数据科学导论第 3 讲：统计学习三要素

王小宁

中国传媒大学数据科学与智能媒体学院

2025 年 03 月 24 日

目录

统计学习方法

统计学习三要素

$$Method = Model + Policy + Algorithm$$

- 模型 (Model)：
  1. 决策函数的集合 $\mathscr{F} = \{f|Y = f(x)\}$。
  2. 参数空间 $\mathscr{F} = \{f|Y = f_\theta(x), \theta \in R^n\}$。
  3. 条件概率集合：$\mathscr{F} = \{P|P(Y|X)\}$。
  4. 参数空间：$\mathscr{F} = \{P|P_\theta(Y|X), \theta \in R^n\}$。

统计学习三要素

- 策略（Policy）
- 损失函数（loss function）：一次预测的好换
- 风险函数：平均意义下模型预测的好坏

1. 0-1 损失函数：

$$L(Y, f(X)) = \begin{cases} 1, Y \neq f(X) \\ 0, Y = f(X) \end{cases}$$

2. 平方损失函数：$L(Y, f(X)) = (Y - f(X))^2$
3. 绝对损失函数：$L(Y, f(X)) = |Y - f(X)|$

统计学习三要素

- 策略（Policy）
- 对数损失函数 (logarithmic loss function)
- 对数似然损失函数 (loglikelihood loss function)
- 损失函数的期望 $L(Y, P(Y|X)) = -logP(Y|X)$,

$$R_{exp}(f) = E_p[L(Y, f(X))] = \int_{X \times Y} L(y, f(x))P(x, y)dxdy$$

- 风险函数 (risk function) 期望损失 (expected loss），由 P(x,y) 可以直接求出 P(x|y), 但不知道 $T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$
- 经验风险 (empirical risk), 经验损失 (empirical loss)

$$R_{emp}(f) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i))$$

统计学习三要素

- 策略（Policy）：经验分线最小化与结构风险最小化
- 经验风险最小化最优模型

$$min_{f \in \mathscr{F}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i))$$

- 当样本容量很小时，经验风险最小化学习的效果未必很好，会产生“过拟合 over-fitting” 结构风险最小化 (structure risk minimization)
- 为防止过拟合提出的策略，等价于正则化（regularization），加入正则化项 (regularize 或罚项 (penalty term)：

$$R_{emp}(f) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f)$$

统计学习三要素

- 求最优模型就是求解最优化问题：

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f)$$

统计学习三要素

- 算法：指解题方案的准确而完整的描述，是一系列解决问题的清晰指令，算法代表着用系统的方法描述解决问题的策略机制。
- 能够对一定规范的输入，在有限时间内获得所要求的输出。如果一个算法有缺陷，或不适合于某个问题，执行这个算法将不会解决这个问题。
- 不同的算法可能用不同的时间、空间或效率来完成同样的任务。
- 一个算法的优劣可以用**空间复杂度**与**时间复杂度**来衡量。

1. 如果最优化问题有显式的解析式，算法比较简单
2. 但通常解析式不存在，就需要**数值计算**的方法

模型评估与模型选择

- 训练误差，训练数据集的平均损失

$$R_{emp}(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

- 测试误差，测试数据集的平均损失

$$r_{test} = \frac{1}{M}\sum_{i=1}^{M} L(y_i, \hat{f}(x_i))$$

- 损失函数是 0-1 损失时：

$$e_{test} = \frac{1}{M}\sum_{i=1}^{M} I(y_i \neq \hat{f}(x_i))$$

- 测试数据集的准确率：

$$r_{test} = \frac{1}{M}\sum_{i=1}^{M} I(y_i = \hat{f}(x_i))$$

模型评估与模型选择

- 过拟合与模型训练

1. 假定给定训练数据集 $T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$

$$f_M(x, w) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M X^m = \sum_{j=1}^{M} w_j x^j$$

2. 经验风险最小

$L(w) = \frac{1}{2}\sum_{j=1}^{N}(f(x_i, w) - y_j)^2$, $L(w) = \frac{1}{2}\sum_{j=1}^{N}(\sum_{j=1}^{M} w_j x^j - y_j)^2$ 其中 $w_j = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^{j+1}}, j = 0, 1, 2 \cdots, M$

