统计学习方法

第一章

梁秋实

1.1 统计学习

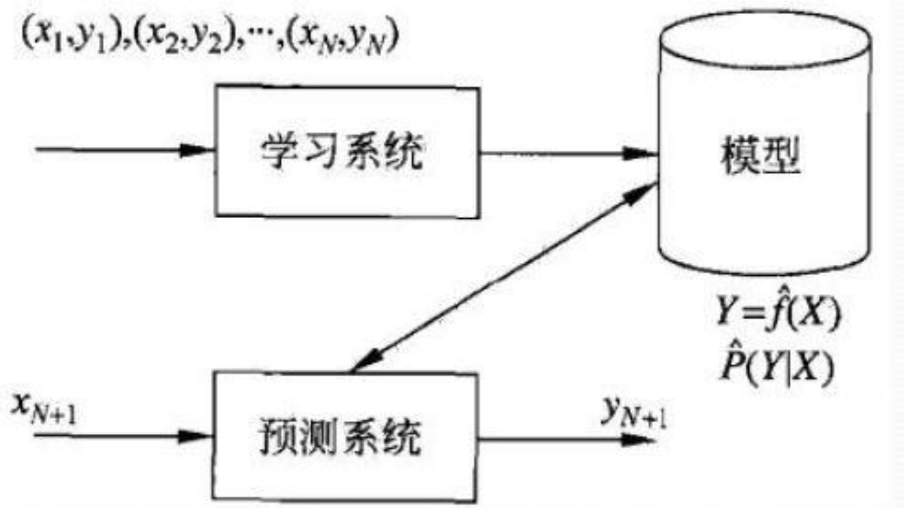
- 对象
  - 数据
  - 基本假设：同类数据具有一定的统计规律性
- 目的：对数据进行预测与分析
- 方法：监督学习、非监督学习、半监督学习、强化学习
- 研究方面：统计学习方法、统计学习理论、统计学习应用
- 重要性
  - 处理海量数据的有效方法；计算机智能化的有效手段；计算机科学发展的一个有效组成部分

1.2 监督学习（1）

- 输入/输出变量、输入/输出空间
- 实例、特征向量、特征空间
  - 实例 $x = \left(x^{(1)}, x^{(2)}, \ldots, x^{(i)}, \ldots, x^{(n)}\right)^T$
  - $x_i = \left(x_i{}^{(1)}, x_i{}^{(2)}, \ldots, x_i{}^{(i)}, \ldots, x_i{}^{(n)}\right)^T$
  - X 、 Y 联合概率分布P(X, Y)
- 样本（样本点）：(x_i, y_i),$i = 1,2, \ldots, N$
- 训练集、测试集
  - $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$
- 回归问题、分类问题、标注问题

1.2 监督学习（2）

- 假设空间
  - 由输入空间到输出空间的映射集合
  - 条件概率分布$P(Y|X)$或决策函数$Y = f(X)$
- 问题的形式化



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

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

1.3 统计学习三要素（1）

- 方法=模型+策略+算法
- 模型：所要学习的条件概率分布或决策函数
- 策略：按照什么样的准则学习或选择最优的模型
  - 损失函数：度量模型一次预测的好坏
  - 风险函数：度量平均意义下模型预测的好坏
- 算法：考虑用什么样的计算方法求解最优模型

1.3 统计学习三要素（2）——策略

- 损失函数
  - 0-1损失函数 $L(Y,f(X))=\begin{cases}1 & Y\neq f(X)\\ 0 & Y=f(X)\end{cases}$
  - 平方损失函数 $L(Y,f(X))=(Y-f(X))^2$
  - 绝对损失函数 $L(Y,f(X))=|Y-f(X)|$
  - 对数损失函数（对数似然损失）$L(Y,P(Y|X))=-\log P(Y|X)$
  - 损失函数的期望（风险函数/期望损失）

