机器学习
Machine Learning
第一讲：导论

张旭东

电子工程系，清华大学

zhangxd@tsinghua.edu.cn

1.课程的主要目的

- 机器学习已经成为一种解决诸多问题的有效工具，是多学科交叉的领域，应用面广泛；
- 考虑清华工科类学生的知识基础，以面向解决实际应用为目标，开设一门通用性和侧重性兼顾的“机器学习”课程。
- 为学生掌握机器学习的本质和算法，以解决实际问题和为开展与本方向相关的研究打下基础。

2.本课程的主要内容

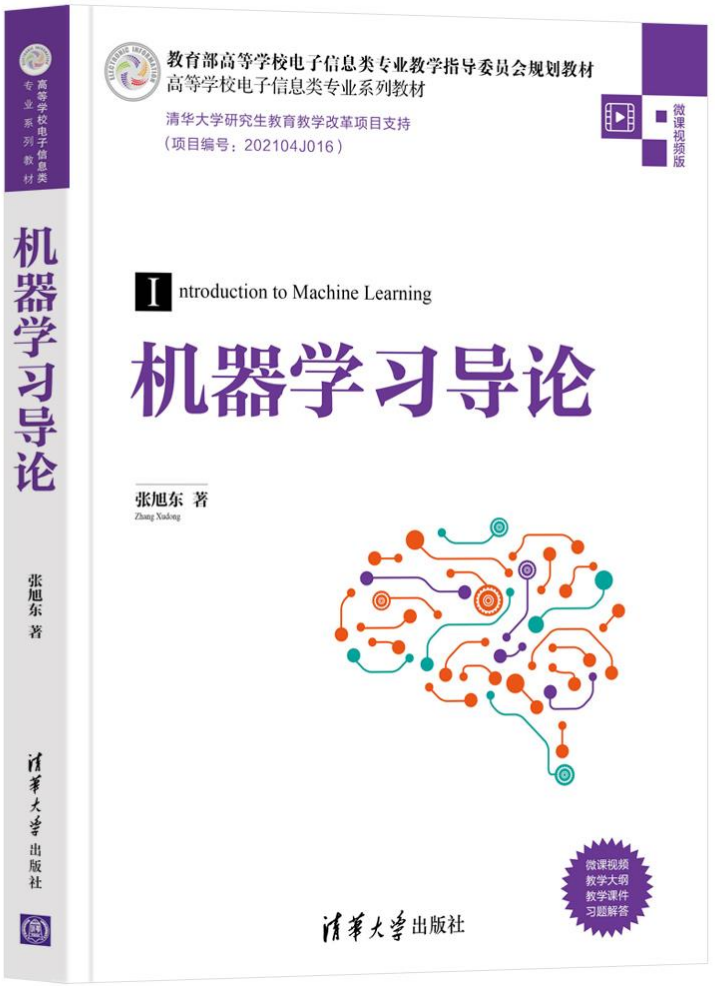
- 课时要求：周3学时，研究生本科贯通课程，秋季学期
- 机器学习的基本知识（15学时）
  - （基本概念、统计基础、回归学习、分类学习的基本算法，机器学习理论简介）
- 核方法和支持向量机（5学时）
- 决策树和集成学习（4学时）
- 神经网络+深度学习（9学时）
- 无监督学习（聚类、EM算法、降维和隐变量学习）（4学时）
- 强化学习+深度强化学习（8学时）

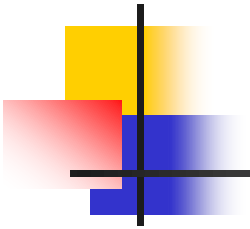
本课程内容的平衡性

- 深度学习很重要，尤其几种商业化应用：计算机视觉、语音识别、自然语言处理、推荐系统等，既有大数据支持，又可以通过大规模计算系统进行训练（学习），取得许多重要进展。但并不是所有应用都有必要使用深度学习，许多问题用传统机器学习已可以取得很好的结果；
- 目前神经网络的第三次复兴（以前有两次衰落过程），但并不能证明深度学习就是智能技术的终极方法，其他方法的延申或新方法仍有可能取得突破；
- 作为机器学习的基础课程，选择在几种主要机器学习方法（包括深度学习作为一种重要方法）之间平衡。

课本

- 教材：
  - 张旭东《机器学习导论》清华大学出版社
  - 或张旭东《机器学习教程》清华大学出版社





参考书

- 参考书
  - C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006
  - I. J. Goodfellow,Y. Bengio, et al, Deep Learning. 2017
  - R. Sutton, G. Barto, Reinforcement Learning, second edition, 2018
  - T. Hastie, et al The Elements of Statistical Learning, Springer, 2009
  - Mehryar Mphri et. al. Foundation of Machine Learning , MIT Press, 2012

扩充阅读（超出技术层面）

- M. Minsky, The Society of Mind，心智社会，机械出版社，2018
- Yann LeCun，Quand La Machine Apprend，杨立昆，科学之路-人、机器与未来，中信出版集团，2021

考核

- 作业：习题+仿真实验报告
- 考试成绩：作业与课程报告（为主）结合

- 预先要求：要求python编程能力，若没有python编程基础，请自学。
- 助教有一个关于编程和project的指导性讲座（线上进行）

“机器学习”课程与其他课程的关系

- 与“模式识别”的关系
  - 狭义讲，“模式识别”是一种任务，“机器学习”是一种通用工具，是目前模式识别用的最多的工具。
- 与“现代信号处理”的关系
  - 现代（自适应）滤波技术与机器学习的回归问题本质上是相同的，盲信号处理与无监督学习的很多工具是相同的。
- 与“现代统计学”的关系
  - 目前机器学习的主要评价准则和目标函数是建立在统计学基础上的。

“机器学习”课程与其他课程的关系

- 与“高等机器学习”的关系
  - “高等”是由MSRA的刘铁岩博士主持和召集的前沿课程，由MSRA的十几位一线研究员分讲各种研究前沿课题，部分系内老师参与讲课。
  - 相比而言本课程是机器学习的专业基础课程，介绍机器学习的基本原理和方法，在基础和前沿方面平衡，是入门性的课程。
  - 修“高等”课需要有机器学习的一定基础。本课程不假设有机器学习基础。

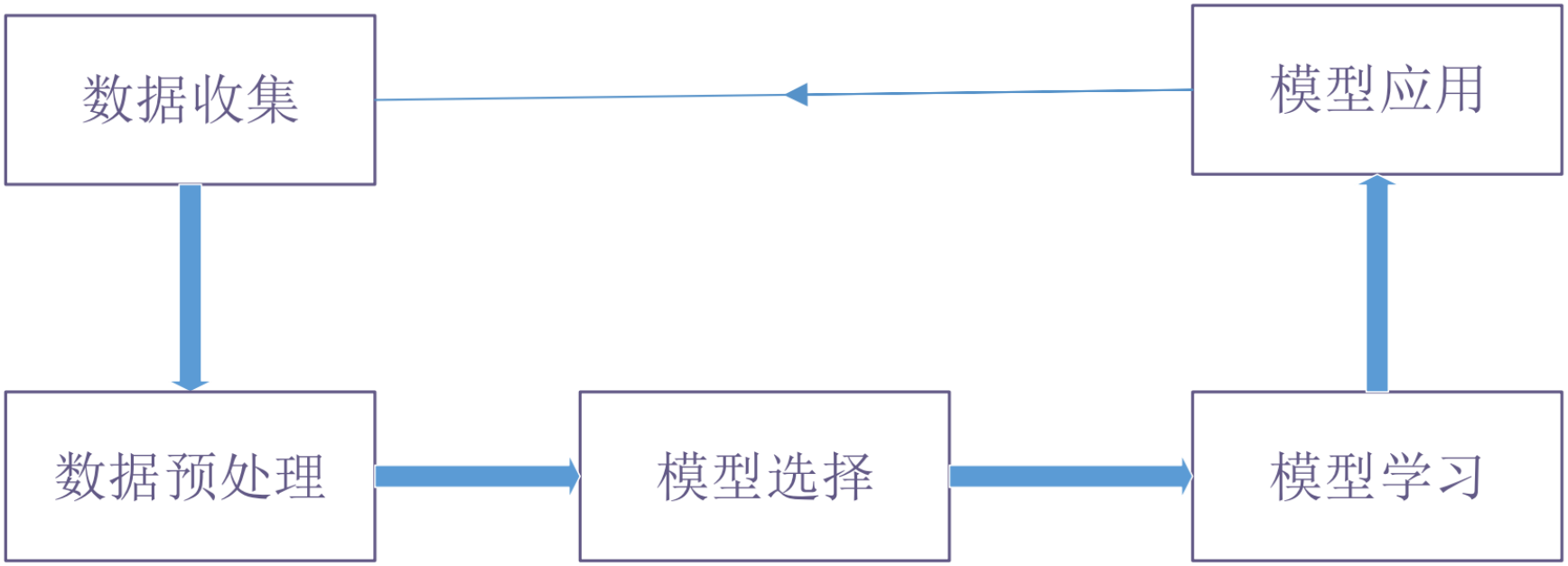
机器学习的重要会议和刊物

- International Conference on Machine Learning, ICML
- Neural Information Processing System, NeurIPS
- Conference On Learning Theory， COLT
- 其他（专业性）：ICLR，IJCAI、AAAI、 ICCV、CVPR、SIGIR、ICASSP、等等
- The Journal of Machine Learning Research
- Machine Learning
- Neural Computation
- IEEE Trans. On Neural Networks and Learning System
- IEEE Trans. On Pattern Analysis and Machine Intelligence

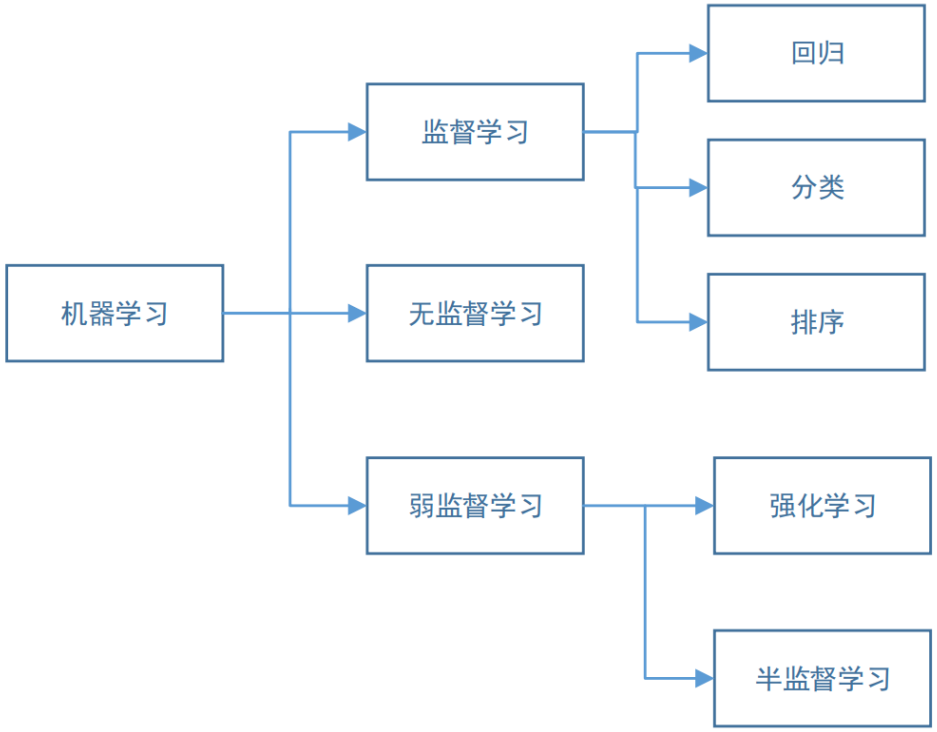
3.机器学习(ML)的参考定义

- 对于某类任务T和性能度量P，一个计算机程序被认为可以从经验E中学习是指，通过经验E的改进后，它在任务T上由性能度量P所衡量的性能有所提高。(Mitchell, 1997)
- Machine learning as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainly. (Murphy, 2012)

一个机器学习系统的基本流程



ML的主要类型
（一种分类方式）



- 监督学习（Supervised learning）
- 无监督学习（Unsupervised learning）
- 弱监督学习
  - 半监督学习
  - 增强（或称强化）学习（Reinforcement learning）

4.监督学习（Supervised learning）

- 数据集（标注集）

$$\mathbb{D} = \left\{ \left(\boldsymbol{x}_i, \boldsymbol{y}_i \right) \right\}_{i=1}^{N} \qquad \Rightarrow \boldsymbol{y} = h\left(\boldsymbol{x} \right)$$

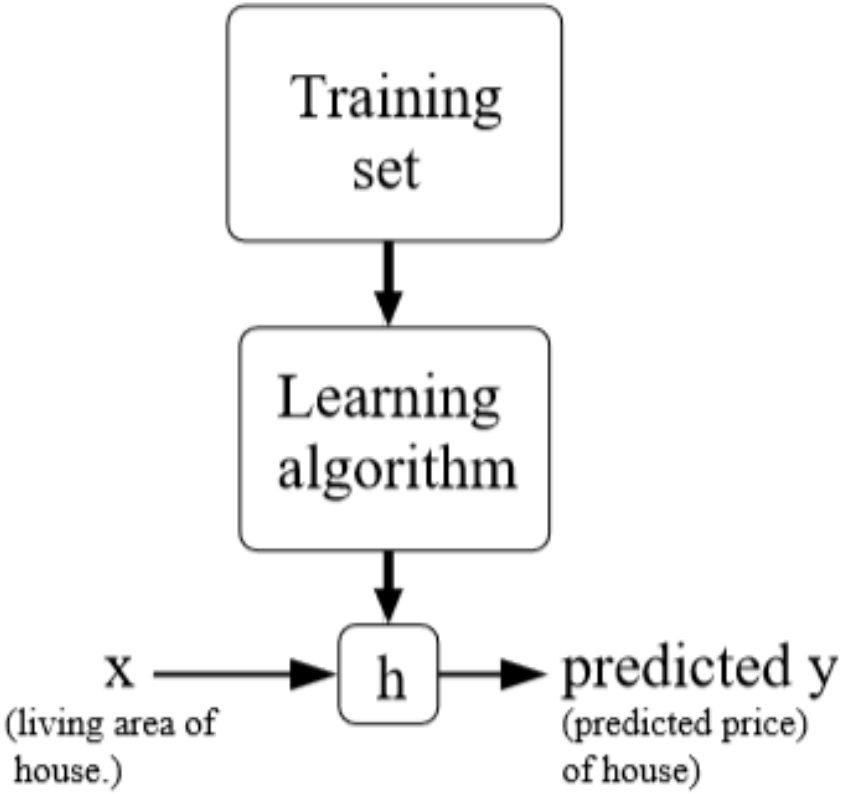
- 分类(Classification) $\quad y \in \{1, \cdots, \mathrm{C}\}$

