《机器学习基础》

机器学习概述

教学内容

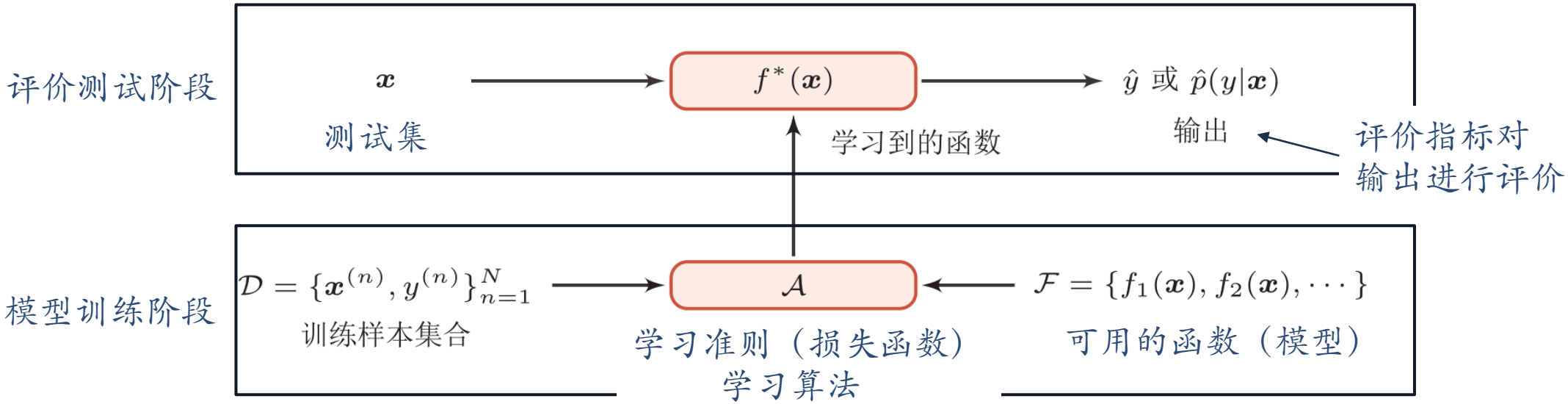
- 机器学习
  - 概念
  - 原理
- 线性回归
  - 定义
  - 经验风险最小化
    - 最小均方误差
  - 结构风险最小化
  - 最大似然估计
  - 最大后验估计
- 机器学习的几个关键点

机器学习的五要素

- 数据
  - 训练集、验证集、测试集
- 模型
  - 线性方法：$f(\mathbf{x}, \theta) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b$
  - 广义线性方法：$f(\mathbf{x}, \theta) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}) + b$
    - 如果$\phi(x)$为可学习的非线性基函数，$f(x,\theta)$就等价于神经网络。
- 学习准则
  - 期望风险 $\mathcal{R}(f) = \mathbb{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y)}[\mathcal{L}(f(\mathbf{x}), y)],$
- 优化算法
  - 梯度下降
- 评价指标
  - 如何衡量模型的准确度

什么是机器学习?

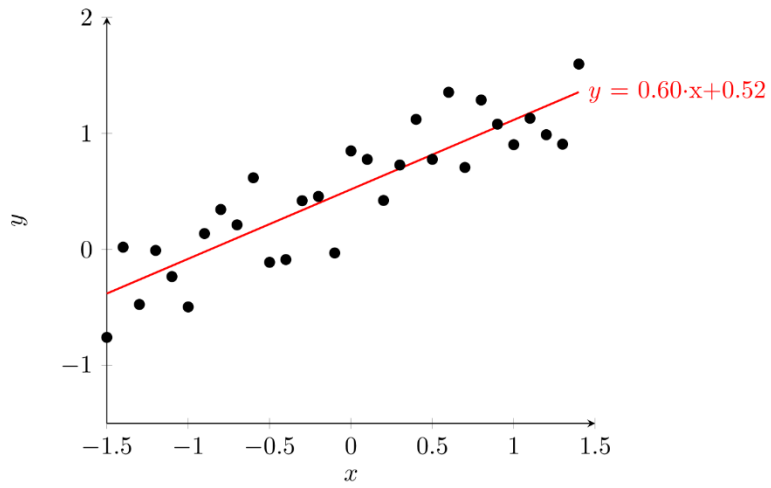
- 机器学习：通过算法使得机器能从大量数据中学习规律从而对新的样本做决策。
  - 规律：决策（预测）函数



模型

- 以线性回归（Linear Regression）为例
- 模型：

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



学习准则

- 损失函数
  - 0-1损失函数

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

  - 平方损失函数

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

学习准则

- 期望风险未知，通过经验风险近似
  - 训练数据：$\mathcal{D} = \{\mathrm{x}^{(\mathrm{n})}, y^{(\mathrm{n})}\}, i \in [1, N]$

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

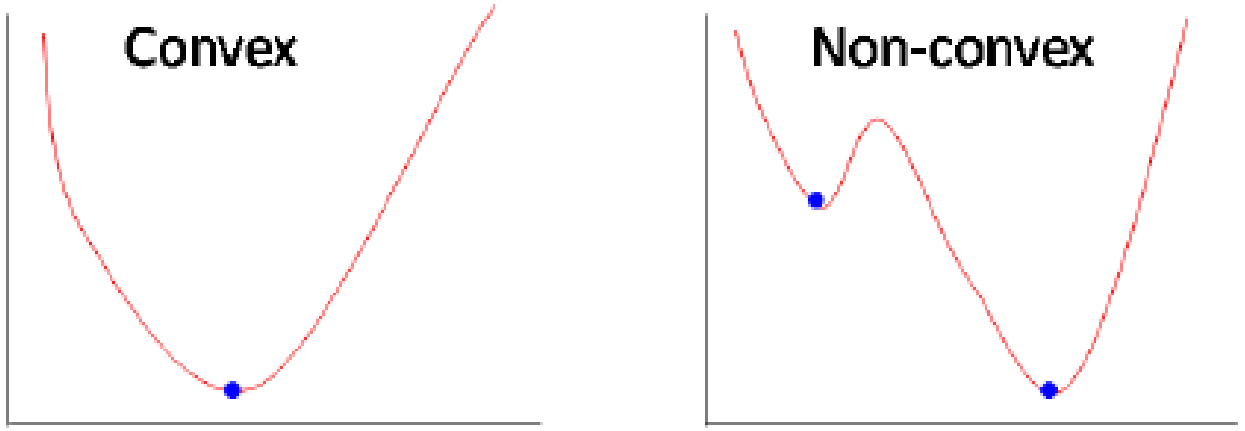
- 经验风险最小化
  - 在选择合适的风险函数后，我们寻找一个参数θ*，使得经验风险函数最小化。

