大数据机器学习

第二讲：机器学习基本概念

袁春
清华大学深圳研究生院
2017/6

提纲

- 基本术语
- 监督学习
- 假设空间
- 学习三要素
- 奥卡姆剃刀定理
- 没有免费的午餐定理
- 训练误差和测试误差
- 正则化
- 泛化能力
- 生成模型与判别模型

基本术语

- Data set
  - 形状=圆形 剥皮=难 味道=酸甜
  - 形状=扁圆形 剥皮=易 味道=酸
  - 形状=长圆形 剥皮=难 味道=甜
  - 。。。。
- Instance/sample
- Attribute value/feature
- Attribute/feature space
- Feature vector

Tsinghua University 清華大學

基本术语

- $D=\{x_1,x_2,\ldots\ldots x_m\}$ m个示例的数据集
- $\boldsymbol{x}_i = (x_{i1}; x_{i2}; \ldots; x_{id})$ 是d维样本空间X的一个特征向量
- training/learning
- training data
- training sample
- Label ((形状=长圆形 剥皮=难 味道=甜), 橙子)
- example

基本术语

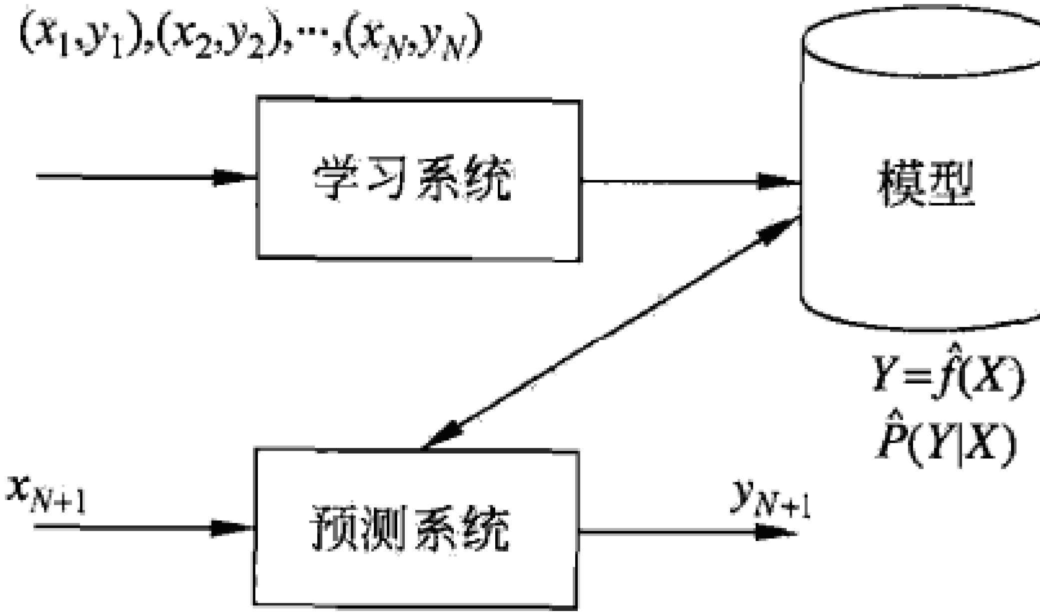
- Classification
- regression
- binary classification
- multi-class classification
- Clustering
- Multi-labeling annotation

监督学习

- 监督学习目的是学习一个由输入到输出的映射，称为模型
- 模型的集合就是假设空间（hypothesis space）
- 模型：
  - 概率模型:条件概率分布P(Y|X),
  - 非概率模型：决策函数Y=f(X)
- 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)

监督学习

- 问题的形式化



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

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

假设空间 hypothesis space

- 学习过程： 搜索所有假设空间，与训练集匹配
  - 形状=圆形 剥皮=难 味道=酸甜 橙
  - 形状=扁圆形 剥皮=易 味道=酸 橘
  - 形状=长圆形 剥皮=难 味道=甜 橙

