



華東師範大學
EAST CHINA NORMAL UNIVERSITY

机器学习概念

数据

□ 数据是经验的另一种说法，也是信息的载体。

■ 结构化数据和非结构化数据 (按数据具体类型划分)

■ 原始数据和加工数据 (按数据表达形式划分)

■ 样本内数据和样本外数据 (按数据统计性质划分)

□ 结构化和非结构化

■ 结构化数据 (structured data) 是由二维表结构来逻辑表达和实现的数据。非结构化数据是没有预定义的数据，不使用数据库二维表来表现的数据。

结构化数据

- ❑ 机器学习模型主要使用的是结构化数据，即二维的数据表。

特征值 feature value		特征 feature			标签 label
	得分	篮板	助攻	比赛结果	
1	27	10	12	赢	
2	33	9	9	输	
3	51	10	8	输	
4	40	13	15	赢	

训练集
training set

训练样例
training example

非结构化数据

□ 非结构化数据包括图片，文字，语音和视频等

图片



语音



文字



第2章 模型评估与选择

2.1 经验误差与过拟合

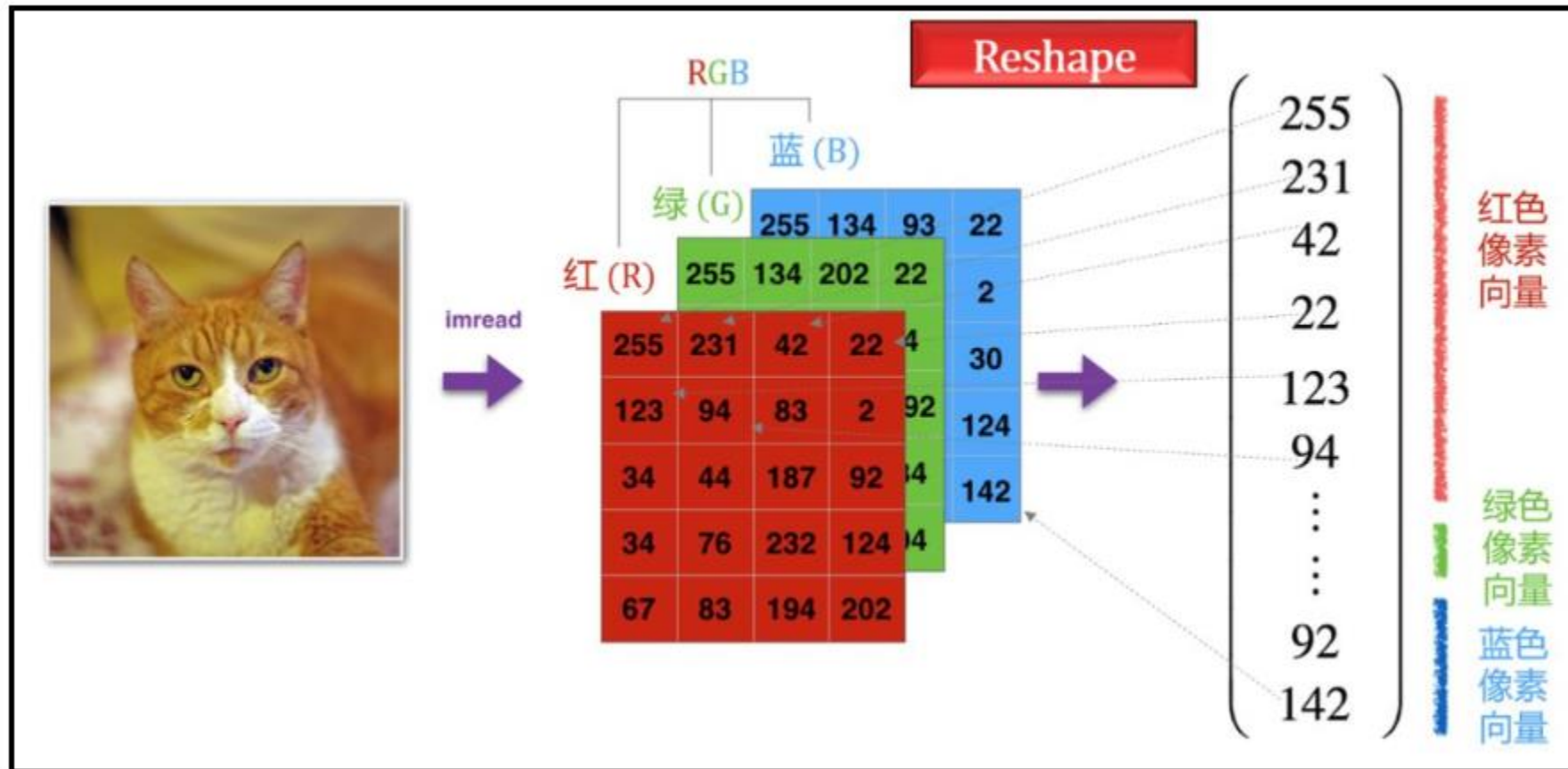
通常我们把分类错误的样本数占样本总数的比例称为“错误率” (error rate), 假如我们在 m 个样本中有 a 个样本分类错误, 则错误率 $E = a/m$, 相应的, $1 - a/m$ 称为“精度” (accuracy), 即“精度 = 1 - 错误率”。更一般地, 我们把学习器的实际预测输出与样本的真实输出之间的总差异称为“误差” (error), 学习器在训练集上的误差称为“训练误差” (training error) 或“经验误差” (empirical error), 在新样本上的误差称为“泛化误差” (generalization error)。显然, 我们希望得到泛化误差小的学习器。然而, 我们事先并不知道新样本是什么样, 实际能做的是努力使训练误差最小化。在通常情况下, 我们可以学得一个训练误差很小, 在训练集上表现很好的学习器, 例如是对所有训练样本都分类正确, 得分类错误率为零, 分类精度为100%, 但这是不是我们想要的学习器呢? 遗憾的是, 这样的学习器在多数情况下都不行。

棋谱



原始数据和加工数据

□ 图像型数据



原始数据和加工数据

□ 文本型数据

ASCII TABLE

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	\$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[ENG OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

原始数据和加工数据

I love python :)



将句子分解成字符，少于
280字符的地方用空格填

$[\text{'I' space 'l' 'o' 'v' 'e' space 'p' 'y' 't' 'h' 'o' 'n' space ':' '}' space space space ... space}]_{1 \times 280}$



根据 ASCII 表
将字符用数字代替

$[49 \text{ 20 } 108 \ 111 \ 76 \ 65 \text{ 20 } 70 \ 79 \ 74 \ 68 \ 111 \ 110 \text{ 20 } 58 \ 41 \text{ 20 } 20 \ 20 \text{ ... 20}]_{1 \times 280}$



独热编码

	0	0	0	0	0	0	...	0
第 20 行	:	:	:	\	\	\	\	\	\	\	\	\	\	:	:	:	\	:
第 49 行	:	1	:	\	\	\	\	\	\	\	\	\	\	:	:	:	\	:
第 108 行	1	:	:	\	\	\	\	\	\	\	\	\	\	:	:	:	\	:
	:	:	1	\	\	\	\	\	\	\	\	\	\	:	:	:	\	:
	:	:	:	\	\	\	\	\	\	\	\	\	\	:	:	:	\	:
	0	0	0	0	0	0	...	0

140×280

原始数据和加工数据

□ 分类型变量

	得分	篮板	助攻	比赛结果
1	27	10	12	赢
2	33	9	9	输
3	51	10	8	输
4	40	13	15	赢

两类变量用「0-1编码」，比如比赛结果 = {赢, 输} 表示成 $y = [1 \ 0 \ 0 \ 1]$ ，1 代表赢，0 代表输。

原始数据和加工数据

	射门	传球	控球	比赛结果
1	9	42	12	赢
2	4	30	9	平
3	6	14	8	赢
4	0	22	15	输

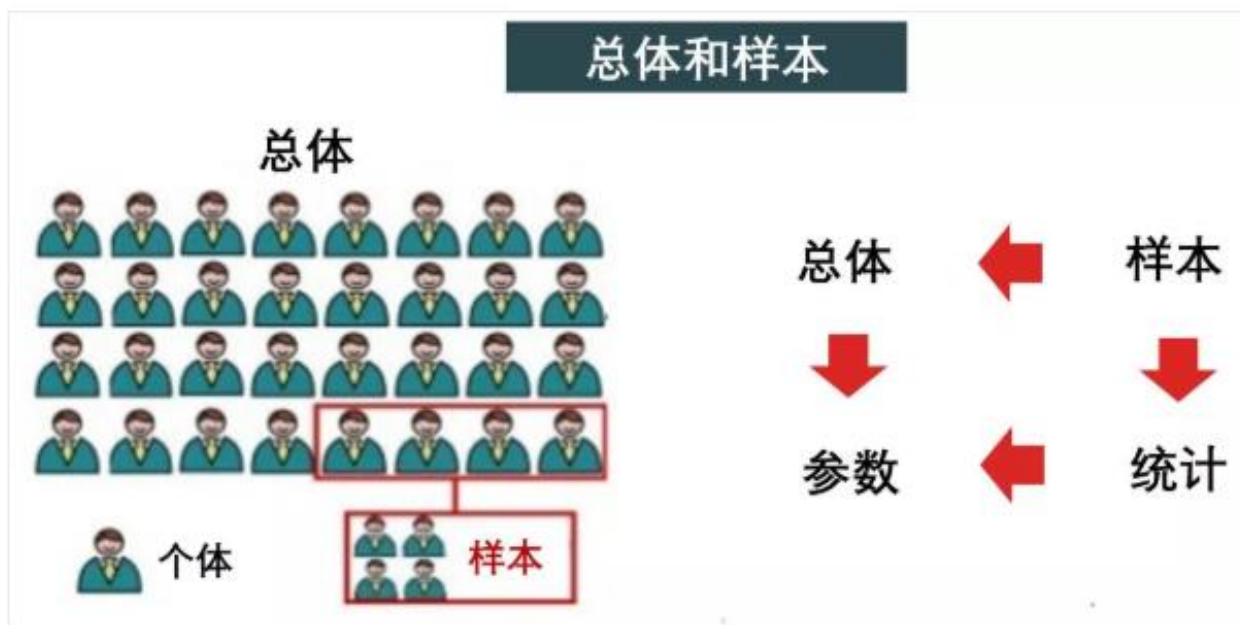
多类变量分别用 0, 1, 2 来表示，那么 $y = [0 \ 1 \ 0 \ 2]$ 。