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

- 机器学习问题转化成为一个最优化问题

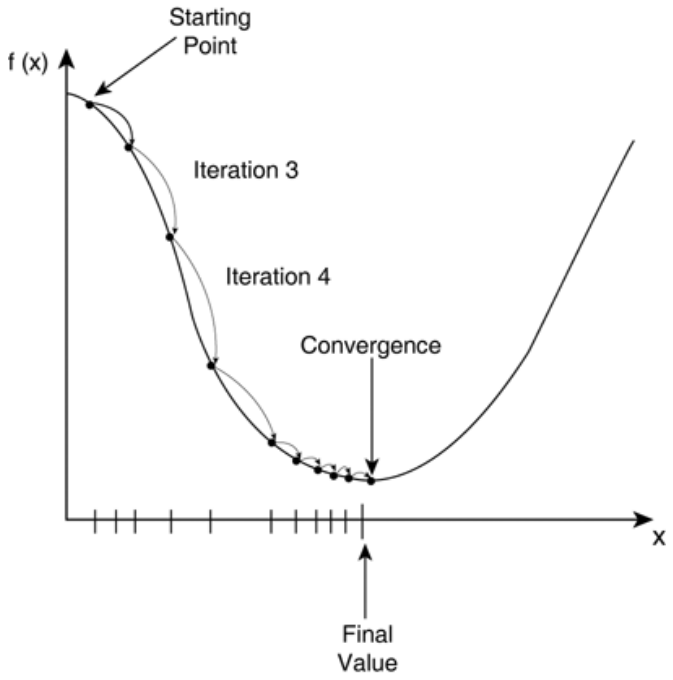
最优化问题

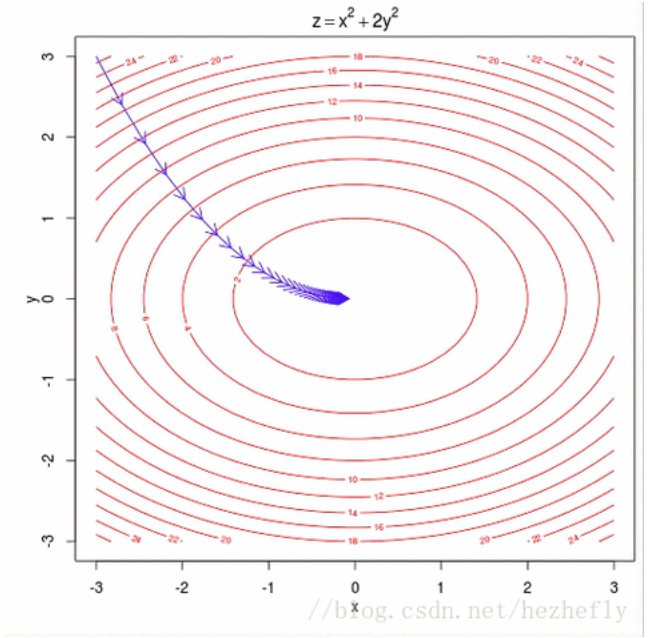
- 机器学习问题转化成为一个最优化问题



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

梯度下降法（Gradient Descent）



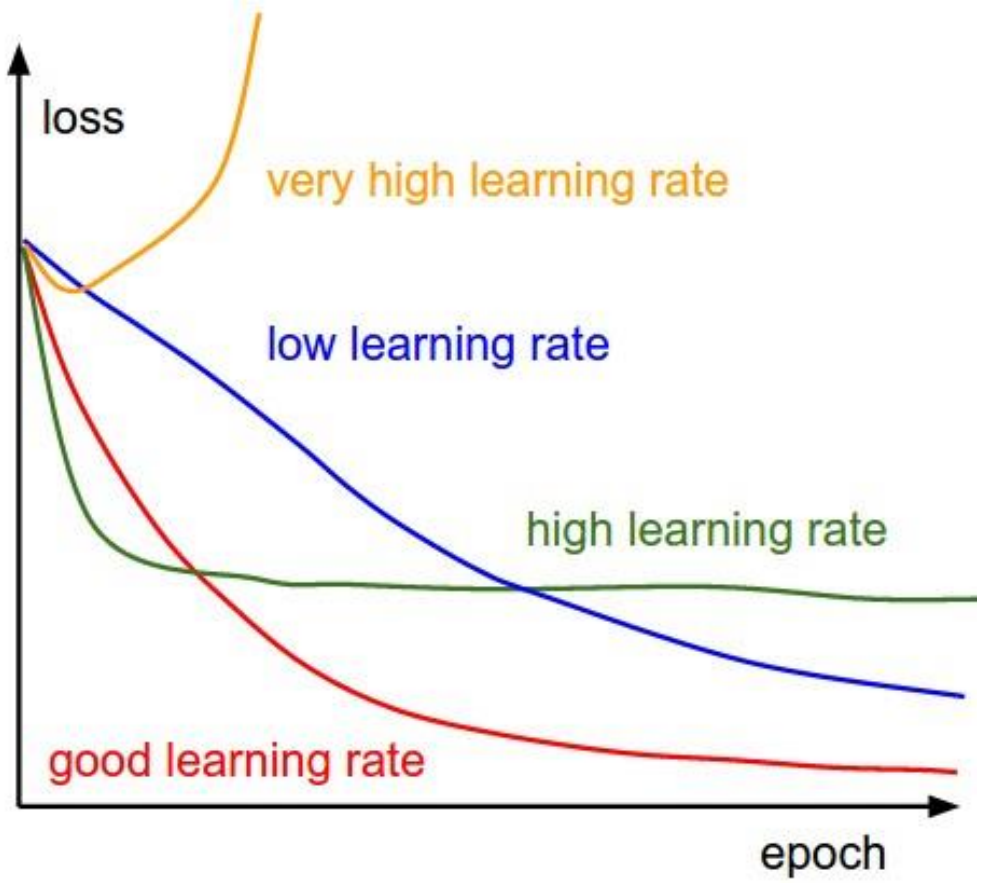


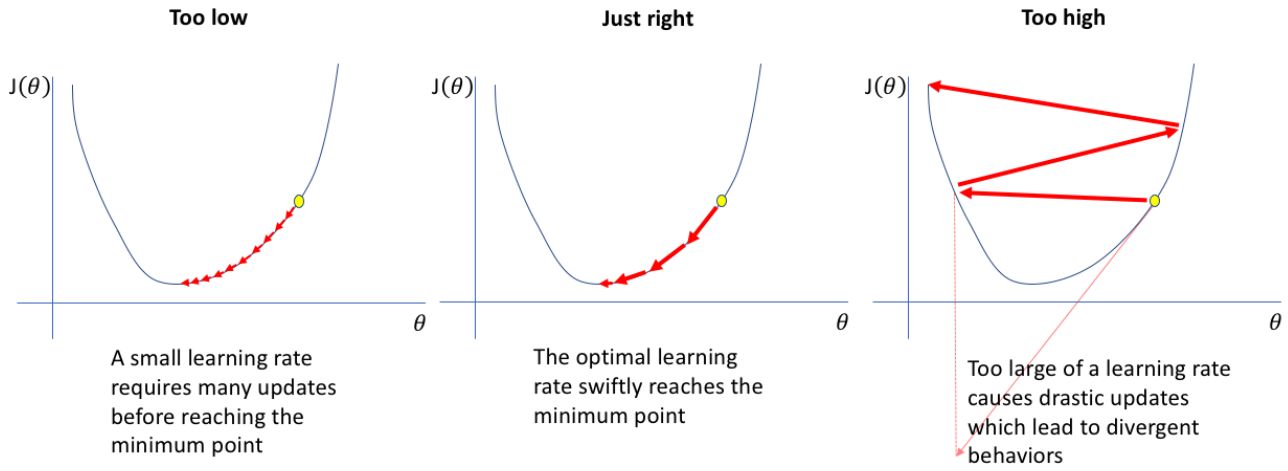
搜索步长α中也叫作学习率（Learning Rate）

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

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

学习率是十分重要的超参数!