$\mathrm{C} = 2$, *binary classification*

$C > 2$, multiclass *classification*

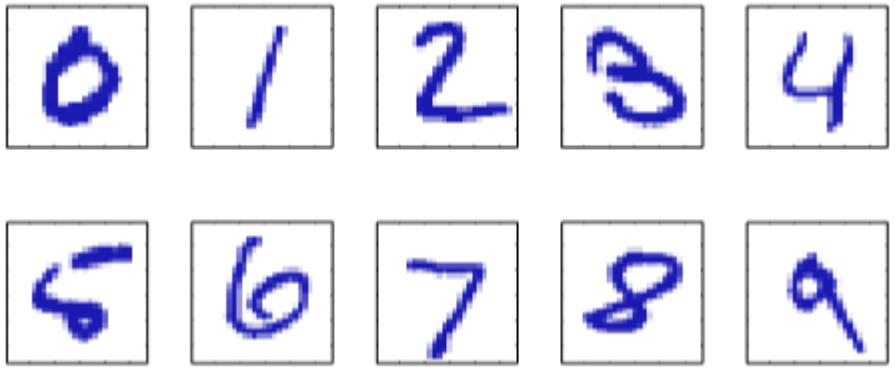
- 回归(Regression) $\quad \boldsymbol{y} \in \mathbb{R}$

Supervised Learning or Predictive



分类(Classification)问题例子

- 垃圾邮件检测（spam email or non-spam email, C=2）
- 手写数字识别 (C>2)



回归(Regression)问题例子

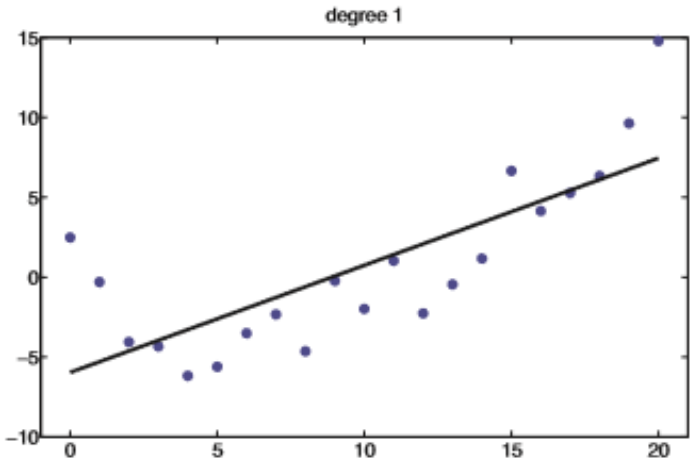
- 股票、房价、人口等建模和预测
- 通信信道建模和预测

$$y = h(x)$$

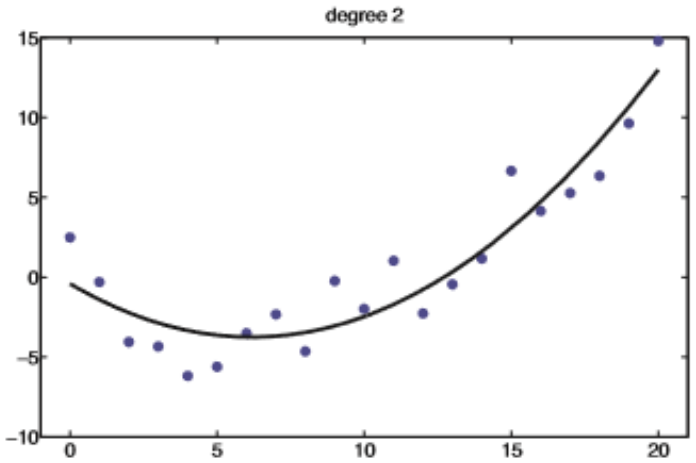
一个房价预测例
（2个自变量，1个输出变量）

Living area (feet^2)	#bedrooms	Price (1000$s)
2104	3	400
1600	3	330
2400	3	369
1416	2	232
3000	4	540
⋮	⋮	⋮

单变量回归问题的图示：线性和多项式



(a)



(b)

(a) Linear regression on some 1d data. (b) Same data with polynomial regression

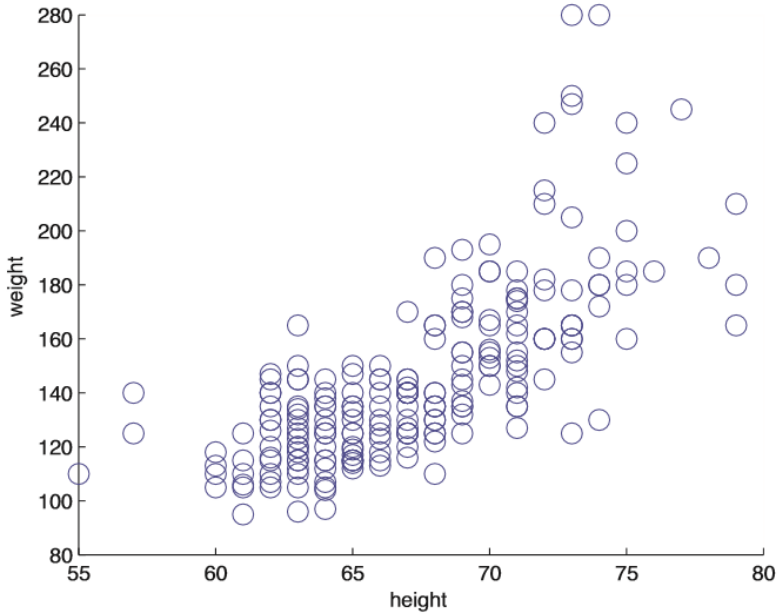
5.无监督学习（Unsupervised learning）

- Descriptive： to discover “interesting structure” in the data。(knowledge discovery)

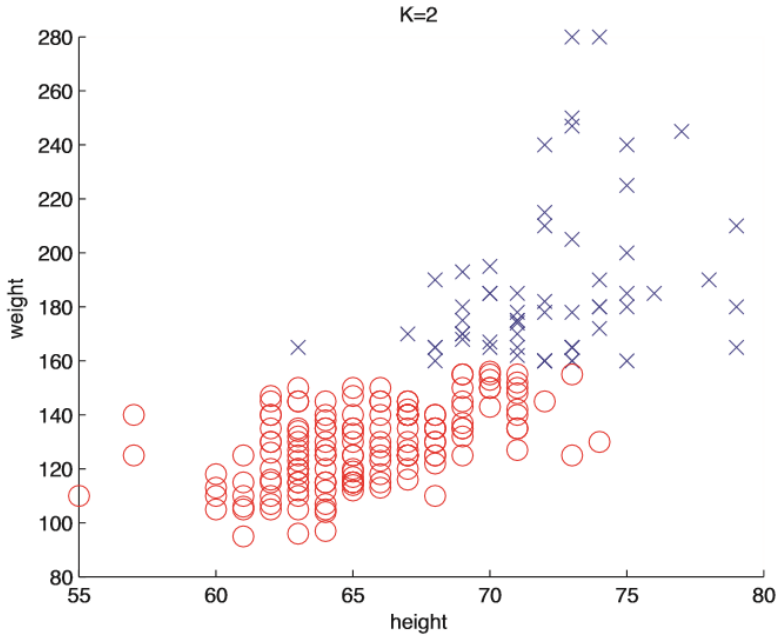
$$\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{N}$$

- examples of unsupervised learning
  - Discovering clusters
  - density estimation
  - Discovering latent factors
    - dimensionality reduction，PCA
    - ICA

聚类的一个实例



(a)



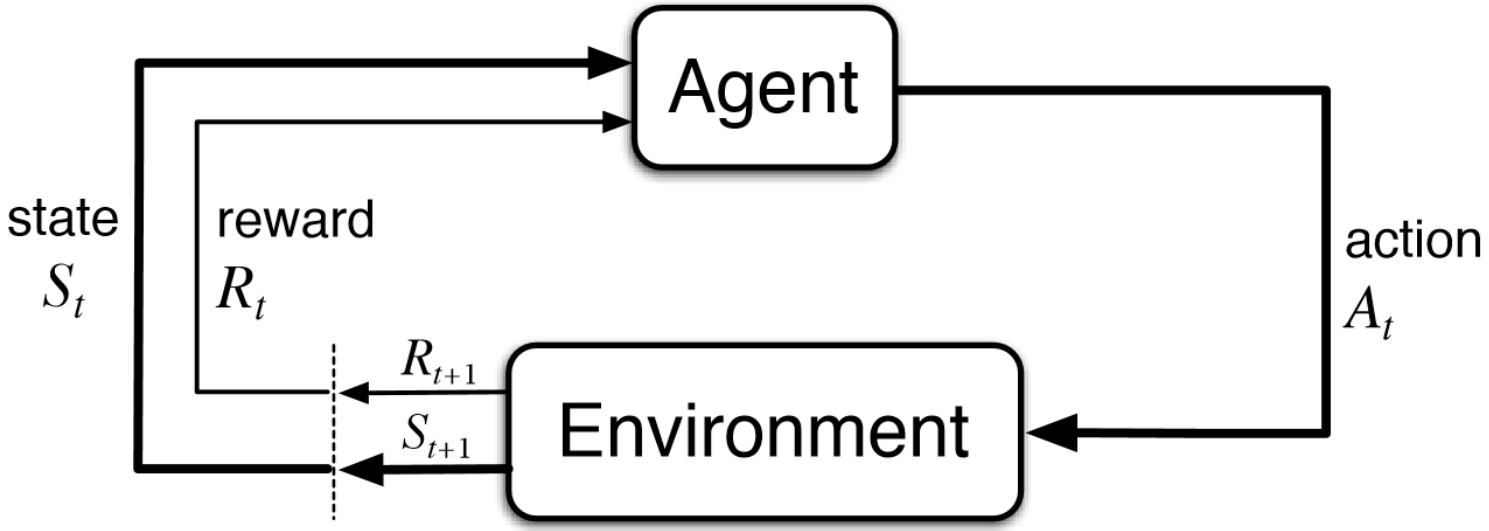
(b)

(a) The height and weight of some people.

(b) A possible clustering using $K = 2$ clusters.

6.强化学习（Reinforcement learning）

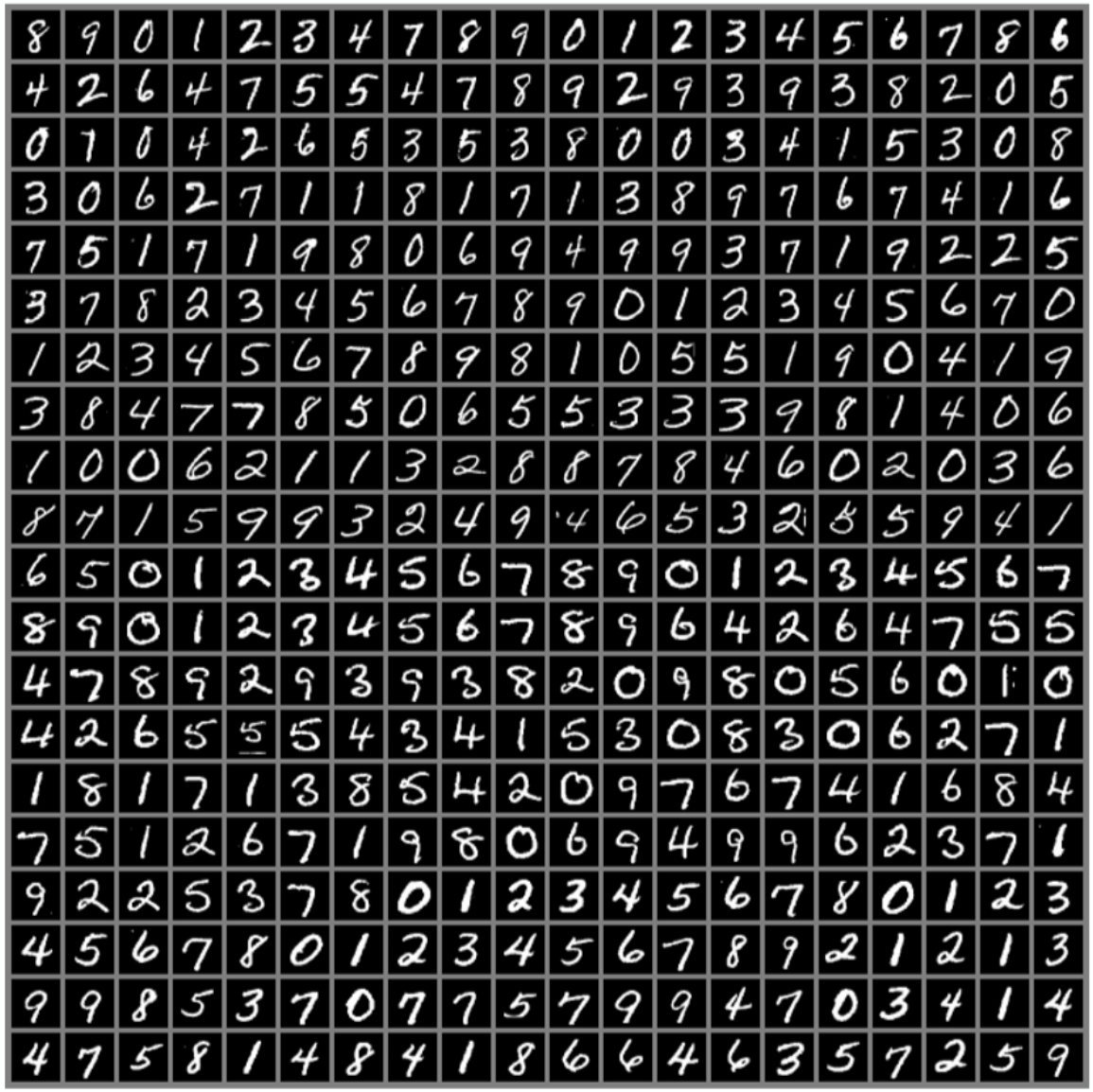
Reinforcement learning problems involve learning what to do—how to map situations to actions—so as to maximize a numerical reward signal.