→

- 形状= * 剥皮=难 味道=* 橙
- 形状=扁圆形 剥皮=易 味道=* 橘

- 假设形状，剥皮，味道 分别有3，2，3 种可能取值，加上取任意值*和空集，假设空间规模4x3x4+1=49
- Version space：

学习三要素

- 学习三要素： 方法＝模型＋策略＋算法
  - 当假设空间F为决策函数的集合： $\mathcal{F} = \{f \mid Y = f(X)\}$
  - F实质为参数向量决定的函数族： $\mathcal{F} = \{f \mid Y = f_\theta(X), \theta \in \mathbf{R}^n\}$
  - 当假设空间F为条件概率的集合： $\mathcal{F} = \{P \mid P(Y \mid X)\}$
  - F实质是参数向量决定的条件概率分布族 $\mathcal{F} = \{P \mid P_\theta(Y \mid X), \theta \in \mathbf{R}^n\}$

学习三要素

- 策略
  - 损失函数和风险函数
  - 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

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

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

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

学习三要素

- 策略
  - 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

  - 经验风险 empirical risk， 经验损失 empirical loss

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

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

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”
  - 为防止过拟合提出的策略，结构风险最小化 structure risk minimization，等价于正则化（regularization），加入正则化项regularizer，或罚项 penalty term：

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

学习三要素

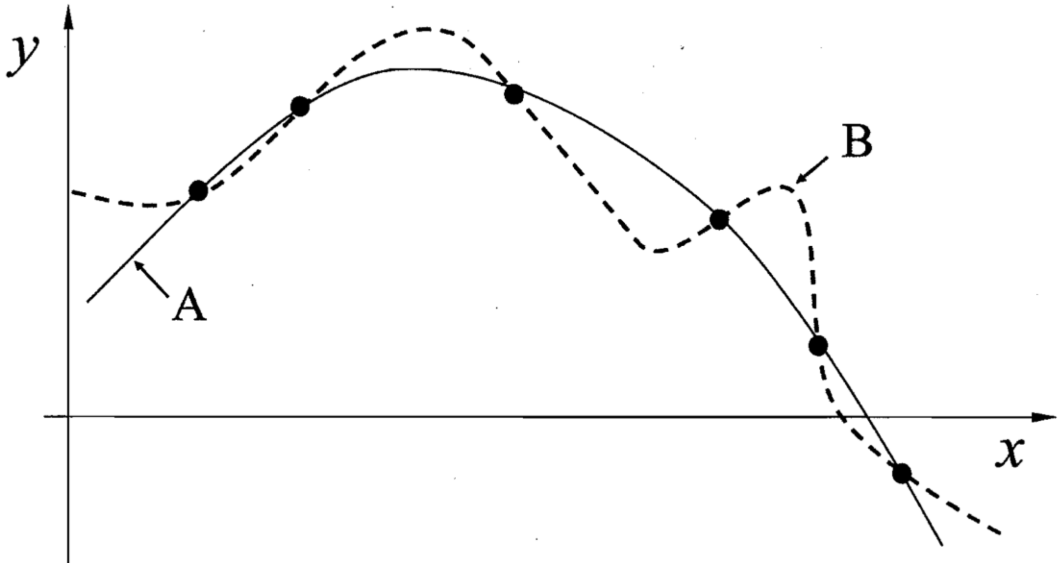
- 方法：求最优模型就是求解最优化问题：

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

- 难点：
  - 全局最优
  - 高效

奥卡姆剃刀Occam’s razor

- 14世纪逻辑学家、圣方济各会修士奥卡姆的威廉（William of Occam，约1285年至1349年）
- 原理称为“如无必要，勿增实体”

