大作业（应用层-机器学习）

一、工具介绍

完成应用层阅读笔记时所选的文章是“Scikit-learn: Machine Learning in Python”，它给出了一个集成了多种机器学习算法的模块scikit-learn。可以从github中获得scikit-learn模块的代码，网址为 https://github.com/scikit-learn/scikit-learn 。

从该链接中下载代码，便能看到scikit-learn模块的具体内容。另外，链接下载的代码还给出了使用说明和一些例程，以方便使用者理解和操作。

本次大作业使用的集成开发环境是PyCharm，编译器版本为Python3.7。

二、实验内容

1.准备工作

按照代码的说明文档，由于之前我已经安装过scikit-learn所需的两个前置模块numpy和scipy，所以只需要在终端中输入指令

```
pip install -U scikit-learn
```

就可以安装scikit-learn模块。

接下来将从scikit-learn模块中选取几个机器学习算法进行操作展示。

2.决策树

使用决策树拟合带有噪声的正弦曲线。

在二维平面空间中随机选取正弦曲线上的一系列点，再选取几个噪声点，共同作为训练样本。使用决策树拟合回归模型(对正弦曲线做线性拟合)。

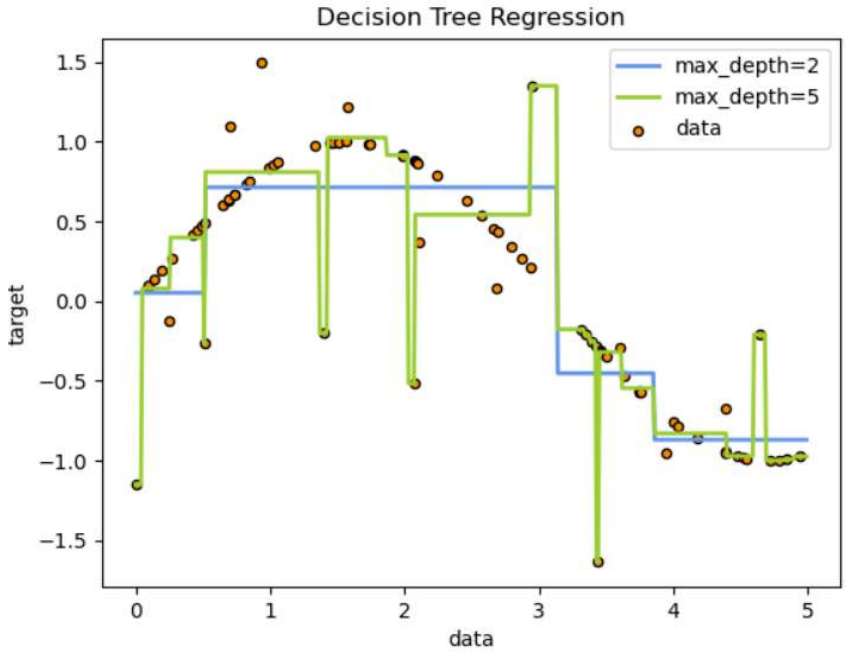

图 1 决策树拟合正弦曲线

使用拟合后生成的决策树做预测。在 x 坐标 0 到 5 之间，以 0.01 为间隔连续取一系列点作为输入，观察输出（即预测结果 y 的值）。在决策树最大深度为 2 和 5 的情况下分别做上述预测，得到的预测结果如图 1 所示。

从图 1 中可以看出，无论是最大深度为 2 还是为 5 都能一定程度地拟合正弦曲线。当最大深度为 2 时，对正弦曲线的拟合效果较差，但是几乎没有受到随机噪声点的影响。当最大深度为 5 时，对正弦曲线的拟合效果较好，但是受随机噪声点的影响很大，若使用该拟合结果进行预测，则在某些位置会有过大的预测误差。

3.支持向量机 SVM（用作分类）

在两种类别的相互分离的数据集中，画出支持向量机模型中的超平面。结果如图 2 所示。图 2 中的实线即为超平面，对二维分类问题来说就是一条分类线。虚线则是过各类中离分类线最近的样本且平行于分类线的直线，虚线和实线（分类线）之间的间隔就是分类间隔。

