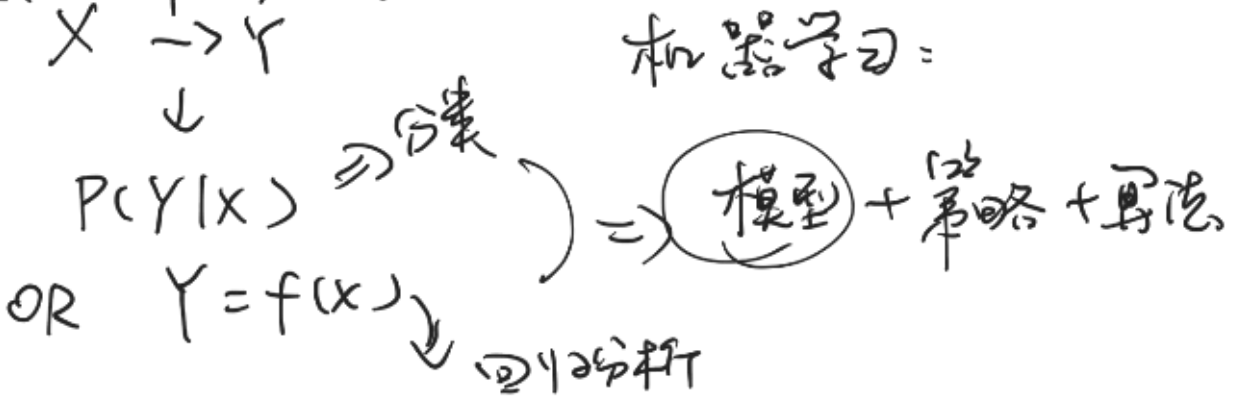
Chap01. 统计方法概论

I. 统计学习要素



1. 模型 => hypothesis space => 一般是无限的

① $\mathcal{F}=\{f \mid Y=f(X)\} \Rightarrow \mathcal{F}=\{f \mid Y=f_\theta(X), \theta \in R^n\}$

② $\mathcal{F}=\{P \mid P(Y|X)\} \Rightarrow \mathcal{F}=\{P \mid P_\theta(Y|X), \theta \in R^n\}$

①：决策函数，表示非概率模型

②条件概率，表示概率模型

2. 策略 => 怎么在假设空间中选取最优模型

(1). 损失函数和风险函数 [模型的好坏]

Loss function & cost function

① 0-1 loss function (LF): $L(Y,f(X))=\begin{cases}1, & Y \neq f(X)\\ 0, & Y=f(X)\end{cases}$

② Quadratic [illegible]

② [illegible] LF : $L(Y,f(x)) = (Y-f(x))^2$

③ absolute LF : $L(Y,f(x)) = |Y-f(x)|$

④ log LF : $L(Y,f(x)) = -\log P(Y|X)$

条件概率模型

Expected Loss :

$$R_{exp} = \iint_{X\times Y} \underbrace{P(x,y)}_{\text{未知}} L(Y,f(x))\,dx\,dy$$

Empirical Loss :

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

$N \to \infty$

(2). Empirical Risk Minimization (ERM)

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

N 较小时 over-fitting 易过

Structural Risk Minimization (SRM) Regularizer

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

$J(f)$ 为模型的复杂度 : f 越复杂, $J(f)$ 越大

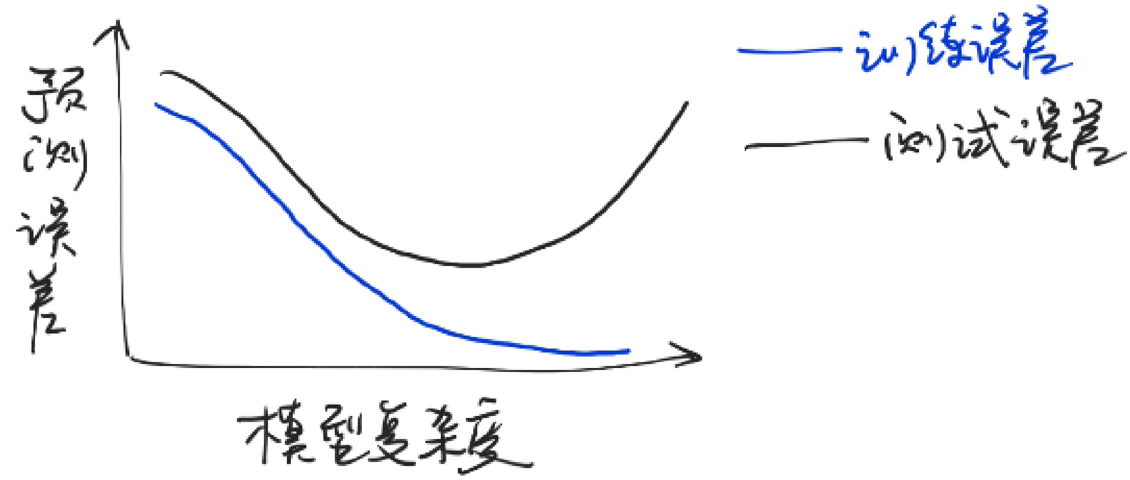
如何表征复杂度?