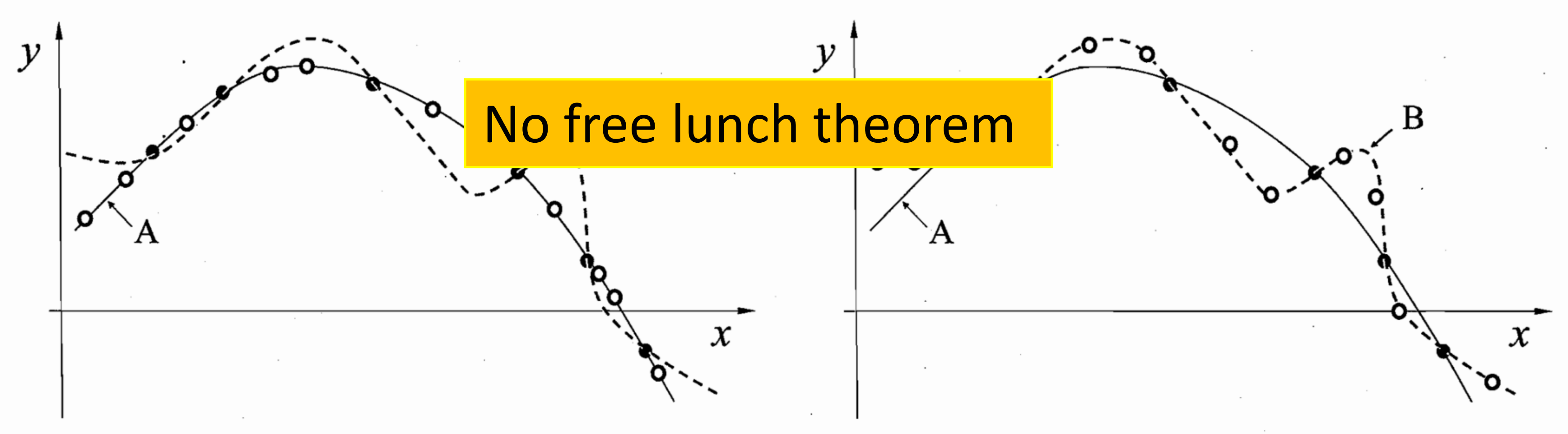


奥卡姆剃刀Occam’s razor

- 疑问一： 哪个更简单？

- 形状= * 剥皮=难 味道=* 橙
- 形状=长圆形 剥皮=* 味道=* 橙

- 疑问二：



(a) A 优于 B　　(b) B 优于 A

No free lunch theorem

No free lunch theorem

- A 好？　B 好？　随机胡猜好?
- 假设样本空间X和假设空间H都是离散的.
- P(h| X , Za)：产生假设 h 的概率
- f(x): 真实目标函数
- “训练集外误差”

$$E_{ote}(\mathfrak{L}_a|X,f)=\sum_{h}\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\,\mathbb{I}(h(\boldsymbol{x})\neq f(\boldsymbol{x}))\,P(h\mid X,\mathfrak{L}_a)$$

No free lunch theorem

- 二分类问题：

总误差竟然与学习算法无关

$$
\begin{aligned}
\sum_f E_{ote}(\mathfrak{L}_a|X,f) &= \sum_f \sum_h \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x})\ \mathbb{I}(h(\boldsymbol{x})\neq f(\boldsymbol{x}))\ P(h\mid X,\mathfrak{L}_a) \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h\mid X,\mathfrak{L}_a) \sum_f \mathbb{I}(h(\boldsymbol{x})\neq f(\boldsymbol{x})) \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h\mid X,\mathfrak{L}_a) \frac{1}{2} 2^{|\mathcal{X}|} \\
&= \frac{1}{2} 2^{|\mathcal{X}|} \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h\mid X,\mathfrak{L}_a) \\
&= 2^{|\mathcal{X}|-1} \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \cdot 1
\end{aligned}
$$

