数据训练分类报告

邓子凡 22231042

摘要

本报告将对一个 3D 的数据集进行分类训练，数据集一共有 1000 个数据，被分成两大类：C0 和 C1.分别用 Decision Trees, AdaBoost + DecisionTrees, 与 SVM 的方法进行训练，并用程序新生成与训练数据同分布的 500 个数据（250 个为 C0 类，250 个数据为 C1 类）来做测试，检测不同算法的分类性能。

方法论

方法 1：决策树（Decision Trees）

通过递归分割数据空间构建树状结构，每个节点代表一个特征的条件判断，叶子节点表示分类结果。

使用基尼不纯度或信息增益作为分割标准，选择使不纯度下降最多的特征进行划分。

对于分类问题，可以选用信息熵作为分类依据，其具体数学表达式为

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

其中，Pk 是数据集中第 k 类样本的比例，熵值越大表示数据越混乱。同时定义信息增益：

$$Gain(D,a) = H(D) - H(D^*)$$

划分前的信息熵减去划分后的信息熵，信息增益越大，则选取该分裂规则。

对于分类和回归问题，一般使用基尼不纯度作为分类依据：

$$Gini(D) = 1 - \sum_{k=1}^{K} p_k^2$$

基尼值越小表示节点纯度越高，比较分裂前的基尼值和分裂后的基尼值减少多少，减少的越多，则选取该分裂规则。

选择特征 A 和分割点 s 使基尼不纯度下降最大：

$$\Delta Gini = Gini(D) - \left(\frac{|D_{left}|}{|D|} Gini(D_{left}) + \frac{|D_{right}|}{|D|} Gini(D_{right}) \right)$$

其中 D_{left} 和 D_{right} 是根据特征 A 的阈值 s 分割后的子集。

方法 2：AdaBoost（Adaptive Boosting）

Adaboost 的核心思路是串行训练多个弱分类器，每个分类器聚焦前序分类错误的样本。

通过调整样本权重和模型权重，逐步改进整体模型。

首先初始化数据权重：

$$w_i^{(1)} = \frac{1}{N}, \quad i = 1, 2, ..., N$$

在第 t 轮迭代时训练弱分类器 ht 并计算加权错误率：

$$\epsilon_t = \sum_{i=1}^{N} w_i^{(t)} \cdot I(y_i \neq h_t(x_i))$$

计算分类器权重：

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \epsilon_t}{\epsilon_t} \right)$$

更新样本权重：

$$w_i^{(t+1)} = w_i^{(t)} \cdot \begin{cases} e^{-\alpha_t}, & \text{若分类正确} \\ e^{\alpha_t}, & \text{若分类错误} \end{cases}$$

