# 第一章

# 统计学习方法概论

袁春 清华大学深圳研究生院
李航 华为诺亚方舟实验室

# 目录

# 一、统计学习

- 统计学习的对象
  - data： 计算机及互联网上的各种数字、文字、图像、视频、音频数据以及它们的组合。
  - 数据的基本假设是同类数据具有一定的统计规律性。
- 统计学习的目的
  - 用于对数据（特别是未知数据）进行预测和分析。

# 统计学习

- 统计学习的方法
  - 分类：
    - Supervised learning
    - Unsupervised learning
    - Semi-supervised learning
    - Reinforcement learning
  - 监督学习：
    - 训练数据 training data
    - 模型 model -------- 假设空间 hypothesis
    - 评价准则 evaluation criterion -------- 策略 strategy
    - 算法 algorithm

# 统计学习

- 统计学习的研究：
  - 统计学习方法
  - 统计学习理论（统计学习方法的有效性和效率和基本理论）
  - 统计学习应用

# 二、监督学习

- Instance， feature vector， feature space
- 输入实例x的特征向量：

$$x=(x^{(1)},x^{(2)},\cdots,x^{(i)},\cdots,x^{(n)})^{\mathrm{T}}$$

- $x^{(i)}$与$x_i$ 不同,后者表示多个输入变量中的第i个

$$x_i=(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(n)})^{\mathrm{T}}$$

- 训练集：

$$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$$

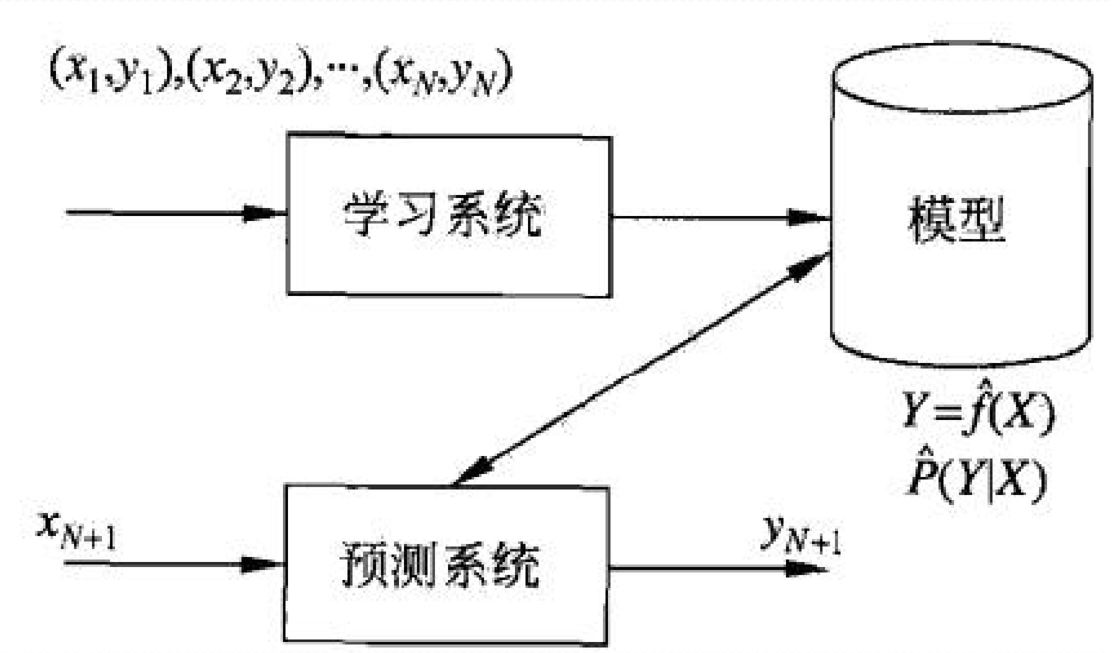
- 输入变量和输出变量：
  - 分类问题、回归问题、标注问题

# 监督学习

- 联合概率分布
  - 假设输入与输出的随机变量X和Y遵循联合概率分布P(X,Y)
  - P(X,Y)为分布函数或分布密度函数
  - 对于学习系统来说，联合概率分布是未知的，
  - 训练数据和测试数据被看作是依联合概率分布P(X,Y)独立同分布产生的。
- 假设空间
  - 监督学习目的是学习一个由输入到输出的映射，称为模型
  - 模式的集合就是假设空间（hypothesis space）
  - 概率模型:条件概率分布P(Y|X), 决策函数：Y=f(X)

