使用不同方法进行点集分类的报告

徐阳
xuyangx@buaa.edu.cn

Abstract

这是使用不同方法进行点集分类的实验报告。本人使用了决策树、集成的决策数算法、支持向量机这三种方法对三维空间中的一组点集进行了分类，并进行比较分析。对于支持向量机的分类方法，本人还使用了不同的核函数进行预测。

Introduction

二分类问题是模式识别的重要议题。通过对三维空间中的点集进行分类练习，可以使得我更加深入地理解模式识别课程中的不同算法。

Methodology

下面是我进行三维空间点集二分类的算法：

M1: 决策树

决策树是一种基于树结构的监督学习算法，主要用于分类和回归任务。它通过一系列规则对数据进行拆分，最终形成一棵树来做出预测。其核心思想是通过一系列“if-else”问题逐步拆分数据，最终到达预测结果（叶子节点）。为了构造决策树，引入信息增益和基尼系数。每次决策将优先使用能够让信息增益变化最大的特征。该算法的优点是直观易解释，缺点是容易过拟合。

M2: AdaBoost决策树

AdaBoost（Adaptive Boosting） 是一种基于决策树的集成学习算法，通过组合多个弱分类器（如简单的决策树，称为“决策树桩”）来构建一个强分类器。其核心思想是逐步调整数据权重，使模型聚焦于难以分类的样本，最终加权投票得到预测结果。其核心思想是迭代训练多个弱分类器（如深度为1的决策树桩），每个分类器针对前一轮分错的样本调整权重。根据错误率动态调整样本权重和模型权重，错误率低的弱分类器获得更高投票权。

M3: 支持向量机

支持向量机（Support Vector Machine, SVM） 是一种强大的监督学习算法，主要用于分类和回归任务。其核心思想是寻找最优超平面，最大化不同类别数据之间的间隔（Margin），从而提升模型的泛化能力。其核心思想是在特征空间中找到一个超平面（如二维空间中的直线），将不同类别的数据分开，且间隔最大化。

对于三维点集的二分类问题，该方法的实质就是在三维空间中寻找一平面，使得该平面可以最大程度地将点集分割。

显然，如果三维点集的分布比较复杂的话，很难找到这样一个满足条件的平面。所以引入核函数，将三维点集映射到高维空间中，在高维空间中寻找超平面，分割两部分点集。

Experimental Studies

下面是不同方法得到的实验结果

Table 1：决策树结果

最大深度	2	3	5	10
决策准确率	0.722	0.725	0.868	0.941
	0.723	0.738	0.845	0.975
	0.710	0.73	0.897	0.923

Table 2：AdaBoost算法结果

规模	深度	2	3	5	10
10		0.733	0.708	0.975	0.968
50		0.707	0.732	0.980	0.970
100		0.797	0.983	0.978	0.960

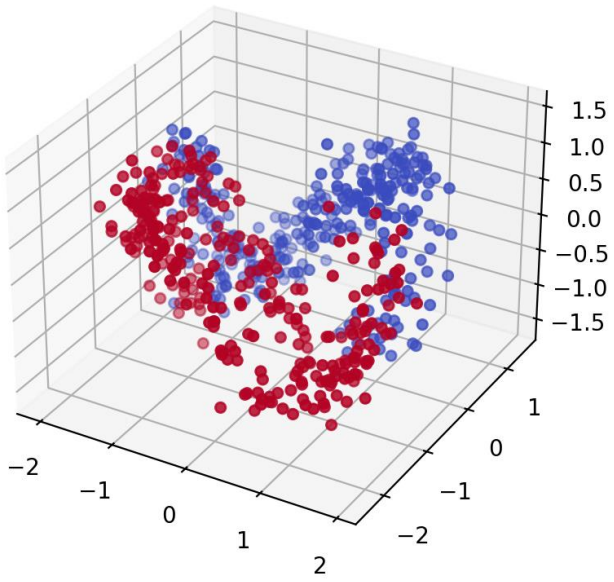

Figure 1：线性核函数svm结果

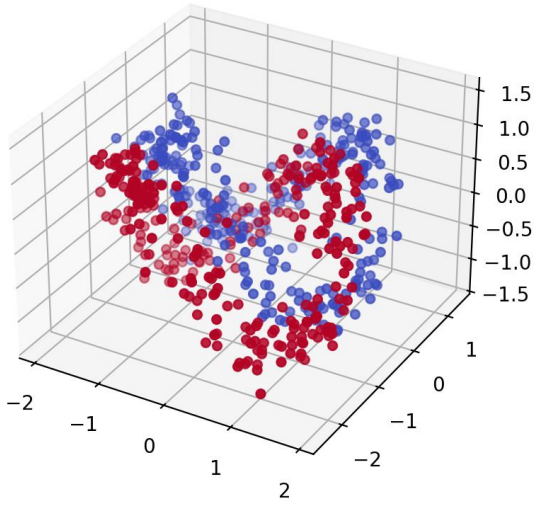

Figure 2：rbf核函数svm结果

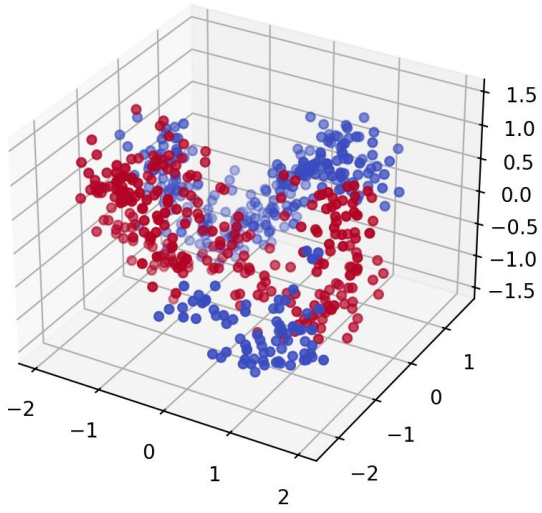

Figure 3：sigmoid核函数svm结果

Table 3: 使用不同核函数的准确性

核函数	Linear	Rbf	Sigmoid
准确率	0.72	0.98	0.6

Conclusions

1、对于决策树算法，在一定范围内，深度越大，预测效果越好

2、对于AdaBoost算法，规模，也即估计器的数量对算法预测效果有很大影响，在一定范围，规模越大，预测效果越好；基础估计器的性能，对应着单个决策树的深度，对算法预测效果也有很大影响，在一定范围内，深度越大，预测效果越好。但是两个因素都有边际效应。当深度过大或者规模过大时，预测准确率随参数增加上升缓慢，甚至有下降的情况出现。