分类算法在 3D-Moon 数据集上的性能比较实验报告

一、摘要

本次作业报告应用 Decision Trees、AdaBoost + DecisionTrees 和 SVM 对 3D 数据集进行分类，对比不同算法的性能差异。通过标准化数据集、可视化分类边界和量化评估指标，验证了 RBF 核 SVM 在非线性可分数据中的最优表现。

二、方法原理

1. **决策树（Decision Trees）**：是一种树形结构的监督学习算法，通过递归划分特征空间实现分类或回归。其核心组成包括：

1) 根节点：代表初始数据集

2) 内部节点：对应特征测试，产生分支决策

3) 叶节点：存储最终预测结果（类别或数值）

基于信息增益递归划分特征空间，作业中设置 max_depth=5 防止过拟合。优势在于直观解释性强，无需数据标准化处理，但是对噪声敏感容易过拟合。

2. **AdaBoost（Adaptive Boosting）**：是一种集成学习算法，通过组合多个弱分类器（如决策树桩）构建强分类器。其核心思想是**迭代调整样本权重**与**分类器权重**，逐步聚焦于难以正确分类的样本，最终通过加权投票提升整体性能。

、核心原理

1. **样本权重初始化**
   初始时所有样本权重相等：$w_i^{(1)} = \frac{1}{N}$
2. **迭代训练弱分类器**
   计算加权错误率：

$$e_m = \sum_{i=1}^{N} w_i^{(m)} \cdot I(y_i \neq G_m(x_i))$$

3. **计算分类器权重**

$$\alpha_m = \frac{1}{2}\ln\left(\frac{1-e_m}{e_m}\right)$$

4. **更新样本权重**

$$w_i^{(m+1)} = \frac{w_i^{(m)}}{Z_m} \cdot \exp\left(-\alpha_m y_i G_m(x_i)\right)$$

5. **加权投票输出**

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

本次作业集成 50 个弱分类器（决策树 max_depth=5），通过动态调整样本权重提升性能。迭代过程中，错误分类样本权重增加，增强模型鲁棒性。

3. **支持向量机（SVM）**：是一种监督学习算法，主要用于分类和回归任务，核心思想是寻找一个最优超平面，将不同类别的数据分开，同时最大化两类数据到超平面的最小距离（称为间隔）。其关键特点包括：

1） 间隔最大化：通过几何间隔定义分类边界，提升泛化能力。

2） 核技巧（Kernel Trick）：将数据映射到高维空间，解决非线性可分问题。

3) 支持向量：超平面的位置仅由距离最近的样本点（即支持向量）决定。

核心原理

1. **线性可分情况**
   - **目标函数**：找到超平面 $\mathbf{w}^T\mathbf{x} + b = 0$，使得间隔最大。
   - **数学表达**：

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$
$$\text{s.t. } y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \quad (i = 1, 2, ..., n)$$

   - **解的形式**：通过拉格朗日对偶问题求解，最终超平面仅依赖支持向量。
2. **非线性可分与核函数**
   - **核函数作用**：将原始特征映射到高维空间，使数据线性可分。
   - **常用核函数**：
     - **线性核**：$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T\mathbf{x}_j$
     - **多项式核**：$K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma\mathbf{x}_i^T\mathbf{x}_j + r)^d$
     - **RBF（高斯）核**：$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2)$
     - **Sigmoid核**：$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma\mathbf{x}_i^T\mathbf{x}_j + r)$
3. **软间隔与松弛变量**
   - **问题**：数据存在噪声或轻微重叠时，严格线性不可分。
   - **解决方案**：引入松弛变量 ξ_i 和惩罚参数 C：

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$
$$\text{s.t. } y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