# 监督学习

- 问题的形式化




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

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

# 三、统计学习三要素

方法＝模型＋策略＋算法

- 模型：
  - 决策函数的集合： $\mathcal{F}=\{f \mid Y=f(X)\}$
  - 参数空间 $\mathcal{F}=\{f \mid Y=f_\theta(X),\theta\in\mathbf{R}^n\}$
  - 条件概率的集合： $\mathcal{F}=\{P \mid P(Y \mid X)\}$
  - 参数空间 $\mathcal{F}=\{P \mid P_\theta(Y \mid X),\theta\in\mathbf{R}^n\}$

# 统计学习三要素

- 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

- 策略
  - 对数损失函数 logarithmic loss function 或对数似然损失函数 loglikelihood loss function

$$L(Y,P(Y|X)) = -\log P(Y|X)$$

  - 损失函数的期望

$$R_{\exp}(f) = E_P[L(Y,f(X))] = \int_{\mathcal{X}\times\mathcal{Y}} L(y,f(x))P(x,y)\mathrm{d}x\mathrm{d}y$$

  - 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

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

# 统计学习三要素

- 策略：经验风险最小化与结构风险最小化
  - 经验风险最小化最优模型

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

  - 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regularizer，或罚项 penalty term：

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

# 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：
  - 如果最优化问题有显式的解析式，算法比较简单
  - 但通常解析式不存在，就需要数值计算的方法

# 四、模型评估与模型选择

- 训练误差，训练数据集的平均损失

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

- 测试误差，测试数据集的平均损失

$$e_{\mathrm{test}}=\frac{1}{N'}\sum_{i=1}^{N'}L(y_i,\hat{f}(x_i))$$

- 损失函数是0-1 损失时：

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

- 测试数据集的准确率：

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

# 模型评估与模型选择

- 过拟合与模型选择
- 假设给定训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$

$$f_M(x,w)=w_0+w_1x+w_2x^2+\cdots+w_Mx^M=\sum_{j=0}^{M}w_jx^j$$

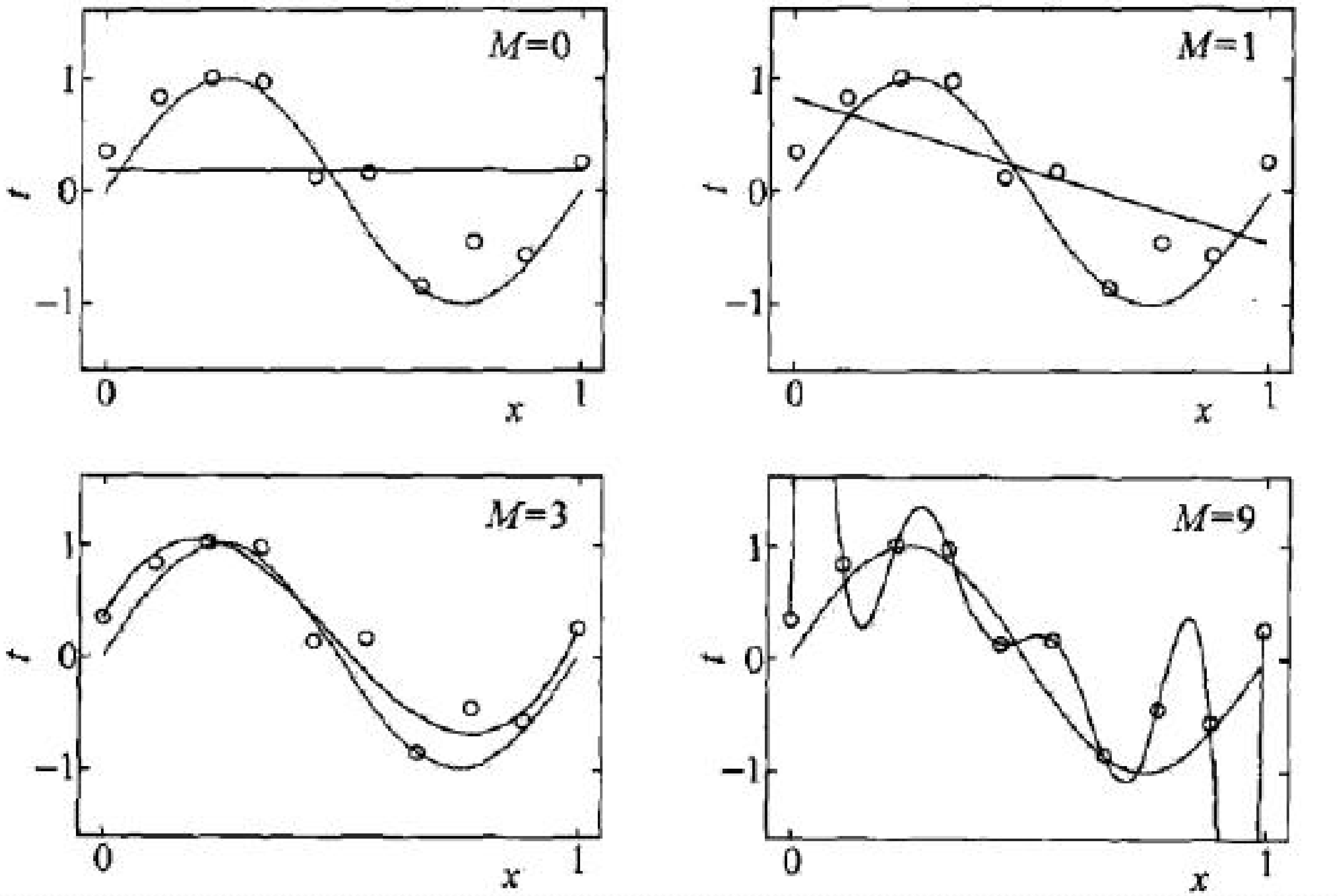
- 经验风险最小：

$$L(w)=\frac{1}{2}\sum_{i=1}^{N}(f(x_i,w)-y_i)^2$$

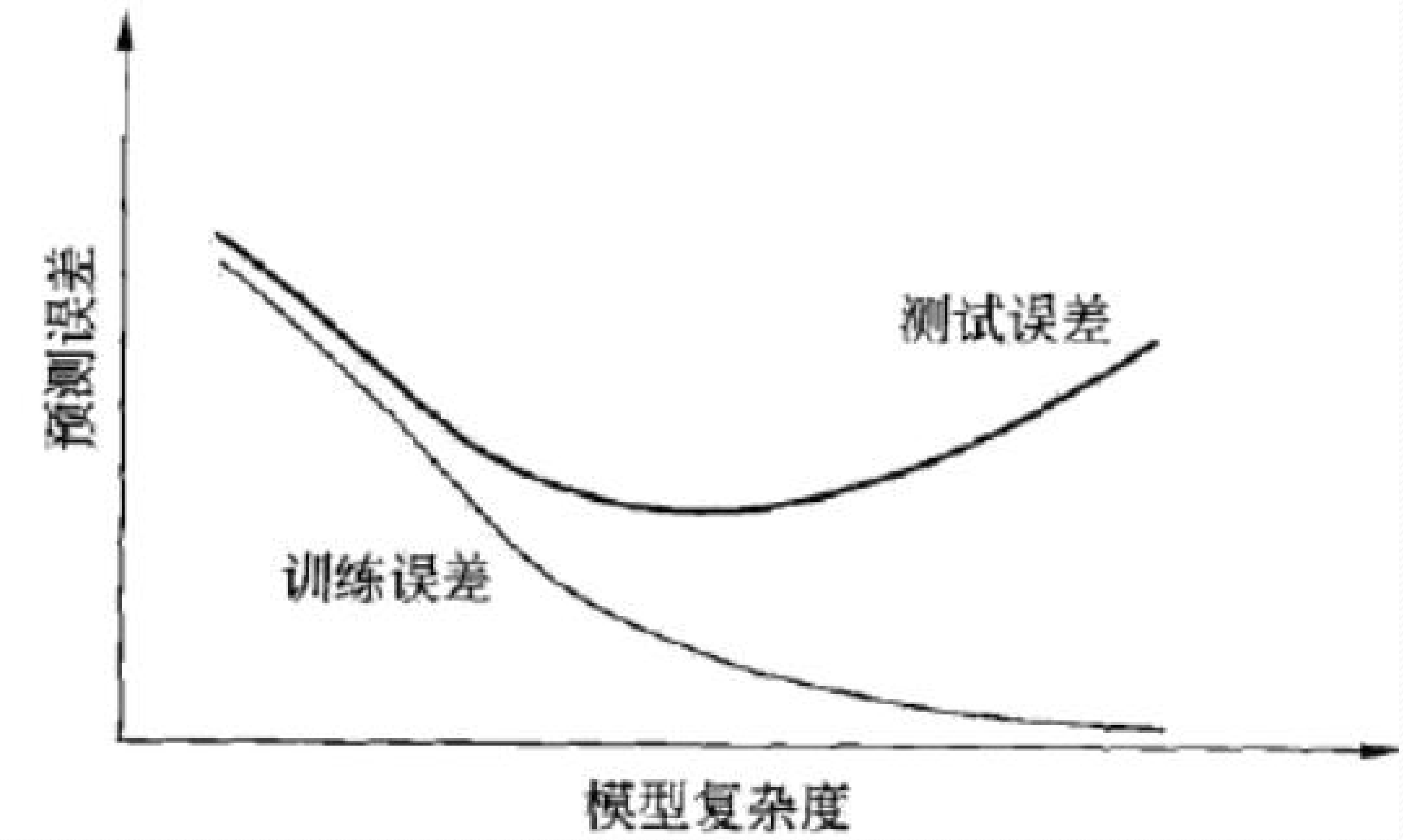
$$L(w)=\frac{1}{2}\sum_{i=1}^{N}\left(\sum_{j=0}^{M}w_jx_i^j-y_i\right)^2$$

$$w_j=\frac{\sum_{i=1}^{N}x_iy_i}{\sum_{i=1}^{N}x_i^{j+1}},\quad j=0,1,2,\cdots,M$$

