

# 分类模型

授课教师：吴翔

邮箱：wuhsiang@hust.edu.cn

March 18-25, 2019

1 统计学习概述

2 基本分类模型

3 聚类模型

4 树模型

5 支持向量机

# 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：

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

- 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

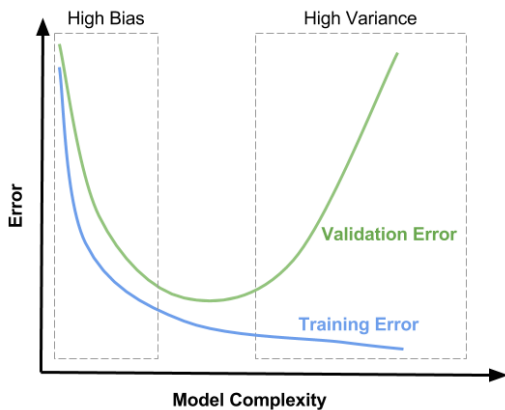


图 1: bias-variance trade-off

## 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

交叉验证将原始数据集分为训练集 (training set) 和验证集 (validation set), 并以验证集的误差率选择最佳模型。

- 留一交叉验证法 (leave-one-out cross validation, LOOCV)
- $k$  折交叉验证法 ( $k$ -fold CV): 将观测集随机分为  $k$  个大小基本一致的组, 或说折 (fold)。每次选取其中一折作为验证集, 而剩余  $k - 1$  折作为训练集。通常, 取  $k = 5$  或  $k = 10$ 。

分类模型验证集误差率:

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^k \text{Err}_k = \frac{1}{k} \sum_{i=1}^k \frac{1}{m_k} \sum_{i=1}^{m_k} I(y_i \neq \hat{y}_i).$$



# 分类模型概述

预测分类响应变量 (categorical response variable):

- ① 基本分类模型 (basic classifier)
- ② 树模型 (tree-based models)
- ③ 支持向量机 (support vector machine, SVM)
- ④ 聚类模型 (clustering models)

# 分类模型的评价

## Confusion Matrix and ROC Curve

		Predicted Class	
		No	Yes
Observed Class	No	TN	FP
	Yes	FN	TP

TN      True Negative  
 FP      False Positive  
 FN      False Negative  
 TP      True Positive

## Model Performance

Accuracy =  $(TN+TP)/(TN+FP+FN+TP)$

Precision =  $TP/(FP+TP)$

Sensitivity =  $TP/(TP+FN)$

Specificity =  $TN/(TN+FP)$

图 2: confusion matrix

# ROC 曲线

# AUC

# 基本分类模型

## 基本分类模型 (basic classifier)

- ① 逻辑斯蒂回归 (logistic regression)
- ② 线性判别分析 (linear discriminant analysis, LDA)
- ③ 二次判别分析 (quadratic discriminant analysis, QDA)
- ④  $K$  最近邻 ( $K$ -nearest neighbor, KNN)

# logistic 回归

给定  $X$  条件下事件  $Y$  发生的概率  $p(X) = \Pr(Y = 1|X)$ , 据此可以将发生比 (odd) 的对数建模为  $X$  的线性函数

$$\log\left[\frac{p(X)}{1 - p(X)}\right] = \beta X.$$

上式左侧称为对数发生比 (log-odd) 或分对数 (logit), 其取值范围在  $(-\infty, \infty)$ 。

当类别  $K \geq 2$  时, 则采用多类别 logistic 回归模型。

# LDA



# QDA

# KNN

# 聚类模型

# 聚类模型 (clustering models)

- ①  $K$  均值聚类 ( $K$ -means clustering)
- ② 系统聚类 (hierarchical clustering)

# $K$ 均值聚类

# 树模型

# 树模型 (tree-based models)

- ① 决策树
- ② 装袋法 (bagging)
- ③ 随机森林 (random forest)
- ④ 提升法 (boosting)

# 决策树



# 支持向量机

# 支持向量机