

数据挖掘与机器学习

潘斌

panbin@nankai.edu.cn

范孙楼227

1

上节回顾

- 概念学习的求解方法
 - Find-S
 - 列表后消除算法
 - 候选消除算法
 - 奥卡姆剃刀原理和没有免费午餐原理
- 模型的评估



本节提要

- 模型的评估与选择
- 过拟合
- 分类的基本概念



- 训练误差 (training error) 和泛化误差 (generalization error)
- 训练误差：关于训练数据集的平均损失

$$R_{\text{emp}}(\hat{f}) = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}(x_i))$$

- 泛化误差：对未知数据预测的误差

$$R_{\text{exp}}(\hat{f}) = E_P[L(Y, \hat{f}(X))] = \int_{x \times y} L(y, \hat{f}(x)) P(x, y) dx dy$$



二分类问题的泛化误差上界

训练数据集 $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ 从联合概率分布 $P(X, Y)$ 独立同分布产生的, $X \in \mathbf{R}^n$, $Y \in \{-1, +1\}$. 设 f 是从 $\mathcal{F} = \{f_1, f_2, \dots, f_d\}$ 中选取的函数. 损失函数是 0-1 损失.

关于 f 的期望风险和经验风险分别是

$$R(f) = E[L(Y, f(X))]$$

$$\hat{R}(f) = \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i))$$



经验风险最小化函数是

$$f_N = \arg \min_{f \in \mathcal{F}} \hat{R}(f)$$

f_N 的泛化能力

$$R(f_N) = E[L(Y, f_N(X))]$$



定理 (泛化误差上界) 对二类分类问题, 当假设空间是有限个函数的集合 $\mathcal{F} = \{f_1, f_2, \dots, f_d\}$ 时, 对任意一个函数 $f \in \mathcal{F}$, 至少以概率 $1 - \delta$, 以下不等式成立:

训练误差小
的模型泛化
误差也会小。

$$R(f) \leq \hat{R}(f) + \varepsilon(d, N, \delta)$$

\mathcal{F} 中包含的函数越多,
泛化误差上界越大。

样本容量 N 越
大, 训练误差
与泛化误差越
接近。

其中,

$$\varepsilon(d, N, \delta) = \sqrt{\frac{1}{2N} \left(\log d + \log \frac{1}{\delta} \right)}$$

Hoeffding 不等式

设 $S_n = \sum_{i=1}^n X_i$ 是独立随机变量 X_1, X_2, \dots, X_n 之和, $X_i \in [a_i, b_i]$, 则对任意 $t > 0$,

随机变量的和
与其期望值偏
差的概率上界

以下不等式成立:

$$P(ES_n - S_n \geq t) \leq \exp \left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2} \right)$$

(1)

影响上界的因
素: 数量多,
差异大

- 泛化误差的估计
 - 用训练误差 (training error) 估计
 - 其他理论方法?
- 验证集 (validating set) 与验证误差 (validating error)
 - 假定从真实分布中独立采样得到
 - 与训练集互斥



ID	X1	X2	Y
1	A	7	T
2	A	7	T
3	B	5	T
4	A	3	F
5	A	2	F
6	A	6	F
7	A	7	F
8	A	7	T
9	B	2	T
10	A	3	F

ID	X1	X2	Y
1	A	7	T
2	A	7	T
4	A	3	F
6	A	6	F
9	B	2	T
10	A	3	F

ID	X1	X2	Y
3	B	5	T
5	A	2	F
7	A	7	F
8	A	7	T

IF X1 = B OR X2 > 6
THEN Y = T

ID	X1	X2	Y
3	B	5	T
5	A	2	F
7	A	7	T
8	A	7	T

验证误差：0.25



准确率的局限性

- 考虑一个分类问题：
 - 类0中的样本数为 9990
 - 类1中的样本数为 10
 - 若模型将所有样例类别预测为类0，则分类准确率为 $9990/10000 = 99.9\%$
 - 模型没有发现任何类1的样例，因而准确率具有误导性（样本不均衡）。



混淆矩阵 (CONFUSION MATRIX)

		PREDICTED CLASS	
		Class=Yes	Class>No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class>No	c (FP)	d (TN)

- a: TP (true positive) 真正，对应于被分类模型正确预测的正样本数
- b: FN (false negative) 假负，对应于被分类模型错误预测为负类的正样本数
- c: FP (false positive) 假正，对应于被分类模型错误预测为正类的负样本数
- d: TN (true negative) 真负，对应于被分类模型正确预测的负样本数



	PREDICTED CLASS	
ACTUAL CLASS	Class=Yes	Class>No
	Class=Yes	a (TP)
Class>No	c (FP)	d (TN)

- 真正率 (TPR) 或称灵敏度 (Sensitivity)
 - 模型正确预测的正样本的比例
 - $TPR = TP / (TP + FN) = a / (a + b)$
- 真负率 (TNR) 或称特指度 (Specificity)
 - 模型正确预测的负样本的比例
 - $TNR = TN / (TN + FP) = d / (c + d)$
- 假正率 (FPR)
 - 被预测为正类的负样本的比例
 - $FPR = FP / (TN + FP) = c / (c + d)$
- 假负率 (FNR)
 - 被预测为负类的正样本的比例
 - $FNR = FN / (TP + FN) = b / (a + b)$



	PREDICTED CLASS		
ACTUAL CLASS	Class=Yes	Class>No	
	Class=Yes	a (TP)	b (FN)
	Class>No	c (FP)	d (TN)

- 精度 p: $Precision (p) = \frac{a}{a + c}$

- 也称查准率，确定在分类器断言为正类的那部分记录中实际为正类的记录所占的比率。

- 精度越高，分类器的假正错误率就越低

- 召回率 r: $Recall (r) = \frac{a}{a + b}$

- 度量被分类器正确预测的正样本的比例，亦称查全率。

- 具有高召回率的分类器很少将正样本误分为负样本

- F_1 度量: $F_1\text{-measure} (F) = \frac{2rp}{r + p} = \frac{2a}{2a + b + c}$

- 精度和召回率的调和平均



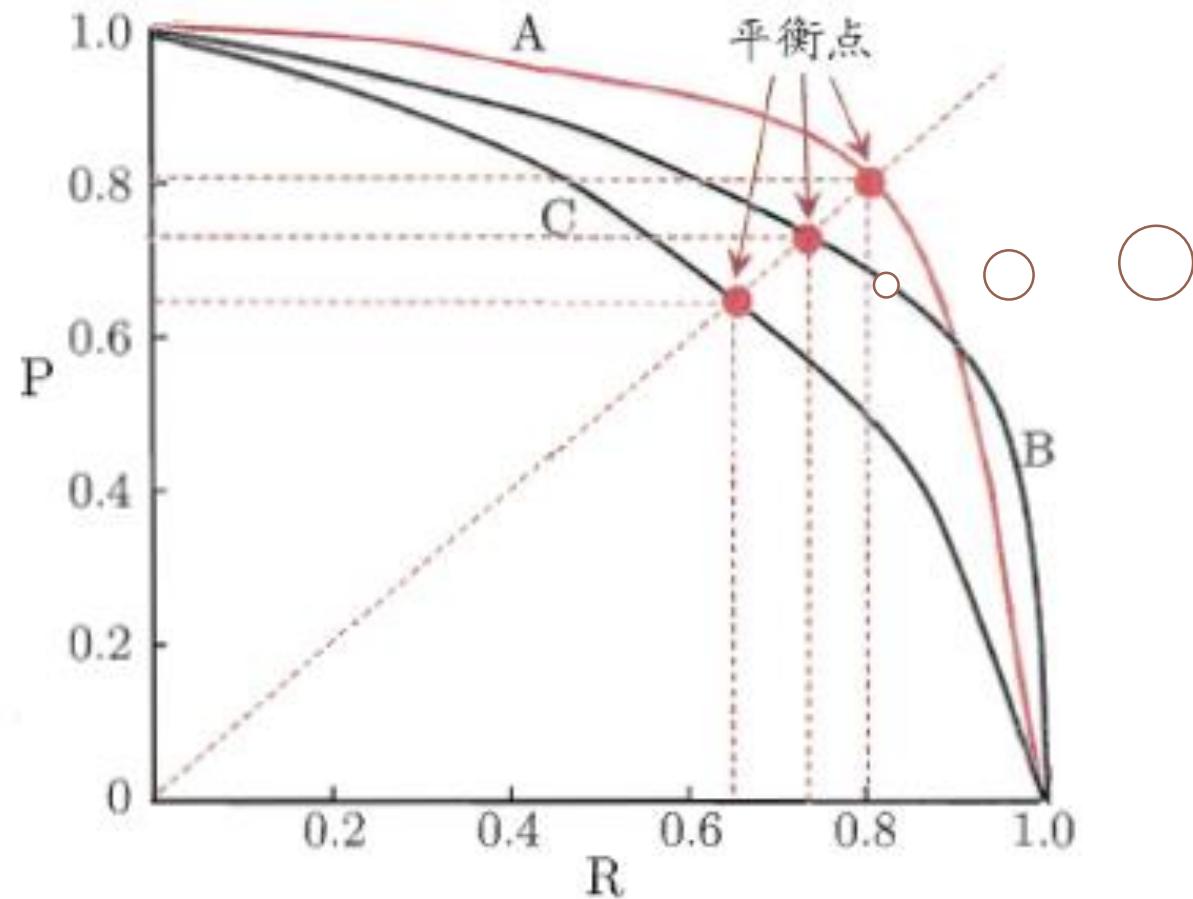
P-R 曲 线 和 P-R 图 - 评估某分类器的能力

- 根据学习器的预测结果对样例进行排序（单学习器）
 - “最可能”是正例的样本排在最前面；
 - “最不可能”是正例的样本排在最后面；
- 按此顺序逐个将样本作为正例进行预测，每次计算 p 、 r 值
- 以 p 为纵轴， r 为横轴作图，即得“P-R 曲线”
- 显示该曲线的图称为“P-R 图”



P-R 曲 线 和 P-R 图

	PREDICTED CLASS	
ACTUAL CLASS	Class=Yes	Class>No
	a (TP)	b (FN)
	c (FP)	d (TN)



P-R曲线与平衡点示意图

学习器A
优于学
习器C

- 精度 = $a/(a+c)$
- 召回率 = $a/(a+b)$



多个二分类混淆矩阵

- 各混淆矩阵上度量的平均值

$$\text{macro-}P = \frac{1}{n} \sum_{i=1}^n P_i ,$$

$$\text{macro-}R = \frac{1}{n} \sum_{i=1}^n R_i ,$$

$$\text{macro-}F1 = \frac{2 \times \text{macro-}P \times \text{macro-}R}{\text{macro-}P + \text{macro-}R} .$$



- 将混淆矩阵对应元素平均后再求指标值

$$\text{micro-}P = \frac{\overline{TP}}{\overline{TP} + \overline{FP}},$$

$$\text{micro-}R = \frac{\overline{TP}}{\overline{TP} + \overline{FN}},$$

$$\text{micro-}F1 = \frac{2 \times \text{micro-}P \times \text{micro-}R}{\text{micro-}P + \text{micro-}R}.$$