后归一化：

$$w_i^{(t+1)} \leftarrow \frac{w_i^{(t+1)}}{\sum w_i^{(t+1)}}$$

最终得到数学模型为：

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

AdaBoost 的损失函数等价于指数损失：

$$L(y, H(x)) = e^{-yH(x)}$$

通过梯度下降法在函数空间中对指数损失进行优化。

方法 3：AdaBoost+决策树

基本的组合原理为使用决策树作为 AdaBoost 的基学习器。浅层决策树（弱学习器）通过 AdaBoost 集成，提升整体表现。

实现方法为弱分类器选择：限制决策树深度（如 max_depth=5），强制成为弱学习器。

权重敏感训练：决策树在拟合时会考虑样本权重 wi，通过加权基尼不纯度进行分裂：

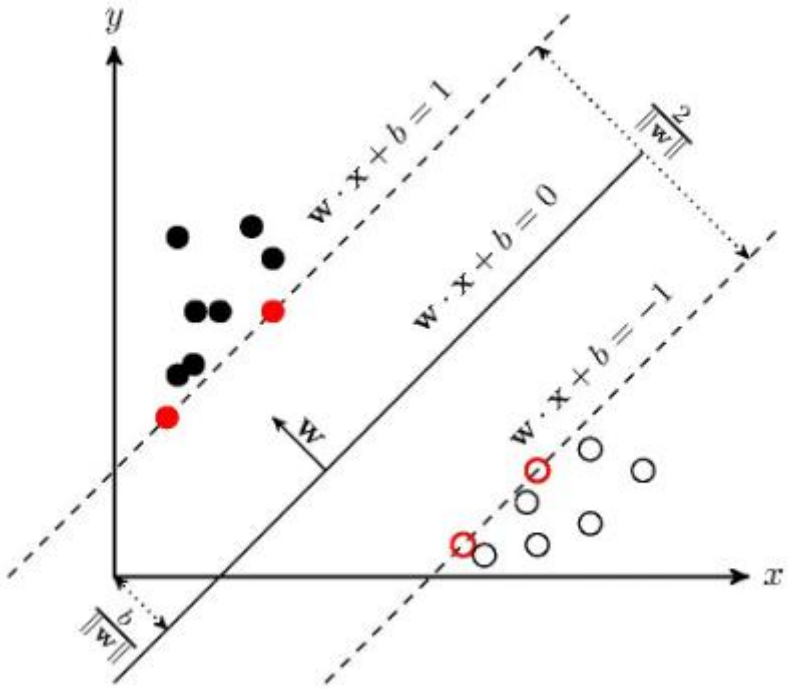
$$Gini_{weighted}(D) = 1 - \sum_{k=1}^{K} \left(\frac{\sum_{i \in C_k} w_i}{\sum_i w_i} \right)^2$$

方法 4：SVM（Support Vector Machine）

支持向量机（Support Vector Machine，SVM）是一种经典的监督学习算法，用于解决二分类和多分类问题。其核心思想是通过在特征空间中找到一个最优的超平面来进行分类，并且间隔最大。

SVM 能够执行线性或非线性分类、回归，甚至是异常值检测任务。它是机器学习领域最受欢迎的模型之一。SVM 特别适用于中小型复杂数据集的分类。

SVM 学习的基本想法是求解能够正确划分训练数据集并且几何间隔最大的分离超平面。如下图所示，



$\boldsymbol{w}\cdot x+b=0$即为分离超平面，对于线性可分的数据集来说，这样的超平面有无穷多个（即感知机），但是几何间隔最大的分离超平面却是唯一的。

SVM 模型的求解最大分割超平面问题可以表示为以下约束最优化问题

$$\min_{w,b}\frac{1}{2}\|\boldsymbol{w}\|^2$$

$$s.t.y_i\left(\boldsymbol{w}\cdot\boldsymbol{x_i}+b\right)\geq 1,i=1,2,\ldots,N$$

这是一个含有不等式约束的凸二次规划问题，可以对其使用拉格朗日乘子法得到其对偶问题（dual problem）。

我们将有约束的原始目标函数转换为无约束的新构造的拉格朗日目标函数

$$L\left(w,b,\alpha\right)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{N}\alpha_i\left(y_i\left(w\cdot x_i+b\right)-1\right)$$

其中α_i为拉格朗日乘子，且$\alpha_i>0$。最终上述优化问题转化为:

$$\max_{\alpha} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \left(\boldsymbol{x}_{\boldsymbol{i}}\cdot\boldsymbol{x}_{\boldsymbol{j}}\right)+\sum_{i=1}^{N}\alpha_i$$

$$s.t.\sum_{i=1}^{N}\alpha_i y_i = 0$$

$$\alpha_i \geq 0, i = 1, 2, \ldots, N$$

将内积替换为核函数$K(x_i, x_j)$，是将原始输入空间映射到新的特征空间，从而，使得原本线性不可分的样本可能在核空间可分。常见核函数有：

名称	表达式	参数
线性核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j$	
多项式核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j)^d$	$d \geqslant 1$ 为多项式的次数
高斯核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2\sigma^2}\right)$	$\sigma > 0$ 为高斯核的带宽(width)
拉普拉斯核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|}{\sigma}\right)$	$\sigma > 0$
Sigmoid 核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(\beta\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j + \theta)$	tanh 为双曲正切函数, $\beta > 0, \theta < 0$

