杭州电子科技大学
HANGZHOU DIANZI UNIVERSITY

笃学力行 守正求新

机器学习 & 深度学习

模型评估与选择

高 飞 Fei Gao

gaofei@hdu.edu.cn

优化问题

机器学习的三要素

- 模型
  - 线性方法 $$f(\mathbf{x},\theta)=\mathbf{w}^{\mathrm{T}}\mathbf{x}+b$$
  - 广义线性方法 $$f(\mathbf{x},\theta)=\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x})+b$$
    - 如果 $\phi(\mathbf{x})$ 为可学习的非线性基函数，则该方法等价于神经网络
- 学习准则/目标函数

$$\mathcal{R}(f)=\mathbb{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y)}[\mathcal{L}(f(\mathbf{x}),y)],$$

- 优化/学习方法
  - 梯度下降等

参数学习

ML·DL

- 期望风险未知，通过经验风险近似
  - 经验风险：训练数据上的风险/误差

$$\mathcal{R}_{\mathcal{D}}^{emp}(\theta) = \frac{1}{N}\sum_{n=1}^{N}\mathcal{L}(y^{(n)}, f(x^{(n)}, \theta))$$

- 经验风险最小化
  - 在选择合适的风险函数后，寻找最优参数，使经验风险最小

$$\theta^* = \arg\min_{\theta} \mathcal{R}_{\mathcal{D}}^{emp}(\theta)$$

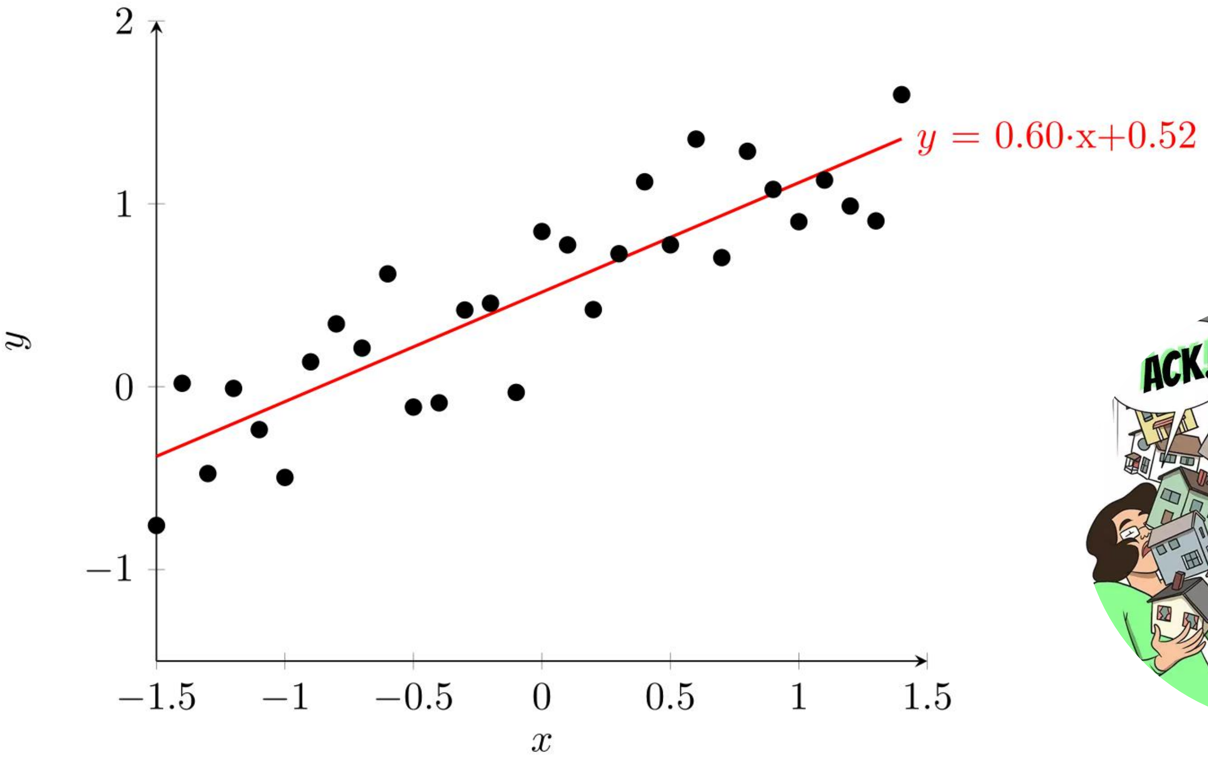
- 机器学习问题转化为最优化问题

线性回归（Linear Regression）

- 模型：

$$f(\mathbf{x}; \mathbf{w}, b) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b$$

  - 增广权重向量和增广特征向量 $f(\mathbf{x}; \hat{\mathbf{w}}) = \hat{\mathbf{w}}^{\mathrm{T}}\hat{\mathbf{x}}$



损失函数（Loss Function）

- 均方损失（Squared Loss）：回归问题

$$\mathcal{L}(y,\hat{y}) = (y - f(x,\theta))^2$$

- 绝对值损失（Absolute Loss）：回归问题

$$\ell(y,\hat{y}) = |y - \hat{y}|$$

- 二值损失（Binary Loss）：分类问题

$$\mathcal{L}(y, f(x,\theta)) = \begin{cases} 0 & \textbf{if } y = f(x,\theta) \\ 1 & \textbf{if } y \neq f(x,\theta) \end{cases}$$

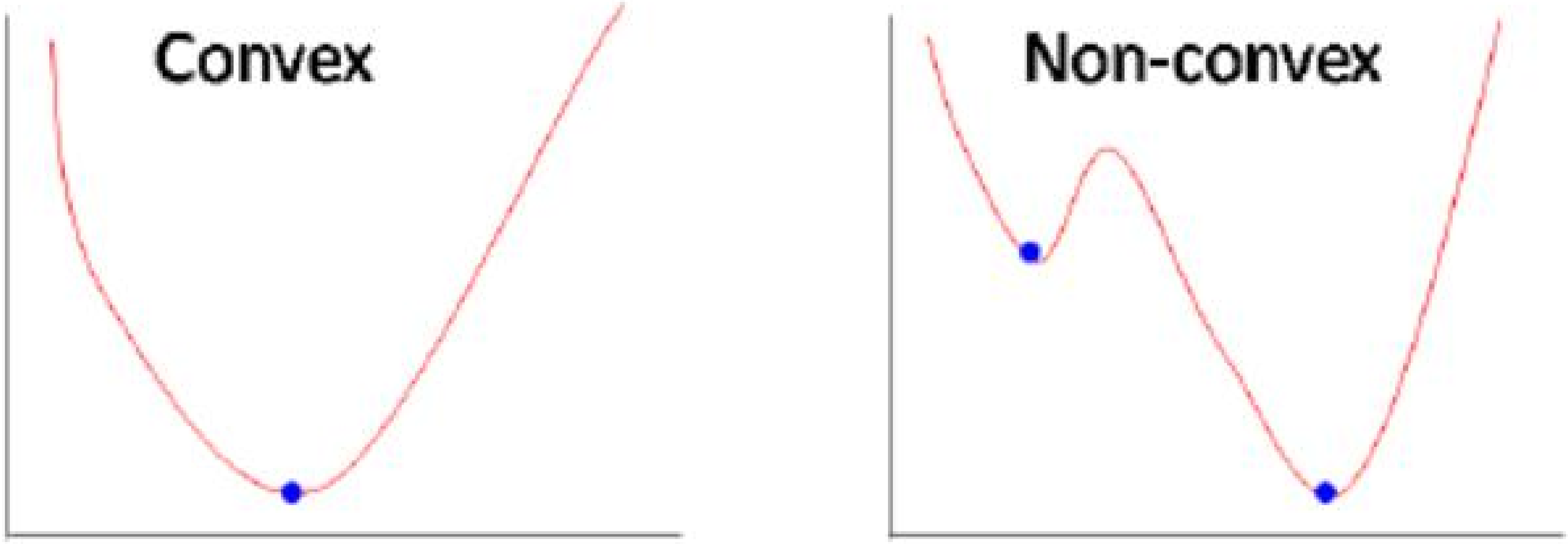
Regression: **squared loss** $\ell(y,\hat{y}) = (y-\hat{y})^2$
or **absolute loss** $\ell(y,\hat{y}) = |y-\hat{y}|$.

Binary Classification: **zero/one loss** $\ell(y,\hat{y}) = \begin{cases} 0 & \text{if } y = \hat{y} \\ 1 & \text{otherwise} \end{cases}$

Multiclass Classification: also zero/one loss.