也可使用独热编码

$$\text{赢} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \text{平} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \text{输} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, y = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

样本内和样本外

在统计中，把研究对象的全体称为总体 (population)，而把组成总体的各个元素称为个体，把从总体中抽取的若干个体称为样本 (sample)。



样本内和样本外



数据规范化

Min-Max规范化:

$$x' = x'_{min} + \frac{x - x_{min}}{x_{max} - x_{min}} \times (x'_{max} - x'_{min})$$

中心化:

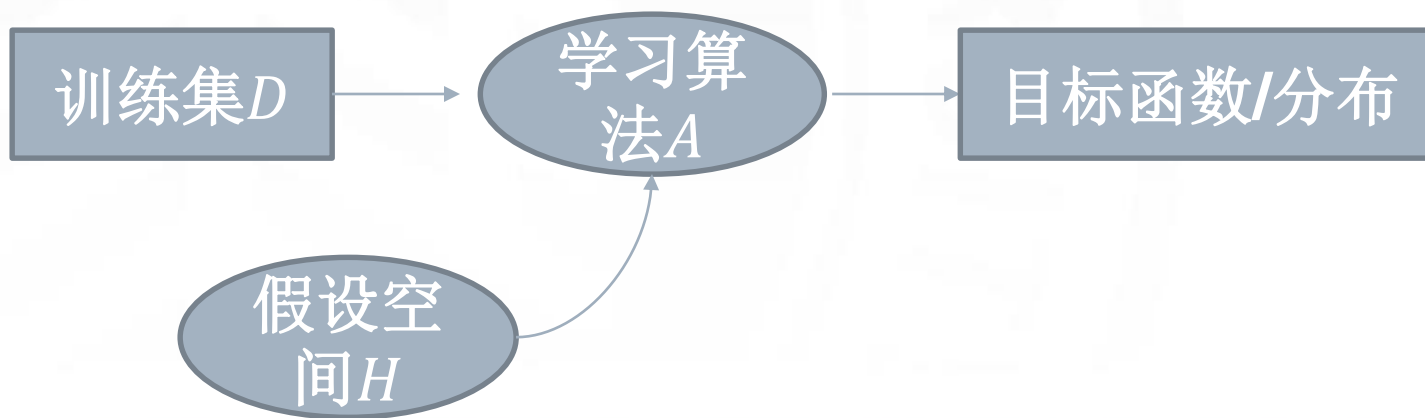
$$x' = x - \bar{x}$$

标准化:

$$x' = \frac{x - \bar{x}}{\sigma}$$

机器学习基本概念

- ❑ 输入: $\mathbf{X} \in \mathcal{X}$, 输出: $\mathbf{Y} \in \mathcal{Y}$, 输出实例: $y \in \mathcal{Y}$
- ❑ 输入实例: $\mathbf{x} = (x_1, x_2, \dots, x_d) \in \mathcal{X}$ 或者 $\mathbf{x} = (x^1, x^2, \dots, x^d) \in \mathcal{X}$
- ❑ 目标函数: $Y = f(\mathbf{X})$; 目标分布: $P(Y|\mathbf{X})$
- ❑ 对具体的输入时: $y = f(\mathbf{x})$ 或 $P(y|\mathbf{x})$
- ❑ 数据集: $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$



机器学习三要素：模型

- 模型：决策函数或者条件概率分布
- 假设空间：决策函数或者条件概率分布的集合
- 决策函数集合： $\mathcal{H} = \{f \mid Y = f(\mathbf{X}; \theta), \theta \in \mathcal{R}^n\}$
- 条件概率的集合： $\mathcal{H} = \{P \mid P(Y|\mathbf{X}; \theta), \theta \in \mathcal{R}^n\}$

机器学习三要素：策略

□ 策略：从假设空间中选取最优模型

□ 损失函数: $L(y, \hat{y})$

■ 0-1损失函数:

$$L(Y, f(\mathbf{X})) = \mathbb{I}[f(\mathbf{X}) \neq Y] = \begin{cases} 1, & Y \neq f(\mathbf{X}) \\ 0, & Y = f(\mathbf{X}) \end{cases}$$

■ 平方损失函数:

$$L(Y, f(\mathbf{X})) = (Y - f(\mathbf{X}))^2$$

■ 绝对损失函数:

$$L(Y, f(\mathbf{X})) = |Y - f(\mathbf{X})|$$

■ 对数损失函数:

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

机器学习三要素：策略

例：在分类数为 M 的分类问题中，设 $p_i(\mathbf{x})$ 为分类器将 \mathbf{x} 预测为类别 i 的概率，则其对数损失函数为？

$$L(y, P(y|\mathbf{x})) = -\log P(y|\mathbf{x}) = -\sum_{i=1}^M \mathbb{I}[y = i] \log p_i(\mathbf{x})$$

更进一步，考虑数据集容量为 N ，则该数据集的平均损失函数（代价函数）为：

$$L(Y, P(Y|\mathbf{X})) = -\log P(Y|\mathbf{X}) = -\frac{1}{N} \sum_{j=1}^N \sum_{i=1}^M \mathbb{I}[y_j = i] \log p_i(\mathbf{x}_j)$$

机器学习三要素：策略

成本（风险）函数：

□ 期望风险：
$$R_{exp}(f) = \int_{\mathcal{X} \times \mathcal{Y}} L(y, f(\mathbf{x})) P(\mathbf{x}, y) d\mathbf{x} dy$$

给定训练集：
$$T = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$$

□ 经验风险：

$$R_{emp}(f) = \frac{1}{N} \sum_{i=1}^N L(y_i, f(\mathbf{x}_i))$$

□ 结构风险：

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

策略：

$$\min_{f \in \mathcal{H}} R_{emp}(f) \quad \text{或者} \quad \min_{f \in \mathcal{H}} R_{srm}(f)$$

机器学习三要素：算法

□ 算法：学习模型的具体算法, 选取最优模型

□ 最优化问题: $\min_{w,b} J(w, b)$

■ 极值问题

■ 梯度下降

■ 牛顿法和拟牛顿法

■ 约束优化问题——拉格朗日乘数法

机器学习一般流程

步骤：

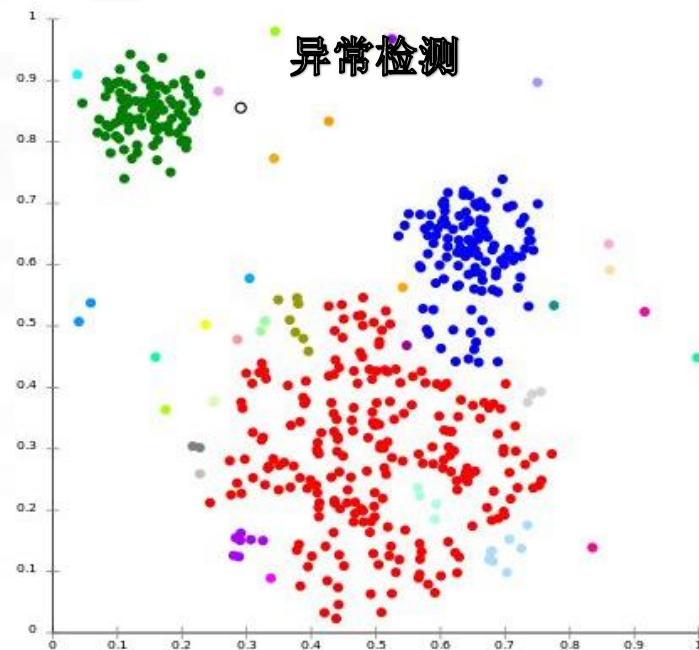
1. 得到一个有限的训练数据集
2. 确定包含所有可能的模型的假设空间，即学习模型的集合
3. 确定模型选择的准则，即学习的策略
4. 实现求解最优模型的算法，即学习的算法
5. 通过学习方法选择最优模型
6. 利用学习的最优模型对新数据进行预测和分析。