$$f(x) \Rightarrow \min_{f\in F} R_{srm}(f)$$

3. 算法 : 学习模型的具体方法, [高效求解全局最优解]

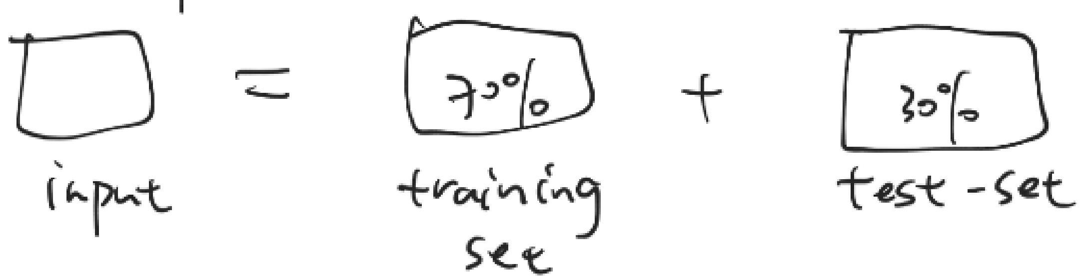
Ⅱ、模型选择与正则化，交叉验证

Over-fitting：所选模型复杂度(参数个数)比真实模型高，导致，虽然测试误差低，但是预测误差大.

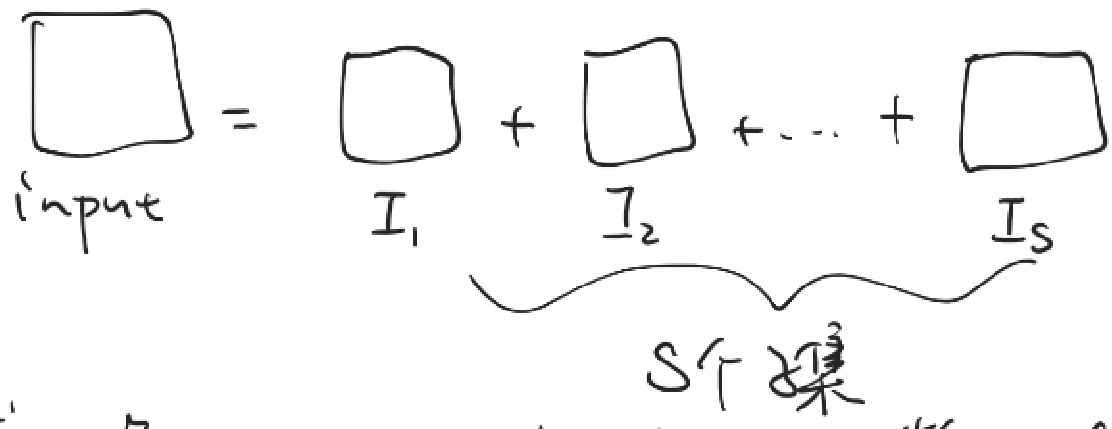


交叉验证：cross-validation

① simple cross-validation



② S-fold cross-validation (应用较多)



依次将 $T_1, T_2, \cdots T_S$ 作为 test set，其余 $S-1$ 个作为 training set. 每次的平均损失为 $L_i[f]$

则最后所选模型 $f \Rightarrow \min_{f \in F} \frac{1}{S} \sum_{i=1}^{S} L_i[f]$

③ leave-one-out cross-validation ($S=N$)

五. 生成模型与判别模型

1. 生成模型：(generative model)

数据学习联合概率分布 $P(X,Y)$，然后求出条件概率分布 $P(Y|X)$ 作为预测的模型.

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

例如：朴素贝叶斯法和隐马尔可夫模型

2. 判别模型：(discriminative model)

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.

例如：K 邻近算法，感知机，决策树，逻辑斯谛回归模型，最大熵模型，支持向量机，提升方法和条件随机场.

IV. 监督学习问题

1. 分类问题 = { Y为有限个离散值
X可为连续或离散
f(x) 或 P(Y|X) → 分类器
0-1 loss function

K-近邻法，感知机，朴素贝叶斯法，决策树，决策列表，logistic 回归模型，支持向量机，提升方法，贝叶斯网络，神经网络，Winnow……

2. 标注问题 { Y为标记序列
X为观测序列
$P(Y^{(1)}, Y^{(2)}, \cdots, Y^{(n)} | X^{(1)}, X^{(2)}, \cdots X^{(n)})$

隐马尔可夫模型，条件随机场

3. 回归问题 ⟺ 函数拟合 ⇒ $Y = f(X)$

Loss function：常用为 MEAN-SQUARED-LOSS
可用最小二乘法求解 $f(x)$

上次修改：22:37