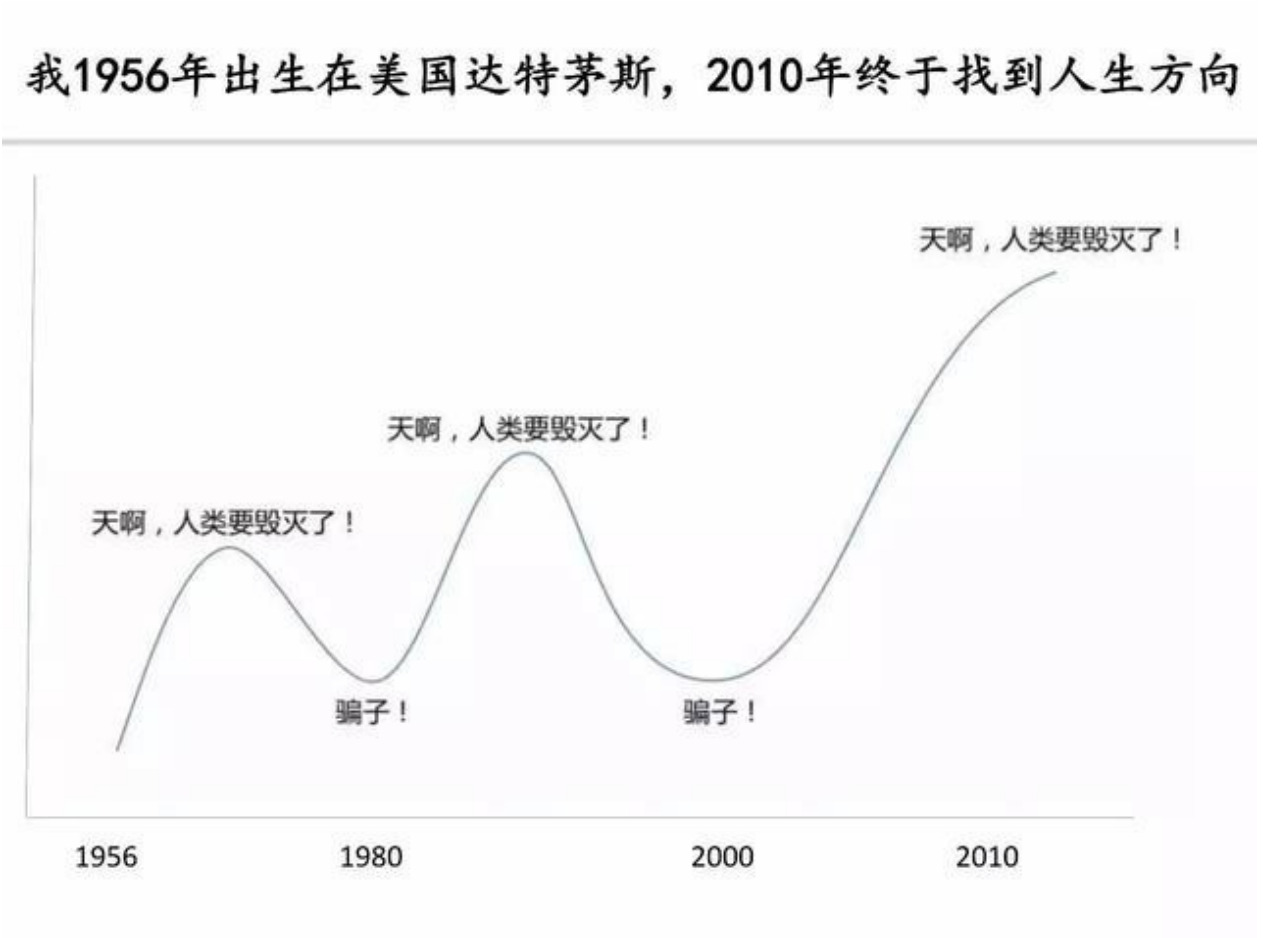
机器学习基础

赵永红

四川师范大学物理系

CSRC-2017

人和人工智能


我1956年出生在美国达特茅斯，2010年终于找到人生方向
天啊，人类要毁灭了！
天啊，人类要毁灭了！
天啊，人类要毁灭了！
骗子！
骗子！
1956
1980
2000
2010

狼来了……

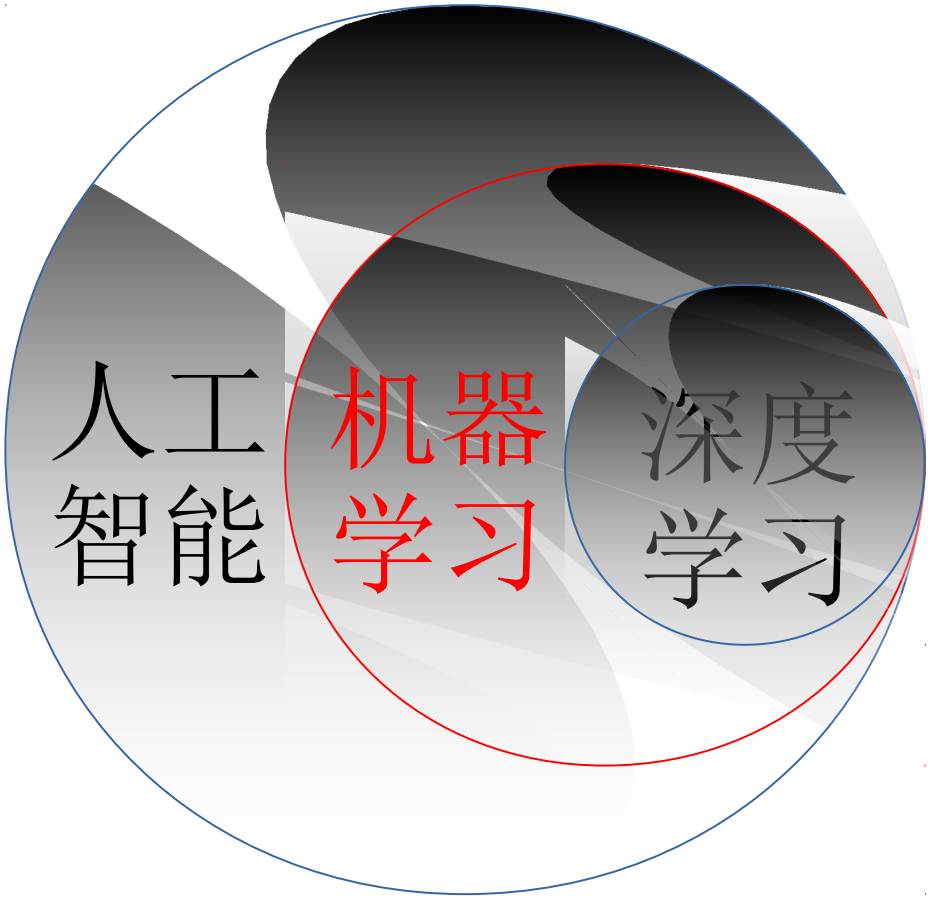
李开复

但是最近，你们吹牛吹得太过火了！

- 无人驾驶
  - 以为计算机视觉就足够做到无人驾驶吗？
  - 要多完美的团队和多少钱才能烧一个完整L4无人驾驶啊？
  - 天价估值吓死宝宝了
- 家用机器人就更别提了，尤其是有手有脚有眼睛有耳朵那种
- 语音识别和语音理解
  - Speech-to-text 不是 speech-to-meaning
  - 不如多看看处理噪音的问题
- 自然语言理解（NLU）
  - 我一次只懂一个领域，跨领域NLU根本不存在！
  - 不！要！问！我！为！什！么！
- 拿不到医疗数据，做个P的AI？
- 人脸识别是中国特色应用
  - 不过四个独角兽啊？太夸张了吧？
- 人工智能平台？至少还要三年！

困难

1. Gradual learning
2. Unsupervised learning
3. Strong generalization
4. Category learning from few examples
5. Learning to learn
6. Compositional learning
7. Learning without forgetting
8. Transfer learning
9. Knowing when you don’t know
10. Learning through action



机器如何学习？ → 人类如何学习？

机器学习？

AI Magazine Volume 18 Number 3 (1997) (© AAAI)

Does Machine Learning Really Work?

Tom M. Mitchell

The Niche for Machine Learning:1997

数据挖掘：从过去的数据预言未来；

处理难以手动编程的事情；

新闻过滤： TF-IDF ；

NewsWeeder: Learning to Filter Netnews

(To appear in ML 95)

Ken Lang
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
akl+@cs.cmu.edu

为什么需要“学习”？

程序员是否万能？

人脸识别、物体识别和语言理解均难以通过手动编程实现；

机器学习程序：可以从过去的数据（经验）中进行学习；

什么时候需要机器学习：图像识别、语音识别、垃圾邮件识别、搜索排名（ google 比百度领先 10 年，在互联网领域意味着什么？）；

“学习”和统计

相同点：揭示数据的规律；依赖于相同的数学工具（微积分、概率论、线性代数和其它算法）

不同点：

统计	给出结论以支持决策	重在解释
机器学习	给出决策过程	重在预言

给机器学习分类

监督学习： data ——> label

无监督学习： data——> pattern

强化学习： data——> maximize long term reward，希望之星！

学习算法

- 学习：完成某项任务（ Task ）的能力（ Performance ）随着经验（ Experience ）而增加；

- 学习算法是什么？如何设计？

固定程序机器人

If 左边有个 10cm 宽的坑； then

左脚向左跨出 10cm ；

右脚向左跨出 10cm ；

else if ……

……

……

end if

任务

- 固定代码难以完成；
- 机器人： Learn to walk VS Program to walk ？
- 学习任务：处理样本的能力；
- 样本：

	特征 1 x_1	特征 2 x_2	特征 3 x_3
样本 1 X1			
样本 2 X2			

任务种类

分类：几何图形识别、人脸识别、空间群预测 [1]；

$$f : \mathbb{R}^n \to \{1, \ldots, k\}$$

回归：预测一个数值；股票价格；

$$f : \mathbb{R}^n \to \mathbb{R}$$

转录： OCR 、语音识别；

翻译： sequence to sequence ；

结构化信息提取：语法分析、图片描述；

……

[1]. Nat. Comm., 8, 14284 (2017);