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

$$\theta_{t+1} = \theta_t - \alpha \frac{\partial \mathcal{L}(\theta_t; x^{(t)}, y^{(t)})}{\partial \theta},$$

- 小批量（Mini-Batch）随机梯度下降法

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

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

  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

线性回归

线性回归（Linear Regression）

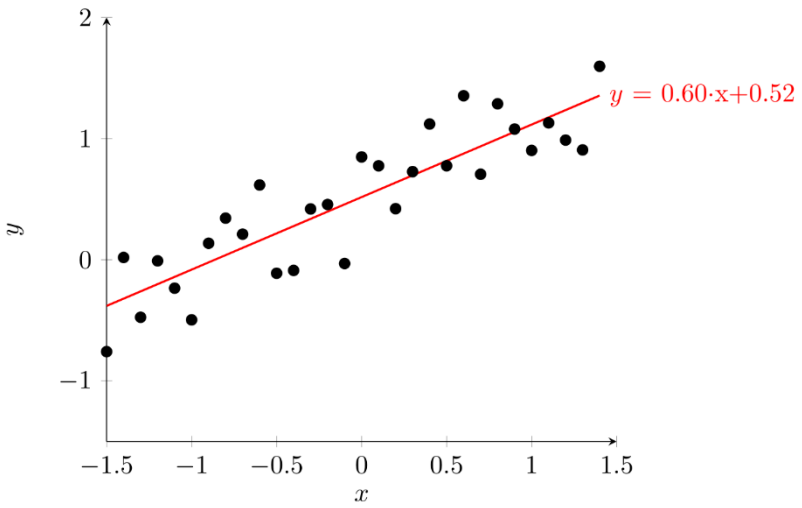
- 模型：

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

- 增广权重向量和增广特征向量

$$\hat{\mathbf{x}}=\mathbf{x}\oplus 1\triangleq\begin{bmatrix}\mathbf{x}\\1\end{bmatrix}=\begin{bmatrix}x_1\\\vdots\\x_k\\1\end{bmatrix},$$

$$\hat{\mathbf{w}}=\mathbf{w}\oplus b\triangleq\begin{bmatrix}\mathbf{w}\\b\end{bmatrix}=\begin{bmatrix}w_1\\\vdots\\w_k\\b\end{bmatrix},$$



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

优化方法

- 经验风险最小化（最小二乘法）
- 结构风险最小化（岭回归）
- 最大似然估计
- 最大后验估计

经验风险最小化

矩阵微积分

- 标量关于向量的偏导数

$$\frac{\partial y}{\partial \boldsymbol{x}} = [\frac{\partial y}{\partial x_1}, \cdots, \frac{\partial y}{\partial x_M}]^{\mathrm{T}}$$

- 向量关于向量的偏导数

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_N}{\partial x_1} \\ \vdots & \vdots & \vdots \\ \frac{\partial y_1}{\partial x_M} & \cdots & \frac{\partial y_N}{\partial x_M} \end{bmatrix}$$

- 向量函数及其导数

$$\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{I},$$
$$\frac{\partial \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{A}^{\mathrm{T}},$$
$$\frac{\partial \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}}{\partial \boldsymbol{x}} = \boldsymbol{A}$$

经验风险最小化

- 模型

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

- 学习准则

$$\begin{aligned}\mathcal{R}(\mathbf{w}) &= \sum_{n=1}^{N}\mathcal{L}(y^{(n)}, f(\mathbf{x}^{(n)};\mathbf{w})) \\ &= \frac{1}{2}\sum_{n=1}^{N}\left(y^{(n)} - \mathbf{w}^{\mathrm{T}}\mathbf{x}^{(n)}\right)^2 \\ &= \frac{1}{2}\|\mathbf{y} - X^{\mathrm{T}}\mathbf{w}\|^2,\end{aligned}$$

经验风险最小化

- 优化

$$\frac{\partial}{\partial \mathbf{w}} \mathcal{R}(\mathbf{w}) = 0$$

$$\begin{aligned} \frac{\partial \mathcal{R}(\mathbf{w})}{\partial \mathbf{w}} &= \frac{1}{2} \frac{\partial \ \|\mathbf{y} - X^{\mathrm{T}}\mathbf{w}\|^2}{\partial \mathbf{w}} \\ &= -X(\mathbf{y} - X^{\mathrm{T}}\mathbf{w}), \end{aligned}$$

结构风险最小化

- 结构风险最小化准则

$$\mathcal{R}(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}\|^2 + \frac{1}{2}\lambda\|\boldsymbol{w}\|^2,$$

- 得到

$$\boldsymbol{w}^* = (\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} + \lambda I)^{-1}\boldsymbol{X}\boldsymbol{y},$$

  - 岭回归（Ridge Regression）

最大似然估计

关于概率的一些基本概念

- 概率（Probability）
  - 一个随机事件发生的可能性大小，为0到1之间的实数。
- 随机变量（Random Variable）
  - 比如随机掷一个骰子，得到的点数就可以看成一个随机变量X，其取值为{1,2,3,4,5,6}。
- 概率分布（Probability Distribution）
  - 一个随机变量X取每种可能值的概率

$$P(X = x_i) = p(x_i), \qquad \forall i \in \{1, \cdots, n\}.$$

  - 并满足

$$\sum_{i=1}^{n} p(x_i) = 1,$$

$$p(x_i) \geq 0, \qquad \forall i \in \{1, \cdots, n\}.$$

概率的一些基本概念

- 伯努利分布（Bernoulli Distribution）
  - 在一次试验中，事件A出现的概率为μ，不出现的概率为1－μ。若用变量X 表示事件A出现的次数，则X 的取值为0和1，其相应的分布为

$$p(x) = \mu^x (1-\mu)^{(1-x)}$$

- 二项分布（Binomial Distribution）
  - 在n次伯努利分布中，若以变量X 表示事件A出现的次数，则X 的取值为{0,… ,n}，$P(X=k) = \binom{n}{k} \mu^k (1-\mu)^{n-k}, \qquad k = 1 \cdots, n$