   - **参数 C**：控制误分类惩罚强度。C 越大，模型越严格（可能过拟合）；C 越小，允许更多误分类（可能欠拟合）。

三、实验结果

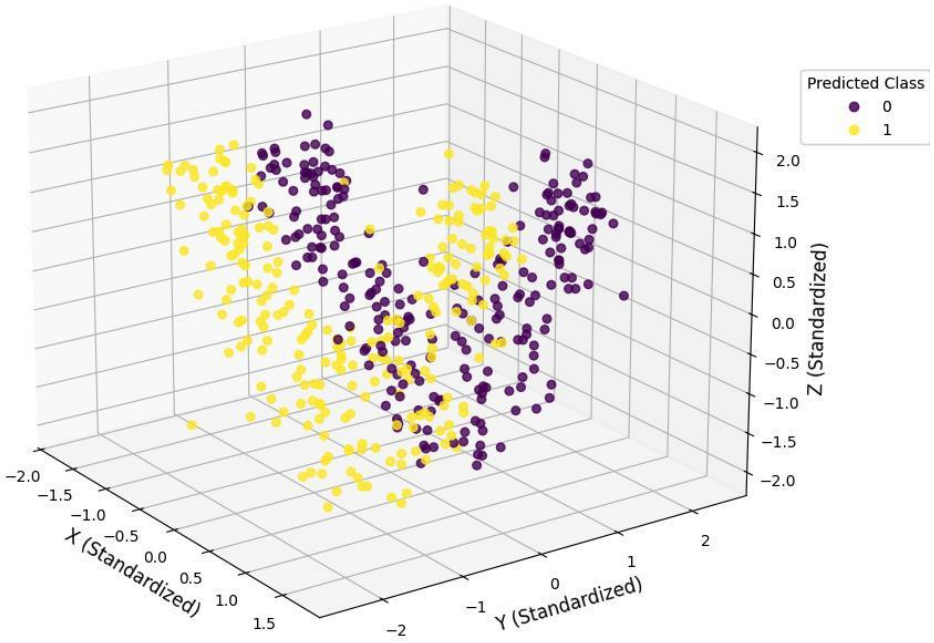
运行程序得到如下结果：

```
----------------------------------------
Model: Decision Tree
Accuracy: 0.9520
              precision    recall  f1-score   support

         0.0       0.95      0.95      0.95       250
         1.0       0.95      0.95      0.95       250

    accuracy                           0.95       500
   macro avg       0.95      0.95      0.95       500
weighted avg       0.95      0.95      0.95       500


======================================
Model: AdaBoost + Decision Tree
Accuracy: 0.9720
              precision    recall  f1-score   support

         0.0       0.97      0.98      0.97       250
         1.0       0.98      0.97      0.97       250

    accuracy                           0.97       500
   macro avg       0.97      0.97      0.97       500
weighted avg       0.97      0.97      0.97       500
```

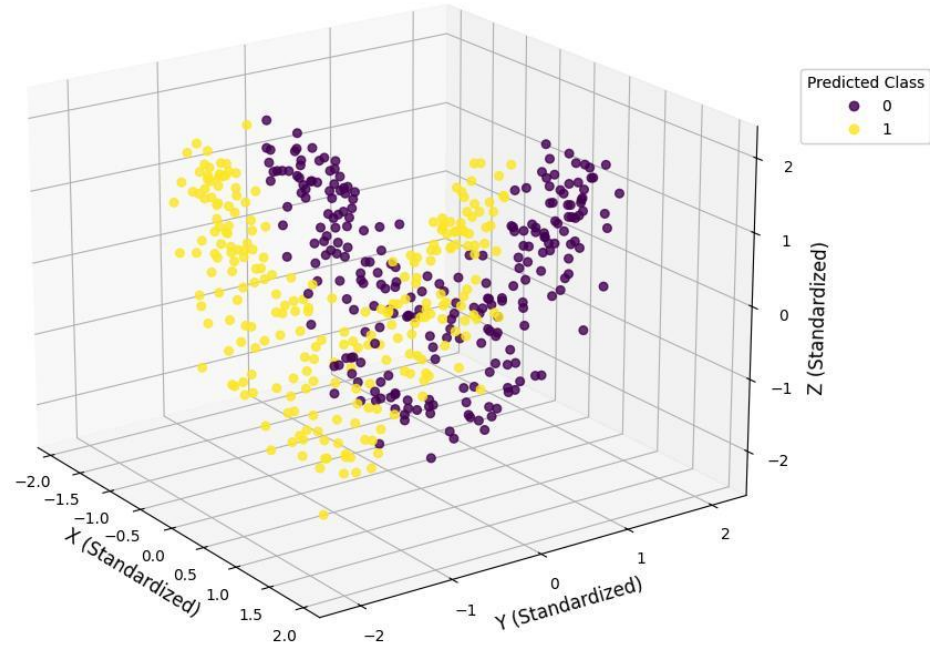
```
----------------------------------------
Model: SVM Linear
Accuracy: 0.6740
              precision    recall  f1-score   support

         0.0       0.67      0.69      0.68       250
         1.0       0.68      0.66      0.67       250

    accuracy                           0.67       500
   macro avg       0.67      0.67      0.67       500
weighted avg       0.67      0.67      0.67       500

----------------------------------
Model: SVM Poly
Accuracy: 0.7520
              precision    recall  f1-score   support

         0.0       0.76      0.74      0.75       250
         1.0       0.75      0.76      0.75       250

    accuracy                           0.75       500
   macro avg       0.75      0.75      0.75       500
weighted avg       0.75      0.75      0.75       500

----------------------------------
```

```
Model: SVM RBF
Accuracy: 0.9880
              precision    recall  f1-score   support

         0.0       0.98      0.99      0.99       250
         1.0       0.99      0.98      0.99       250

    accuracy                           0.99       500
   macro avg       0.99      0.99      0.99       500
weighted avg       0.99      0.99      0.99       500
```