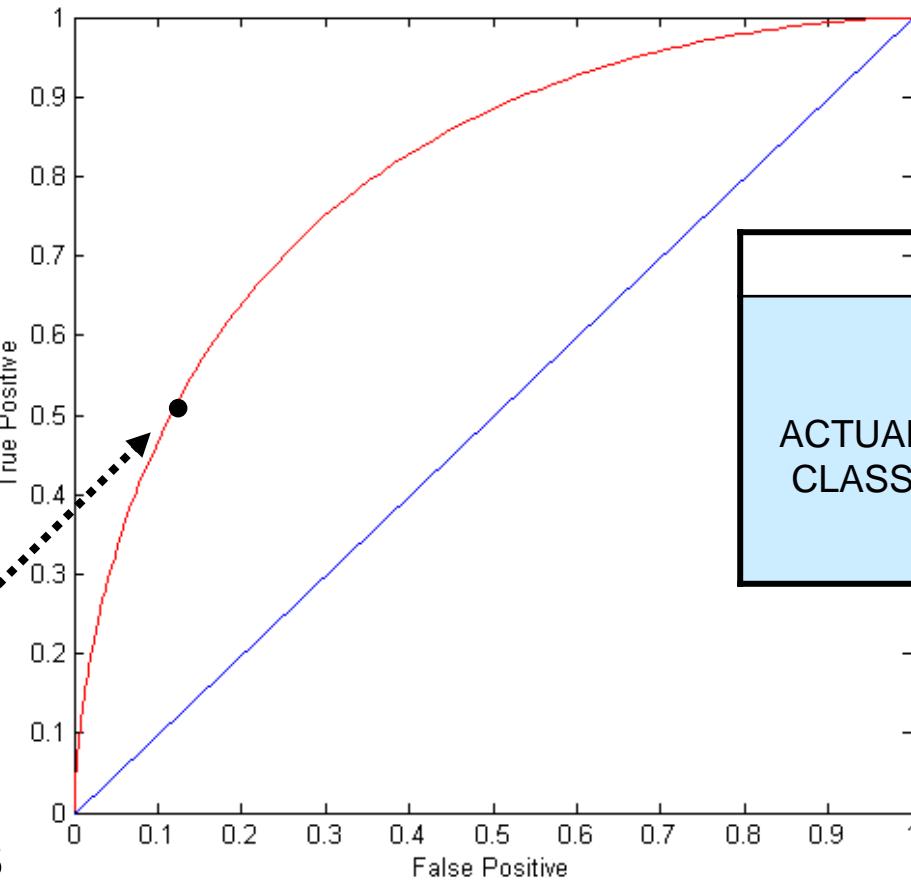
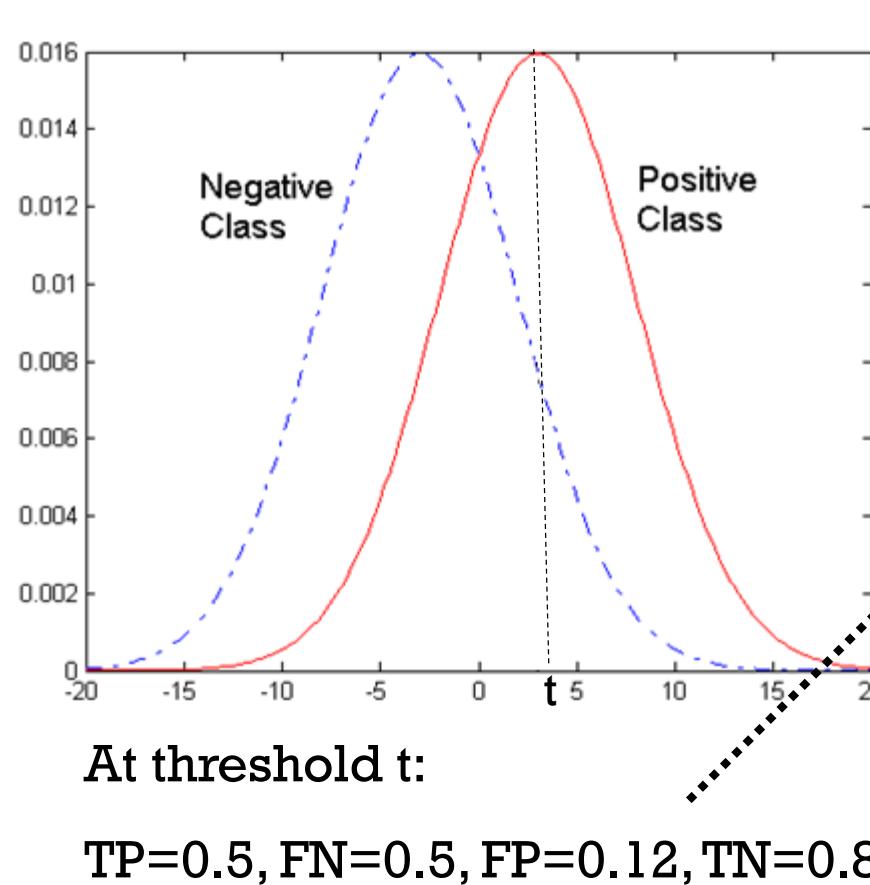


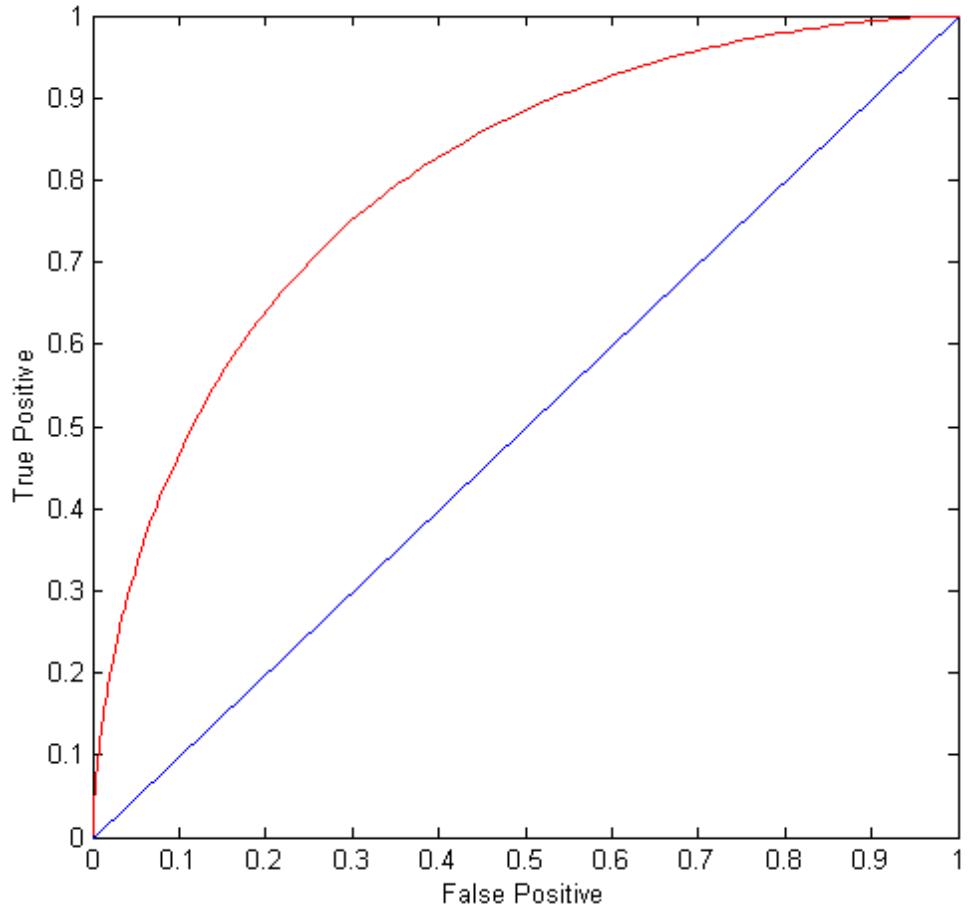
接受者操作特征 (RECEIVER OPERATING CHARACTERISTIC, ROC) 曲线 - 评估某模型能力

- 显示分类器真正率和假正率之间折中的一种图形化方法
- 真正率 (TPR) 沿 y 轴绘制 (召回率)
- 假正率 (FPR) 沿 x 轴绘制
- 沿着曲线的每个点对应于一个分类器归纳的模型



- 1-dimensional data set containing 2 classes (positive and negative)
- any points located at $x > t$ is classified as positive





■ ROC 曲线

- ($TPR=0, FPR=0$): 把每个实例都预测为负类的模型
- ($TPR=1, FPR=1$): 把每个实例都预测为正类的模型
- ($TPR=1, FPR=0$): 理想模型
- 对角线: 随机猜测的模型
- 靠近图左上角的分类器是优秀的



■ 如何绘制 ROC 曲线

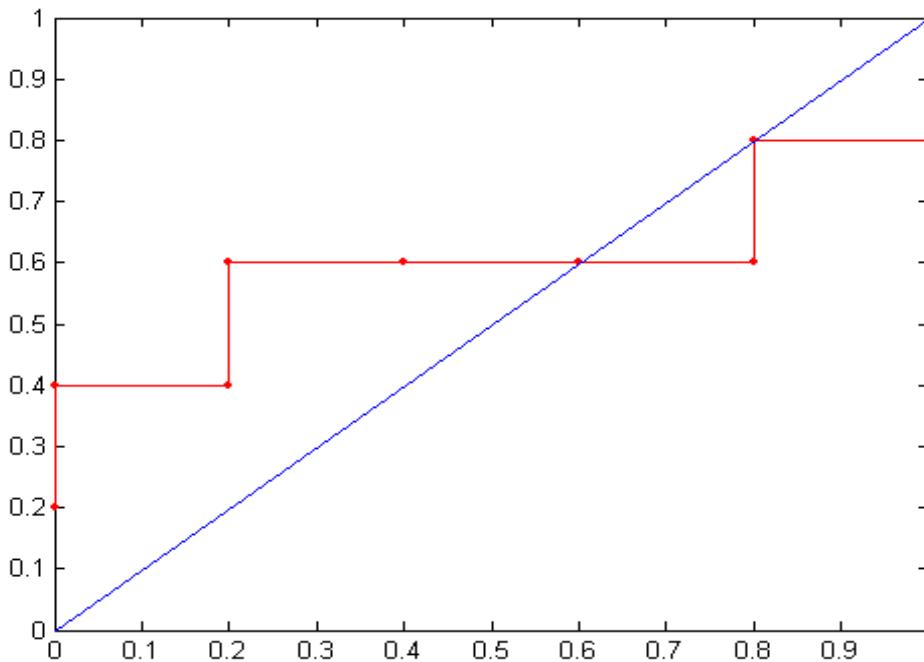
- (1) 假定为正类定义了连续值输出，对检验记录按它们的输出值递增排序。
- (2) 选择秩最低的检验记录（即输出值最低的记录），把选择的记录以及那些秩高于它的记录指派为正类。这种方法等价于把所有的检验实例都分为正类。因为所有的正检验实例都被正确分类，而所有的负测试实例都被误分，因此 $TPR=FPR=1$ 。
- (3) 从排序列表中选择下一个检验记录，把选择的记录以及那些秩高于它的记录指派为正类，而把那些秩低于它的记录指派为负类。通过考察前面选择的记录的实际类标号来更新 TP 和 FP 计数。如果前面选择的记录为正类，则 TP 计数减少而 FP 计数不变。如果前面选择的记录为负类，则 FP 计数减少而 TP 计数不变。
- (4) 重复步骤 3 并相应地更新 TP 和 FP 计数，直到最高秩的记录被选择。
- (5) 根据分类器的 FPR 画出 TPR 曲线。



- 对十个样本用不同阈值进行分类，5正5负

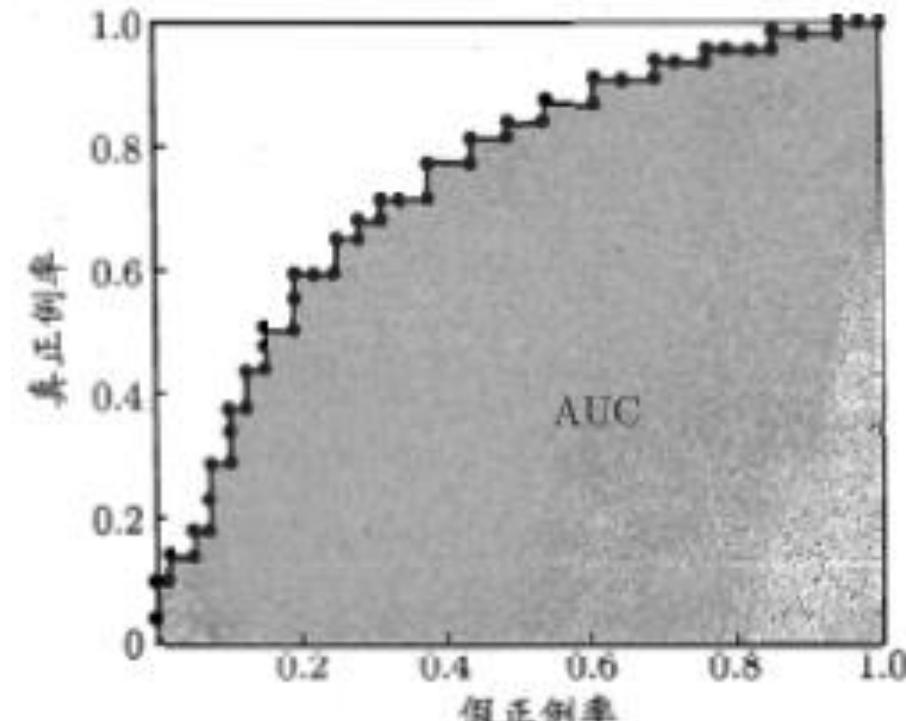
Class	+	-	+	-	-	-	+	-	+	+	+	
阈值	0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95	1.00	
TP	5	4	4	3	3	3	3	2	2	1	0	0
FP	5	5	4	4	3	2	1	1	0	0	0	0
TN	0	0	1	1	2	3	4	4	5	5	5	5
FN	0	1	1	2	2	2	2	3	3	4	5	5
TPR	1	0.8	0.8	0.6	0.6	0.6	0.6	0.4	0.4	0.2	0	
FPR	1	1	0.8	0.8	0.6	0.4	0.2	0.2	0	0	0	