二项式系数，表示从n个元素中取出k个元素而不考虑其顺序的组合的总数。

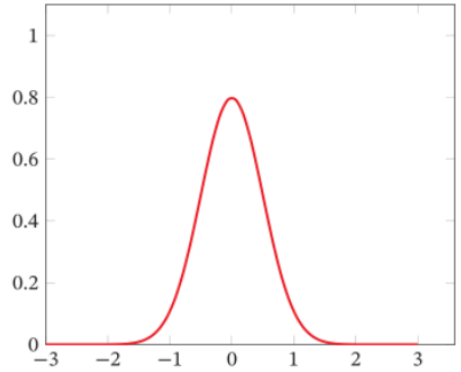
概率的一些基本概念

- 连续随机变量 Y 的概率分布一般用概率密度函数（Probability Density Function， PDF）$p(x)$ 来描述。

$$\int_{-\infty}^{+\infty} p(x)\mathrm{d}x = 1$$

- 高斯分布（Gaussian Distribution）

$$X \sim \mathcal{N}(\mu, \sigma^2) \qquad p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



概率的一些基本概念

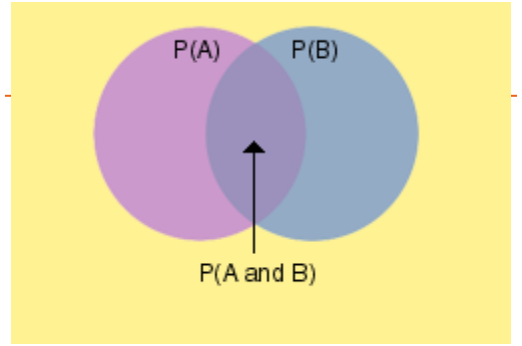
- 条件概率（Conditional Probability）
  - 对于离散随机向量(X,Y)，已知X = x的条件下，随机变量Y = y的条件概率为：

$$p(y|x) = P(Y = y|X = x) = \frac{p(x, y)}{p(x)}$$

- 贝叶斯公式
  - 两个条件概率p(y|x)和p(x|y)之间的关系

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

例子



Marginal Probability

性别\行业	计算机	教育	
男	0.4	0.1	0.5
女	0.1	0.4	0.5
	0.6	0.4	

Marginal Probability

p(男|计算机)=

似然（Likelihood）

贝叶斯公式：

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

$$p(w|X) \propto p(X|w)p(w)$$

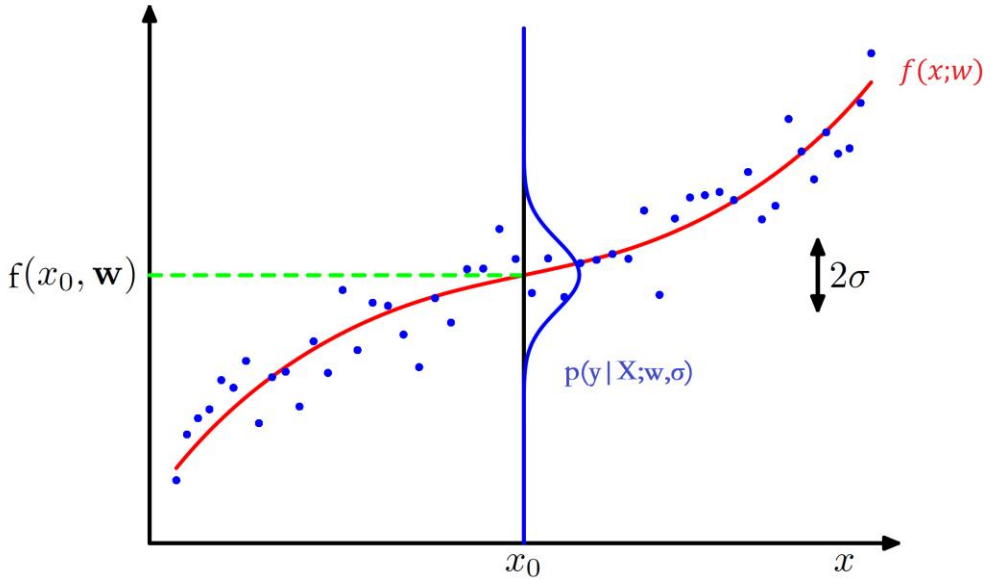
后验 posterior　　似然 likelihood　　先验 prior

posterior ∝ likelihood × prior

从概率角度来看线性回归

- 假设标签y为一个随机变量，其服从以均值为$f(x;w) = w^T x$，方差为σ^2 的高斯分布。

$$p(y|\mathbf{x};\mathbf{w},\sigma) = \mathcal{N}(y;\mathbf{w}^{\mathrm{T}}\mathbf{x},\sigma^2)$$
$$= \frac{1}{\sqrt{2\pi}\sigma}\exp\Big(-\frac{(y-\mathbf{w}^{\mathrm{T}}\mathbf{x})^2}{2\sigma^2}\Big).$$



线性回归中的似然函数

- 参数w在训练集D上的似然函数（Likelihood）为

$$
\begin{aligned}
p(\mathbf{y}|X;\mathbf{w},\sigma) &= \prod_{n=1}^{N} p(y^{(n)}|\mathbf{x}^{(n)};\mathbf{w},\sigma) \\
&= \prod_{n=1}^{N} \mathcal{N}(y^{(n)};\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(n)},\sigma^2)
\end{aligned}
$$

最大似然估计

- 最大似然估计（Maximum Likelihood Estimate，MLE）
  - 是指找到一组参数w使得似然函数p(y|X;w,σ)最大

$$\text{令}\ \frac{\partial \log p(\mathbf{y}|X;\mathbf{w},\sigma)}{\partial \mathbf{w}} = 0$$

$$\mathbf{w}^{ML} = (XX^{\mathrm{T}})^{-1}X\mathbf{y}.$$

最大后验估计

最大后验估计

$$p(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y};\nu,\sigma) = \frac{p(\boldsymbol{w},\boldsymbol{y}|\boldsymbol{X};\nu,\sigma)}{\sum_{\boldsymbol{w}} p(\boldsymbol{w},\boldsymbol{y}|\boldsymbol{X};\nu,\sigma)}$$
$$\propto p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w};\sigma)p(\boldsymbol{w};\nu),$$

后验 posterior　　似然 likelihood　　先验 prior

$$p(\boldsymbol{w};\nu) = \mathcal{N}(\boldsymbol{w};\boldsymbol{0},\nu^2 I)$$

$$\log p(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y};\nu,\sigma) \propto \log p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w};\sigma) + \log p(\boldsymbol{w};\nu)$$
$$\propto -\frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(y^{(n)} - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^{(n)}\right)^2 - \frac{1}{2\nu^2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w},$$
$$= -\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{w}\|^2 - \frac{1}{2\nu^2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}.$$

