深度学习 week5

数学科学学院 杨睿涵 17307130276

2021/4/2

一、实现思路

1.鸢尾花数据分析（iris_analysis.py)

利用 matplotlib.pyplot 中的 Violinplot、Pointplot、Pairplot 工具，从 数据分布和斜率观察特征与品种的关系，以及特征关系的矩阵图

利用 seaborn 中的 heatmap 工具，展现不同特征之间的相关性的热力图

程序实现：

从 datasets 导入 iris 矩阵类型的数据以及 dataframe 类型数据 load_data 函数用于初步查看数据结构，实际运行将被注释掉 iris_information 函数用于输出数据的基本信息，包括鸢尾花种 类、特征、数据数量

iris_mean_cov_R 函数输出不同鸢尾花特征的均值、协方差矩阵 以及相关系数

violin_point_pair_plot 函数用于绘制特征的小提琴图、特征的折 线置信区间图以及特征关系的矩阵图

heatmapplot 用于绘制不同特征的热力图 main 函数用于连接上述函数

2.多项式回归分析

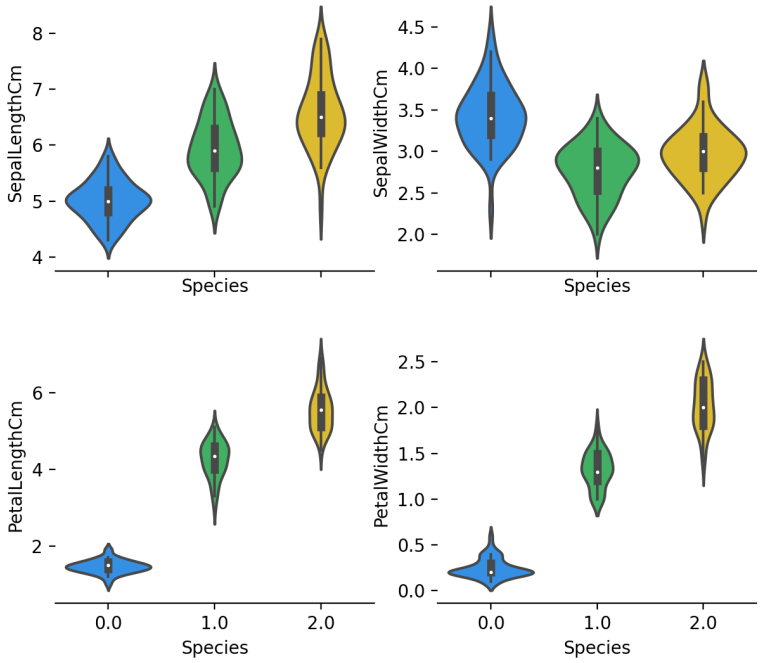
利用线性回归进行拟合(linear_regression.py)