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

1.3 统计学习三要素（3）——策略

- 经验风险（经验损失）
  - 给定一个训练数据集$T=\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$，模型关于训练数据集的平均损失，记作$R_{emp}$。
  - $R_{emp}(f)=\frac{1}{N}\sum_{i=1}^{N}L(y_i,f(x_i))$
  - 经验风险最小化（ERM）——经验风险最小的模型就是最优的模型

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

1.3 统计学习三要素（4）——策略

- 结构风险
  - 在经验风险上加上表示模型复杂度的正则化项或罚项
  - $R_{srm}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f)$ （$\lambda \geq 0$）
  - 结构风险最小化——结构风险最小的模型是最优的模型

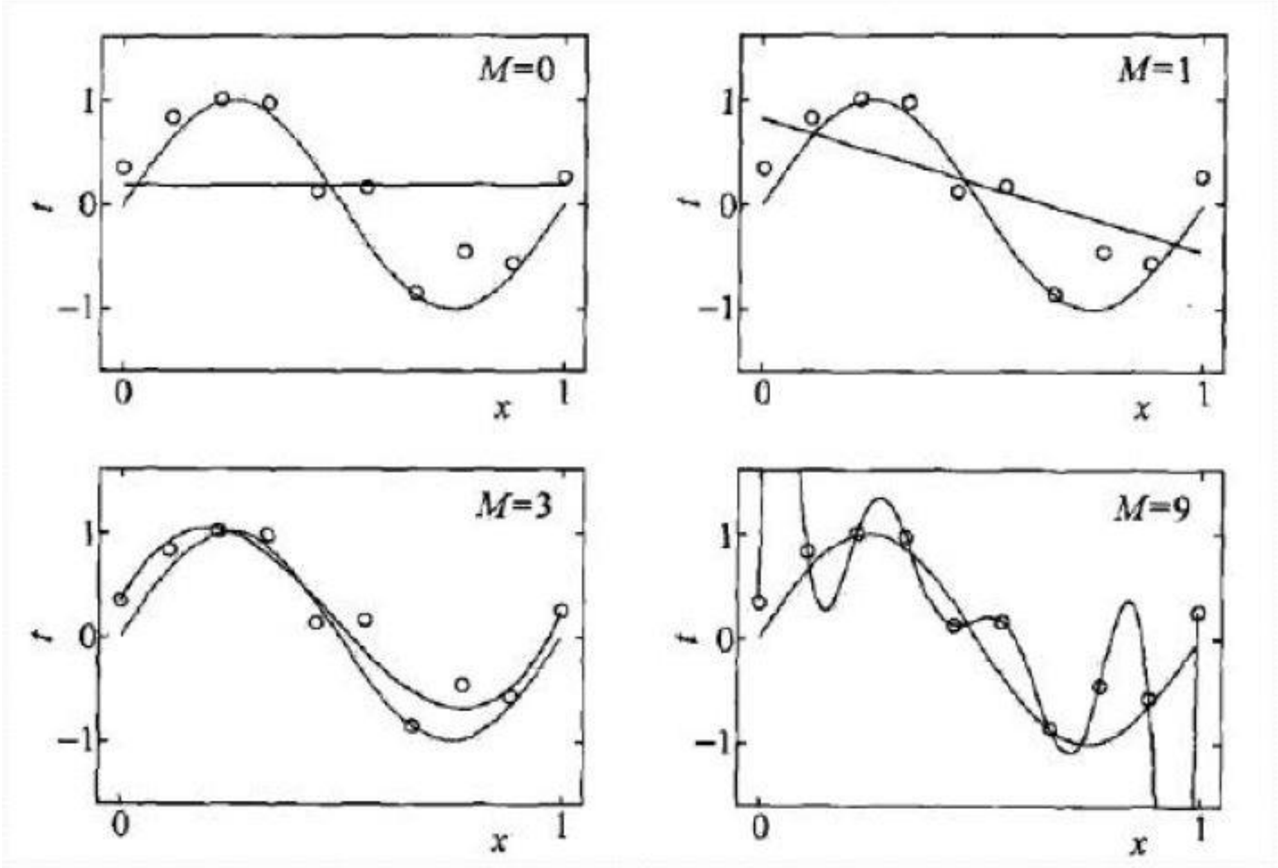
$$\min_{f \in \mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f)$$