算法性能 Performance

0-1 loss：结果正确为 1，结果错误为 0；分类、转录等任务；

对数概率：计算密度分布函数类任务；

主要困难：

1、损失函数需要根据具体问题来定义；

2、损失函数难以度量；

经验 Experience 和数据集

经验：学习（训练）所使用的数据集；

	未标注	已标注
数据集	无监督学习	监督学习

学习函

联合概率：$p(\mathbf{x}) = \prod_{i=1}^{n} p(\mathrm{x}_i \mid \mathrm{x}_1, \ldots, \mathrm{x}_{i-1})$

（无监督学习可以用监督学习展开）

半监督学习（部分标注）、强化学习（环境依赖）；

人类 vs. 机器

	人类	机器
可同时处理的数据集数目	多	单
数据格式	色彩、形状、美丑	数字 一切都需要数字化 异构数据可以放在一起学习
输出结果精度	模糊	精确

智文

word2vec 完成了文字数字化。

如何处理异构数据（例如，美的判断），是热点之一。

学习算法：线性回归

任务： $\hat{y} = \boldsymbol{w}^\top \boldsymbol{x}$ 模型（参数）： w

控制每个分量 x_i 对结果 y 的贡献

性能： $\mathrm{MSE}_{\mathrm{test}} = \frac{1}{m} \sum_i (\hat{\boldsymbol{y}}^{(\mathrm{test})} - \boldsymbol{y}^{(\mathrm{test})})_i^2$

学习的是训练集，而性能是有测试集来衡量。

算法： $\nabla_{\boldsymbol{w}} \mathrm{MSE}_{\mathrm{train}} = 0$

$$\boldsymbol{w} = \left(\boldsymbol{X}^{(\mathrm{train})\top} \boldsymbol{X}^{(\mathrm{train})} \right)^{-1} \boldsymbol{X}^{(\mathrm{train})\top} \boldsymbol{y}^{(\mathrm{train})}$$

推广： $\hat{y} = \boldsymbol{w}^\top \boldsymbol{x} + b$

想要测试误差最小，但是只有训练数据，因此首先要求训练误差“最”小。

数据集与误差

按照功能可分为：

训练 (train) 集，用来训练模型；

$$\frac{1}{m^{(\text{train})}}||\boldsymbol{X}^{(\text{train})}\boldsymbol{w}-\boldsymbol{y}^{(\text{train})}||_2^2$$

测试 (test) 集，用来测试模型性能；

$$\text{MSE}_{\text{test}}=\frac{1}{m}\sum_i(\hat{\boldsymbol{y}}^{(\text{test})}-\boldsymbol{y}^{(\text{test})})_i^2$$

验证 (validation) 集，用来确定超参数；

超参数这个概念广泛存在，就如 DFT，优化的是电荷密度，而 k 点和 ecut 就是超参数。

实际应用

推广误差：过拟合、欠拟合

成立条件：训练样本和测试样本均满足相同的独立全同分布。

分两步最小化测试误差：

1. 最小化训练误差；否则 → 欠拟合

2. 最小化训练误差和测试误差的差值；否则 → 过拟合

（测试误差一般大于训练误差）

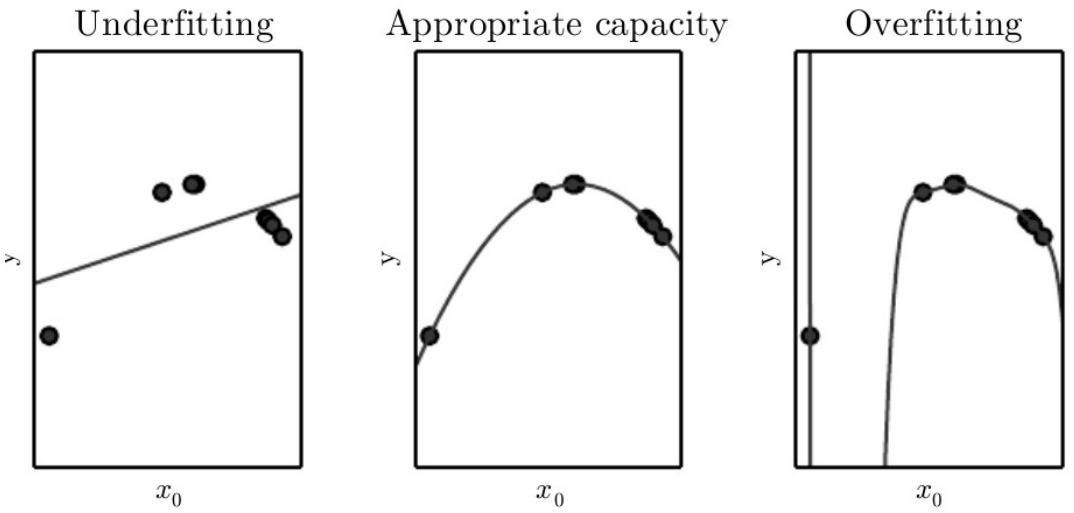
如果所有样本均完全随机独立，则无法构建学习算法。

模型的容量 capacity

$$\hat{y} = b + wx$$

$$\hat{y} = b + w_1 x + w_2 x^2$$

y 相对 w 依然为线性



类似基函数的选择，例如平面波截断。

如何选择容量？

Occam 剪刀原理：在能够描述观测现象的所有模型中，应选取尽可能简洁的那一个。

VC(Vapnik-Chervonenkis) 尺度 [1]：一个二分类器能够有效正确分类的样点的个数的最大值。

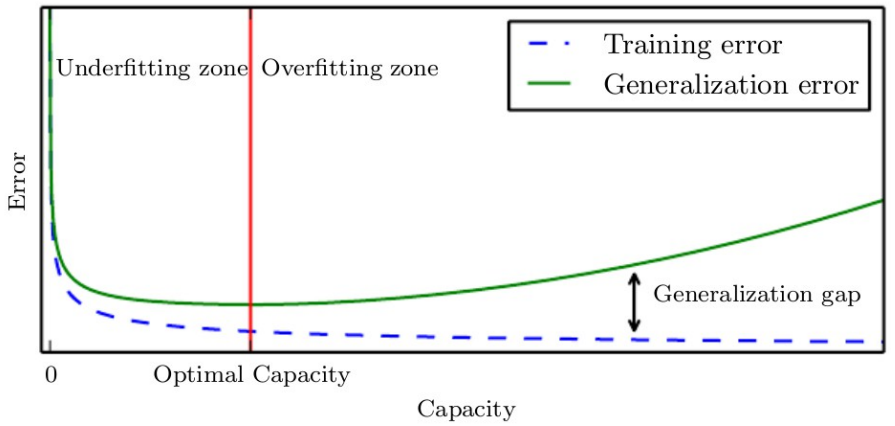
深度学习模型的容量难以定义：

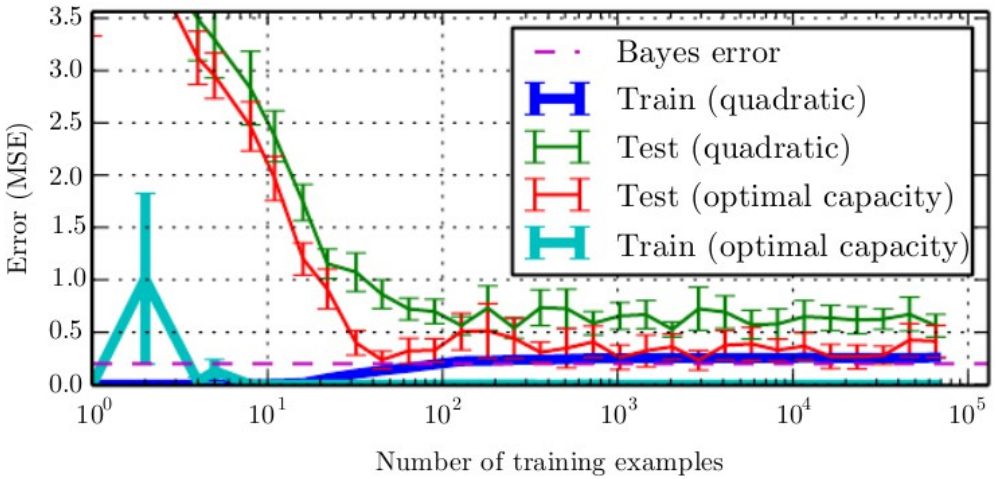
1、优化算法对结果的影响和模型本身的容量同等重要；

2、算法的性能跟优化算法相关，但是缺乏解决非凸优化问题的数学手段；

[1]、 Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative
frequencies of events to their probabilities. Theory of Probability and Its Applications, 16, 264–280. 114

