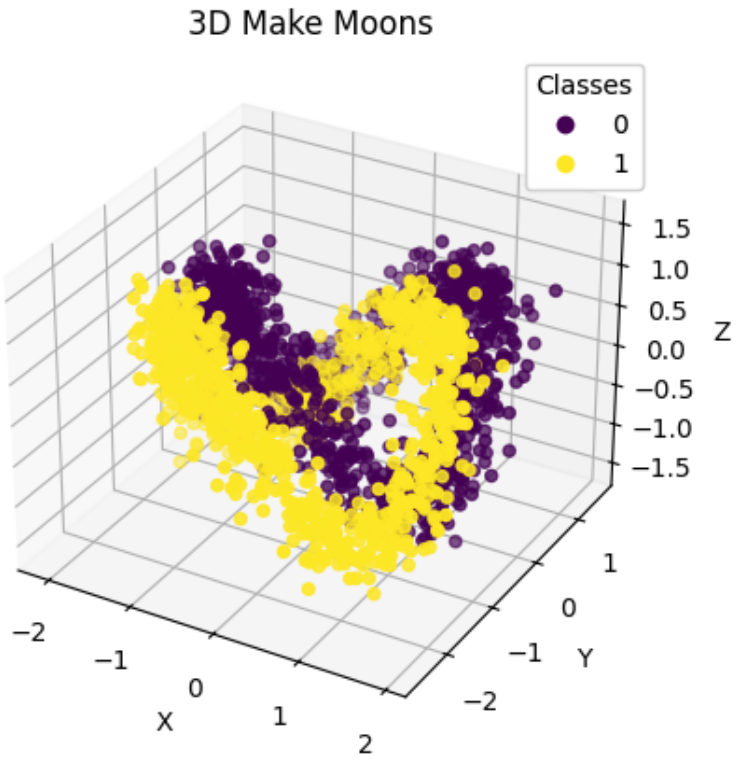
Assignment 2

Abstract

本实验选用了三种分类方法：**决策树（Decision Tree）**、**AdaBoost集成决策树（AdaBoost + Decision Trees）** 以及**支持向量机（SVM）**，其中SVM使用了四种核函数：线性核（linear）、多项式核（poly）、径向基核（rbf）和sigmoid核函数。分别用这些方法在生成的1000个数据上进行训练，并在500个数据的测试集上完成测试，并比较实验结果。

原始生成的训练数据



Introduction

在本实验中，目标是比较多种分类算法在三维非线性数据集上的表现。

利用数据生成函数生成了两个类别的数据（C0 和 C1），每类分别包含 n=500个样本。训练集包含 1000 个样本，测试集为重新生成的 500 个新样本（两类均匀分布）。

实验的目标是训练一个分类器 f(x)，使其能正确预测样本的类别标签。

$$y \in \{0, 1\}$$

相当于在寻找一个判别函数，满足决策规则（主要针对SVM）：

$$f(x) = \text{sign}(\mathbf{w}^T \phi(x) + b)$$

其中可以通过不同的kernel function将输入投影到高维空间。

Methodology

1. 决策树（Decision Tree）

决策树是通过**最大信息增益**划分数据。

信息增益（Information Gain）：

节点熵定义为：

$$H(D) = -\sum_{k=1}^{K} p_k \log_2 p_k$$

其中p表示概率。

对特征A的信息增益为：

$$G(D, A) = H(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} H(D^v)$$

希望最大化信息增益，最大程度增加每一类别中数据的纯度。

2. AdaBoost（Adaptive Boosting）

AdaBoost通过加权组合多个弱分类器来构造强分类器。

弱分类器输出：

第 m 轮的弱分类器，其误差为：

$$\text{err}_t = \frac{\sum_{i=1}^{N} w_i \cdot \mathbb{I}(y_i \neq h_t(x_i))}{\sum_{i=1}^{N} w_i}$$

弱分类器权重：

$$\alpha_m = \frac{1}{2} \ln \left(\frac{1 - \text{err}_t}{\text{err}_t} \right)$$