机器学习分类

□ 按有无标签分类

- 监督学习：垃圾邮件分类、房价预测
- 非监督学习：异常检测
- 半监督学习：标注语音
- 强化学习：Alpha GO

垃圾邮件分类



机器学习分类

□ 按输出空间分类

- 二分类：垃圾邮件分类
- 多分类：图像分类
- 回归：房价预测
- 结构化学习：机器翻译、语音识别、聊天机器人



→ *Dog*

图像分类



→ *Cat*

机器学习分类

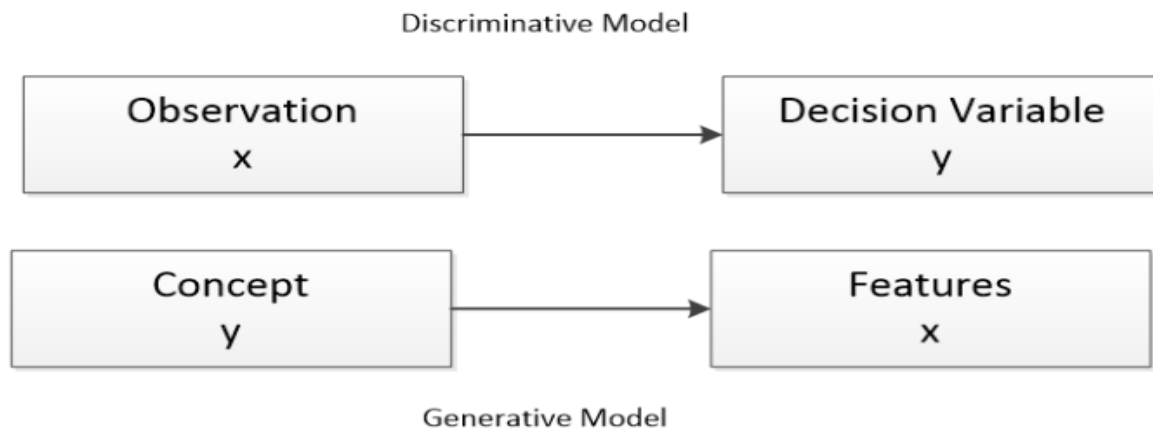
□ 按模型分类

■ 生成式模型：GAN

- 先确定 $P(x, y)$
- 然后利用贝叶斯定理： $P(y|x) = \frac{P(x, y)}{P(x)}$

■ 判别式模型：决策树、支持向量机

- 直接确定 $P(y|x)$
- 或 $f(x)$



机器学习分类

□ 按算法分类

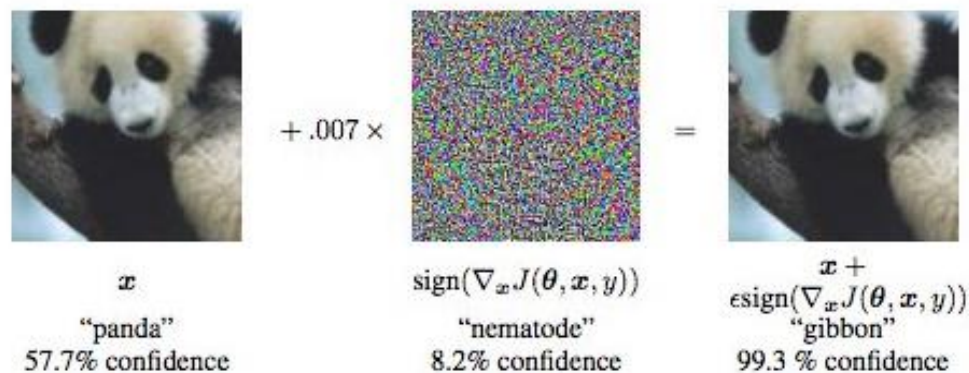
- 批量学习：一次性批量输入给学习算法，可以被形象的称为填鸭式学习
- 在线学习：按照顺序，循序的学习，不断的去修正模型，进行优化
- 主动学习：通过某种策略找到未进行类别标注的样本数据中最有价值的数据，交由专家进行人工标注后，将标注数据及其类别标签纳入到训练集中迭代优化分类模型，改进模型的处理效果

机器学习挑战

□ 模型的预测效果

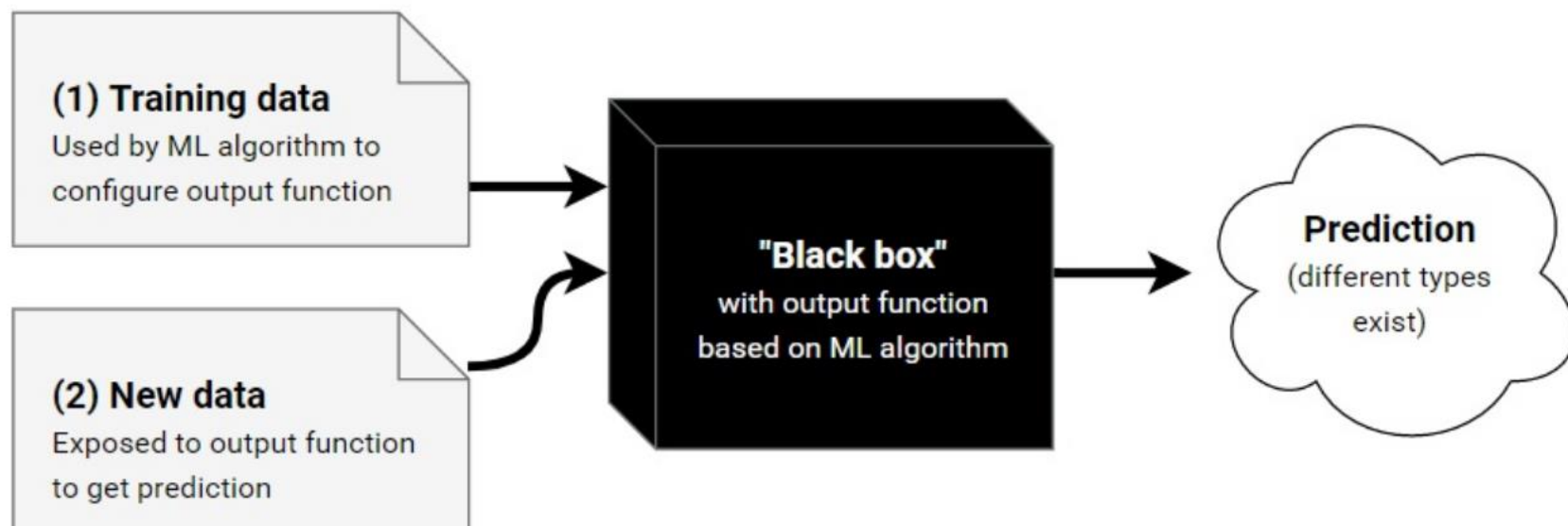
□ 模型的稳定性

- 对抗样本：攻击者通过在源数据上增加人类难以通过感官辨识到的细微改变，但是却可以让机器学习模型接受并做出错误的分类决定。
- 一个典型的场景就是图像分类模型的对抗样本，通过在图片上叠加精心构造的变化量，在肉眼难以察觉的情况下，让分类模型产生误判。



机器学习挑战

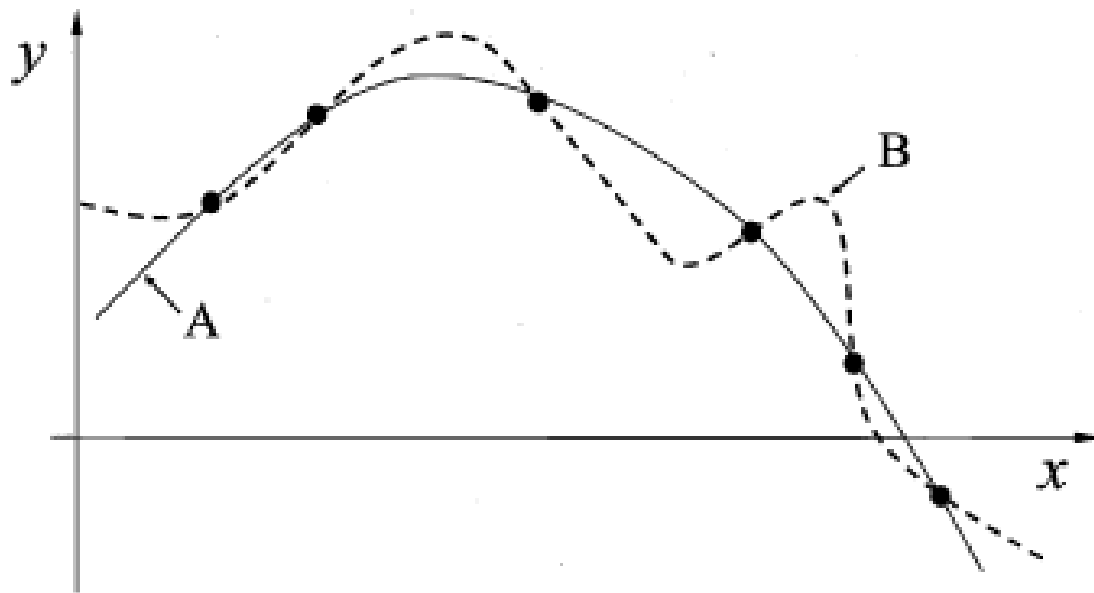
- 模型结果的可解释性
- 算法歧视



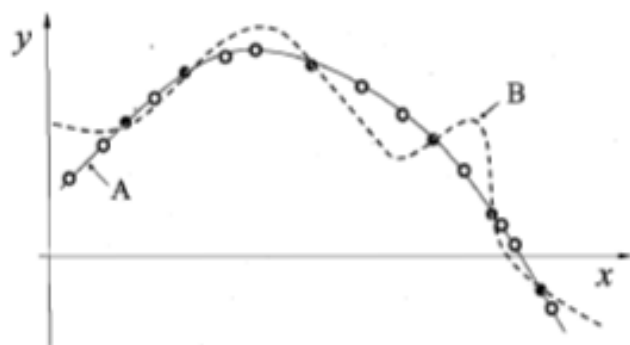
奥卡姆剃刀

奥卡姆剃刀 (Occam's razor)

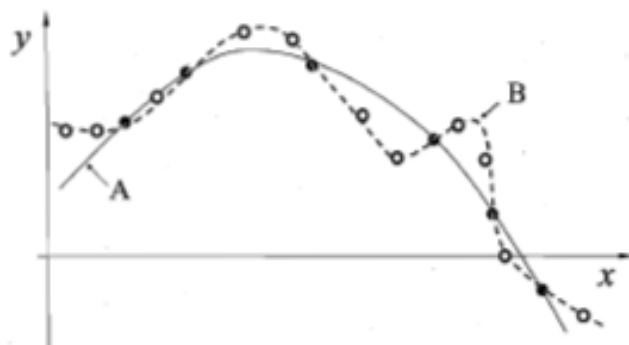
若有多于一个假设与观察一致，则选最简单的那个



没有免费的午餐



(a) A 优于 B



(b) B 优于 A

没有免费的午餐

假设样本空间 \mathcal{X} 和假设空间 \mathcal{H} 都是离散的。令 $P(h|X, \mathcal{L}_a)$ 表示算法 \mathcal{L}_a 基于训练数据 X 产生假设 h 的概率，再令 f 代表我们希望学习的真实目标函数。则算法 \mathcal{L}_a 在训练集之外的所有样本上的误差为：

$$E_{ote}(\mathcal{L}_a|X, f) = \sum_h \sum_{\mathbf{x} \in \mathcal{X} - X} P(\mathbf{x}) \mathbb{I}[h(\mathbf{x}) \neq f(\mathbf{x})] P(h|X, \mathcal{L}_a)$$

没有免费的午餐

考虑二分类问题，且真实目标函数可以是任何函数

$$f: \mathcal{X} \rightarrow \{0, 1\}$$

对所有可能的函数按照均匀分布对误差求和，有

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

也就是：

$$\sum_f E_{ote}(\mathcal{L}_a | X, f) = \sum_f E_{ote}(\mathcal{L}_b | X, f)$$

没有一种机器学习算法是适用于所有情况的

频率派和贝叶斯派

$$X_{N \times p} = (x_1, x_2, \dots, x_N)^T, x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$$