程序实现：直接用线性回归的公式（正则化后公式）求解

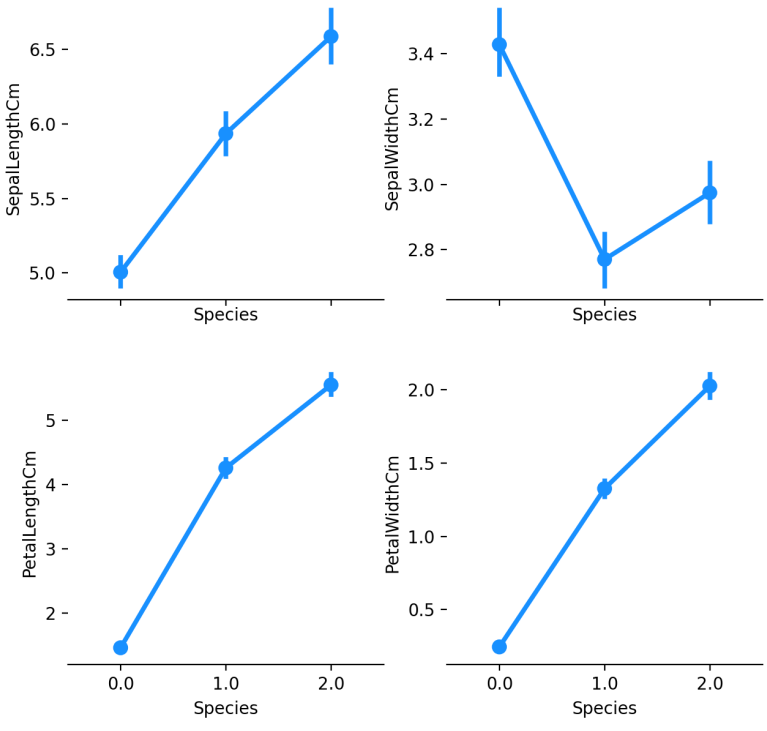
二、运行结果与思考

鸢尾花数据集结果图

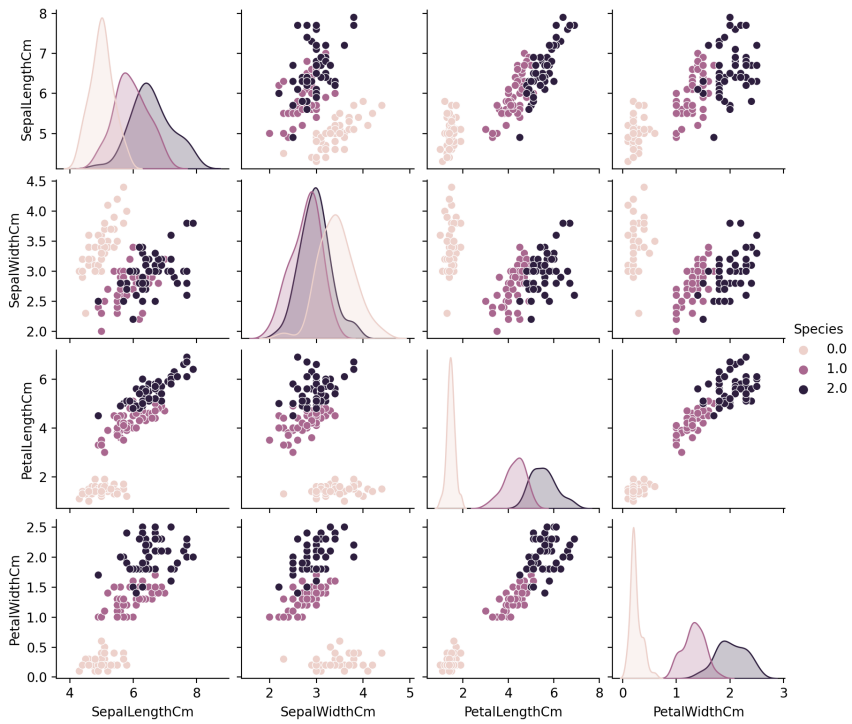
1）特征的小提琴图



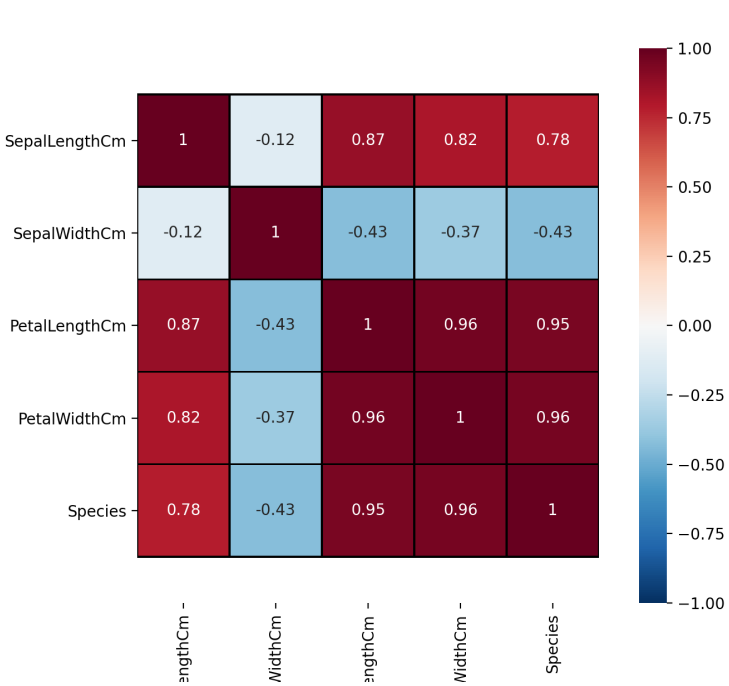
2）特征的折线置信区间图