样本权重更新：

$$w_i^{(m+1)} = w_i^{(m)} \cdot e^{-\alpha_m y_i h_m(x_i)}$$

再进行归一化以保证权重和为1

最终分类结果：

$$H(x) = \text{sign} \left(\sum_{m=1}^{M} \alpha_m h_m(x) \right)$$

3. 支持向量机（SVM）

SVM 的目标是寻找一个超平面，使得分类间隔最大。原始形式的优化目标如下：

线性可分情况：

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{subject to } y_i(\mathbf{w}^T x_i + b) \geq 1$$

对应的决策函数:

$$f(x) = \text{sign}(\mathbf{w}^T x + b)$$

核函数（Kernel Function）：

对于非线性可分数据，引入核函数

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$

用来将输入空间映射到高维空间，使数据线性可分。

使用的核函数：

线性核：

$$K(x, x') = x^T x'$$

多项式核：

$$K(x, x') = (x^T x' + c)^d$$

RBF核（高斯核）：

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

Sigmoid核：

$$K(x, x') = \tanh(\alpha x^T x' + c)$$

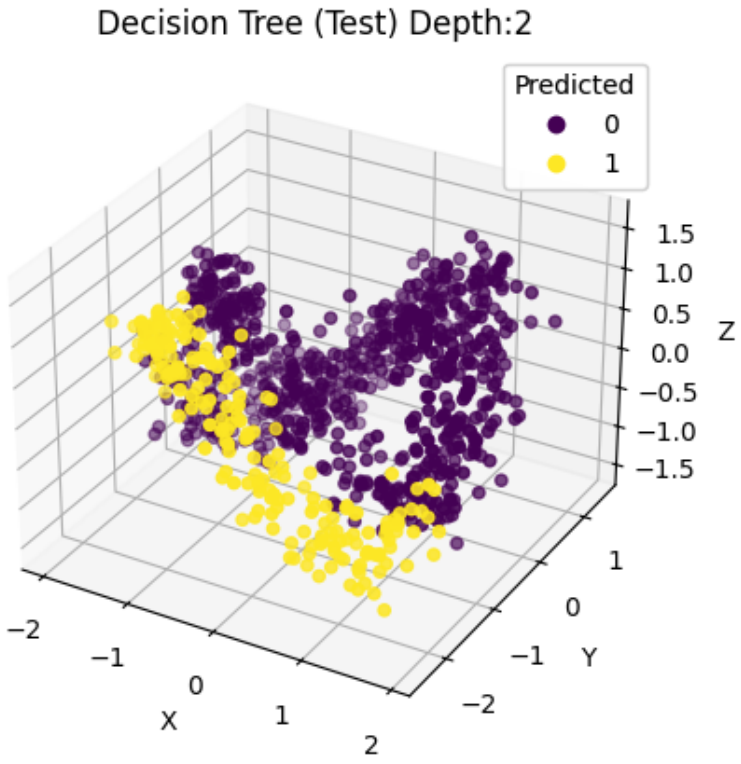
Experimental Studies

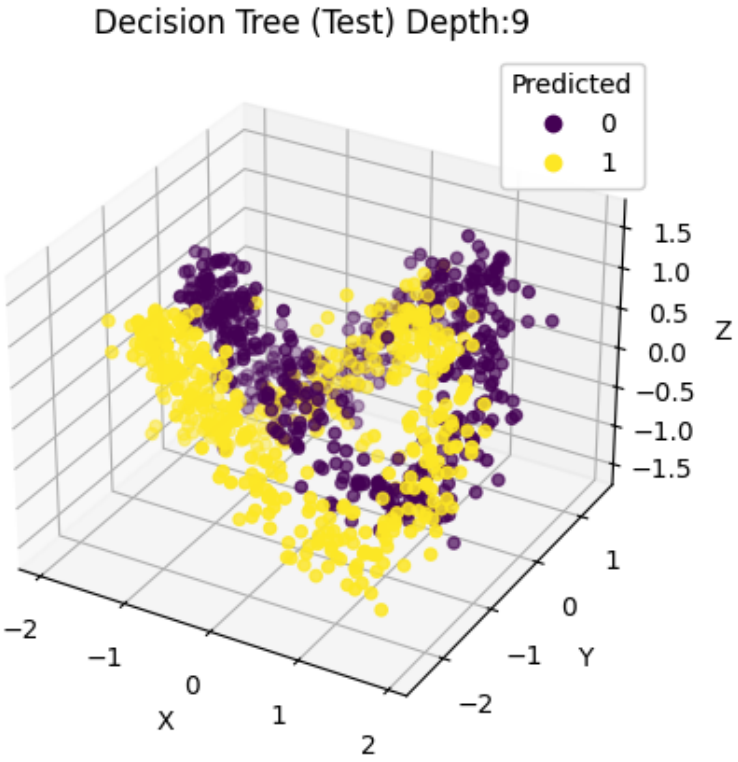
1. 决策树（Decision Tree）

通过设置不同的允许深度测试算法在测试数据上的误差情况。

可以看到在深度为9时，测试误差基本不再下降，后面出现过拟合的情况，前面有欠拟合的情况。因此应当针对数据情况选择合适的决策树最大深度，得到较好的结果。

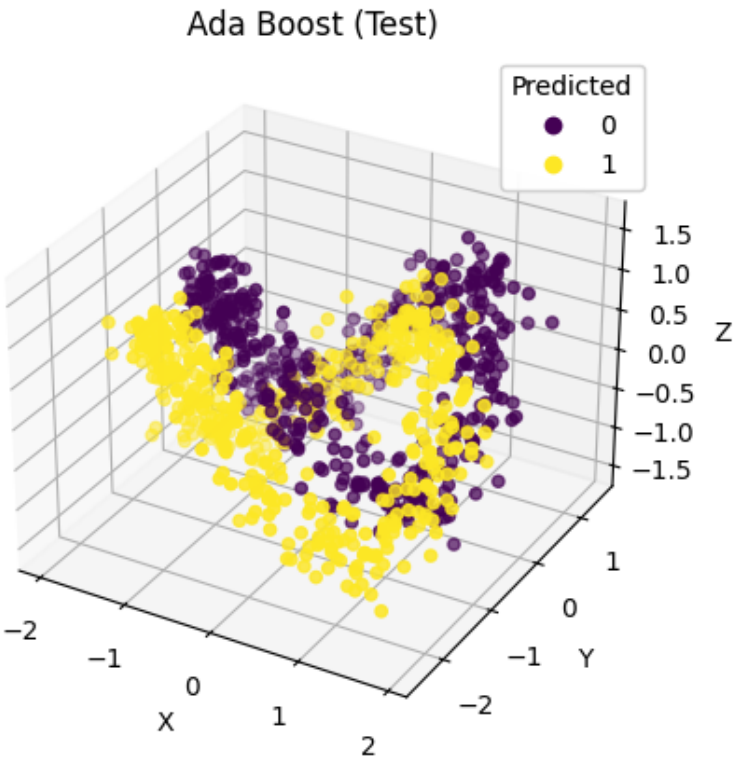
下面是在决策树最大深度为2和9时的测试结果。





2. AdaBoost（Adaptive Boosting）