模型评估与模型选择

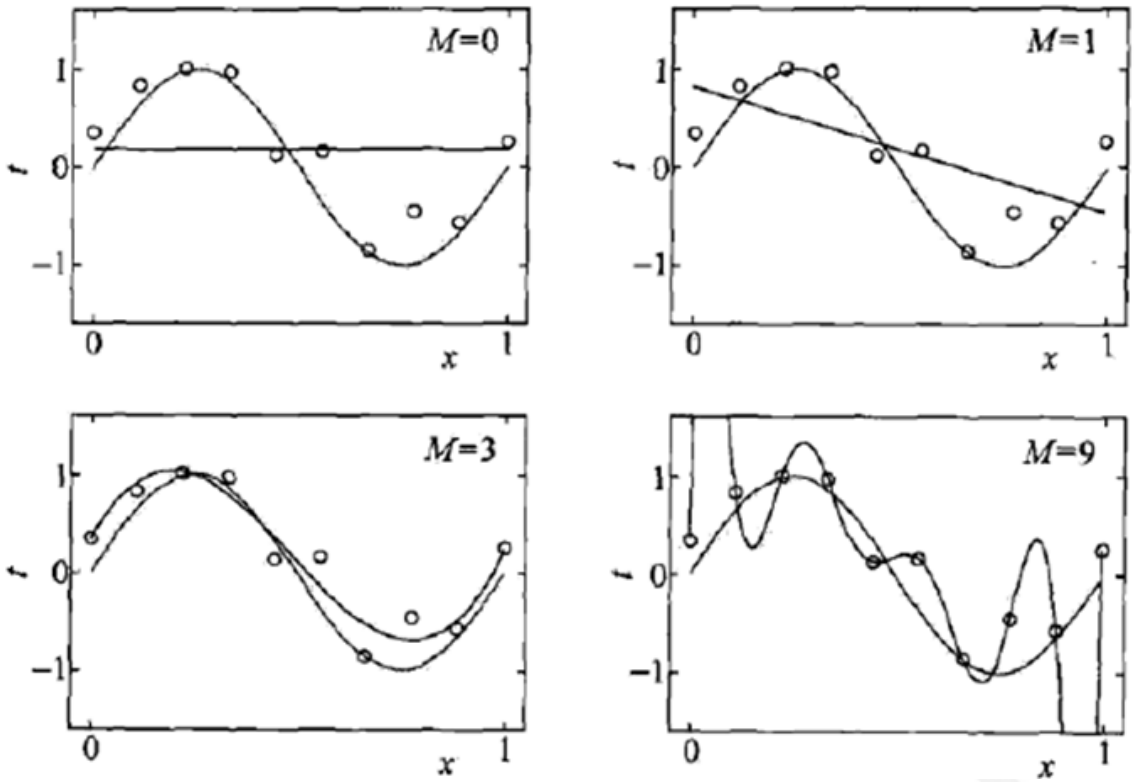

图 1: Model selection

模型评估与模型选择

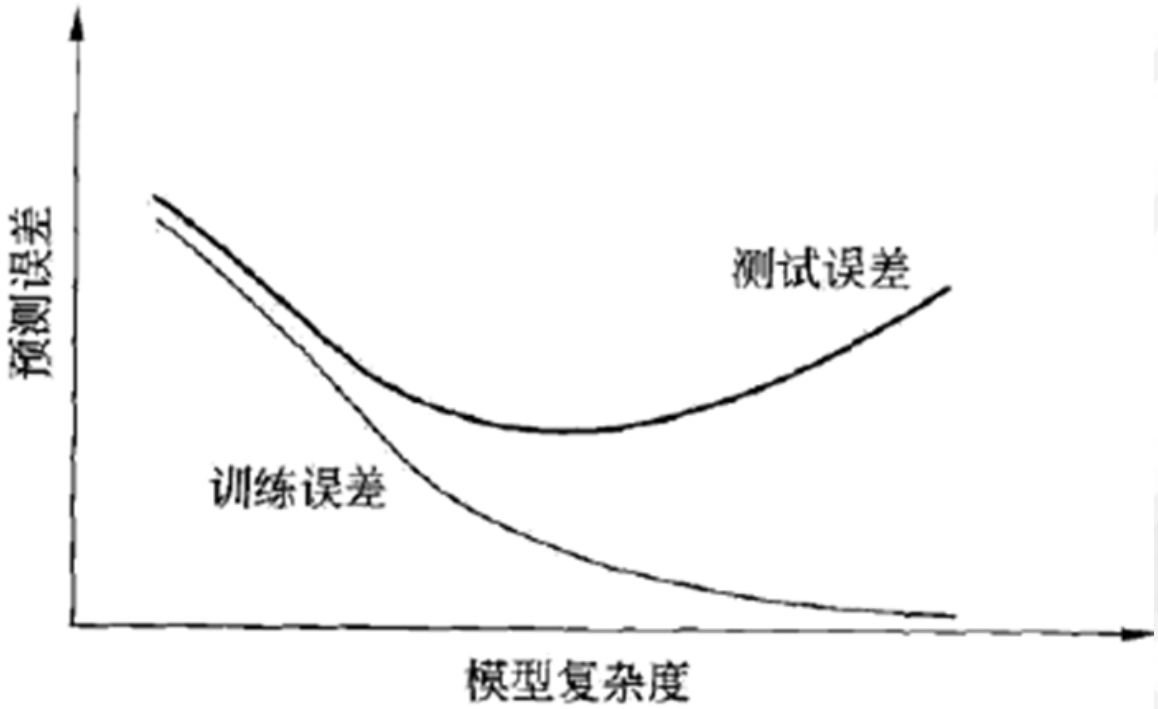

图 2: Model selection

正则化与交叉验证

- 正则化一般形式：

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f)$$

- 回归问题中：

1. 岭回归（Ridge Regression）

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \frac{\lambda}{2} ||w||^2$$

2. Lasso 回归（Lasso Regression）

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \frac{\lambda}{2} ||w||_1$$

交叉验证

- 训练集 (training set): 相当于教材，用于训练模型
- 验证集 (validation set): 相当于练习册，用于模型选择
- 测试集 (test set): 相当于期末考试，用于最终对学习方法的评估

1. 简单交叉验证
2. S 折交叉验证
3. 留一交叉验证

泛化能力 (generalization ability)

- 泛化误差（generalization error)

$$R_{exp}(\hat{f}) = E_p[L(Y, \hat{f}(X))] = \int_{X \times Y} L(y, \hat{f}(x)) P(x, y) dxdy$$

- 泛化误差上界
  1. 比较学习方法的泛化能力——比较泛化误差上界
  2. 性质：样本容量增加，泛化误差趋于 0，假设空间容量越大，泛化误差越大

泛化能力

- 经验风险最小化函数：

$$f_N = \underset{f \in \mathcal{F}}{argmin} \hat{R}(f)$$

- 泛化能力：$R(f_N) = E[L(Y, f_N(X))]$
- 定理：泛化误差上界，二分类问题
  1. 当假设空间是有限个函数的结合
  2. 对任意一个函数 f，至少以概率 $1-\delta$，以下不等式成立：

$$R(f) \leq \hat{R}(f) + \varepsilon(d, N, \delta)$$

其中 $\varepsilon(d, N, \delta) = \sqrt{\frac{1}{2}(logd + log(\frac{1}{\delta}))}$

泛化能力

- 经验风险最小化函数：