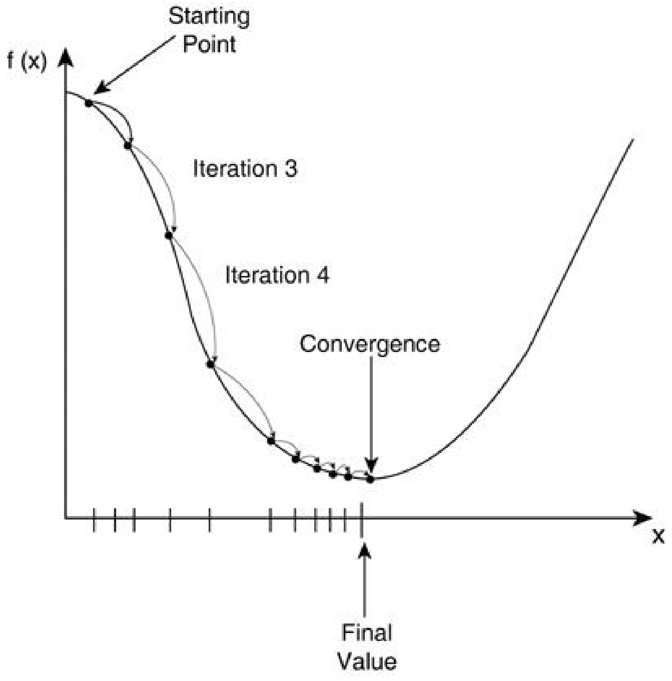
最优化问题

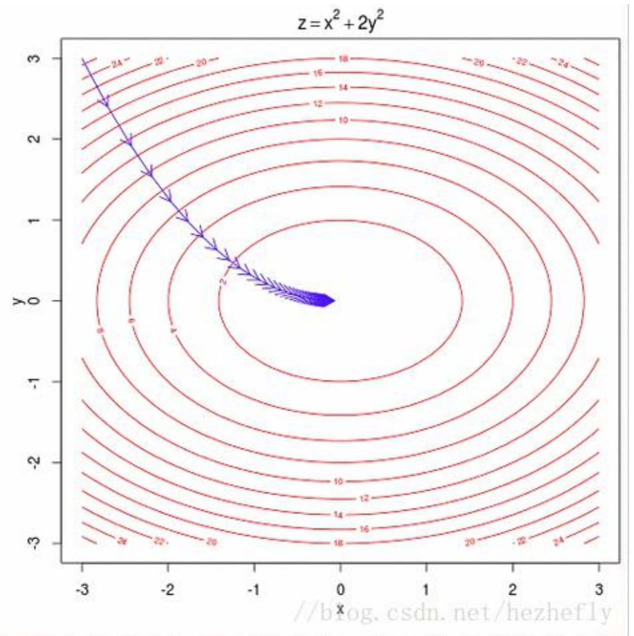
- 机器学习问题转化成为一个最优化问题



$$\min_{\mathbf{x}} \quad f(\mathbf{x})$$

梯度下降法（Gradient Descent）





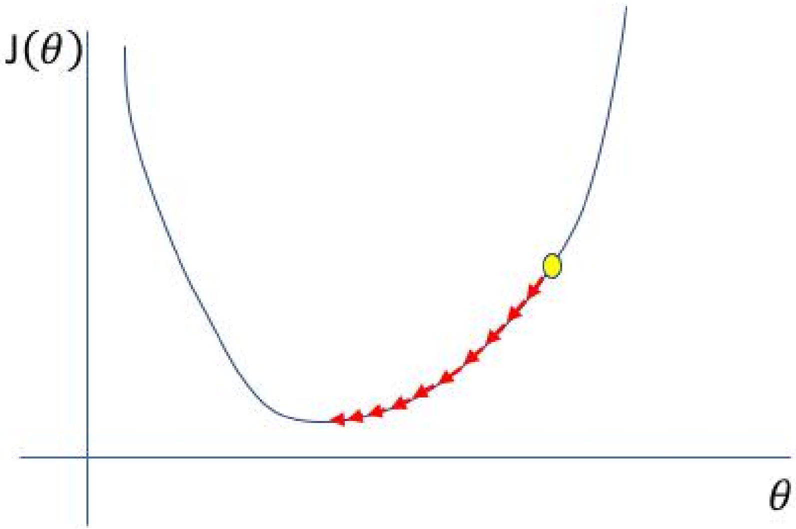
$$\theta_{t+1} = \theta_t - \alpha \frac{\partial \mathcal{R}(\theta)}{\partial \theta_t}$$

$$= \theta_t - \alpha \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \mathcal{L}(\theta_t; x^{(i)}, y^{(i)})}{\partial \theta}.$$

搜索步长α中也叫作学习率（Learning Rate）

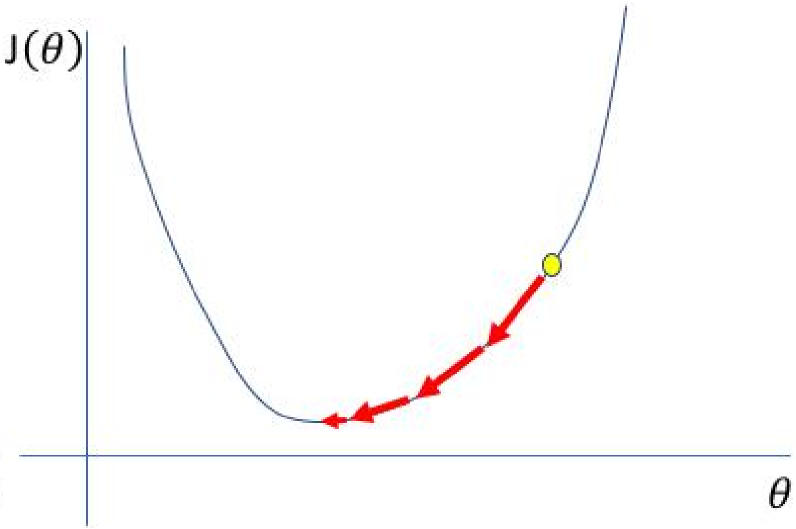
学习率是十分重要的超参数！

Too low



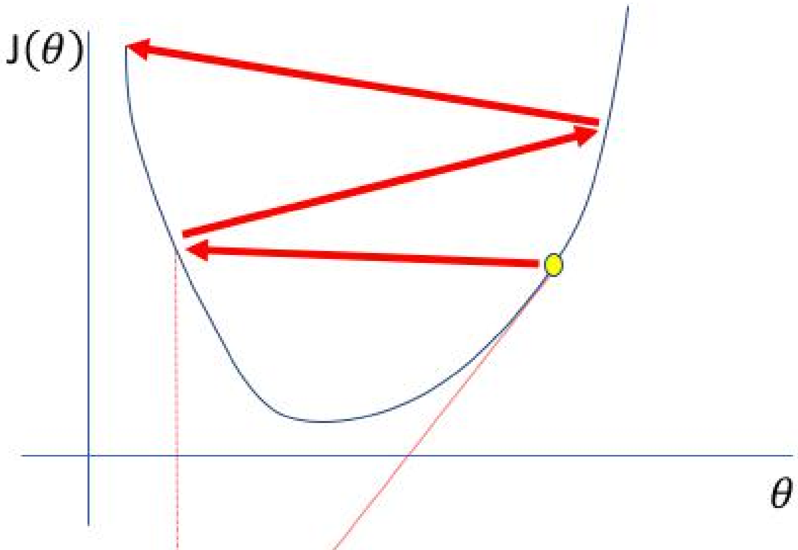
A small learning rate requires many updates before reaching the minimum point

Just right



The optimal learning rate swiftly reaches the minimum point

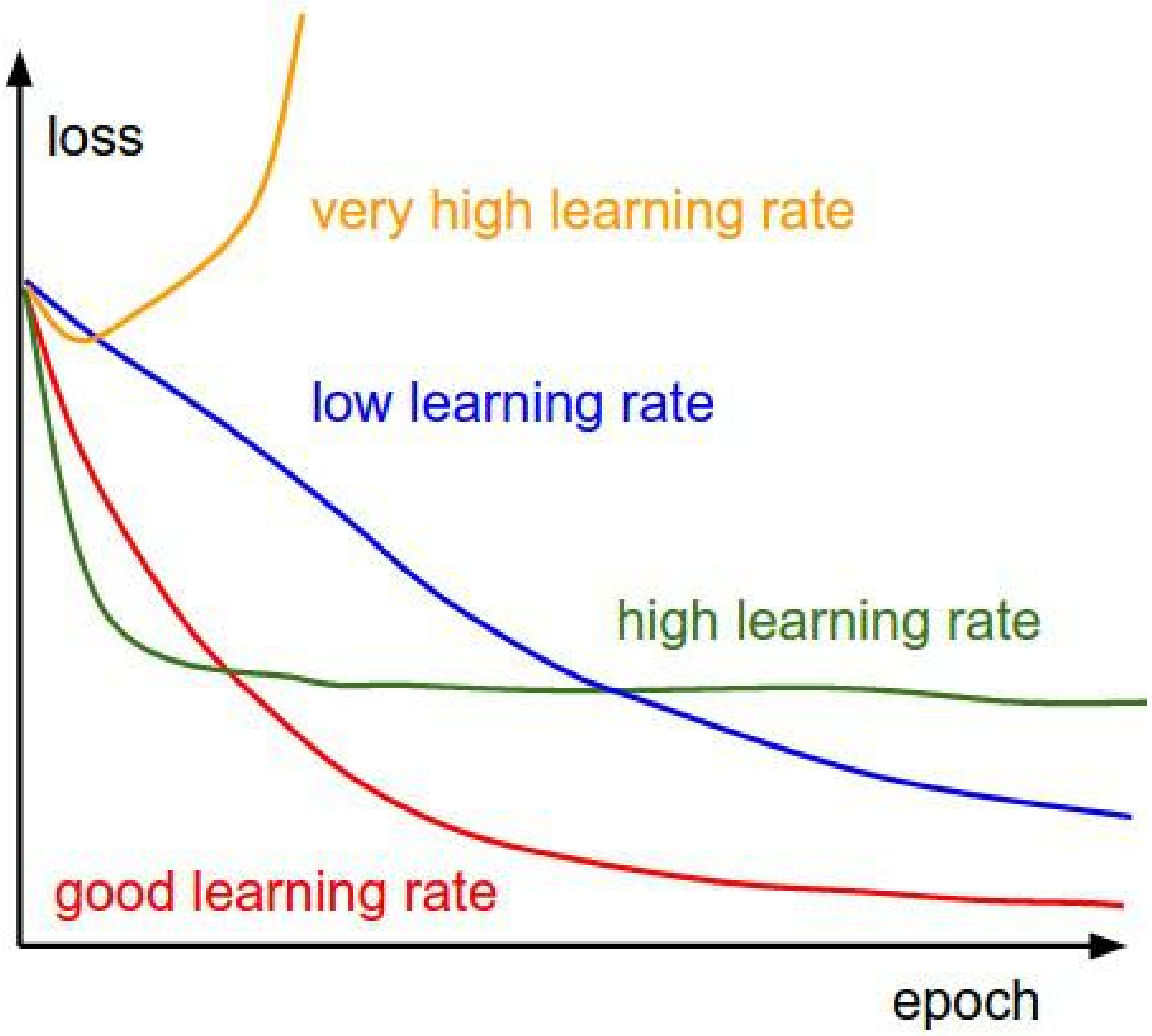
Too high



Too large of a learning rate causes drastic updates which lead to divergent behaviors

学习率是十分重要的超参数！

ML·DL



随机梯度下降法

- 随机梯度下降法（Stochastic Gradient Descent，SGD）也叫增量梯度下降，每个样本都进行更新

$$\theta \leftarrow \theta - \alpha \frac{\partial \mathcal{L}\left(\theta; x^{(n)}, y^{(n)}\right)}{\partial \theta}$$