假设每个观测都是由 $p(x|\theta)$ 生成的

频率派: θ 是一个常量

- 以客观世界为研究主体, 为了找到客观世界的某个规律
- 不关心参数空间的所有细节, 相信数据都是在这个空间里的"某个"参数值下产生的

对于 N 个观测来说, 观测集的概率为

$$p(X|\theta) \stackrel{iid}{=} \prod_{i=1}^N p(x_i|\theta)$$

为了求 θ 的大小, 我们采用最大对数似然MLE的方法

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}} \log p(X|\theta) \stackrel{iid}{=} \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^N \log p(x_i|\theta)$$

频率派和贝叶斯派

贝叶斯派： θ 是一个满足预设的先验分布 $\theta \sim p(\theta)$

- 探究的是我们对某一事件发生的相信程度，且这种相信程度会因为观测到的客观事件而改变
- 关心参数空间里的每一个值，参数空间里的每个值都有可能是真实模型使用的值。

根据贝叶斯定理，依赖观测集的参数后验为：

$$p(\theta|X) = \frac{p(X|\theta) \cdot p(\theta)}{p(X)} = \frac{p(X|\theta) \cdot p(\theta)}{\int_{\theta} p(X|\theta) \cdot p(\theta) d\theta}$$

为了求 θ 的大小，我们采用最大参数后验MAP的方法

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}} p(\theta|X) = \underset{\theta}{\operatorname{argmax}} p(X|\theta) \cdot p(\theta)$$

频率派和贝叶斯派

贝叶斯估计:

$$p(\theta|X) = \frac{p(X|\theta) \cdot p(\theta)}{\int_{\theta} p(X|\theta) \cdot p(\theta) d\theta}$$

贝叶斯预测:

$$p(x_{new}|X) = \int_{\theta} p(x_{new}|\theta) \cdot p(\theta|X) d\theta$$

贝叶斯派: 概率图模型

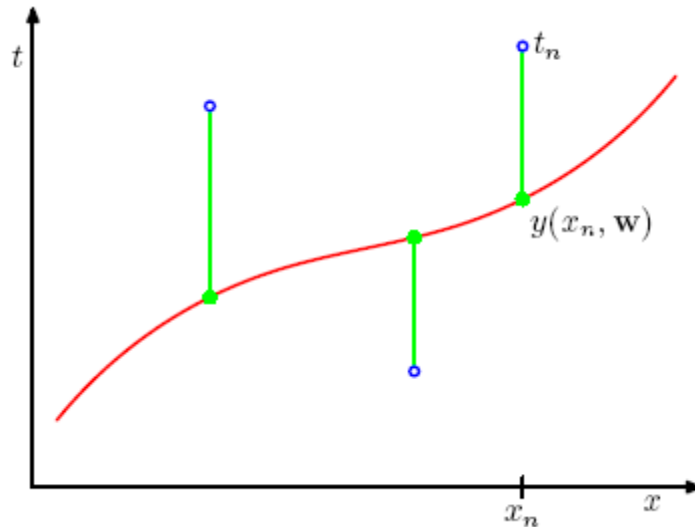
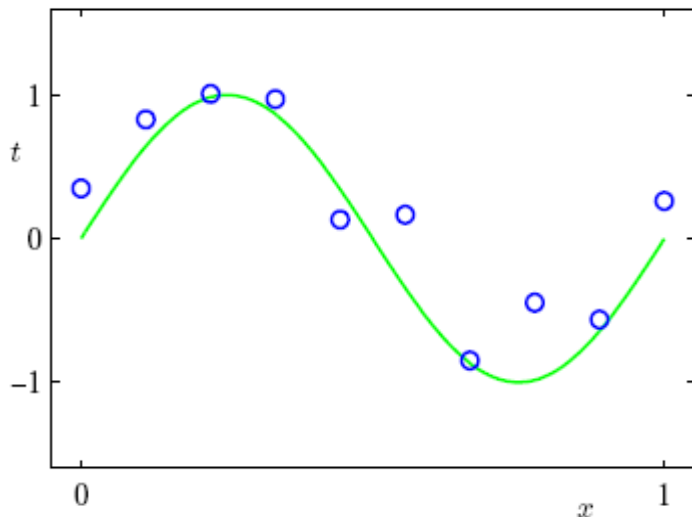
- 求积分问题: MCMC

频率派: 统计机器学习的优化问题:

- 1) 建立模型、概率
- 2) 定义损失函数
- 3) 梯度下降/牛顿法求解

例：多项式曲线拟合问题

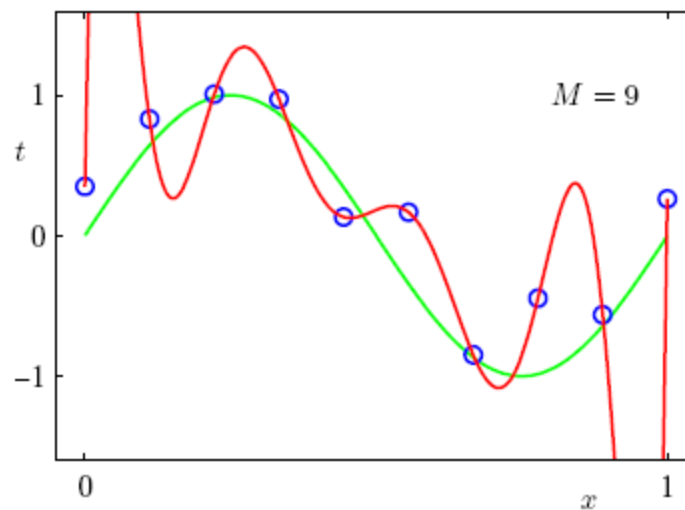
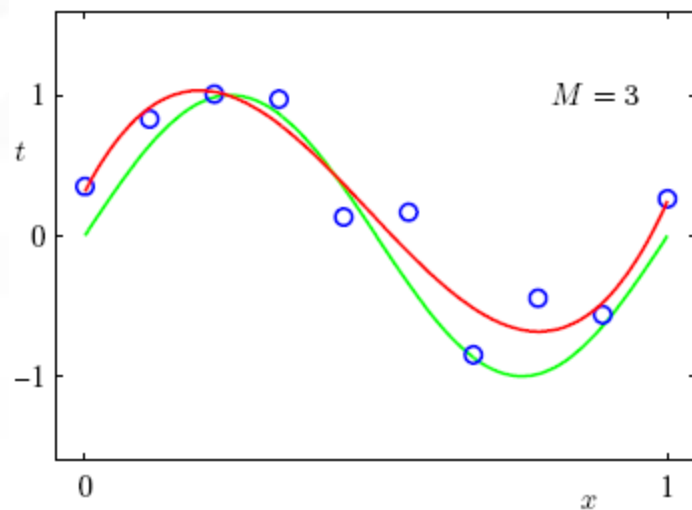
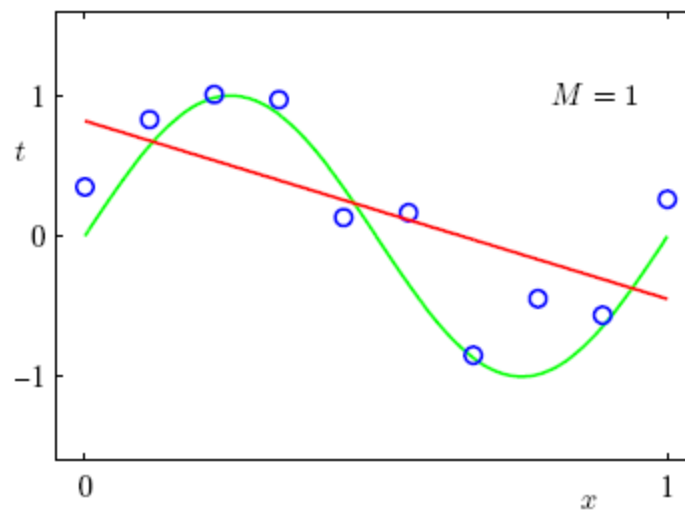
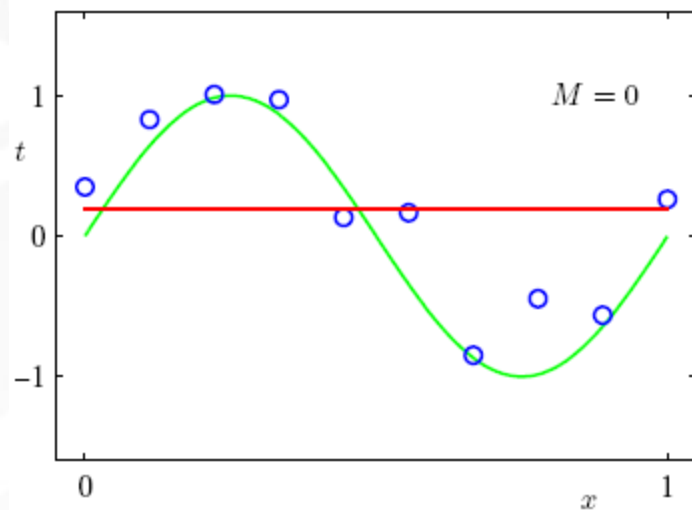
□ 例：给定一个训练集，输入为 $(x_1, x_2, \dots, x_N)^T$ ，输出为 $(y_1, y_2, \dots, y_N)^T$ ， $N=10$.