No free lunch theorem

- N F L 定理前提条件：
  - 所有“问题”出现的机会相同，或所有问题同等重要
  - 假设真实函数 f 的均匀分布。

> - 形状= * 剥皮=难 味道=* 橙
> - 形状=长圆形 剥皮=* 味道=* 橙

  - N F L 寓意：脱离具体问题，空谈“什么方法好”毫无意义。

训练误差和测试误差

- 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：

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

过拟合

- 过拟合与模型选择 - 多项式曲线拟合的例子
- 假设给定训练数据集

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

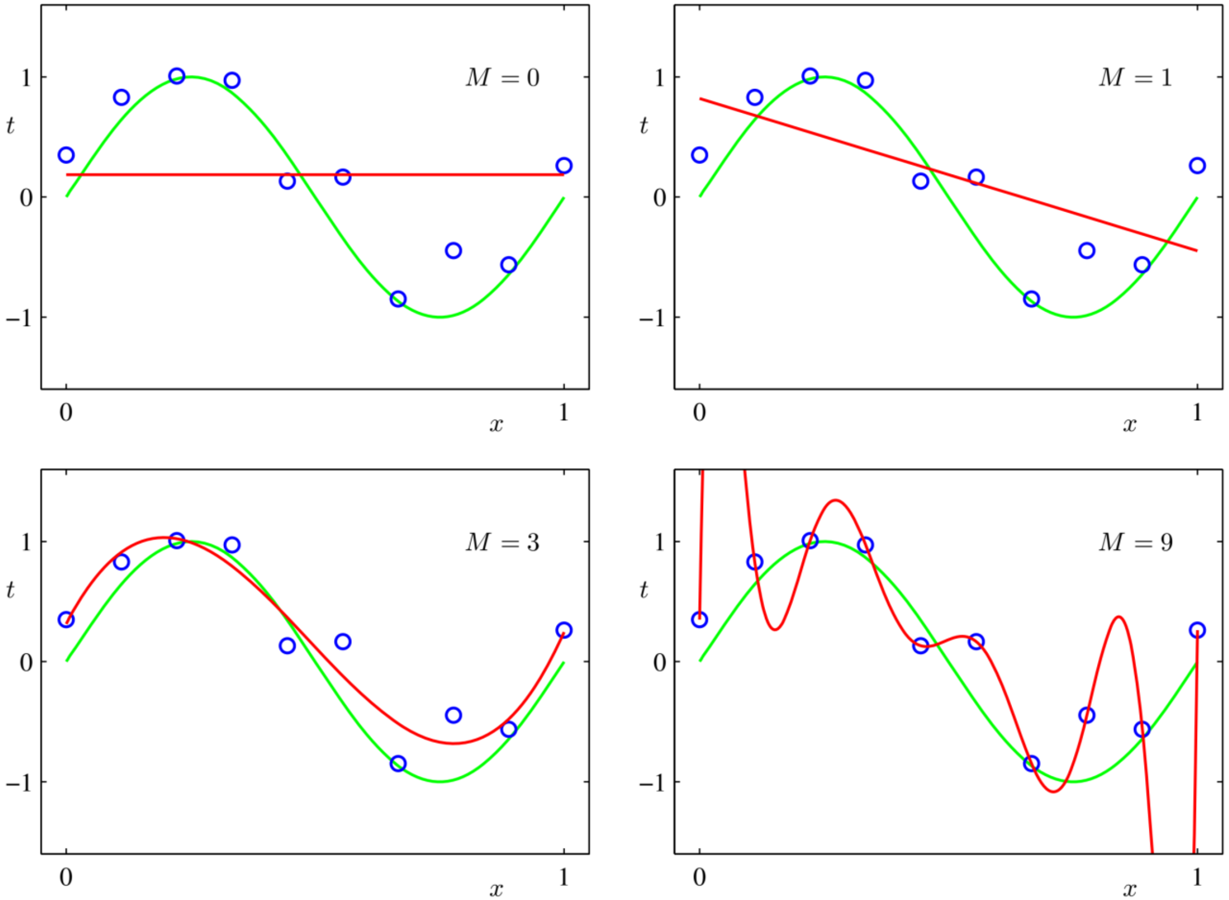
$$f_M(x,w)=w_0+w_1x+w_2x^2+\cdots+w_Mx^M=\sum_{j=0}^{M}w_jx^j$$