# 模型评估与模型选择



# 模型评估与模型选择



# 五、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：

$$L(w) = \frac{1}{N}\sum_{i=1}^{N} (f(x_i; w) - y_i)^2 + \frac{\lambda}{2} \| w \|^2$$

$$L(w) = \frac{1}{N}\sum_{i=1}^{N} (f(x_i; w) - y_i)^2 + \lambda \| w \|_1$$

# 正则化与交叉验证

- 交叉验证：
  - 训练集 training set：　用于训练模型
  - 验证集 validation set：　　用于模型选择
  - 测试集　test　set：　用于最终对学习方法的评估

  - 简单交叉验证
  - S折交叉验证
  - 留一交叉验证

# 六、泛化能力 generalization ability

- 泛化误差 generalization  error

$$R_{\exp}(\hat{f}) = E_P[L(Y,\hat{f}(X))] = \int_{\mathcal{X}\times\mathcal{Y}} L(y,\hat{f}(x))P(x,y)\mathrm{d}x\mathrm{d}y$$

- 泛化误差上界
  - 比较学习方法的泛化能力------比较泛化误差上界
  - 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

- 二分类问题 $X \in \mathbf{R}^n$，$Y \in \{-1,+1\}$

- 期望风险和经验风险

$$R(f) = E[L(Y,f(X))] \qquad \hat{R}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i,f(x_i))$$

# 泛化能力 generalization ability

- 经验风险最小化函数：

$$f_N = \arg\min_{f\in\mathcal{F}} \hat{R}(f)$$

- 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，当假设空间是有限个函数的结合 $\mathcal{F}=\{f_1, f_2, \cdots, f_d\}$，对任意一个函数f，至少以概率1-δ，以下不等式成立：