1.4 模型评估与模型选择（1）

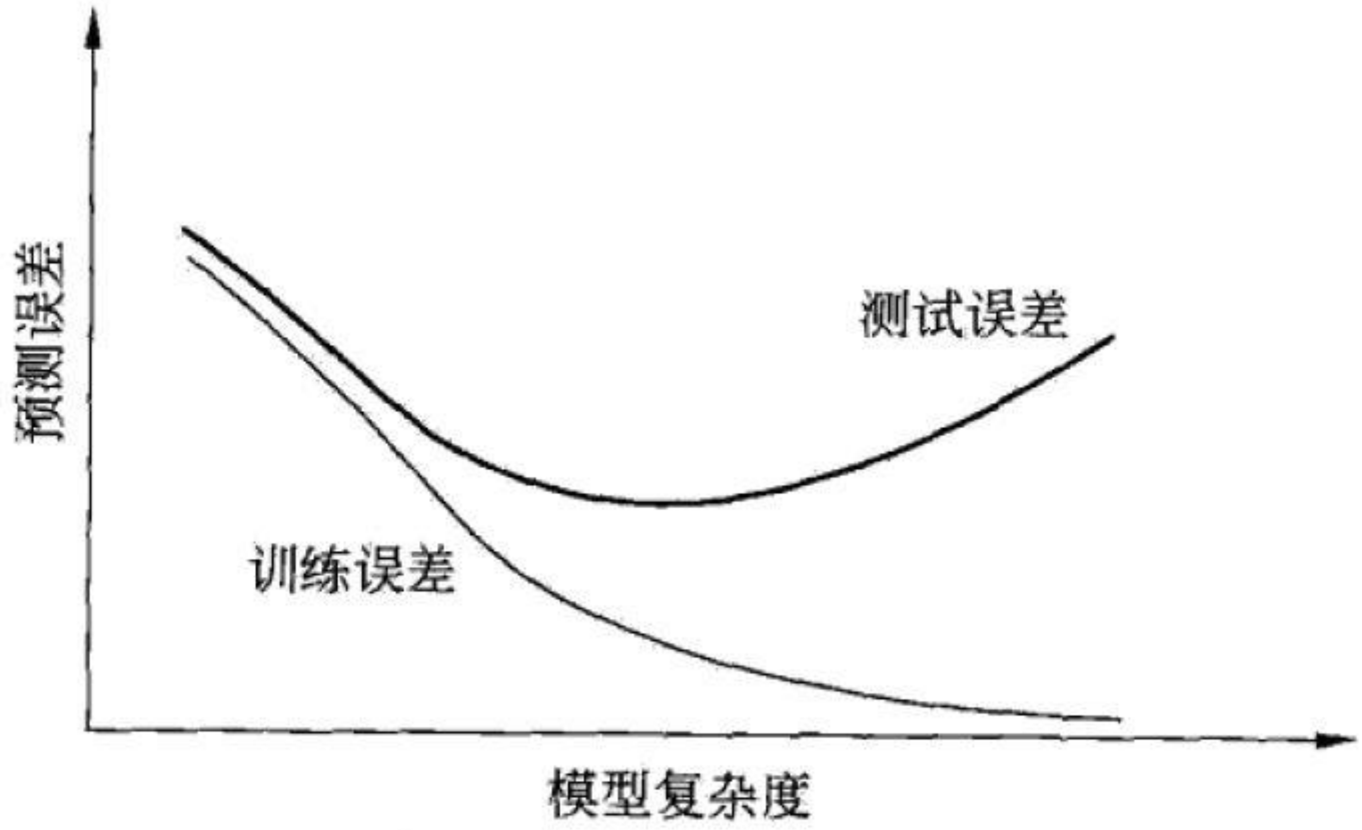
- 训练误差
  - 对判断给定问题是不是一个容易学习的问题是有意义的
- 测试误差
  - 反映了学习方法对未知的测试数据集的预测能力
- 泛化能力：学习方法对未知数据的预测能力
- 过拟合：决策函数对于训练集几乎全部拟合，但是对于测试集拟合效果过差的现象
- 欠拟合：模型拟合程度不高，数据距离拟合曲线较远，或指模型没有很好地捕捉到数据特征，不能够很好地拟合数据