  - 随机梯度下降相当于在批量梯度下降的梯度上引入了随机噪声。当目标函数非凸时，反而可以使其逃离局部最优点。

- 小批量（Mini-Batch）随机梯度下降法

$$\theta_{t+1} \leftarrow \theta_t - \alpha \cdot \frac{1}{K} \sum_{(\mathbf{x},y)\in \mathcal{I}_t} \frac{\partial \mathcal{L}\Big(y, f(\mathbf{x}, \theta)\Big)}{\partial \theta}.$$

K 通常不会设置很大，一般在 $1 \sim 100$ 之间。在实际应用中为了提高计算效率，通常设置为 2 的 n 次方。

随机梯度下降法

算法 2.1: 随机梯度下降法

输入: 训练集 $\mathcal{D}=\{(\mathbf{x}^{(n)},y^{(n)})\}_{n=1}^{N}$，验证集 $\mathcal{V}$，学习率 α

随机初始化 θ;

repeat

  对训练集 $\mathcal{D}$ 中的样本随机重排序;

  for $n=1\cdots N$ **do**

    从训练集 $\mathcal{D}$ 中选取样本 $(\mathbf{x}^{(n)},y^{(n)})$;

    // 更新参数

    $\theta \leftarrow \theta - \alpha\dfrac{\partial\mathcal{L}\big(\theta;x^{(n)},y^{(n)}\big)}{\partial\theta}$;

  end

until 模型 $f(\mathbf{x},\theta)$ 在验证集 $\mathcal{V}$ 上的错误率不再下降;

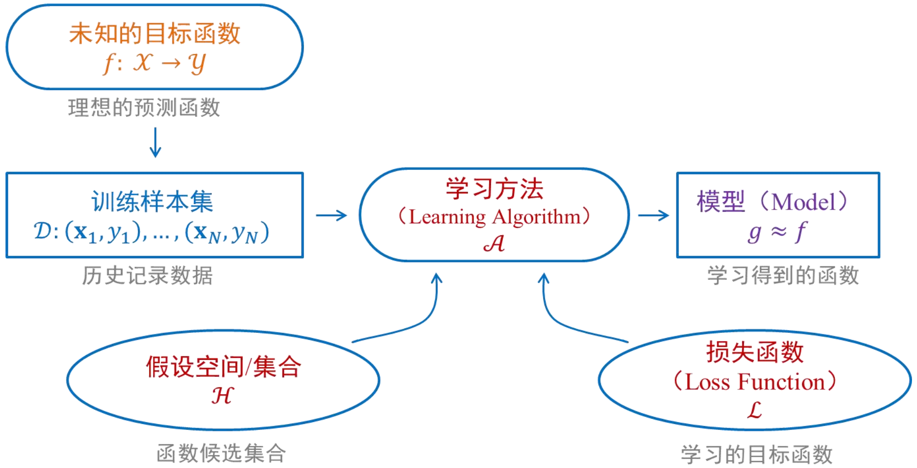
输出: θ

- 优化（Optimization）
  - 理想情形：最小化期望误差（Expected Error）

$$\epsilon \triangleq \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(y,f(x))] = \sum_{(x,y)} \mathcal{D}(x,y)\ell(y,f(x))$$

  - 学习过程，最小化训练误差（Training Error）

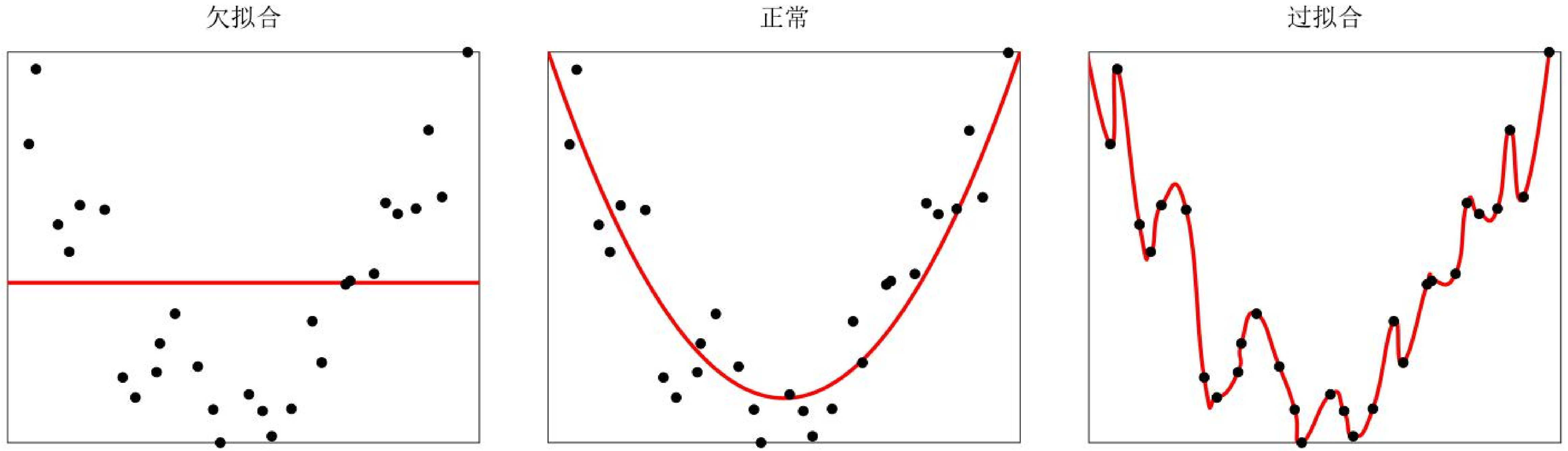
$$\hat{\epsilon} \triangleq \frac{1}{N}\sum_{n=1}^{N} \ell(y_n, f(x_n))$$



  - 问题
    - 过拟合（Overfitting）
    - 欠拟合（Underfitting）

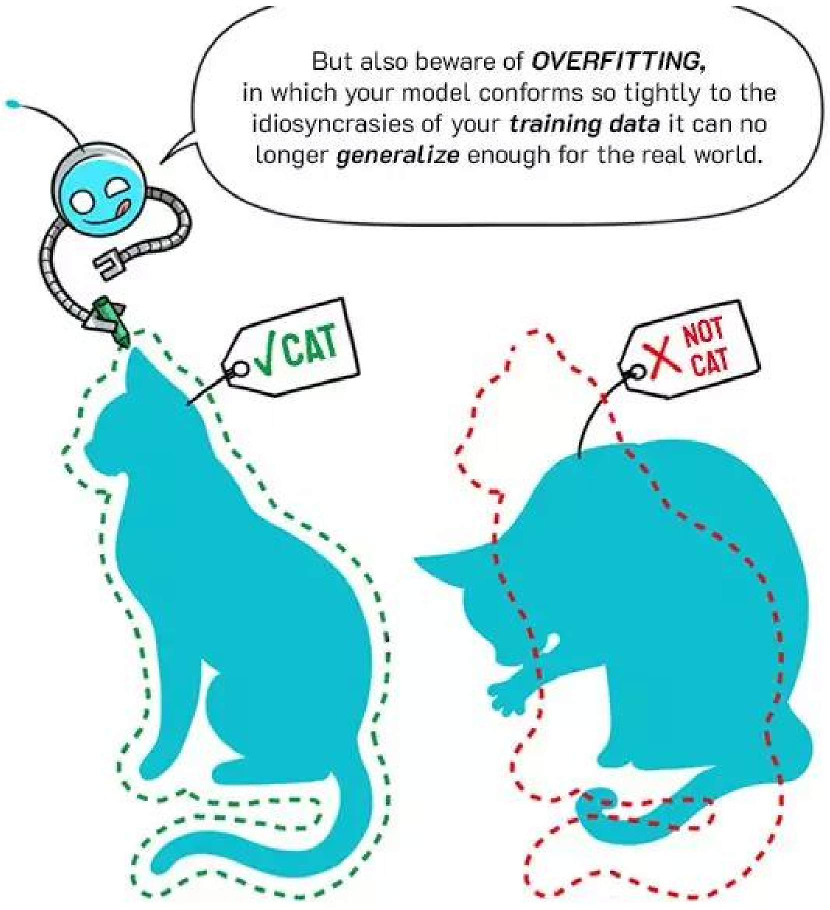
机器学习 = 优化?

