清華大學
Tsinghua University

大数据机器学习

第一章：统计学习及监督学习概论

Chun Yuan
2019/09/19

统计学习

- 统计学习的特点
  - statistical learning：是关于计算机基于数据构建概率统计模型并运用模型对数据进行预测与分析的一门学科。
  - 统计学习也称为统计机器学习(statistical machine learning）。

统计学习

- 统计学习的特点
  - （1）统计学习以计算机及网络为平台，是建立在计算机网络上的；
  - （2）统计学习以数据为研究对象，是数据驱动的学科；
  - （3）统计学习的目的是对数据进行预测与分析；
  - （4）统计学习以方法为中心，统计学习方法构建模型并应用模型进行预测与分析；
  - （5）统计学习是概率论、统计学、信息论、计算理论、最优化理论及计算机科学等多个领域的交叉学科，并且在发展中逐步形成独自的理论体系与方法论。

Herbert A. Simon：如果一个系统能够通过执行某个过程改进它的性能，这就是学习。

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。
  - 对数据的预测与分析是通过构建概率统计模型实现的。

统计学习

- 统计学习的方法
  - 分类:
    - Supervised learning
    - Unsupervised learning
    - Semi-supervised learning
    - Reinforcement learning

统计学习

- 从给定的、有限的、用于学习的训练数据（training data）集合出发，假设数据是独立同分布产生的;
- 假设要学习的模型属于某个函数的集合，称为假设空间（hypothesis space）；
- 应用某个评价准则（evaluation criterion），从假设空间中选取一个最优模型，使它对已知的训练数据及未知的测试数据（test data）在给定的评价准则下有最优的预测;
- 最优模型的选取由算法实现;
- 统计学习方法包括模型的假设空间、模型选择的准则以及模型学习的算法。称其为统计学习方法的三要素，简称为模型（model）、策略（strategy）和算法（algorithm）。

统计学习方法的步骤

- （1）得到一个有限的训练数据集合
- （2）确定包含所有可能的模型的假设空间，即学习模型的集合;
- （3）确定模型选择的准则，即学习的策略;
- （4）实现求解最优模型的算法，即学习的算法;
- （5）通过学习方法选择最优模型;
- （6）利用学习的最优模型对新数据进行预测或分析。;

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

- 监督学习 Supervised learning
- 无监督学习 Unsupervised learning
- 强化学习 Reinforcement learning
- 半监督学习 Semi-supervised learning
- 在线学习　Online learning
- 主动学习　Active learning
- 弱监督学习 weakly supervised learning

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：

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

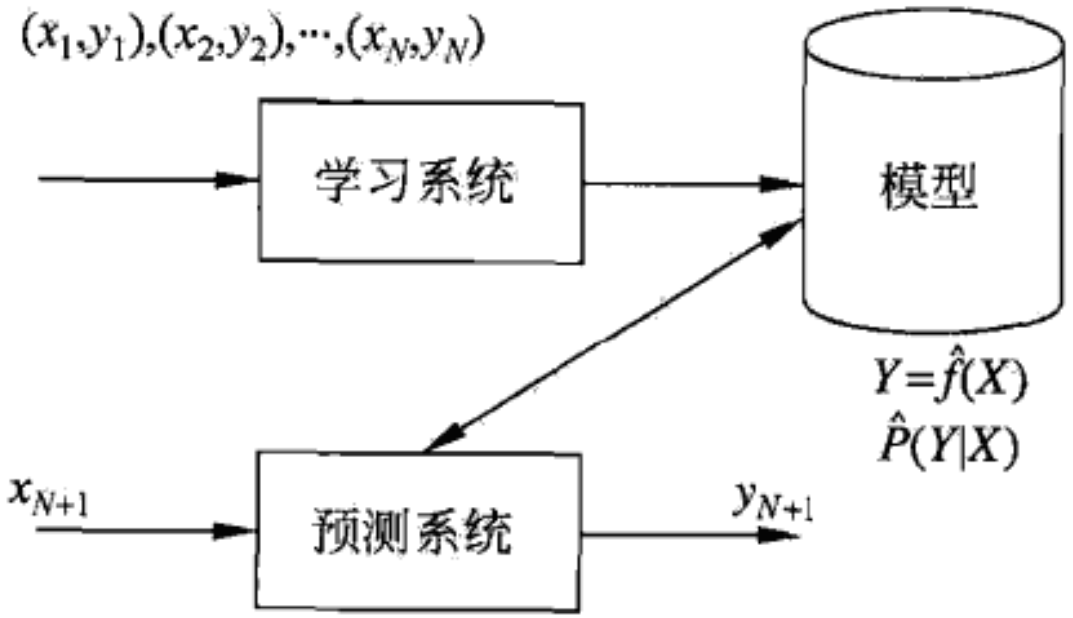
- 输入变量和输出变量：
  - 分类问题、回归问题、标注问题

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）
  - 概率模型和非概率模型：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



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

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

无监督学习（从无标注数据中学习）

- 训练集:

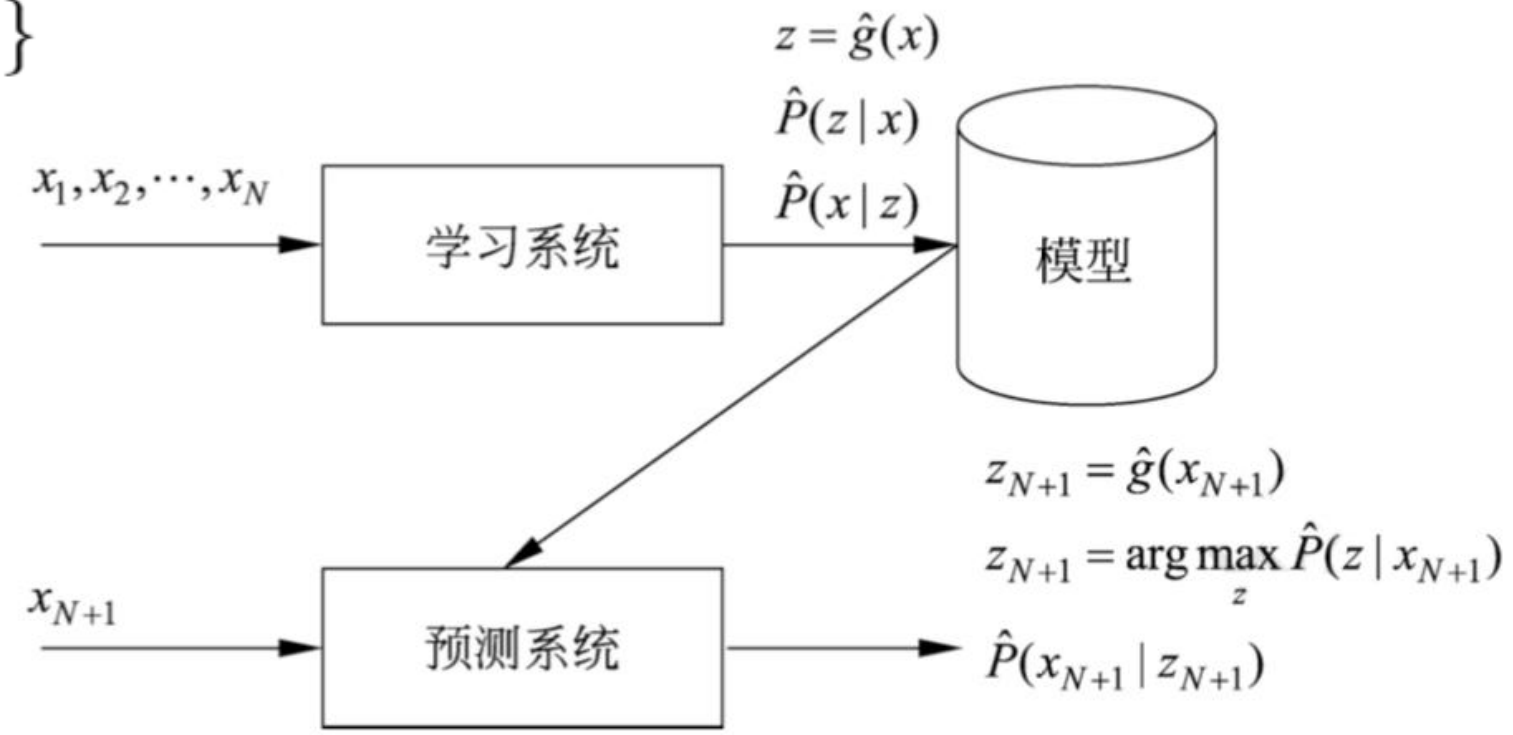
$$U = \{x_1, x_2, \cdots, x_N\}$$

- 模型函数:

