数据挖掘与机器学习

1

潘斌
panbin@nankai.edu.cn
范孙楼227

上节回顾

- 特征
  - 颜色特征
  - 形状特征
  - 纹理特征
  - 中层特征

本节提要

- 不同模型的性能比较
  - 假设检验方法
  - 偏差-方差分解
- 分类任务
  - 特征
  - 分类模型与学习方法

模型评估与选择

Model Assessment and Selection

损失函数（LOSS FUCTION）

（1）0-1 损失函数（0-1 loss function）

- $$L(Y,f(X))=\begin{cases}1, & Y\neq f(X)\\ 0, & Y=f(X)\end{cases}$$

（2）平方损失函数（quadratic loss function）

- $$L(Y,f(X))=(Y-f(X))^2$$

（3）绝对损失函数（absolute loss function）

$$L(Y,f(X))=|Y-f(X)|$$

（4）对数损失函数（logarithmic loss function）或对数似然损失函数（log-likelihood loss function）

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

风险函数（RISK FUNCTION）

- 理论上f(X)基于联合分布P(X,Y)的平均意义下的损失，即**期望损失**（Expected Loss）、**风险函数**
- 风险函数可理解为样本总体的损失

$$R_{\exp}(f) = E_P[L(Y, f(X))] = \int_{\mathcal{X}\times\mathcal{Y}} L(y, f(x))P(x, y)\,\mathrm{d}x\mathrm{d}y$$

未知

经验风险（EMPIRICAL RISK）

- f(x)关于训练数据集$T=\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$的平均损失

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

模型选择策略

- 经验风险最小化（ Empirical Risk Minimization, ERM ） 求解最优化问题：

$$\min_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i))$$

- 结构风险最小化（ Structural Risk Minimization, SRM ） 求解最优化问题：

$$\min_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f)$$

模型性能度量方法

- 错误率（error rate）和准确度（accuracy）
  - 错误率：分类问题中，误分类样本的比例
  - 准确度：正确分类样本的比例，即正确率
  - 对样本集D

$$E(f;D) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f\left(\boldsymbol{x}_i\right) \neq y_i\right) .$$

$$\begin{aligned}\operatorname{acc}(f;D) &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f\left(\boldsymbol{x}_i\right) = y_i\right) \\ &= 1 - E(f;D) .\end{aligned}$$

$$E(f;\mathcal{D}) = \int_{\boldsymbol{x}\sim\mathcal{D}}\mathbb{I}\left(f\left(\boldsymbol{x}\right) \neq y\right)p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} ,$$

$$\begin{aligned}\operatorname{acc}(f;\mathcal{D}) &= \int_{\boldsymbol{x}\sim\mathcal{D}}\mathbb{I}\left(f\left(\boldsymbol{x}\right) = y\right)p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ &= 1 - E(f;\mathcal{D}) .\end{aligned}$$

- 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

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

  - 泛化误差：对未知数据预测的误差

$$R_{\exp}(\hat{f})=E_P[L(Y,\hat{f}(X))]=\int_{\mathcal{X}\times\mathcal{Y}}L(y,\hat{f}(x))P(x,y)\mathrm{d}x\mathrm{d}y$$

二分类问题的泛化误差上界

训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$ 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 $\mathcal{F}=\{f_1,f_2,\cdots,f_d\}$ 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 $\mathcal{F}=\{f_1,f_2,\cdots,f_d\}$ 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：

$\mathcal{F}$中包含的函数越多，泛化误差上界越大。

训练误差小的模型泛化误差也会小。

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

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 $X_1,X_2,\cdots,X_n$ 之和，$X_i\in[a_i,b_i]$，则对任意 $t>0$，以下不等式成立：

随机变量的和与其期望值偏差的概率上界

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

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

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

ID	X1	X2	Y
3	B	5	T
5	A	2	F
7	A	7	F
8	A	7	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	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	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		
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

