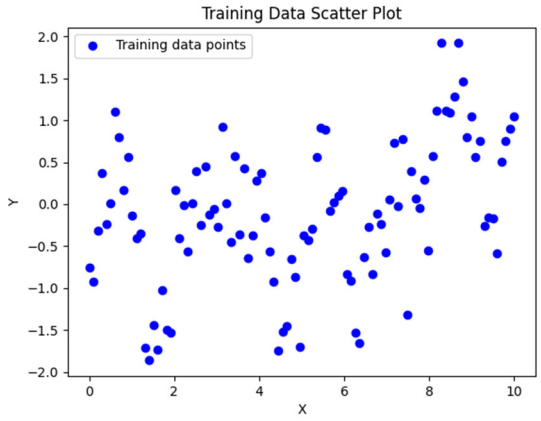
模式识别第一次作业

一、数据分析

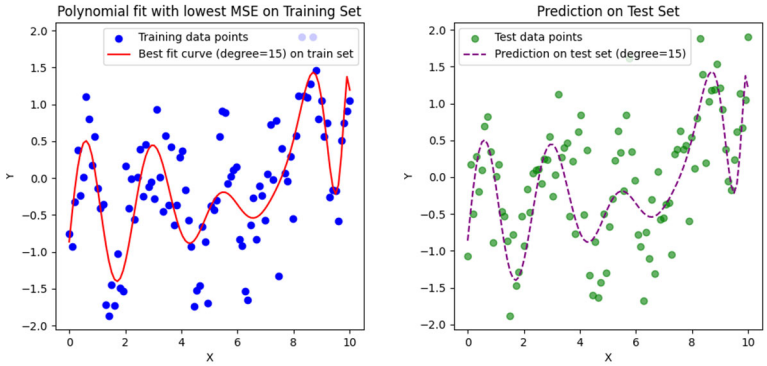
绘制所给数据集的散点图如下：



训练集数据呈非线性分布，首先考虑使用多项式方法以及决策树方法，并用均方差作为模型的度量。

二、多项式

遍历（1-30）的不同次数，得出 degree=15 时，在训练集上得到的 MSE 值最小，为 0.29073，进而在 degree=15 时，在测试集上得到的 MSE 值为 0.33310，所得图像分别如下所示：

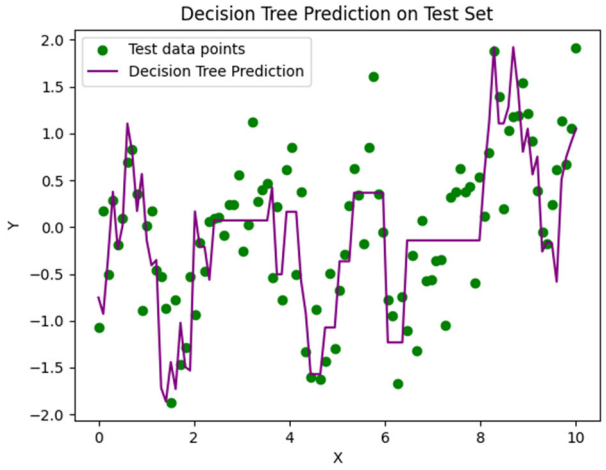


结合图像我们可以观察得到，虽然在这种情况的 MSE 值较小，整体的拟合结果还算不错，但是在 4-7 范围内的数据拟合结果偏差较大，考虑噪声或异常值的情况，该拟合结果是可以接受的。

三、决策树

根据所学知识，对于呈现出非线性特征的数据，可以采用决策树方法进行拟合，在实际操作过程中，如果 depth 设置过大，会出现过拟合现象，进而模型在

测试集上的表现极差。限制层数在一定范围内进行遍历，得到最佳 MSE 结果：0.3197792763688166。所得图像如下图所示：

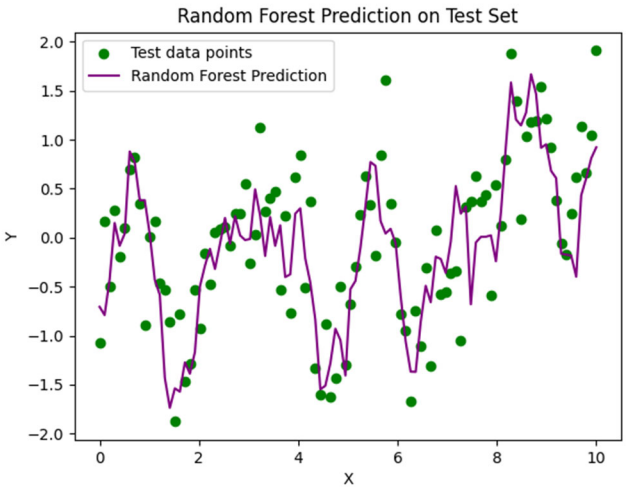


但是在实际问题中，可能没有测试集这种条件供我们调试，所以该方法存在一定问题，或者说我对于决策树方法还没有学习到一定的程度，后续仍需对该方法进行进一步的学习。

四、随机森林

在完成决策树的构建后，考虑随机森林，随机森林是基于树的机器学习方法，该算法运用多棵决策树的力量来进行决策，这也是其被称为随机森林的原因，这种方法相对于决策树有着一定的优势，防止过拟合便是其中的一点。

使用随机森林算法得到在训练集上的 MSE 为 0.04576，在测试集上的 MSE 为 0.27645，相较于决策树算法有了一定的提升。



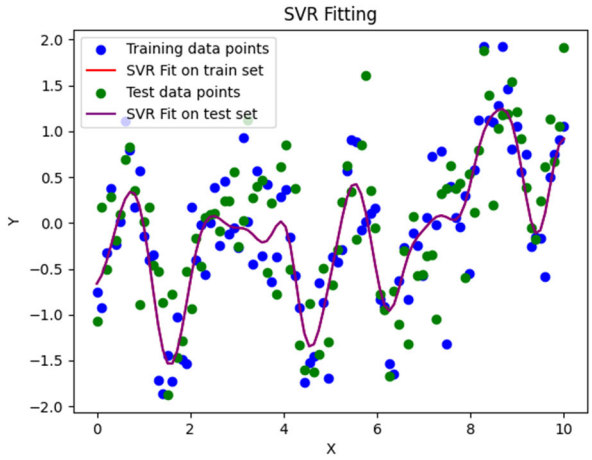
五、SVM

完成上述几个模型后，接着考虑使用 SVM 完成拟合，支持向量机（support

vector machine，SVM，又名支持向量网络）是在分类与回归分析中分析数据的监督式学习模型与相关的学习算法。简单的来讲，SVM 是一种二分类模型。

由于训练数据线性不可分，考虑使用 RBF（径向基函数）为核函数，先利用试错与经验框定超参数的范围，再对这个范围进行网格搜索来得到较优的参数取值，进而对数据进行拟合。

根据选定的模型，最终的测试集拟合结果如下图所示：



训练集与测试集上的 MSE 值分别为 0.17975 与 0.24919。

六、总结

通过上述几种方法的比较，可以很清晰地看到这些方法的优缺点，最终使用 SVM 方法对数据进行拟合，得到的效果是相对较好的，在训练集与测试集上得到的均方误差值都较小，并且没有像决策树、随机森林算法那样在训练集上出现过拟合现象。通过对复杂非线性数据的拟合，也可以体会到不同算法的魅力以及机器学习本身的乐趣。