$$z = g(x)$$

- 条件概率分布:

$$P(z|x)$$

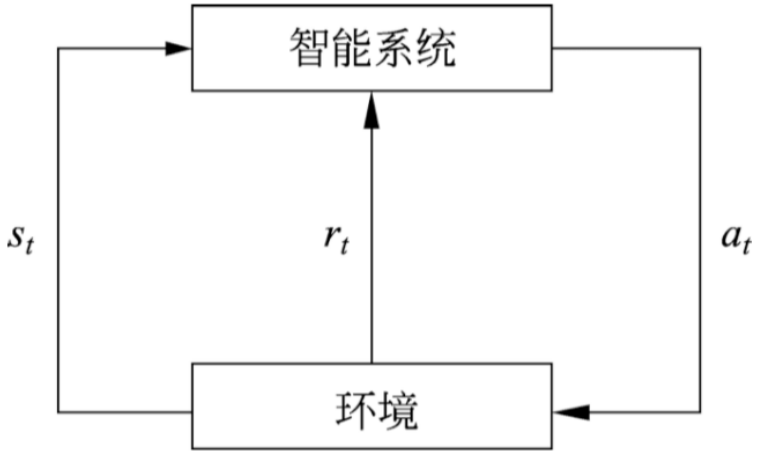


强化学习

强化学习的马尔可夫决策过程是状态、奖励、动作序列上的随机过程，由五元组 $\langle S, A, P, r, \gamma \rangle$ 组成。

- S 是有限状态（state）的集合
- A 是有限动作（action）的集合
- P 是状态转移概率（transition probability）函数:

$$P(s'|s,a) = P(s_{t+1} = s'|s_t = s, a_t = a)$$



- r 是奖励函数（reward function）：$r(s,a) = E(r_{t+1}|s_t = s, a_t = a)$
- γ 是衰减系数（discount factor）：$\gamma \in [0,1]$

强化学习

- 状态转移概率函数: $P(s'|s,a)$
- 奖励函数: $r(s,a)$
- 策略π: 给定状态下动作的函数 $a = f(s)$ 或者条件概率分布 $P(a|s)$
- 状态价值函数: $v_\pi(s) = E_\pi[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots | s_t = s]$
- 动作价值函数: $q_\pi(s,a) = E_\pi[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots | s_t = s, a_t = a]$

强化学习方法

- 无模型（model-free）
  - 基于策略（policy-based）：求解最优策略π^*
  - 基于价值（value-based）：求解最优价值函数
- 有模型（model-based）
  - 通过学习马尔可夫决策过程的模型，包括转移概率函数和奖励函数
  - 通过模型对环境的反馈进行预测
  - 求解价值函数最大的策略π^*

半监督学习

- 少量标注数据，大量未标注数据
- 利用未标注数据的信息，辅助标注数据，进行监督学习
- 较低成本

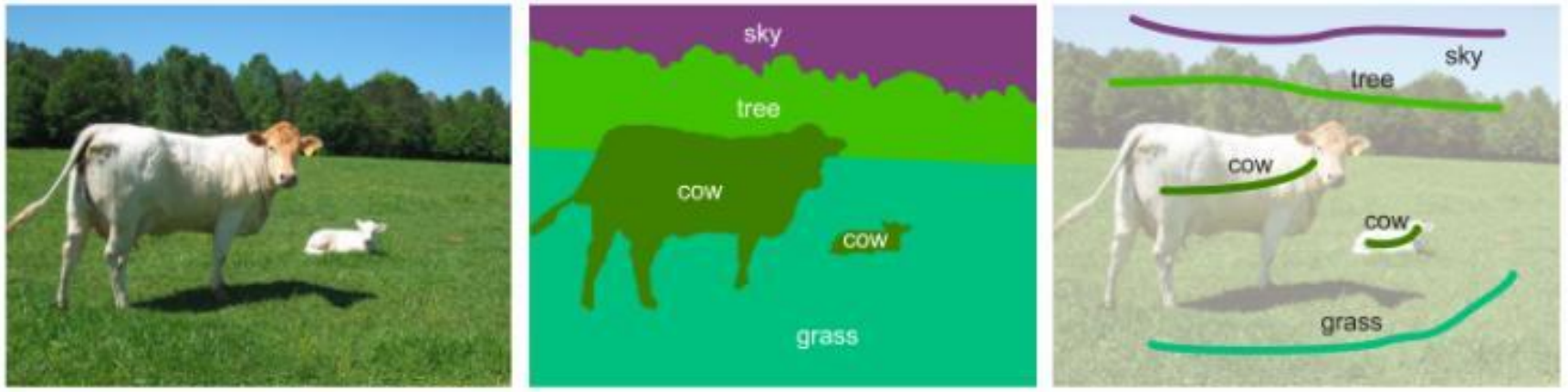
主动学习

- 机器主动给出实例，教师进行标注
- 利用标注数据学习预测模型

弱监督学习

使用比目标任务更弱级别的监督标签

以计算机视觉任务为例：仅使用图像级的类别标签作为监督得到各类别对应的判定区域、定位框、(语义/实例)分割图，



(a) image　(b) mask annotation　(c) scribble annotation

➢监督信息粒度：

- Image_wise_label
- Point
- Scribble
- Bounding_box
- Sementic_segmentation
- Instance_segmentation

- 相关领域：

➢卷积神经网络的可视化：可视化每一层卷积对应提取的特征模式

➢全监督语义分割：FCN+CRF

➢Saliency detection: 检测显著性物体（前景）

➢弱监督：使用比目标任务更弱级别的监督标签

How to get more labeled training data?

Traditional Supervision: Have subject matter experts (SMEs) hand-label more training data

Too expensive!

Active Learning: Estimate which points are most valuable to solicit labels for

Semi-supervised Learning: Use structural assumptions to automatically leverage unlabeled data

***Weak* Supervision:** Get lower-quality labels more efficiently and/or at a higher abstraction level

Get cheaper, lower-quality labels from non-experts

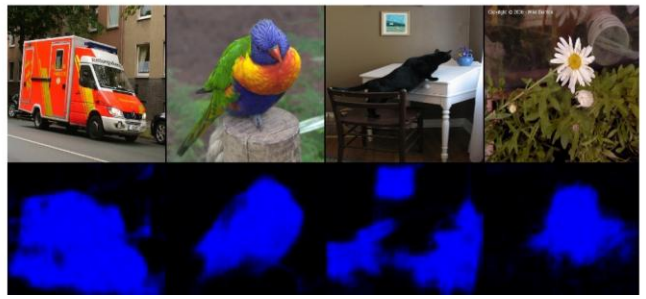
Get higher-level supervision over unlabeled data from SMEs

- *Heuristics*
- *Distant Supervision*
- *Constraints*
- *Expected distributions*
- *Invariances*