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

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

# 七、生成模型与判别模型

- 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：

$$P(Y \mid X)$$

- 生成方法Generative approach 对应生成模型：generative model，

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

- 朴素贝叶斯法和隐马尔科夫模型

# 生成模型与判别模型

- 判别方法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 approach对应discriminative model

$$Y = f(X)$$

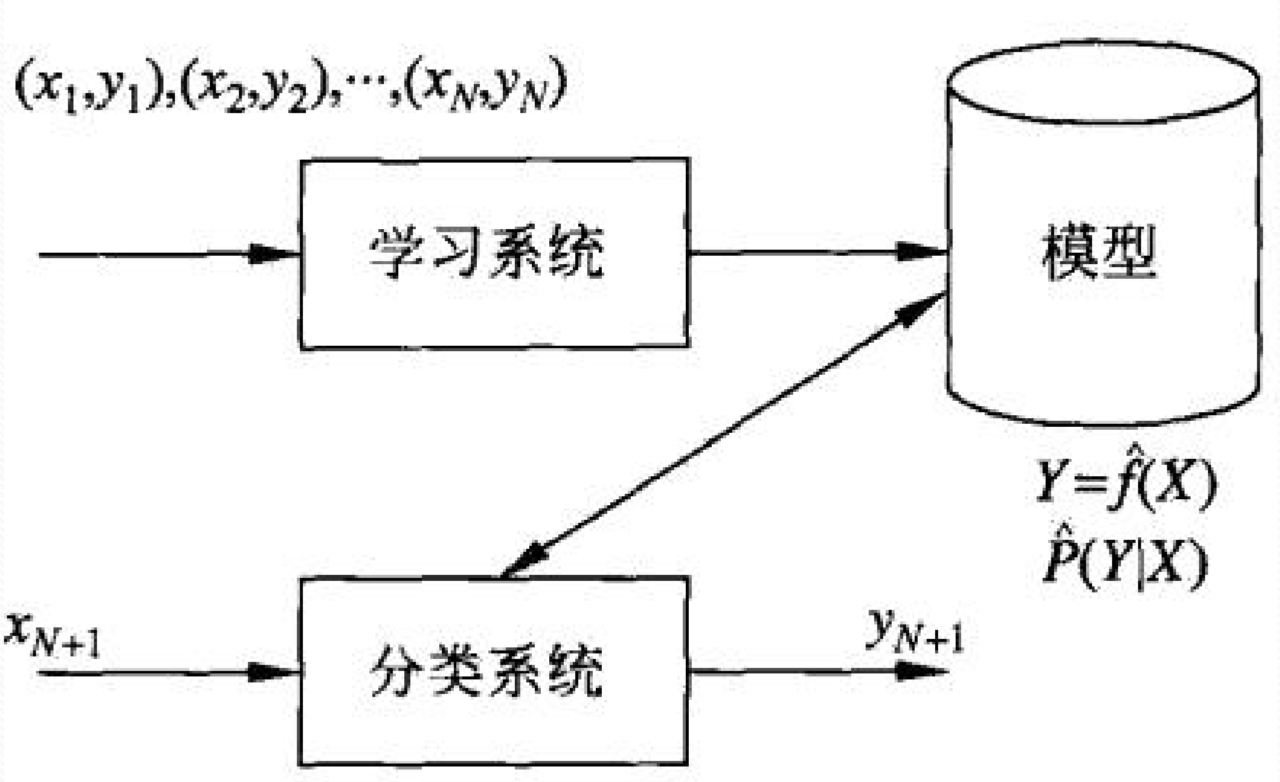
$$P(Y \mid X)$$

- 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。

# 生成模型与判别模型

- 各自优缺点：
  - 生成方法：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：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。