ROC Curve:



AUC (AREA UNDER ROC CURVE)

- 即ROC曲线下的面积
- 可以评价模型的性能



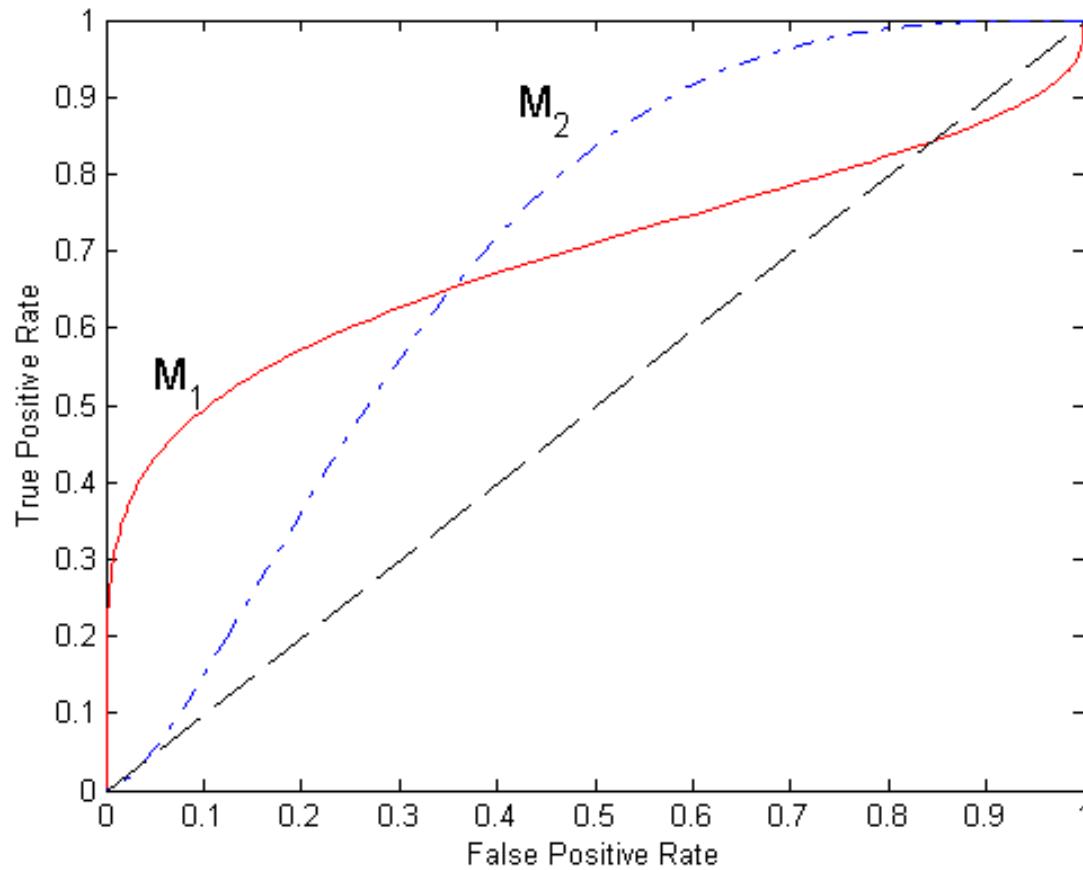
基于有限样例绘制的 ROC 曲线与 AUC

假定 ROC 曲线是由坐标为 $\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ 的点按序连接而形成($x_1 = 0, x_m = 1$)

$$AUC = \frac{1}{2} \sum_{i=1}^{m-1} (x_{i+1} - x_i) \cdot (y_i + y_{i+1}) .$$



- Using ROC for Model Comparison



- No model consistently outperform the other
 - M₁ is better for small FPR
 - M₂ is better for large FPR
- AUC



PR 曲 线 与 ROC 曲 线

- 例：计算下述二分类器的TPR, FPR, P, R

M1	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class>No
	Class=Yes	90	10
	Class>No	10	1,999,890

M2	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class>No
	Class=Yes	90	10
	Class>No	910	1,998,990

M1: TPR=0.9, FPR=0.00000500025, P=0.9
R=0.9

M2: TPR=0.9,
FPR=0.00045502275,
P=0.09
R=0.9



代价矩阵

		PREDICTED CLASS		
		C(i j)	Class=Yes	Class>No
ACTUAL CLASS	Class=Yes	C(Yes Yes)	C(No Yes)	
	Class>No	C(Yes No)	C(No No)	

$C(i|j)$: 预测 (误分类) – 一个j类记录为i类的代价 (医院诊断)



犯假负错误的代价是
犯假正错误的100倍

■ 计算分类代价

尽管模型M₂改善了准确率，但仍然较差。因这些改善是建立在增加代价更高的假负错误之上的。

ACTUAL CLASS	Cost Matrix	PREDICTED CLASS	
		C(i j)	+
+	-1	100	
-	1	0	

标准的准确度量趋向于M₂优于M₁

ACTUAL CLASS	Model M ₁	PREDICTED CLASS	
		+	-
+	220	30	
-	70	180	

$$\begin{aligned} \text{Accuracy} &= 80\% \\ \text{Cost} &= 2850 \end{aligned}$$

ACTUAL CLASS	Model M ₂	PREDICTED CLASS	
		+	-
+	210	40	
-	10	240	

$$\begin{aligned} \text{Accuracy} &= 90\% \\ \text{Cost} &= 3800 \end{aligned}$$



如何基于PR和AOC，开展模型的检验？

—— 留出法（HOLDOUT）

- 将原始数据集D划分为二不相交集合，一个作为训练集S，一个作为验证集T
- 在训练集上训练出模型
- 在验证集上评估模型的性能
- 二者的划分常依分析者的判断
 - 50-50
 - 2/3作为训练集，1/3作为验证集

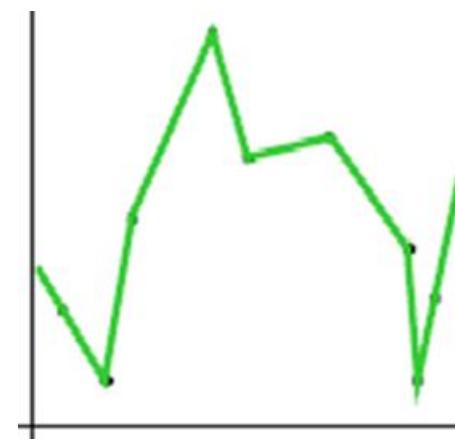
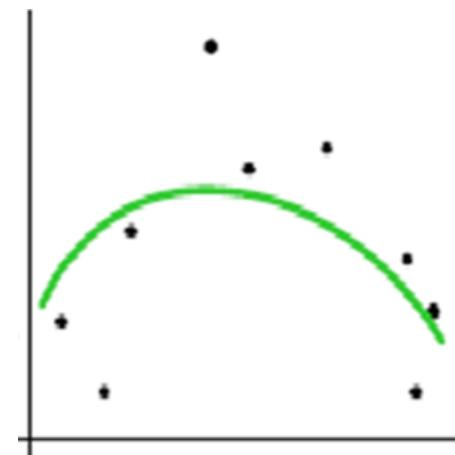
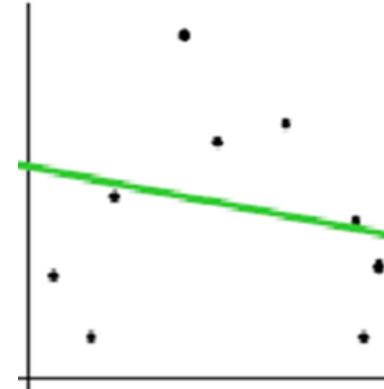
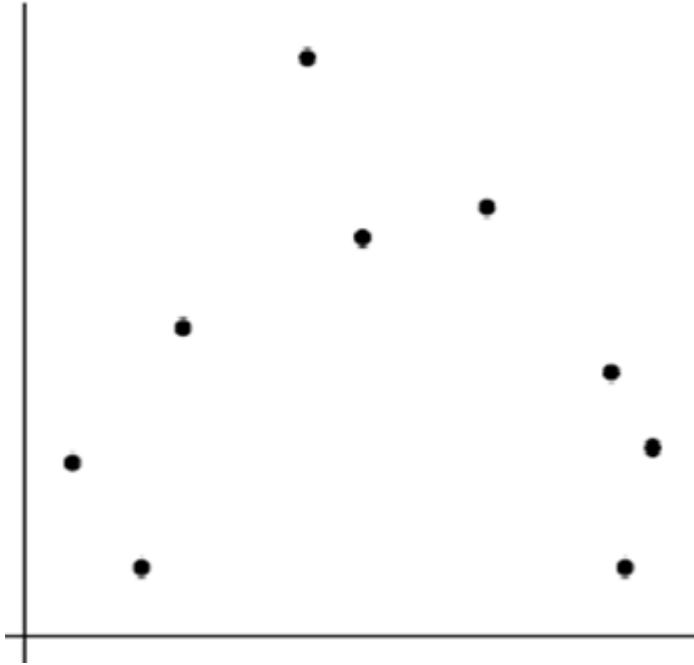


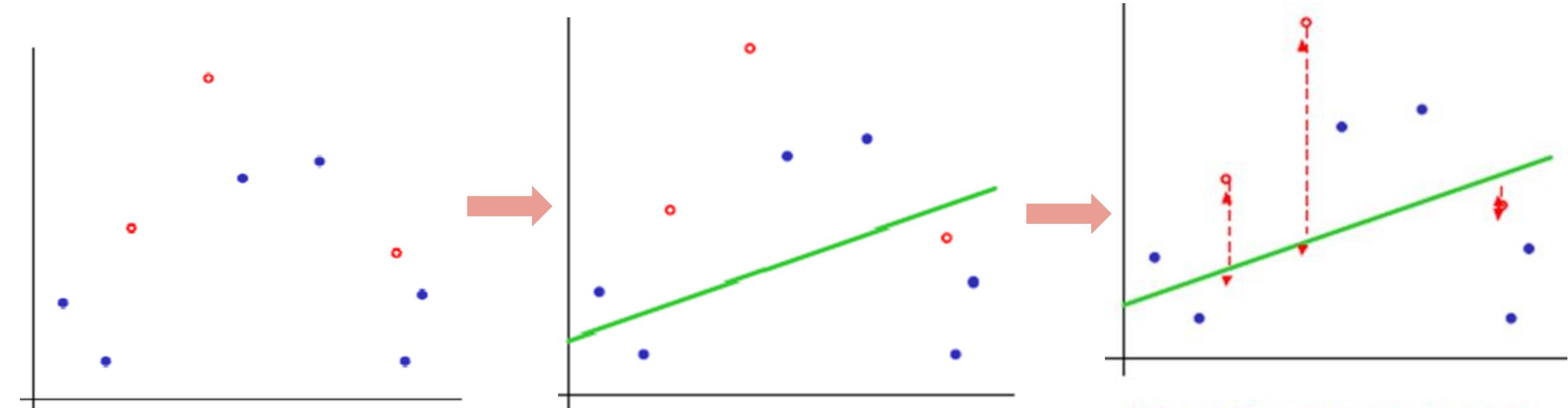
以二分类任务为例, 假定 D 包含 1000 个样本, 将其划分为 S 包含 700 个样本, T 包含 300 个样本, 用 S 进行训练后, 如果模型在 T 上有 90 个样本分类错误, 那么其错误率为 $(90/300) \times 100\% = 30\%$, 相应的, 精度为 $1 - 30\% = 70\%$.