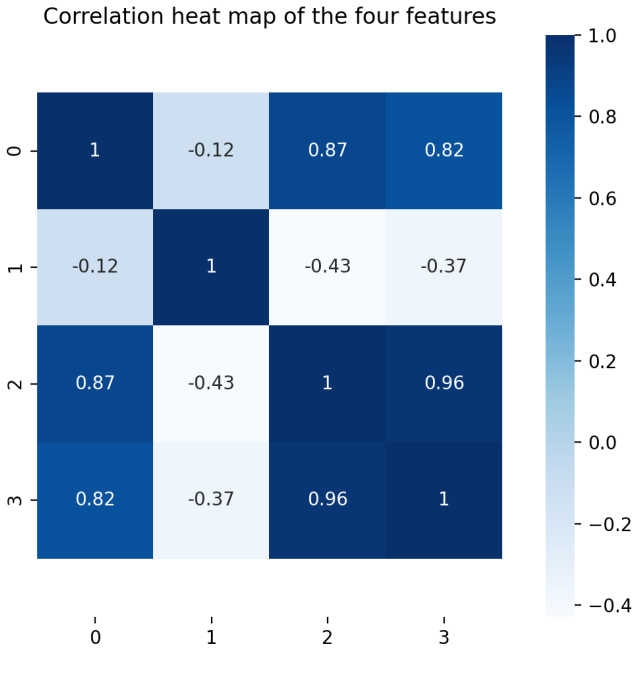


3）特征之间的矩阵图



4）特征相关的热力图



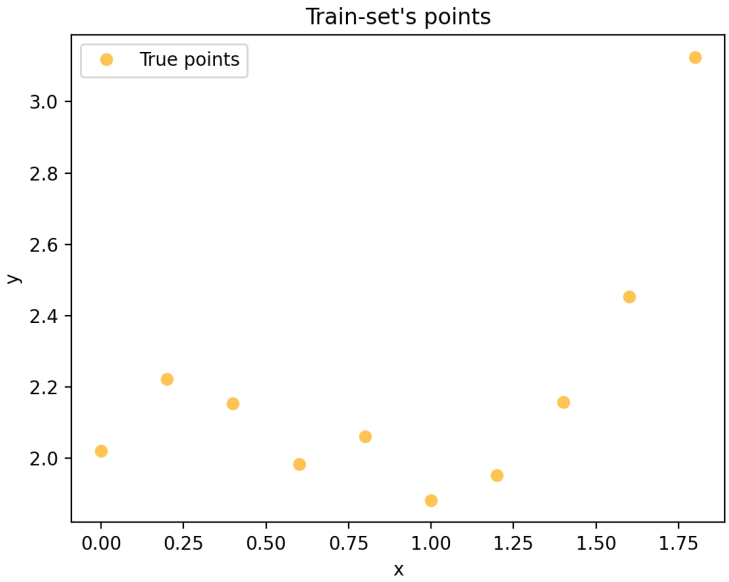


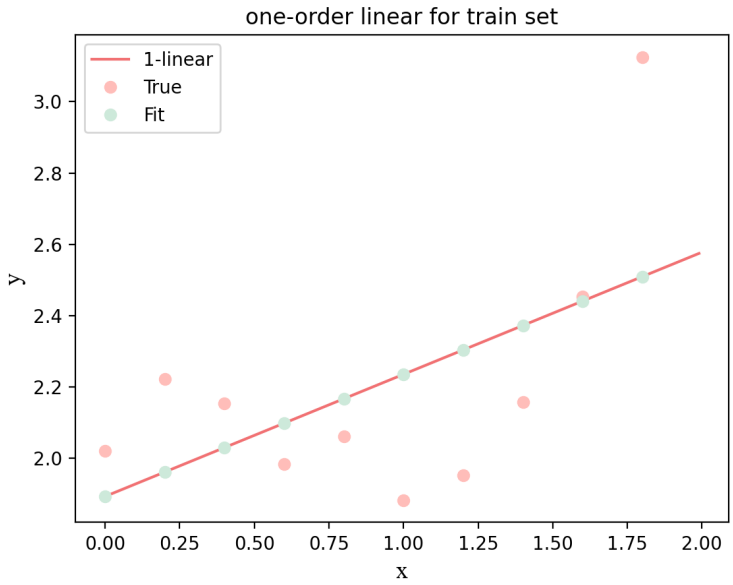
思考：容易发现鸢尾花花瓣的长与宽具有很强的正相关性，而花萼的长与宽只有较弱的负相关性，另外花萼长与花瓣长，花萼宽都有较明显的正相关性。