线性核和多项式核：

1）这两种核的作用也是首先在属性空间中找到一些点，把这些点当做 base，核函数的作用就是找与该点距离和角度满足某种关系的样本点。

2）样本点与该点的夹角近乎垂直时，两个样本的欧式长度必须非常长才能保证满足线性核函数大于 0；而当样本点与 base 点的方向相同时，长度就不必很长；而当方向相反时，核函数值就是负的，被判为反类。即，它在空间上划分出一个梭形，按照梭形来进行正反类划分。

RBF 核：

1）高斯核函数就是在属性空间中找到一些点，这些点可以是也可以不是样本点，把这些点当做 base，以这些 base 为圆心向外扩展，扩展半径即为带宽，即可划分数据。

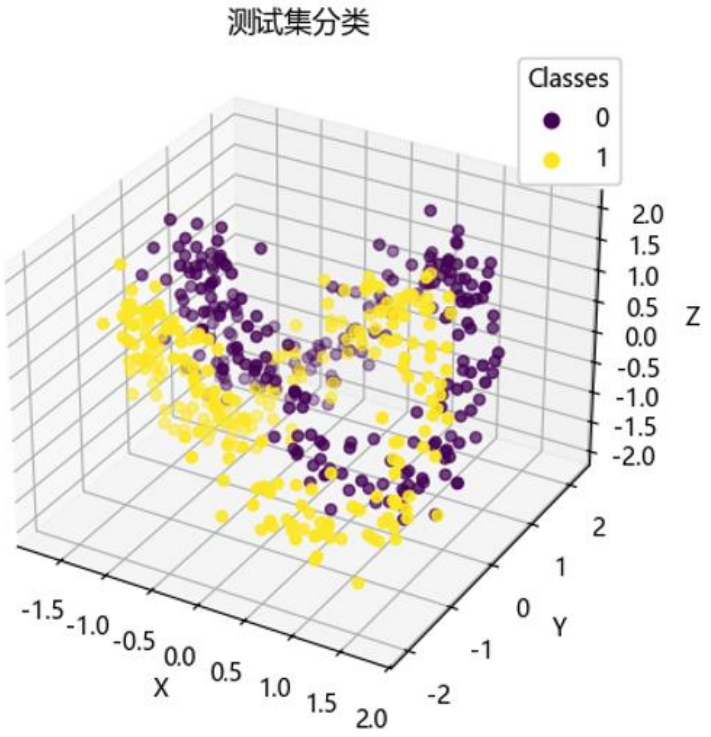
2）换句话说，在属性空间中找到一些超圆，用这些超圆来判定正反类。