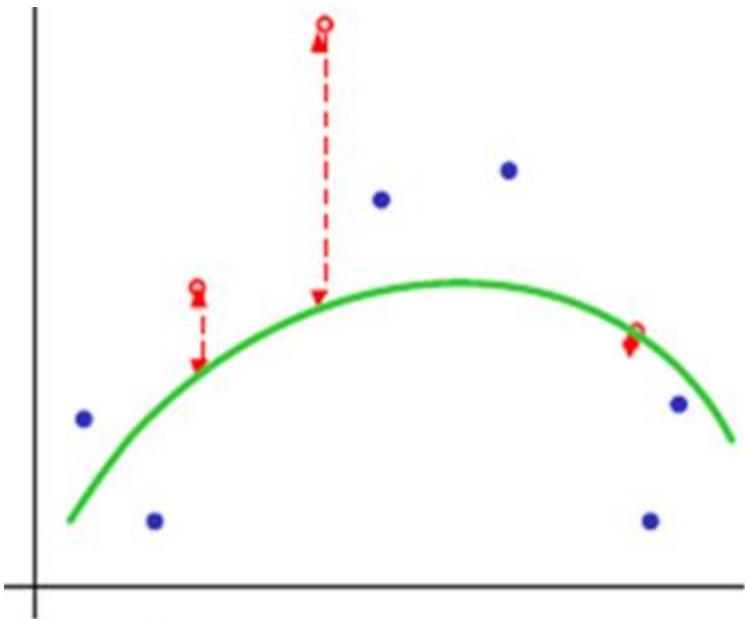
- 划分要尽可能保持原始数据分布的一致性
- 一般采用多次随机划分、重复评估后取平均值



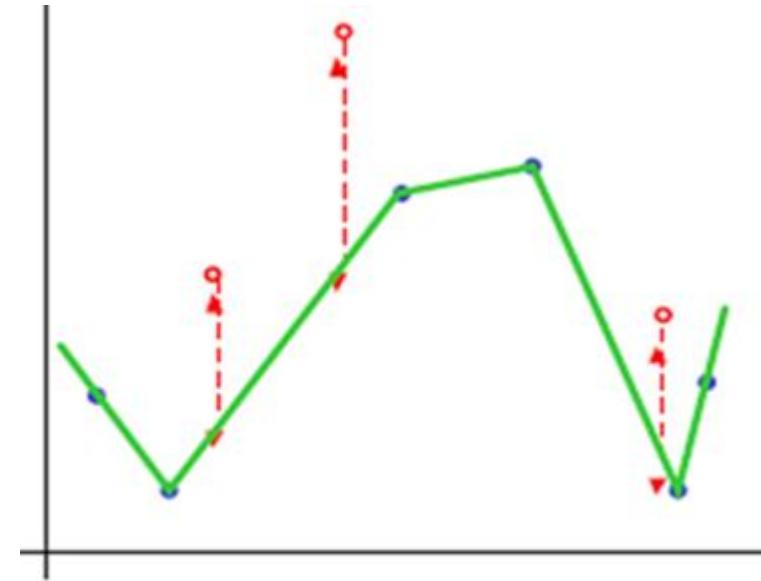
例：回归问题







Mean Squared Error = 0.9



Mean Squared Error = 2.2

■ 留一法 (Leave-One-Out Cross Validation)

- 记 L 为整个训练样本集的大小。
- 对于第 i 个训练样本，将其取出，对剩下 $L-1$ 个样本进行训练，得到模型，并用第 i 个训练样本对该模型的性能进行测试
- 该过程重复 L 次
- 用这 L 个模型的准确率的平均值作为评价模型性能的指标



$MSE_{LOOCV} = 2.12$

$MSE_{LOOCV} = 0.962$

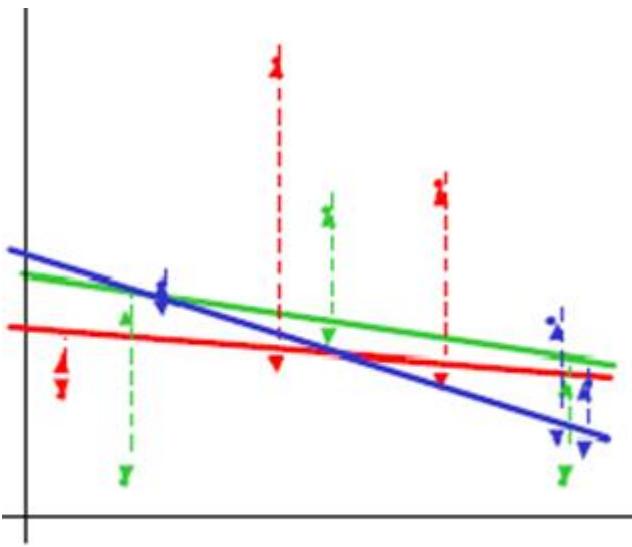
$MSE_{LOOCV} = 3.33$



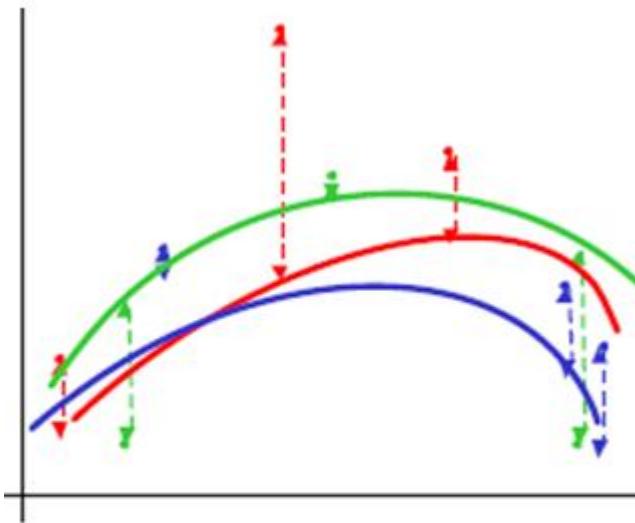
■ K 折交叉验证法 (K-fold Cross Validation)

- 将训练样本集随机分为 K 个集合，通常分为 K 等份
- 对其中的 $K-1$ 个集合进行训练，得到模型，并用剩下的一个集合对该模型的性能进行测试。
- 该过程重复 K 次，取 K 次过程中的测试错误的平均值作为评价模型性能的指标。

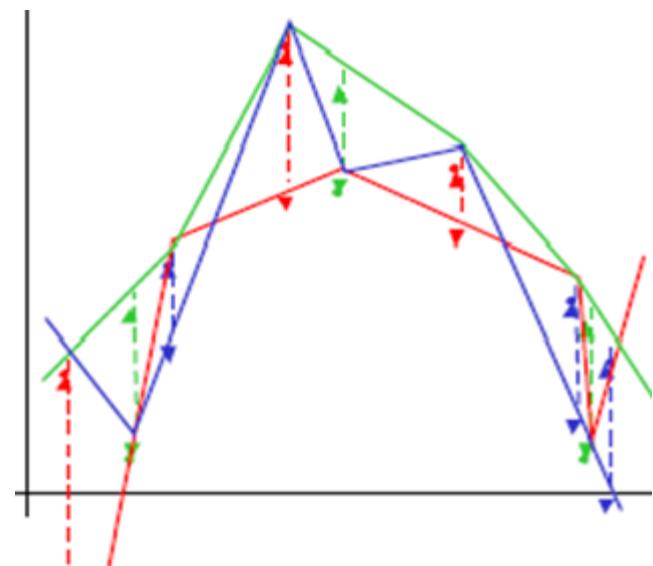




$MSE_{3FOLD}=2.05$



$MSE_{3FOLD}=1.11$



$MSE_{3FOLD}=2.93$



开展模型的检验 ——自助法（BOOTSTRAPPING）

- 以自助采样法为基础
 - 数据集D中包含m个样本
 - 每次随机从D中挑选一个样本到D'
 - 该样本被放回初始数据集D中
 - 重复m次得到包含m个样本的数据集D'
 - D'中不重复样本占多少比例？



PR/ROC+交叉验证是否足够了? ——不同模型（学习器）性能的比较

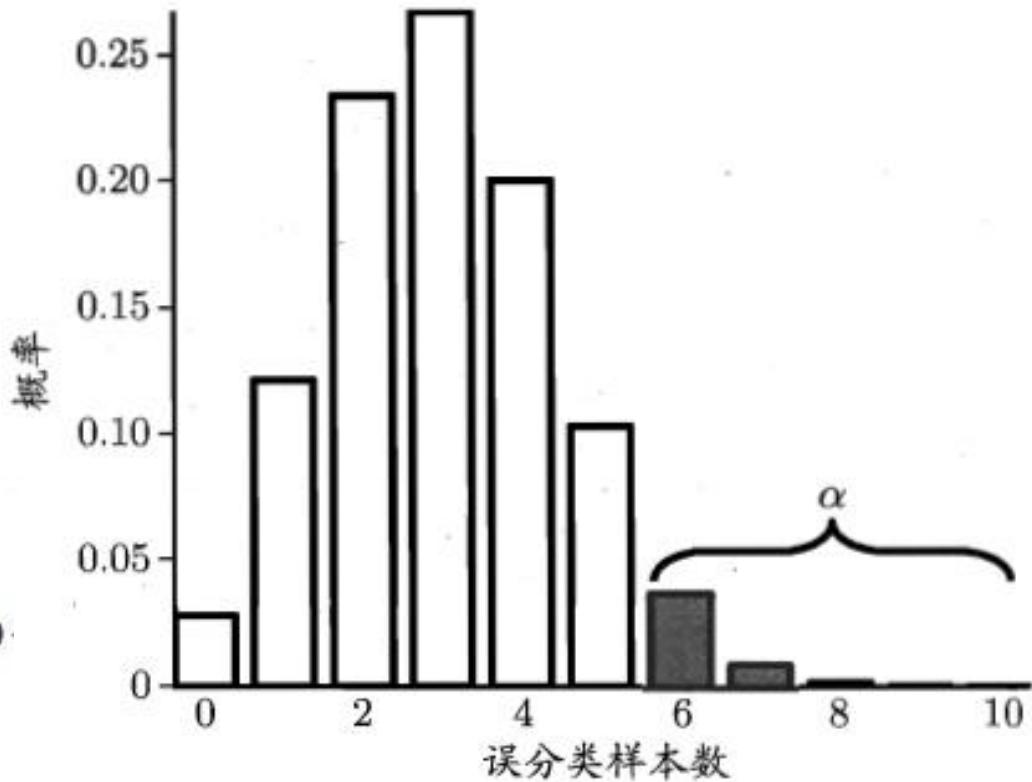
- 考虑以下二个模型：
 - 模型M1: accuracy = 85%, tested on 30 instances
 - 模型M2: accuracy = 75%, tested on 5000 instances
- 模型M1是否优于模型M2?



模型检验

- $H_0: \varepsilon \leq \varepsilon_0$
- 二项检验

二项分布 $B(n, p)$ 在 n 很大时逼近正态分布 $N(np, np(1-p))$



二项分布示意图 ($m = 10, \varepsilon_0 = 0.3$)



模型检验

- $H_0: \varepsilon \leq \varepsilon_0$
- 近似正态检验

$$\frac{\frac{x}{m} - \varepsilon_0}{\sqrt{\varepsilon_0(1-\varepsilon_0)/m}} \sim AN(0,1)$$



模型检验

- 单总体情况。这种情况下 t 统计量的定义为

$$t = \frac{\bar{X} - u_0}{\sigma / \sqrt{N}}$$

式中 \bar{X} 为样本的均值, u_0 为总体的均值, σ 为总体标准差, N 为样本个数, 由于总体标准差无法得知, 因此一般用样本标准差 S 来估计总体标准差。

- $H_0: \varepsilon \leq \varepsilon_0$

k 个测试错误率, $\hat{\epsilon}_1, \hat{\epsilon}_2, \dots, \hat{\epsilon}_k$,

- t 检验 (交叉验证法)

$$\mu = \frac{1}{k} \sum_{i=1}^k \hat{\epsilon}_i ,$$

$$\sigma^2 = \frac{1}{k-1} \sum_{i=1}^k (\hat{\epsilon}_i - \mu)^2 .$$

$$\tau_t = \frac{\sqrt{k}(\mu - \varepsilon_0)}{\sigma} \text{ 服从自由度为 } k-1 \text{ 的 } t \text{ 分布}$$



- 准确率的置信区间

- X : 模型正确预测的记录数

- p : 模型真正的准确率

- 当 N 充分大时

$$P\left(\left|\frac{acc - p}{\sqrt{p(1-p)/N}}\right| \leq Z_{1-\alpha/2}\right) = 1 - \alpha$$



