模式识别第二次作业

一、数据集分析

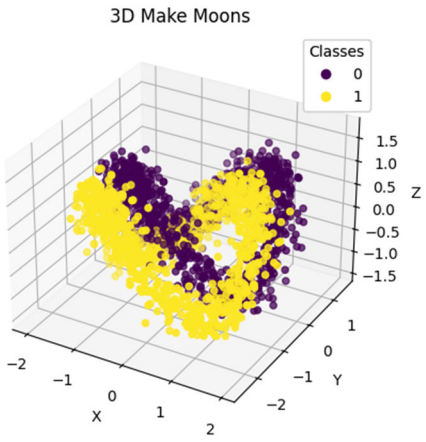

图 1 数据集

在本次作业中，我们使用一个给定函数 make_moons_3d 来生成数据集，该函数生成了类似于三维空间中的“月亮”的数据集，本次作业在该数据集的基础上进行，对 Logistic Regression, SVM 与 XGBoost 三种算法进行分析与比较。

二、实验过程

利用给定函数生成训练集与测试集，分别使用 Logistic Regression, SVM 与 XGBoost 三种算法，根据结果比较分类性能。

2.1 Logistic Regression

逻辑回归（Logistic Regression）是统计学和机器学习中常用的一种预测分析算法，尤其适用于二分类问题。该算法计算效率高，相较于其他算法，训练和预测的速度较快，下图为可视化结果。

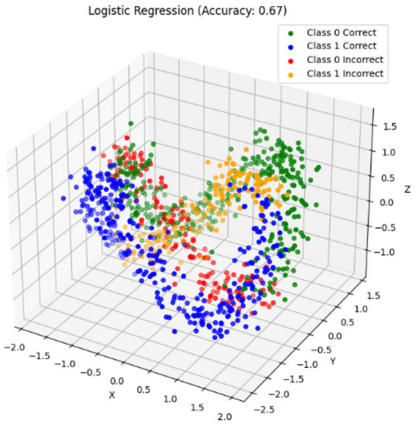

图 2 逻辑回归结果

2.2 SVM

支持向量机（Support Vector Machine，简称 SVM）是一种强大的监督学习算法，用于分类、回归甚至异常检测。SVM 在很多实际应用中都显示出较低的泛化错误率，是一个健壮的分类器。

在原始特征空间中线性不可分的情况下，SVM 可以通过引入一个核函数将数据映射到一个更高维的空间，使得数据在新的空间中变得线性可分。常见的核函数包括线性核、多项式核、径向基函数（RBF）核和 sigmoid 核等。

在本次实验中，我选用了线性核、多项式核、径向基函数核三种核函数来进行实验。对应的可视化结果如下：

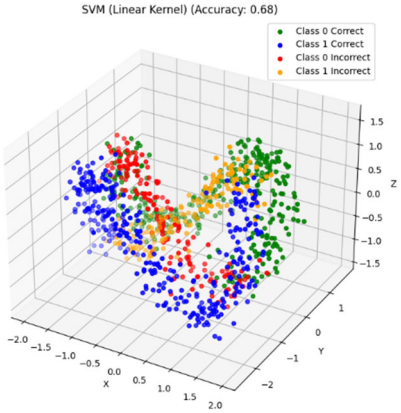

图 3 线性核结果

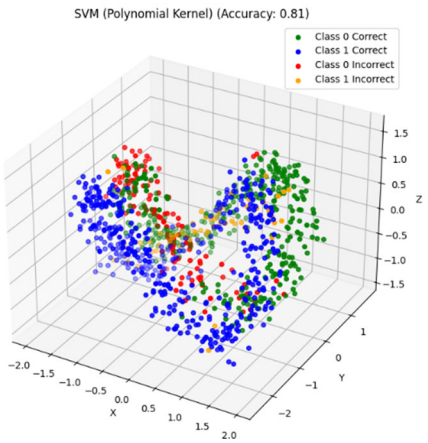

图 4 多项式核结果

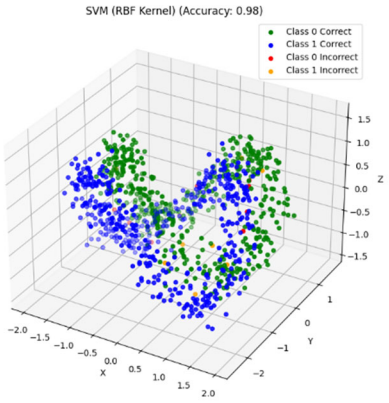

图 5 RBF 核结果

2.3 XGBoost

XGBoost（eXtreme Gradient Boosting）是一个高效的机器学习算法，广泛用于竞赛和工业环境中，特别是在处理结构化数据的分类和回归问题上表现出色。XGBoost 使用梯度提升算法，这是一种通过迭代地添加弱预测模型（通常是决策树）以解决分类和回归问题的技术。在每一步，XGBoost 添加新的模型，尽量减少上一轮预测的残差。下图为可视化结果：