正则化系数 $\lambda = \sigma^2/\nu^2$

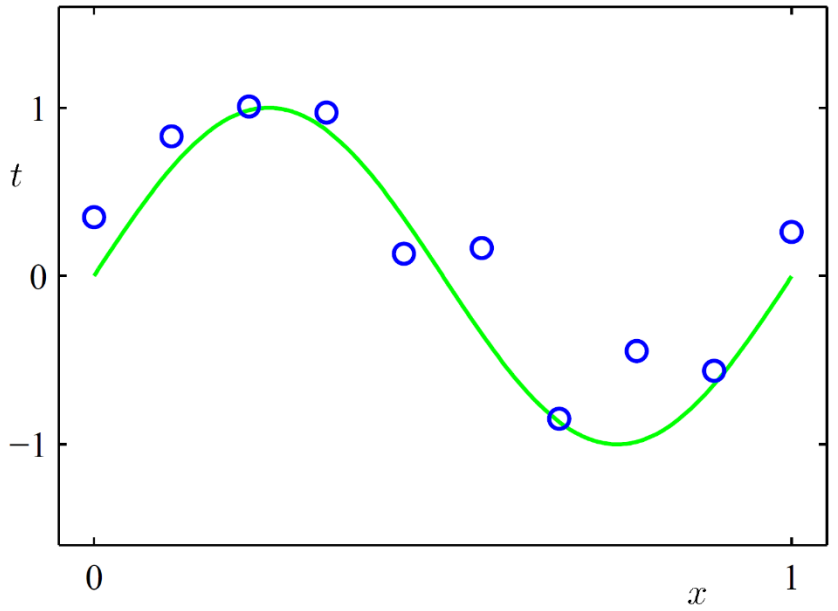
总结

	无先验	引入先验
平方误差	经验风险 最小化	结构风险 最小化
概率	最大似然估计	最大后验估计
	$\boldsymbol{w}^{ML} = (\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})^{-1}\boldsymbol{X}\boldsymbol{y}$	$\boldsymbol{w}^{*} = (\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} + \lambda I)^{-1}\boldsymbol{X}\boldsymbol{y}$

多项式回归

From chapter 1 of Bishop's PRML.

一个例子：Polynomial Curve Fitting



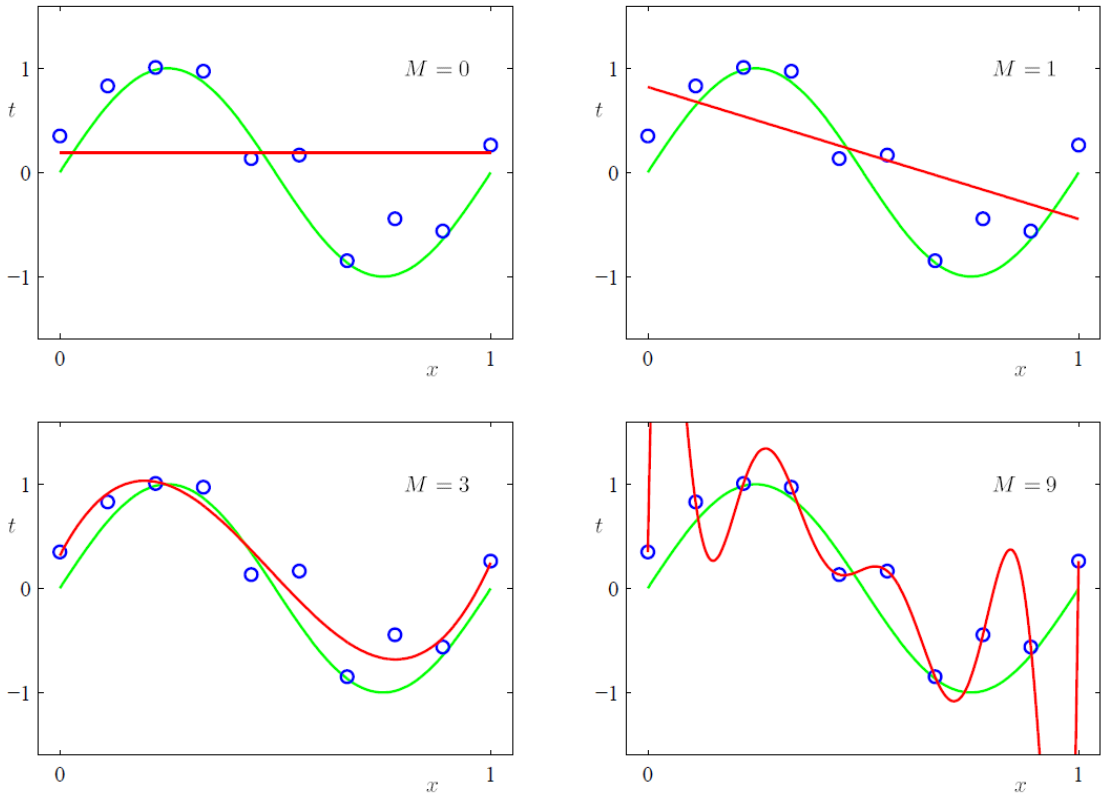
模型

$$y(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M$$

损失函数

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

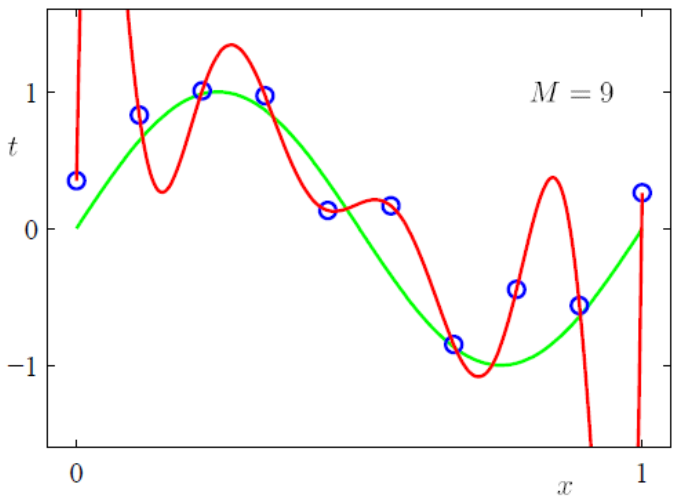
Which Degree of Polynomial?



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

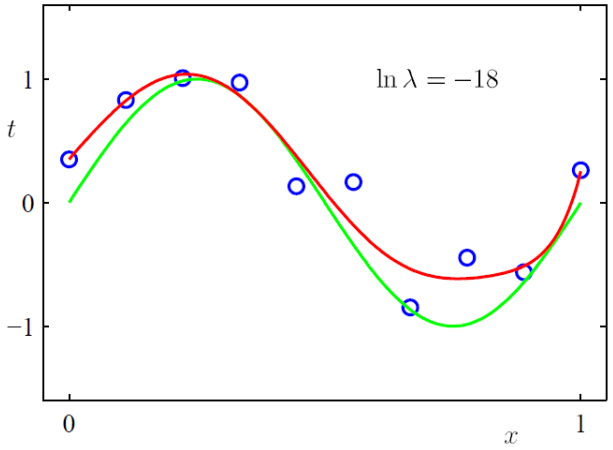
As order of polynomial M increases, so do coefficient magnitudes!