Transfer Learning: Use models already trained on a different task

Use one or more (noisy / biased) pre-trained models to provide supervision

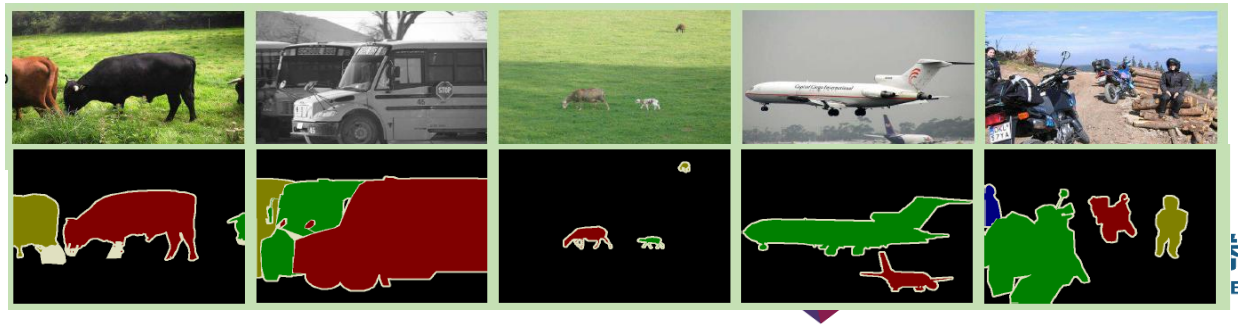
- 弱监督定位



- 弱监督目标检测

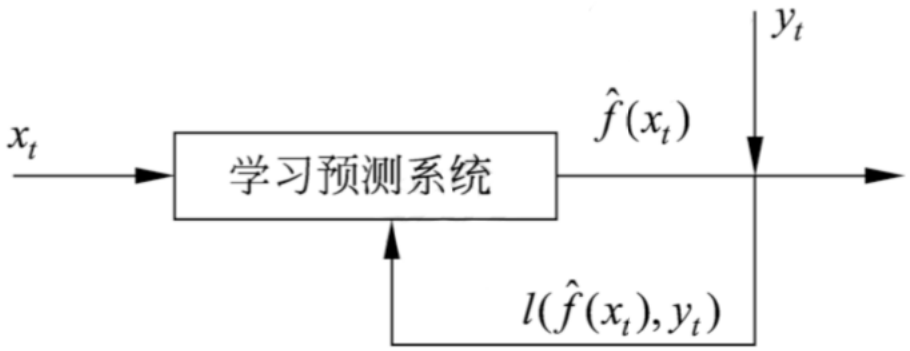
- 弱监督语义分割

- 弱监督实例分割



统计学习

- 按算法分类:
  - 在线学习（online learning）



  - 批量学习（batch learning）

统计学习

- 按模型分类:
- 概率模型与非概率模型:
  - *P(z|x) 或P(x|z)，　y=f(x)*
  - 逻辑斯谛回归既可看作是概率模型，又可看作是非概率模型。
  - 相互转换
- 线性模型与非线性模型
- 参数化模型与非参数化模型

统计学习

- 按技巧分类:
  - 贝叶斯学习（Bayesian learning）

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)}$$

模型估计时，估计整个后验概率分布 $P(\theta|D)$。如果需要给出一个模型，通常取后验概率最大的模型。

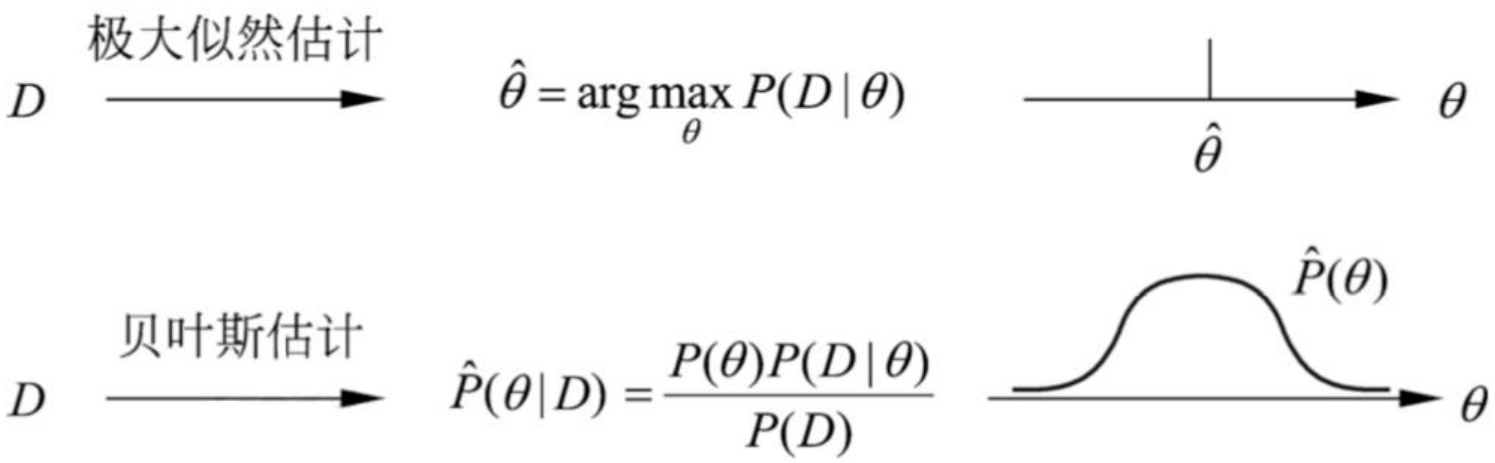
预测时，计算数据对后验概率分布的期望值：

$$P(x|D) = \int P(x|\theta, D)P(\theta|D)\mathrm{d}\theta$$

这里 x 是新样本。

统计学习

- 按技巧分类:
  - 贝叶斯学习（Bayesian learning）



统计学习

- 按技巧分类:
  - 核方法（Kernel method）
    - 使用核函数表示和学习非线性模型，将线性模型学习方法扩展到非线性模型的学习
    - 不显式地定义输入空间到特征空间的映射，而是直接定义核函数，即映射之后在特征空间的内积
    - 假设x_1，x_2是输入空间的任意两个实例，内积为$<x_1, x_2>$，输入空间到特征空间的映射为φ，核方法在输入空间中定义核函数 $K(x_1, x_2)$，使其满足 $K(x_1, x_2) = < \varphi(x_1), \varphi(x_2)>$

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

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

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：

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

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

模型评估与模型选择

- 训练误差，训练数据集的平均损失

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

- 测试误差，测试数据集的平均损失

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

- 损失函数是0-1 损失时：

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

- 测试数据集的准确率：

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

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

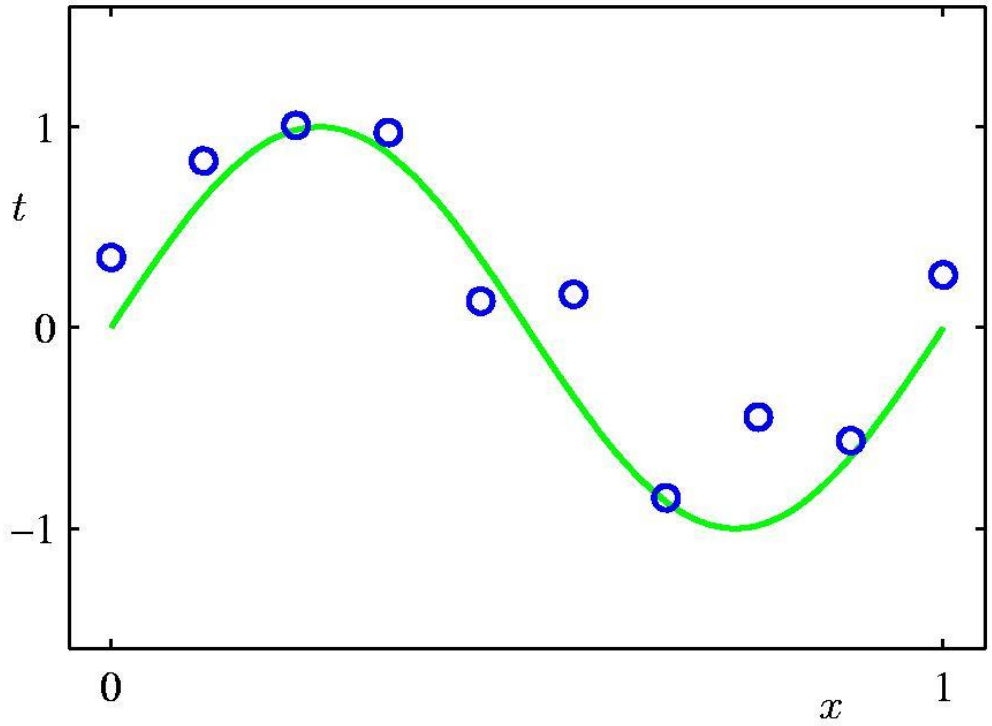
$$f_M(x,w)=w_0+w_1x+w_2x^2+\cdots+w_Mx^M=\sum_{j=0}^{M}w_jx^j$$

- 经验风险最小:

$$L(w)=\frac{1}{2}\sum_{i=1}^{N}(f(x_i,w)-y_i)^2 \qquad L(w)=\frac{1}{2}\sum_{i=1}^{N}\left(\sum_{j=0}^{M}w_jx_i^j-y_i\right)^2 \qquad w_j=\frac{\sum_{i=1}^{N}x_iy_i}{\sum_{i=1}^{N}x_i^{j+i}},\quad j=0,1,2,\cdots,M$$