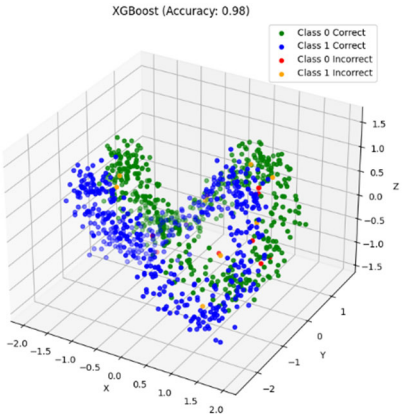

图 6 XGBoost 结果

三、总结与讨论

通过实验，我们可以看到，使用 Logistic Regression 的效果是最差的，在测试集上的准确率为 0.67。

针对 SVM 算法的不同核函数，得到的结果也有较大的差距，由差到好分别为线性核、多项式核、RBF 核，其准确率分别为 0.68/0.81/0.98。

使用 XGBoost 方法效果较好，得到的准确率为 0.98。

针对实验使用的不同方法，对结果进行分析。

首先分析 Logistic Regression，这种算法通常在线性可分数据集上表现快速良好，但是我们要处理的数据集具有复杂的非线性模式，逻辑回归的性能就会下降，因为它无法捕捉这种复杂关系。通过投影图我们可以清晰的看到，交界处的数据

处理结果很差。同样的，针对 SVM 方法中的线性核，也无法对于该类复杂非线性数据进行较好的处理。

针对 SVM 方法的另外两个核函数：多项式核与 RBF 核。多项式核函数通过原始特征的高次项组合增加了数据特征的维度，使得它可以较好地处理高维度非线性可分数据。RBF 核同理，可以映射到高维的空间，进而处理各种复杂非线性数据，所以相比于线性方法，这两种方法的表现更好。

这也提醒我们，在使用 SVM 方法时，要注意灵活选择核函数。

最后是 XGBoost，这种方法通过构建一系列的决策树，并对这些树进行学习，使得模型能够准确捕捉到输入数据中的复杂非线性关系，使得模型的拟合能力较好，并且其多种正则化技术，也可以避免过拟合。

要注意的是，在大规模数据集上，尽管 XGBoost 提供了高效的计算，但在资源有限的情况下训练时间可能仍是一个问题。我们应该根据实际问题，来选择不同的方法。

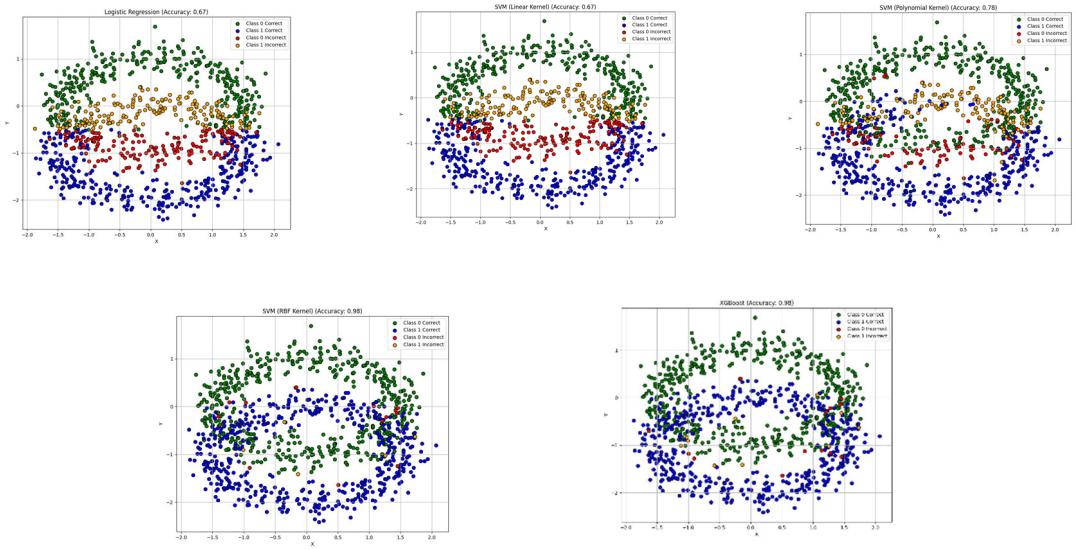

图 7 Z 方向投影图