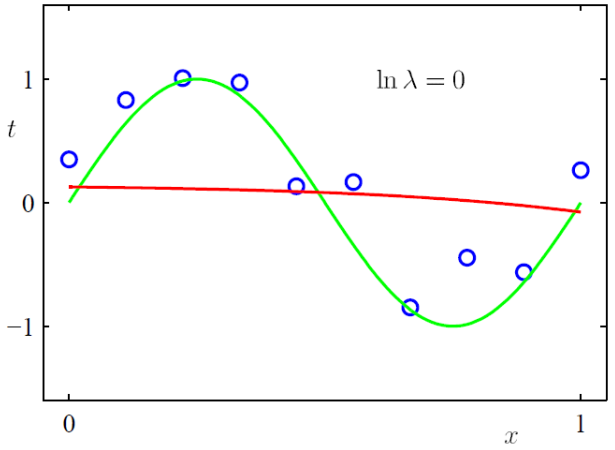
$$\tilde{E}(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{\lambda}{2}||\boldsymbol{w}||^2$$

对大的系数进行惩罚

Controlling Overfitting: Regularization

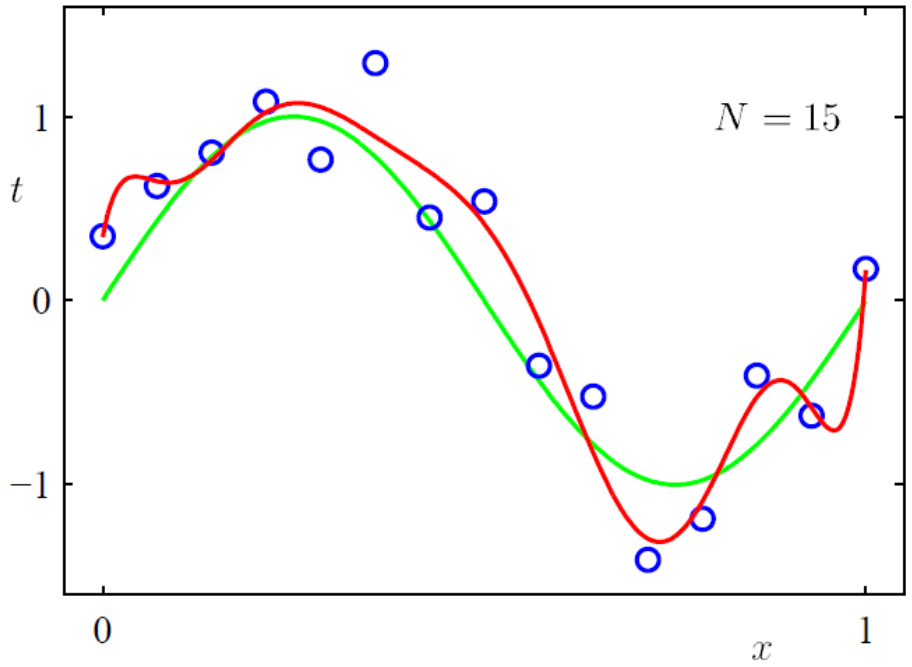
$$\tilde{E}(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{\lambda}{2}||\boldsymbol{w}||^2$$

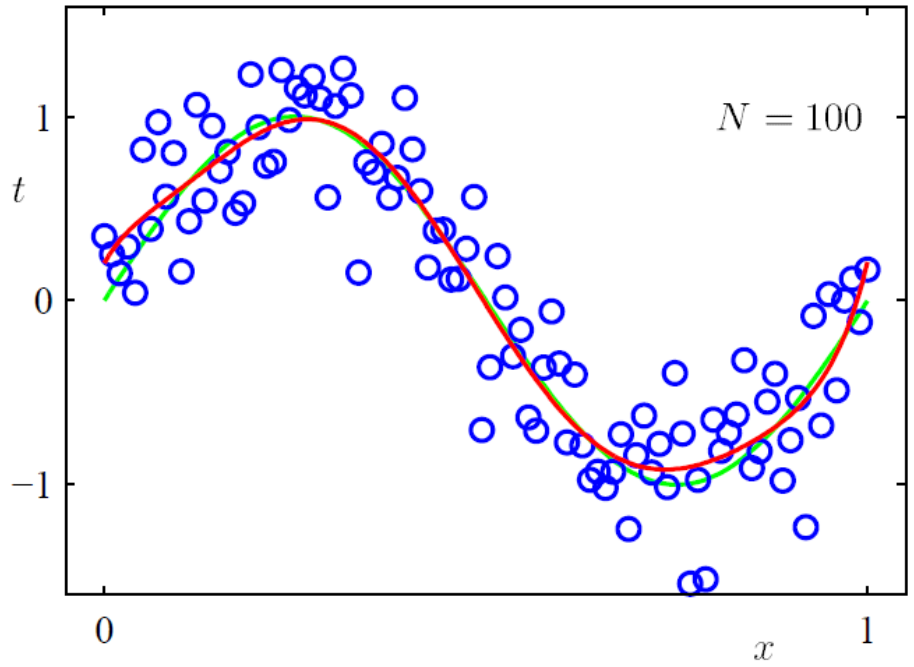




	$\ln\lambda = -\infty$	$\ln\lambda = -18$	$\ln\lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18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

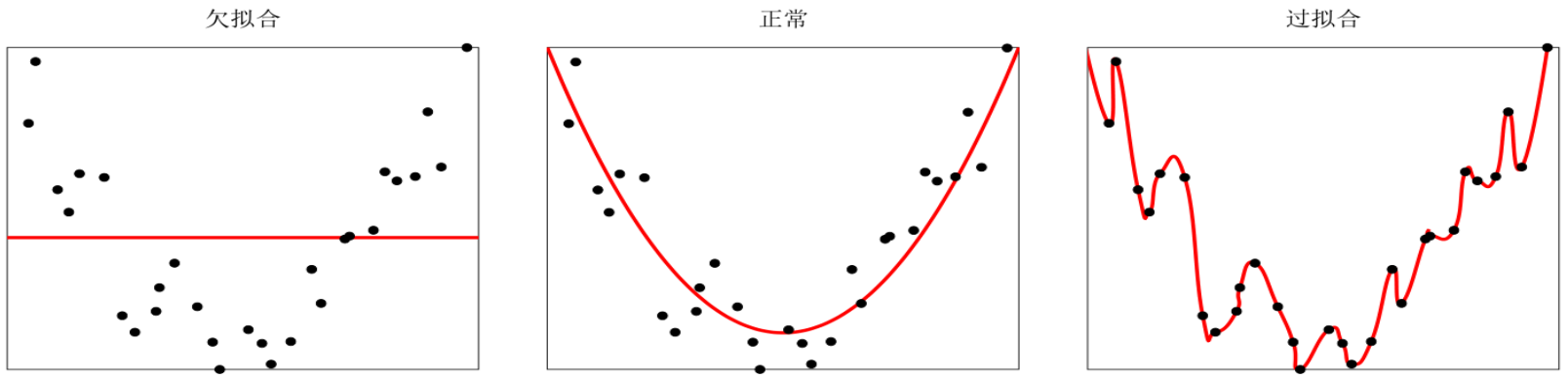




机器学习的几个关键点

机器学习 = 优化?

机器学习 = 优化? NO!