Polynomial Curve Fitting

sin(2πx)+noise



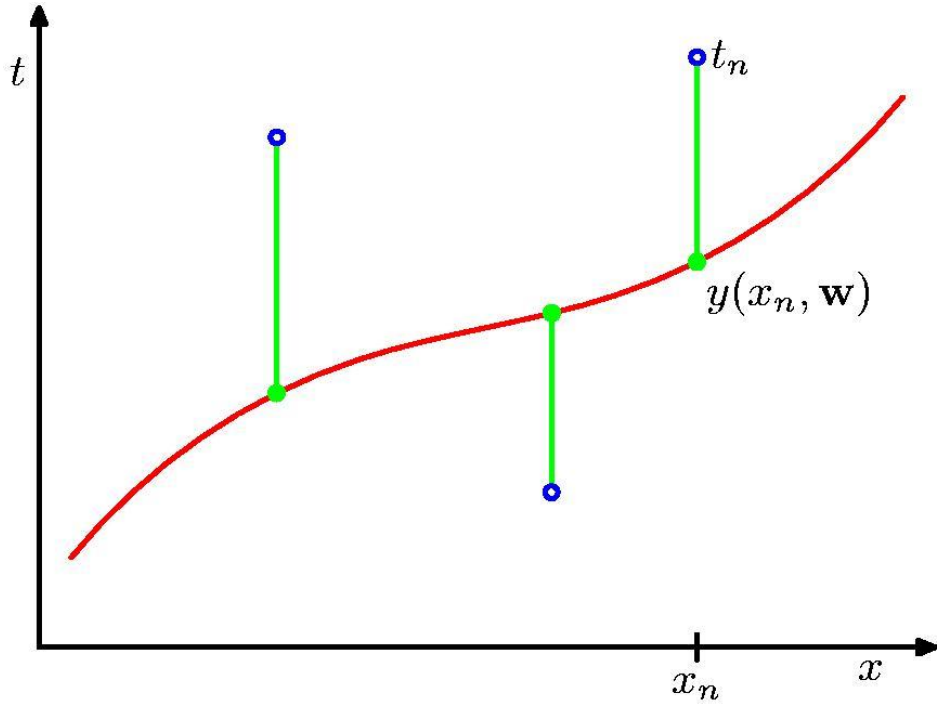
$$\mathbf{x} \equiv (x_1, \ldots, x_N)^{\mathrm{T}}$$

$$\mathbf{t} \equiv (t_1, \ldots, t_N)^{\mathrm{T}}$$

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Sum-of-Squares Error Function

$\mathbf{x} \equiv (x_1, \ldots, x_N)^{\mathrm{T}}$

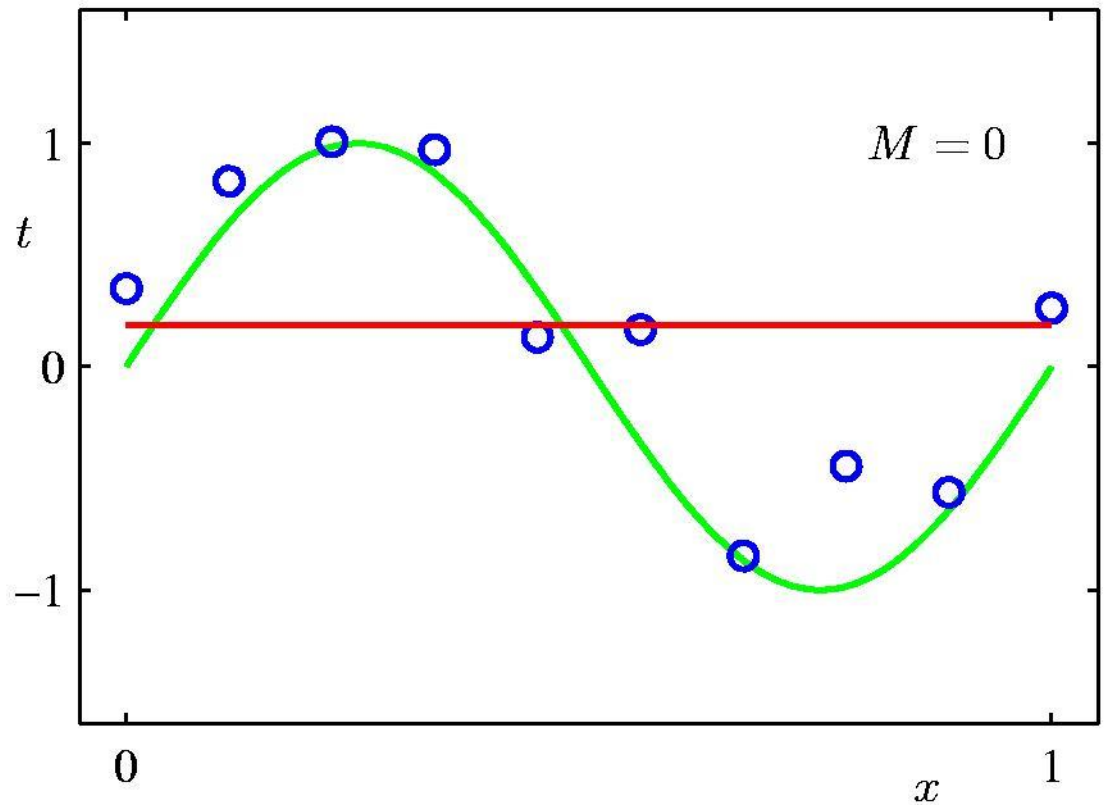


$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

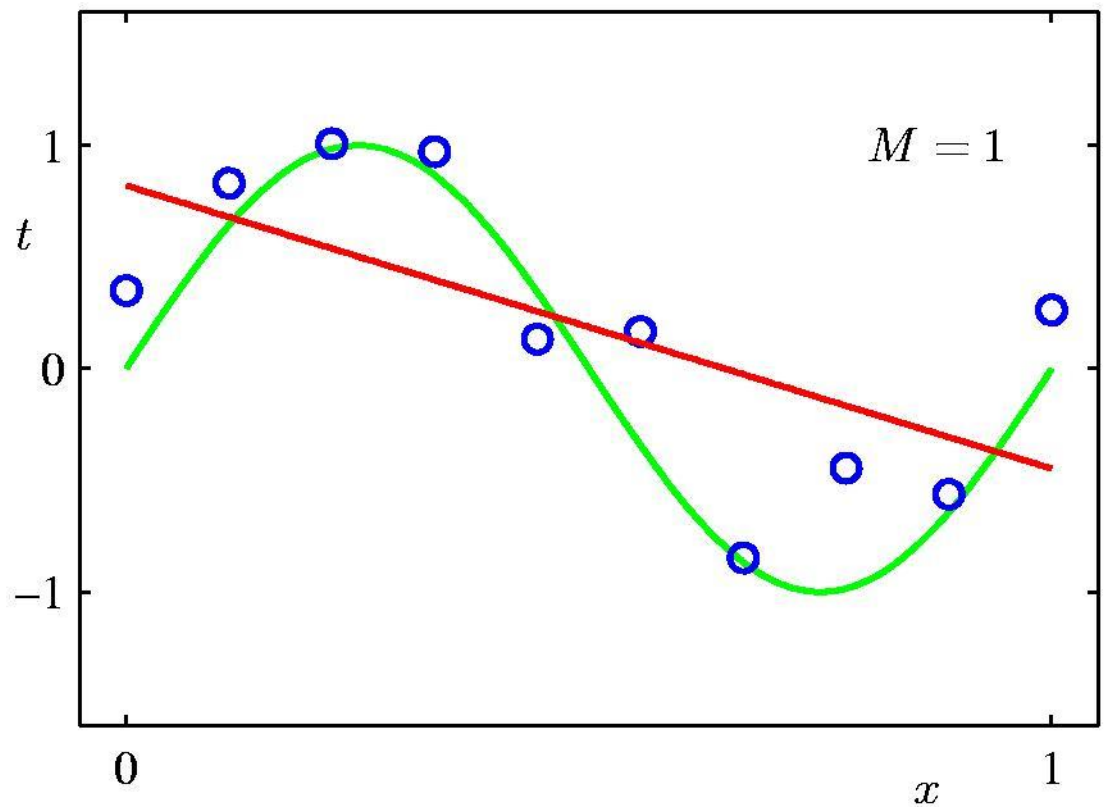
$\mathbf{x} \equiv (x_1, \ldots, x_N)^{\mathrm{T}}$