- 过拟合:
  - 经验风险最小化原则很容易导致模型在训练集上错误率很低，但是在未知数据上错误率很高。



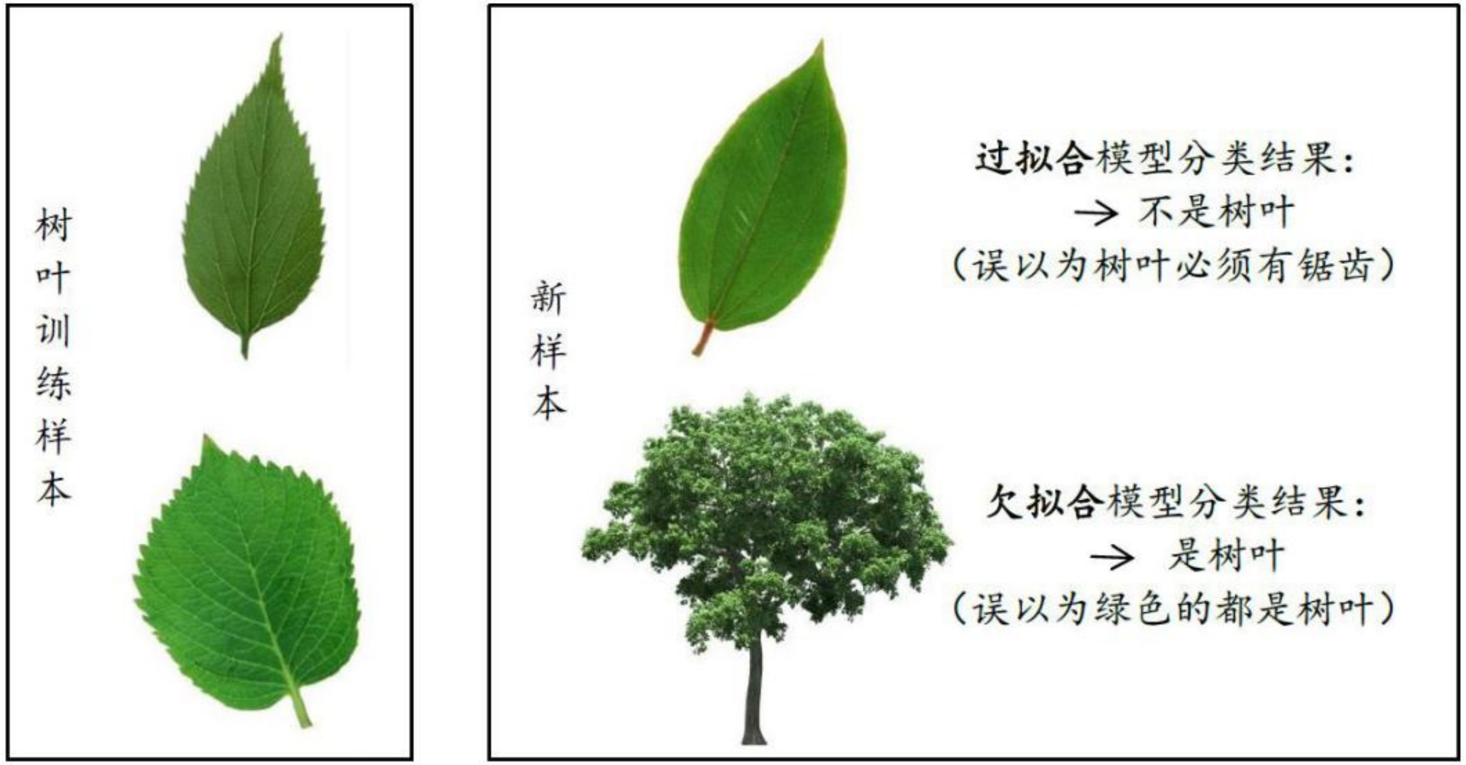
学习方法

- 例：聚类，欠拟合 vs 过拟合



ML·DL

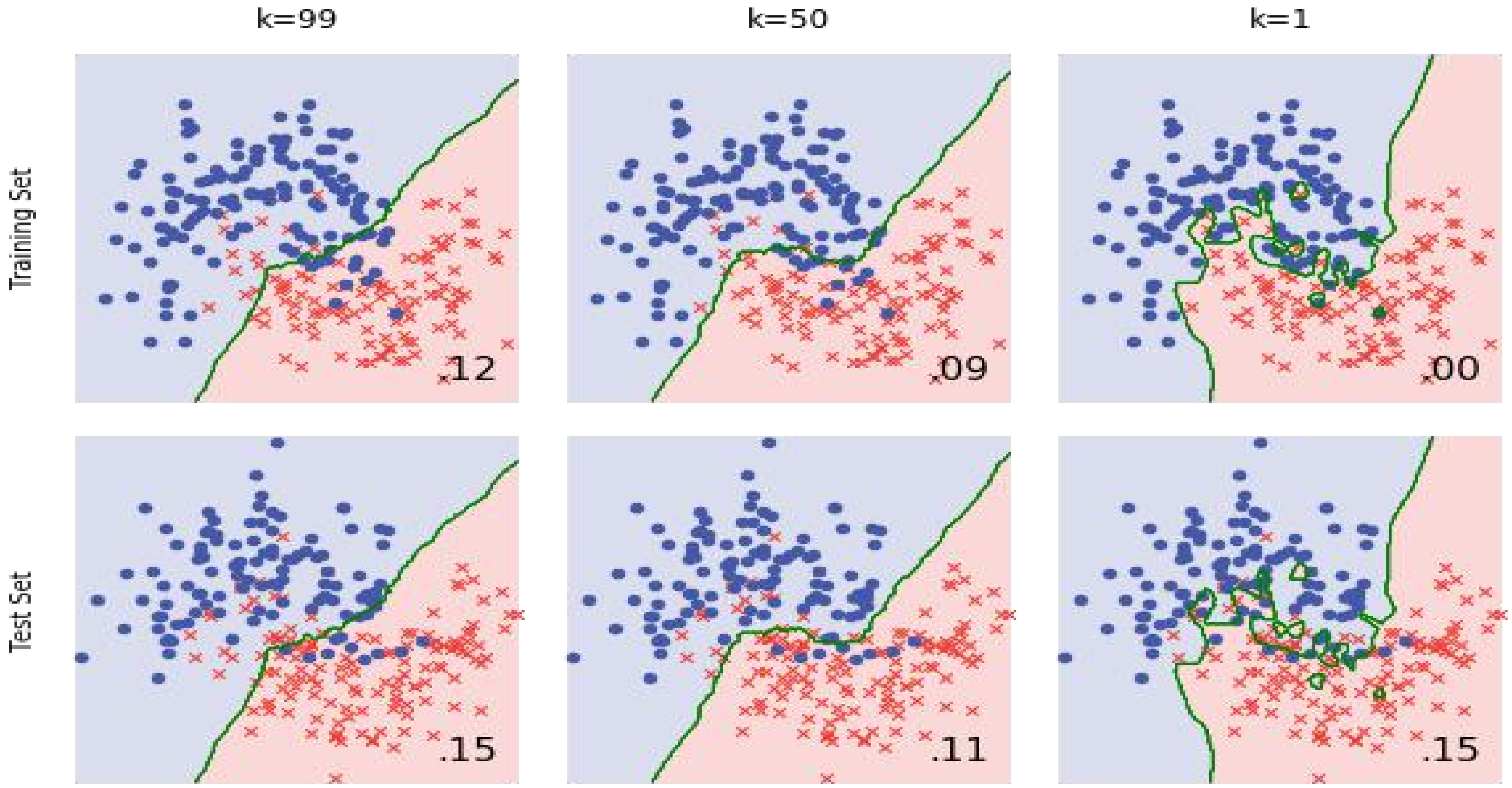
- 例：聚类，欠拟合 vs 过拟合



过拟合、欠拟合的直观类比

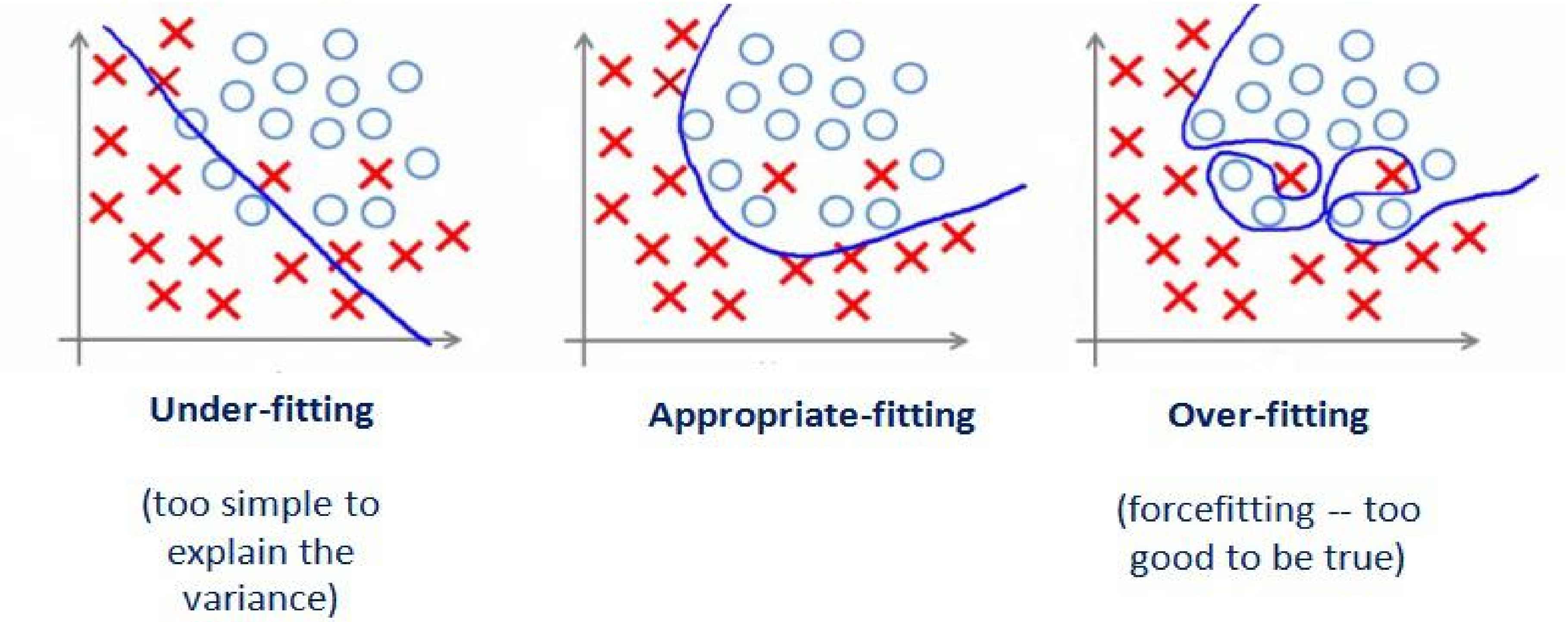
学习方法

- 例：聚类，欠拟合 vs 过拟合



过拟合

- 过拟合：经验风险最小化原则很容易导致模型在训练集上错误率很低，但是在未知数据上错误率很高。
  - 过拟合问题往往是由于训练数据少和噪声等原因造成的。

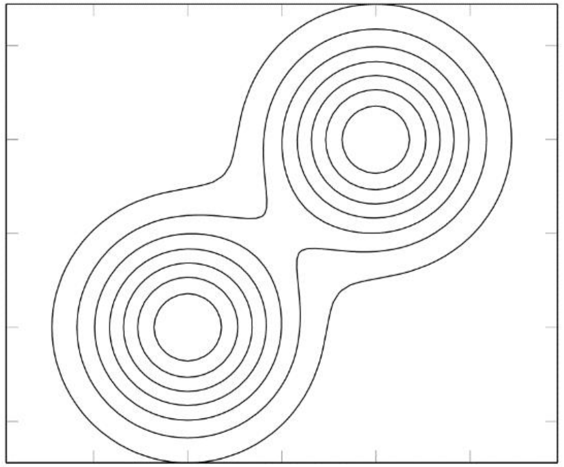


泛化错误

期望风险

$$\mathcal{R}(f) = \mathbb{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y)}[\mathcal{L}(f(\mathbf{x}), y)],$$