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

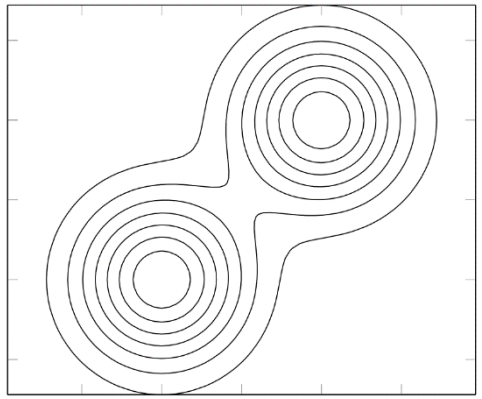
期望风险

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

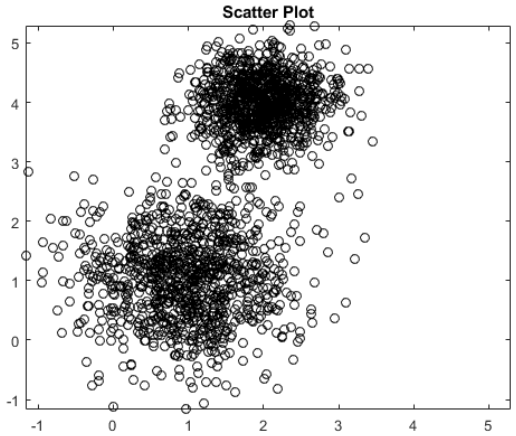
经验风险

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

真实分布 p_r



$\neq$



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

泛化错误

如何减少泛化错误?

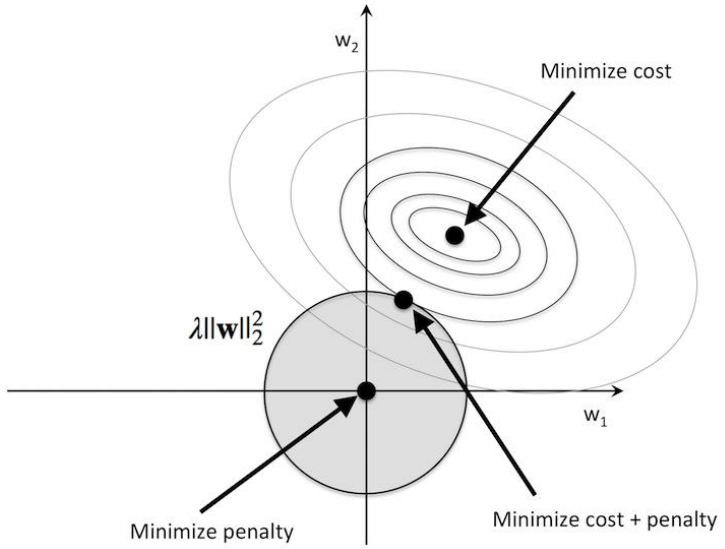
优化
经验风险最小

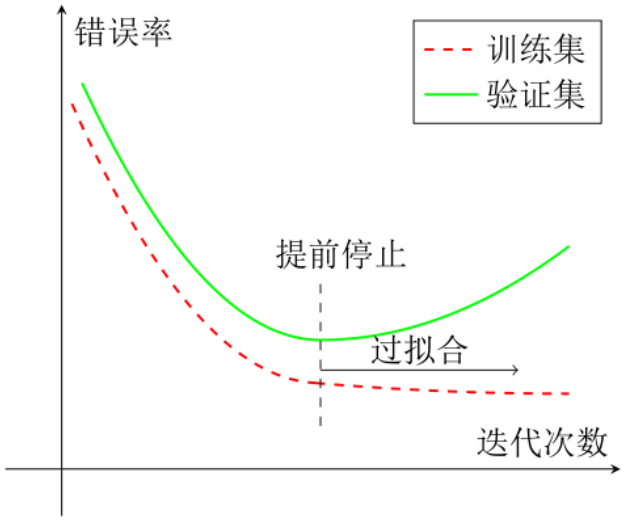
正则化
降低模型复杂度

正则化（regularization）

所有损害优化的方法都是正则化。

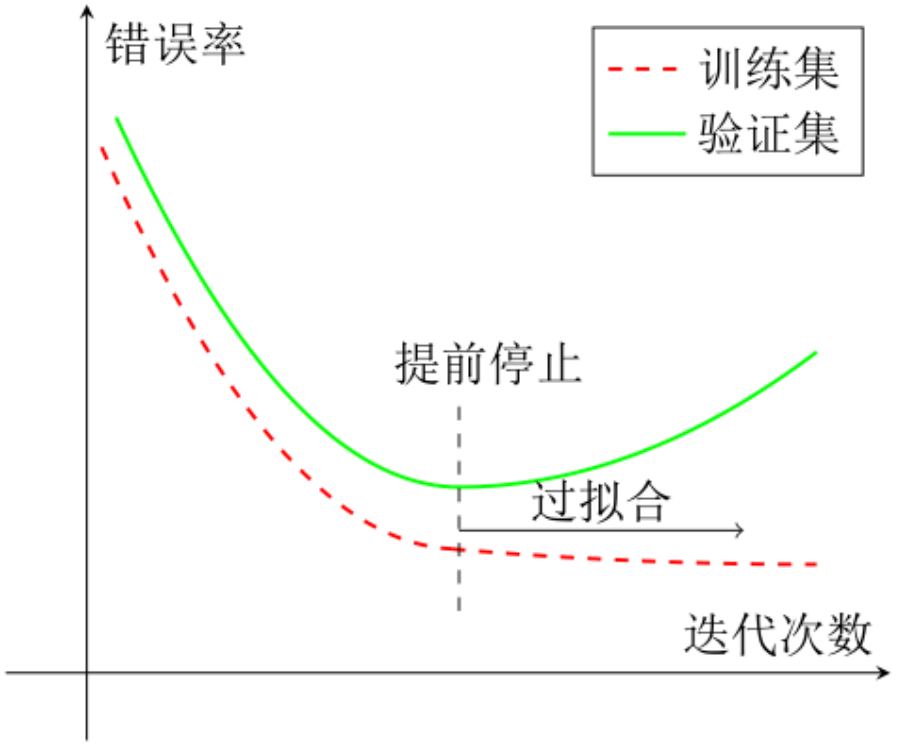
- 增加优化约束：L1/L2约束、数据增强
- 干扰优化过程：权重衰减、随机梯度下降、提前停止





提前停止

- 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。



常见的机器学习类型

	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?