□ 拟合多项式

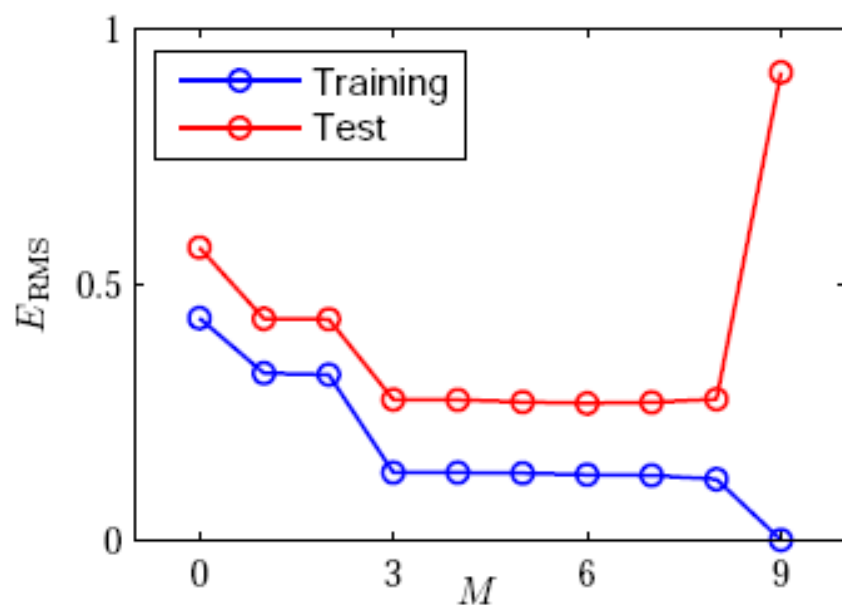
$$f(x) = \omega_0 + \omega_1 x + \omega_2 x^2 + \dots + \omega_M x^M$$

例：多项式曲线拟合问题



例：多项式曲线拟合问题

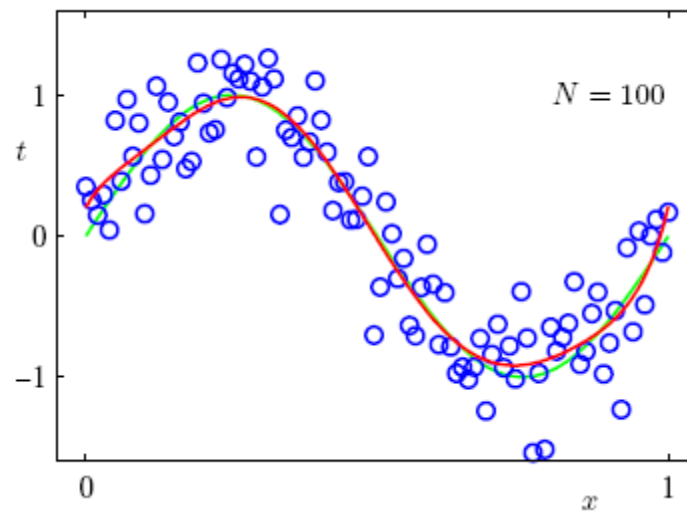
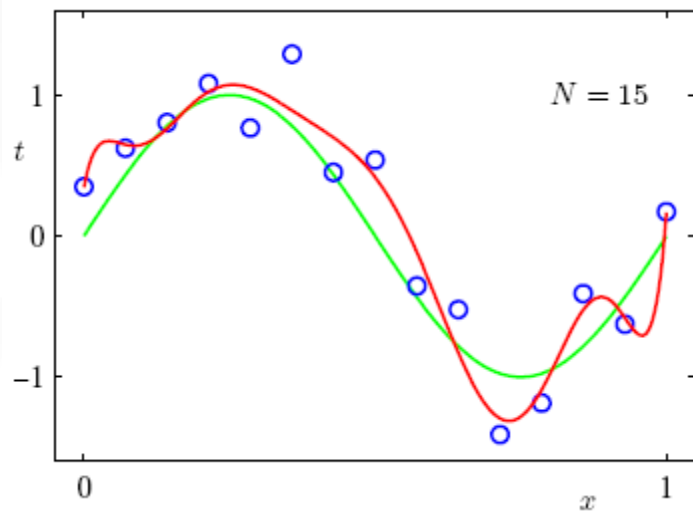
□ 过拟合



	$M = 0$	$M = 1$	$M = 3$	$M = 9$
w_0^*	0.19	0.82	0.31	0.35
w_1^*		-1.27	7.99	232.37
w_2^*			-25.43	-5321.83
w_3^*			17.37	48568.31
w_4^*				-231639.30
w_5^*				640042.26
w_6^*				-1061800.52
w_7^*				1042400.18
w_8^*				-557682.99
w_9^*				125201.43

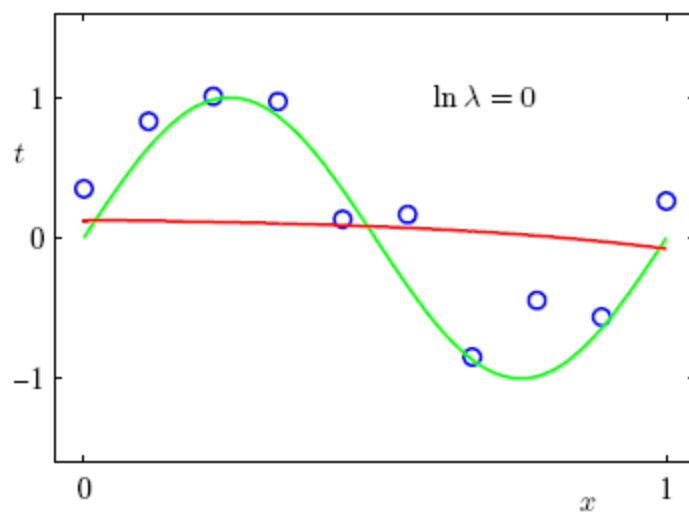
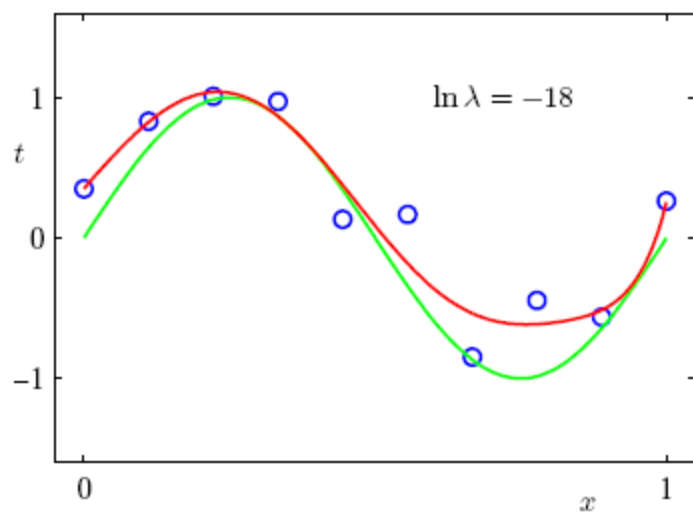
例：多项式曲线拟合问题

□ 增加数据量



例：多项式曲线拟合问题

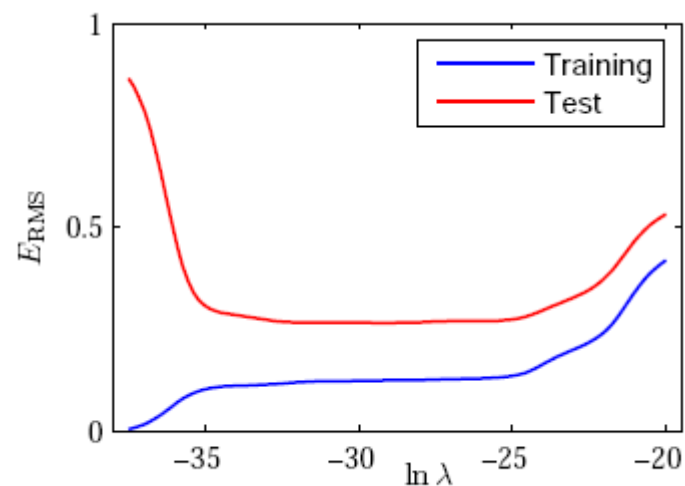
□ 正则化：引入惩罚项



例：多项式曲线拟合问题

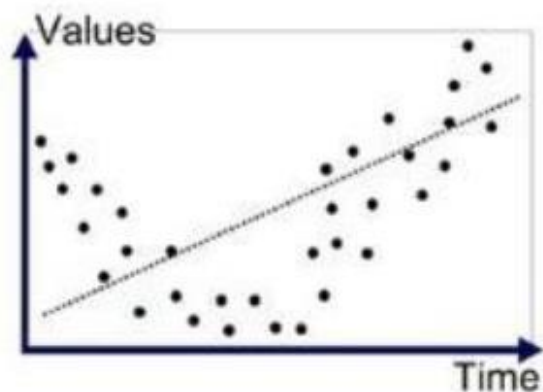
□ 引入惩罚项

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

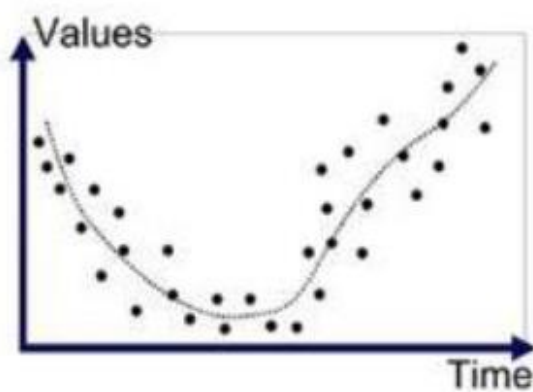


过拟合和欠拟合

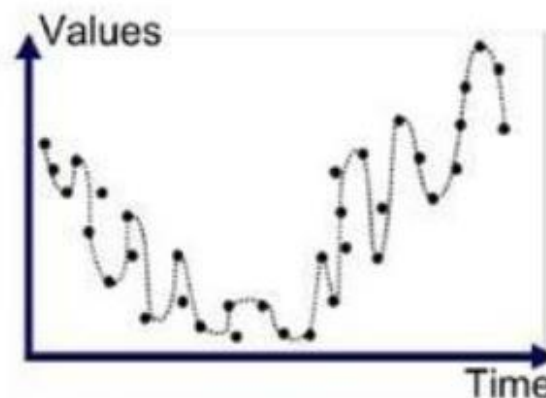
- ❑ 过拟合（泛化能力弱）：训练误差低，测试误差高
- ❑ 欠拟合：训练误差高