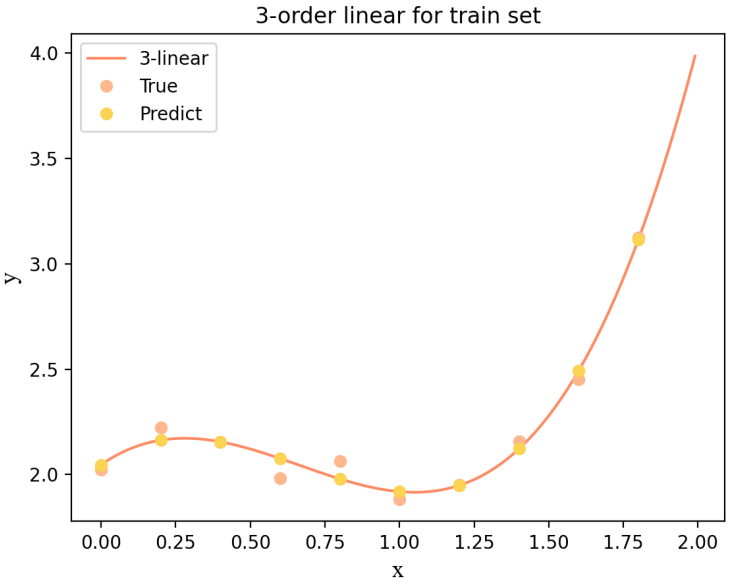
另一方面，花瓣的长宽与种类间有很明显的关系，而两种花的花萼信息之间虽然有明显的交叉，但是仍然有一定的区别。