- 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

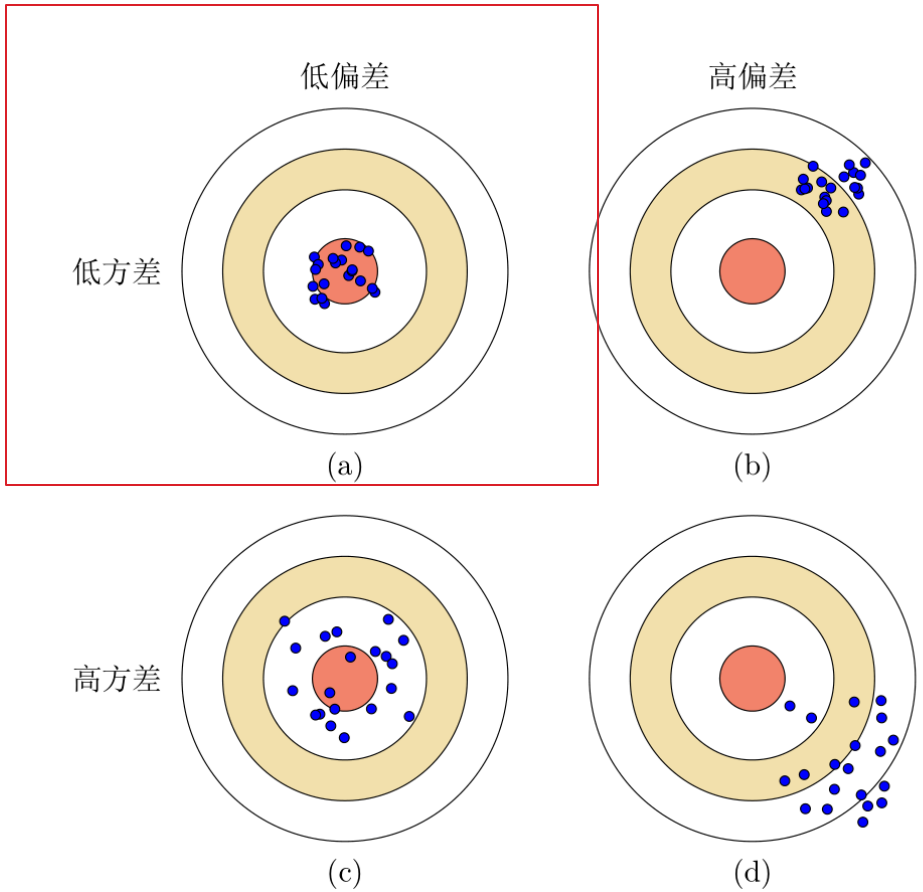
$$\mathcal{R}(f) = (\text{bias})^2 + \text{variance} + \varepsilon.$$

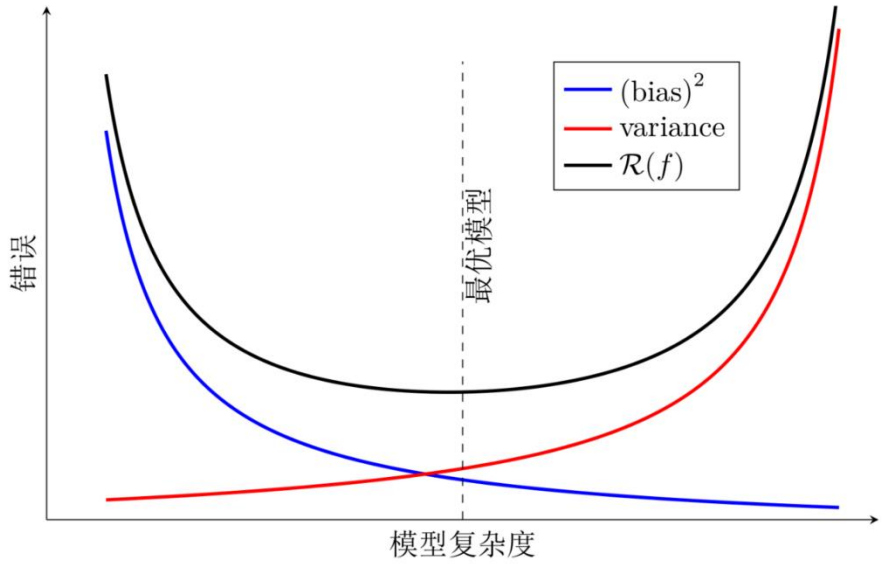
$$\mathbb{E}_{\mathbf{x}}\left[\left(\mathbb{E}_{\mathcal{D}}\left[f_{\mathcal{D}}(\mathbf{x})\right] - f^*(\mathbf{x})\right)^2\right]$$

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

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

模型选择：偏差与方差





集成模型：有效的降低方差的方法

PAC学习

- PAC： Probably Approximately Correct

- 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

$$P\left((\mathcal{R}(f) - \mathcal{R}_{\mathcal{D}}^{emp}(f)) \leq \epsilon\right) \geq 1 - \delta$$

近似正确，$0 < \epsilon < 0.5$

可能，$0 < \delta < 0.5$

样本复杂度

- 如果固定ϵ,δ，可以反过来计算出样本复杂度为

$$n(\epsilon,\delta) \geq \frac{1}{2\epsilon^2}(\ln|\mathcal{F}| + \ln\frac{2}{\delta})$$

  - 其中$|\mathcal{F}|$为假设空间的大小，可以用Rademacher复杂性或VC维来衡量。

- PAC学习理论可以帮助分析一个机器学习方法在什么条件下可以学习到一个近似正确的分类器。

- 如果希望模型的假设空间越大，泛化错误越小，其需要的样本数量越多。

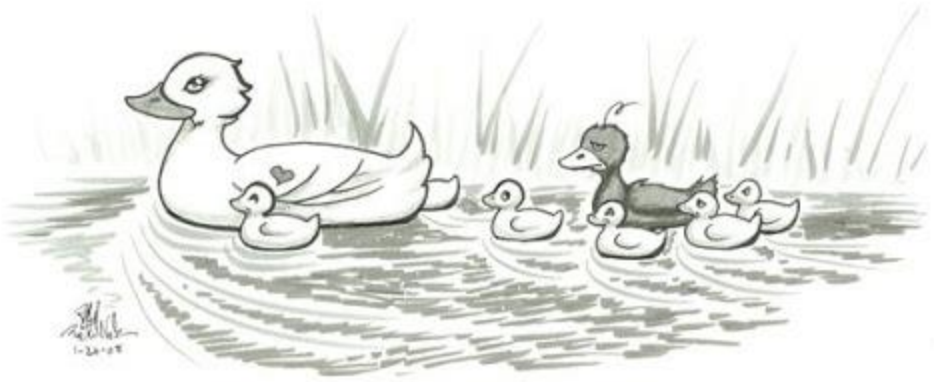
常用的定理

常用的定理

- 没有免费午餐定理（No Free Lunch Theorem，NFL）
  - 对于基于迭代的最优化算法，不存在某种算法对所有问题（有限的搜索空间内）都有效。如果一个算法对某些问题有效，那么它一定在另外一些问题上比纯随机搜索算法更差。

常用的定理

- 丑小鸭定理(Ugly Duckling Theorem)
  - 丑小鸭与白天鹅之间的区别和两只白天鹅之间的区别一样大.



常用的定理

- 奥卡姆剃刀原理(Occam's Razor)
  - 如无必要，勿增实体

课后作业

- 掌握知识点
  - 矩阵微分
  - 概率论
  - 信息论
  - 约束优化
- 编程练习
  - chap2_linear_regression