Underfitted



Good Fit/Robust



Overfitted

模型选择：正则化

□ 正则化：降低模型复杂度，减少测试误差

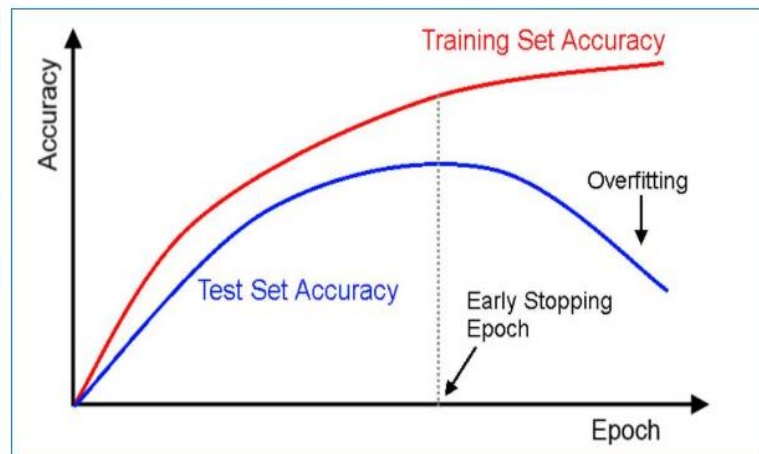
- 惩罚项、训练集增强
- Dropout、earlystopping

□ 范数：满足以下条件的函数

$f: \mathbf{R}^n \rightarrow \mathbf{R}$ 称为范数

- (1) 非负的： $\forall \mathbf{x} \in \mathbf{R}^n, f(\mathbf{x}) \geq 0$
- (2) 正定的： $f(\mathbf{x}) = 0 \Leftrightarrow \mathbf{x} = 0$
- (3) 齐次的： $\forall \mathbf{x}, t, f(t\mathbf{x}) = |t|f(\mathbf{x})$
- (4) 三角不等式： $\forall \mathbf{x}, \mathbf{y} \in \mathbf{R}^n, f(\mathbf{x} + \mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$

范数是欧式空间向量长度的推广，使用 $f(\mathbf{x}) = \|\mathbf{x}\|$



模型选择：正则化

□ 向量范数

l_p 范数 ($p \geq 1$):
$$\|\mathbf{x}\|_p = \sqrt[p]{\sum_{i=1}^n |x_i|^p}$$

l_1 范数 (曼哈顿距离):
$$\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$$

l_2 范数 (欧氏距离):
$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$$

l_∞ 无穷范数:
$$\|\mathbf{x}\|_\infty = \max_{i \in 1..n} |x_i|$$

l_0 零范数:
$$\|\mathbf{x}\|_0 = \#(i | x_i \neq 0)$$



$p = \infty$



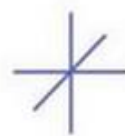
$p = 2$



$p = 1$



$0 < p < 1$



$p = 0$

模型选择：正则化

□ 矩阵范数： $\|A\| = \max_{\|\mathbf{x}\| \neq 0} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$

l_1 范数：

$$\|A\|_1 = \max_j \sum_i |a_{ij}|$$

l_2 范数：

$$\|A\|_2 = \sqrt{\max_i |\lambda_i|}, \text{ 其中 } \lambda_i \text{ 为 } A^T A \text{ 的特征值}$$

l_∞ 范数：

$$\|A\|_\infty = \max_i \sum_j |a_{ij}|$$

l_0 范数：

$$\|A\|_0 = \sum_i \sum_j \mathbb{I}[a_{ij} \neq 0]$$

F-范数：

$$\|A\|_F = \sqrt{\sum_i \sum_j a_{ij}^2}$$

模型选择：正则化

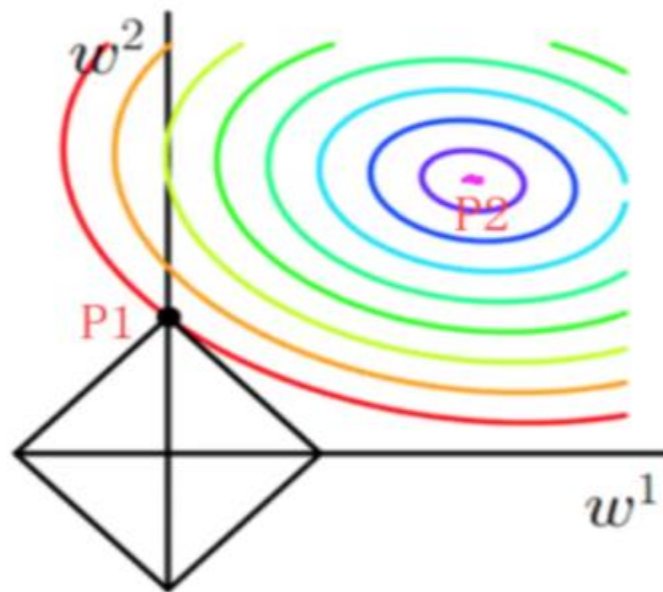
□ 目标函数： $\bar{J}(\omega, b) = J(\omega, b) + \frac{\lambda}{2N} \Omega(\omega)$

□ L^1 正则化： $\Omega(\omega) = \|\omega\|_1$

假设数据只有两个特征即， ω_1, ω_2

$$\bar{J}(\omega_1, \omega_2) = J(\omega_1, \omega_2) + \frac{\lambda}{2N} (|\omega_1| + |\omega_2|)$$

$$\begin{aligned} L^1 \text{ 正则化 } \omega_1 &:= \omega_1 - \alpha \frac{d\bar{J}}{d\omega_1} \\ &= \omega_1 - \frac{dJ}{d\omega_1} - \frac{\alpha\lambda}{2N} \text{sign } \omega_1 \end{aligned}$$

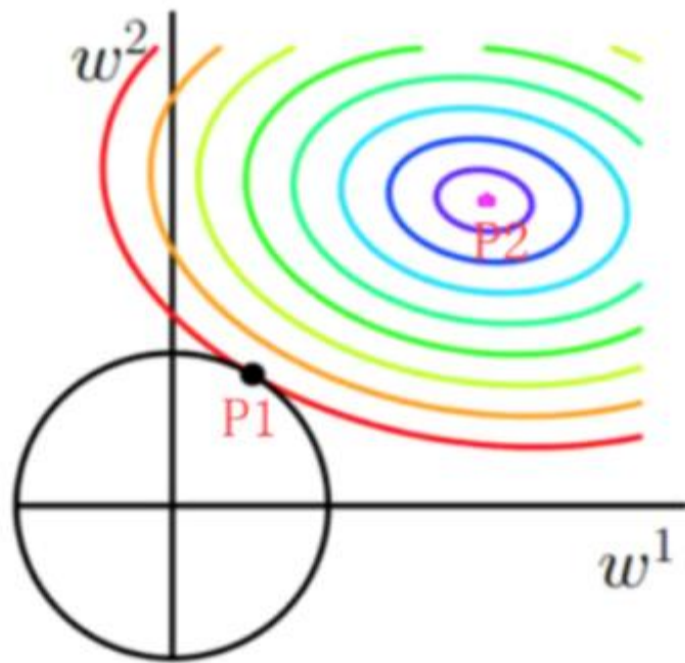


模型选择：正则化

□ L^2 正则化 : $\Omega(\omega) = \|\omega\|_2^2$

$$\bar{J}(\omega_1, \omega_2) = J(\omega_1, \omega_2) + \frac{\lambda}{2N}(\omega_1^2 + \omega_2^2)$$

$$\begin{aligned} L^2 \text{ 正则化 } \omega_1 &:= \omega_1 - \alpha \frac{d\bar{J}}{d\omega_1} \\ &= \omega_1 - \frac{dJ}{d\omega_1} - \frac{\alpha\lambda}{N} \omega_1 \\ &= \left(1 - \frac{\alpha\lambda}{N}\right) \omega_1 - \frac{dJ}{d\omega_1} \end{aligned}$$

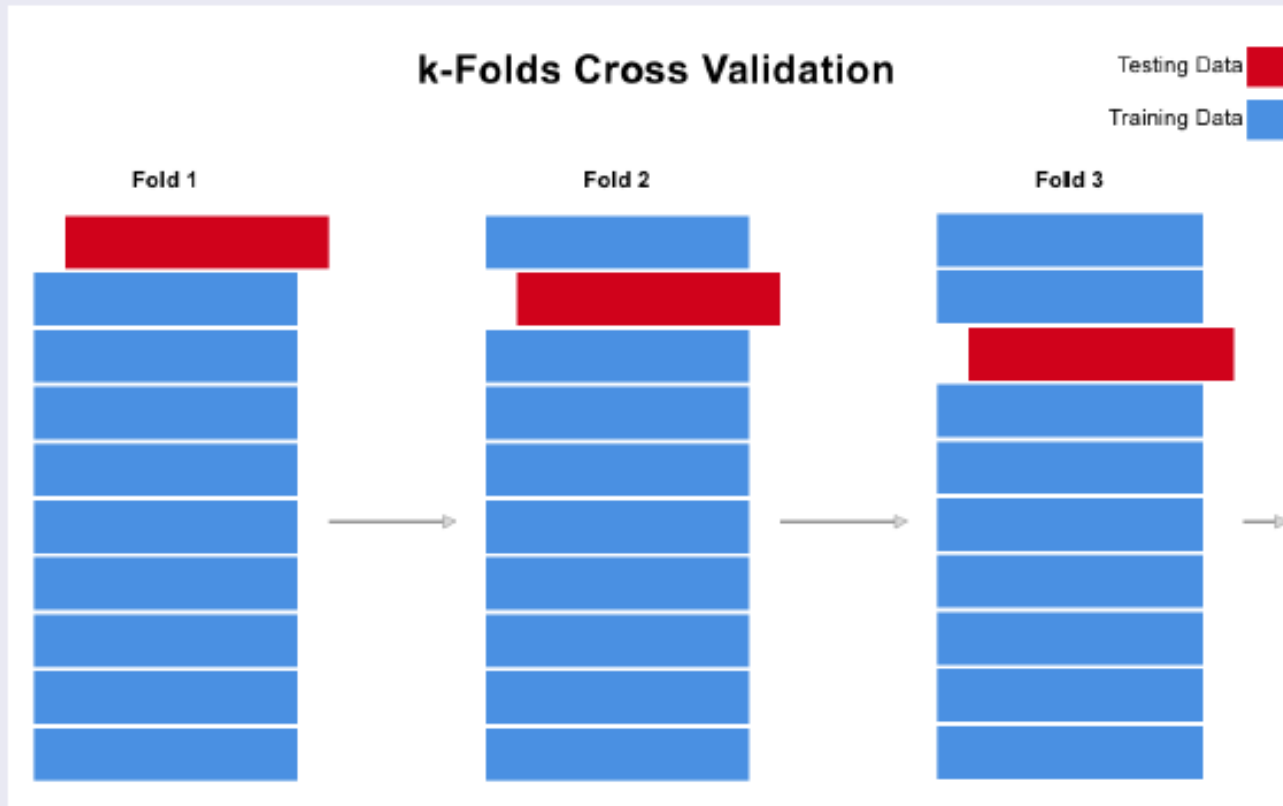


模型选择：交叉验证

- 将数据分成训练集、验证集、测试集
- 交叉验证：重复利用数据
 - 简单交叉验证：从训练集中随机分成训练数据和测试数据，对不同的模型进行测试，选择最优的。
 - K折交叉验证
 - 随机将已给数据分成K个互不相交、大小相同的子集
 - 利用K-1个子集的数据训练模型，余下的子集测试模型；
 - 将这一过程对可能的K种选择重复进行
 - 最后选个K次评测中平均测试误差最小的模型
 - 留一交叉验证：K=N