7.构建机器学习算法：基本元素

- 特定的数据集（从数据学习）
- 代价函数（评价函数、风险函数）
- 模型
  - （不同类型、层次的各种模型：线性模型、非线性模型、参数模型、神经网络、深度神经网络。。。）
- 优化过程、优化算法

7.1 数据集示例：
手写数字识别MNIST数据集



美国国家标准与技术研究所发布

60000个训练集图像及其标注

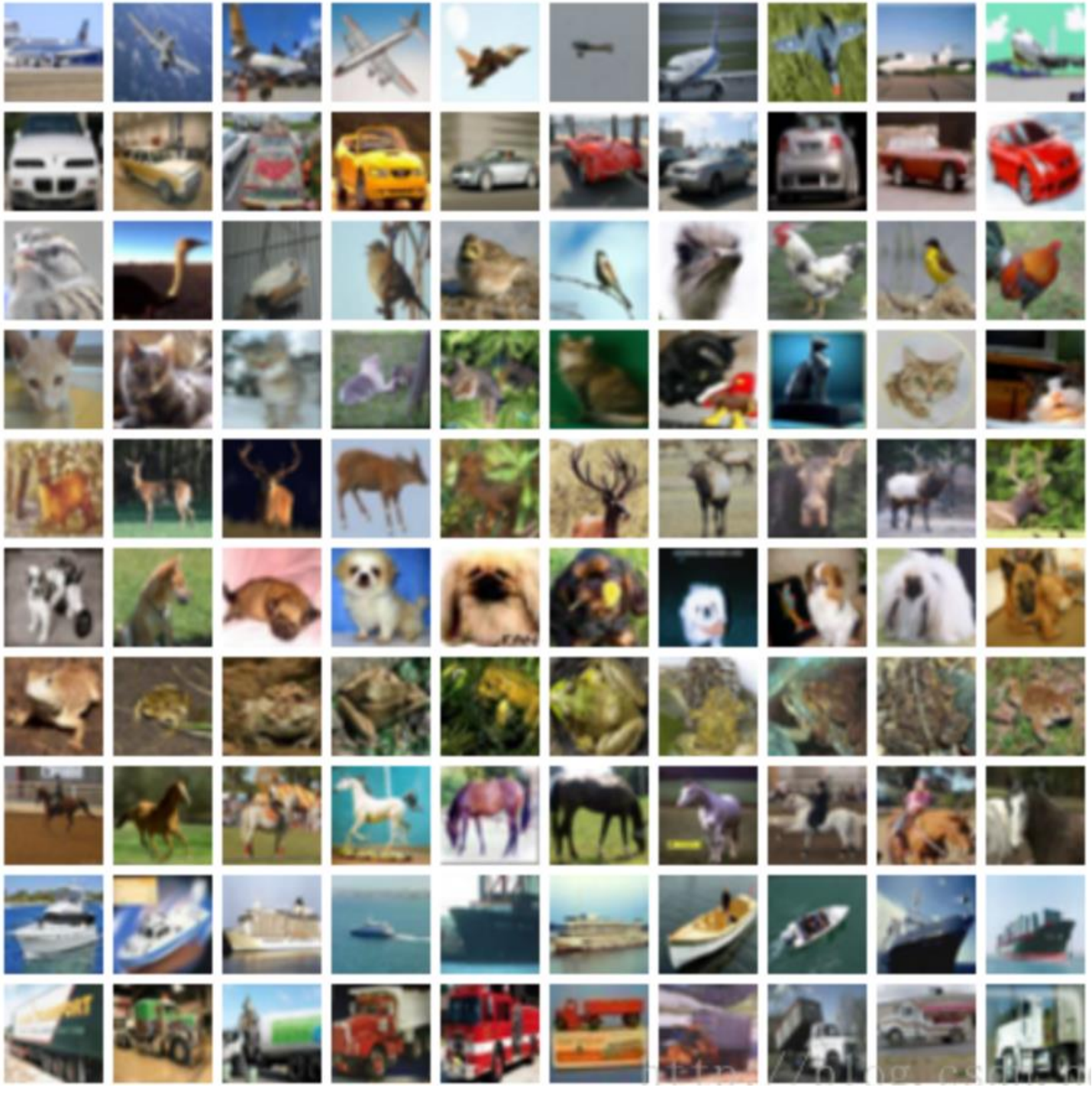
10000个测试集图像及其标注

28*28像素的手写数字图像

CIFAR-10数据集

- CIFAR-10 是一个包含60000张图片的数据集。其中每张照片为32*32的彩色照片，每个像素点包括RGB三个数值，数值范围 0 ~ 255。
- 所有照片分属10个不同的类别，分别是 'airplane', 'automobile', 'bird', 'cat', 'deer', 'dog', 'frog', 'horse', 'ship', 'truck'
- 其中五万张图片被划分为训练集，剩下的一万张图片属于测试集。

CIFAR-10
数据集
示例



7.2 目标函数

- 风险函数和经验风险函数

风险函数

$$J^*(\boldsymbol{\theta}) = \mathrm{E}_{(\boldsymbol{x}, y)\sim p_{data}} \left\{ L\left(f(\boldsymbol{x}; \boldsymbol{\theta}), y \right) \right\}$$

p_{data} 表示数据的生成分布

$L\left(f(\boldsymbol{x}; \boldsymbol{\theta}), y \right)$ 每个样本的损失函数

经验风险函数（训练集代价函数）

$$J(\boldsymbol{\theta}) = \mathrm{E}_{(\boldsymbol{x}, y)\sim \hat{p}_{data}} \left\{ L\left(f(\boldsymbol{x}; \boldsymbol{\theta}), y \right) \right\}$$

$$= \frac{1}{\mathrm{N}} \sum_{i=1}^{N} L\left(f(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}), y^{(i)} \right)$$

$\hat{p}_{data}$ 表示训练集的经验分布

或N表示训练集的样本数目

ML的训练和泛化

- 由于p_{data}实际中未知，ML的训练是以经验风险最小替代风险最小。
- 经验风险最小若推知风险最小，称为泛化能力强。
- 实际中，经验风险优化可能带来过拟合问题，使得泛化性能差。
- 经验风险结合正则化等技术进行优化，可提升泛化能力。

7.3 模型、优化

- 不同类型、层次的各种模型
- 参数模型和非参数模型
- 线性模型、非线性模型
  - 逻辑回归、决策树、SVM、神经网络、深度神经网络
- 优化算法（一般是相对独立的模块）

7.4 机器学习的一些基本术语

- 训练集、测试集
- 训练误差、测试误差
- 欠拟合（Under-fitting）
- **过拟合（Over-fitting）**
- 泛化性能（Generalization）
- **正则化（ Regularization）**

8.机器学习系统的实际性能评价

- 在样本集上进行实际性能评价

$$\mathbf{D}=\left\{\left(\boldsymbol{x}_i,y_i\right)\right\}_{i=1}^{N}$$

- 回归的基本评估指标：均方误差

$$E_{mse}\left(h\right)=\frac{1}{N}\sum_{i=1}^{N}\left(h(\boldsymbol{x}_i)-y_i\right)^2$$

- 分类的基本指标：分类错误率和分类准确率

$$E=\frac{1}{N}\sum_{i=1}^{N}I\left(h(\boldsymbol{x}_i)\neq y_i\right)$$

$$Acc=\frac{1}{N}\sum_{i=1}^{N}I\left(h(\boldsymbol{x}_i)=y_i\right)=1-E$$

精度和查全率（针对二分类定义）

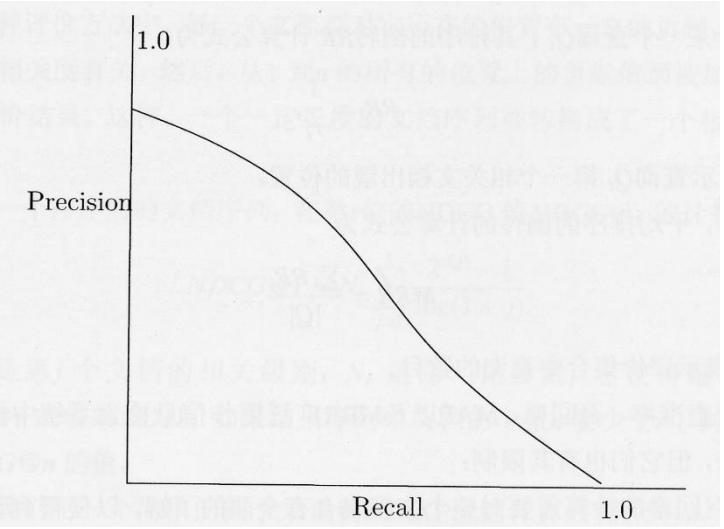
标注的真实类型	分类器返回的类型	
	正类	负类
正类	N_{TP}	N_{FN}
负类	N_{FP}	N_{TN}

精度（Precision）

$$\mathbf{Pr} = \frac{N_{TP}}{N_{TP} + N_{FP}}$$

查全率（Recall）

$$\mathbf{Re} = \frac{N_{TP}}{N_{TP} + N_{FN}}$$



样本集：300正类、9700负类
该例子中，若将所有样本
判为负例，则*Acc*=0.97

癌症的例子

标注的真实类型	分类器返回的类型	
	正类	负类
正类	210/260	90/40
负类	200/400	9500/9300

“/”之上侧所示

$$\mathrm{Pr} \approx 0.51, \quad \mathrm{Re} = 0.7 \qquad Acc = 0.971$$

“/”下侧的数据

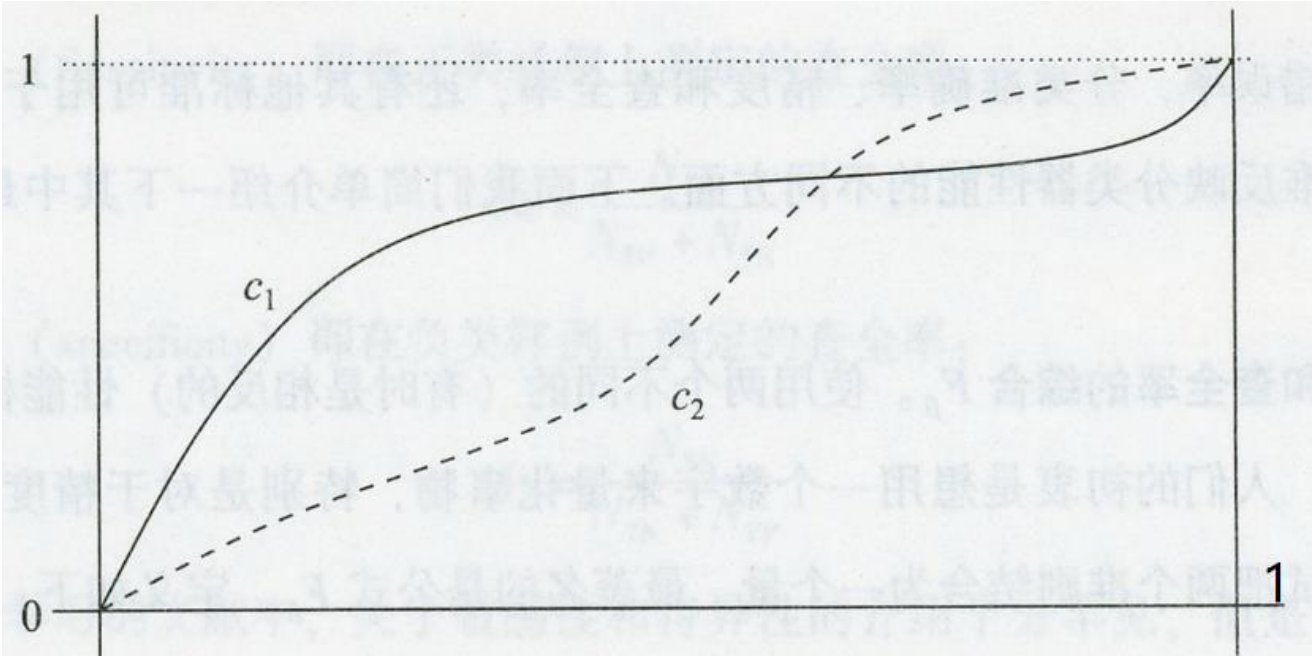
$$\mathrm{Pr} \approx 0.39, \quad \mathrm{Re} \approx 0.87 \qquad Acc = 0.956$$