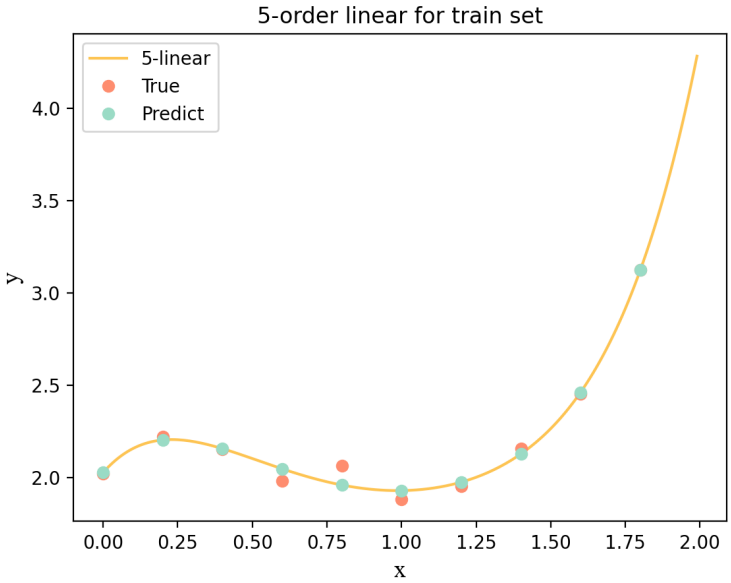
多项式回归结果图

以下依次为训练集真实数据，一次，三次，五次的拟合结果

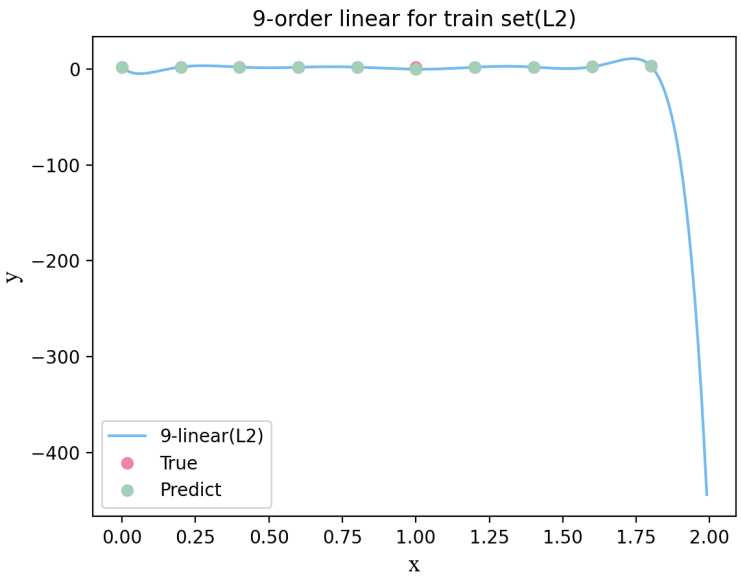


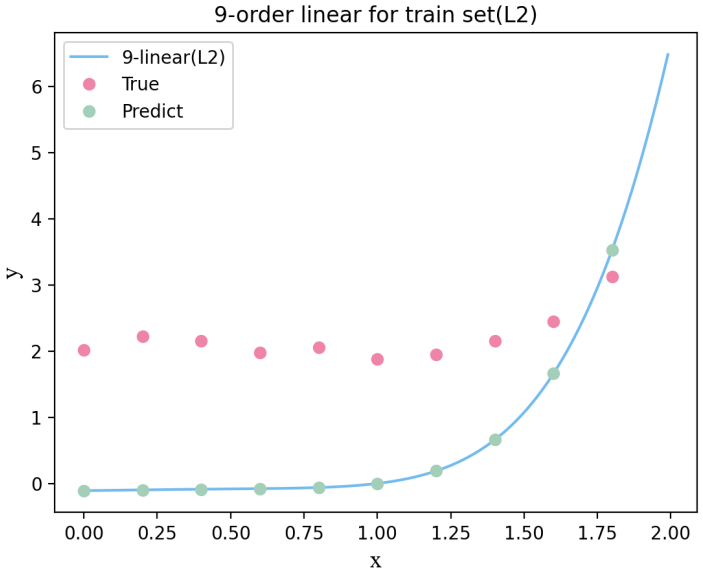


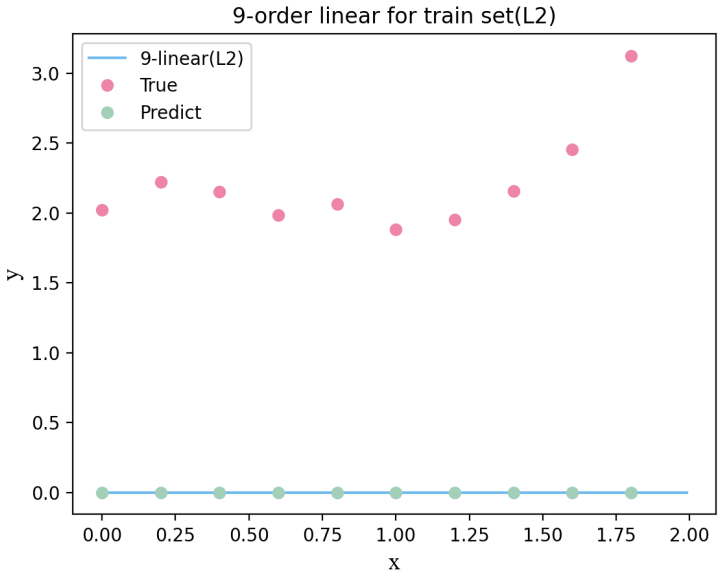




以下依次为九次，九次+L2（较小的超参数），九次+L2(中等大小超参数），九次+L2(（较大超参数）的拟合结果，以及误差结果








```
一次多项式估计的均方误差为：41.631508
3次多项式估计的均方误差为：0.370237
5次多项式估计的均方误差为：150.527248
9次多项式估计的均方误差为：573297901.824864
9次多项式+L2正则项后估计的均方误差为：117079327650.818344
9次多项式+L2正则项后估计的均方误差为：260.896282
9次多项式+L2正则项后估计的均方误差为：80.198559
```

思考：

从上述结果可以发现：

1次多项式不能很好的拟合训练集数据，属于欠拟合，从而在测试集上的结果有较大误差。

3次多项式能较好的拟合训练机数据，并且有较好的范化能力，测试集上的均方误差较小。（0.37）

5次多项式，9次多项式对于训练集数据的拟合能力比3次多项式都要好，但是泛化能力随着模型的复杂度升高而降低，在测试集上的结果误差很大，属于过拟合。

而9次多项式加了正则项后：

较小的正则化参数不能有效提高范化能力，误差依旧很大。

中等大小的正则化参数对于范化能力有较大提升，可以看到拟合的曲线接近3次多项式，但是最终测试集测试结果上依旧有较大误差。

正则化参数较大时，拟合曲线为一条直线，由于正则项的权重过大，导致模型空间的复杂程度下降，不能很好的拟合训练集，也相当于欠拟合。

所以一开始对于模型空间的选取不应过大，对于过大的模型空间，就算加上了正则化参数，改善效果也是有限的。

欠拟合的产生原因：1.模型空间较小，拟合能力差

2.在样本较少的情况下，特征过多

改进方法：1.扩大模型空间

2.减少特征个数

过拟合的产生原因：1.模型空间过大，范化能力差

改进方法：减小模型空间，加正则项（正则项相当于平衡了bias和variance，不加正则项仅让bias减小，而模型空间的增大会导致variance上升，从而范化能力下降，从贝叶斯的观点来看就是对于参数的分布进行先验估计）