K折交叉验证

k -fold cross-validation



过拟合

学习	开车
过拟合	车祸
VC维太大（模型过于复杂）	车开太快
数据有noise	道路崎岖
数据量不够	对路况的了解程度不够
从简单模型开始	先慢慢开
数据清洗/修剪	使用更加精确的道路信息
数据提示	利用更多的道路信息
正则化	踩刹车
验证	观察仪表盘
特征转换	踩油门

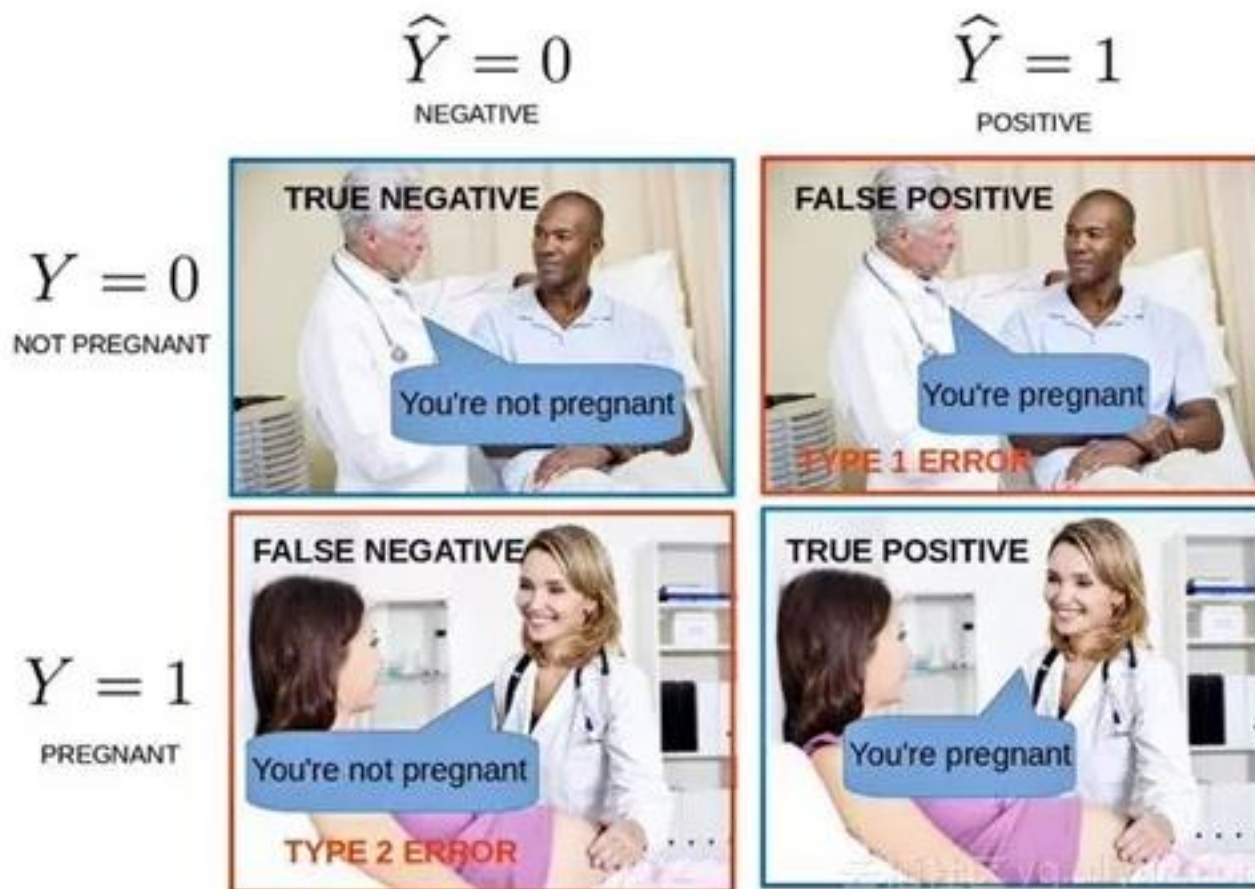
模型评估指标：分类问题

□ 混淆矩阵

- 真正 (True Positive, TP)
- 假正 (False Positive, FP)
- 假负 (False Negative, FN)
- 真负 (True Negative, TN)

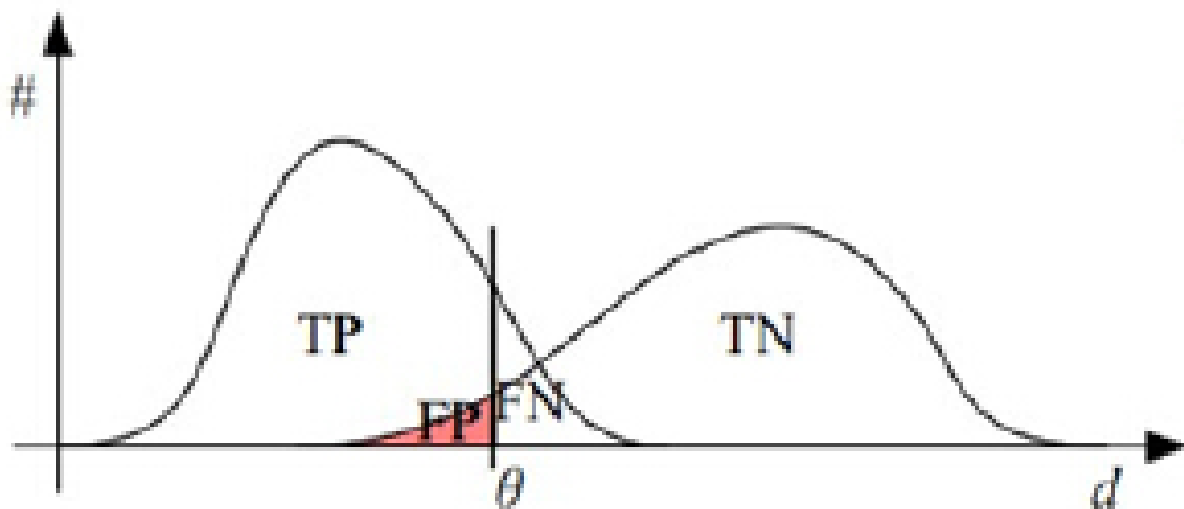
预测值 实际值	Positive	Negative
正	TP	FN
负	FP	TN

模型评估指标：分类问题



模型评估指标：分类问题

- ❑ 真正率 (True Positive Rate) $TPR = TP / (TP + FN)$
- ❑ 假正率 (False Positive Rate) $FPR = FP / (FP + TN)$
- ❑ 假负率 (False Negative Rate) $FNR = FN / (TP + FN)$
- ❑ 真负率 (True Negative Rate) $TNR = TN / (FP + TN)$



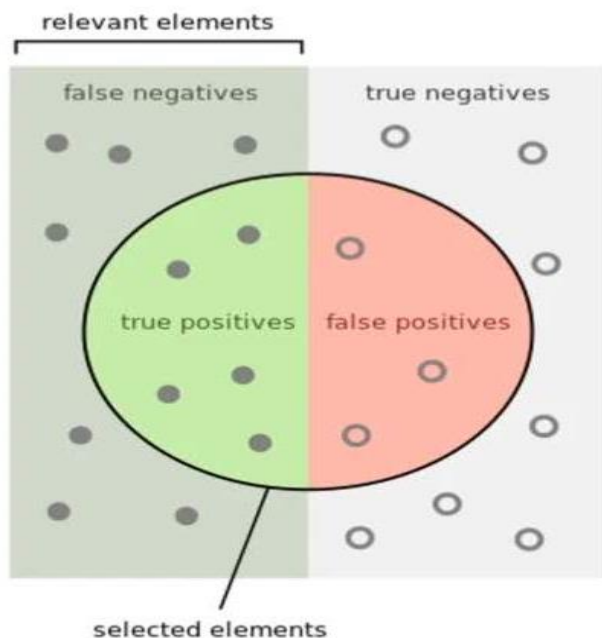
模型评估指标：分类问题

❑ 查全率、召回率 (Recall)、灵敏度

■ $\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$

❑ 查准率、精确率 (Precision)

■ $\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$



How many selected items are relevant?

Precision = $\frac{\text{true positives}}{\text{true positives} + \text{false positives}}$

How many relevant items are selected?