ROC 曲线

正样分类准确率 $P_{Ac} = \dfrac{N_{TP}}{N_{TP} + N_{FN}}$

负样错误率 $N_e = \dfrac{N_{FP}}{N_{TN} + N_{FP}}$

以 P_{Ac} 为纵轴，以 N_e 为横轴 画出一条曲线



一种比较总体优劣的方法是采用AUG参数，一个分类器AUG参数（Area Under ROC Curve）表示为其ROC曲线之下和坐标横轴之间的面积

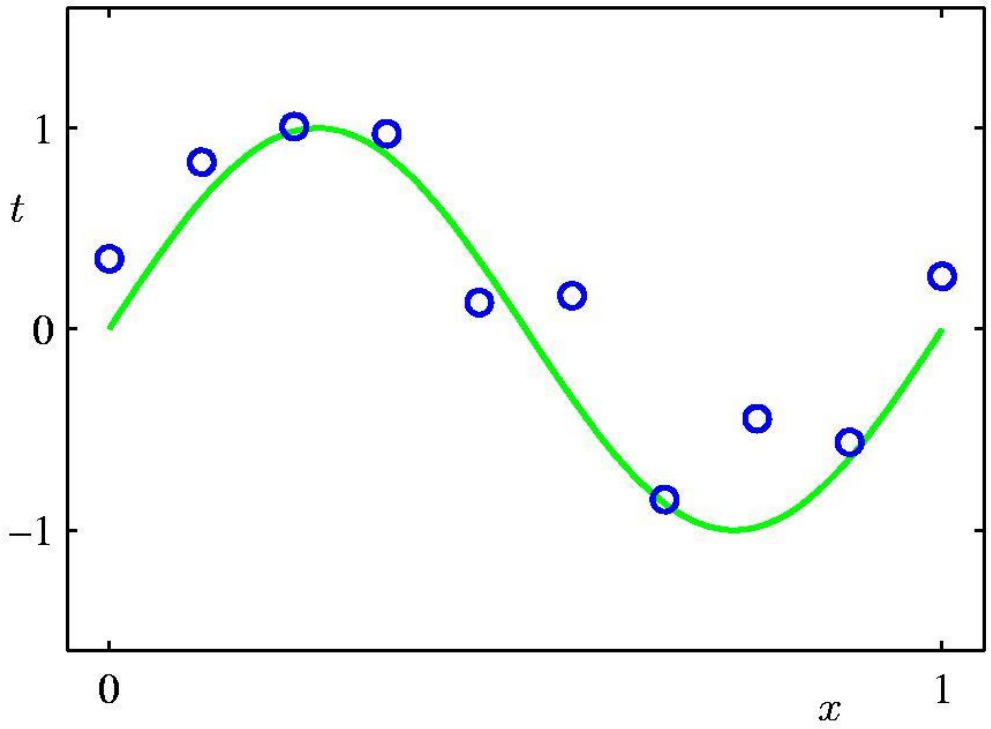
9.一个回归综合例子：多项式曲线拟合

Polynomial Curve Fitting

参数方法

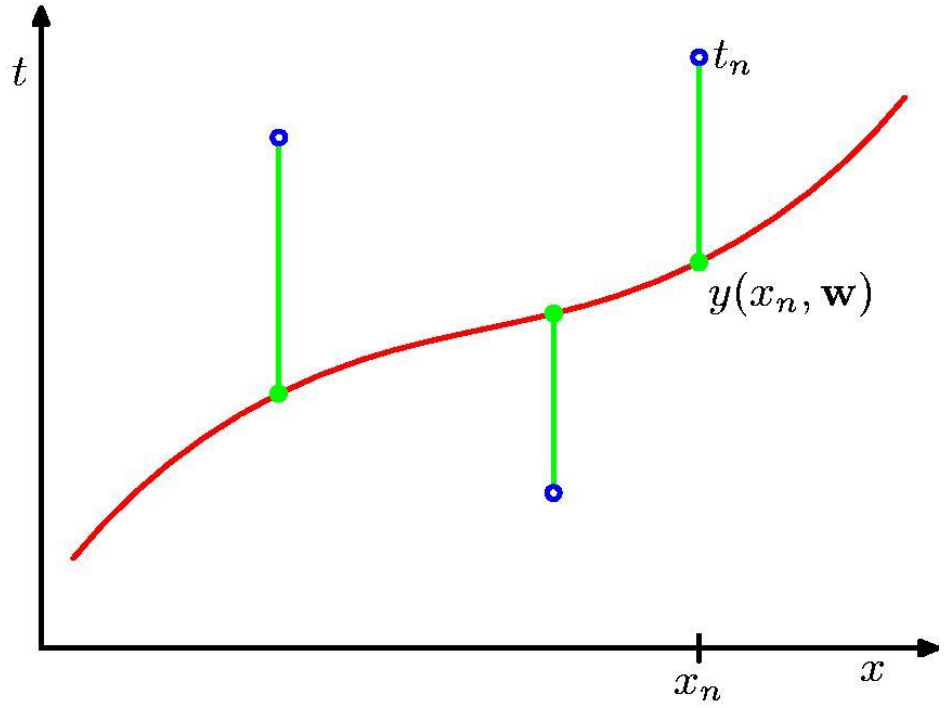
样本集为

$$\{x_n, t_n\}\Big|_{n=1}^{10}$$



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

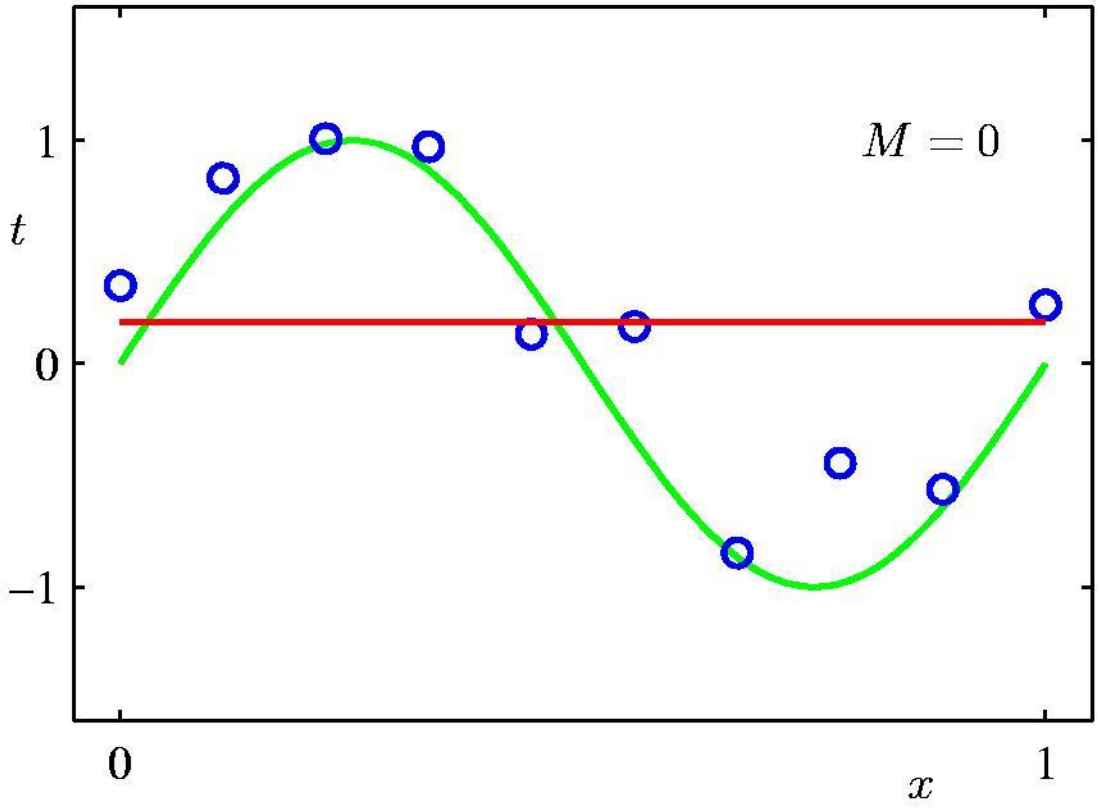
选择平方误差和作为评价函数
Sum-of-Squares Error Function



