Coursera Machine Learning Week6

Coursera Machine Learning Lunar's note

`MachineLearning` `Coursera`

评估学习算法

评估假设函数Evaluating a Hypothesis

将数据集分为训练集m_{train}和测试集m_{test}(如7:3),针对训练集算出参数。然后根据所得参数计算training error 和 test error。若training error很小而test error很大那么说明过拟合(overfitting)了。

模型选择 model selection

1. 多项式次数-d
   将数据集分为训练集，交叉验证集(Cross Vadidation Set)和测试集(典型划分 3：1：1)。对于不同的d,使用训练集拟合参数d，在CV集中测试其表现，选择有最小交叉测试误差的参数d。这样就可以避免过度拟合训练集。

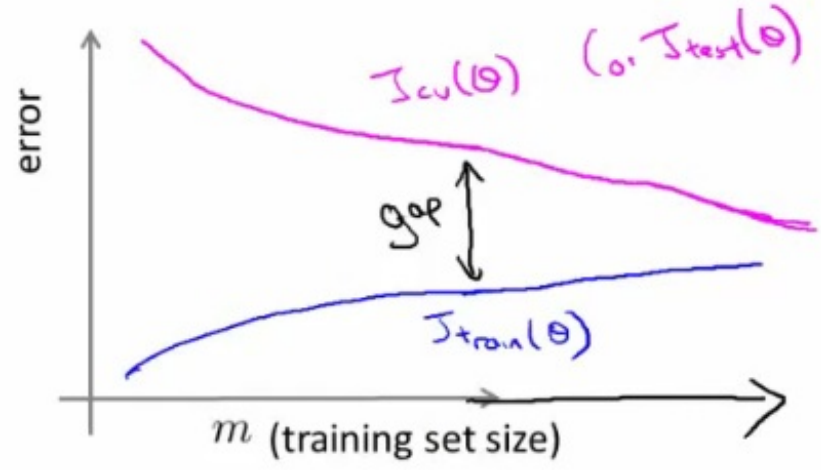
偏差和方差 Bias vs. Variance

- Bias - underfit J_{train}和J_{cv}都过大。
- Variance - overfit
  J_{train}较小，J_{cv}较大。
- 正则化(Regularization)和偏差/方差
  过大的λ会导致underfit，而过小的λ会导致overfit。选择正则化参数λ的方法类似与上面的选的多项式次数d的方法，只是在计算J_{train}和J_{cv}时不引入正则化。

学习曲线 Learning Curves

一个横轴为m（训练集大小），竖轴为error的曲线。

- High bias时，J_{train}和J_{cv}在m越来越大时会相当接近并接近平行，但error都比较高。也就是说High bias时无法通过加大训练集样本数来提高准确度。
- High variance时，J_{train}和J_{cv}中间会有较大间隔，通过增加m可以减小间隔。



总结：改进算法的方法

1. 使用更多的训练样本
2. 减小特征值个数
3. 尝试额外的特征值
4. 提高特征值的幂
5. 增大正则化参数

6. 减小正则化参数

   如何**选择这些方法呢？**

   - High variance (overfitting)

     【1】【2】【5】
   - High bias (underfitting)

     【3】【4】【6】

     本章重点在于交叉验证(CV)和Bias&Variance

误差分析 Error Analysis

机器学习的推荐方法：
1. 选用一个简单的模型并快速实现它。利用交叉检验测试它。
2. 画出学习曲线，判定High bias/variance
3. 误差分析，手工查看结果并分类，分析什么类型的样本会导致错误，并改进它。
通过误差度量值来决定优化方法而不仅是通过直觉。

倾斜类的误差分析

当样本分类比例悬殊时会导致误差度量值的失衡。
例如，癌症预测样本中只有0.5%罹患癌症，那么通过一般的度量值比如预测准确率来衡量，那么99.4%提升到99.5%并不一定是算法真正得到优化，因为简单的预测所有人都没有癌症也可以得到相同的结果。
对于这种“倾斜类(Skewed Classes)”我们要使用不同的误差度量值。
+ 查准率(Precision)
预测m个样本为A，这些样本中有n个实际上为A。那么Precision=n/m。
查准率越高越好。
+ 召回率(Recall)
一个样本集合中实际上有m个样本为A，其中有n个被预测为A。那么Recall = n/m。
召回率越高越好。

Precision 和　Recall的平衡

- 提高置信度的阈值会导致 高Precision 低 Recall。降低置信阈值则相反。

- 在应用中如果需要“宁可错杀三千，也不能放过一个”这种确信的话，可以适当提高召回率。
- 如果追求预测某一类的精准度时，可以提高那类的查准率。
- 通常使用 F score :$2\frac{PR}{P+R}$来进行查准率和召回率的平衡。F socre越高越好。

机器学习中的数据

通常来说数据量越大，得到的机器学习算法越好。
甚至有人说:

> The man who owns more data wins instead of who has better algorithm.

但是需要在以下条件下，大量数据的价值才能得以体现
1. 特征值要包含足够的信息。比如仅仅知道房子的面积而不给其地段的话，再多的数据也白搭。
2. 机器学习模型过于简单，比如只用简单线性回归。