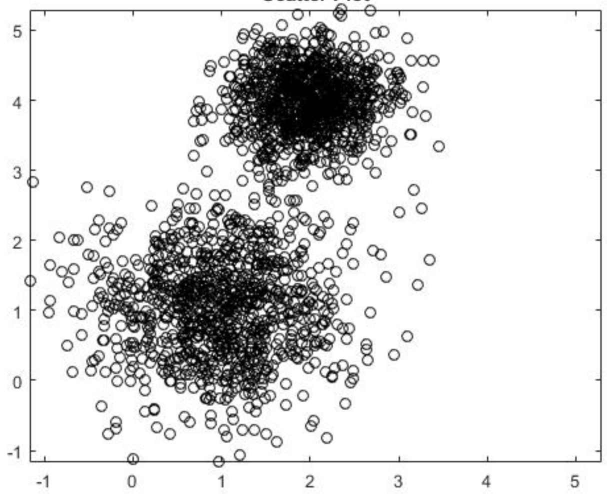
真实分布 p_r



$\neq$

经验风险

$$\mathcal{R}_{\mathcal{D}}^{emp}(\theta) = \frac{1}{N}\sum_{n=1}^{N}\mathcal{L}(y^{(n)}, f(x^{(n)}, \theta))$$



$$\mathcal{G}_{\mathcal{D}}(f) = \mathcal{R}(f) - \mathcal{R}_{\mathcal{D}}^{emp}(f)$$

泛化错误（Generalization Error）

机器学习基石，林轩田，B站

如何减少泛化错误？

优化

经验风险最小

正则化

降低模型复杂度

正则化（regularization）

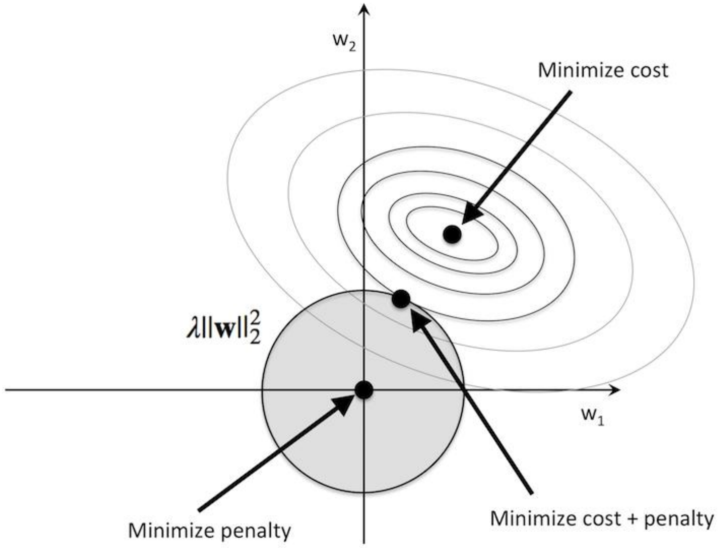
所有损害优化的方法都是正则化。

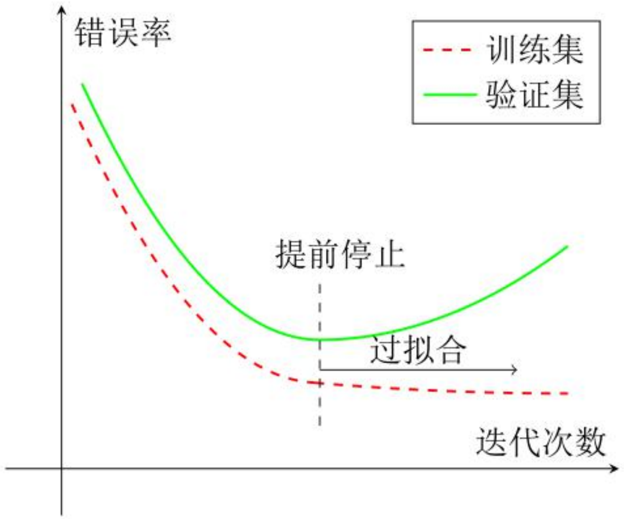
增加优化约束

L1/L2约束、数据增强

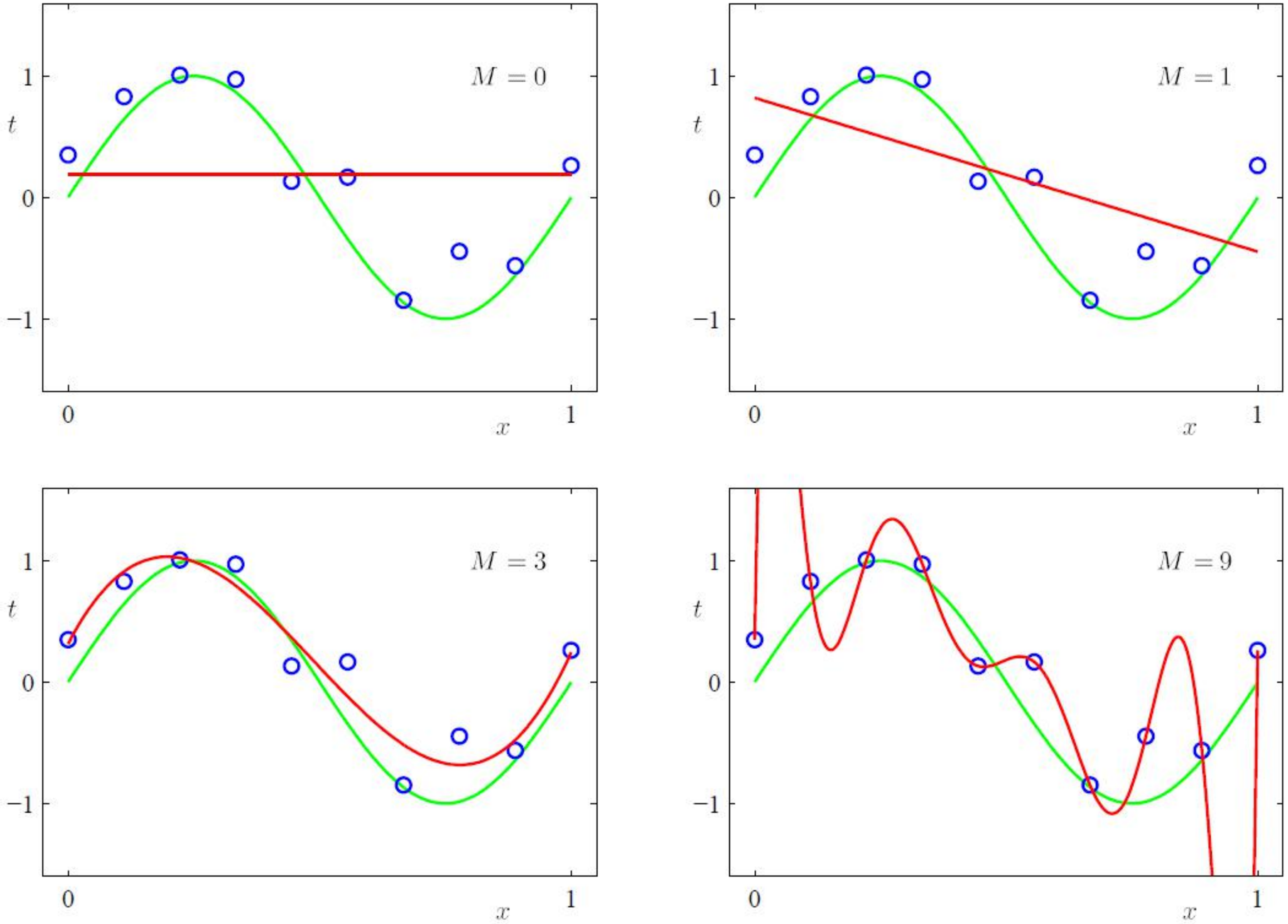
干扰优化过程

权重衰减、随机梯度下降、提前停止、梯度惩罚





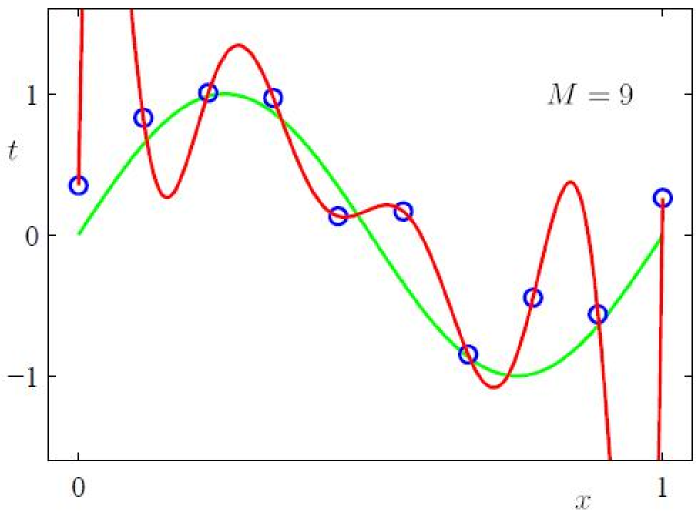
Which Degree of Polynomial?



A model selection problem

M = 9 → E(w) = 0: This is overfitting

Controlling Overfitting: Regularization



	$M=0$	$M=1$	$M=3$	$M=9$
$w_0^\star$	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

As order of polynomial M increases, so do coefficient magnitudes!