- 经验风险最小：

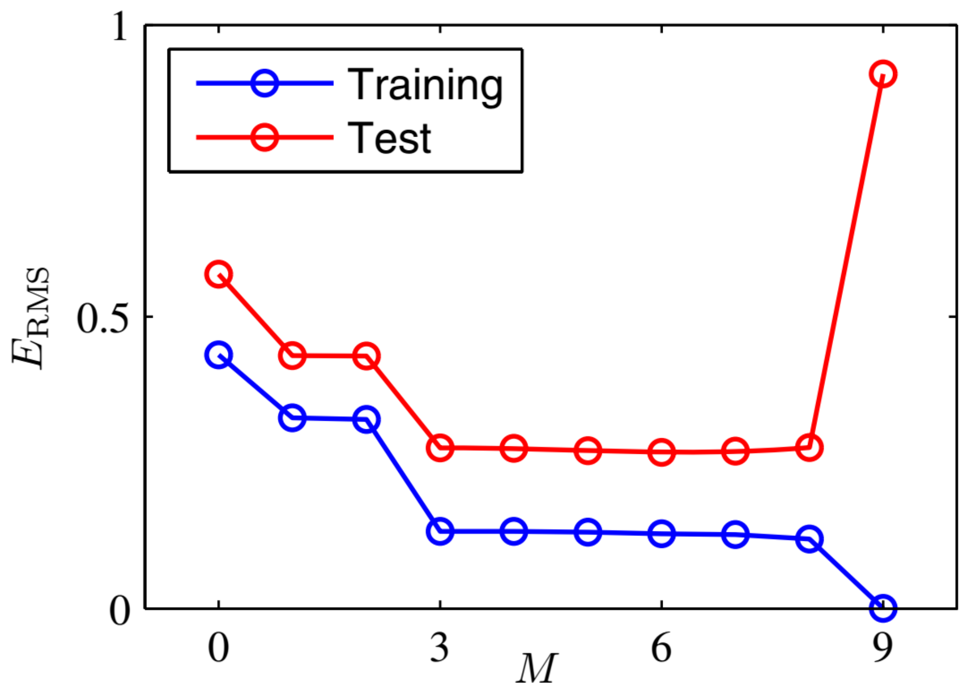
$$L(w)=\frac{1}{2}\sum_{i=1}^{N}(f(x_i,w)-y_i)^2 \qquad L(w)=\frac{1}{2}\sum_{i=1}^{N}\left(\sum_{j=0}^{M}w_jx_i^j-y_i\right)^2$$