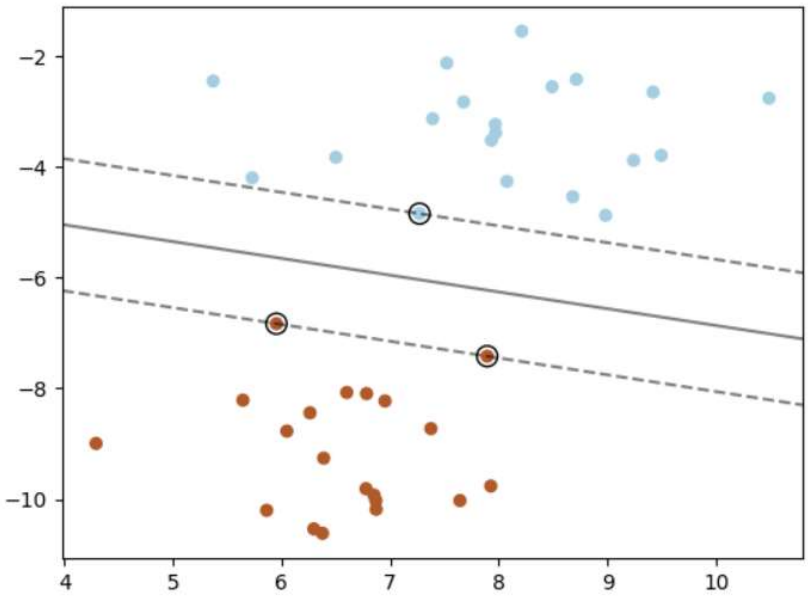

图 2 作出超平面

4. 一类支持向量机（one-class SVM）

使用一类支持向量机（one-class SVM）来做新奇检测（novelty detection）。与一般的 SVM 不同，一类支持向量机是无监督的学习，它只是划出数据集边界以判断新的数据是否与原训练集为同一类别，也就是将新的数据分为与训练集相似或不同两种情况。

其演示结果如图 3 所示。图 3 中，200 个白色点是训练样本点，使用 SVM 进行训练，得到红色的边界线。

用它来做预测时，如果待预测的点落在红色的边界线内，则被判断为常规点，与训练样本集是一个类别的；如果落在红色边界外，则被判断为反常点（新奇点），与训练样本集不

是一个类别。图 3 中，40 个紫色的常规点和 40 个黄色的反常点是类别已知的测试点。可以看到，对紫色的常规点，大部分都落在红色边界内，预测正确，只有少数（7 个）点预测错误；对黄色的反常点，全部都落在红色边界外，全部预测正确。

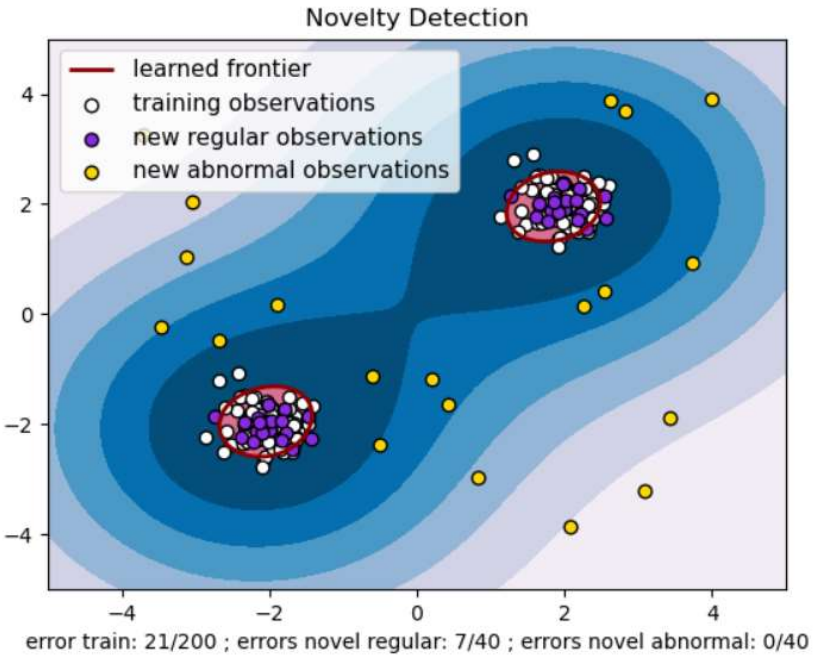

图 3 新奇检测

三、总结与感想

以上选取了 scikit-learn 中所集成算法对 3 个问题做了简单的操作展示。除此之外 scikit-learn 模块还能用于许多其他的有监督和无监督机器学习问题。它的操作相对简单，获得的结果清晰明了，而且便于扩展，也很适合非专业人士使用。

而对于我自己，经过亲自动手实践，一方面我阅读说明，操作例程，观察和分析结果，对机器学习有了更直观更深入的了解；另一方面通过配置 Python 编程环境，从 github 上下载运行代码，也加强了我管理和使用计算机程序代码等相关的技能。这也使我具备了一定的使用机器学习算法来分析解决相关问题的能力。

通过这次自主的训练实践，我收获的不仅仅是对 scikit-learn 模组的使用和对机器学习的进一步理解，我还产生了对机器学习和计算机编程更浓厚的兴趣，我更是学到了使用 github 等平台寻找所需代码以解决更多的实际问题和自我学习的方法。