北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

统计学习及监督学习概论

DSSC
Data Science and Service Center

分享人：zyj
2019/6/20

预备知识

- 统计学习定义：关于计算机基于**数据**构建**统计模型**并运用模型对数据进行**预测和分析**的学科。
- 统计学习对象：数据
- 统计学习假设：数据具有一定的**统计规律性**
- 统计学习目的：对数据的预测与分析

预备知识

- 分类：
  - Supervised Learning
  - Unsupervised Learning
  - Semi-supervised Learning
  - Reinforcement Learning

监督学习

- 监督学习
  - 训练集

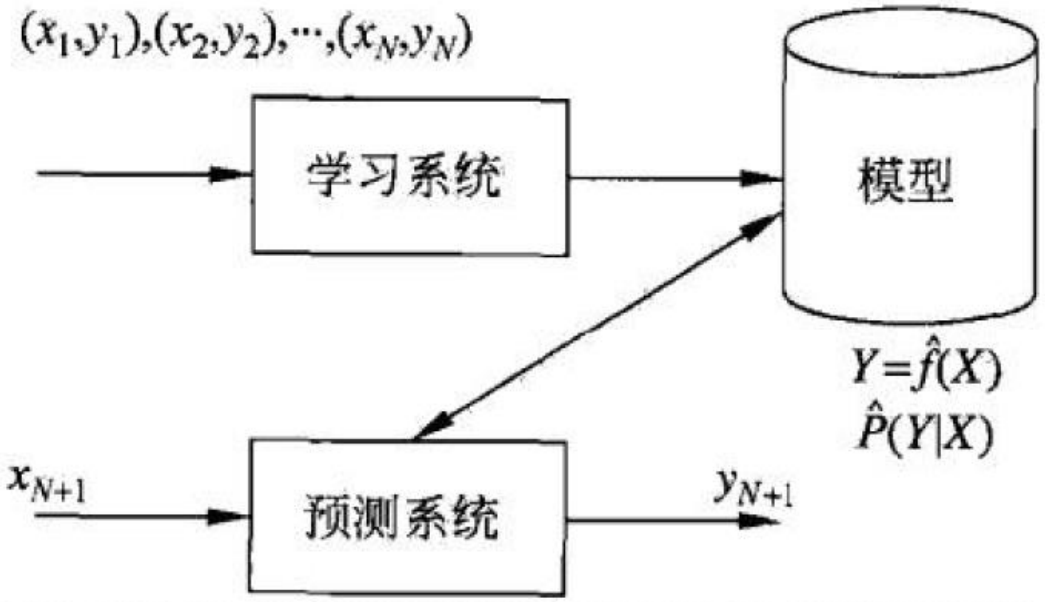
$$\mathrm{T} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

  - 输入空间、输出空间
  - 假设：
    - 训练数据与测试数据依照联合概率分布 $P(X, Y)$ 独立同分布产生的。
  - 目的：学习一个输入空间到输出空间的映射，即模型。
  - 概率模型 $P(y|x)$ 决策函数 $\mathrm{y} = f(x)$
  - 假设空间：即所有模型的集合。

- 监督学习
  - 分类问题、回归问题、标注问题。

监督学习

- 问题形式化



$$y_{N+1} = \arg\max_{y_{N+1}} \hat{P}(y_{N+1} \mid x_{N+1})$$

$$y_{N+1} = \hat{f}(x_{N+1})$$

统计学习三要素

- 统计学习方法=模型+策略+算法
- 模型
  - 决策函数集合 $\mathcal{F} = \{f | f_\theta(X), \theta \in R^n\}$
  - 条件概率集合 $\mathcal{F} = \{P | P_\theta(Y|X), \theta \in R^n\}$