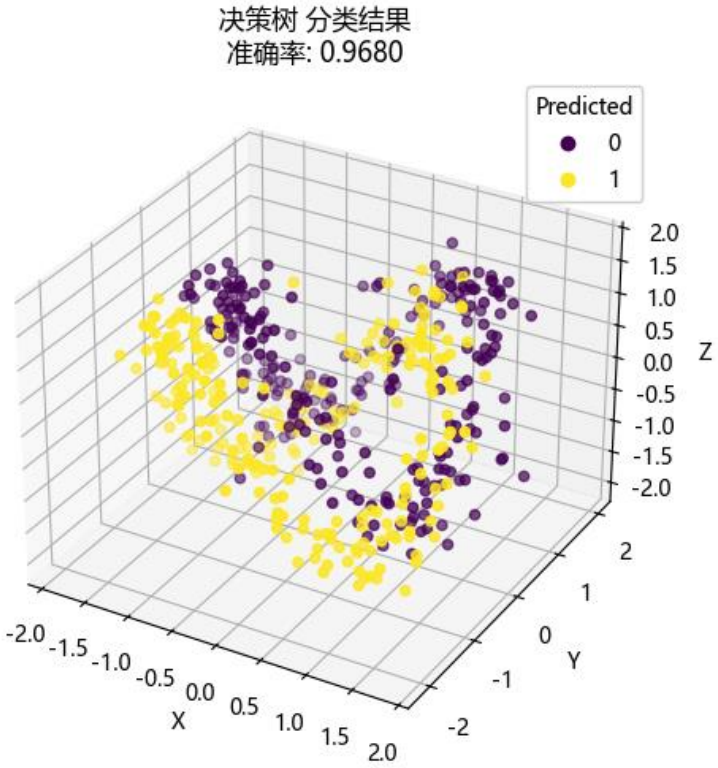
Sigmoid 核：

1）同样地是定义一些 base，

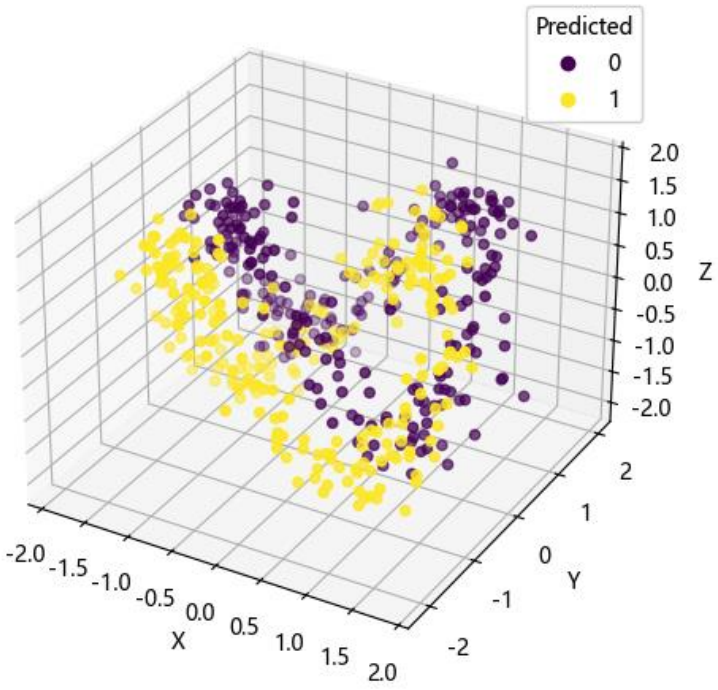
2）核函数就是将线性核函数经过一个 tanh 函数进行处理，把值域限制在了-1 到 1 上。

实验分析

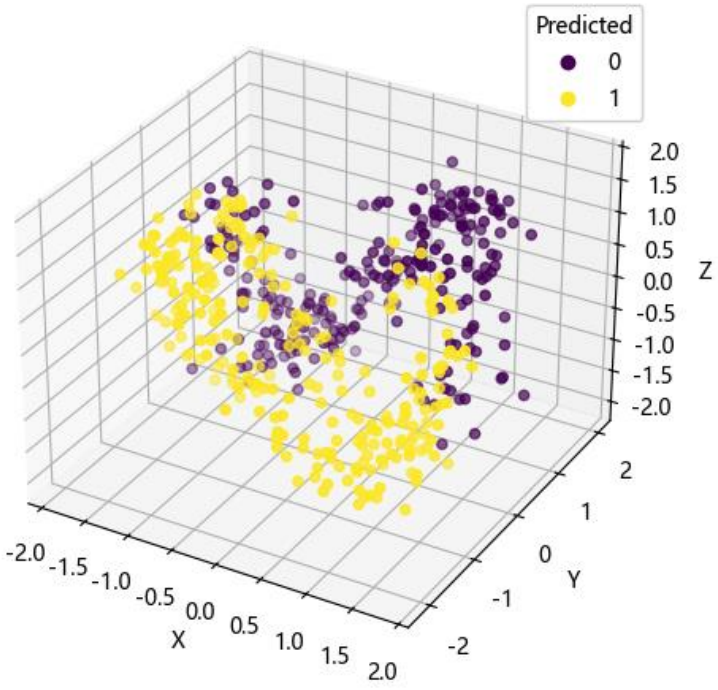

测试集分类
Classes
0
1
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
Z
2
1
0
-1
-2
Y
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
X