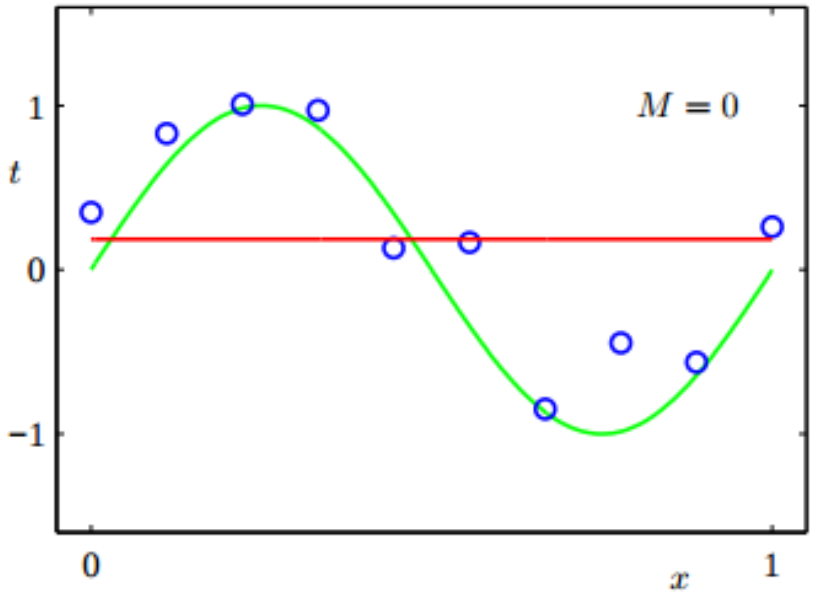
典型的训练误差和测试误差

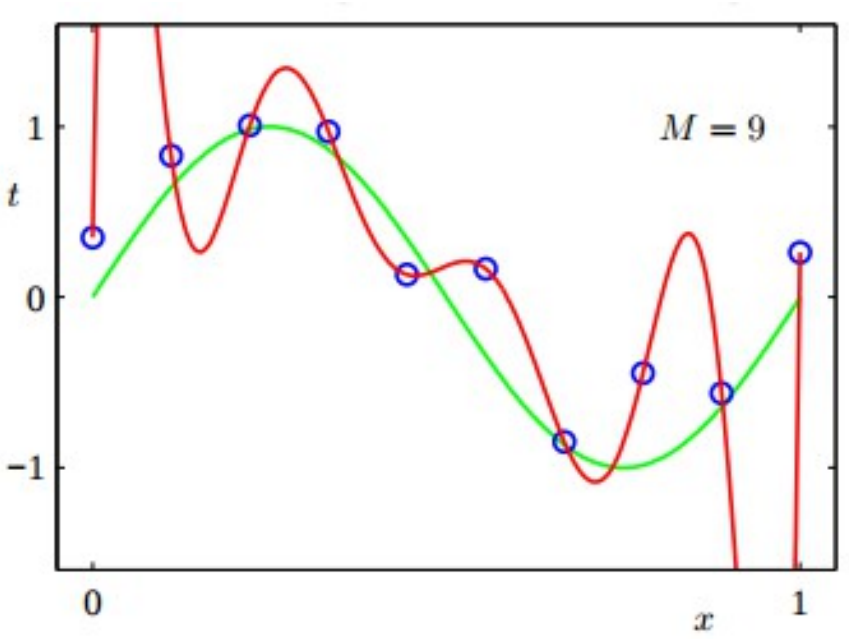




减小推广误差的两种方法：增加容量、增加训练样本数

不是单纯减少训练误差！


M = 0
1
t
0
−1
0
x
1


M = 9
1
t
0
−1
0
x
1

数据集有限  非逻辑上严格对所有样本正确；

如果对所有数据产生的概率分布函数平均，则任何分类方法都没有区别。

“没有免费的午餐”定理

没有最优算法，只有最适合当前数据集的算法

通用智能很困难

Wolpert, D. H. (1996). The lack of a priori distinction between learning algorithms. Neural Computation, 8 (7), 1341–1390. 116

正则化

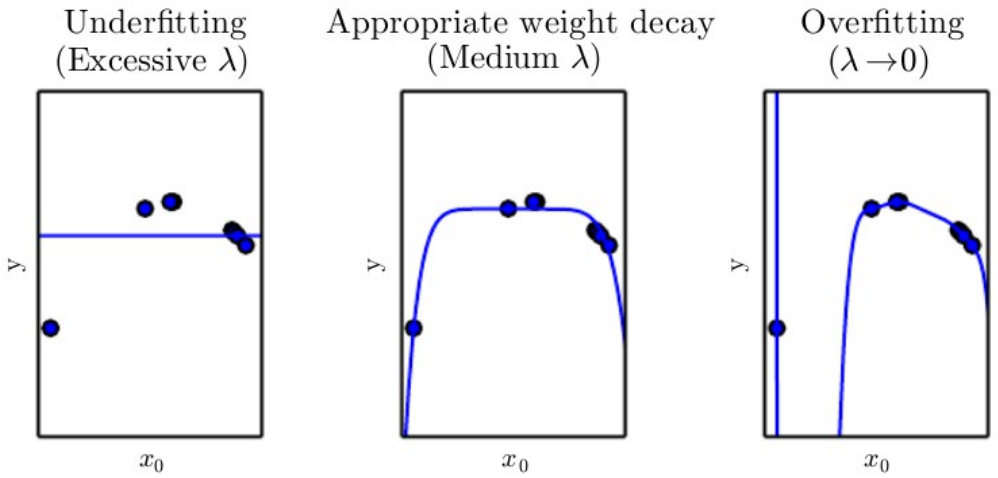
修正项：降低广义误差，不改变训练误差；

根据具体问题选择（边界条件往往是物理意义之所在）

$$J(\boldsymbol{w}) = \mathrm{MSE}_{\mathrm{train}} + \lambda \boldsymbol{w}^{\top}\boldsymbol{w}$$

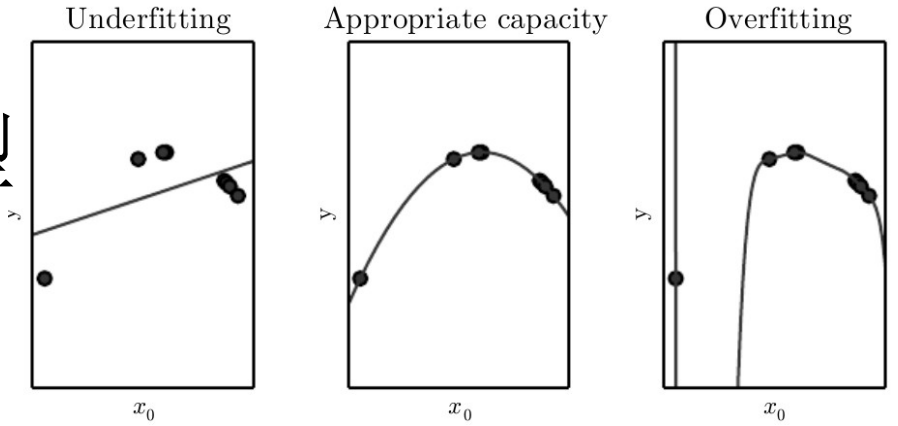
控制相对大小，可以有多项。

与 Lagrange multiplier 类似， J(w) 是由约束条件描述的系统的拉格朗日量。



九阶多项式
不同的正则
化系数

不同的模型
容量



超参数、验证数据集

超参数：用来定义模型的种类，比如神经网络的层数、神经元个数、惩罚因子等；

验证数据集：用来确定超参数的数据集；

嵌套模型：先确定超参数，再通过训练学习出模型里的参数；

交叉验证：数据集较小的时候，可以随机多次抽取验证集，剩下的作为训练集；

增加随机性可弥补样本空间大小的不足；

机器学习算法可以看作是一种基于统计学的复杂函数计算方法；

训练集、测试集和验证集之间的独立全同性很重要；

基本概念

评价器（ estimator ）：用来估算某个参量数值的统计函数；

偏差（ bias ）：期望值与真实值之差；

方差（ variance ）：$Var(\theta) = E(\theta - E(\theta))^2$

用来衡量估算值在全体样品上的整体性偏差；

标准差（ standard error ）：方差的平方根；

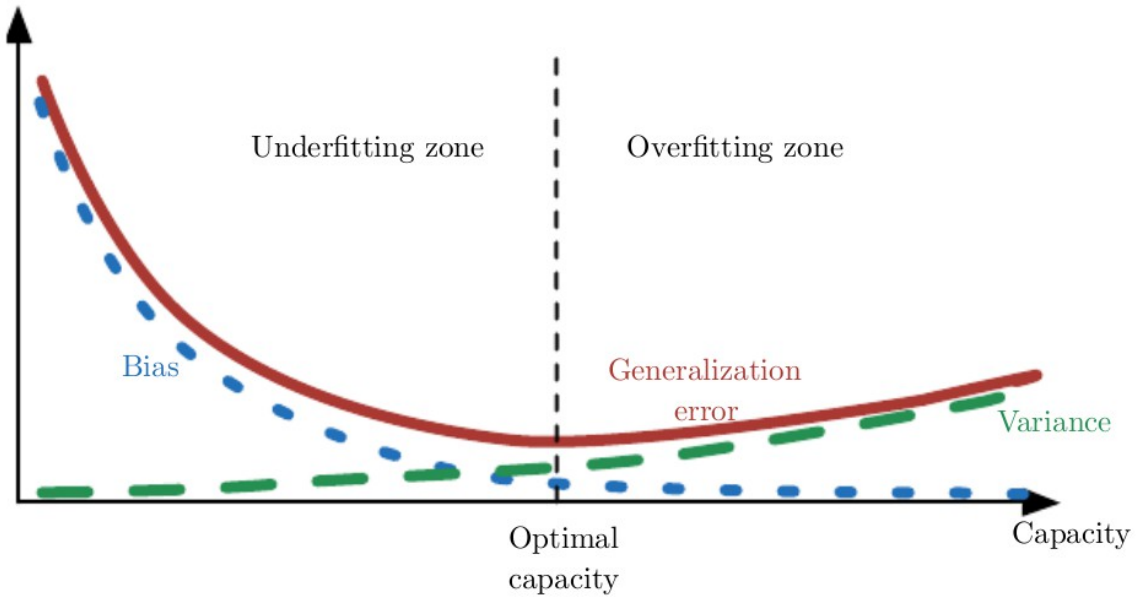
偏差

$$\mathrm{bias}(\hat{\boldsymbol{\theta}}_m) = \mathbb{E}(\hat{\boldsymbol{\theta}}_m) - \boldsymbol{\theta}$$

无偏差： $\mathbb{E}(\hat{\boldsymbol{\theta}}_m) = \boldsymbol{\theta}$.

