第六周

1. 模型选择

问题：如何选择模式，选择多项式的项数。

1. 将总样本数据划分为训练集60%，测试集40%，用训练集对多种模型进行训练，选择误差最小的对测试集的数据进行预测。存在测试集恰好符合的嫌疑，所选择的模型正好也符合测试集，而不一定符合新样本。
2. 将总样本数据划分为训练集60%，交叉验证集20%，测试集20%，先用训练集数据对多模型进行训练，再用验证集对模型进行验证，选择误差最小的，最后用测试集进行预测。

训练集误差方程：



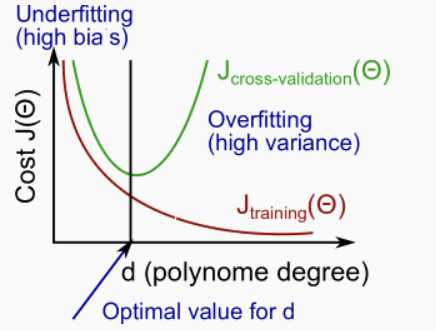
验证集误差方程：



测试集误差方程：



2.高偏差和高方差



高偏差：多项式次数小，训练误差大，验证误差大。

高方差：多项式次数大，训练误差小，验证误差大。

3.正则化参数

参数较小时，对参数的惩罚小，训练误差小，验证集误差大。

参数较大时，对参数的惩罚大，训练误差大，验证集误差大。

步骤如下：

1. 创建一个列表，以2倍的增速。（0.01、0.02、……、10.24）
2. 创建具有不同多项式次数的模型集合。
3. 用结合不同的模型训练出。
4. 使用训练验证集，不使用。
5. 选择验证集误差最小的作为预测。

4.样本数量m

对于高偏差情况（）：

训练集误差会随着m增大而增大，验证集会随着m增大而减小，最终会趋于一致，但误差很大。

对于高方差情况：

训练集误差会随着m增大而增大，验证集会随着m增大而减小，最终会趋于一致，但误差很小。

结论：对于高偏差情况，增大样本数量没有太大效果；对于高方差情况，增大样本数量稍有效果。

综述：总结上面的情况得出

1. 训练集误差和验证集误差都很大时，说明欠拟合，过大，多项式次数d太小。
2. 训练集误差很小，验证集误差很大时，说明过拟合，太小，多项式次数d太大。
3. 样本数量m视情况而定。

5.改善模型：

1）获得更多的样本数据—高方差

2）减少样本特征—高方差

3）获得更多的样本特征—高偏差

4）增加多项式项数—高偏差

5）减小值—高偏差

6）增大值—高方差

6.机器学习设计流程：

1）快速使用算法实现。

2）画出学习曲线，对方差、偏差进行分析，并对误差进行调整。

3）进行误差分析，通过人工检测发现系统不足，增加特征量来改进。

7.数值评估度量：

1）斜偏率是指样本结果一类与另一类的比例差距太大。例如：一个预测的准确率为99%，而直接预测一种结果的准确率为99.5%。需要查准率和召回率来度量。

2）查准率：预测这个结果，最终也是这个结果的概率。

3）召回率：最终是这个结果的样本，我预测是这个结果的概率。

4）希望在预测时，有高的准确率，但同时希望有高的查准率和召回率。所以直接预测结果的方法不合适。

5）F值：是查准率和召回率的权衡值。F值=2\*（PR/（P+R）），P为查准率，R为召回率。

8.大数据原理：

1. 数据量越大，高方差，越不容易出现过拟合。
2. 特征量越多，高偏差，越不容易出现欠拟合。

学习总结：

本周主要学习了如何选择模型、关于方差偏差以及学习曲线的诊断法和机器学习的数值评估标准（查准率、召回率、F值），通过分析改进系统，本周主要是对系统的改进方法。通过本周的编程作业，首先进行了线性回归，然后画出了学习曲线，之后进行多项式回归，并画出学习曲线，最后比较了不同的在训练集和交叉验证集上的效果。