**1、过拟合、欠拟合分析**

（1）数据的划分：1）训练集（60%） + 验证集（20%） + 测试集（20%）

2）训练集（70%） + 测试集（30%）

避免使用相同的数据，即当成训练集又当成测试集，使得模型的泛化能力差。

（2）当出现模型的泛化能力差时，也就是出现过拟合或欠拟合问题。

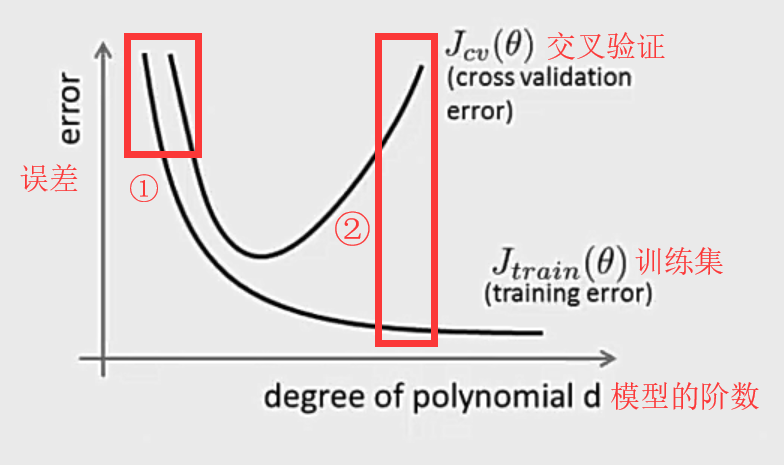
需要考虑的问题：（有可能高偏差和高方差同时存在）。

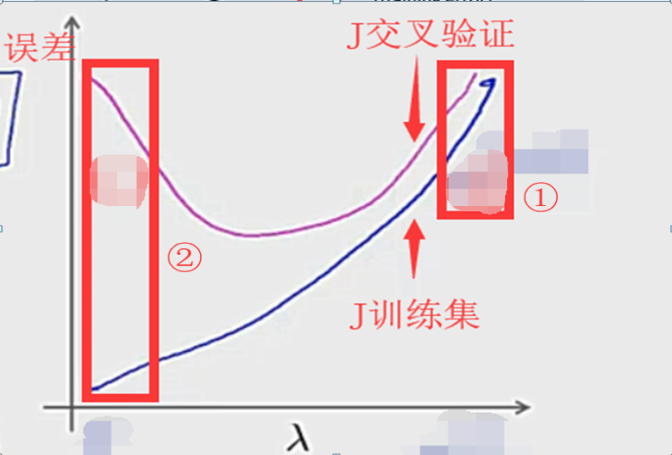
1）**高偏差**问题：模型的阶数d或 过小，出现**欠拟合。**

当，一定出现了欠拟合，如下图①区域。

2）**高方差**问题：模型的阶数d或过大，出现**过拟合。**

当 ，一定出现了过拟合，如下图②区域。

模型阶数d的取值，对模型的影响：

正则化的取值，对模型的影响：

**总结**：1）训练集误差和交叉验证集误差近似时：高偏差/欠拟合

2）交叉验证集误差远大于训练集误差时：高方差/过拟合

（3）学习曲线：（横坐标训练集m的数量）