渐进偏差： $\lim_{m\to\infty} \mathbb{E}(\hat{\boldsymbol{\theta}}_m) = \boldsymbol{\theta}$

方差和标准差



$$\mathrm{SE}(\hat{\mu}_m) = \sqrt{\mathrm{Var}[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}]} = \frac{\sigma}{\sqrt{m}}$$

平衡：偏差和方差

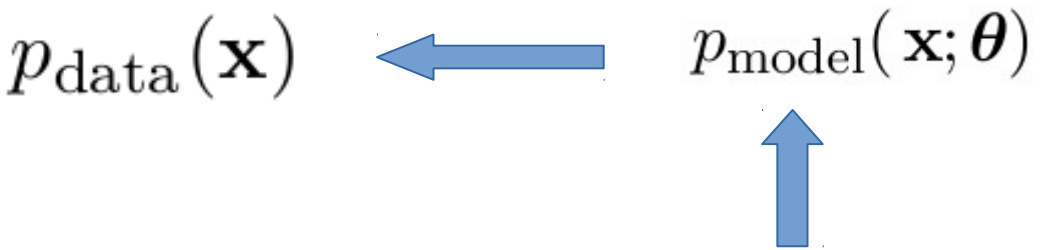
交叉验证；

均方差（ MSE ）：

$$\begin{aligned}\mathrm{MSE} &= \mathbb{E}[(\hat{\theta}_m - \theta)^2] \\ &= \mathrm{Bias}(\hat{\theta}_m)^2 + \mathrm{Var}(\hat{\theta}_m)\end{aligned}$$

统计学的意义： MSE 是机器学习中最常用的损失函数；

最大似然估计



$$\boldsymbol{\theta}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} p_{\text{model}}(\mathbb{X};\boldsymbol{\theta})$$

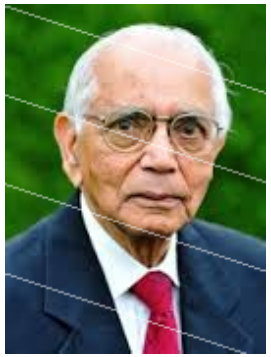
$$= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{m} p_{\text{model}}(\boldsymbol{x}^{(i)};\boldsymbol{\theta})$$

最大似然：使得模型概率分布尽可能符合实际的概率分布。

条件最大似然：$\boldsymbol{\theta}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{Y} \mid \boldsymbol{X};\boldsymbol{\theta})$

最大似然的性质

渐进最优：收敛性随样本数而增加；



条件 1：真实分布必须对应于某一个模型分布 $p_{\text{model}}(\cdot;\boldsymbol{\theta})$

条件 2：真实分布必须对应于唯一模型分布；

对大样本数目极限，最大似然统计给出最小

MSE；

Rao, C. (1945). Information and the accuracy attainable in the estimation of statistical parameters. Bulletin of the Calcutta Mathematical Society, 37 , 81–89. 135 , 295

"a living legend whose work has influenced not just statistics, but has had far reaching implications for fields as varied as economics, genetics, anthropology, geology, national planning, demography, biometry, and medicine."

概率是什么？

频率学派：概率是“长期性”的频率，例如投掷硬币次数越多，单面朝上概率趋于0.5；（唯物主义）

贝叶斯学派：跟过去经验有关，例如硬币被做了手脚而投掷的不知道；（唯心主义）

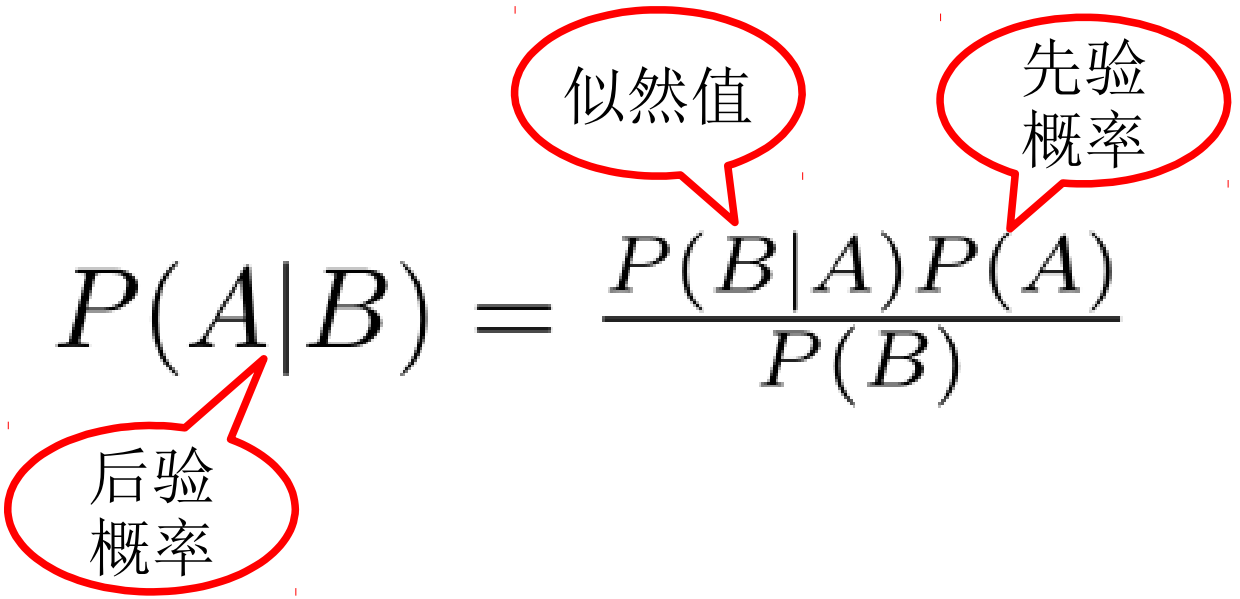
贝叶斯：近乎神？

Nathaniel Read Silver (born January 13, 1978) is an American statistician and writer who analyzes baseball (see sabermetrics) and elections (see psephology). He is the editor-in-chief of ESPN's *FiveThirtyEight* and a Special Correspondent for ABC News. Silver first gained public recognition for developing PECOTA,[3] a system for forecasting the performance and career development of Major League Baseball players, which he sold to and then managed for Baseball Prospectus from 2003 to 2009.[4]

这日，曾留下武林秘籍 All of Statistics: A Concise Course in Statistical Inference 的大侠 Larry Wasserman (Department of Statistics, Department of Machine Learning, Carnegie Mellon University) 突发感想，在自己博客 Normal Deviate 中写下大号书评：

Nate Silver is a Frequentist: Review of “the signal and the noise”

贝叶斯法则



贝叶斯鸡汤：

What gets us into trouble is not what we don't know. It's what we know for sure that just ain't so.

贝叶斯统计

	频率论统计	贝叶斯统计
概率函数	单个参数	所有参数
θ	确定	随机
单个样本取值	随机	确定

算命都得懂点贝叶斯统计。

贝叶斯模型适用于训练数据集较小的情况。

最大似然和贝叶斯推理

区别 1 ：从全体参数预测

$$p(\boldsymbol{\theta} \mid x^{(1)}, \ldots, x^{(m)}) = \frac{p(x^{(1)}, \ldots, x^{(m)} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(x^{(1)}, \ldots, x^{(m)})}$$

$$p(x^{(m+1)} \mid x^{(1)}, \ldots, x^{(m)}) = \int p(x^{(m+1)} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid x^{(1)}, \ldots, x^{(m)}) \, d\boldsymbol{\theta}$$

区别 2 ：如何看待过去的经验

善知过去未来，争议很大！

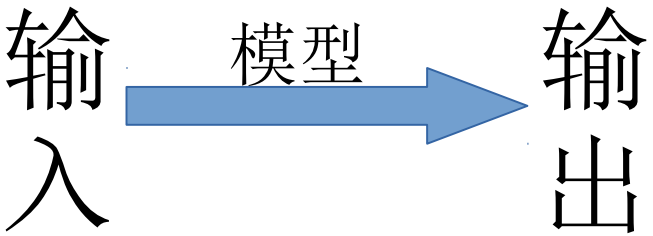
最大后验估计
Maximum a posteriori (MAP)

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log P(\boldsymbol{y}^{(i)} \mid \boldsymbol{x}^{(i)}; \boldsymbol{\theta})$$

$$\boldsymbol{\theta}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \boldsymbol{x}) = \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{x} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$$

监督（ supervised ）学习

已知　　未知　　已知



$$p(y \mid \boldsymbol{x})$$

$$p(y \mid \boldsymbol{x}; \boldsymbol{\theta})$$

线性回归：模型定义

模型：

$$y = wx + b$$

损失函数：

$$\mathcal{L}(y, t) = \frac{1}{2}(y - t)^2$$

价格函数：

$$\mathcal{E}(w, b) = \frac{1}{2N}\sum_{i=1}^{N}\left(y^{(i)} - t^{(i)}\right)^2$$

$$= \frac{1}{2N}\sum_{i=1}^{N}\left(wx^{(i)} + b - t^{(i)}\right)^2$$

(损失函数的算术平均)