AdaBoost中设置弱分类器为决策树，最大深度为5，总分类器个数为100，学习率为1。训练结果如下图：

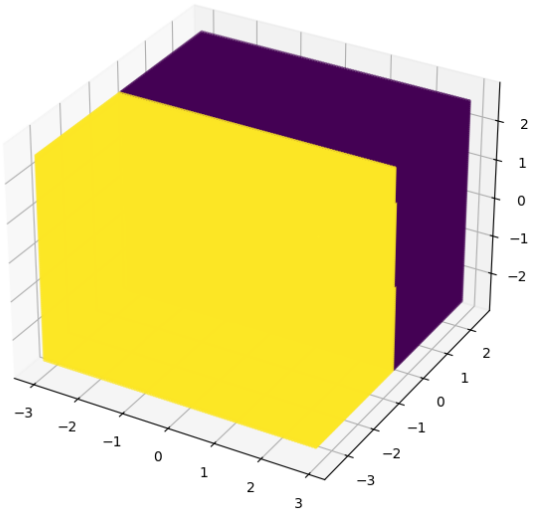


平均正确率在97%-98%左右。

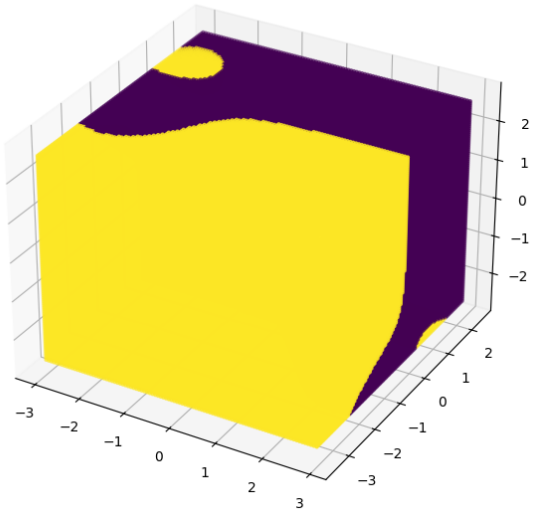
3. 支持向量机（SVM）

四种不同的核函数分类边界和测试结果如下：

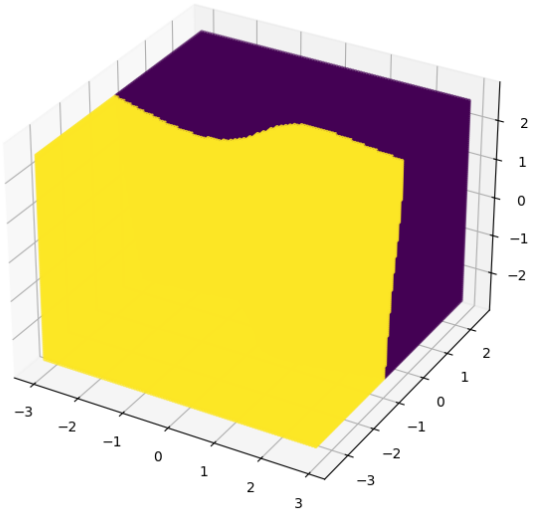
Linear Kernel - Accuracy: 0.67



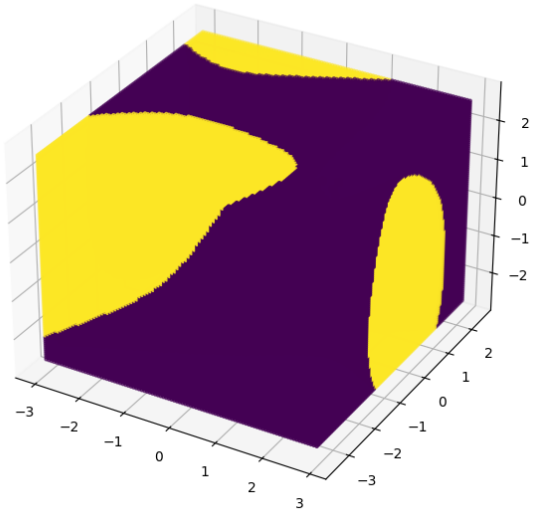
Poly Kernel - Accuracy: 0.87

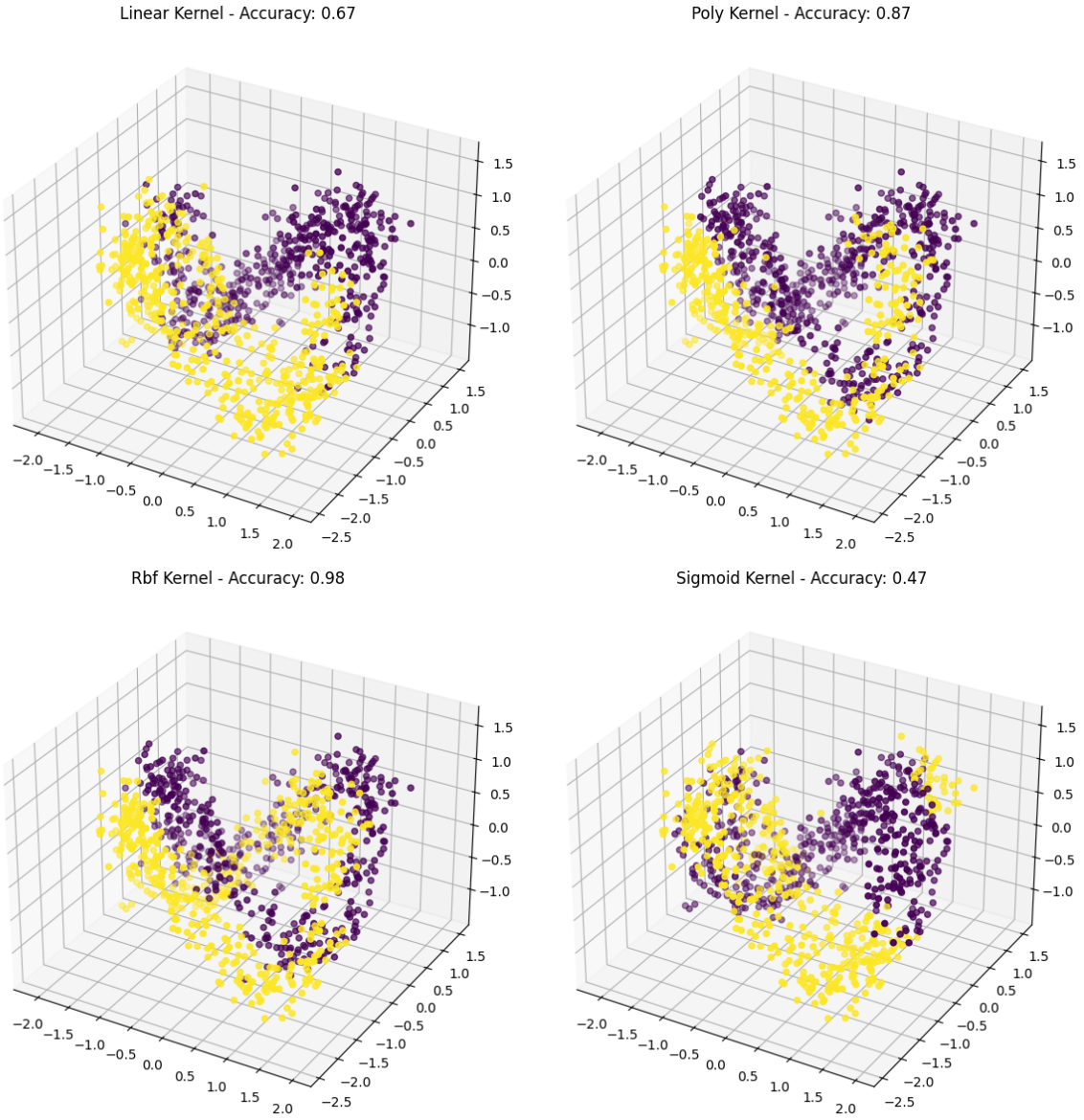


Rbf Kernel - Accuracy: 0.98



Sigmoid Kernel - Accuracy: 0.59





Conclusion