$$f_N = \underset{f \in \mathcal{F}}{argmin} \hat{R}(f)$$

- 泛化能力：$R(f_N) = E[L(Y, f_N(X))]$
- 定理：泛化误差上界，二分类问题
  1. 当假设空间是有限个函数的结合
  ，
  2. 对任意一个函数 f，至少以概率 $1-\delta$，以下不等式成立：

$$R(f) \leq \hat{R}(f) + \varepsilon(d, N, \delta)$$

其中 $\varepsilon(d, N, \delta) = \sqrt{\frac{1}{2}(logd + log(\frac{1}{\delta}))}$

生成模型

- 监督学习的目的就是学习一个模型
- 决策函数：$Y = f(X)$
- 条件概率分布：$P(Y|X)$
- 生成方法 (Generative approach) 对应生成模型 (generative model)
- 朴素贝叶斯法和隐马尔科夫模型

$$P(Y|X) = \frac{P(X,Y)}{P(X)}$$

判别模型

- 判别方法 (Discriminative approach) 由数据直接学习决策函数 $f(X)$ 或条件概率分布 $P(Y|X)$ 作为预测的模型，即判别模型 (discriminative model)

$$Y = f(X)$$

- K 近邻法、感知机、决策树、logistic 回归模型、最大熵模型、支持向量机、提升方法和条件随机场。

$$P(Y|X)$$

生成模型与判别模型

- 各自优缺点：

1-1. 生成方法：可还原出联合概率分布 P(X,Y)

1-2. 判别方法不能。生成方法的收敛速度更快，当样本容量增加的时候，学到的模型可以更快地收敛于真实模型；当存在隐变量时，仍可以使用生成方法，而判别方法则不能用。

2-1. 判别方法：直接学习到条件概率或决策函数，直接进行预测，往往学习的准确率更高；

2-2. 由于直接学习 Y=f(X) 或 P(Y|X)，可对数据进行各种程度上的抽象、定义特征并使用特征，因此可以简化学习过程

R

需要安装软件

- 要使用这一模板，需要安装以下软件：
  - Texlive (为了能编译出 beamer)
  - R
  - Rstudio (为了使用 Rmarkdown)
- 此外，在 R 中还需要安装以下包：
  - knitr (为了编译 Rmarkdown)
  - rticles (支持中文)
  - tinytex (轻量级的 LaTeX)

本周推荐

1. 2 本书：《女士品茶-20 世纪统计怎样变革了科学》，中国统计出版社，2004；《模型思维》，浙江人民出版社
2. 一部电影：《Infinite Secrets:The Genius of Archimedes(阿基米德的秘密)》

End!