选择合适的 w 和 b，最小化价格函数。

$$\frac{\partial \mathcal{E}}{\partial w_j} = \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)}(y^{(i)} - t^{(i)})$$

$$\frac{\partial \mathcal{E}}{\partial b} = \frac{1}{N}\sum_{i=1}^{N} y^{(i)} - t^{(i)}.$$

线性回归：直接求解

$$\frac{\partial \mathcal{E}}{\partial w_j} = \frac{1}{N}\sum_{j'=1}^{D}\left(\sum_{i=1}^{N} x_j^{(i)} x_{j'}^{(i)}\right) w_{j'} - \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)} t^{(i)} = 0$$

$$\mathbf{w} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{t}$$

关于权重因子 w 的线性方程组。

线性回归：梯度下降

$$w_j \leftarrow w_j - \alpha \frac{\partial \mathcal{E}}{\partial w_j}$$

$$w_j \leftarrow w_j - \alpha \frac{1}{N} \sum_{i=1}^{N} x_j (y^{(i)} - t^{(i)})$$

$$\mathbf{w} \leftarrow \mathbf{w} - \frac{\alpha}{N} \mathbf{X}^\top (\mathbf{y} - \mathbf{t})$$

线性分类器： NOT， AND

$$z = \mathbf{w}^T \mathbf{x} + b$$

$$y = \begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases}$$

x_1	t
0	*1*
1	*0*

$w_1 = -2,\ b = 1$ NOT

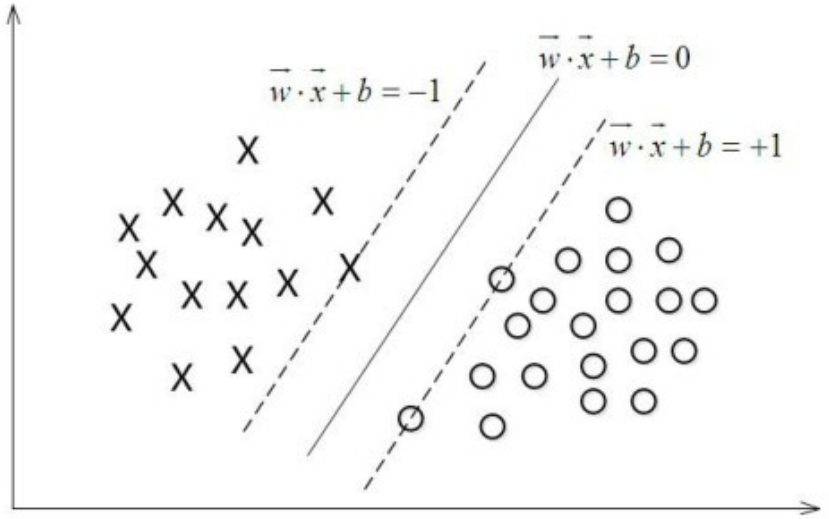
x_1	x_2	t
0	*0*	*0*
0	*1*	*0*
1	*0*	*0*
1	*1*	*1*

$$b < 0$$
$$w_2 + b < 0$$
$$w_1 + b < 0$$
$$w_1 + w_2 + b > 0$$

$b = -1.5, w_1 = w_2 = 1$ AND

逻辑回归：线性分类器

处理二分类问题。



二分类器：线性可分

$$p(y=1 \mid \boldsymbol{x};\boldsymbol{\theta}) = \sigma(\boldsymbol{\theta}^\top \boldsymbol{x})$$

特征映射：从线性到非线性

$$\boldsymbol{\phi}(x) = \begin{pmatrix} 1 \\ x \\ x^2 \\ x^3 \end{pmatrix}$$

$$y = \mathbf{w}^\top \boldsymbol{\phi}(x)$$

非线性：异或

$$\begin{bmatrix} 0 & 0 \\ 1 & 1 \\ 1 & 1 \\ 2 & 2 \end{bmatrix} \rightarrow \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \quad ?$$

异或：表示（象）学习

$$\phi_1(\mathbf{x}) = x_1$$
$$\phi_2(\mathbf{x}) = x_2$$
$$\phi_3(\mathbf{x}) = x_1 x_2$$

$\phi_1(\mathbf{x})$	$\phi_2(\mathbf{x})$	$\phi_3(\mathbf{x})$	t
0	0	0	0
0	1	0	1
1	0	0	1
1	1	1	0

$$z = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) + b$$

$$b = -0.5 \qquad w_1 = 1 \qquad w_2 = 1 \qquad w_3 = -2$$

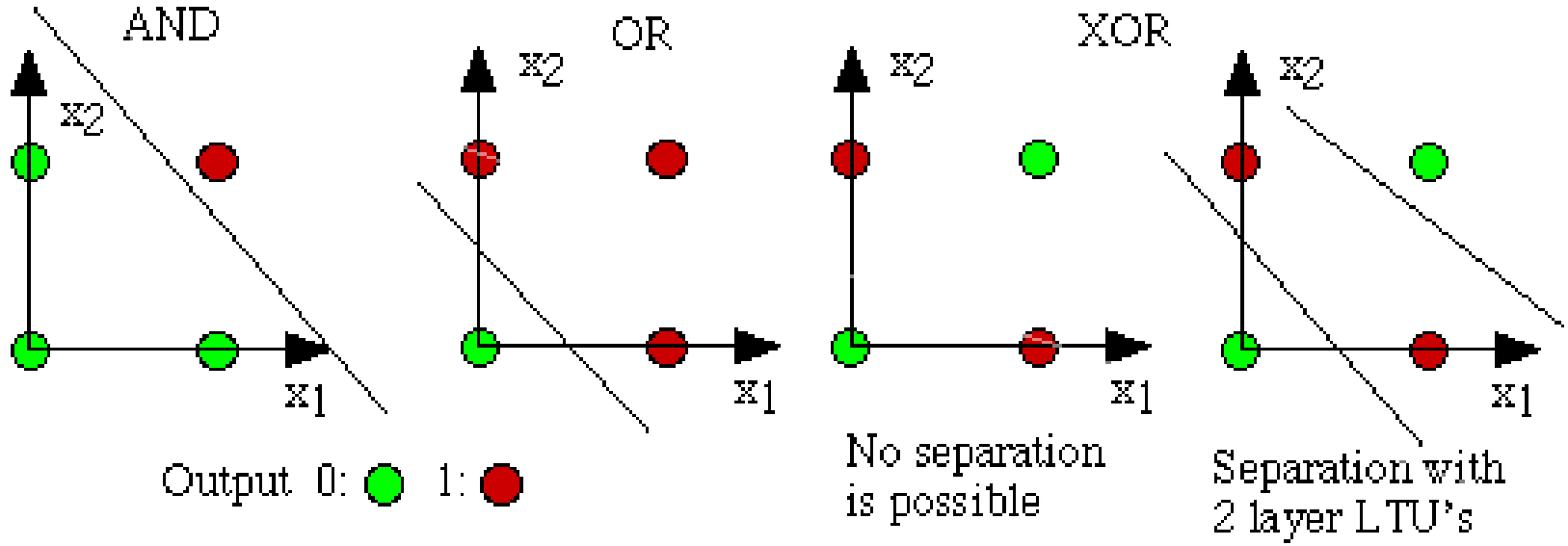
缺点：

1、无法预先知道特征；

2、在高维空间，特征数巨大；

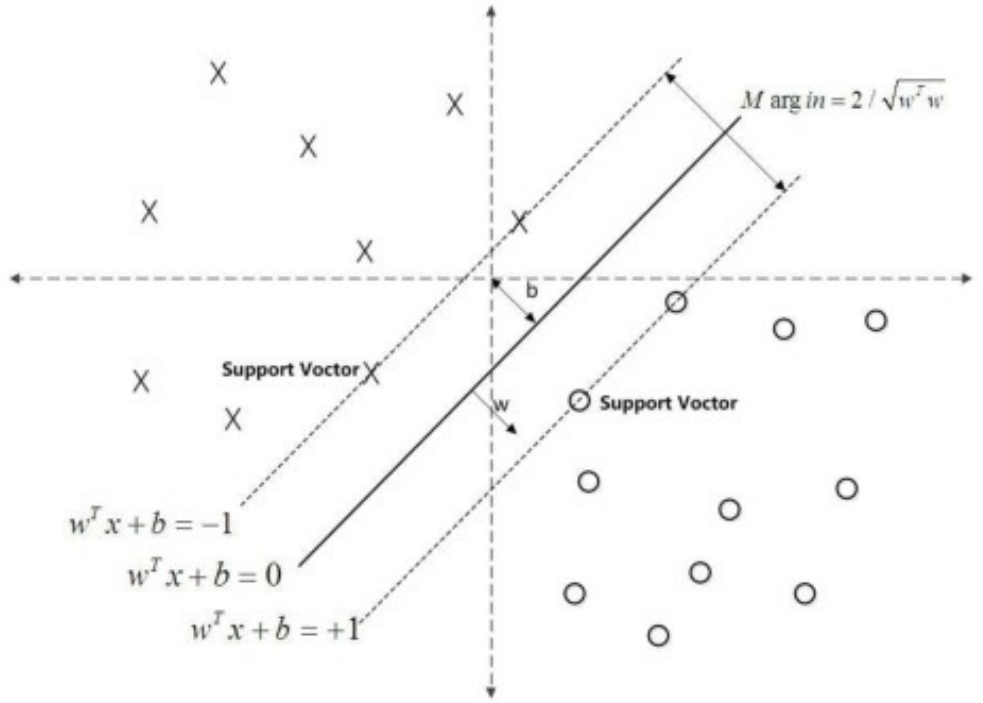
（核算法和支持向量机的起源。）

几何含义



Minsky, Marvin, and Papert Seymour. “Perceptrons.” (1969)

支持向量机：线性分类器



如何求解 w & b ?