1.4 模型评估与模型选择（2）

- 设M次多项式为$f_M(x,w) = w_0 + w_1 x + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$。



1.4 模型评估与模型选择（3）



1.5 正则化与交叉验证

- 正则化
  - 正则化项：一般是模型复杂度的单调递增函数
  - 奥卡姆剃刀原理（简单有效原理）：如无必要，勿增实体
- 交叉验证
  - 简单交叉验证
    - 随机地将数据分为训练集和测试集（例如，70%为训练集，30%为测试集）
  - S折交叉验证
    - 随机地将数据切分为S个互不相交的大小相同的子集，利用S-1个子集的数据训练模型，余下的子集测试模型；将这一过程对可能的S种选择重复进行
  - 留一交叉验证：S=N
    - 往往在数据缺乏的情况下使用

1.6 泛化能力（1）

- 泛化误差

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

- 泛化误差上界
  - 定理：对二类分类问题，当假设空间是有限个函数的集合$\mathcal{F} = \{f_1, f_2, \dots, f_d\}$时，对任意一个函数$f \in \mathcal{F}$，至少以概率$1-\delta$，以下不等式成立：

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

其中，

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

1.6 泛化能力（2）

Hoeffding不等式：设S_n是独立随机变量$X_1, X_2, \dots, X_n$之和，$X_i \in [a_i, b_i]$，则对任意$t > 0$，以下不等式成立：

$$P(ES_n - S_n \geq t) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

令

$$t = N\varepsilon$$

$$P(R(f) - \hat{R}(f) \geq \varepsilon) \leq \exp(-2N\varepsilon^2)$$

由于$\mathcal{F} = \{f_1, f_2, \dots, f_d\}$是一个有限集合，故

$$P\left(\exists f \epsilon \mathcal{F}: R(f) - \hat{R}(f) \geq \varepsilon\right) = P\left(\bigcup_{f \epsilon \mathcal{F}}\{R(f) - \hat{R}(f) \geq \varepsilon\}\right)$$