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

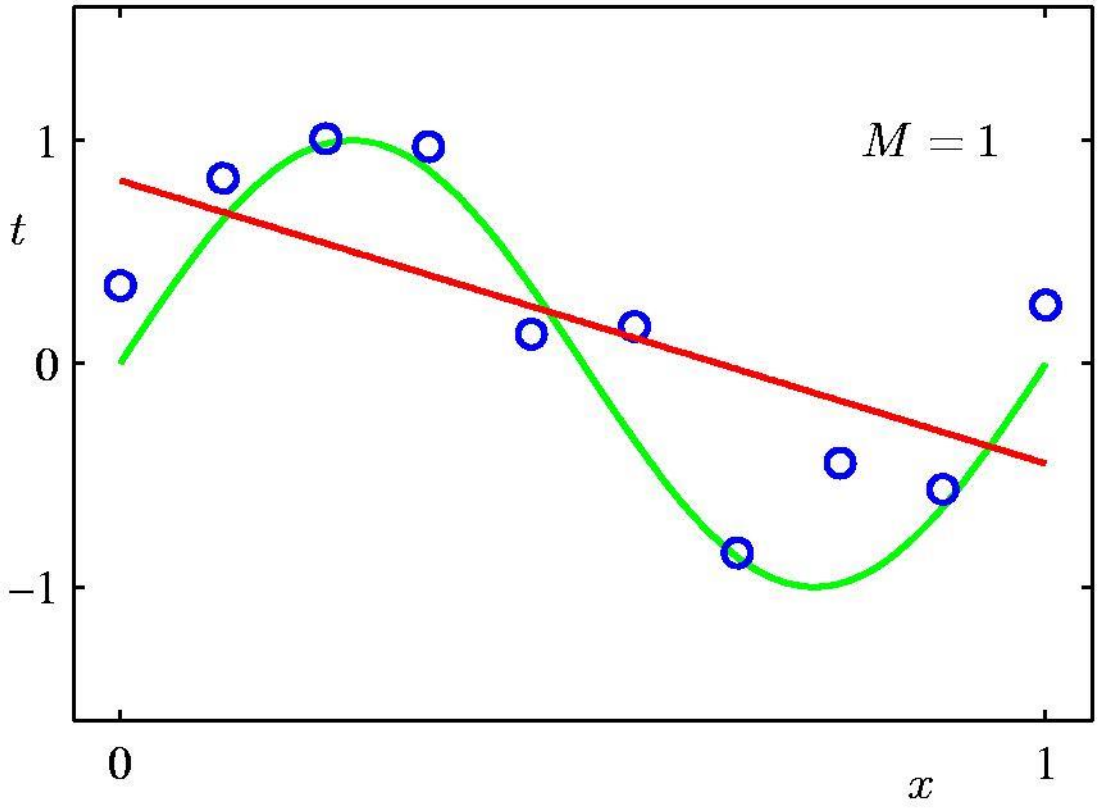
0^{th} Order Polynomial

欠拟合

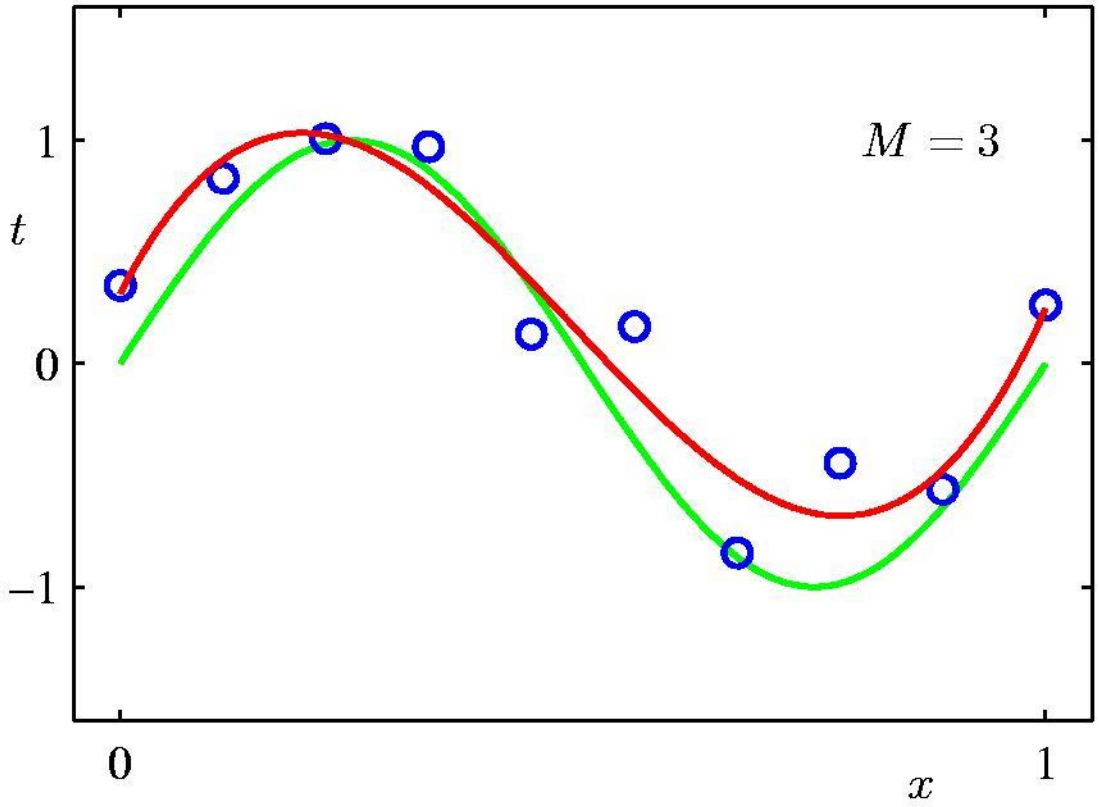


1st Order Polynomial

欠拟合

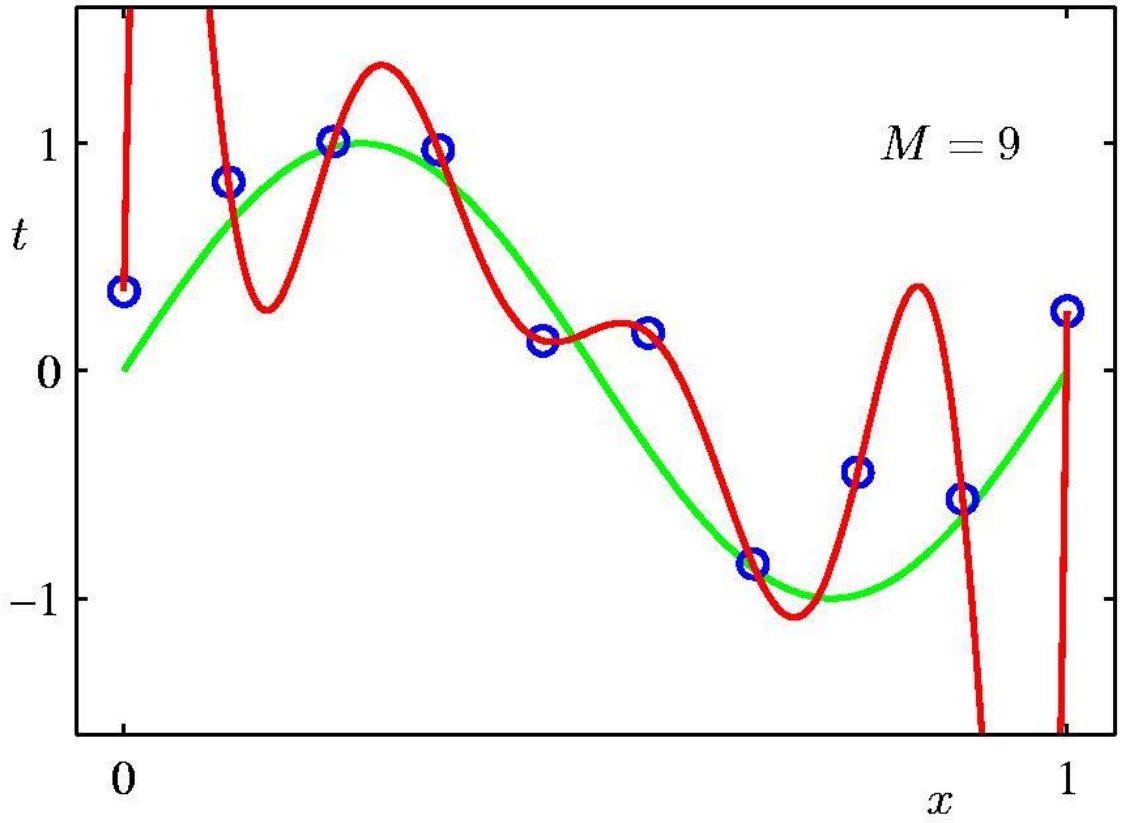


3rd Order Polynomial

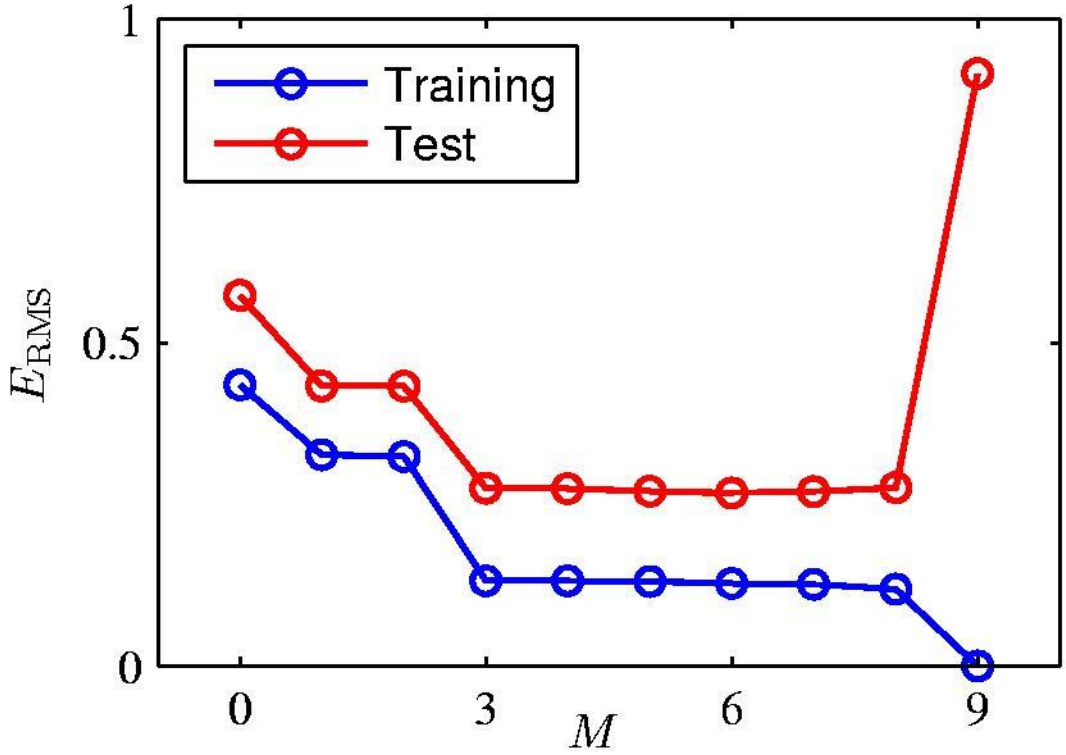


9th Order Polynomial

过拟合



不同阶的训练误差和测试误差（注意Over-fitting情况）



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

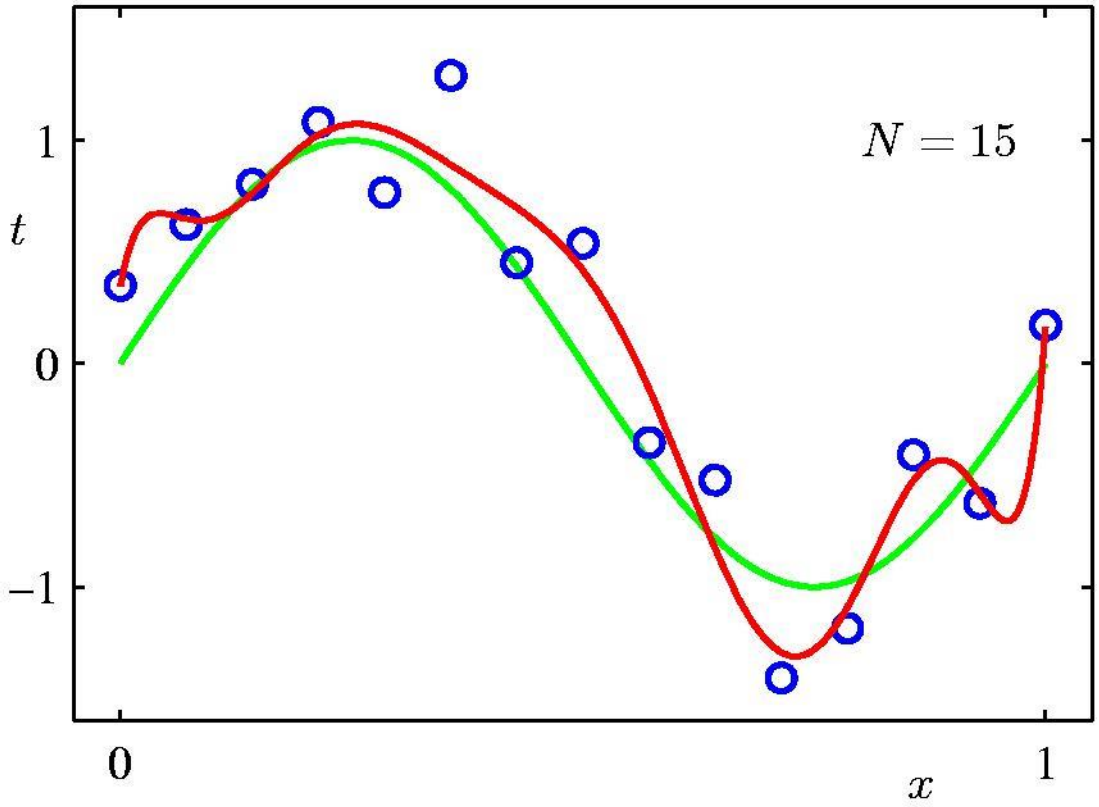
Polynomial Coefficients

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

增加训练数据集规模降低过拟合

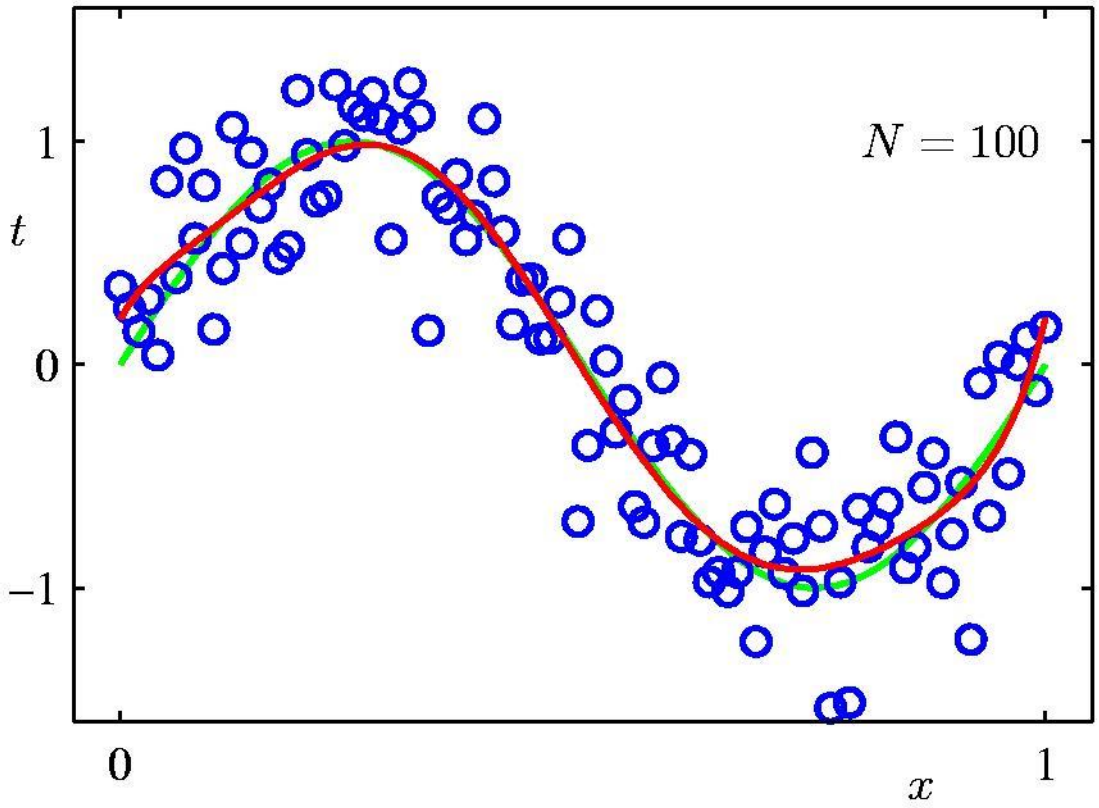
Data Set Size: $N = 15$

9th Order Polynomial



增加训练数据集规模降低过拟合
Data Set Size: $N = 100$

9th Order Polynomial



正则化（Regularization）

保持训练集规模条件下的降低过拟合方法

Penalize large coefficient values

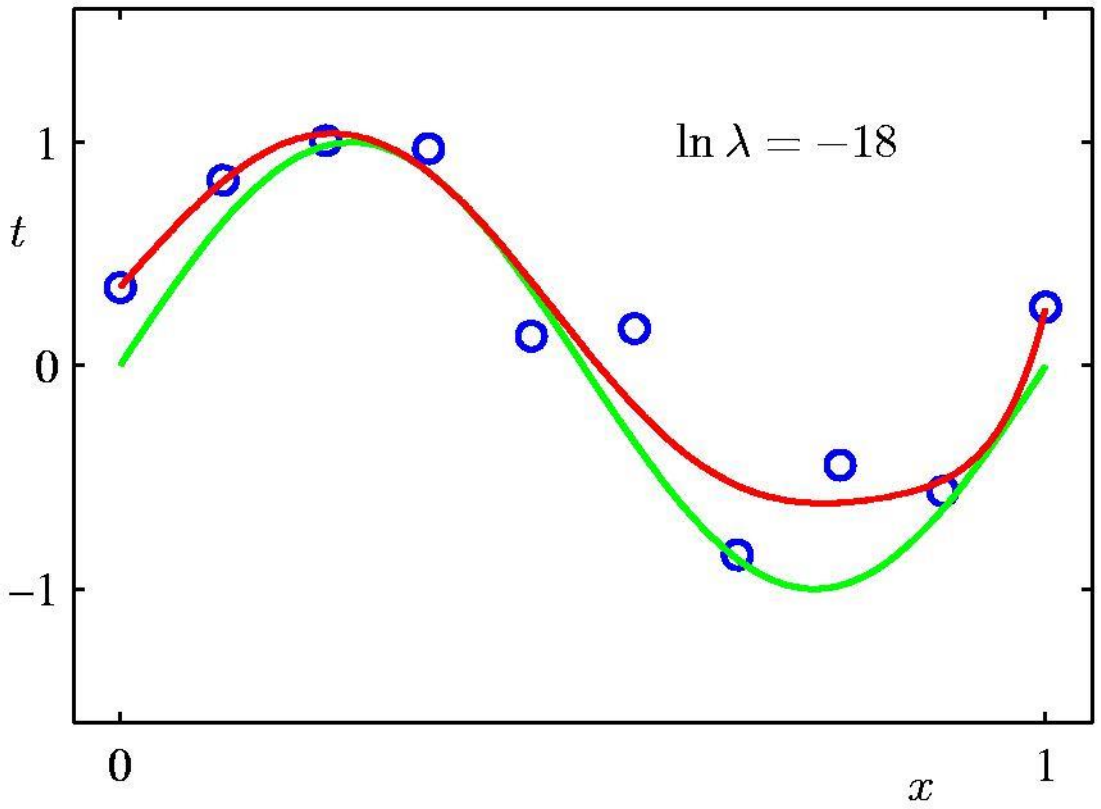
$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