拉格朗日乘子

$$\mathcal{L}(w,b,\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{n}\alpha_i(y_i(w^Tx_i+b)-1)$$

对 w, b 求极小：

$$\frac{\partial L}{\partial w}=0\Rightarrow w=\sum_{i=1}^{n}\alpha_i y_i x_i$$

$$\frac{\partial L}{\partial b}=0\Rightarrow \sum_{i=1}^{n}\alpha_i y_i=0$$

$$\Rightarrow \mathcal{L}(w,b,\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

对 α 求极大：

$$w^*=\sum_{i=1}^{n}\alpha_i y_i x_i$$

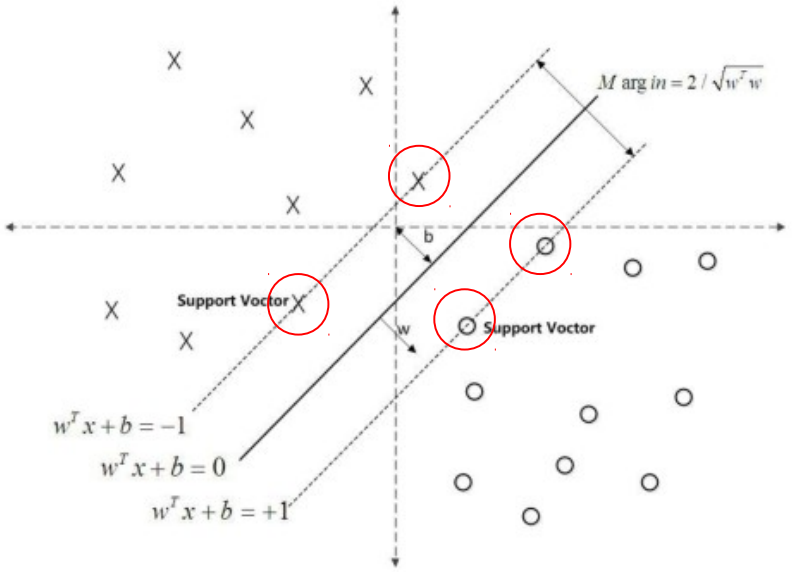
$$b^*=-\frac{\max_{i:y_i=-1} w^{*T}x_i+\min_{i:y_i=1} w^{*T}x_i}{2}$$

线性分类器：支持向量机

$$f(x) = w^T x + b$$

$$= (\sum_{i=1}^{n} \alpha_i y_i x_i)^T x + b$$

$$= \sum_{i=1}^{n} \alpha_i y_i \langle x_i, x \rangle + b$$

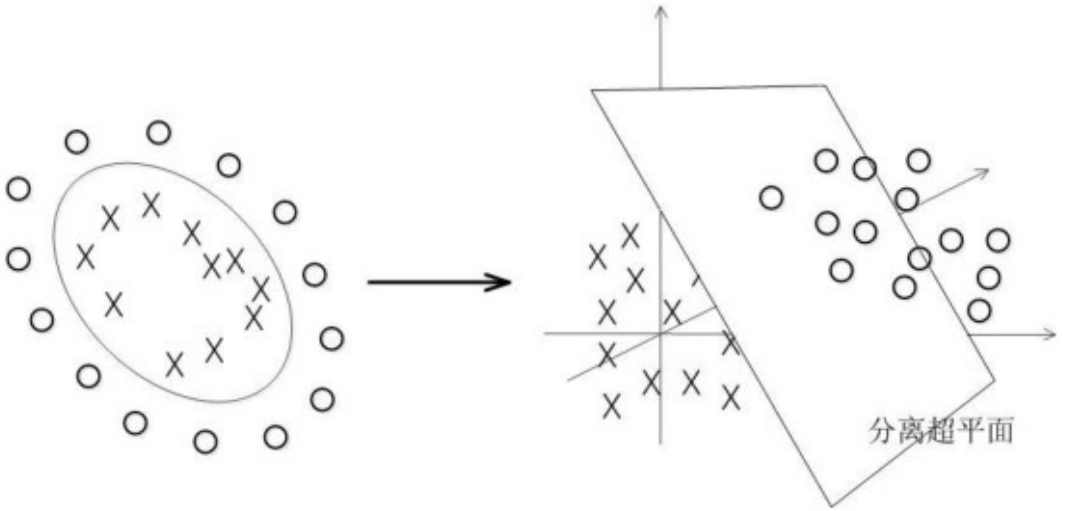


内积
核方法的基础

决定分类器的只有有限的几个支持向量。

支持向量机：非线性分类器和核模型

非线性 —核模型→ 线性



低维 —核模型→ 高维

1. 将数据从低维映射到高维；
2. 在高维空间进行分类；

核模型：内积 → 核函数

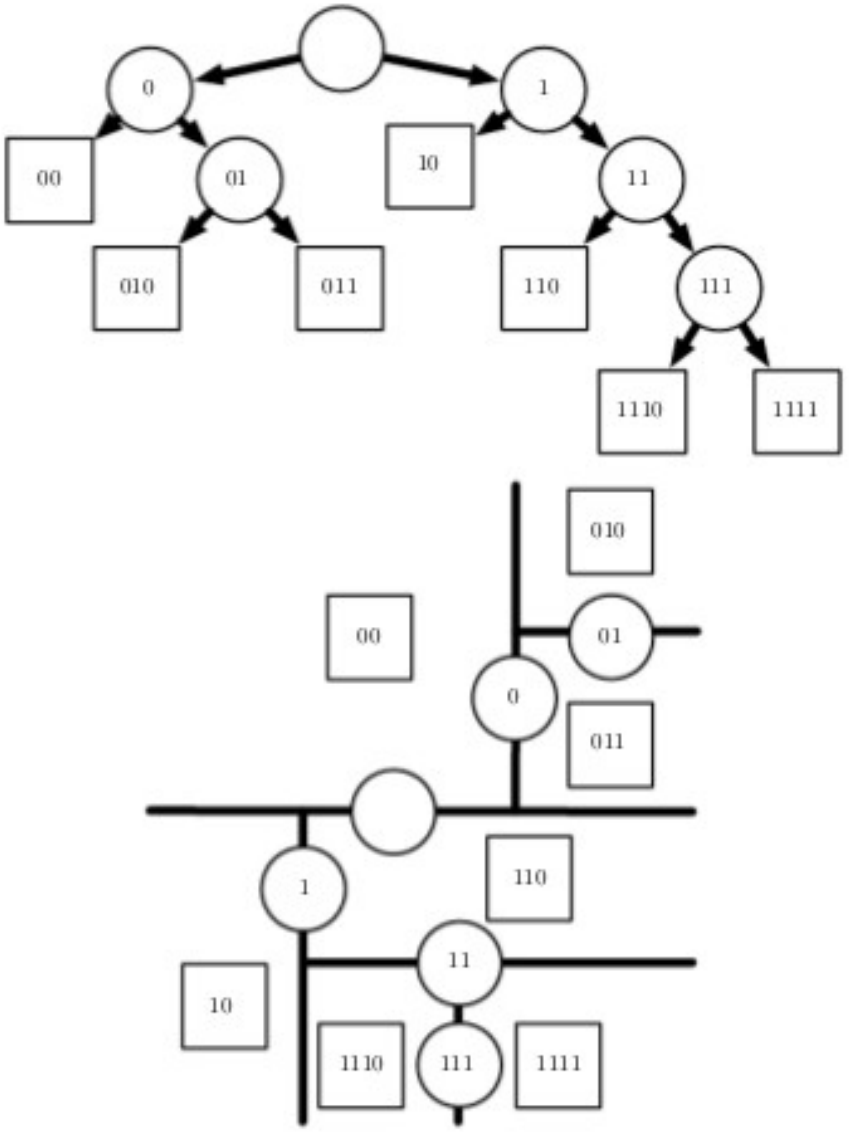
核函数

多项式核： $K(x_1, x_2) = (\langle x_1, x_2 \rangle + R)^d$

线性核： $K(x_1, x_2) = \langle x_1, x_2 \rangle$

高斯核： $K(x_1, x_2) = exp(-\|x_1 - x_2\|^2 / 2\sigma^2)$

决策树

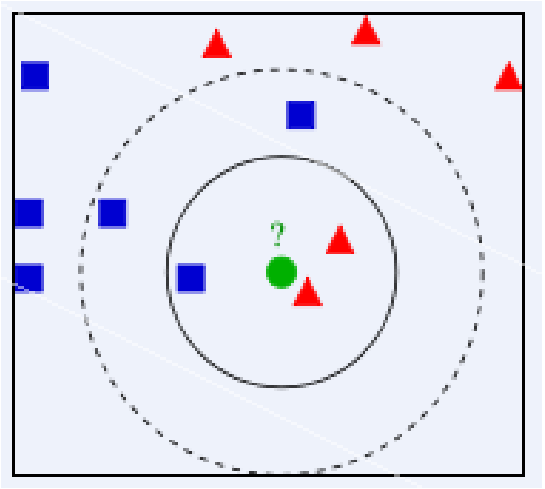

0
1
00
01
10
11
010
011
110
111
1110
1111
010
00
01
0
011
1
110
11
10
1110
111
1111