$\mathbf{t} \equiv (t_1, \ldots, t_N)^{\mathrm{T}}$

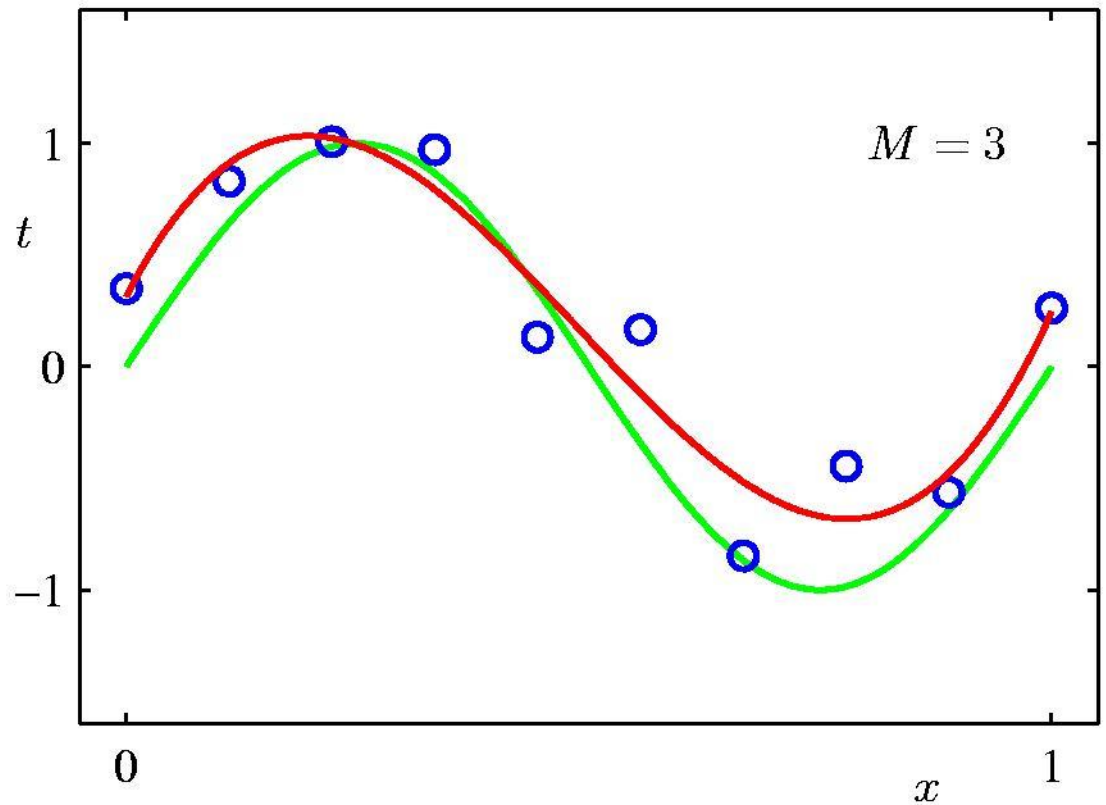
0^{th} Order Polynomial



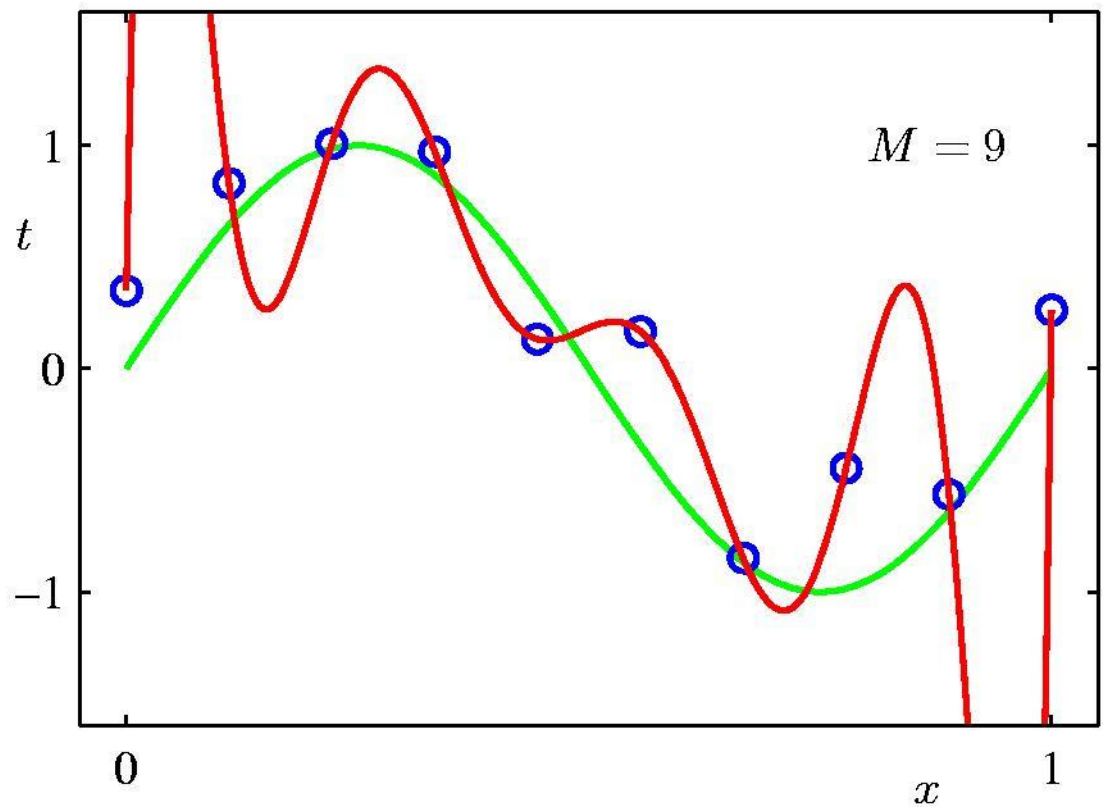
1st Order Polynomial



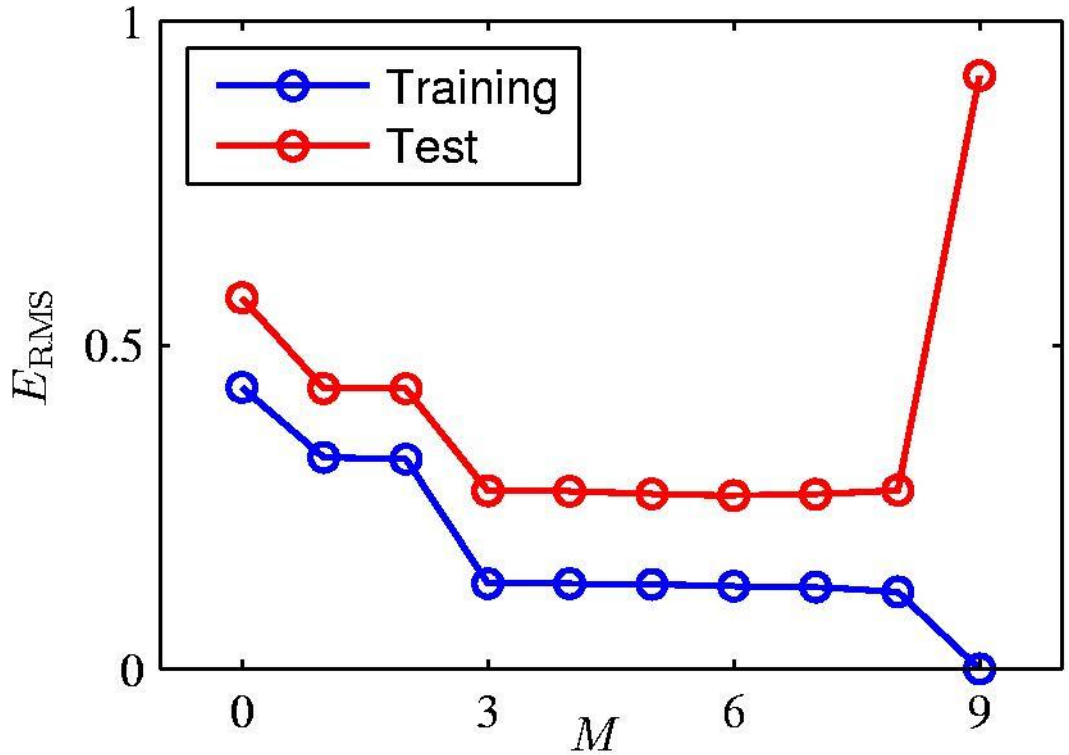
3rd Order Polynomial



9th Order Polynomial



Over-fitting



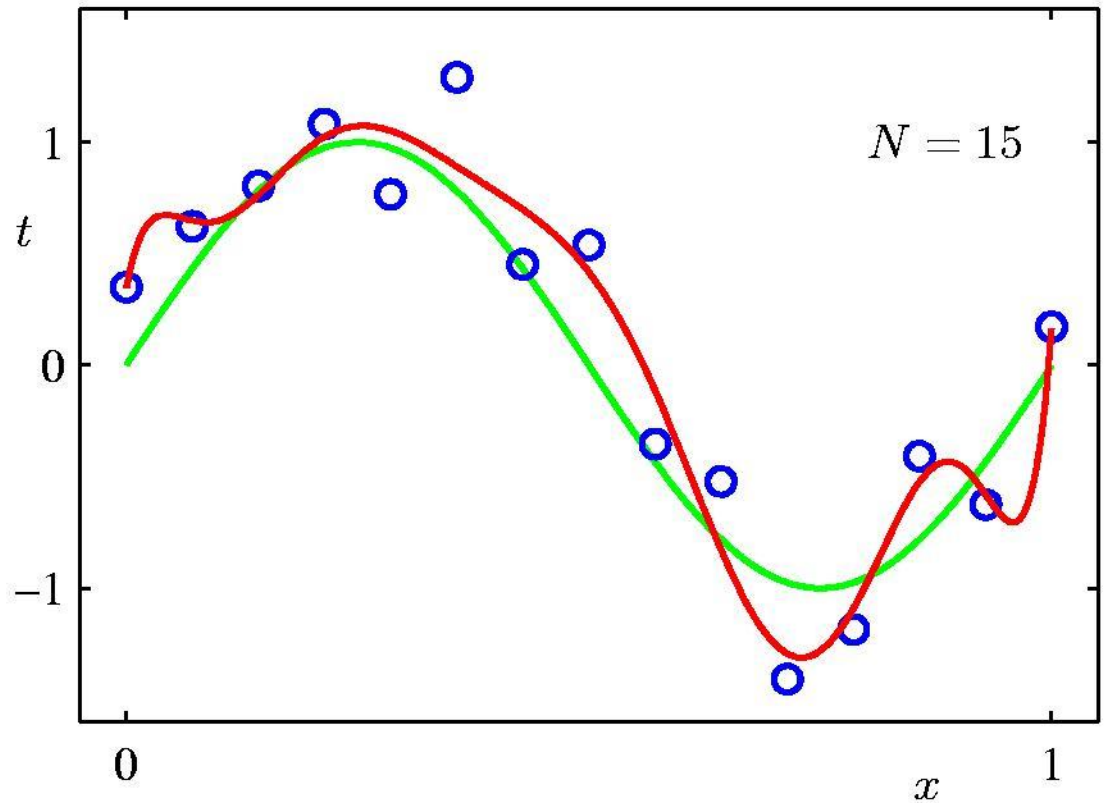
Root-Mean-Square (RMS) Error: $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^{\star})/N}$