例 考虑一个模型，它在 100 个检验记录上具有 80% 的准确率。在 95% 的置信水平下，模型的真实准确率的置信区间是什么？

随着记录数 N 的增大所产生的置信区间：

N	20	50	100	500	1000	5000
置信 区间	0.584	0.670	0.711	0.763	0.774	0.789
	-0.919	-0.888	-0.867	-0.833	-0.824	-0.811

- 从置信区间角度比较模型性能

- Given two models, say M1 and M2, which is better?
 - M1 is tested on D1 (size=n1), found error rate = e_1
 - M2 is tested on D2 (size=n2), found error rate = e_2
 - Assume D1 and D2 are independent
 - If n_1 and n_2 are sufficiently large, then

$$e_1 \sim N(\mu_1, \sigma_1^2)$$

$$e_2 \sim N(\mu_2, \sigma_2^2)$$

- Approximate:
$$\hat{\sigma}_i^2 = \frac{e_i(1 - e_i)}{n_i}$$



- To test if performance difference is statistically significant: $d = e1 - e2$
 - $d \sim N(d_t, \sigma^2_t)$ where d_t is the true difference
 - Since D1 and D2 are independent, their variance adds up:

$$\sigma_t^2 = \sigma_1^2 + \sigma_2^2 \approx \hat{\sigma}_1^2 + \hat{\sigma}_2^2$$

$$= \frac{e1(1 - e1)}{n1} + \frac{e2(1 - e2)}{n2}$$

- At $(1-\alpha)$ confidence level, $d_t = d \pm Z_{1-\alpha/2} \hat{\sigma}_t$



- An Illustrative Example

- Given: M1: n1 = 30, e1 = 0.15
M2: n2 = 5000, e2 = 0.25

- $d = |e_2 - e_1| = 0.1$ (2-sided test)

$$\hat{\sigma}_d^2 = \frac{0.15(1 - 0.15)}{30} + \frac{0.25(1 - 0.25)}{5000} = 0.0043$$

- At 95% confidence level, $Z_{1-\alpha/2} = 1.96$

$$d_t = 0.100 \pm 1.96 \times \sqrt{0.0043} = 0.100 \pm 0.128$$

- => Interval contains 0 => difference may not be statistically significant



- 直接利用假设检验方法比较模型性能

- 交叉验证配对t检验

- 两个学习器A、B

- K折交叉验证法得测试集上的误差为 $\epsilon_1^A, \epsilon_2^A, \dots, \epsilon_k^A$ 和 $\epsilon_1^B, \epsilon_2^B, \dots, \epsilon_k^B$

- 若A、B性能相同，则在相同训练/测试集上的误差应一样，即

$$\Delta_i = \epsilon_i^A - \epsilon_i^B = 0$$

- 计算差值的均值和方差 $\bar{\Delta} = \frac{\sum_{i=1}^K \Delta_i}{K}$ $S^2 = \frac{\sum_{i=1}^K (\Delta_i - \bar{\Delta})^2}{K-1}$

- $H_0 : \varepsilon^A - \varepsilon^B = 0$

$$\frac{\sqrt{K\bar{\Delta}}}{S} \sim t(K-1)$$



■ McNemar 检验

■ 考察二学习器

两学习器分类差别列联表

算法 B	算法 A	
	正确	错误
正确	e_{00}	e_{01}
错误	e_{10}	e_{11}

■ 若二学习器性能相同，则 $e_{01} = e_{10}$

$$\tau_{\chi^2} = \frac{(|e_{01} - e_{10}| - 1)^2}{e_{01} + e_{10}} \sim \chi^2(1)$$

修正系数
(《机器学习》有误)



■ McNemar 检验

		宣讲后		合计
		有必要	无必要	
宣讲前	有必要	28	6	34
	无必要	49	17	66
合计		77	23	100

- 计算统计量；查表找临界值



■ Friedman 检验 和 Nemenyi 检验

- 用于多算法的比较 (H_0 : 所有算法性能相同)
- 基于算法排序
 - 使用交叉验证法得到每个算法在每个数据集上的测试结果
 - 在每个数据集上根据测试性能好坏排序，并赋序值 1, 2,。若算法性能相同，则平分序值
- 计算平均序值

算法比较序值表

数据集	算法 A	算法 B	算法 C
D_1	1	2	3
D_2	1	2.5	2.5
D_3	1	2	3
D_4	1	2	3
平均序值	1	2.125	2.875



- 若算法性能相同，则平均序值应相同。

- k: 算法个数

- N: 数据集个数

- r_i : 第*i*个算法的平均序数

$$\begin{aligned}\tau_{\chi^2} &= \frac{k-1}{k} \cdot \frac{12N}{k^2-1} \sum_{i=1}^k \left(r_i - \frac{k+1}{2} \right)^2 \\ &= \frac{12N}{k(k+1)} \left(\sum_{i=1}^k r_i^2 - \frac{k(k+1)^2}{4} \right)\end{aligned}$$

在 k 和 N 都较大时，服从自由度为 $k-1$ 的 χ^2 分布。

$$\tau_F = \frac{(N-1)\tau_{\chi^2}}{N(k-1) - \tau_{\chi^2}} \sim F(k-1, (k-1)(N-1))$$



- 若 H_0 被拒绝，则算法性能显著不同。需要进一步区分各算法。

Nemenyi 检验计算出平均序值差别的临界值域

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}},$$

若两个算法的平均序值之差超出了临界值域 CD , 则以相应的置信度拒绝“两个算法性能相同”这一假设.

q_α 的值可以查看下表获得:

α	算法个数 k									
	2	3	4	5	6	7	8	9	10	
0.05	1.960	2.344	2.569	2.728	2.850	2.949	3.031	3.102	3.164	
0.1	1.645	2.052	2.291	2.459	2.589	2.693	2.780	2.855	2.920	



■ 例 (续)

■ $K = 3, N = 4, \alpha = 0.05$

$$\tau_F = \frac{(N-1)\tau_{\chi^2}}{N(k-1) - \tau_{\chi^2}} \sim F(k-1, (k-1)(N-1))$$

 F 检验的常用临界值 $\alpha = 0.05$

数据集 个数 N	算法个数 k	
	2	3
4	10.128	5.143
5	7.709	4.459
8	5.591	3.739
10	5.117	3.555
15	4.600	3.340
20	4.381	3.245

$$\tau_F = 24.429$$

数据集	算法 A	算法 B	算法 C
D_1	1	2	3
D_2	1	2.5	2.5
D_3	1	2	3
D_4	1	2	3
平均序值	1	2.125	2.875

Nemenyi 检验中常用的 q_α 值

α	算法个数 k		
	2	3	4
0.05	1.960	2.344	2.569
0.1	1.645	2.052	2.291

拒绝原假设！

$$CD = 1.657$$

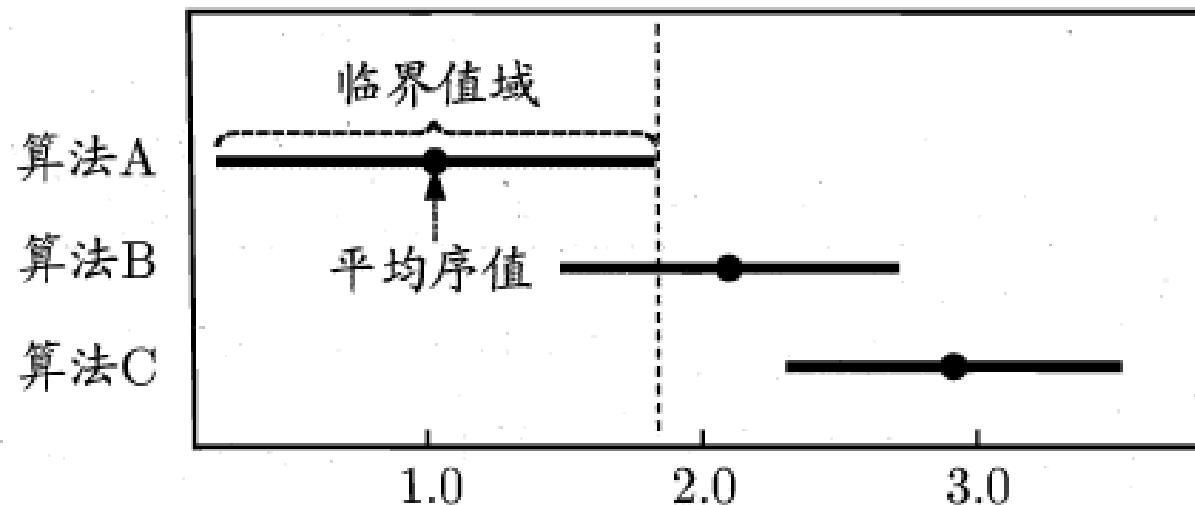
$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$$

A、C 性能显著不同，AB\BC 否



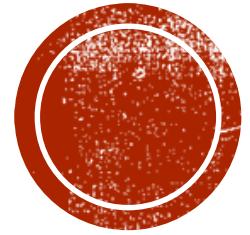
■ Friedman 检验图

- 横轴：平均序值；纵轴：各算法
- 点：每个算法的平均序值；横线段：临界值域
- 若两个算法的横线段有交叠，则算法无显著差别；否则说明有显著差别。



Friedman 检验图

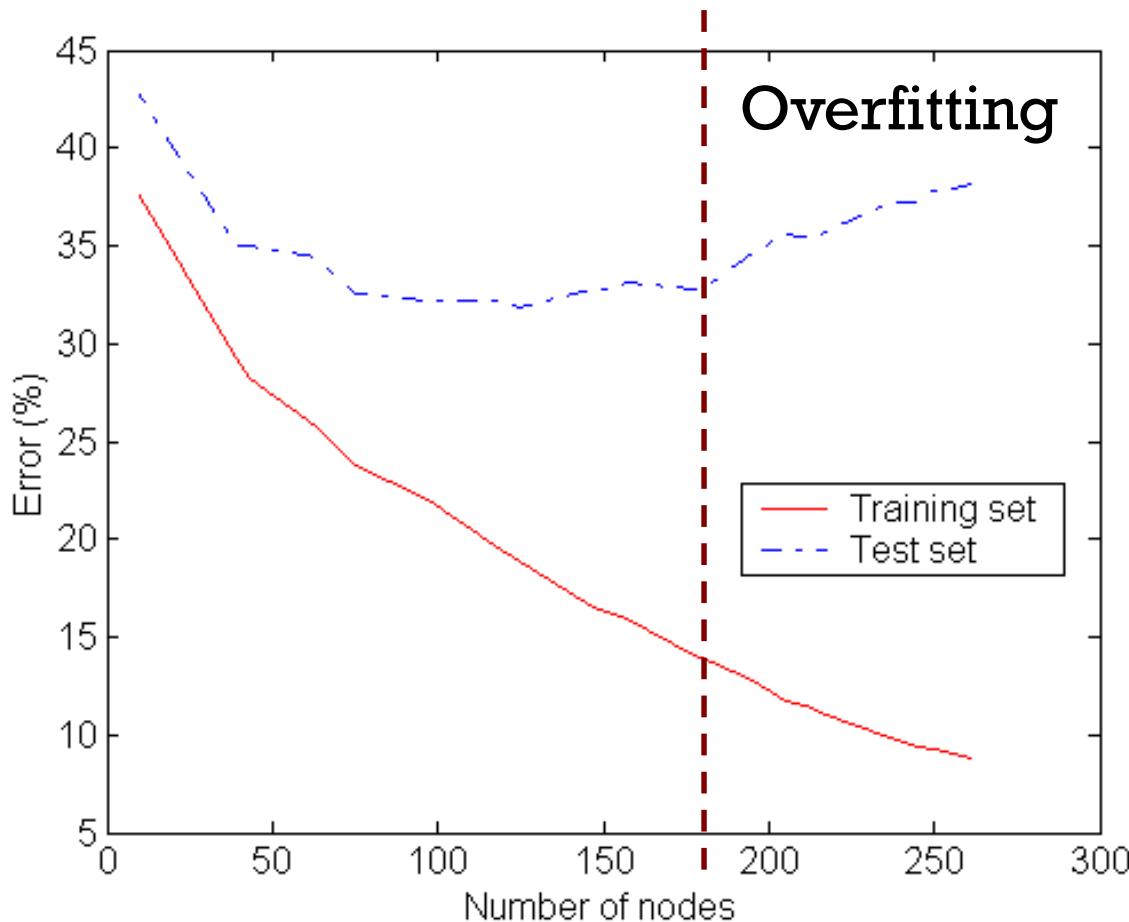




偏差-方差分解



■ 过拟合与欠拟合



Underfitting: when model is too simple, both training and test errors are large



模型M1：锯齿 \cap 绿色 \rightarrow 树叶



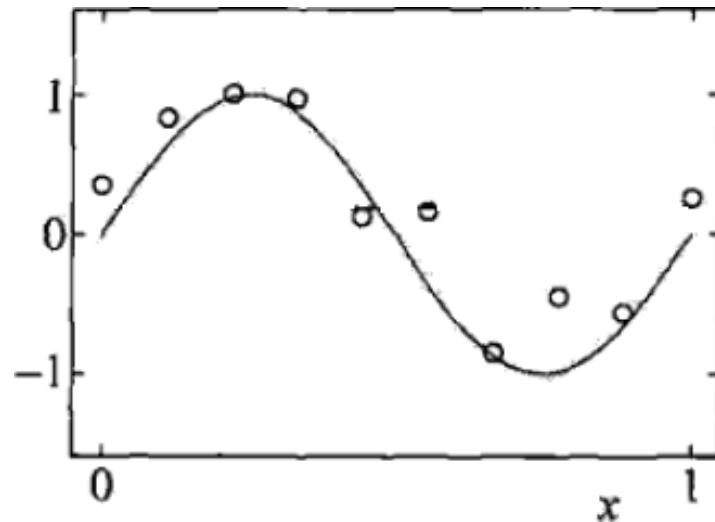
模型M2：绿色 \rightarrow 树叶



例 假设给定一个训练数据集：

$$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$$

在 M 次多项式函数中选择一个对已知数据以及未知数据都有很好预测能力的函数.



