作业2 总结

本次作业完成情况较好，主要有以下几方面问题，总结如下

代码实践

本次作业的本意是希望大家基于线性回归/非线性回归的方法对电力系统的稳定裕度进行回归分析。sklearn包中可以直接调用线性回归的模型，在作业批阅中发现部分同学没有直接调包，有以下几种方式

1. 利用解析公式计算：线性回归（包括Lasso和Ridge回归）有解析的计算公式，部分同学利用解析的计算表达式计算进行了计算，但是都发现了在实际的应用中，会遇到矩阵维度太高、矩阵病态等问题，包装好的线性回归函数对这些问题有一定的工程处理，感兴趣的同学可以查看函数手册。
2. 利用梯度下降法求解：线性/非线性回归本质上也是一个最优化问题，所以可以利用梯度下降等优化算法寻找最好的参数。但是自己手写的优化算法可能也会遇到数值问题。**同时，同学们可以思考，课堂中非线性回归是通过把非线性项看作一个单独变量，将非线性回归当作线性回归进行处理的。如果直接建立优化模型求解非线性回归的最优参数，这两种方法得到的结果是否相同？**
3. 利用神经网络模块求解：因为自己构造的优化算法可能存在数值问题，所以有些同学利用神经网络的构造了线性回归/非线性回归的模型，利用神经网络反向传播的方式优化参数。

同学们发现引入二次项之后，拟合精度会大幅提高。引入三次/四次项后，拟合精度进一步提升，但是训练耗时增加，不合理设置正则化项，可能存在过拟合现象。有个别同学从稳定裕度的底层物理机理，开展了一定的思考，判断几次模型更合适。有同学说，对于电路公式一般最多到二次，因此没必要选择更高次数的拟合公式。这些思考都是有益的，利用物理机理构建数据驱动方法，可以让模型构建更加准确、更加简洁、解释性更强。

另外，很多同学认为题目中说的回归是指广义的回归问题，所以使用了随机森林/SVR等方法进行回归，有这些尝试也很好。但是大部分同学都没有对随机森林的超参数进行调优，下节课会介绍随机森林方法，同学们可以更清楚随机森林有哪些超参数可以调整。

理论误区

从理论层面，本次作业要求大家判断是否发生了欠/过拟合现象。很多同学尝试用课堂中所教的方差和偏差的理论进行解释。这种尝试很好，但是用得似乎并不正确。

什么是方差？方差一定是针对于某一随机变量的，一定是针对某一概率分布的。那么是哪个随机变量的方差呢？张老师课堂上说的是“假设有无穷多个平行宇宙，在其他平行宇宙模型表现得如何”，这一解释很生动。我们看到的数据背后服从了一个特定的规律（由于有误差的存在，更具体地来说是一个概率分布），我们构建了学习器来学习这个规律。但是我们看到的只是一组实例，我们的学习器在这组实例上学习，会得到一个学习误差，一般会比较小。**那么如果我们再生成一组服从这样的特定规律的数据，我们的学习器在其上的拟合误差是怎么样的？**如果拟合误差大的话，证明对于同一规律产生的不同数据集，我们的学习器表现参差很大，这才是我们所说的方差的概念。在这种情况下，我们可以认为学习器发生了过拟合，过于关注了数据的局部规律，忽视了整体的大规律。

从上面的解释也可以看到，方差的计算需要在不同数据集上测试得到，所以在实践中，一般而言，我们无法获得方差的计算结果。也就不能用方差-偏差理论来解释学习器是否发生了过拟合。实践中，更多的是比较验证集上误差和训练集上误差来判断是否存在过拟合现象，这相当于是对方差-偏差理论的一个简化。

其他

最后，看得出来部分同学利用了GPT等生成模型生成了参考代码，或者说让GPT推荐了如何进行数据驱动完成我们这门课程的作业。作为一门以大数据为主题的课程，我们不反对同学们使用GPT这类先进的大数据工具。但是善用和滥用的边界在哪里？这是一个目前没有答案的问题。

简单的代码可以由GPT指导同学们完成，这是没有问题的，这省去了同学们查阅资料的时间成本。但是数据驱动方法的思想是需要同学们自己获得的，希望同学们能保持独立思考，数据之间存在哪些关系？用什么样的模型更合适？为什么模型表现不好？怎么调整参数能避免模型的过拟合或欠拟合？这些问题可能是在科研中，同学们要实际面对的问题。对于更多的未知问题，GPT不会总告诉你答案，但是你自己会。希望同学们能在我们这门课上掌握基本的机器学习模型，并对如何利用数据驱动方法解决电力系统实际问题有一定基础的思考。大模型是新的起点，而不是思考的终点；大模型强大的检索和生成能力，是打开了世界，还是封闭了世界，只有使用者才能决定。

同学们在之后的作业中，如果利用了大模型生成代码，可以额外指出哪些代码是大模型生成的，哪些是自己的独立思考。