**学习曲线绘制：**（一共两条曲线）

①纵坐标：**训练集误差**（训练集的代价函数值）和**交叉验证集**误差（交叉验证集的代价函数值）

②横坐标：**训练集数据**数量：**m**。

**作用：**

①用于改进算法；

②用于判断是否出现高偏差、高方差。

1）**高偏差**：（学习曲线：左图），出现**欠拟合**现象。

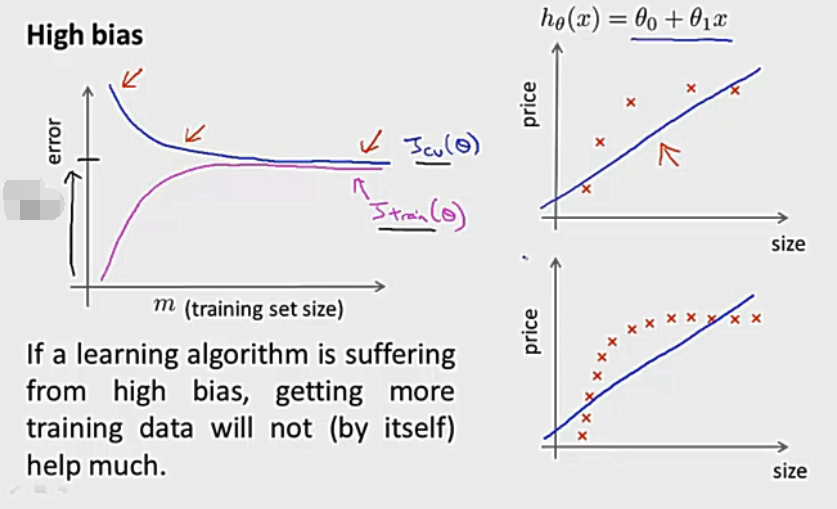
①修正方法：1、增加特征数量；

2、增加多项式阶数；

3、减小；

②增加训练集数量m，不能减小误差：

由图像可知，m增加→误差依然很大（终值不会减小），故**不需要**增加训练集m的数量。



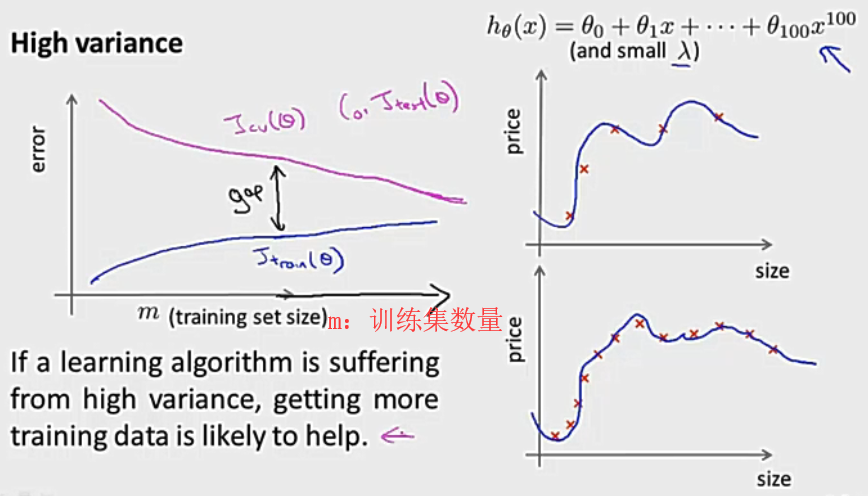
2）**高方差**：（学习曲线：左图），出现**过拟合**现象。

修正方法：1、增加训练集数量m，有助于减小误差，

m增加→误差下降→减小过拟合现象。

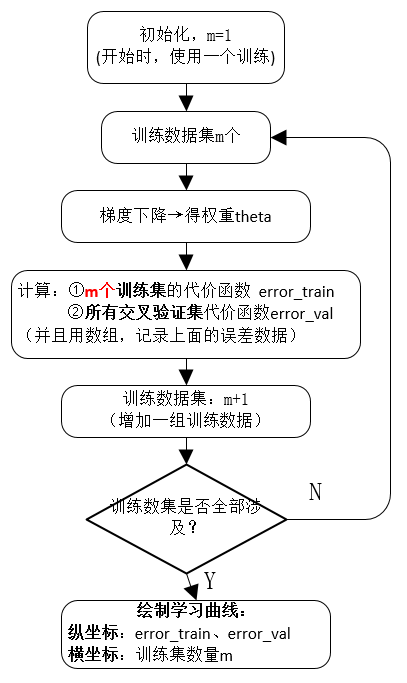
2、减少特征数量；

3、增加；

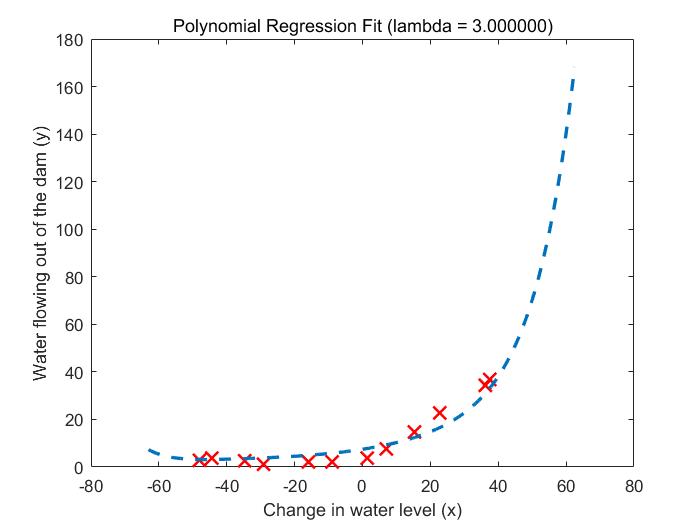