设 M 次多项式为

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

求以下经验风险最小化:

$$L(w) = \frac{1}{2} \sum_{i=1}^N (f(x_i, w) - y_i)^2 = \frac{1}{2} \sum_{i=1}^N \left(\sum_{j=0}^M w_j x_i^j - y_i \right)^2$$

平方
损失



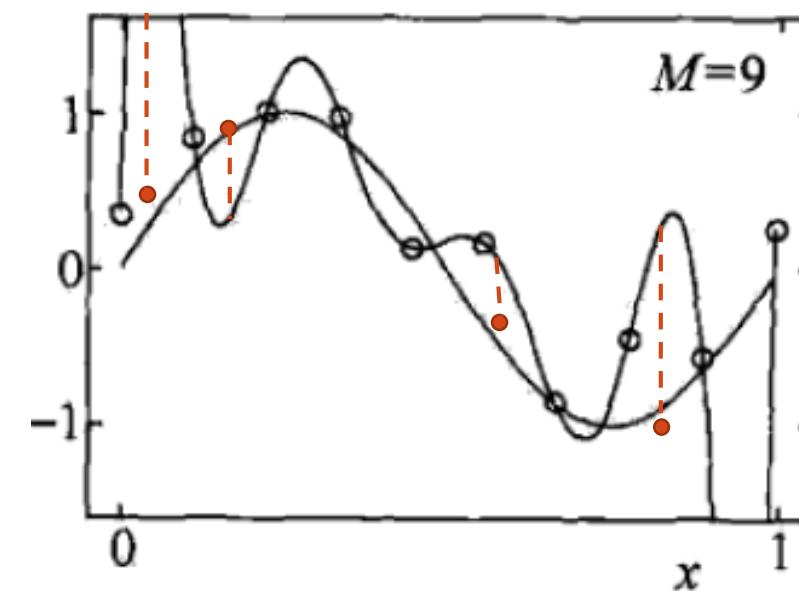
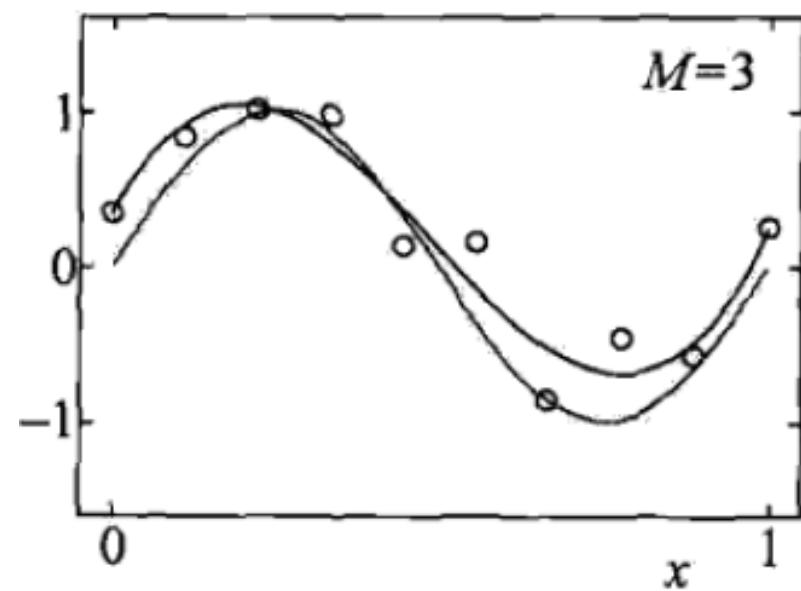
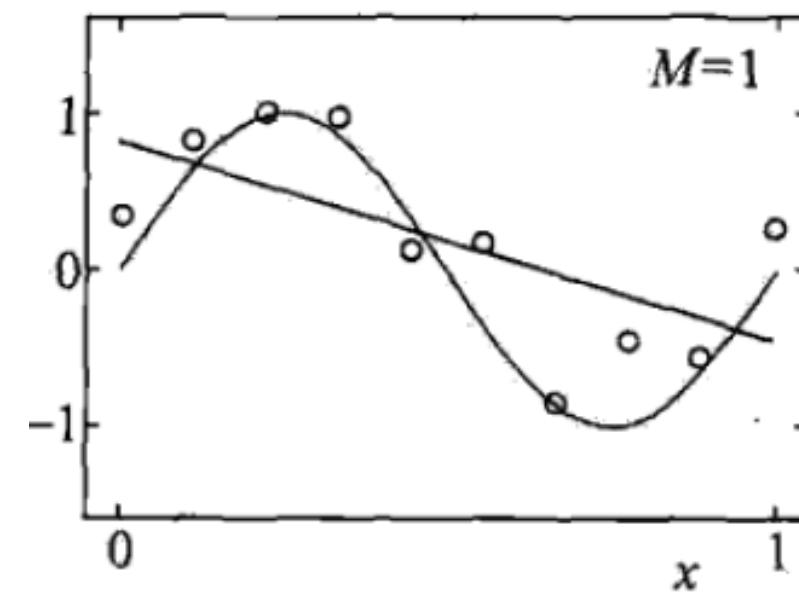
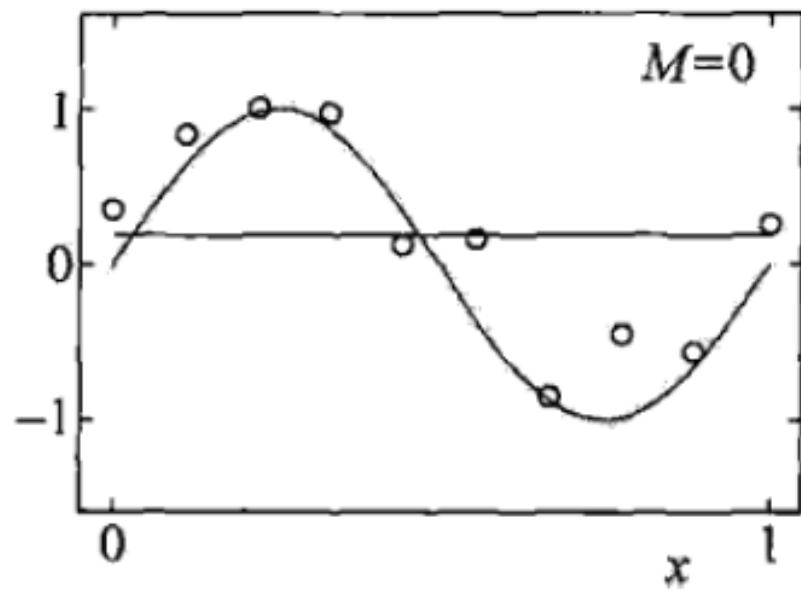
对 w_j 求偏导数并令其为 0， 可得

$$w_j = \frac{\sum_{i=1}^N x_i y_i}{\sum_{i=1}^N x_i^{j+1}}, \quad j = 0, 1, 2, \dots, M$$

于是求得拟合多项式系数 $w_0^*, w_1^*, \dots, w_M^*$.



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



噪 声 (标注错误) 期 望 为 0

噪 声 与 模 型 无 关

• 偏 差 - 方 差 分 解

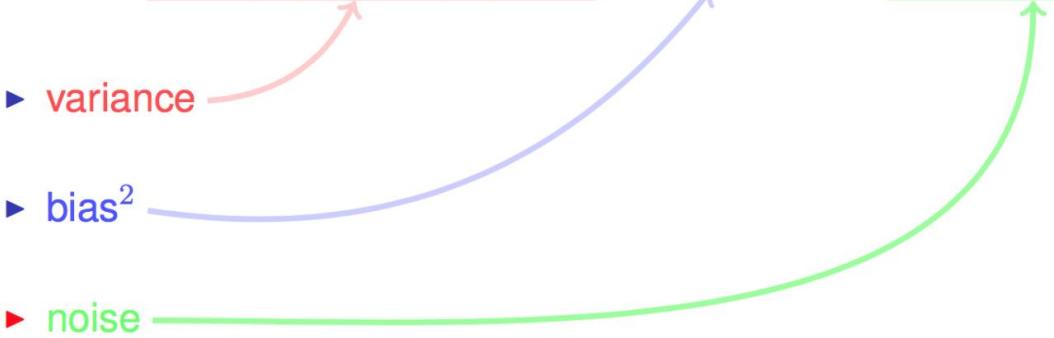
假定 $\mathbb{E}_D[y_D - y] = 0$. $\bar{f}(\mathbf{x}) = \mathbb{E}_D[f(\mathbf{x}; D)]$

$$\begin{aligned}
 E(f; D) &= \mathbb{E}_D [(f(\mathbf{x}; D) - y_D)^2] \\
 &= \mathbb{E}_D [(f(\mathbf{x}; D) - \bar{f}(\mathbf{x}) + \bar{f}(\mathbf{x}) - y_D)^2] \\
 &= \mathbb{E}_D [(f(\mathbf{x}; D) - \bar{f}(\mathbf{x}))^2] + \mathbb{E}_D [(\bar{f}(\mathbf{x}) - y_D)^2] \\
 &\quad + \mathbb{E}_D [2(f(\mathbf{x}; D) - \bar{f}(\mathbf{x}))(\bar{f}(\mathbf{x}) - y_D)] \\
 &= \mathbb{E}_D [(f(\mathbf{x}; D) - \bar{f}(\mathbf{x}))^2] + \mathbb{E}_D [(\bar{f}(\mathbf{x}) - y_D)^2] \\
 &= \mathbb{E}_D [(f(\mathbf{x}; D) - \bar{f}(\mathbf{x}))^2] + \mathbb{E}_D [(\bar{f}(\mathbf{x}) - y + y - y_D)^2] \\
 &= \mathbb{E}_D [(f(\mathbf{x}; D) - \bar{f}(\mathbf{x}))^2] + \mathbb{E}_D [(\bar{f}(\mathbf{x}) - y)^2] + \mathbb{E}_D [(y - y_D)^2] \\
 &\quad + 2\mathbb{E}_D [(\bar{f}(\mathbf{x}) - y)(y - y_D)] \\
 &= \mathbb{E}_D [(f(\mathbf{x}; D) - \bar{f}(\mathbf{x}))^2] + \mathbb{E}_D [(\bar{f}(\mathbf{x}) - y)^2] + \mathbb{E}_D [(y - y_D)^2] \\
 &= Variance + Bias + Noise
 \end{aligned}$$

符 号	涵义
\mathbf{x}	测试样本 (总 体)
D	数据集 (多 个)
y_D	\mathbf{x} 在数据集中的标记
y	\mathbf{x} 的真实标记
f	训练集 D 学得的模型
$f(\mathbf{x}; D)$	由训练集 D 学得的模型 f 对 \mathbf{x} 的预测输出
$\bar{f}(\mathbf{x})$	模型 f 对 \mathbf{x} 的 期望 预测 输出