$$\leq \sum_{f \epsilon \mathcal{F}} P(R(f) - \hat{R}(f) \geq \varepsilon)$$

$$\leq d \exp(-2N\varepsilon^2)$$

1.6 泛化能力（3）

等价地，对任意$f\epsilon\mathcal{F}$，有

$$P\left(R(f)-\hat{R}(f)<\varepsilon\right)\geq 1-d\exp(-2N\varepsilon^2)$$

令

$$\delta=d\exp(-2N\varepsilon^2)$$

则

$$P\left(R(f)<\hat{R}(f)+\varepsilon\right)\geq 1-\delta$$

即至少以概率$1-\delta$有

$$R(f)<\hat{R}(f)+\varepsilon$$

1.6 泛化能力（4）

$$\delta = d\exp(-2N\varepsilon^2)$$

$$\log d + \log\frac{1}{\delta} = 2N\varepsilon^2$$

$$\varepsilon = \sqrt{\frac{1}{2N}\left(\log d + \log\frac{1}{\delta}\right)}$$

- 泛化误差上界性质:
  - 它是样本容量的函数，样本容量增加，泛化误差上界趋于0
  - 它是假设空间容量的函数，假设空间容量越大， 泛化误差上界越大

1.7 生成模型与判别模型（1）

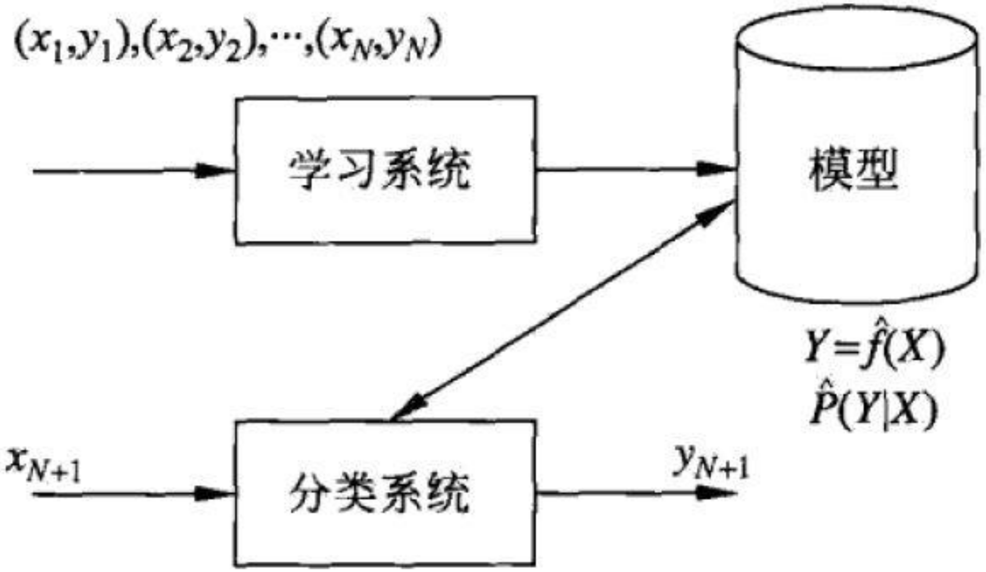
- 生成方法
  - 由数据学习联合概率分布$P(X,Y)$，然后求出条件概率分布$P(Y|X)$作为预测的模型，即生成模型。
  - 典型的生成模型：朴素贝叶斯法和隐马尔可夫模型
- 判别方法
  - 由数据直接学习决策函数$f(X)$或者条件概率分布$P(Y|X)$作为预测的模型，即判别模型。
  - 典型的判别模型：K近邻法、感知机、决策树、logistic回归模型、最大熵模型、支持向量机、提升方法和条件随机场

1.7 生成模型与判别模型（2）

- 生成方法的特点：
  - 可还原出联合概率分布$P(X,Y)$, 而判别方法不能
  - 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
- 判别方法的特点：
  - 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
  - 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

1.8 分类问题（1）

- 分类器：监督学习从数据中学习的一个分类模型或分类决策函数，称为分类器
- 分类：分类器对新的输入进行输出的预测
- 类：可能的输出
- 多类分类问题/二类分类问题

