分类问题报告

摘要

本次作业报告是对课上所学的决策树、集成学习和核方法知识的应用，目标是使用 Decision Trees、AdaBoost + DecisionTrees 和 SVM 对 3D 数据集进行分类，比较不同算法的性能，并分析原因。

方法简介

一、Decision Trees

决策树是基于树形结构的监督学习算法，用于分类或回归任务。它通过递归分裂数据集形成决策节点和叶节点。优点是直观易懂、预处理要求低，但容易过拟合。实验中，设置 max_depth=5，random_state=42。

二、AdaBoost

AdaBoost 是集成学习算法，通过组合多个弱学习器（如浅层决策树）构建强学习器。优点是鲁棒性强，但对异常值敏感且训练较慢。在此实验中，令决策树 max_depth=3，random_state=42。算法如下图所示：

1. **初始化样本权重**:
   对训练集中的每个样本赋予初始权重 $w_i = \frac{1}{N}$（N 为样本数）。
2. **迭代训练弱分类器**（共 T 轮）：
   - **步骤1**：用当前样本权重训练弱分类器 $h_t(x)$。
   - **步骤2**：计算分类错误率 $\epsilon_t = \sum_{i=1}^{N} w_i \cdot I(y_i \neq h_t(x_i))$。
   - **步骤3**：计算弱分类器权重 $\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$。
   - **步骤4**：更新样本权重：

     $$w_i \leftarrow w_i \cdot e^{-\alpha_t y_i h_t(x_i)}$$

     并重新归一化（使 $\sum w_i = 1$）。
3. **组合强分类器**:
   最终模型为所有弱分类器的加权投票：

   $$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

三、SVM

SVM 通过最大化间隔超平面进行分类，适合高维数据。核函数可处理非线性问题，但计算复杂且参数调整困难。对于非线性可分的数据，SVM 通过核函数将

原始特征空间映射到高维特征空间，在高维空间中寻找线性超平面。常用核函数有：

线性核：$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$

多项式核：$\kappa(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + c)^d$

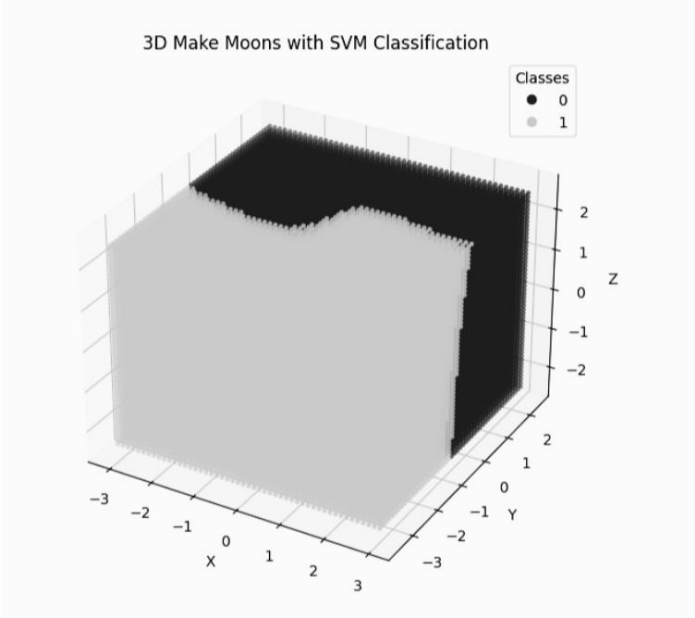
高斯核（RBF 核）：$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$

本次作业使用 Decision Trees、AdaBoost + DecisionTrees 和 SVM 三种方法对三维数据进行分类，并比较它们的分类效果。考虑到数据是三维的，预计采用合适核函数的 SVM 方法可能会有最好的表现。

实验结果

实验结果如下图：

```
决策树分类器的准确率: 0.9590
AdaBoost+决策树分类器的准确率: 0.9754
SVM(线性核)分类器的准确率: 0.6756
SVM(RBF核)分类器的准确率: 0.9809
SVM(多项式核)分类器的准确率: 0.8630
SVM(sigmoid核)分类器的准确率: 0.5814
```



SVM 分类结果示意图

结果分析：

在本实验中，针对具有 3D“月亮”形状分布的非线性数据集（其边界复杂且难以划分），各类算法的表现呈现显著差异。其中，RBF 核 SVM 以最高分类准确率脱颖而出，验证了其通过高维映射处理非线性数据、捕捉复杂边界的核心优势；AdaBoost+决策树紧随其后，得益于集成学习框架下样本权重动态调整与决策树局部拟合能力的协同作用，既强化了非线性边界的识别精度，又有效缓解了过拟合风险；决策树单独应用时虽位列第三，但其基于特征分割的递归建模机制天然适配非线性数据，尽管高维场景下易陷入过拟合，仍通过多级划分实现了对“月亮”形状的局部拟合；多项式核 SVM 因核函数复杂度限制（多项式次数选择对性能敏感），对复杂非线性边界的拟合能力弱于 RBF 核，故排名居中；而线性核 SVM 与 Sigmoid 核 SVM 表现垫底，前者受限于线性假设无法处理非线性分布，后者则因高维映射中的数值不稳定性和参数敏感性，导致性能显著下降。这一结果充分印证了核函数选择与算法适配性对非线性数据建模的关键影响。

总结：

1. SVM（RBF 核）准确率最高，因为 RBF 核能处理非线性边界。
2. AdaBoost 通过集成降低过拟合，表现第二。
3. 决策树直接拟合非线性数据，但容易过拟合。
4. 线性核和 Sigmoid 核不适合复杂数据分布。

结论

在本次利用在决策树、集成学习和核方法部分学习到的三种算法 Decision Trees, AdaBoost + DecisionTrees 和 SVM 来解决对一个 3D“月亮”形数据进行分类的问题过程中，我对于这三种算法的特点有了更深的理解，且认识到了对于不同分布的数据集应当如何选择合适的分类算法。