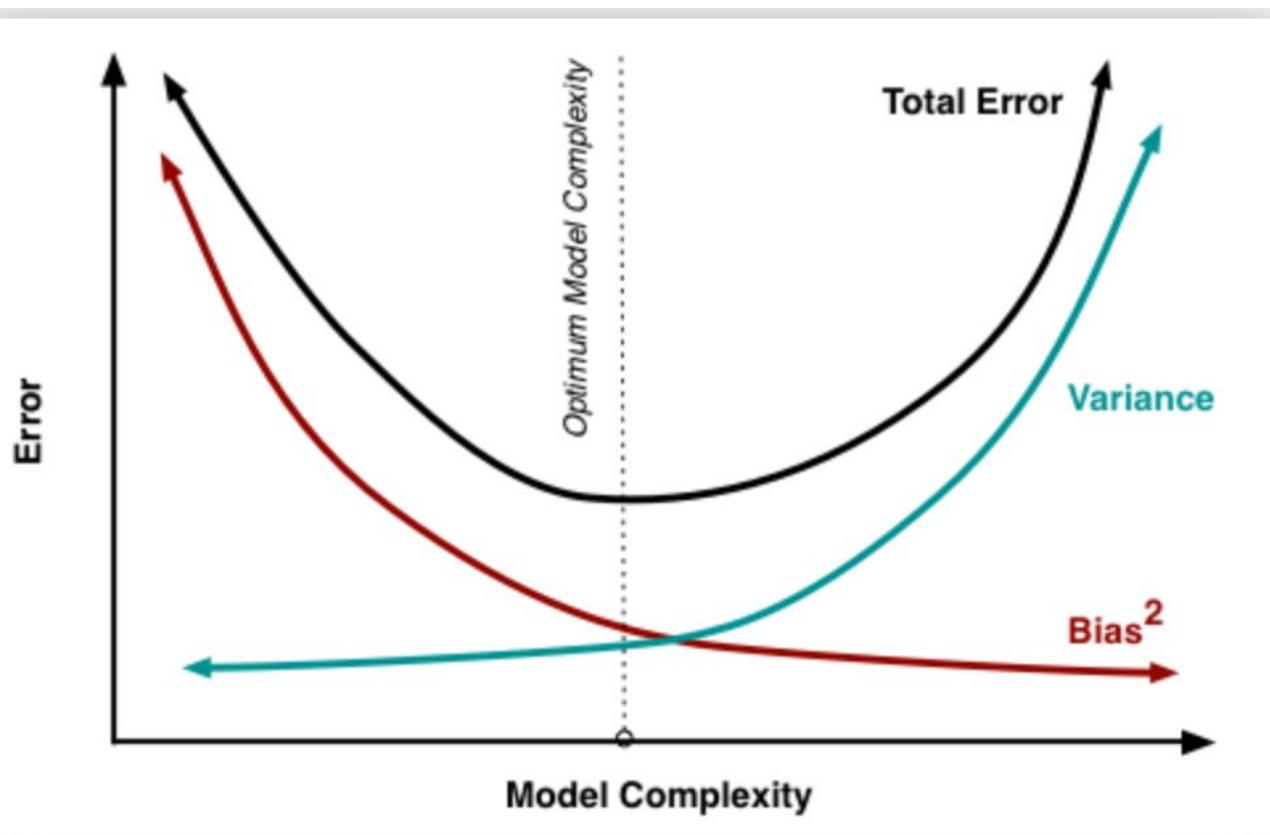


$$Err(\mathbf{x}) = \mathbb{E}_D \left[(f(\mathbf{x}; D) - \bar{f}(\mathbf{x}))^2 \right] + (\bar{f}(\mathbf{x}) - y)^2 + \mathbb{E}_D \left[(y_D - y)^2 \right]$$



■ 欠拟合

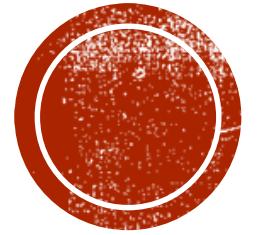
- 拟合能力不足
- 偏差占主导
- 训练数据的扰动不足以使学习器发生变化
- 增加模型参数



■ 过拟合

- 拟合能力过强
- 方差占主导
- 训练数据的轻微扰动会导致模型变化
- 减少参数，正则化
- 集成学习





分 类

Classification

新的问题

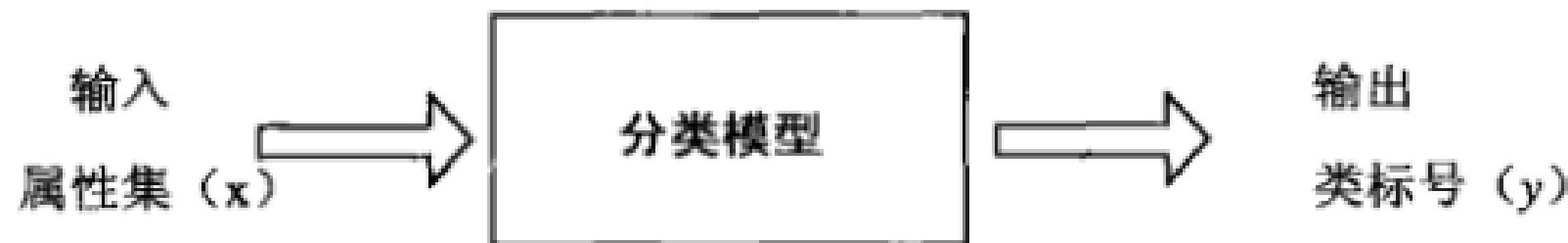
脊椎动物的数据集

名字	体温	表皮覆盖	胎生	水生动物	飞行动物	有腿	冬眠	类标号
人类	恒温	毛发	是	否	否	否	否	哺乳类
蟒蛇	冷血	鳞片	否	否	否	否	是	爬行类
鲑鱼	冷血	鳞片	否	是	否	否	否	鱼类
鲸	恒温	毛发	是	半	否	否	否	哺乳类
青蛙	冷血	无	否	否	否	是	是	两栖类
巨蜥	冷血	鳞片	否	否	是	是	是	爬行类
蝙蝠	恒温	毛发	是	否	是	是	是	哺乳类
鸽子	恒温	羽毛	否	否	是	否	否	鸟类
猫	恒温	软毛	是	否	否	否	否	哺乳类
豹纹鲨	冷血	鳞片	否	是	否	否	否	鱼类
海龟	冷血	鳞片	否	否	否	否	否	爬行类
企鹅	恒温	羽毛	是	半	否	否	是	鸟类
豪猪	恒温	刚毛	否	半	否	否	否	哺乳类
鳗	冷血	鳞片	是	否	否	否	否	鱼类
蝾螈	冷血	无	否	半	否	否	是	两栖类



分 类

- 分类任务就是通过学习得到一个目标函数 f , 把每个属性集 X 映射到一个预先定义的类标号 y
 - 目标函数也称分类模型



分类器的任务是根据输入属性集 x 确定类标号 y



解决分类问题的一般方法

- 使用一种学习算法 (learning algorithm) 确定分类模型
 - 拟合输入数据中的类标号和属性集之间的联系
 - 正确预测未知样本的类标号
 - 决策树、基于规则的分类法、KNN、神经网络、支持向量机、贝叶斯分类.....
- 提取有效的样本特征



特征的分类：

根据人类视觉系统对场景的感知，特征分为**低层特征**和**高层语义特征**两个方面。

低层特征包括：外观特征、边缘特征、纹理特征、形状特征等。

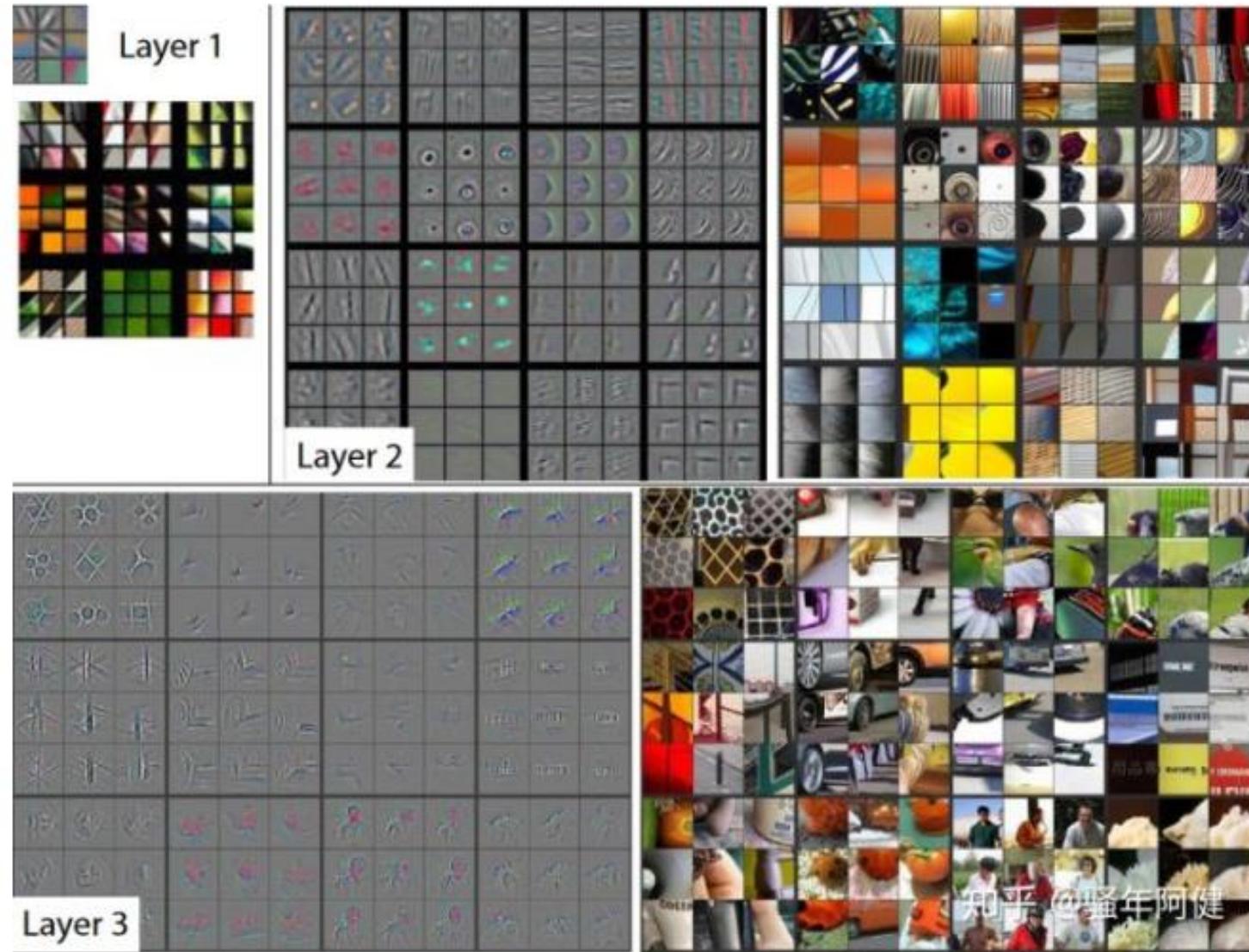
高层语义特征包括：场景语义、行为语义和情感语义等3个层次。

在底层图像特征与高层语义特征之间，往往存在着“语义鸿沟”，视觉物体的**中层语义表示**是连接低层与高层语义的纽带。

低层特征描述的是图像的视觉特征，中层语义是由视觉特征推导得到的中间语义特征，对应于**视觉词袋**和**中间语义（主题语义）**两种表示方法。



低层、中层和高层特征



知乎 @ 骚年阿健



有效的特征与特征描述举例：

- 统计特征（统计直方图、方差、均值、熵）
- 几何特征（包括面积、周长、位置、方向、距离）
- 边界特征（链码、傅立叶描述子）
- 形状描述（区域描述、图像矩）
- 灰度与颜色描述（梯度特征、灰度直方图特征、梯度方向直方图、特征直方图、颜色直方图和颜色直方图距离、颜色不变量特征）
- 角点特征（Harris、Susan、Fast、Brisk）
- 变换特征（傅里叶、小波）
- 纹理特征（灰度共生矩阵、LBP特征，结构相似性特征，Gabor特征）
- 局部不变特征描述子（SIFT、SURF、GLOH、WLD、D-nets） ...



