Lecture 16　Three Learning Principles

整理者: LobbyBoy* 2020年2月29日

1. Occam's Razor

The simplest model that fits the data is also the most plausible.

法则一：在保证拟合训练数据效果的前提下，越简单的模型越好。这一经验来源于著名的“奥卡姆剃刀”：Entities must not be multiplied beyond necessity——William of Occam(1287-1347)。

什么是简单的模型(simple model)？有两种阐述：

- simple hypothesis h：某个hypothesis看起来很简单，说明其中的参数较少。例如，低维度的perceptron；
- simple model $\mathcal{H}$：某个假说集合看起来很简单，说明其VC dimension低，包含small number of effective hypotheses。

为什么“simple is better”？因为：简单的$\mathcal{H}$ →小的$m_{\mathcal{H}}(N)$ →less likely to fit data perfectly，完全拟合的概率为$\frac{m_{\mathcal{H}}(N)}{2^N}$ →对于噪声，对于“乱乱的资料”，很难拟合得不错→如果拟合得不错，说明资料的确是有某种规律→显著性。

2. Sampling Bias

If the data Is sampled in a biased way,
learning will produce a similarly biased outcome.

也就是说，如果训练资料是从分布$P_1(\mathbf{x}, y)$中generate出来的，而测试资料是从分布$P_2(\mathbf{x}, y) \neq P_1$中generate出来的，那么VC bound就不成立了——VC的一大假设是：data and testing both iid from P。

*本笔记根据台湾大学林轩田教授于线上教育平台Coursera开设的“机器学习基石”课程整理而成(课程内容见: https://www.coursera.org/learn/ntumlone-mathematicalfoundations/home/welcome)。笔记内的大多数图片来自于林老师的课程slides。感谢林老师能够将如此精彩的课程通过线上平台同所有人分享，thanks!

例如，在构建电影推荐系统的问题中，训练资料与测试资料的产生方式如下：选一批人8年的观影资料，将他们前6年的资料做成training data，将后2年的资料做成testing data。若此，training data与testing data就不再是从同一个分布中独立取样得到的，而是有一个先后顺序，即：用较久远的数据训练，用较新的数据测试。

因此，我们的解决方法是：①在训练中，给时间上靠后的样本以更高的权重；②使用一批时间上靠后的数据作为validation set。

3. Data Snooping

> **If a data set has affected any step in the learning process, its ability to assess the outcome has been compromised.**

在数据预处理的时候，我们经常会对数据进行中心化或标准化这样的放缩处理(shifting-scaling)。如果我们在放缩时，用到了testing set的数据，那么就发生了data snooping。例如，在标准化时，用的均值和标准差都是在training set+testing set上计算出来的。正确做法是，用training set上计算出的均值和标准差进行标准化，在测试时，将从training set上计算出的均值和标准差同样地运用于testing set上进行标准化。

4. Power of Three

与Machine Learning有关的三个概念：

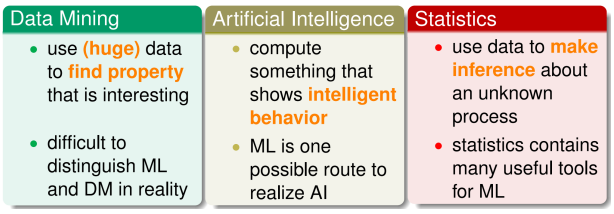

图 1: Three Related Fields

三种Bounds：

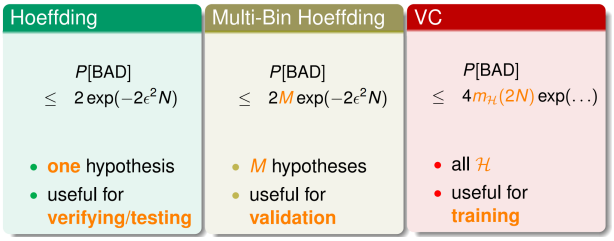

图 2: Three Theoretical Bounds

三种线性模型：

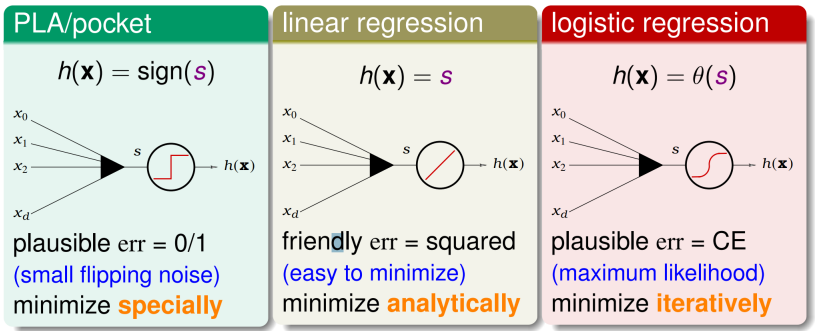

图 3: Three Linear Models

三种补充工具：

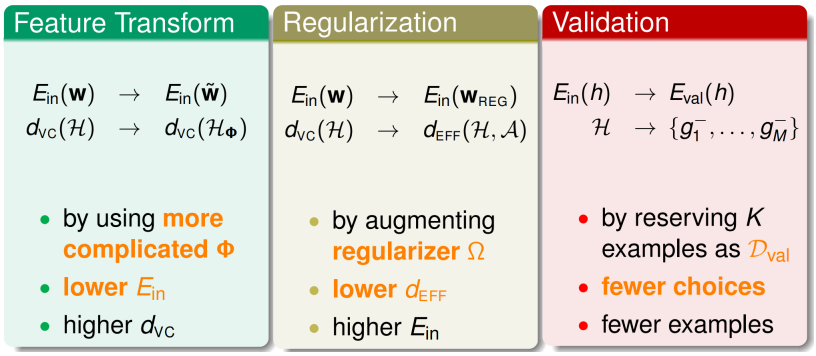

图 4: Three Key Tools