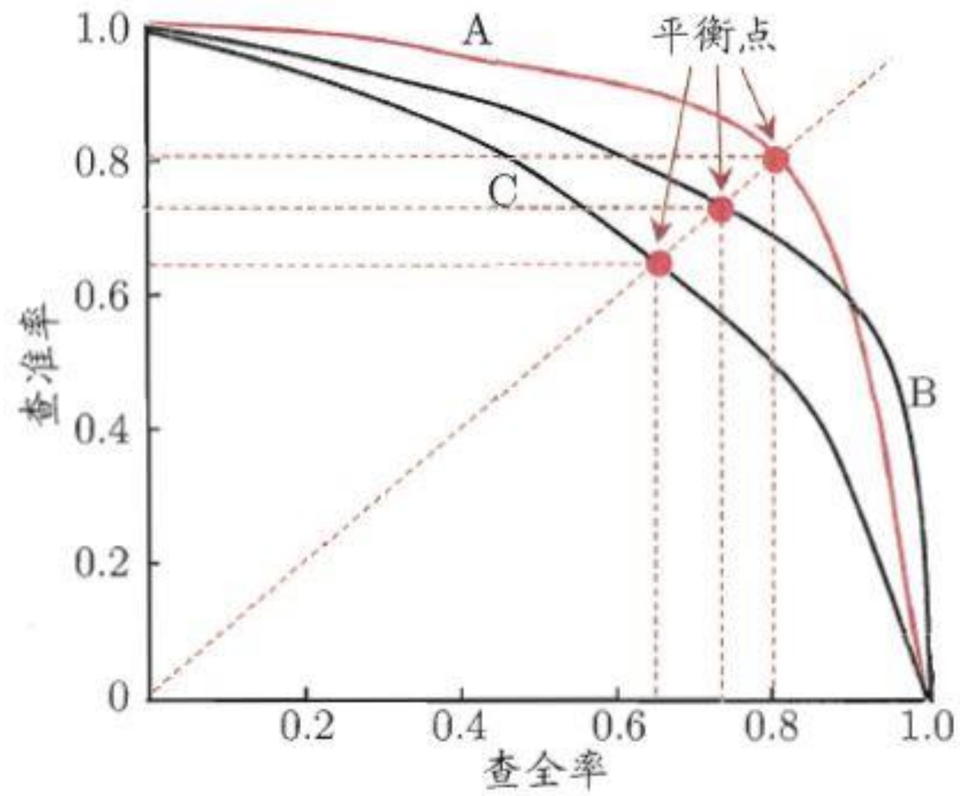


1.8 分类问题（2）

- 对于二类分类问题，通常以关注的类为正类，其他类为负类
- 4种情况（的总数）：
  - TP：将正类预测为正类数 FN：将正类预测为负类数
    FP：将负类预测为正类数 TN：将负类预测为负类数
- 准确率（accuracy）
  - 对于给定的测试数据集，分类器正确分类的样本数与总样本数之比
- 错误率
  - 与准确率相反，描述被分类器错分的比例