这是正则化约束的一种例子

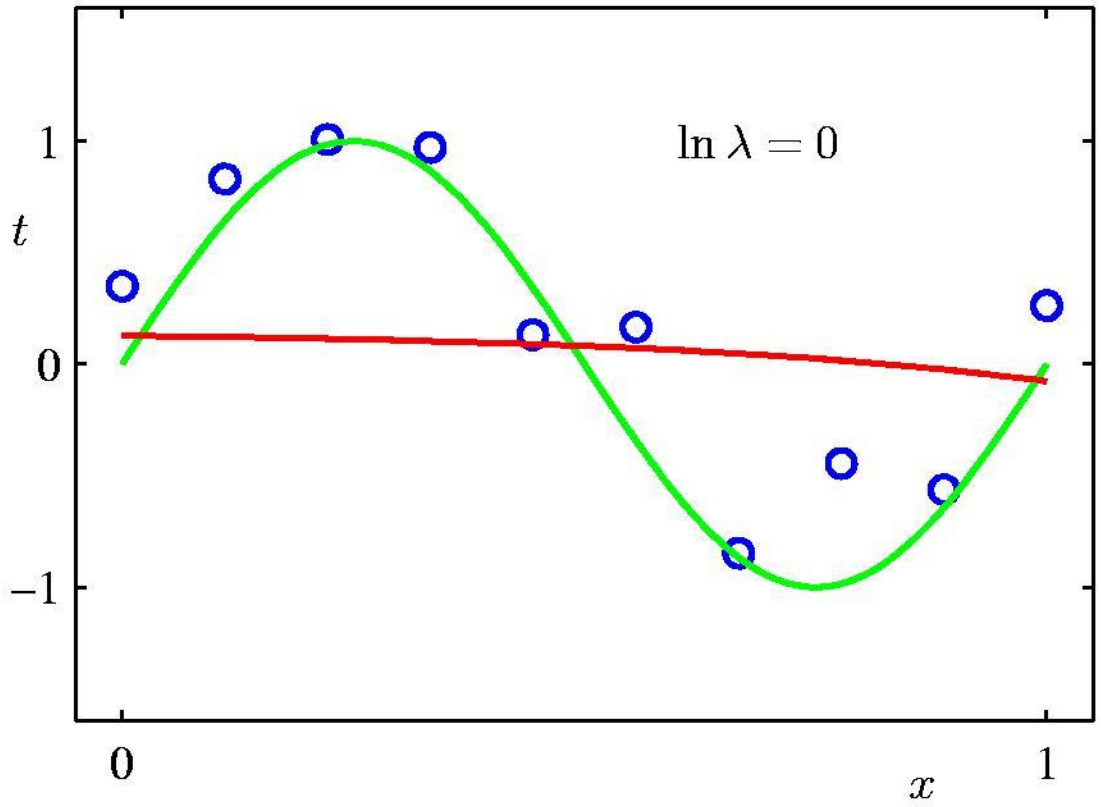
一种正则化对应一种“偏爱”选择，“权范数平方”对应的是偏爱范数小的权向量。

Regularization: $\ln \lambda = -18$

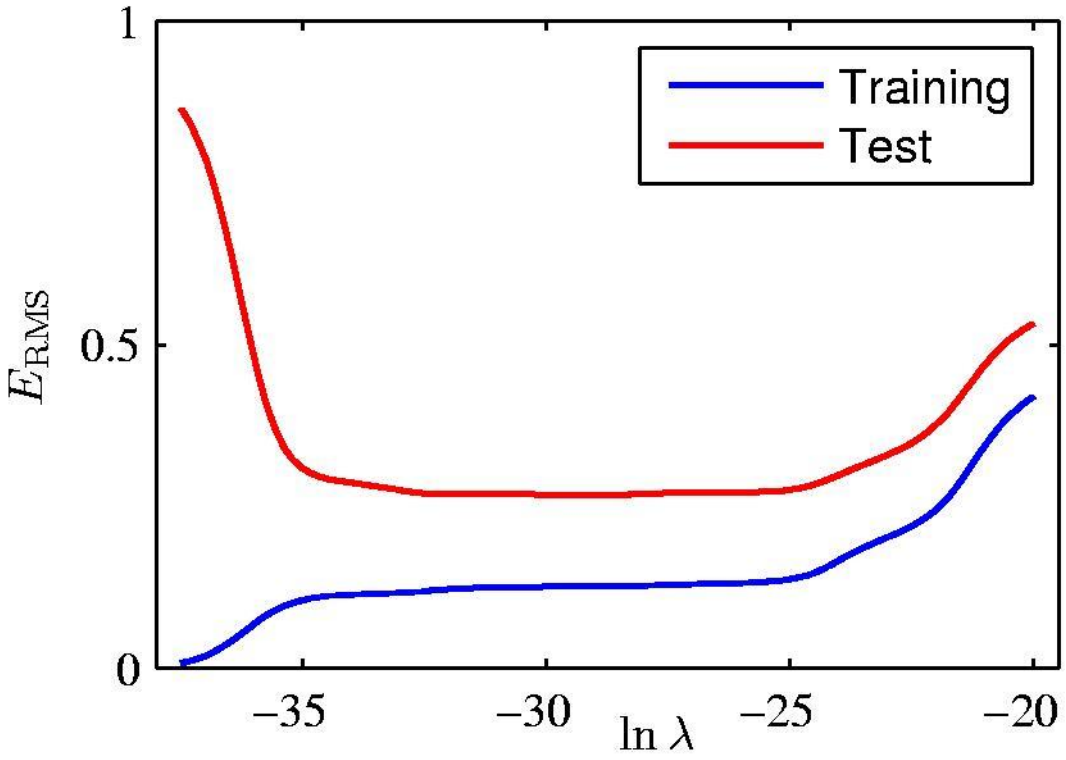
9th Order Polynomial



Regularization: $\ln \lambda = 0$



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

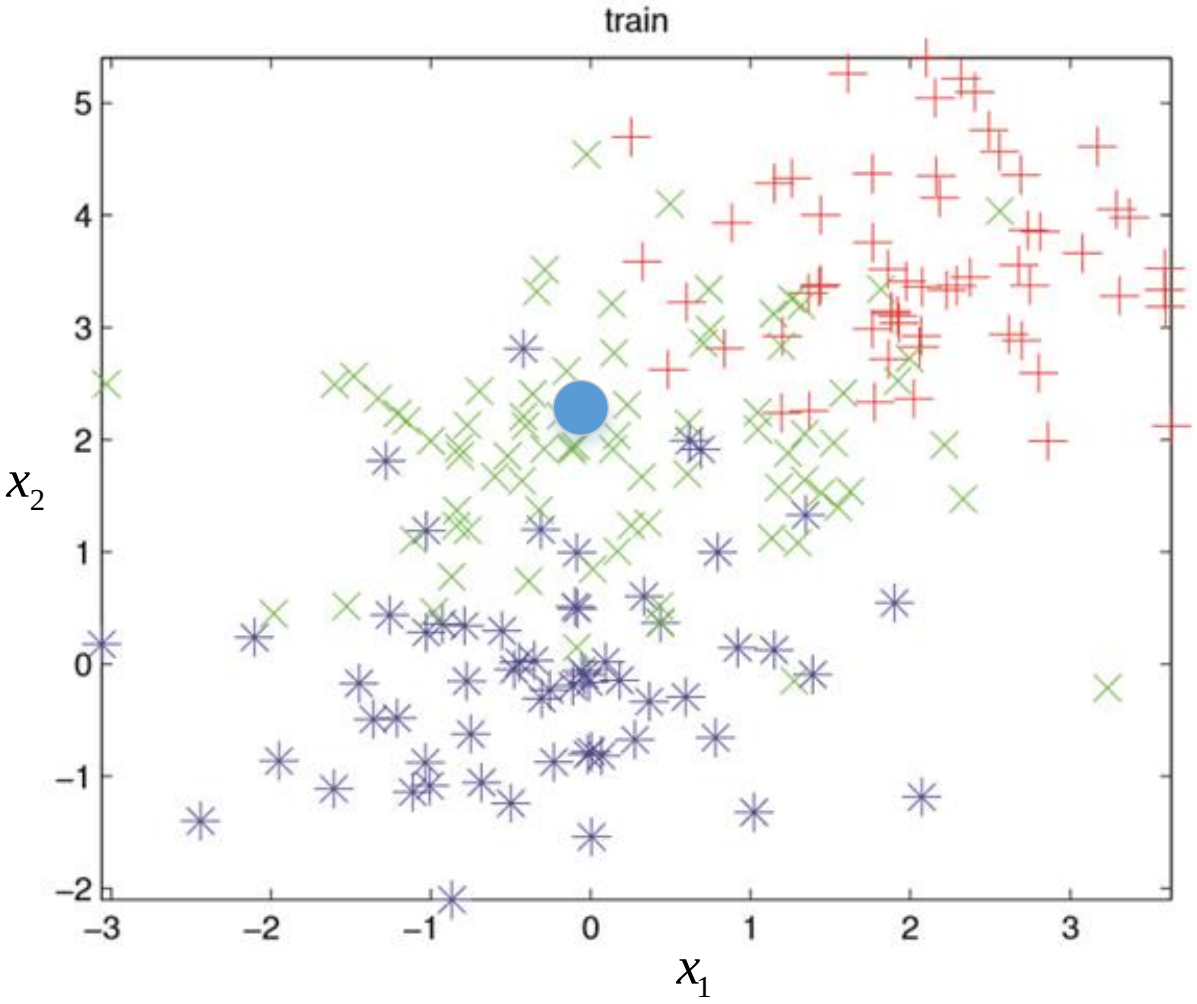


Polynomial Coefficients

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

10.一个分类例子：KNN

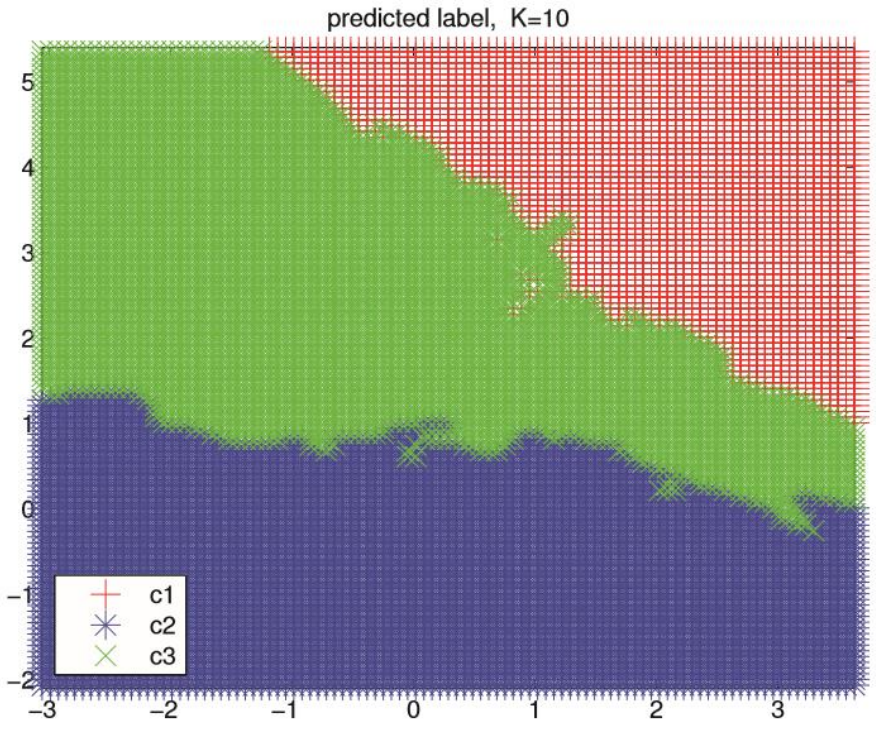
K-nearest neighbors：KNN

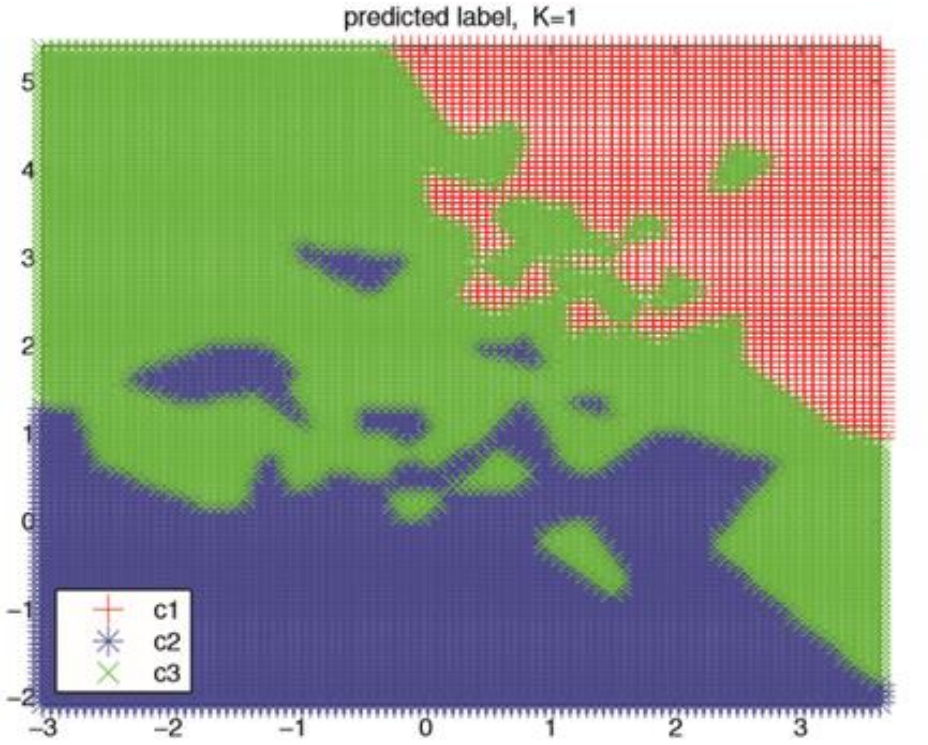


三种类型，
K=10
MAP准则，
分类区间如下图

C=1，红
C=2，绿
C=3，蓝

本例的分类边界





KNN算法简述

$$\mathbb{I}(e) = \begin{cases} 1 & \text{if } e \text{ is true} \\ 0 & \text{if } e \text{ is false} \end{cases}$$

对于给定样本，其取类c的概率

$$p(y = c|\mathbf{x}, \mathcal{D}, K) = \frac{1}{K} \sum_{i \in N_K(\mathbf{x}, \mathcal{D})} \mathbb{I}(y_i = c)$$

$N_K(\mathbf{x}, \mathcal{D})$ 是K个近邻训练样本集合

分类器输出（MAP）

$$\hat{y}(\mathbf{x}) = \text{argmax}_c \, p(y = c|\mathbf{x}, \mathcal{D})$$

KNN算法讨论

- 对于这个例子，若设K=1，则训练集分类误差为0，但分类边界复杂，泛化性能差（overfit）；
- 取K=10，则分类边界更光滑，但训练集分类误差不为0.
- K的选择是一个重要问题。在训练集误差和泛化性之间平衡。

11.训练集、测试集和验证

- 超参数：模型阶（例：多项式拟合的阶M，KNN的参数K，正则项参数λ）。一般不能通过训练过程获得，而是需要通过验证过程确定。
- 模型选择、超参数选择等
- 泛化性能测试，在测试集误差逼近泛化误差。

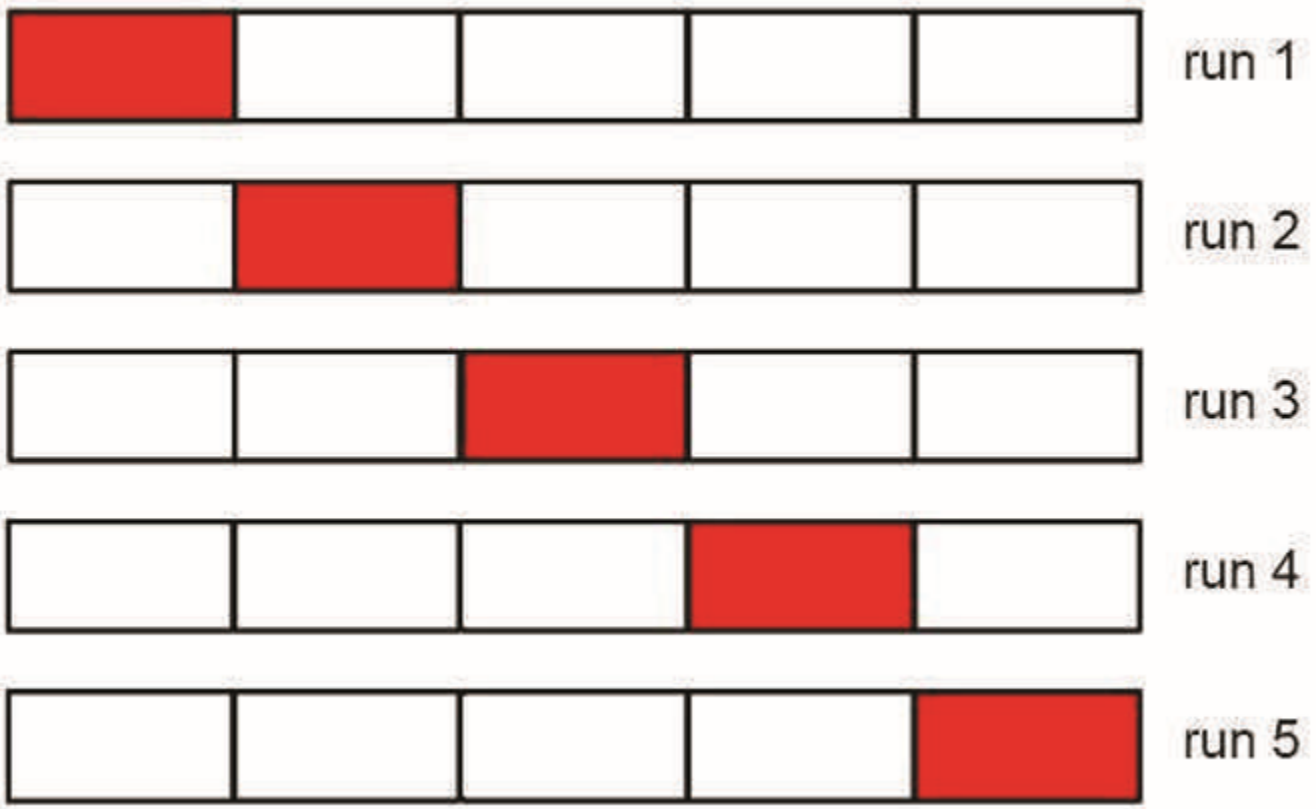
验证和测试

- 情况一：独立的训练集和测试集，两个集合独立地产生自同一个数据生成分布，训练集训练模型，通过测试误差逼近泛化误差，这是理想情况。
- 情况二：独立的训练集和测试集。将训练集分成两部分，一部分（例如80%），用于训练；另一部分（例如20%），用作确定超参数，并称为验证集（validation set），用该验证集确定超参数。超参数确定和训练获得模型参数后，再用测试集测试性能（逼近泛化误差）。