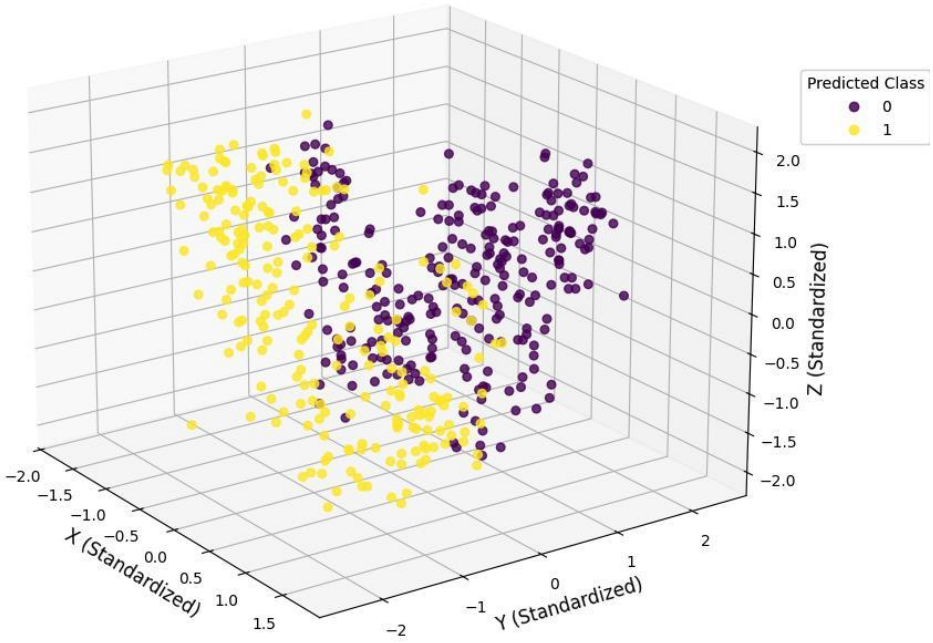
Classification by Decision Tree
(Test Set Results)



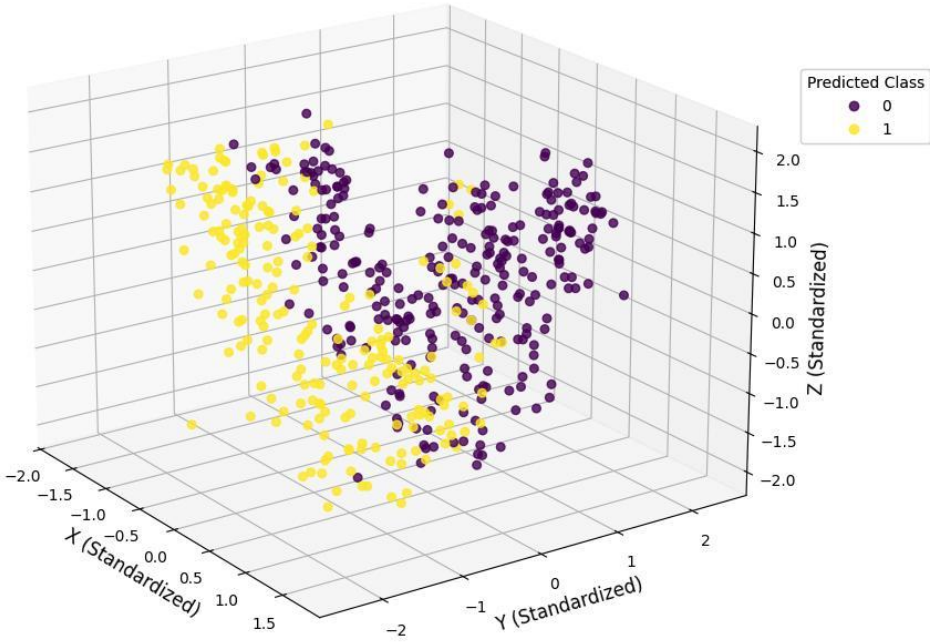
Classification by AdaBoost + Decision Tree
(Test Set Results)