统计学习三要素

- 方法=模型+策略+算法
- 策略

损失函数：一次预测的好坏

风险函数：平均意义下模型预测的好坏

0-1损失函数 0-1 loss function

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

平方损失函数 quadratic loss function

$$L(Y, f(X)) = (Y - f(X))^2$$

绝对损失函数 absolute loss function

$$L(Y, f(X)) = |Y - f(X)|$$

统计学习三要素

- 方法=模型+策略+算法
- 策略
  - 风险函数（期望损失）

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

  - 经验风险（经验损失）

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

统计学习三要素

- 方法=模型+策略+算法
- 策略：经验风险最小化与结构风险最小化
  - 经验风险最小化

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

    - 过拟合
  - 结构风险最小化

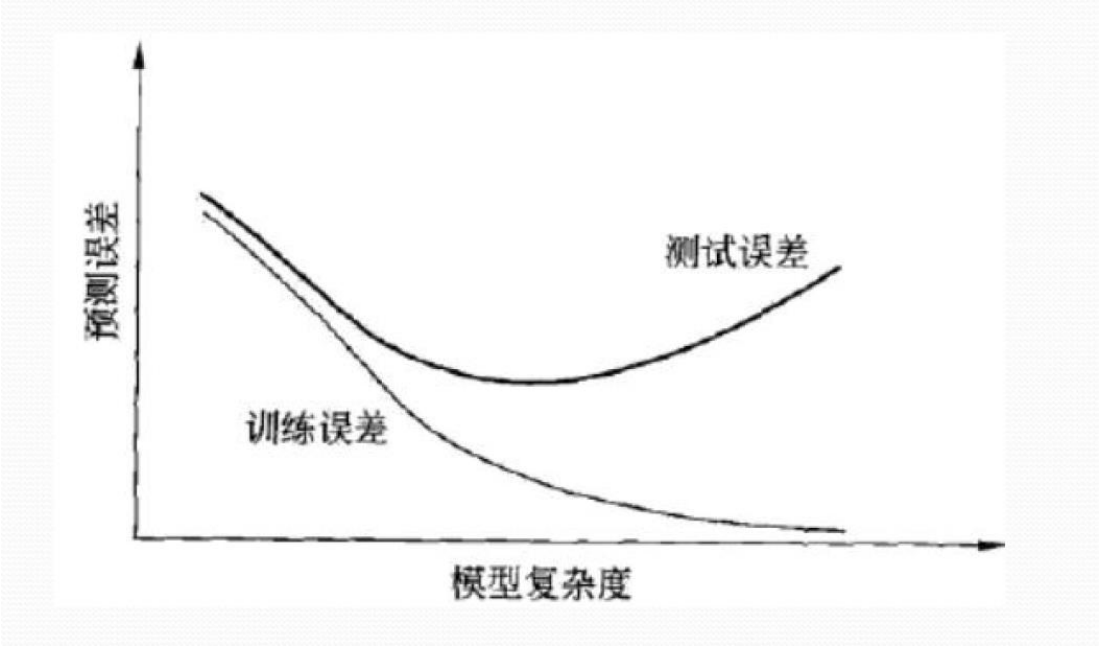
$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f)$$

统计学习三要素

- 方法=模型+策略+算法
- 算法
  - 求解模型即求解最优化问题
    - 存在解析解
    - 不存在解析解

模型评估

- 训练误差、测试误差



交叉验证

- 数据集划分
  - 训练集 training set
  - 验证集 validation set
  - 测试集 test set
- 交叉验证
  - 简单交叉验证
  - S折交叉验证
  - 留一交叉验证

泛化能力

- 泛化误差

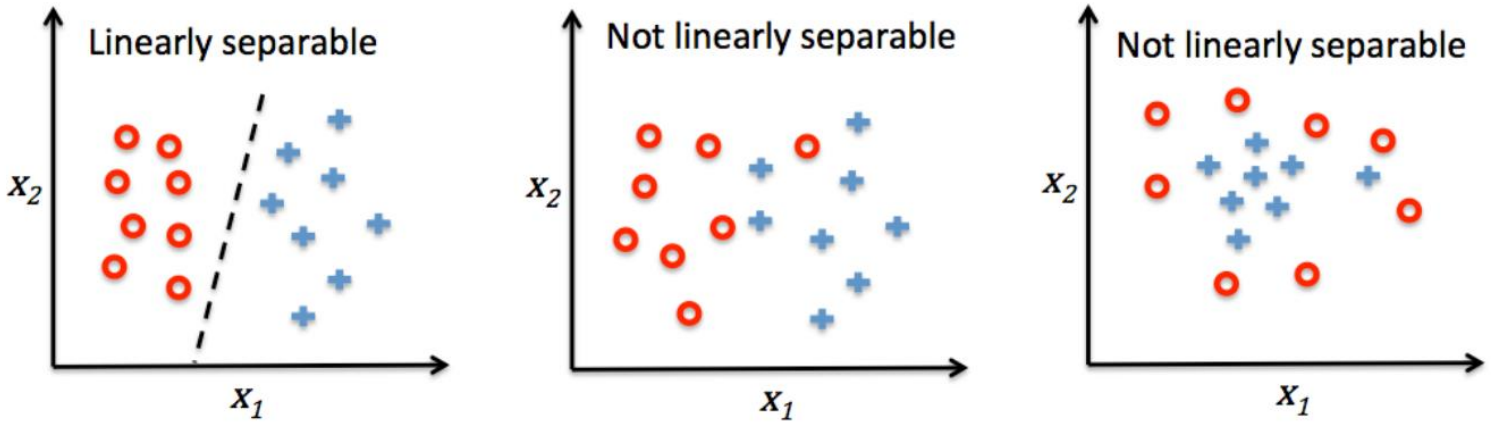
$$R_{exp}(f) = E_p[L(Y, f(x)] = \int_{X \times Y} L\big(y, f(x)\big) P(x, y)\, \boldsymbol{d}x\, \boldsymbol{d}y$$

- 泛化误差上界定理
  - 当假设空间是有限个函数的集合$\mathcal{F} = \{f_1, f_2, \dots, f_d\}$,对任意一个函数 f，至少以概率$1 - \delta$，以下不等式成立：

$$R_{exp}(f) \leq R_{emp}(f) + \epsilon(d, N, \delta)$$

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

- 线性可分数据集


Linearly separable
x_2
x_1
Not linearly separable
x_2
x_1
Not linearly separable
x_2
x_1

北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

谢谢！

DSSC
Data Science and Service Center