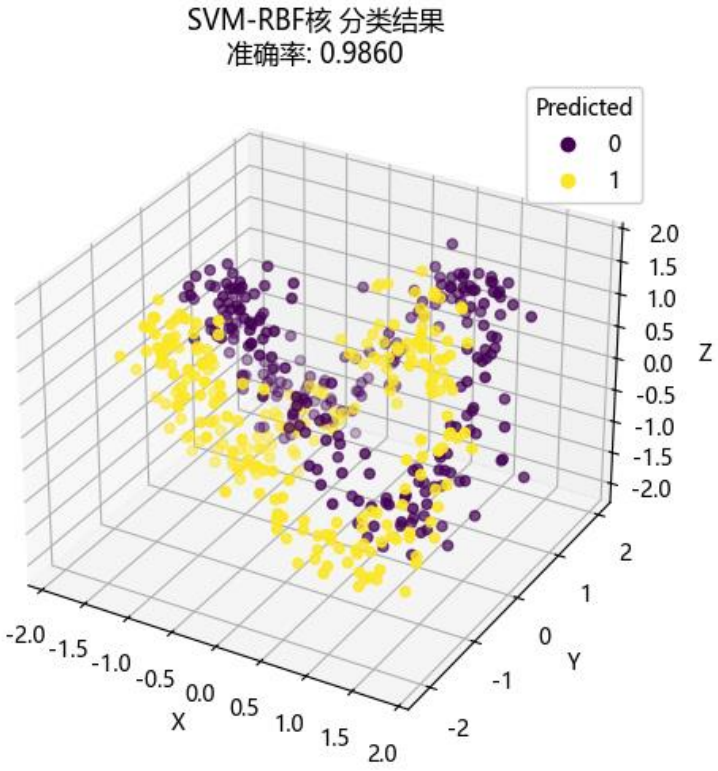

决策树 分类结果
准确率: 0.9680
Predicted
0
1
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
Z
2
1
0
-1
-2
Y
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
X

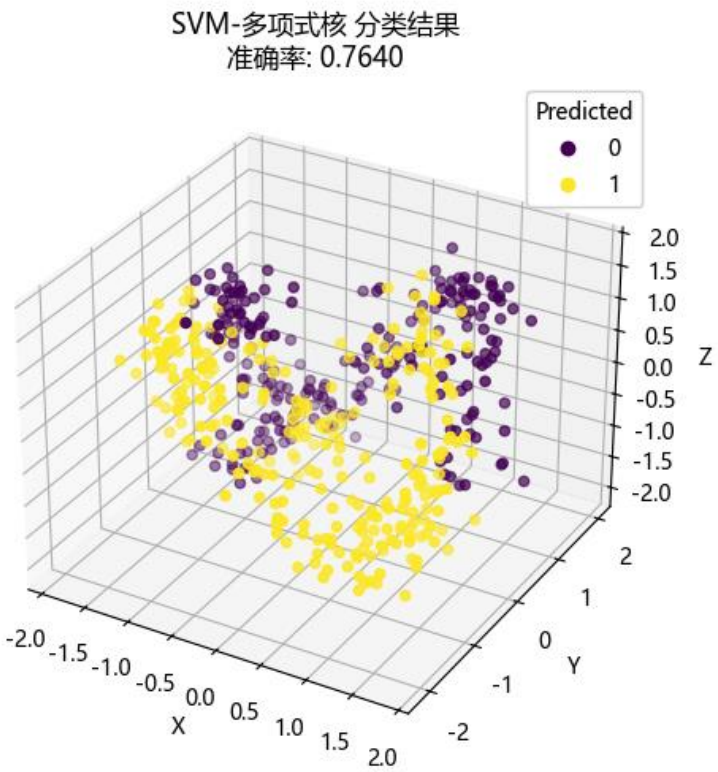
AdaBoost+决策树 分类结果
准确率: 0.9840



SVM-线性核 分类结果
准确率: 0.6820




SVM-RBF核 分类结果
准确率: 0.9860
Predicted
0
1
X
Y
Z


SVM-多项式核 分类结果
准确率: 0.7640
Predicted
0
1
X
Y
Z

实验结论

针对于三维螺旋数据分类，SVM 使用 RBF 核因能通过高维映射精确拟合复杂螺旋结构表现最佳，AdaBoost 集成决策树通过权重调整聚焦局部特征次之，而纯决策树因轴对齐分裂限制及线性 SVM 因无法处理非线性分离表现较差，体现了算法归纳偏置与数据几何特性的匹配决定性作用。