SVM with RBF Kernel 和 AdaBoost + DecisionTree 整体效果最好。RBF核（高斯分布）在处理非线性数据（如本任务中3D弯曲的moons）时适应性强，可以很好的划分非线性边界。AdaBoost能通过组合多个弱分类器（层数小的决策树）纠正错误分类，使得最后在确定结果时每个数据点都有大部分的决策树给出正确结果，最终综合统计结果正确率高。在此非线性任务中，多个决策树统计结果，相较单一决策树更稳定和准确。

SVM with Linear Kernel 和 SVM with Sigmoid Kernel 整体表现效果较差。由于数据是高度非线性分布，线性核无法找到好的分割平面，因此Linear Kernel正确率较低。而Sigmoid Kernel输出结果是(-1, 1) 可能导致核矩阵中的信息变“压缩”，分类边界变得不清晰。同时在高维非线性但结构相对明确的空间中，Sigmoid 很难学出一个有效的边界，这个函数在低维度线性情况下课能表现更好。

SVM with Polynomial Kernel 表现介于Linear 与 RBF 之间。多项式核可以建模一定程度的非线性，但次数过高容易过拟合，次数不足容易欠拟合，效果不如RBF稳定。单个 Decision Tree 深度过深时会出现明显的过拟合，选择合适的深度可以在测试数据上达到更好的效果。