# 八、分类问题



# 分类问题

- 二分类评价指标
  - TP　true positive
  - FN　false negative
  - FP　false positive
  - TN　true  negative
- 精确率

$$P = \frac{TP}{TP + FP}$$

- 召回率

$$R = \frac{TP}{TP + FN}$$

- F1值

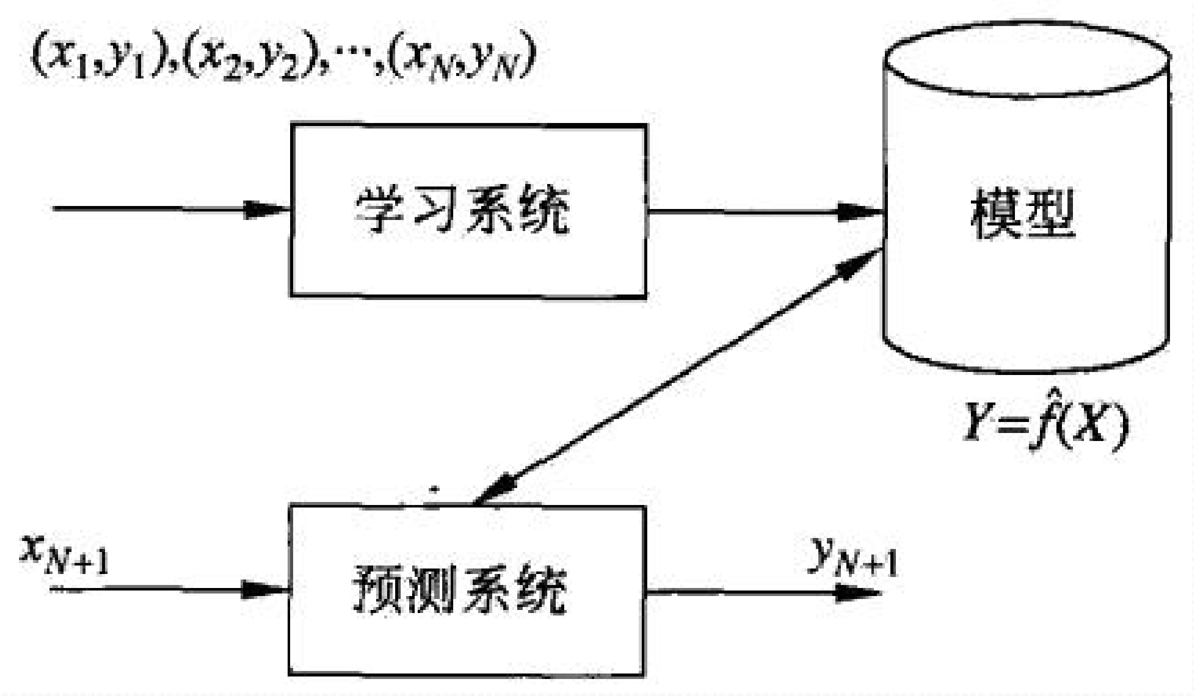
$$\frac{2}{F_1} = \frac{1}{P} + \frac{1}{R} \qquad F_1 = \frac{2TP}{2TP + FP + FN}$$

# 九、标注问题

- 标注：tagging， 结构预测：structure prediction
- 输入：观测序列， 输出：标记序列或状态序列
- 学习和标注两个过程
- 训练集： $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$
- 观测序列： $x_i=(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(n)})^{\mathrm{T}}$， $i=1,2,\cdots,N$
- 输出标记序列： $y_i=(y_i^{(1)},y_i^{(2)},\cdots,y_i^{(n)})^{\mathrm{T}}$
- 模型：条件概率分布 $P(Y^{(1)},Y^{(2)},\cdots,Y^{(n)}\mid X^{(1)},X^{(2)},\cdots,X^{(n)})$

# 十、回归问题

- 回归模型是表示从输入变量到输出变量之间映射的函数. 回归问题的学习等价于函数拟合。
- 学习和预测两个阶段
- 训练集： $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$



# 回归问题

- 例子：
- 标记表示名词短语的“开始”、“结束”或“其他”（分别以B, E, O表示)
- 输入：At Microsoft Research, we have an insatiable curiosity and the desire to create new technology that will help define the computing experience.

- 输出：At/O Microsoft/B Research/E, we/O have/O an/O insatiable/6 curiosity/E and/O the/O desire/BE to/O create/O new/B technology/E that/O will/O help/O define/O the/O computing/B experience/E.

# 回归问题

- 回归学习最常用的损失函数是平方损失函数，在此情况下，回归问题可以由 著名的最小二乘法(least squares)求解。
- 股价预测

Q & A