Recall = $\frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$

模型评估指标：分类问题

□ Precision-Recall curve

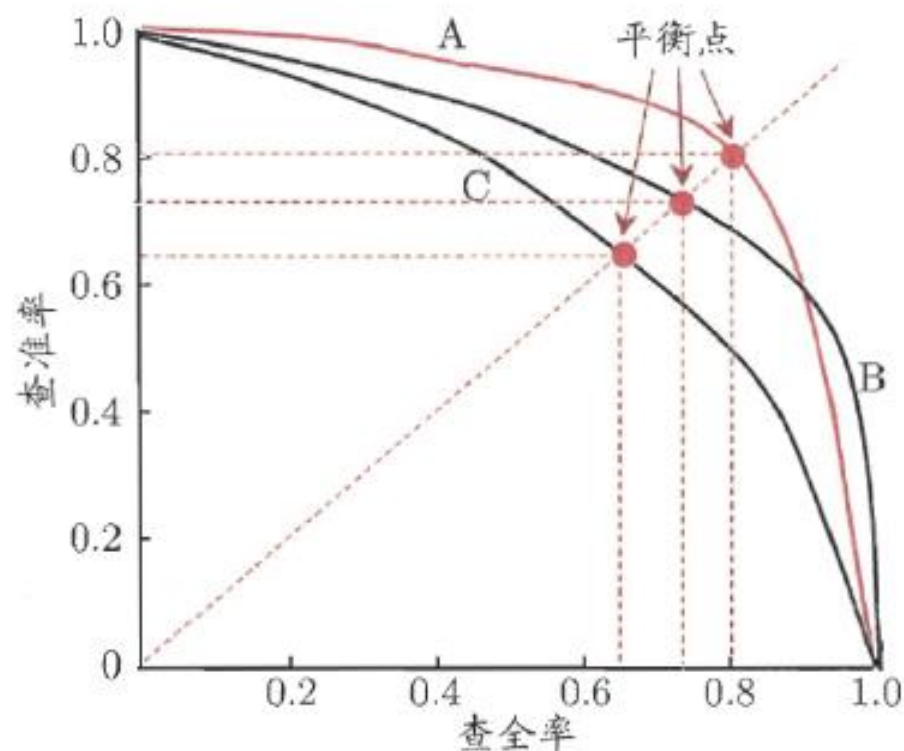
□ 漏警率：MA=1-R

□ 虚警率：FA=1-P

□ F1值和 F_β 值：

■ $F_1 = \frac{2 \times P \times R}{P + R}$

■ $F_\beta = \frac{(1 + \beta^2) \times P \times R}{\beta^2 P + R}$



模型评估指标：分类问题

□ 准确率 (Accuracy)

■ $ACC = (TP+TN)/(TP+TN+FP+FN)$

□ 错误率 (Error rate)

■ $Error = 1 - ACC$

□ 计算速度：分类器训练和预测需要的时间

□ 鲁棒性：处理缺失值和异常值的能力

□ 可扩展性：处理大数据集的能力

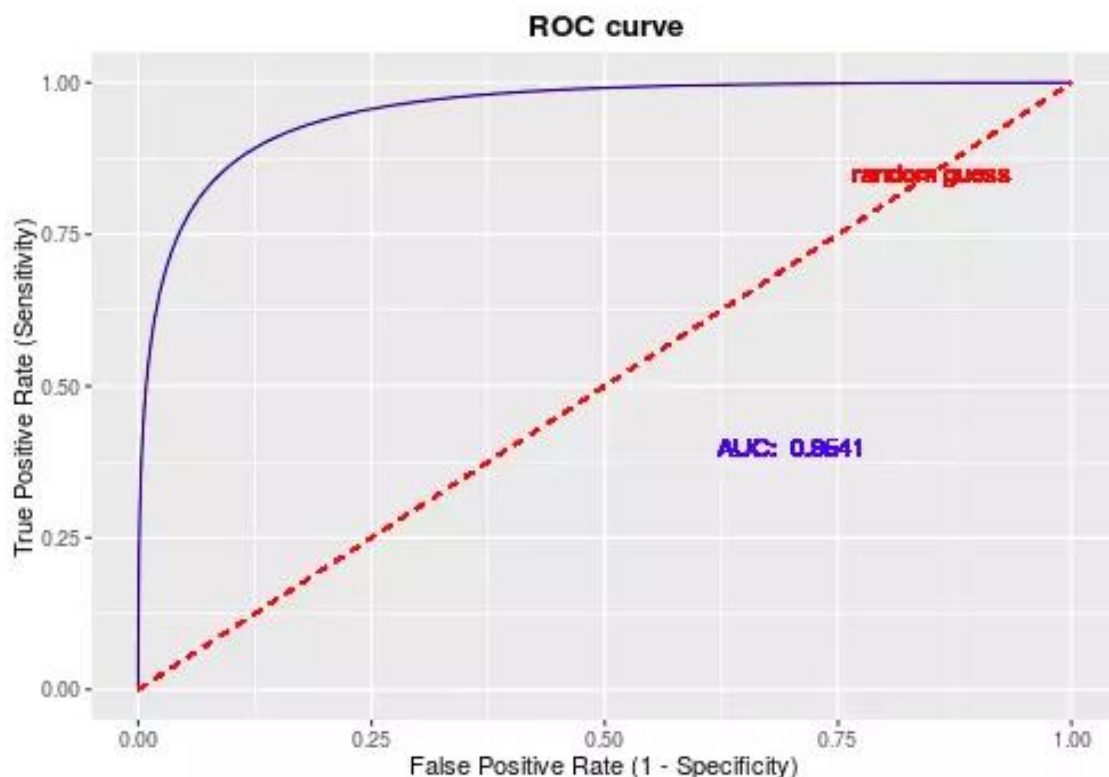
□ 可解释性：分类器的预测标准的可理解性

模型评估指标：分类问题

□ ROC (Receiver operation characteristic) 曲线

■ $TPR = TP / (TP + FN)$: 灵敏度 (Sensitivity)

■ $FPR = FP / (TN + FP)$: 1-特异度 (Specificity)



模型评估指标：分类问题

□ AUC值：ROC曲线下的面积

□ Wilcoxon-Mann-Witney Test

- Score：表示每个测试样本属于正样本的概率
- 任意给一个正样本和负样本，正样本的Score大于负样本Score的概率。（AUC值）
- AUC值越大，正样本的Score值越有可能大于负样本的值，从而能够更好地分类

□ 计算

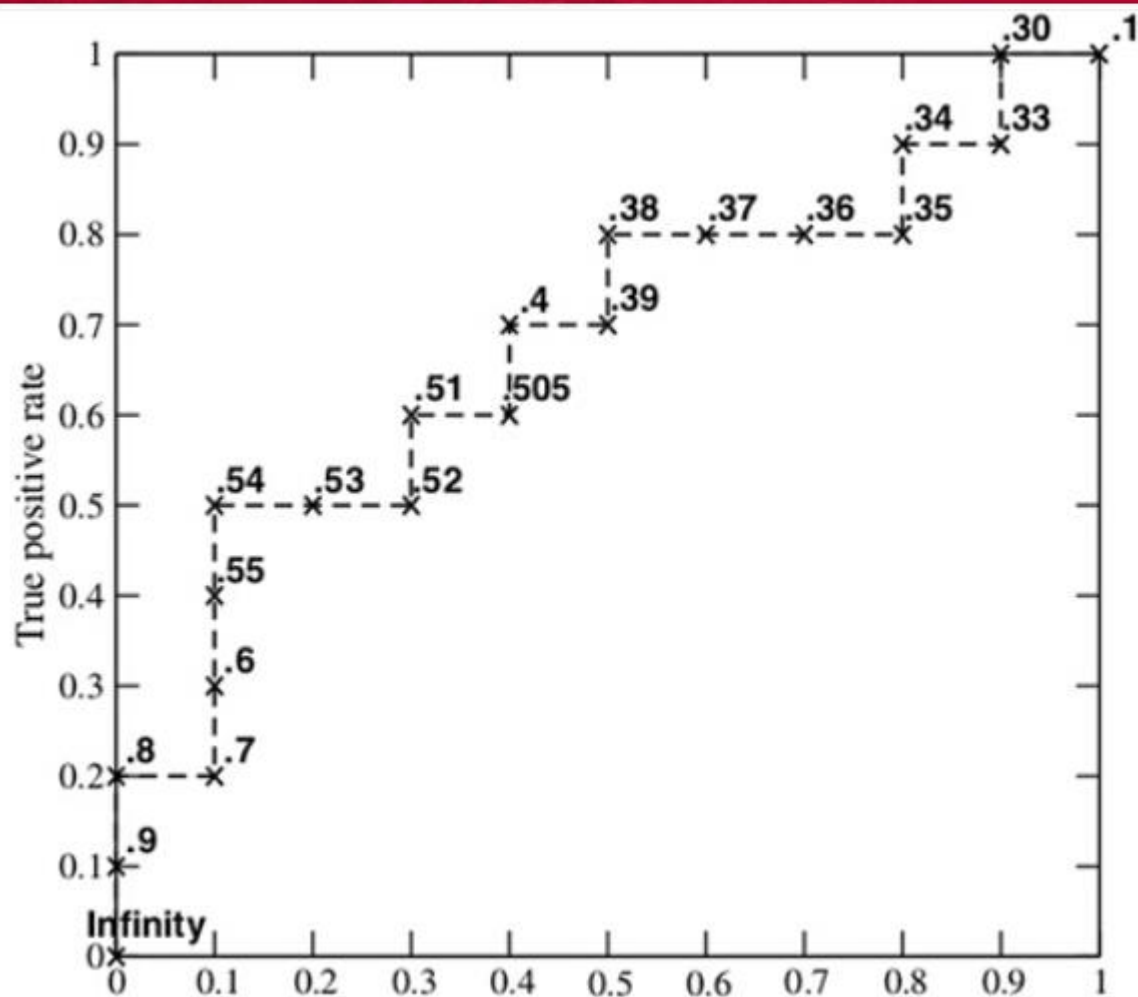
- 对Score从大到小排序，然后令最大Score对应的sample的rank为n，第二大score对应sample的rank为n-1，以此类推

$$AUC = \frac{\sum_{i \in \text{positiveClass}} \text{rank}_i - \frac{M(1+M)}{2}}{M \times N}$$

模型评估指标：分类问题

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

模型评估指标：分类问题



□ AUC=0.68

模型评估指标：回归问题

□ 平均绝对误差 (MAE)

■ $MAE = \frac{1}{N} \sum_i |\hat{y}_i - y_i|$

□ 均方误差MSE

■ $MSE = \frac{1}{N} \sum_i (\hat{y}_i - y_i)^2$

□ 均方根误差RMSE

■ $RMSE = \sqrt{\frac{1}{N} \sum_i (\hat{y}_i - y_i)^2}$



THE END