分类模型

授课教师：吴翔

邮箱：wuhsiang@hust.edu.cn

March 18-25, 2019

统计学习概述

统计学习方法

统计机器学习（statistical machine learning）可分为：

- 有监督学习（supervised learning）vs 无监督学习（unsupervised learning）：聚类分析即为典型的无监督学习
- 参数方法（parametric methods）vs 非参数方法（non-parametric methods）
- 回归（regression）问题 vs 分类（classification）问题：分别针对连续变量和分类变量

测试均方误差的分解

测试均方误差的期望值（expected test MSE）可以分解为如下三个部分：

$$E(y-\hat{f}(x))^2 = \underbrace{\mathrm{Var}(\hat{f}(x))}_{\text{variance}} + \underbrace{[\mathrm{Bias}(\hat{f}(x))]^2}_{\text{bias}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{irreducible}} .$$

- 模型方差（variance）：针对不同的训练数据，$\hat{f}$ 的变化程度。
- 模型偏误（bias）：通过相对简化的模型来**近似**真实世界的问题时所引入的误差。

权衡模型偏误与方差

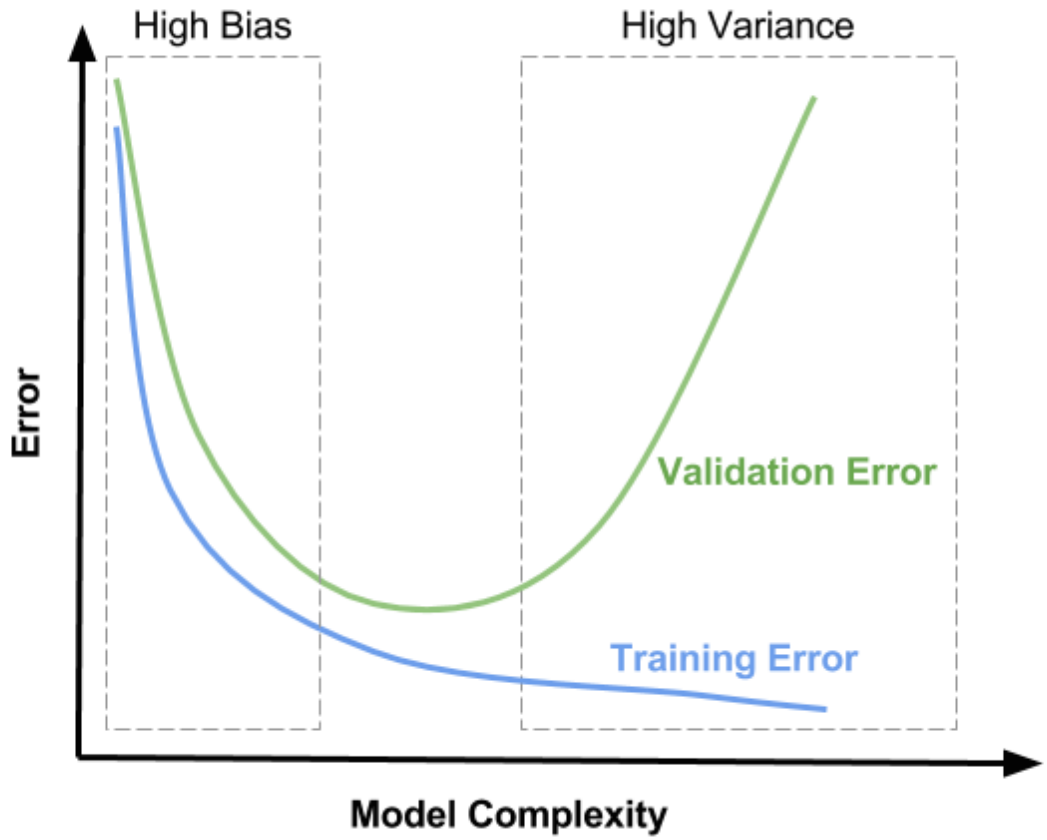

图 1: bias-variance trade-off

如何选择统计模型?

- 传统统计模型的局限：线性回归模型等统计模型通常最小化训练数据的均方误差，但是其测试均方误差（test MSE）却较大。换言之，传统统计模型执着于寻求“真实规律”，以致于将一些随机因素**误判**为 f 的真实性质。
- 权衡模型偏误与方差（bias-variance trade-off）：随着模型灵活性（或自由度）的增加，模型方差随之增大，但模型偏误则相应减小（过度拟合问题）。通过交叉验证（cross-validation）来实现两者的权衡。
- 权衡预测精度与可解释性（accuracy-interpretability trade-off）：诸如 bagging、boosting、support vector machines 等非线性模型具有很高的预测精度，但不易解释；linear models 等易于解释，但预测精度不高。两者的权衡取决于研究目的。

交叉验证

分类模型概述

1. 基本分类模型
   - 逻辑斯蒂回归
   - 线性判别分析（linear discriminant analysis, LDA）：包括 LDA 和 QDA
   - K 最近邻（$K-$nearest neighbor）
2. 树模型（tree-based models）
   - 决策树
   - 装袋法（bagging）
   - 随机森林（random forest）
   - 提升法（boosting）
3. 支持向量机（support vector machine, SVM）
4. 聚类模型（clustering models）
   - K 均值聚类（$K-$means clustering）
   - 系统聚类（hierarchical clustering）

基本分类模型

逻辑斯蒂回归

树模型

决策树

支持向量机

支持向量机

聚类模型

K 均值聚类