Polynomial Coefficients

	$M=0$	$M=1$	$M=3$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

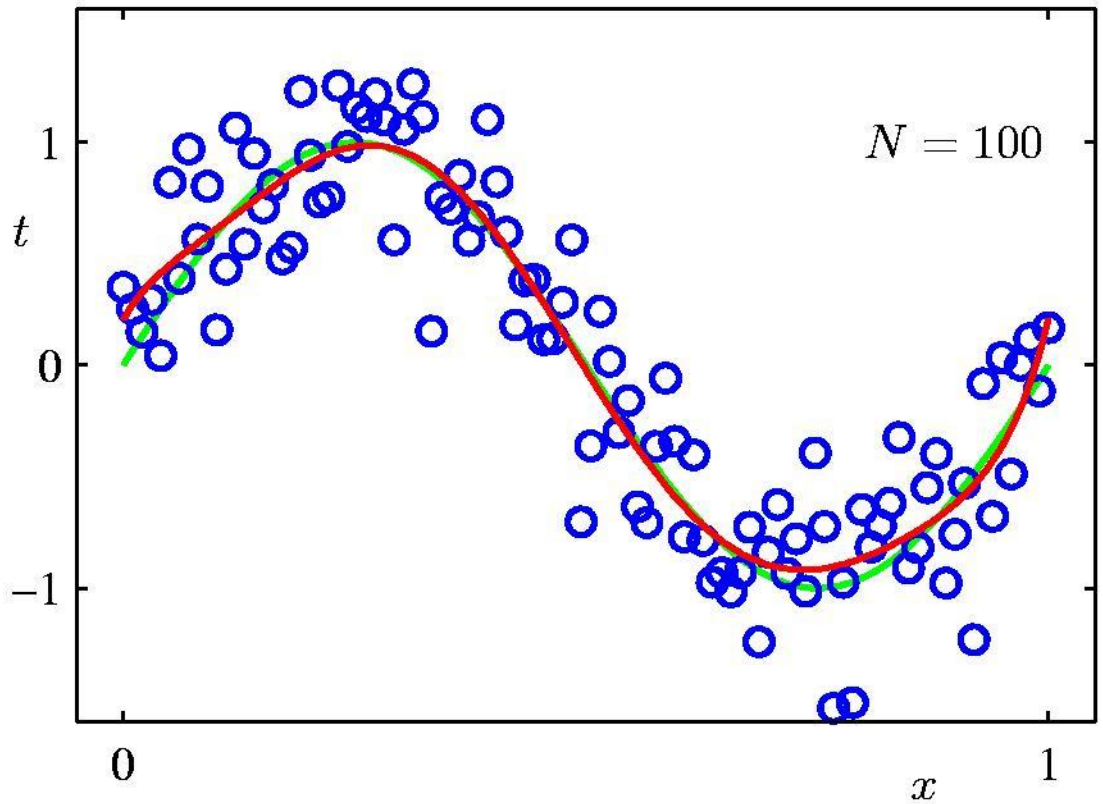
Data Set Size: $N = 15$

9th Order Polynomial



Data Set Size: $N = 100$

9th Order Polynomial

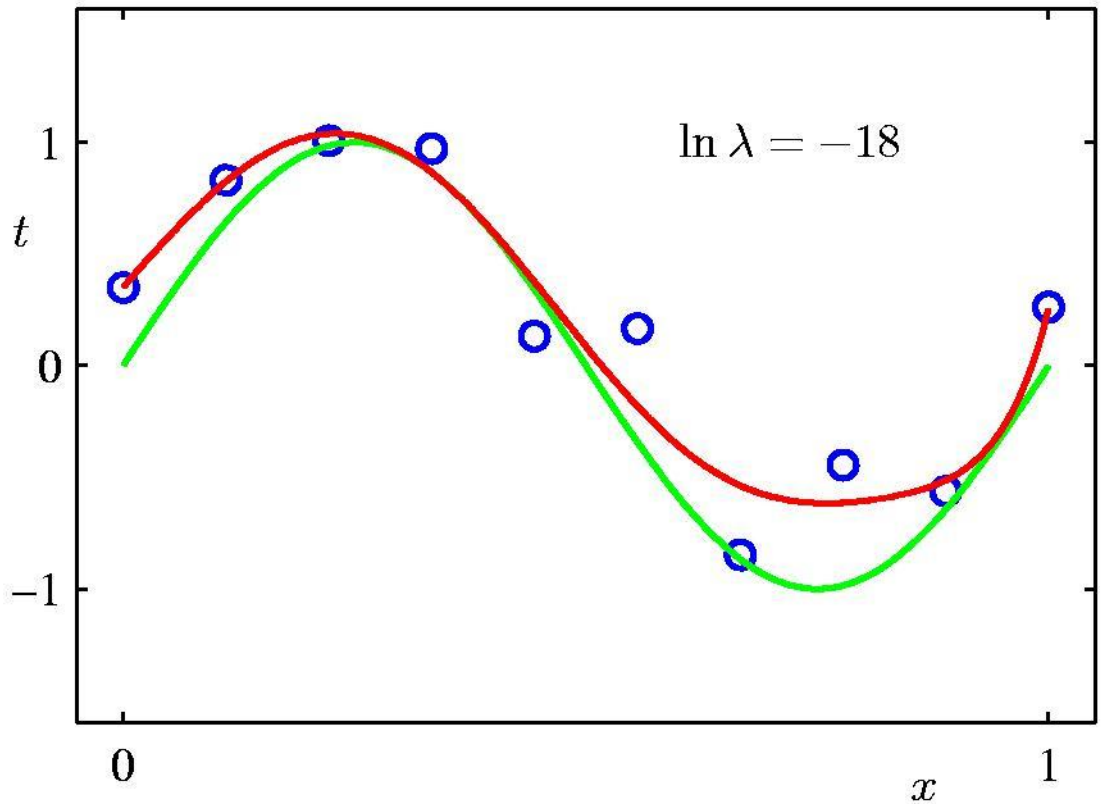