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

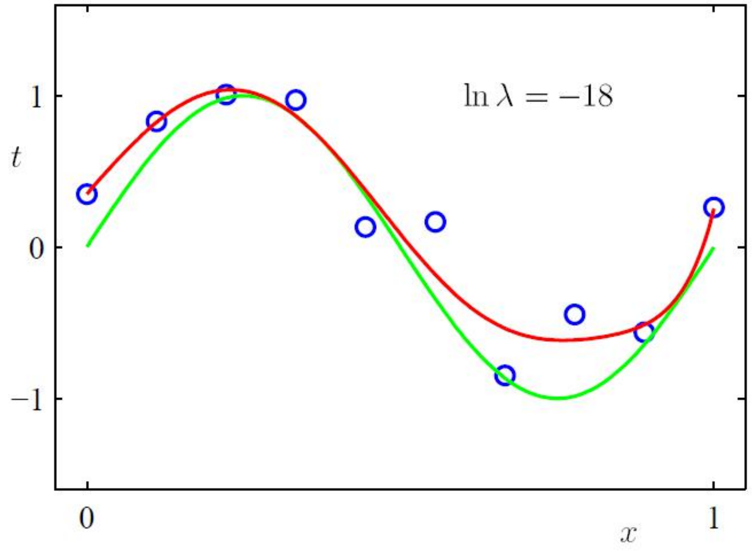
对大的系数进行惩罚

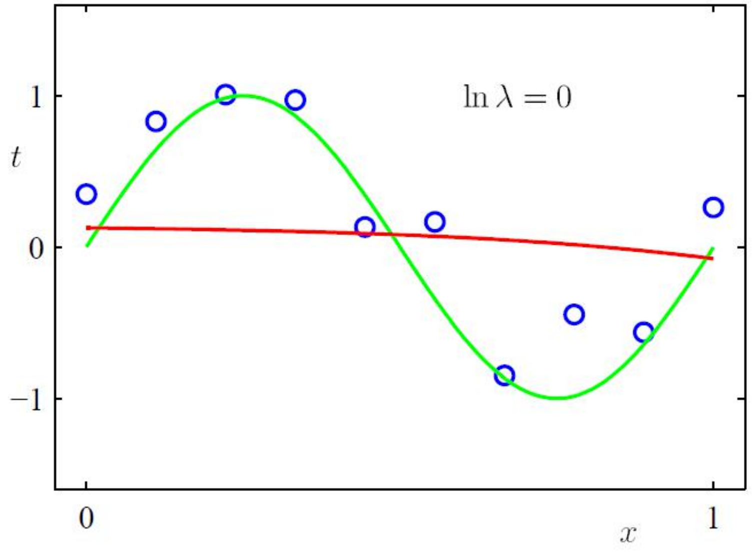
Controlling Overfitting: Regularization

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

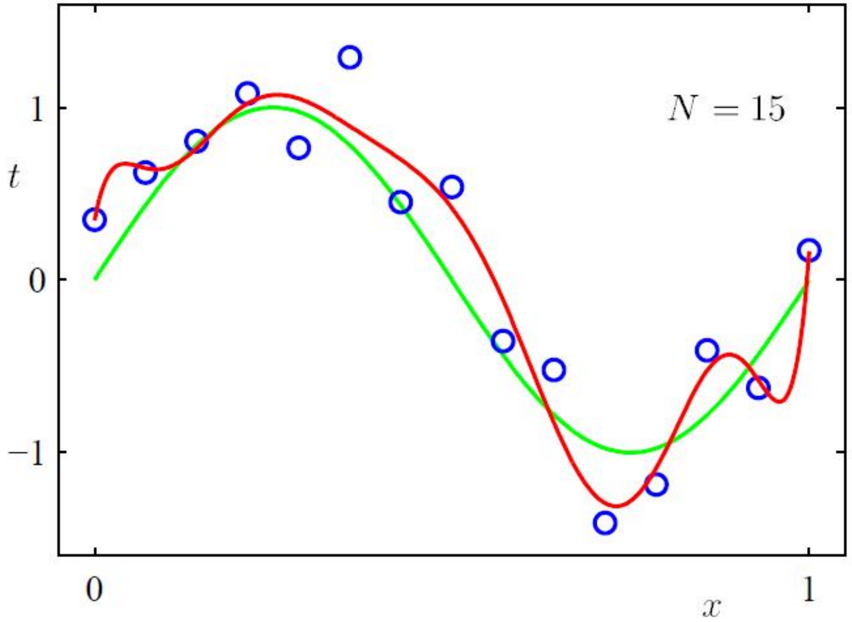
结构风险最小化准则（岭回归）

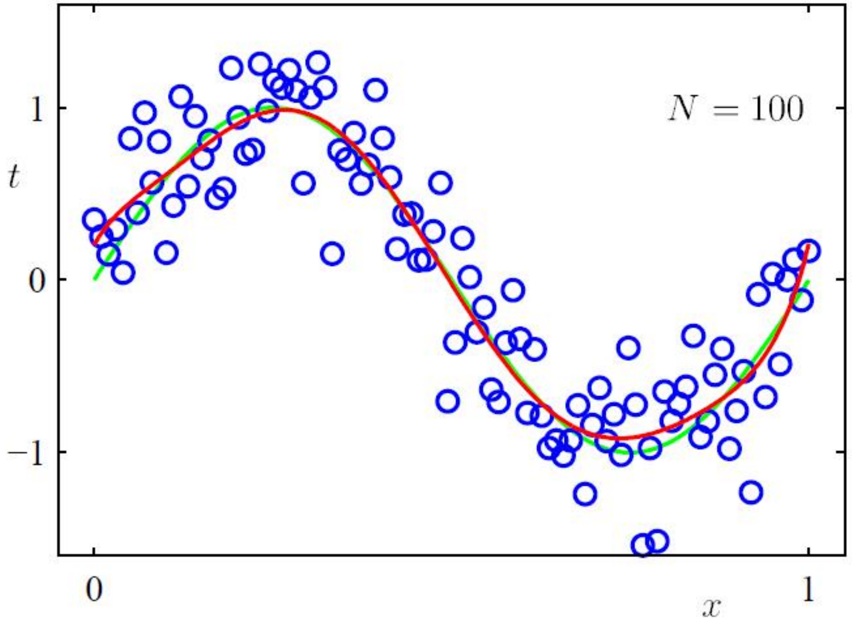
	$\ln\lambda = -\infty$	$\ln\lambda = -18$	$\ln\lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18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





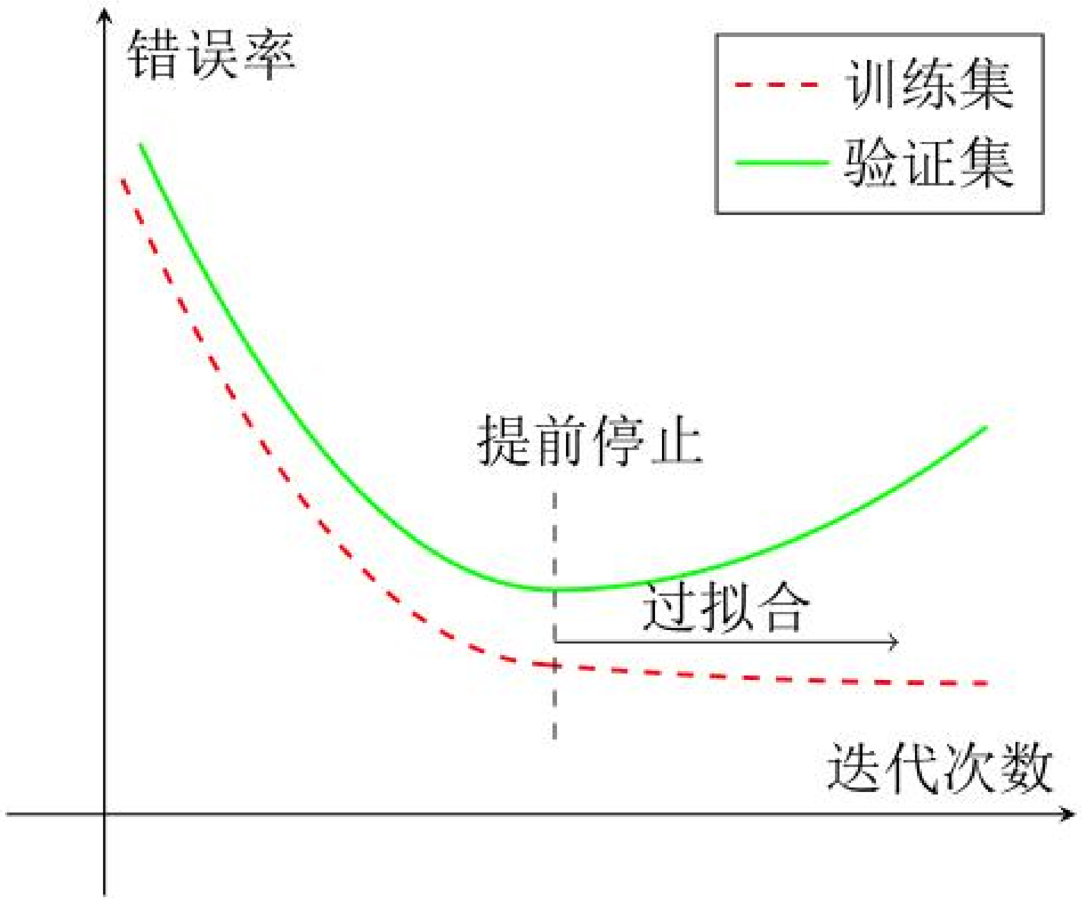
Controlling Overfitting: Dataset size





提前停止

- 验证集（Validation Dataset）
  - 我们使用一个验证集（Validation Dataset）来测试每一次迭代的参数在验证集上是否最优。
  - 如果在验证集上的错误率不再下降，就停止迭代。



模型选择

常见的机器学习类型

ML·DL

	监督学习	无监督学习	强化学习
训练样本	训练集 $\{(\mathbf{x}^{(n)}, y^{(n)})\}_{n=1}^{N}$	训练集 $\{\mathbf{x}^{n}\}_{n=1}^{N}$	智能体和环境交互的轨迹 τ 和累积奖励 G_τ
优化目标	$y = f(\mathbf{x})$ 或 $p(y\|\mathbf{x})$	$p(\mathbf{x})$ 或带隐变量 $\mathbf{z}$ 的 $p(\mathbf{x}\|\mathbf{z})$	期望总回报 $\mathbb{E}_\tau[G_\tau]$
学习准则	期望风险最小化 最大似然估计	最大似然估计 最小重构错误	策略评估 策略改进

如何选择一个合适的模型？

ML·DL

- 模型选择
  - 拟合能力强的模型一般复杂度会比较高，容易过拟合。
  - 如果限制模型复杂度，降低拟合能力，可能会欠拟合。
- 偏差与方差分解
  - 期望错误可以分解为

$$\mathcal{R}(f) = (\text{bias})^2 + \text{variance} + \varepsilon.$$

$$\mathbb{E}_{(\mathbf{x},y)\sim p_r(\mathbf{x},y)}\left[\left(y - f^*(\mathbf{x})\right)^2\right]$$