**总结：**想要获得低偏差、低方差模型：1）确保特征数量够多（保证低偏差）

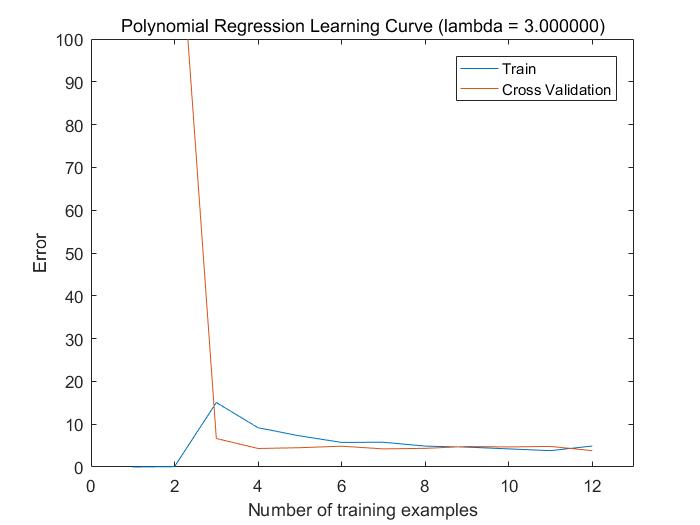
2）确保训练及数量够多（保证低方差）

附：学习曲线绘制步骤：训练样本：第一个 → 最后一个

附：（例程）学习曲线

1）最合适的拟合曲线，选取lambda=3

2）对应的学习曲线

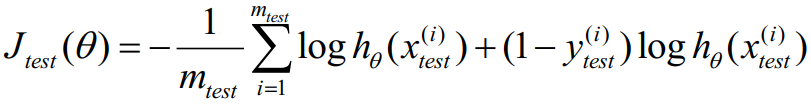


（4）误差计算

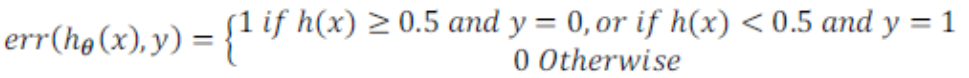
有两种计算误差的方法：

1、线性回归模型：使用测试数据计算代价函数

2、逻辑回归模型：①测试数据计算代价函数：



②使用误分类的比例：计算每一个测试集数据，



然后计算其平均值。

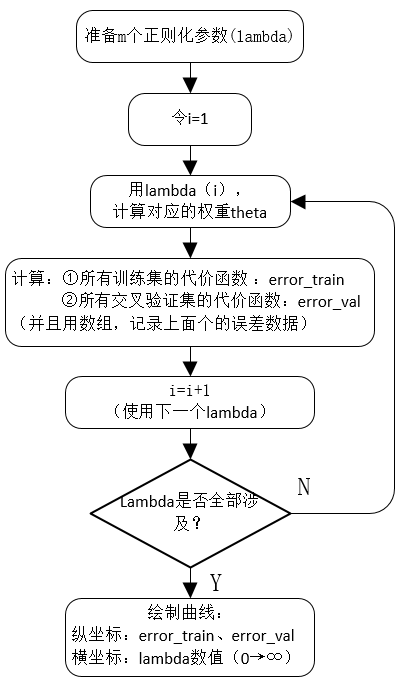
（5）使用交叉验证数据集，选取最优的正则化参数lambda

作用：正则化参数lambda，关系到最终模型的拟合程度

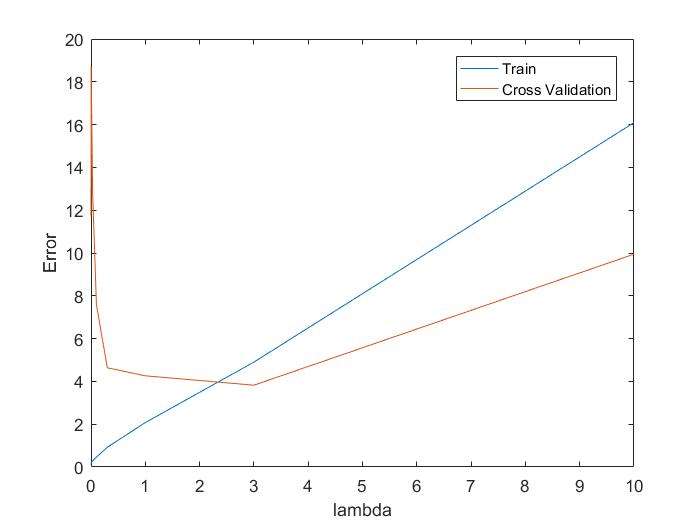
1、lambda过大：导致欠拟合

2、lambda过小：导致过拟合

