过拟合问题

高偏差：欠拟合

高方差：过拟合；表现为，曲线扭曲太多，曲线尽力地通过每个样本点，泛化能力差

导致过拟合的原因：特征数大于样本数；使用了高阶多项式容易造成过拟合

解决方案：

1. 减少特征数，可以人工减少，也可以算法减少（PCA），缺点：丢失部分信息

2. 正则化，保留了所有特征，但减少量级和调整超参，该方法非常有效

什么是正则化？对目标函数是高价多项式函数时，添加的惩罚项lamda(一个随意的很大的数，比如1000，注意不要太大！容易欠拟合)，使得参数（该参数为高阶变量的系数）很小，小到接近0，相当于对应的变量的权重很小很小。

作用：简化模型。由于参数值越小，曲线也就越光滑，就不容易出现过拟合的问题。

如何应用正则化：由于难以确定哪些特征是高阶项，就不知道该缩小哪些特征，所以直接对所有特征都缩小

判断高偏差与高方差

绘制训练集误差曲线与验证集误差曲线（x轴为lamda，y轴为误差）。左边同高的为高偏差，右边训练误差曲线低、验证误差曲线高的为高方差。

绘制学习曲线（x轴为样本数，y轴为误差）知道，高偏差的时候，训练曲线先增大后趋于水平；验证曲线先下降后趋于水平，且接近训练曲线，但比训练曲线高；两者随着样本数增加而变化不大。高方差的时候，训练曲线先增大后趋于水平；验证曲线先下降后趋于水平，且接近训练曲线，但比训练曲线高；两者随着样本数增加而逐渐靠近，但仍有空隙。

由此可分析得到解决高偏差与高方差的方法：

1. 使用更多的样本可以有效减少高方差，而对高偏差无效（这个关系到需不需要花大量时间去找样本，成本问题）
2. 减少特征对高方差有效（特征过多，模型太复杂）
3. 增加特征对高偏差有效（因为特征少，模型太简单）
4. 增加lamda值，对高方差有效（因为模型太复杂，要抑制参数拟合效果）
5. 减少lamda值，对高偏差有效

总的来说，高偏差是因为模型太简单，所以要增加模型复杂度，比如增加特征，减少lamda值提高参数的作用，但与样本数关系不大。高方差是因为模型太复杂，要减少模型复杂度，比如减少特征，增加lamda以减少参数的作用，或者增加样本数