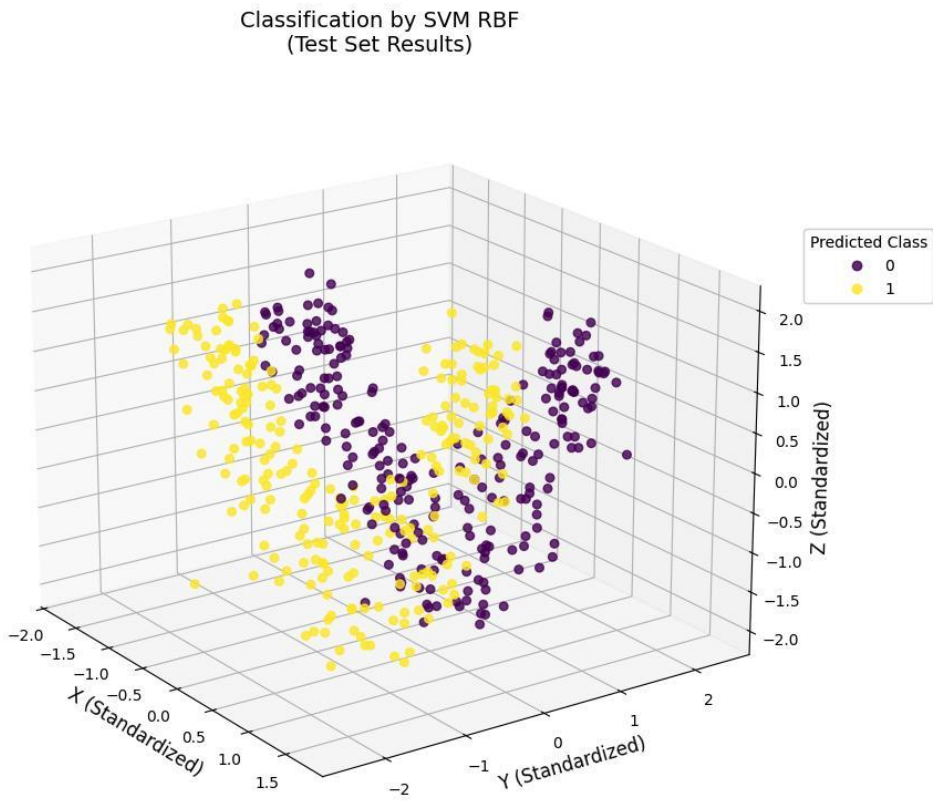


Classification by SVM Linear
(Test Set Results)



Classification by SVM Poly
(Test Set Results)





四、性能差异分析

1. 决策树：

表现：准确率较高（95.2%）。由于数据中存在噪声，决策树容易过拟合训练数据中的局部模式，导致泛化能力下降。

原因：尽管数据在 z 轴上有明显的分界（正类 z 值多为正，负类多为负），但噪声导致决策树需要复杂的分支，增加了错误分割的风险。

2. AdaBoost + 决策树：

表现：准确率显著提升至 97.2%。通过集成多个弱分类器（决策树桩），AdaBoost 减少了方差，提升了鲁棒性。

原因：弱分类器的组合能够有效捕捉数据的全局结构，避免单棵决策树对噪声的敏感。

3. SVM：

线性核：准确率 67.4%。由于数据在 3D 空间中接近线性可分（z 轴提供关键区分），线性核能部分分离类别，但噪声和非线性残余结构限制了性能。

多项式核：准确率 75.2%。3 次多项式核捕捉了部分非线性关系，但参数未调优（如 degree 可能不匹配真实数据分布），表现中等。

RBF 核：准确率最高（98.8%）。RBF 核通过非线性映射处理复杂边界，完美适应数据的螺旋结构，噪声容忍度高。

五、结论

1. RBF 核 SVM 表现最优，因其能建模复杂的非线性决策边界。
2. AdaBoost 通过集成提升了决策树的泛化能力，表现比决策树更优。
3. 线性 SVM 和多项式核 SVM 受限于模型假设，未能充分捕捉数据结构，表现较差。