k 近邻：监督学习

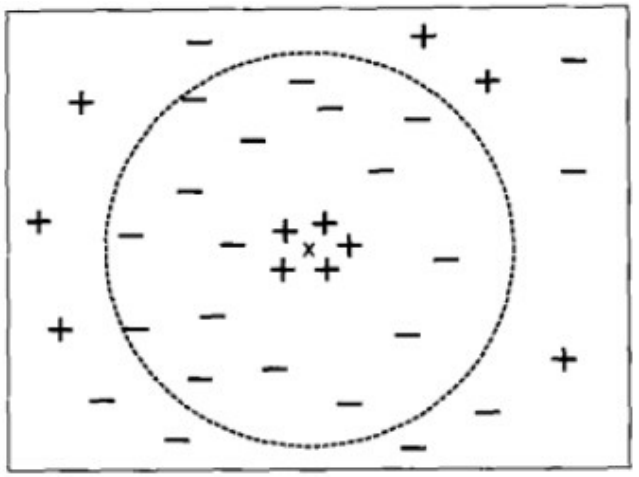
无参数、大容量、计算量大、无法处理特征分布不均匀的情况；

分类：由 k 近邻“多数表决”确定类别；

回归：由 k 近邻的平均值确定取值；



k 值决定论

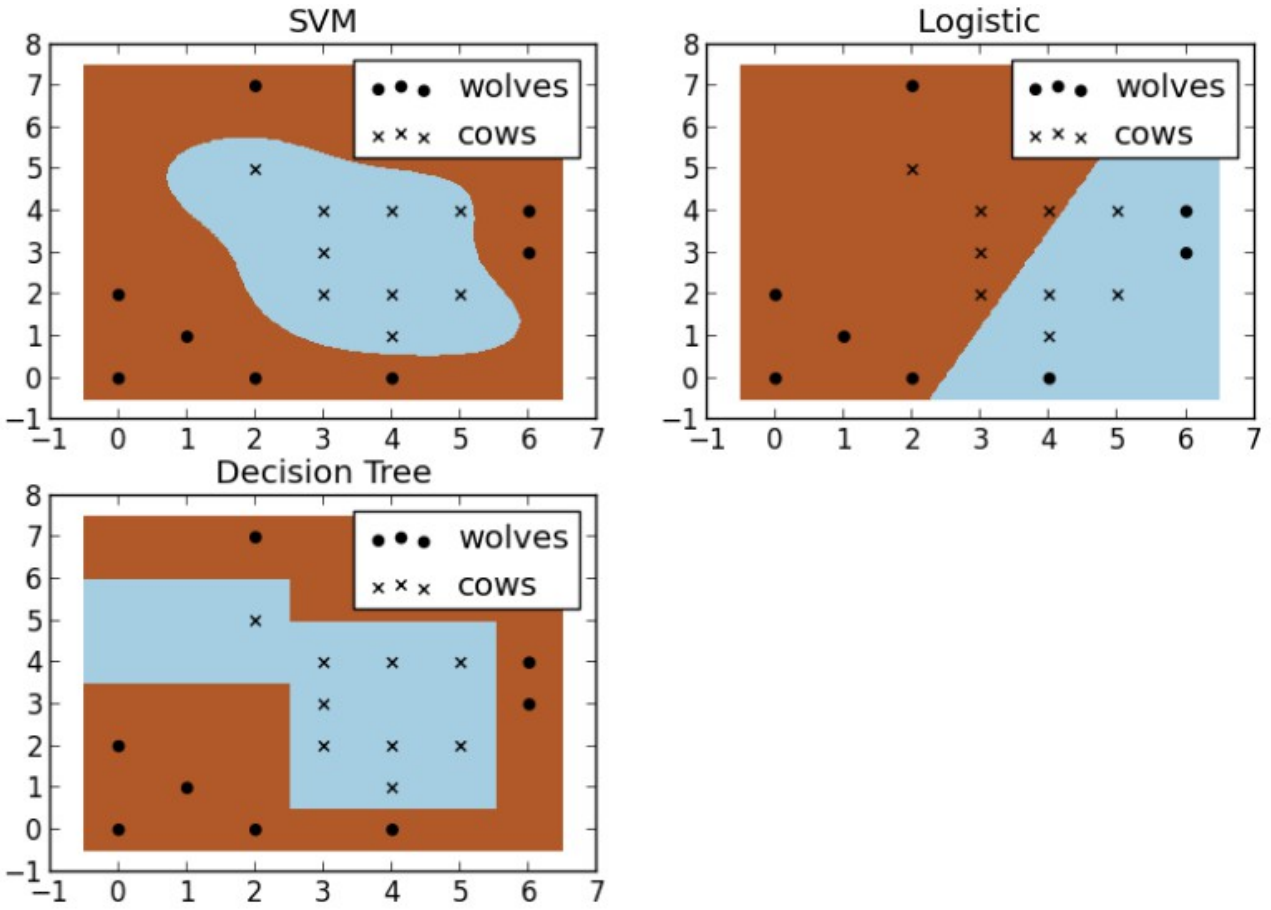


k 值不能太大

	A	B	C	D	E	F
1		K-Nearest Neighbor for Time Series				
2						
3		K	2			
4						
5						Nearest Neighbor
6		X	Y		distance	Value
7	Data	1	23		5.5	
8		1.2	17		5.3	
9		3.2	12		3.3	
10		4	27		2.5	27
11		5.1	8		1.4	8
12		6.5	?			
13						
14						
15	prediction					
16					result	
17					KNN prediction	17.5

高维数据的最近邻：
最大堆排序

羊圈：逻辑回归、支持向量机和决策树



无监督学习

	特征	标签	适用
监督学习	✓	✓	分类
无监督学习	✓	X	提取特征 标记样本 表象变换

标准：低维、稀疏、独立；

主成分分析：幺正变化、去关联化

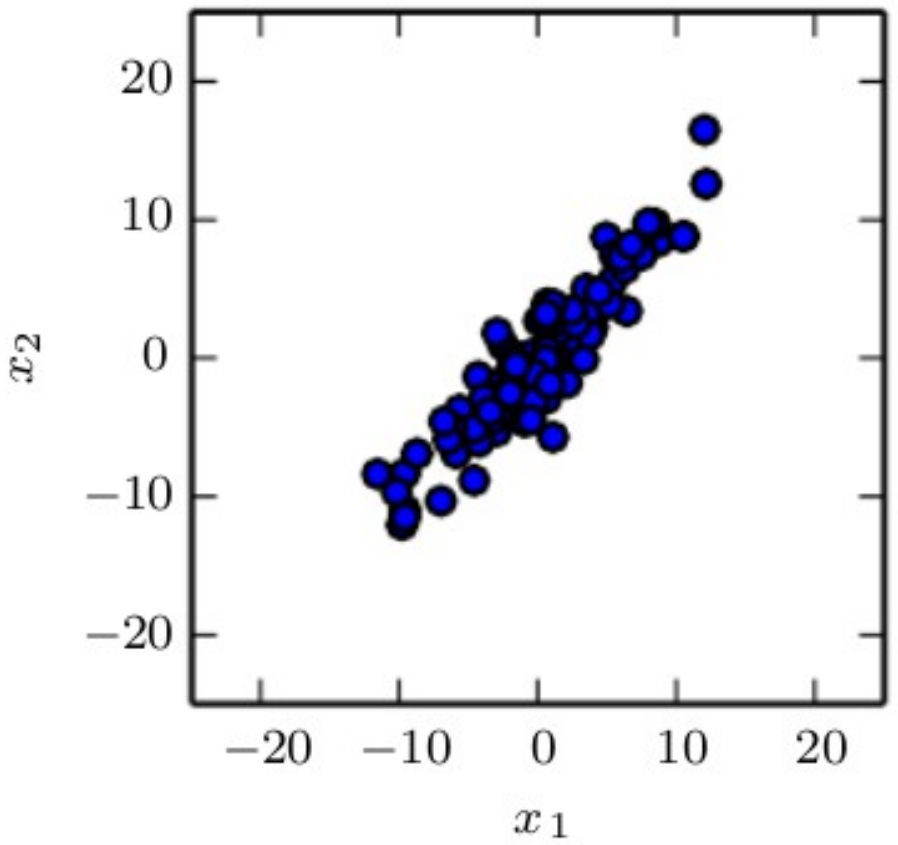
线性变换： $\boldsymbol{z} = \boldsymbol{x}^\top \boldsymbol{W}$ $\mathrm{Var}[\boldsymbol{x}] = \frac{1}{m-1}\boldsymbol{X}^\top \boldsymbol{X} = \frac{1}{m-1}\boldsymbol{W}\boldsymbol{\Sigma}^2\boldsymbol{W}^\top$

对角化： $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top$ $\mathrm{Var}[\boldsymbol{z}] = \frac{1}{m-1}\boldsymbol{\Sigma}^2$ 对角化 去关联

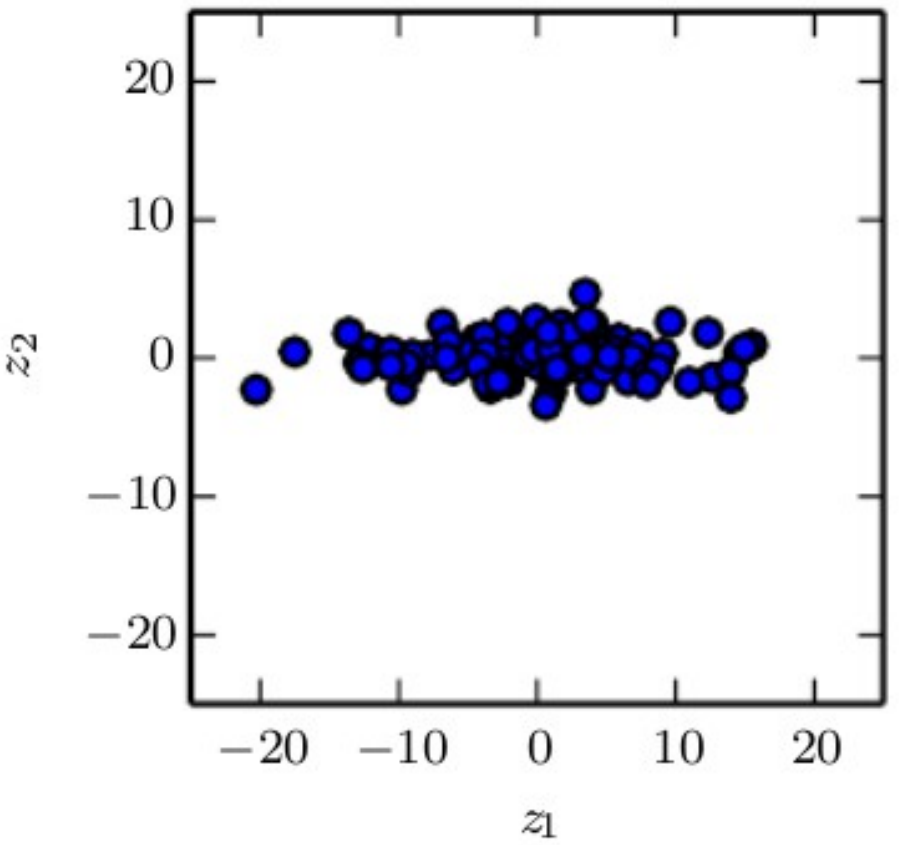
$\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}$ $\boldsymbol{W}\boldsymbol{W}^\top = \boldsymbol{I}$ PCA(X) = eigen($\boldsymbol{X}^\top\boldsymbol{X}$)