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

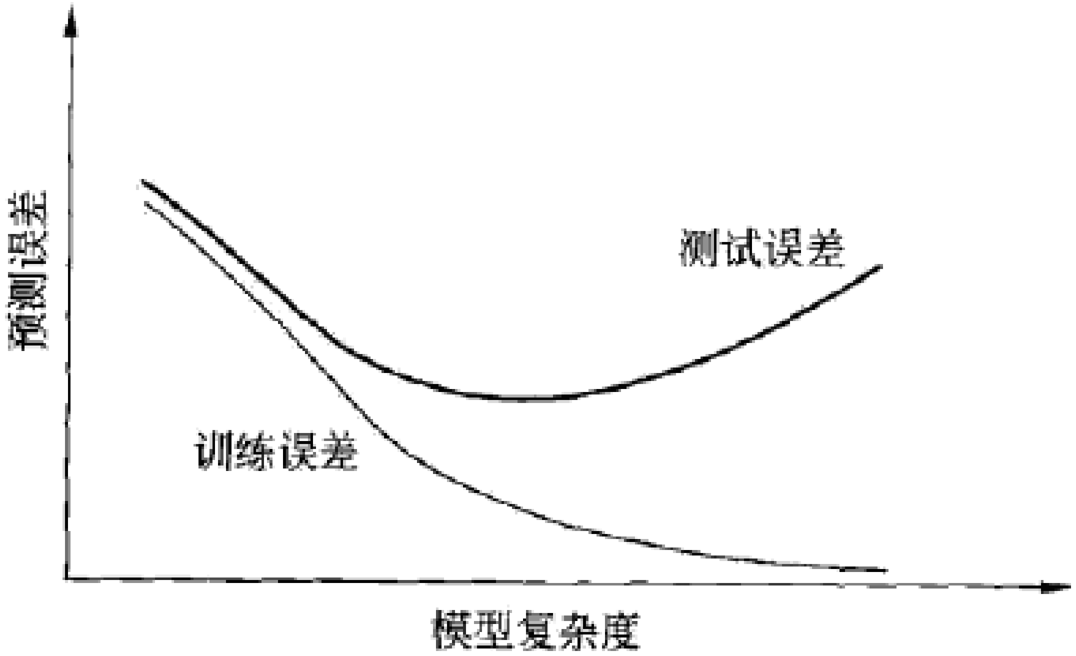
过拟合



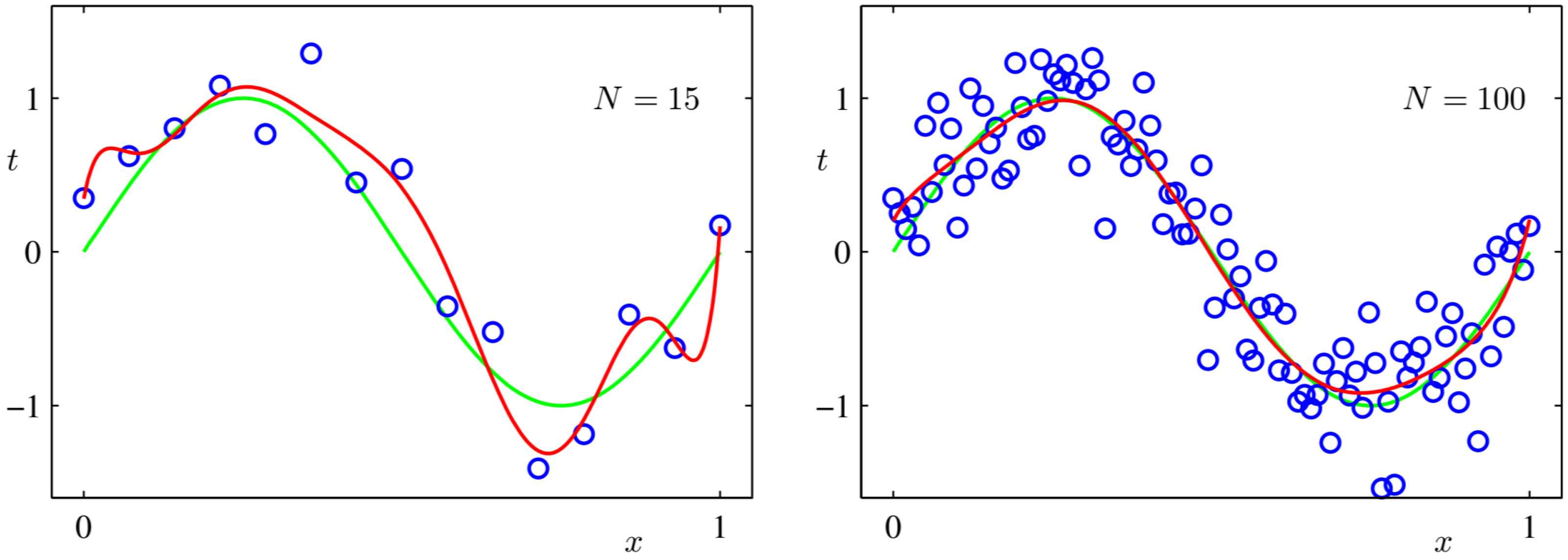
过拟合



过拟合



过拟合



正则化

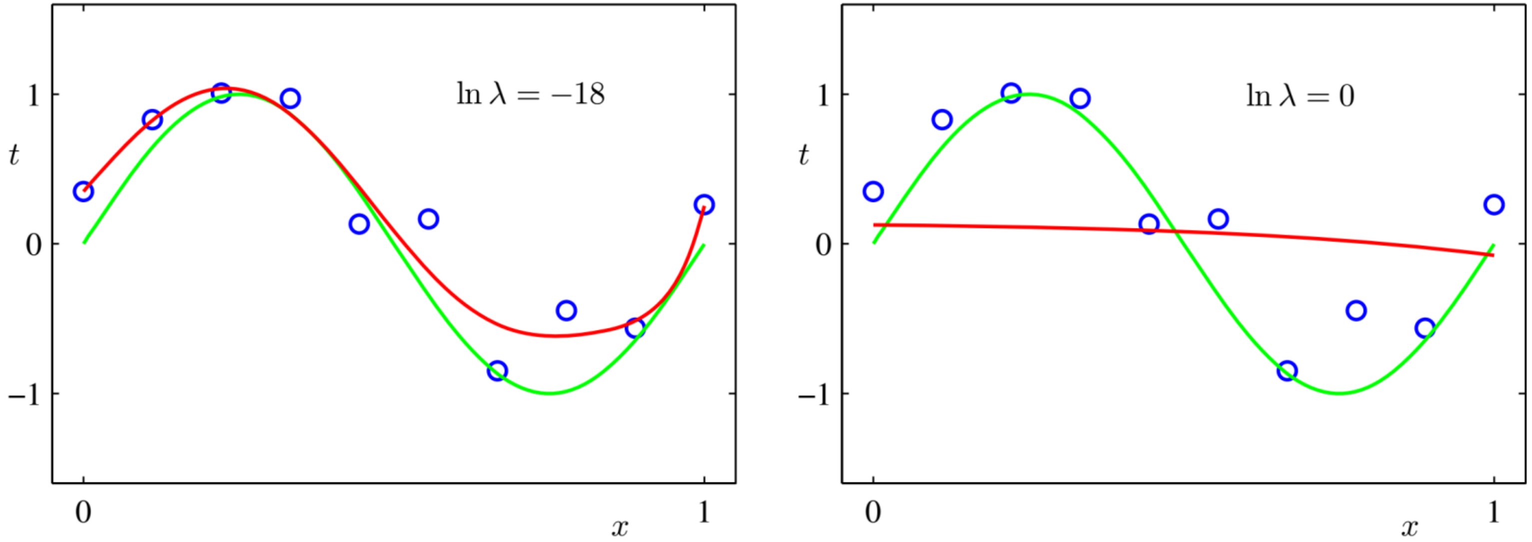
- 正则化一般形式：
$$\min_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f)$$

- 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

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



正则化

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

	$M=0$	$M=1$	$M=6$	$M=9$
$w_0^\star$	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

	$\ln\lambda=-\infty$	$\ln\lambda=-18$	$\ln\lambda=0$
$w_0^\star$	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

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 $R(f) = E[L(Y,f(X))]$
  - 假设空间F为有限集合

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

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
- 判别方法discriminative approach对应判别模型：discriminative model，
  - K近邻，感知机，决策树，logistic 回归等

生成模型与判别模型

- 二者各有优缺点
- **生成模型**：
  - 还原联合概率，而判别模型不能；
  - 学习收敛速度快，当样本容量增加时，学到的模型可以更快收敛；
  - 当存在隐变量时，可以使用生成模型，而判别模型不行。
- **判别模型**：
  - 直接学习决策函数或条件概率，学习的准确率更高；
  - 可以对数据进行抽象，定义特征和使用特征，可以简化学习问题。

Q&A?