模型选择：偏差与方差

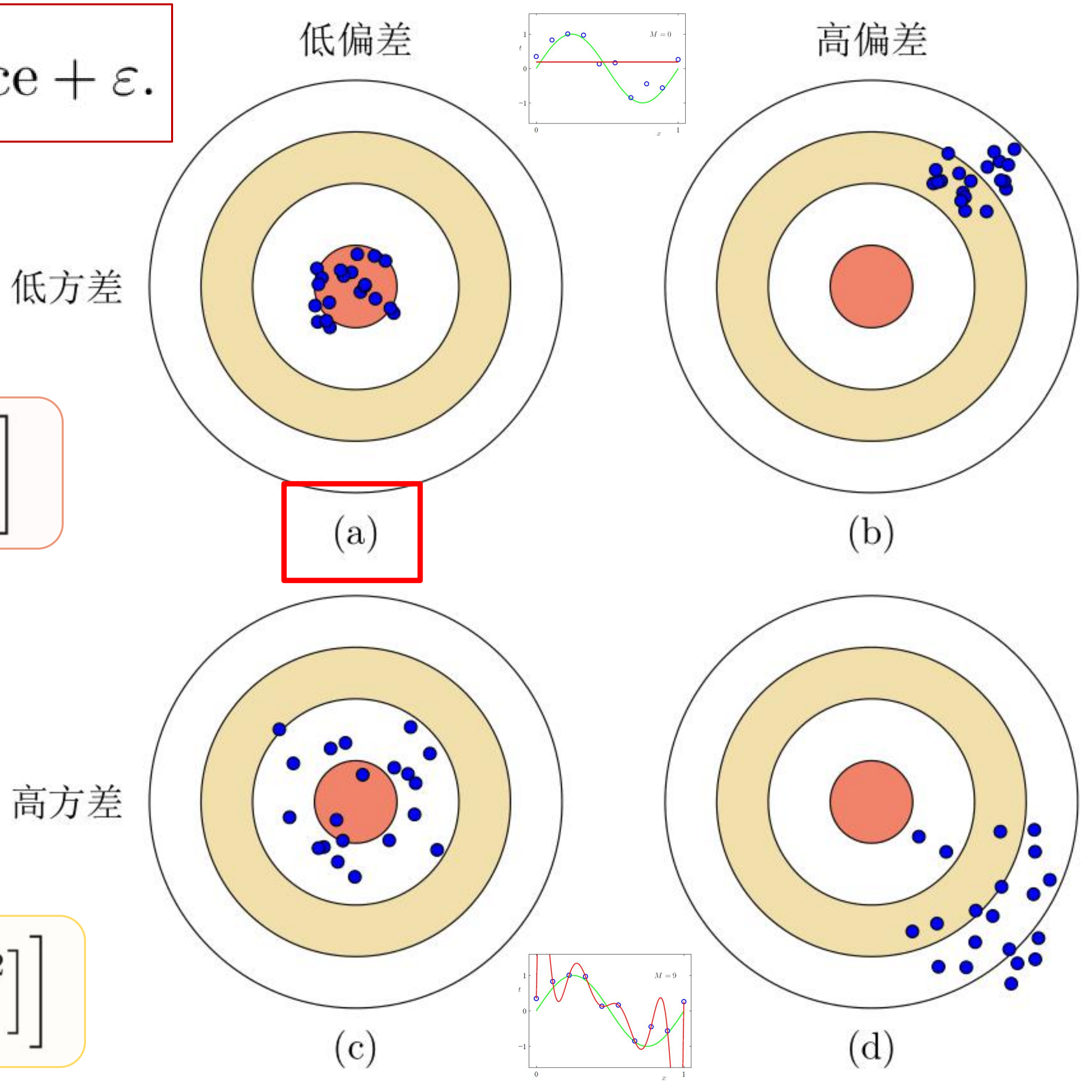
$$\mathcal{R}(f) = (\text{bias})^2 + \text{variance} + \varepsilon.$$

bias

$$\mathbb{E}_{\mathbf{x}}\left[\left(\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})] - f^*(\mathbf{x})\right)^2\right]$$

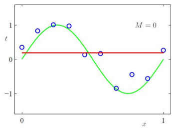
variance

$$\mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(f_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})]\right)^2\right]\right]$$



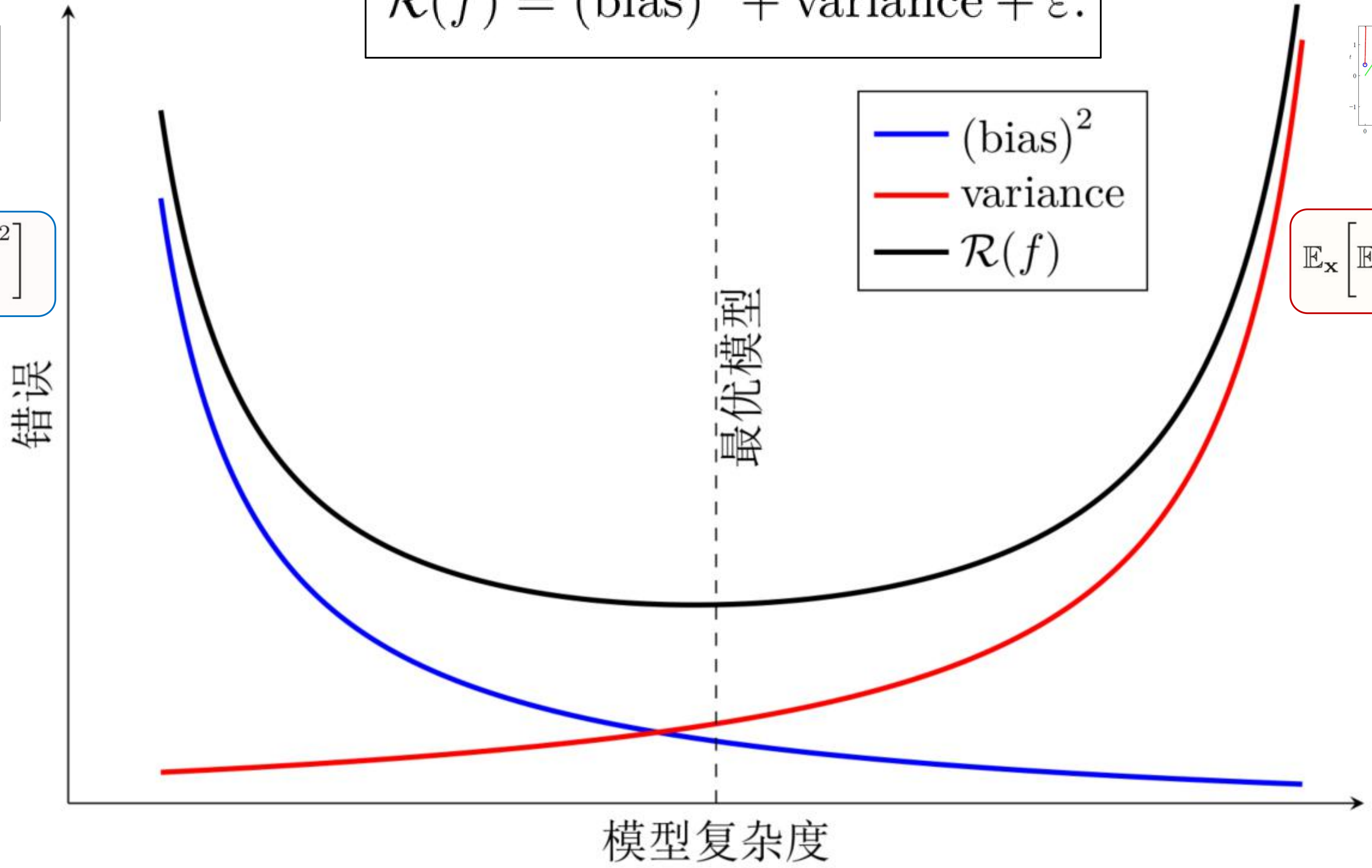
模型选择：偏差与方差

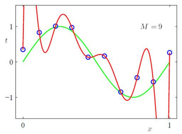
$$\mathcal{R}(f) = (\text{bias})^2 + \text{variance} + \varepsilon.$$



bias

$$\mathbb{E}_{\mathbf{x}}\left[\left(\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\mathbf{x})\right] - f^*(\mathbf{x})\right)^2\right]$$





variance

$$\mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(f_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\mathbf{x})\right]\right)^2\right]\right]$$

集成模型：有效的降低方差的方法

- 集成模型

$$f^{(c)}(\mathbf{x}) = \frac{1}{M}\sum_{m=1}^{M} f_m(\mathbf{x})$$

  - 通过多个高方差模型的平均来降低方差。

- 集成模型的期望错误大于等于所有模型的平均期望错误的1/M，小于等于所有模型的平均期望错误。

$$\bar{\mathcal{R}}(f) \geq \mathcal{R}(f^{(c)}) \geq \frac{1}{M}\bar{\mathcal{R}}(f)$$

ML·DL

- 最基础的理论就是可能近似正确（Probably Approximately Correct, PAC）学习理论。
  - 根据大数定律，当训练集大小|D|趋向无穷大时，泛化错误趋向于0，即经验风险趋近于期望风险。

$$\lim_{|\mathcal{D}|\to\infty} \mathcal{R}(f) - \mathcal{R}_{\mathcal{D}}^{emp}(f) = 0$$

- PAC学习

$$P\Big(\underbrace{(\mathcal{R}(f) - \mathcal{R}_{\mathcal{D}}^{emp}(f)) \leq \epsilon}_{\text{近似正确，} 0<\epsilon<0.5}\Big) \geq 1 - \delta$$

近似正确，0 < ϵ <0.5

可能，0 < δ <0.5

机器学习基石，林轩田，B站

归纳偏好（inductive bias）

算法在学习过程中对某种类型假设的偏好，称为“归纳偏好”(inductive bias)，或简称为“偏好”.