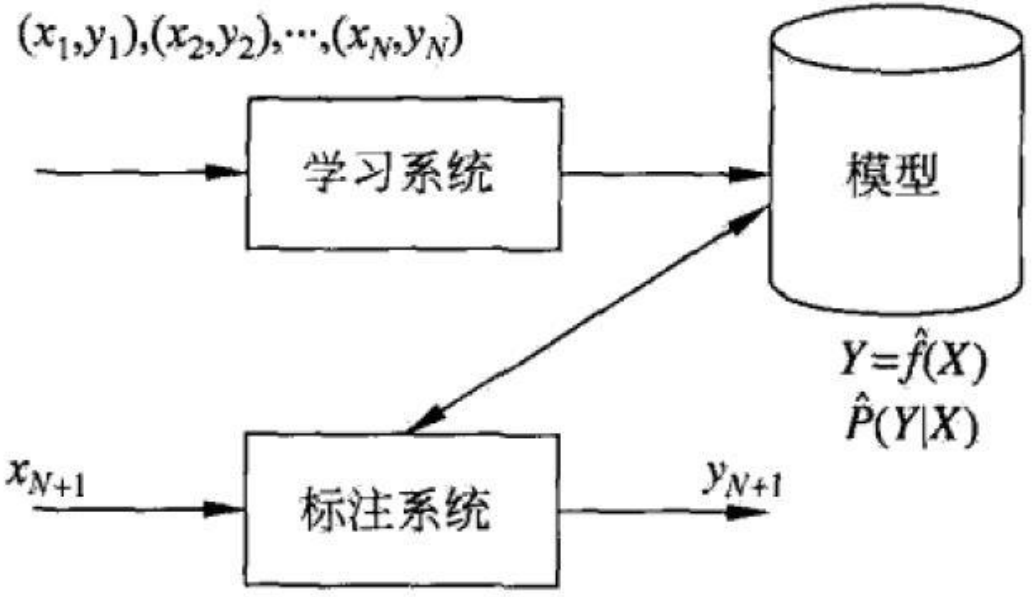
1.8 分类问题（3）

- 精确率（查准率，precision）
  - $P = \frac{TP}{TP+FP}$
- 召回率（查全率）
  - $R = \frac{TP}{TP+FN}$
- P-R曲线、P-R图
- 平衡点
- 调和平均数/加权调和平均数（$\beta > 0$）
  - $\frac{1}{F_1} = \frac{1}{2} \cdot \left(\frac{1}{P} + \frac{1}{R}\right)$
  - $\frac{1}{F_\beta} = \frac{1}{1+\beta^2} \cdot \left(\frac{1}{P} + \frac{\beta^2}{R}\right)$



1.9 标注问题

- 输入：观测序列，输出：标记序列或状态序列
- 标注问题分为学习和标注两个过程

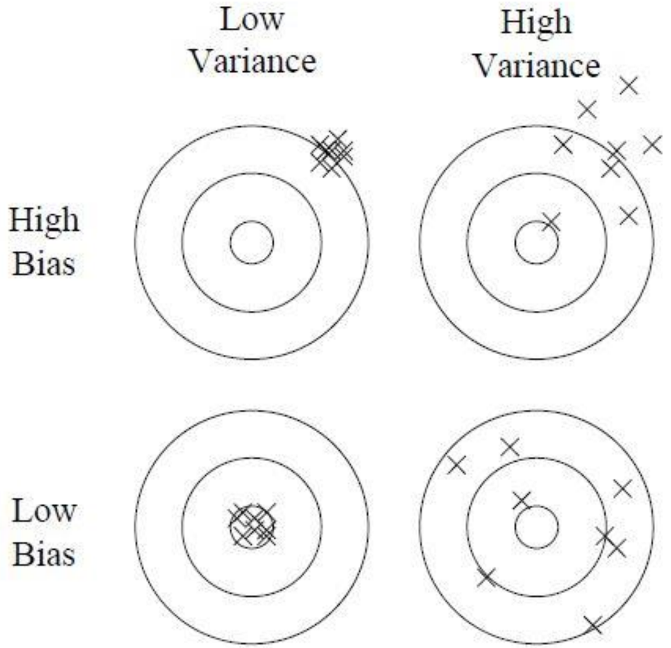


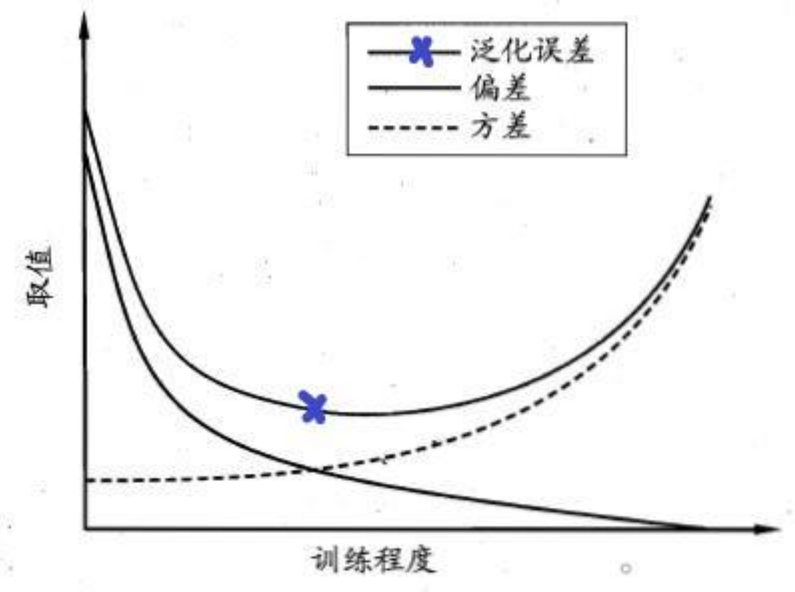
1.10 回归问题

- 回归问题的学习等价于函数拟合
- 分为学习和预测两个过程
- 分类
  - 按照输入变量的个数，分为一元回归和多元回归
  - 按照输入变量和输出变量之间关系的类型即模型的类型，分为线性回归和非线性回归
- 最小二乘法