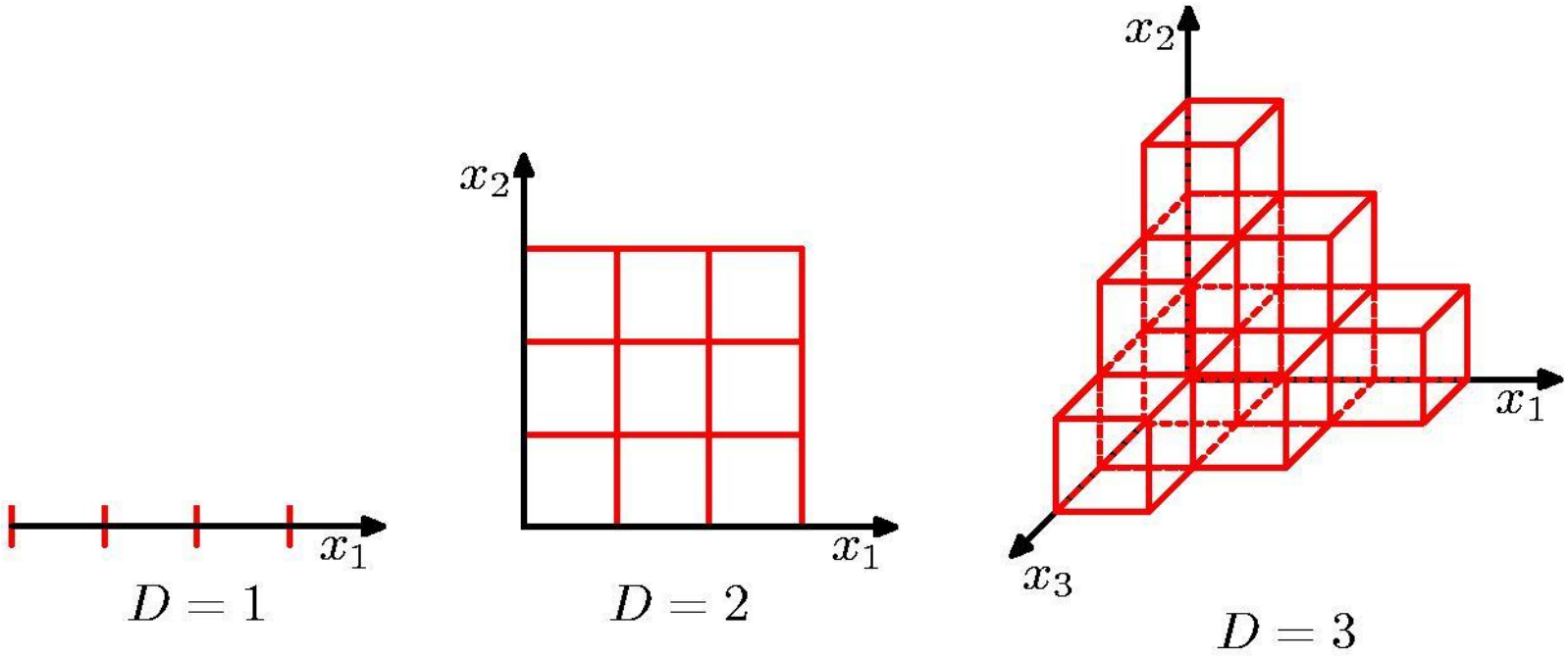
测试集的几种构成和测试方法（续）

- 情况三：交叉验证（cross validation (CV)）。数据集规模有限，将训练集分为K折（K folds），每次训练留出一折作为验证集，其余作为训练集，进行一次训练和验证，然后循环操作（见下页图）。取每次验证集的误差平均，作为验证误差。K=5的示例如下页图。
- 情况四：K=N，每次只有一个样本作为验证集，称为留一验证（leave-one out cross validation, or LOOCV）。

交叉验证示意



12.维数灾难
（The curse of dimensionality）



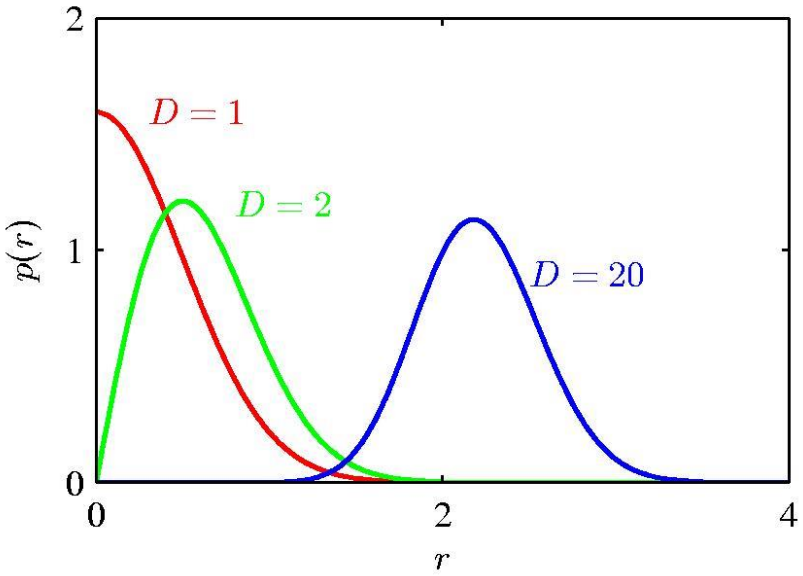
划分一维空间，需要K格，D维空间，需要 K^D

Curse of Dimensionality

Polynomial curve fitting, M = 3

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij} x_i x_j + \sum_{i=1}^{D}\sum_{j=1}^{D}\sum_{k=1}^{D} w_{ijk} x_i x_j x_k$$

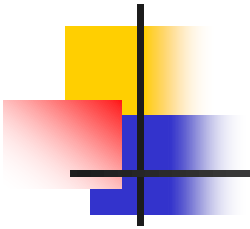
极坐标下D维幅度分布
Gaussian Densities in higher dimensions



manifold embedded

13.没有免费午餐定理
No free lunch theorem

- 对于一个特殊问题，我们可以通过交叉验证这类方法实验地选择最好的模型，然而，没有一个最好的通用模型。 (Wolpert 1996).
- 正因为如此，需要发展各种不同类型的模型以适用于现实世界的各类数据。



奥卡姆剃刀原理

解决实际问题时不是选择越先进、越复杂的算法越好;

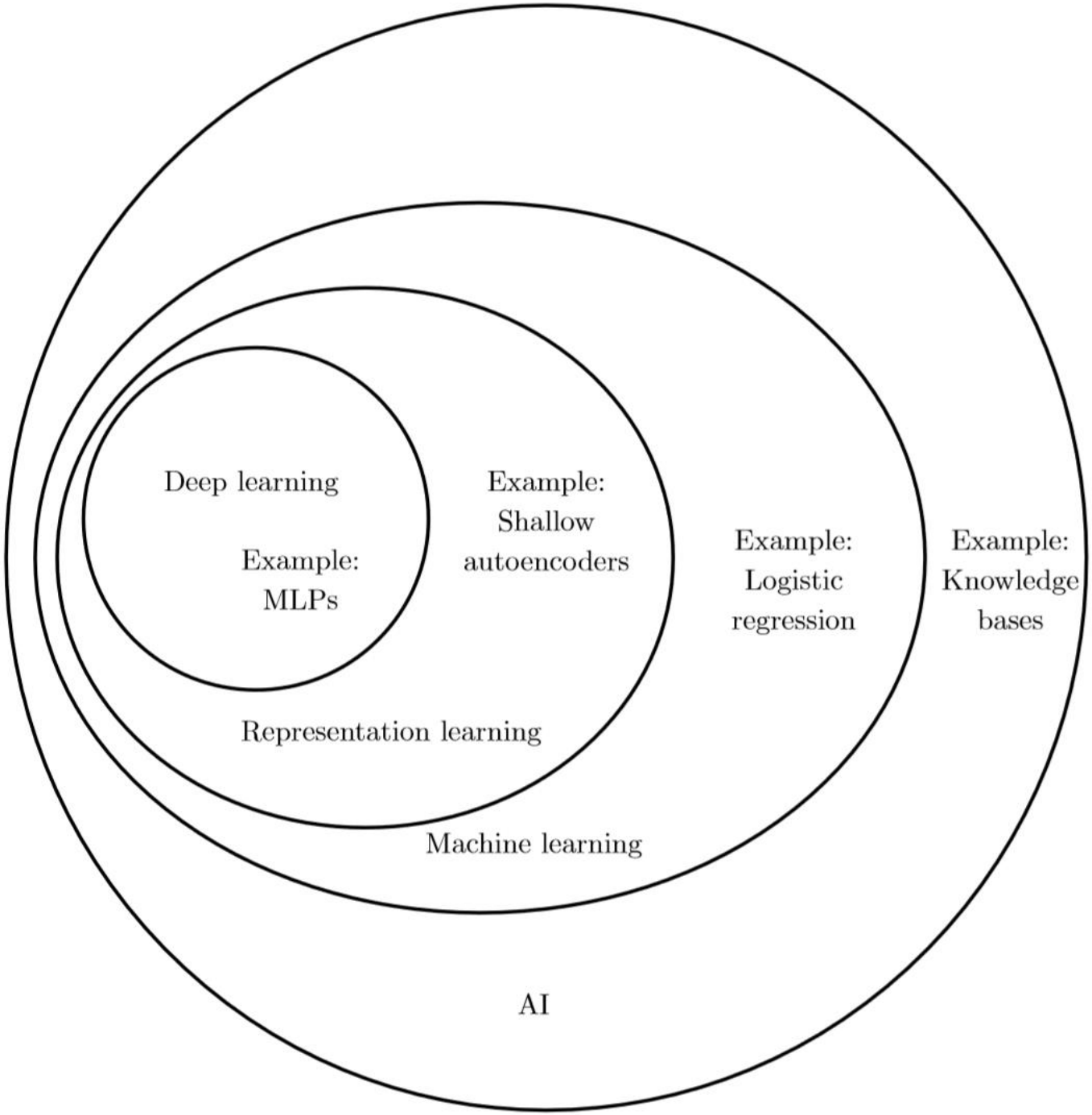
Occam剃刀原理：该原理叙述为：除非必要，“实体”不应该随便增加;
或：设计者不应该选用比“必要”更加复杂的系统。

这个问题也可表示为方法的“适宜性”，即在解决一个实际问题中，选择最适宜的算法。

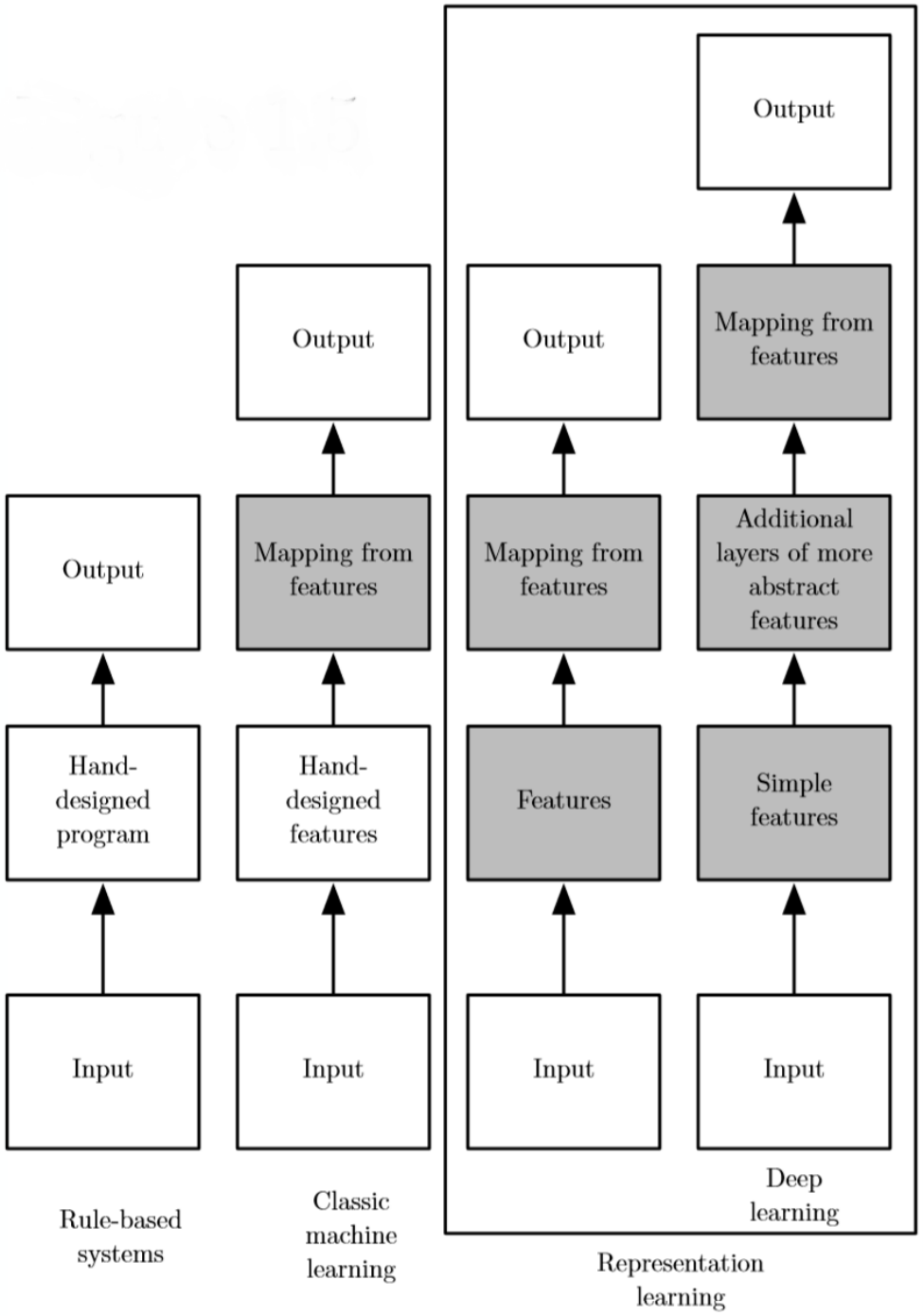
14.深度学习（Deep Learning）

- 起始约2006年，神经网络（Neural Network：NN）方法的第三次复兴；
- 主要结构MLP，CNN，RNN等是传统NN和ML中已存在的；新结构：GAN、Transformer等。
- DL主要依靠大规模训练数据、大型计算集群、和改进的优化方法和专门的训练方法。
- 尽管目前是最活跃的分支，但深度学习是机器学习的一种。目的之一是改善传统ML的泛化能力。