Regularization

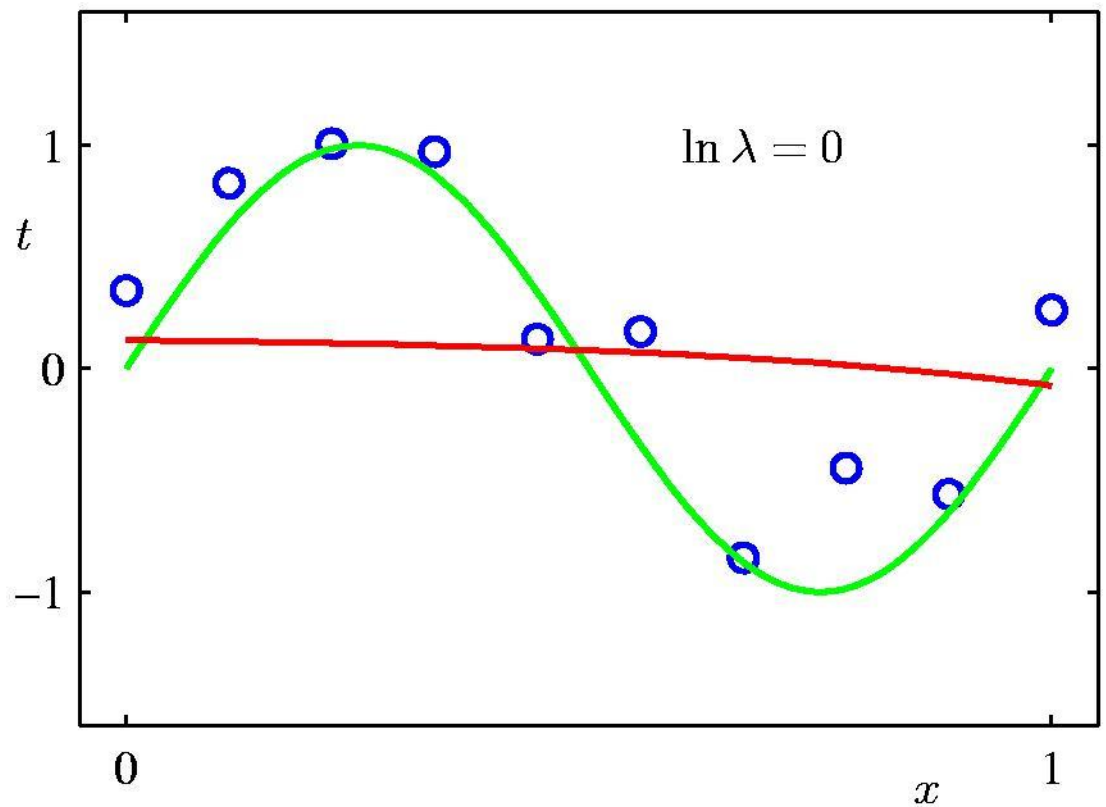
Penalize large coefficient values

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

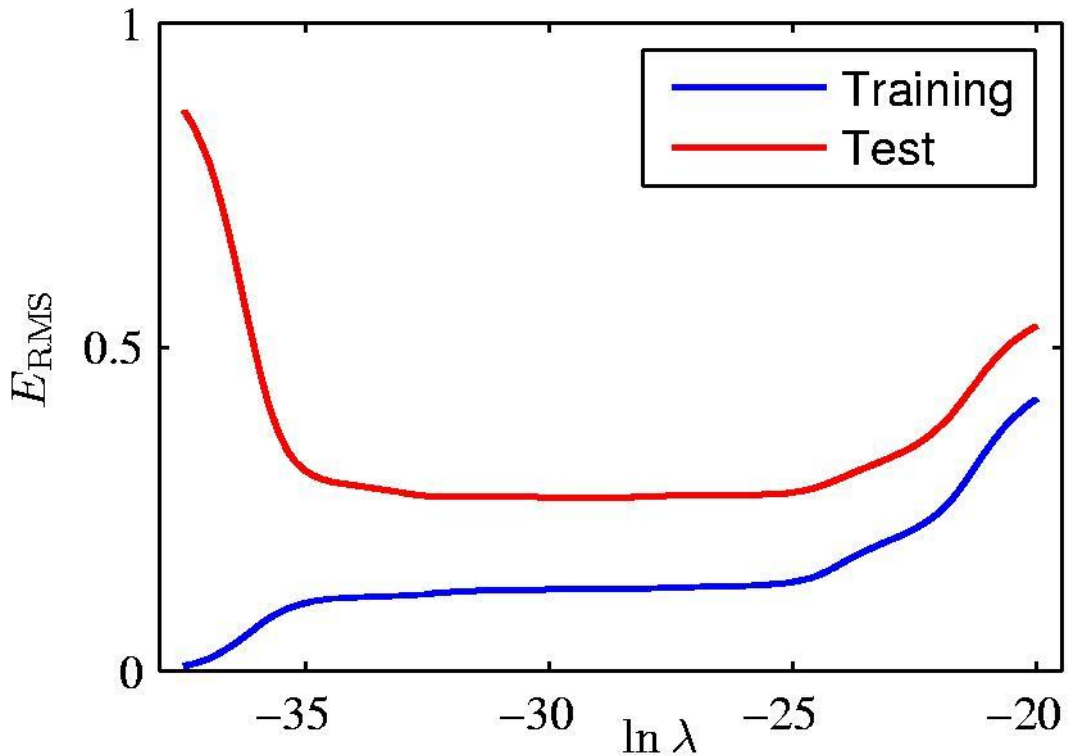
Regularization: $\ln \lambda = -18$



Regularization: $\ln \lambda = 0$



Regularization:

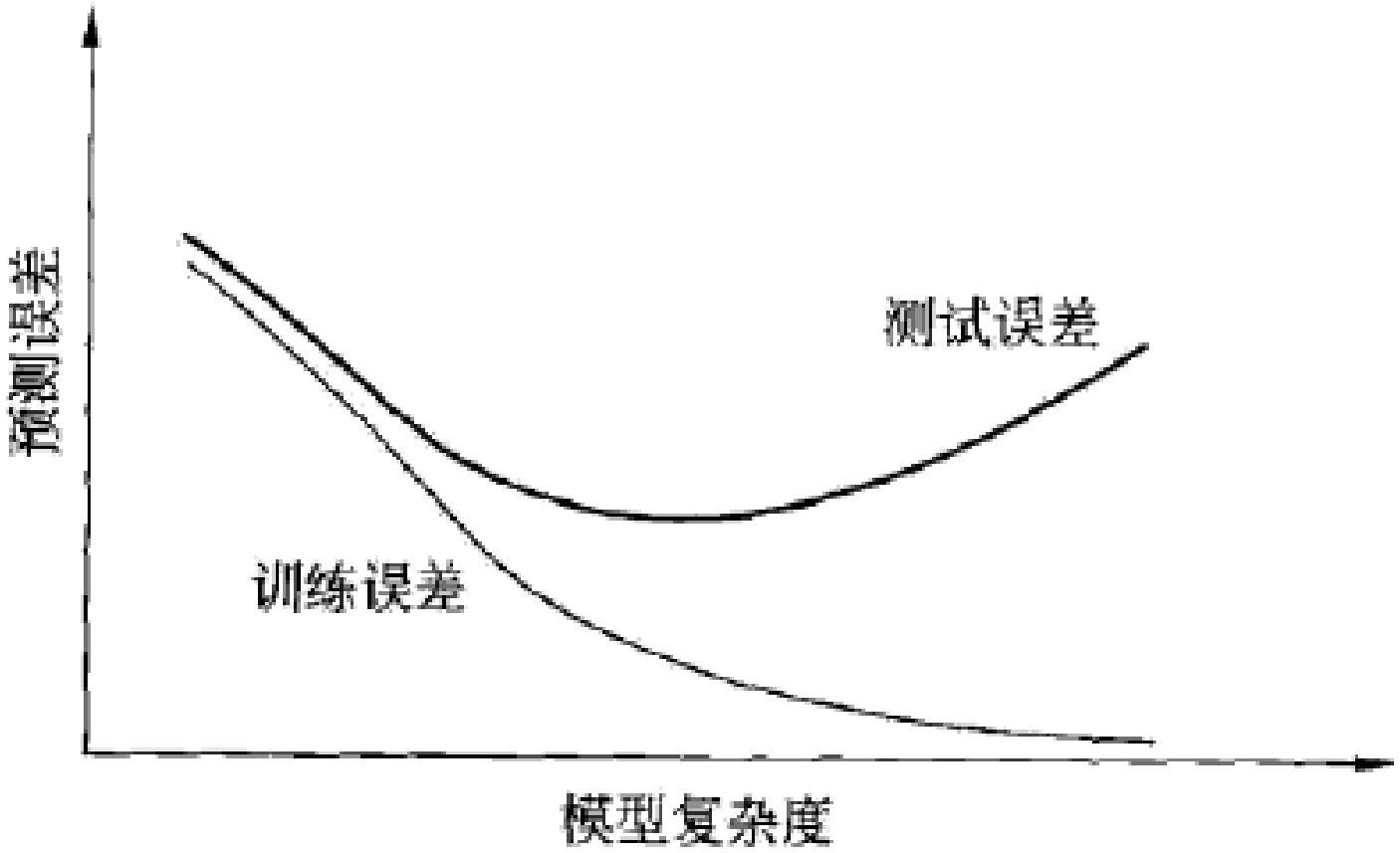


E_{RMS} vs. $\ln \lambda$

Polynomial Coefficients

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01

模型评估与模型选择



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

泛化能力 generalization ability

- 泛化误差 generalization error $R_{exp}(\hat{f}) = E_P[L(Y, \hat{f}(X))] = \int_{x \times y} L(y, \hat{f}(x)) P(x, y) \mathrm{d}x\mathrm{d}y$

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

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：

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

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

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

$$Y = f(X)$$

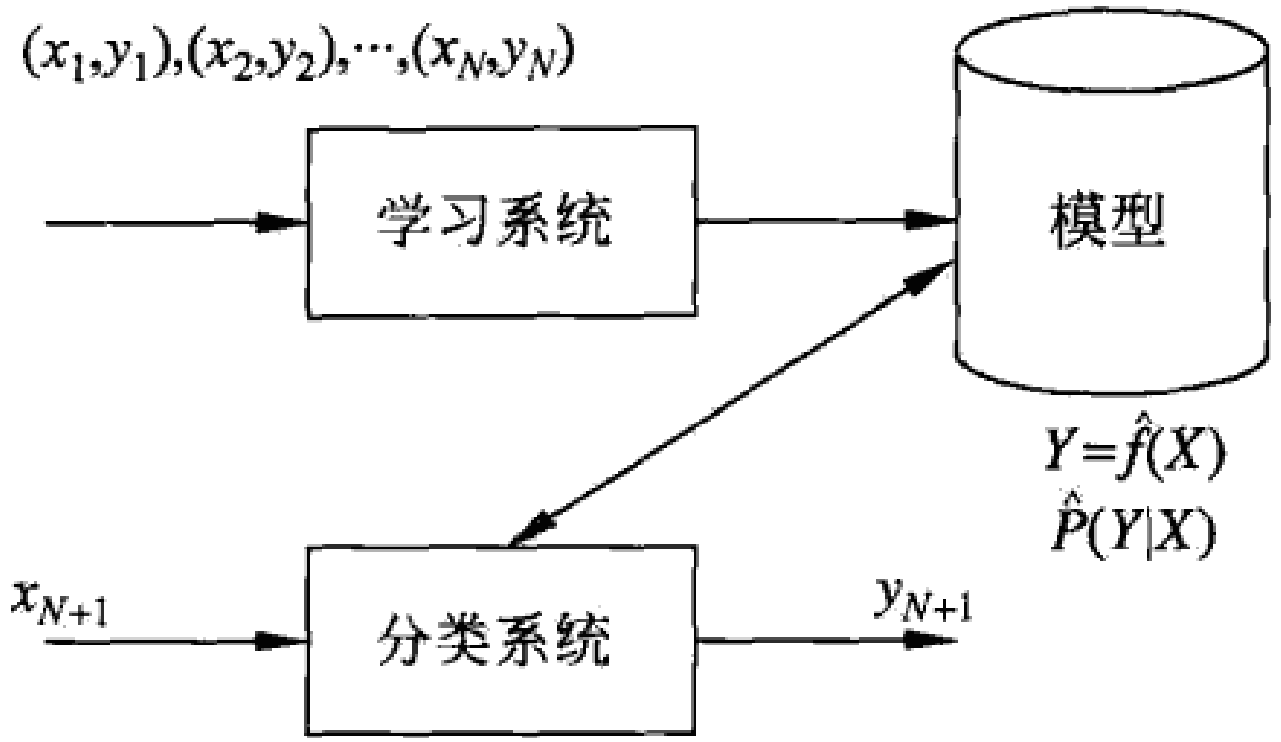
$$P(Y \mid X)$$

- 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。

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 $$R = \frac{TP}{TP + FN}$$
- F_1值 $$\frac{2}{F_1} = \frac{1}{P} + \frac{1}{R}$$

$$F_1 = \frac{2TP}{2TP + FP + FN}$$

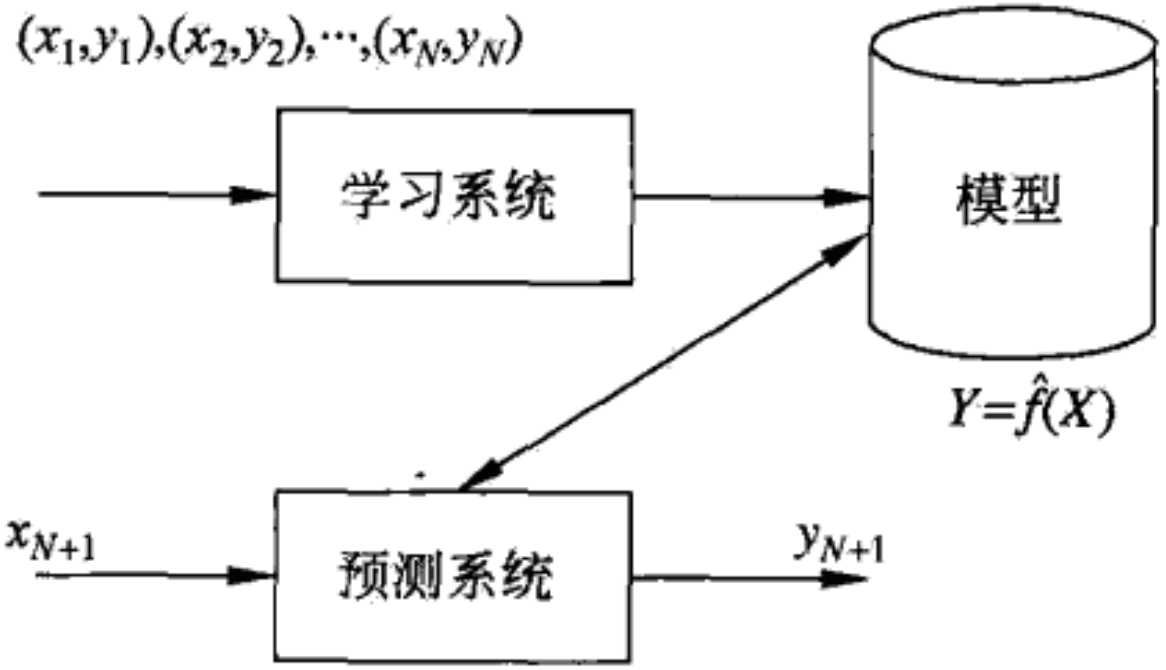
标注问题

- 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 $P(Y^{(1)},Y^{(2)},\cdots,Y^{(n)}\mid X^{(1)},X^{(2)},\cdots,X^{(n)})$

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?