- 1 基本概念
- 2 颜色特征描述
- 3 形状特征描述
- 4 图像的纹理分析技术
- 5 小结



◆ 目的

让计算机具有认识或者识别图像的能力，即图像识别。特征选择是图像识别中的一个关键问题。特征选择和提取的基本任务是如何从众多特征中找出最有效的特征。

◆ 特征形成

根据待识别的图像，通过计算产生一组原始特征，称之为特征形成。



◆ 特征提取

原始特征的数量很大，或者说原始样本处于一个高维空间中，通过映射或变换的方法可以将高维空间中的特征描述用低维空间的特征来描述，这个过程就叫特征提取。

◆ 特征选择

从一组特征中挑选出一些最有效的特征以达到降低特征空间维数的目的，这个过程就叫特征选择。

选取的特征应具有如下特点：

可区别性

可靠性

独立性好

数量少

8.2 颜色特征描述

8.2.1 颜色矩

8.2.2 颜色直方图



8.2.1 颜色矩

- ◆ 颜色矩是以数学方法为基础的，通过计算矩来描述颜色的分布。
- ◆ 颜色矩通常直接在RGB空间计算
- ◆ 颜色分布的前三阶矩表示为：

$$\mu_i = \frac{1}{N} \sum_{j=1}^N p_{i,j}$$

$$\sigma_i = \left(\frac{1}{N} \sum_{j=1}^N (p_{i,j} - \mu_i)^2 \right)^{\frac{1}{2}}$$

$$s_i = \left(\frac{1}{N} \sum_{j=1}^N (p_{i,j} - \mu_i)^3 \right)^{\frac{1}{3}}$$





8.2.2 颜色直方图

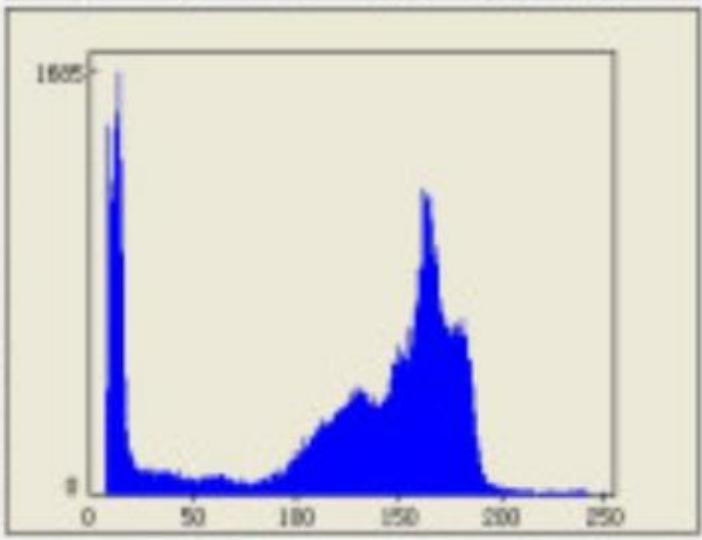
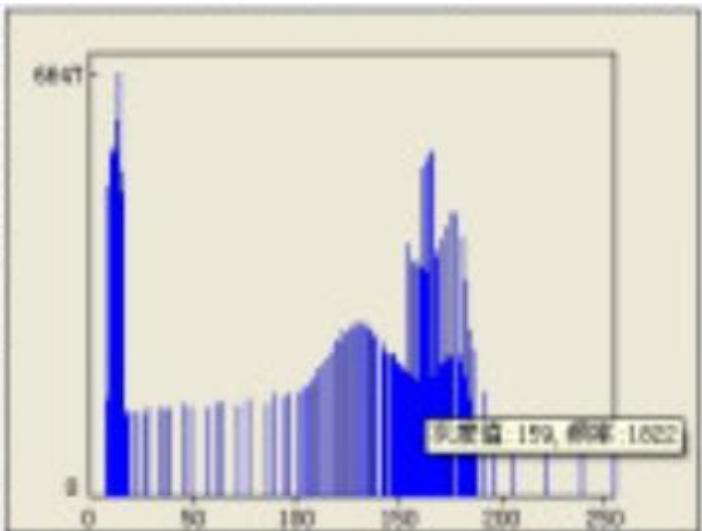
- ◆ 设一幅图像包含M个像素，图像的颜色空间被量化成N个不同颜色。颜色直方图H定义为：

$$p_i = h_i \quad (8-4)$$

h_i 为第i种颜色在整幅图像中具有的像素数。

- ◆ 归一化为：

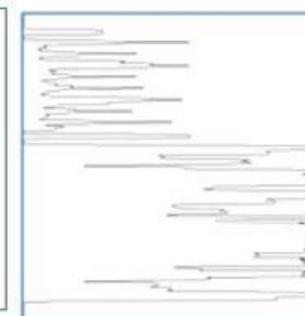
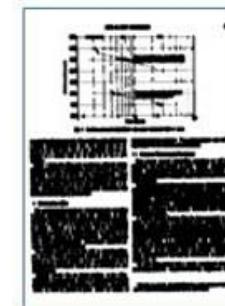
$$p_i = h_i / M \quad (8-5)$$



投影直方图

- 对于一个二值图像 $f(x, y)$, 其垂直和水平投影直方图可以分别记为 $p_v(y)$ 和 $p_h(x)$, 其中

- $p_v(y) = \sum_{x=0}^{w(f)} f(x, y), \quad p_h(x) = \sum_{y=0}^{h(f)} f(x, y),$
- $w(f)$ 和 $h(f)$ 分别为 $f(x, y)$ 的像素宽度和高度。



形状特征提取

- 基于图像内物体形状的检索是基于内容检索当中一个**最具挑战性**的问题之一，因为寻找符合人眼感知特性的形状特征不是一件简单的工作。
- 首要的**困难**是要将不同物体从图像中分割出来，这是计算机视觉的困难问题之一。
- 形状的描述也是困难的问题，常用的方法有**傅立叶描述子**、**矩不变量**、各种简单的**形状因子**（如**面积**、**圆度**、**偏心度**、**主轴方向**）等。
- 除了这些全局特征以外，有时也用一些局部特征（如**直线段**、**圆弧**、**角点**、**高曲率点**等），以解决遮挡问题。

福

寿



形状特征提取

Hu不变矩

Zernike不变矩

Hough变换

Radon变换

傅里叶描述子



HU不变矩

不变矩

◎ $f(x, y)$ 的 $j + k$ 阶矩定义为：

■ $m_{jk} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x^j y^k f(x, y) dx dy,$

◎ $f(x, y)$ 的 $j + k$ 阶中心规格矩定义为：

■ $\mu_{jk} = \frac{1}{m_{00}} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \bar{x})^j (y - \bar{y})^k f(x, y) dx dy,$

○ 其中 $\bar{x} = \frac{m_{10}}{m_{00}}$, $\bar{y} = \frac{m_{01}}{m_{00}}$, $j + k = 2, 3, 4, \dots$ 。

◎ 中心规格矩同时具有位置无关性和大小无关性。

◎ 使得二阶中心矩 μ_{11} 为最小的旋转角 θ 可以由如下公式得出

■ $\tan 2\theta = \frac{2\mu_{11}}{\mu_{20} - \mu_{02}}.$

○ 相对于主轴计算的中心规格矩，在物体缩放、平移和旋转时保持不变。

HU不变矩

- $\emptyset_1 = \mu_{20} + \mu_{02}$
- $\emptyset_2 = (\mu_{20} - \mu_{02})^2 + 4\mu_{11}^2$
- $\emptyset_3 = (\mu_{30} - 3\mu_{12})^2 + (3\mu_{21} - \mu_{03})^2$
- $\emptyset_4 = (\mu_{30} + \mu_{12})^2 + (\mu_{21} + \mu_{03})^2$
- $\emptyset_5 = (\mu_{30} - 3\mu_{12})(\mu_{30} + \mu_{12})[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2] + (3\mu_{21} - \mu_{03})(\mu_{21} + \mu_{03})[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2]$
- $\emptyset_6 = (\mu_{20} - \mu_{02})[(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2] + 4\mu_{11}(\mu_{30} + \mu_{12})(\mu_{21} + \mu_{03})$
- $\emptyset_7 = (3\mu_{21} - \mu_{03})(\mu_{30} + \mu_{12})[(\mu_{30} + \mu_{12})^2 - 3(\mu_{21} + \mu_{03})^2] + (\mu_{30} - 3\mu_{12})(\mu_{21} + \mu_{03})[3(\mu_{30} + \mu_{12})^2 - (\mu_{21} + \mu_{03})^2]$



HU不变矩



(a)



(b)



(c)



(d)

	\emptyset_1	\emptyset_2	\emptyset_3	\emptyset_4	\emptyset_5	\emptyset_6	\emptyset_7
(a)	0.001286	0.000000	0.000000	0.000000	-0.000000	0.000000	0.000000
(b)	0.411764	0.001129	0.000195	0.004179	0.000004	0.000120	-0.000002
(c)	0.001615	0.000000	0.000000	0.000000	0.000000	0.000000	-0.000000
(d)	0.310856	0.002382	0.005049	0.000931	-0.000001	-0.000009	0.000002



ZERNIKE不变矩

◎ Zernike不变矩是来自数学领域的概念，其定义是一组多项式序列，每个多项式在单位圆上是正交的。

◎ 存在奇数和偶数Zernike多项式，其中偶数多项式可以定义为

- $Z_n^m(\rho, \varphi) = R_n^m(\rho) \cos(m\varphi)$,

- 而奇数多项式则定义为

- $Z_n^{-m}(\rho, \varphi) = R_n^m(\rho) \sin(m\varphi)$,

- 其中， m 和 n 为非负整数，且 $n \geq m$ ， φ 是以弧度为单位的方位角， ρ 为规范化的径向距离，径向多项式 R_n^m 则定义为

- $$\odot R_n^m(\rho) = \sum_{k=0}^{(n-m)/2} \frac{(-1)^k (n-k)!}{k! \left(\frac{n+m}{2} - k\right)! \left(\frac{n-m}{2} - k\right)!} \rho^{n-2k}.$$



ZERNIKE不变矩

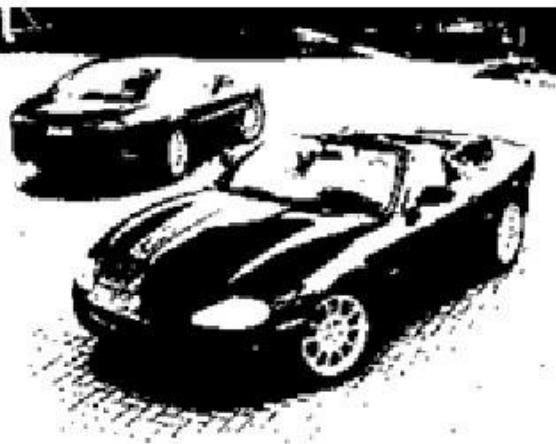
- ◎ 在Zernike不变矩的基础上，伪Zernike不变矩被提出，并具有更好的鲁棒性和对图像噪声的不敏感性。
 - 寻找目标像素所占据的外接矩形区域；
 - 根据所设定的阶数 n ，进行重复次数 l 的计算，从而进一步得到所需计算的不变矩系数的个数；
 - 针对每一对阶数 n 和重复次数 l ，进行如下四个子步骤的计算：
 - 将步骤1所得到的外接矩形区域归一化为一个单位圆；
 - 将所有图像像素从笛卡尔坐标系转换至以单位圆圆心为原点的极坐标系中；
 - 在极坐标系下计算每一个像素离开原点的径向距离；
 - 统计所有像素的径向距离得到在当前阶数 n 和重复次数 l 下的不变矩系数；
 - 将所有的不变矩系数组合起来得到一个系数序列，便是一幅二值图像所对应的伪Zernike不变矩的计算结果。



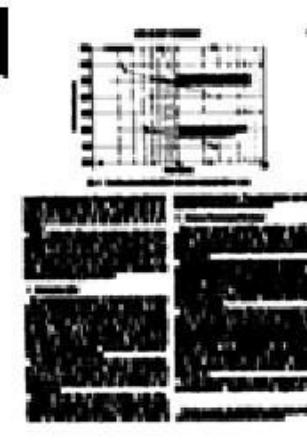
伪ZERNIKE不变矩计算



(a)



(b)



(c)



(d)

	A_{00}	A_{11}	A_{22}	A_{20}	A_{33}	A_{31}
(a)	0.318224	0.083217	0.289287	0.034964	0.163702	0.004639
(b)	0.318229	0.045778	0.423569	0.037042	0.074096	0.080350
(c)	0.318310	0.142310	0.385979	0.073231	0.185505	0.038667
(d)	0.318289	0.039561	0.186443	0.094189	0.019000	0.054490





THE END!