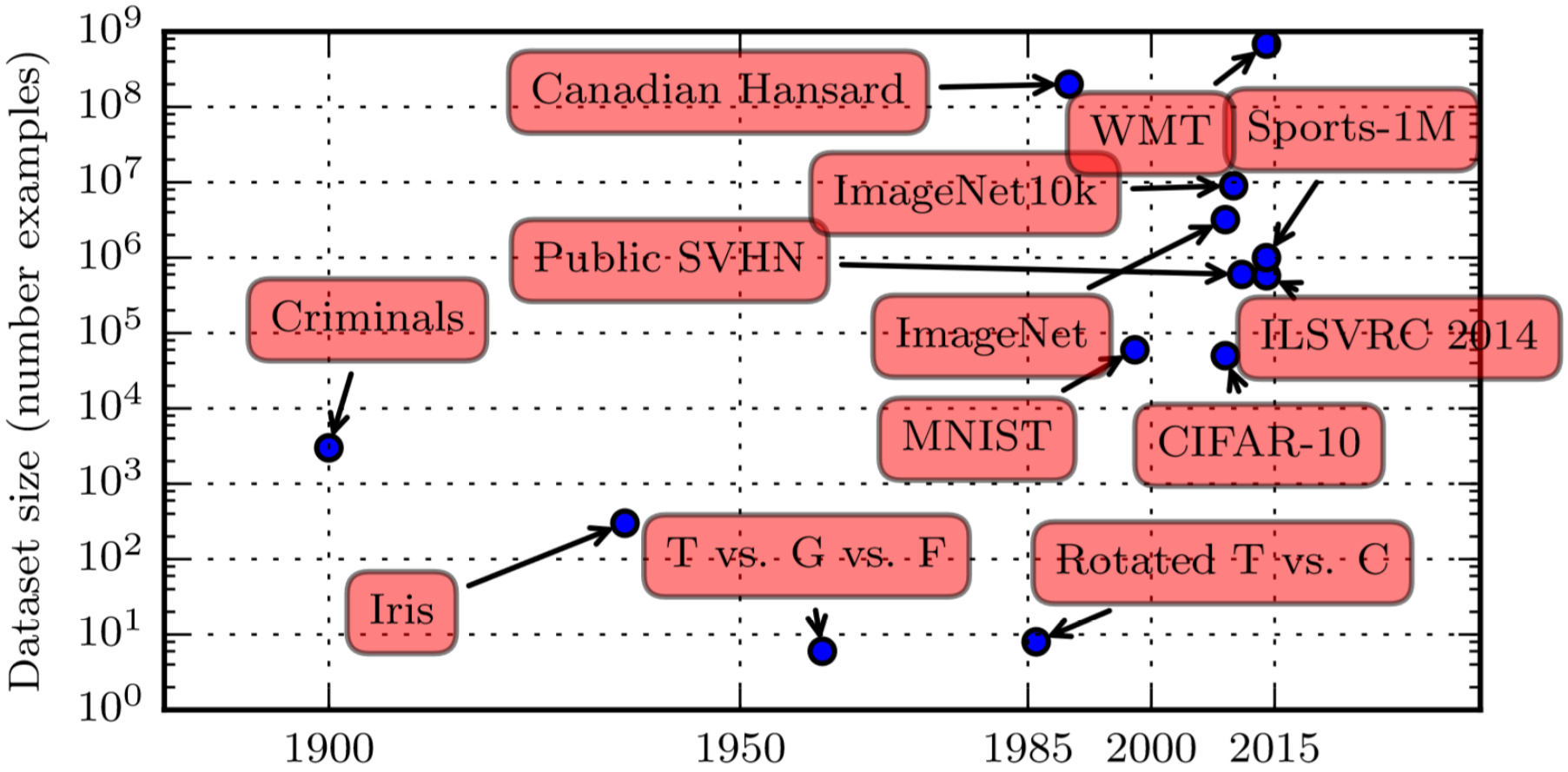
深度学习是机器学习的一种



经典方法
机器学习
深度学习
之间关联

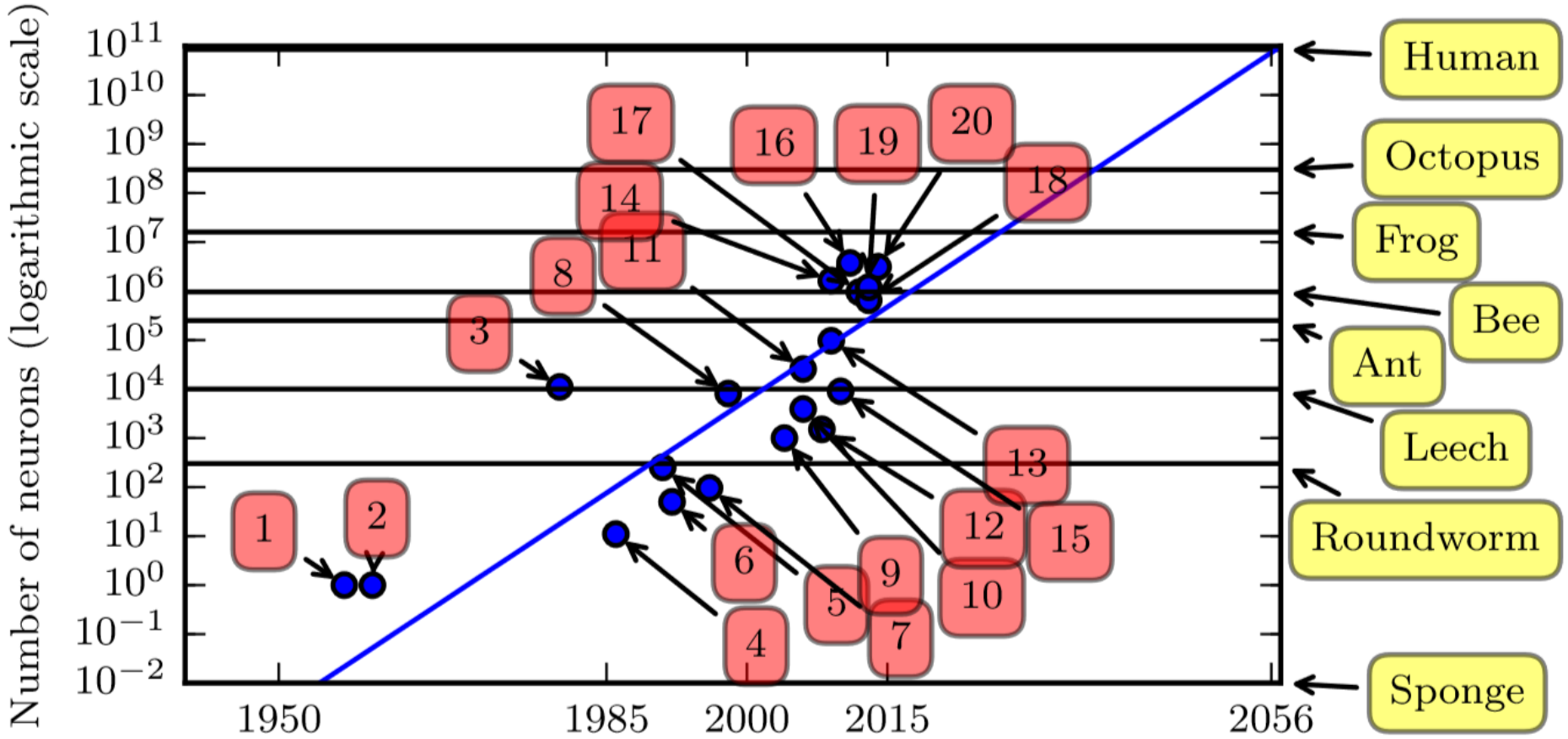


数据集的增长规模



大的数据集是解决问题的一个方法，
很多应用没有大的数据集怎么办?

深度学习系统神经元增长趋势



规模增长是解决问题的一个方法，
所有问题都能靠规模增长解决吗?

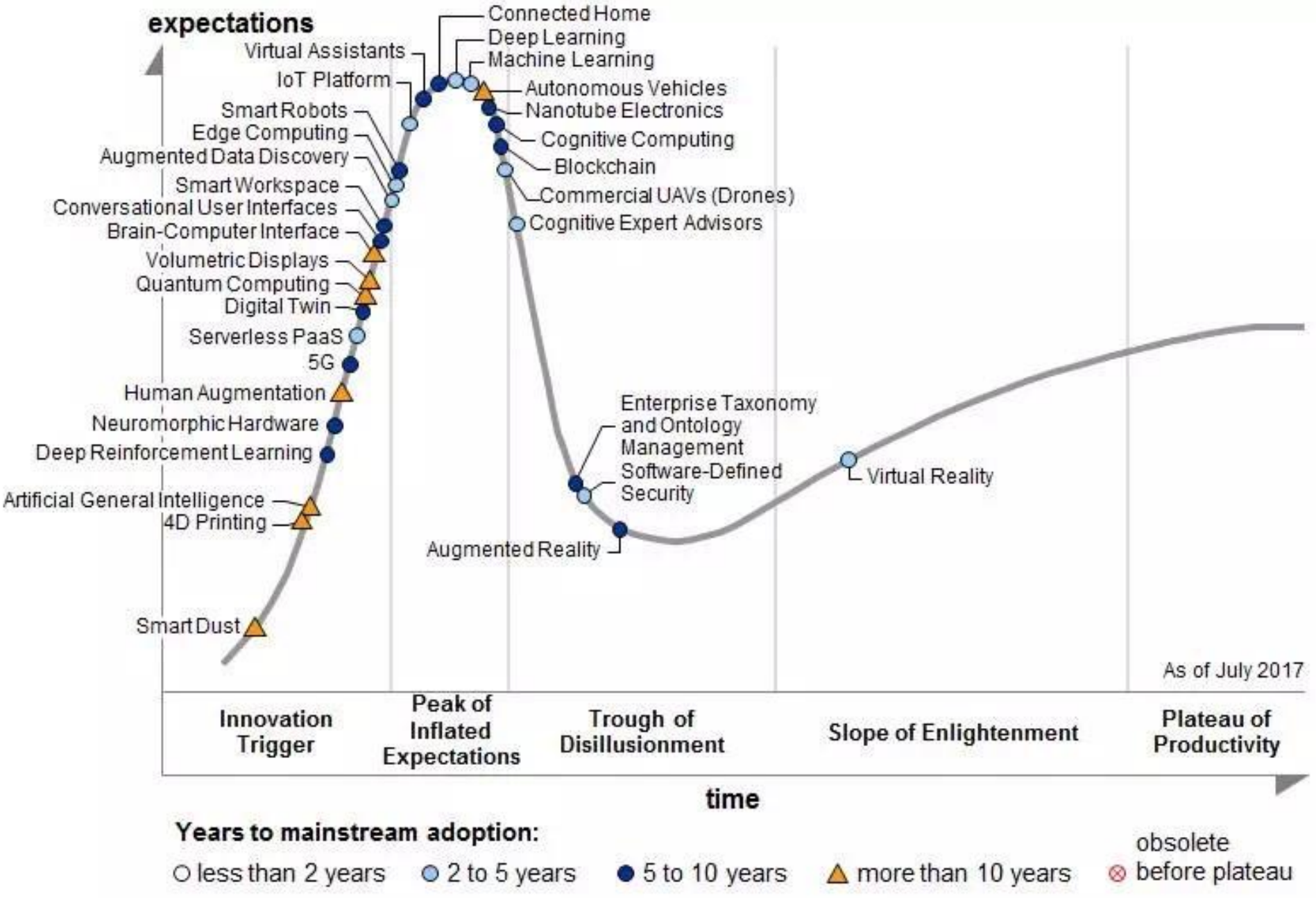
15.一些应用领域

- 图像分类和识别、计算机视觉
- 语音识别
- 自然语言处理
- 推荐系统
- 网络搜索引擎
- 通信信道建模
- 通信、雷达等的信号分类和识别
- 智能机器人
- 无人驾驶汽车
- 无人机自主导航
- 各种专业领域的特殊应用，等等

16. 机器学习的简短历史

- 图灵：1950年发表“计算机器和智能”（Computing Machinery and Intelligence)，图灵测试、机器学习、遗传算法、强化学习；
- 1956年夏天达特茅斯会议（J. McCarthy, M. Minsky, C. Shannon, N. Rochester)
- McCullon-Pitts神经元（1943）；Minsky于1951年开发第一个硬件神经网；F. Rosenblatt（1957）提出感知机，Widrow（1960）提出Adaline（LMS，随机梯度）
- 1980年夏，卡内基-梅隆大学第一届机器学习研讨会
- 1980-90年代，神经网络复兴，BP算法
- 1990年代-2000年代，统计学习成为主流，SVM、核方法、图模型决策树、集成学习等
- 2006年—至今，神经网络再次复兴，深度学习
- 强化学习（RL）也一直持续进展，80年代TD算法，Q学习，直到近期DRL（核心思想MDP，POMDP）

技术创新的预期图


expectations
Connected Home
Deep Learning
Machine Learning
Virtual Assistants
IoT Platform
Autonomous Vehicles
Nanotube Electronics
Smart Robots
Edge Computing
Cognitive Computing
Augmented Data Discovery
Blockchain
Smart Workspace
Commercial UAVs (Drones)
Conversational User Interfaces
Cognitive Expert Advisors
Brain-Computer Interface
Volumetric Displays
Quantum Computing
Digital Twin
Serverless PaaS
5G
Human Augmentation
Enterprise Taxonomy and Ontology Management
Neuromorphic Hardware
Deep Reinforcement Learning
Software-Defined Security
Virtual Reality
Artificial General Intelligence
4D Printing
Augmented Reality
Smart Dust
As of July 2017
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time
Years to mainstream adoption:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
obsolete before plateau

本章思考题

- 什么是机器学习？（或怎样理解机器学习？）
- 机器学习的类型？
- 构造机器学习系统的基本元素？
- 模型的类型（参数模型、非参数模型等）
- 怎样理解深度学习。
- 几个主要名词（过拟合、泛化、正则性等）
- 模型的参数和超参数？超参数的作用，怎样确定超参数。
- 什么是交叉验证？