附：具体执行步骤：取曲线最低点对应的lambda（即：最优参数）



附：示例结果图片**（选择lambda=3，，为最优解）**



**2、误差分析**

（1）偏斜类

类偏斜情况：表现为我们的训练集中有非常多的同一种类的实例，只有很少或没有其他类的实例。

（2）查准率、召回率

当遇到类偏斜问题时，这两个参数，作为模型的评估度量值。

**结论：**当一个模型具有高查准率、高召回率时，模型的分类性能好

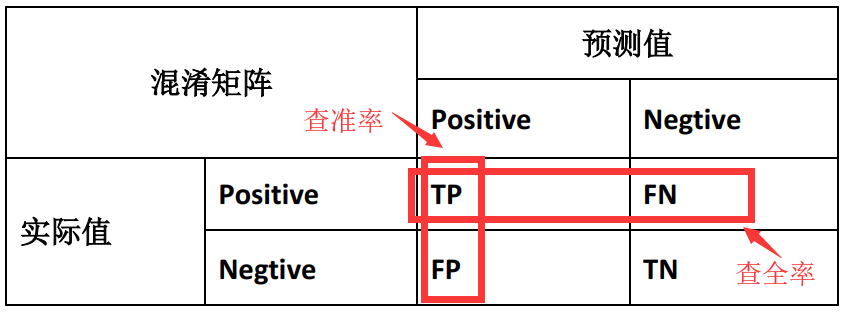
1）查准率

例：在所有我们预测有恶性肿瘤的病人中，实际上有恶性肿瘤的病人的百分比，越高越好。

2）召回率

例，在所有实际上有恶性肿瘤的病人中，成功预测有恶性肿瘤的病人的百分比，越高越好。





（3）权衡查准率和查全率

1）使模型的输出为：

2）调整阈值，eg：0.5、0.7…….，也就是有多大几率，确定y=1

3）将不同阀值情况下，绘制：查全率与查准率的图表，（曲线的形状根据数据的不同而不同）；

4）套用公式： ，选择使F最大的P、R。（P：precision、R：Recall）

附：查准率、查全率的关系图