Σ 是对角化矩阵，当 Σ 为非方矩阵的时候，该变换称为奇异值分解。其中奇异值较大的特征，被称为主成分。

用来减少特征数，降维。



二维数据



一维数据

k 均值：聚类

聚类：在没有明确标定的情况下，对个体按潜在的类型进行分类。

1、 随机选取k个聚类质心点（cluster centroids）为$\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$。

2、 重复下面过程直到收敛 {

对于每一个样例i，计算其应该属于的类

$$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2.$$

对于每一个类j，重新计算该类的质心

$$\mu_j := \frac{\sum_{i=1}^m 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^m 1\{c^{(i)} = j\}}.$$

}

随机梯度下降法

价格函数 $J(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x},\mathrm{y}\sim\hat{p}_{\mathrm{data}}} L(\boldsymbol{x}, y, \boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m}\left(-\log p(y^{(i)} \mid \boldsymbol{x}^{(i)}; \boldsymbol{\theta})\right)$

梯度运算 $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{x}^{(i)}, y^{(i)}, \boldsymbol{\theta})$

m= 千万以上 ⟶ m' =100

在深度学习
部分讲述！

$$\boldsymbol{g} = \frac{1}{m'}\nabla_{\boldsymbol{\theta}}\sum_{i=1}^{m'} L(\boldsymbol{x}^{(i)}, y^{(i)}, \boldsymbol{\theta})$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \epsilon \boldsymbol{g},$$

GGA
……

学习
速率

设计一个完整的机器学习算法

数据

+ 价格（ cost ）函数

（损失函数的算术平均值）

+ 优化流程

+ 模型

= 机器学习算法

困境：维度灾难

传统 AI 的核心问题：语音识别和物体识别

计算机学家最感兴趣的问题

与高维积分类似， N^d ；

理解高维积分：把每个点看作一个样本，其坐标作为向量 x，函数值作为标签 y，这样高维积分转换为机器学习问题。

流形：不严格的数学

非欧空间的局域欧氏子空间；

n 维流形在任意点和 n 维欧氏空间局域同胚；

把高维数据在低维流形上表示，因此是一种降维的方法；

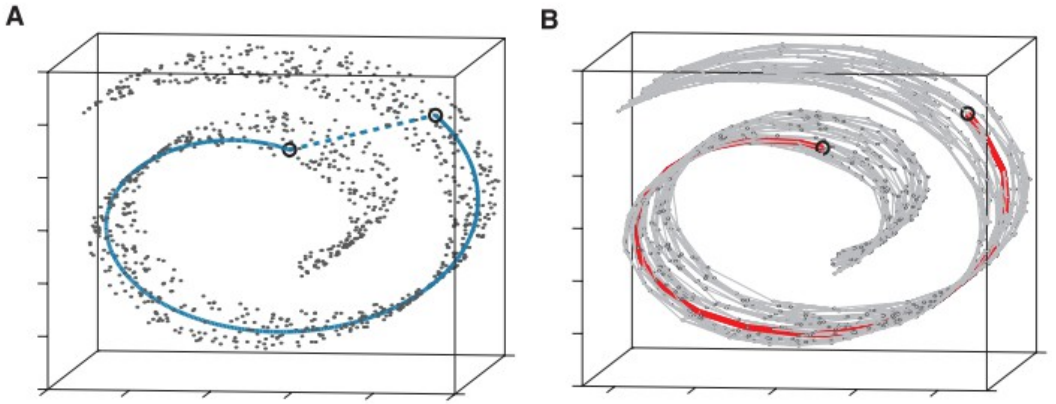
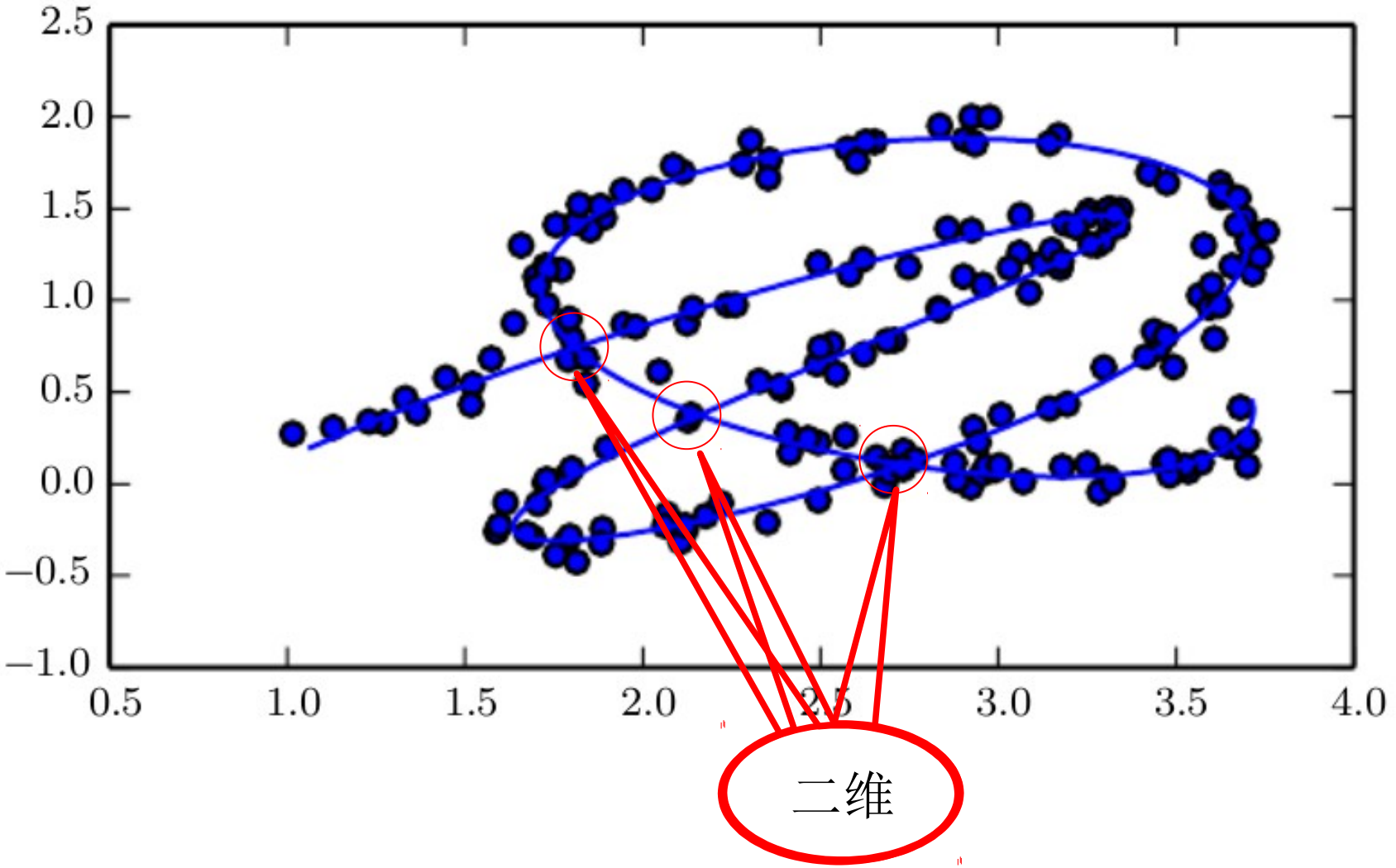


Fig. 3. The "Swiss roll" data set, illustrating how Isomap exploits geodesic paths for nonlinear dimensionality reduction. (**A**) For two arbitrary points (circled) on a nonlinear manifold, their Euclidean distance in the high-dimensional input space (length of dashed line) may not accurately reflect their intrinsic similarity, as measured by geodesic distance along the low-dimensional manifold (length of solid curve). (**B**) The neighborhood graph G constructed in step one of Isomap (with $K = 7$ and $N =$ 1000 data points) allov geodesic path to be co path in G. (**C**) The two- step three, which best neighborhood graph (o now represent simpler paths than do the corre

A Global Geometric Framework for Nonlinear Dimensionality Reduction

Joshua B. Tenenbaum,[1]* Vin de Silva,[2] John C. Langford[3]

学习函数的分与合



总结

1、传统机器学习算法的数学结构清晰；

2、应用层面最大的问题是如何针对具体问题提取特征；

3、传统机器学习有不可逾越的困难，尤其是在计算机科学家感兴趣的图像识别和语音识别领域；

4、物理和数学上的很多问题可以用传统机器学习进行理解。

什么是人工智能？

没有标准答案！

Scikit-learn 见！

谢谢！