《机器学习的那些事儿》 （1）

1. 过拟合有多种形式
   - 偏差-方差分解
   - 偏差-方差窘境



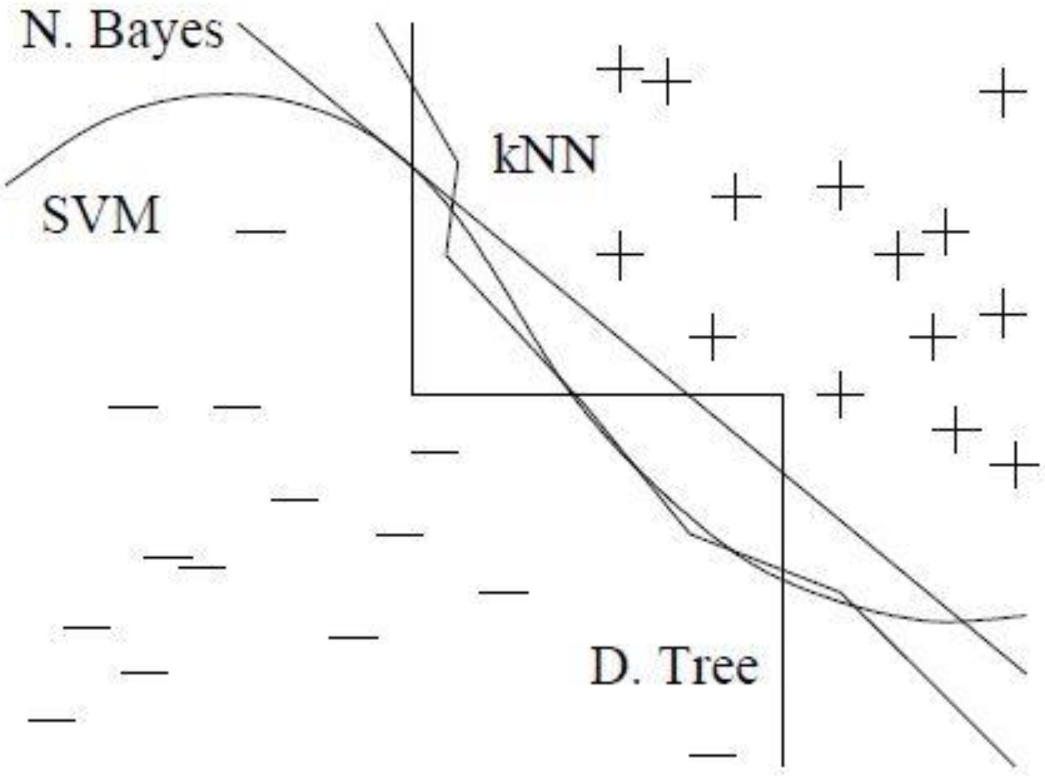


《机器学习的那些事儿》 （2）

- 2. 直觉不适用于高维空间
  - 维度灾难
  - 非均匀性祝福
- 3. 理论保证与看上去的不一样
  - 边界保证
  - 渐进保证

《机器学习的那些事儿》（3）

- 4. 数据多胜过算法聪明
- 5. 要学习很多模型，而不仅仅是一个
- 6. 可表示并不意味着可学习



“A few useful things to know about machine learning,” Communications of the ACM, vol. 55, no. 10, pp. 78--87, 2012.