“奥卡姆剃刀”(Occam's razor)

“若有多个假设与观察一致，

则选最简单的那个”

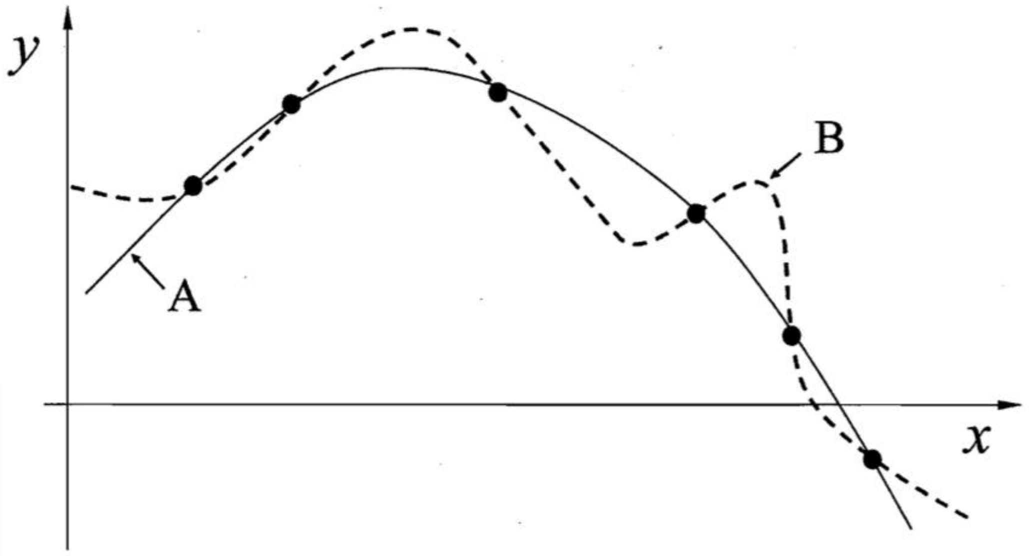
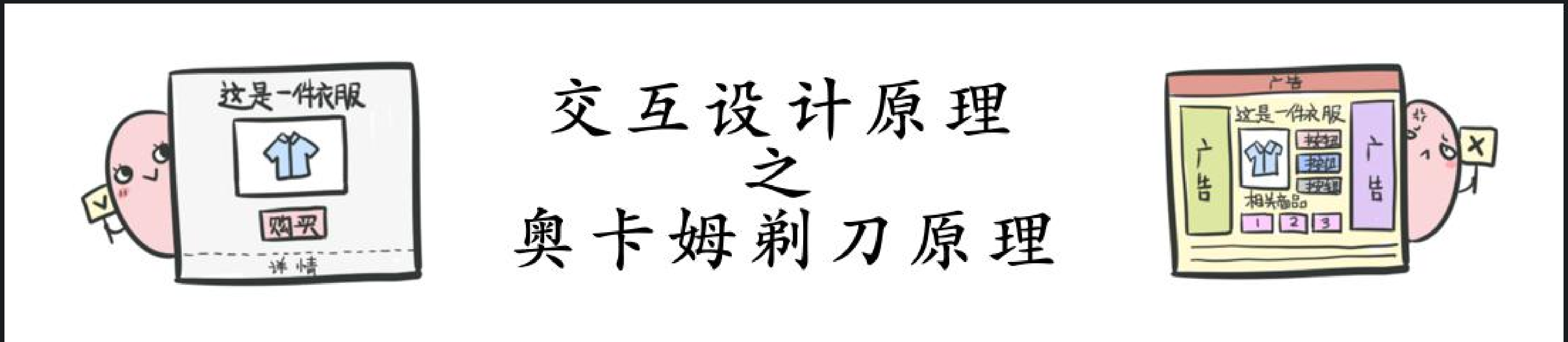


图 1.3 存在多条曲线与有限样本训练集一致

奥卡姆剃刀定律

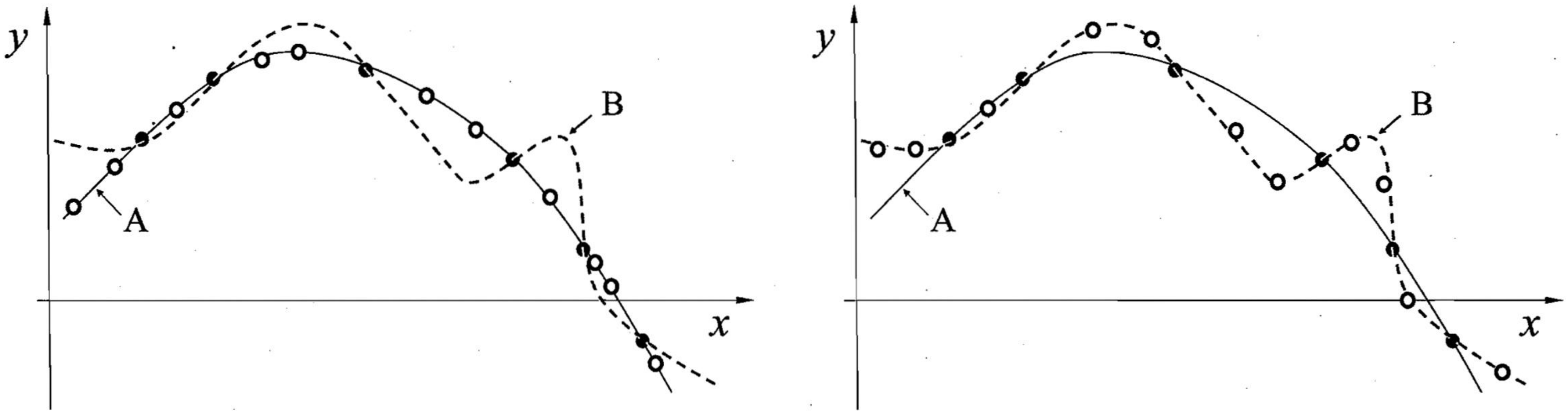
Occam's Razor

奥卡姆剃刀定律又称“奥康的剃刀”，它是由14世纪英格兰的逻辑学家、圣方济各会修士奥卡姆的威廉（William of Occam，约1285年至1349年）提出。这个原理称为“如无必要，勿增实体”，即“简单有效原理”。



没有免费午餐定理（No Free Lunch Theorem，NFL）

- 不存在一种机器学习算法适合于任何领域或任务。
  - 对于一个学习算法$\mathfrak{L}_a$，若它在某些问题上比$\mathfrak{L}_b$好，则必定存在另一些问题，在那里$\mathfrak{L}_b$比$\mathfrak{L}_a$好。这个结论对任何算法均成立。



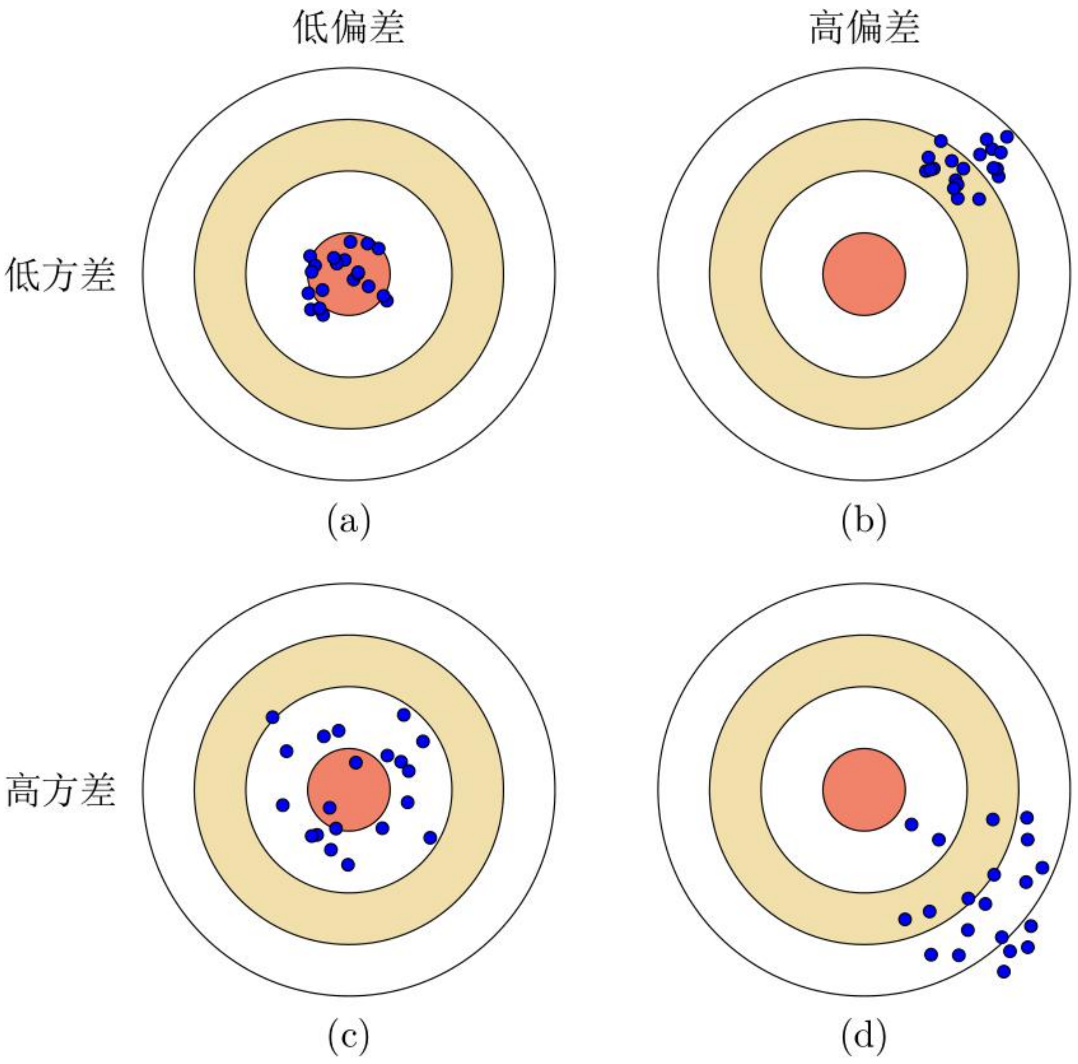
(a) A 优于 B　　(b) B 优于 A

图 1.4　没有免费的午餐. (黑点: 训练样本; 白点: 测试样本)

小结

小结

- 基本术语
  - 过拟合/欠拟合
  - 正则化
- 学习理论
  - 偏差与方差
  - 可能近似正确（PAC）学习理论
  - 归纳偏好
  - 奥卡姆剃刀准则
  - 没有免费午餐定理



机器学习 & 深度学习

杭州电子科技大学
HANGZHOU DIANZI UNIVERSITY

笃学力行 守正求新