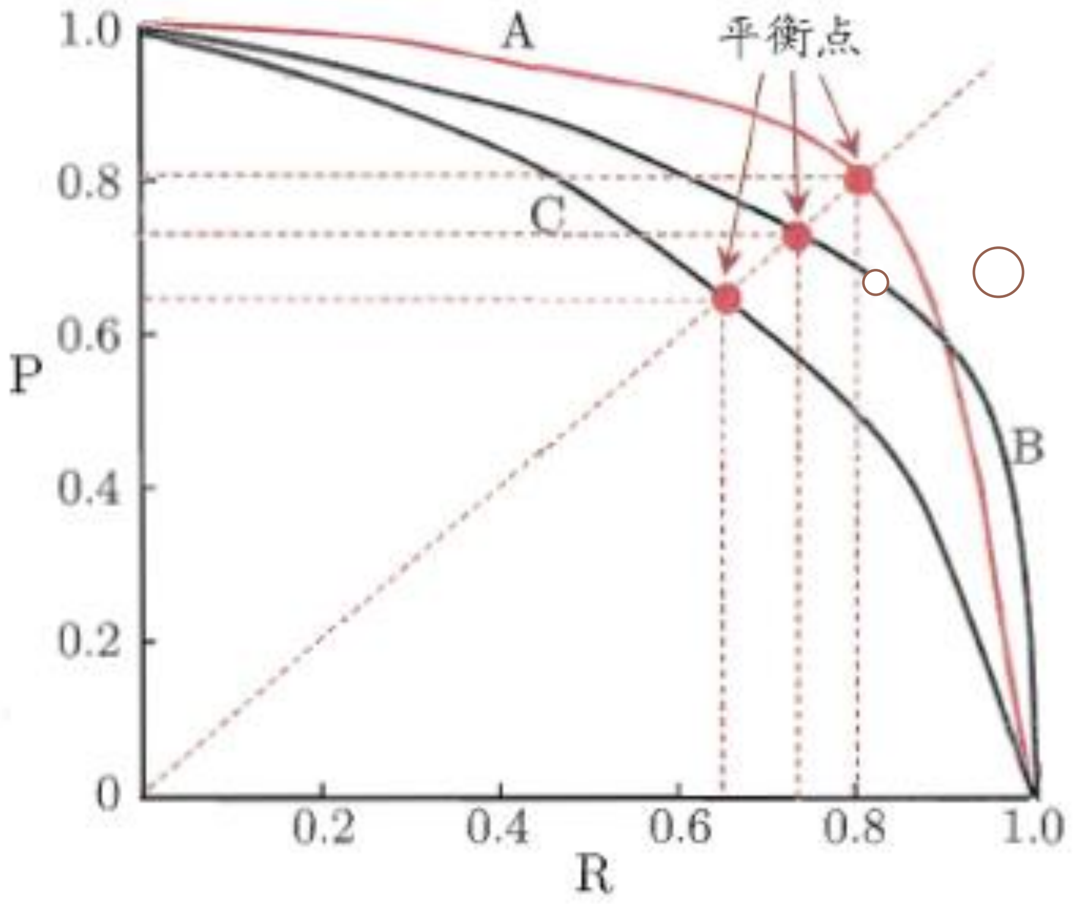
- 精度p：$\text{Precision (p)} = \frac{a}{a+c}$
  - 也称查准率，确定在分类器断言为正类的那部分记录中实际为正类的记录所占的比率。
  - 精度越高，分类器的假正错误率就越低
- 召回率r：$\text{Recall (r)} = \frac{a}{a+b}$
  - 度量被分类器正确预测的正样本的比例，亦称查全率。
  - 具有高召回率的分类器很少将正样本误分为负样本
- F_1度量：$F_1\text{−measure (F)} = \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$
  - 精度和召回率的调和平均

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”（举例展示）

P-R曲 线 和P-R图

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
	Class=No	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

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

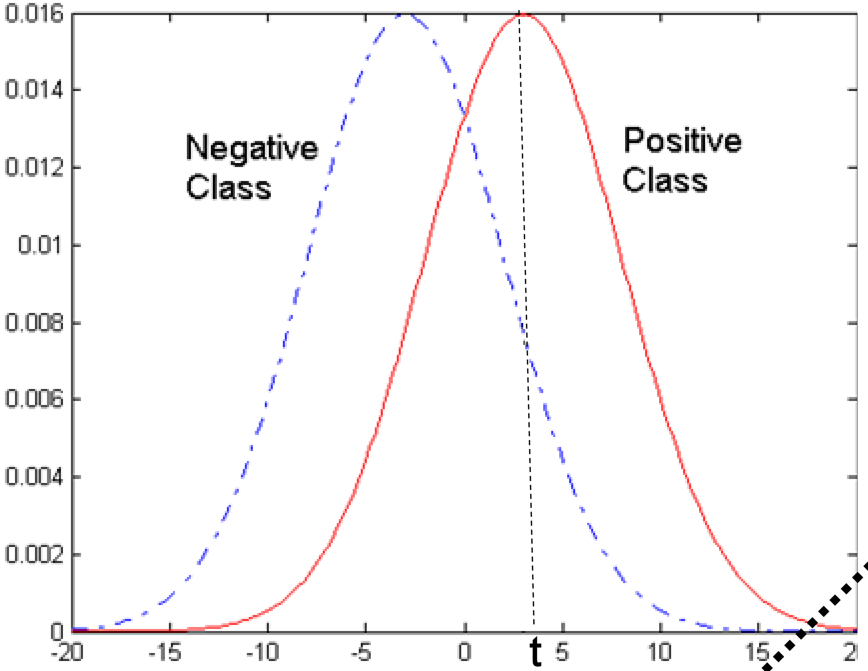
$$\text{micro-}R = \frac{\overline{TP}}{\overline{TP} + \overline{FN}},$$

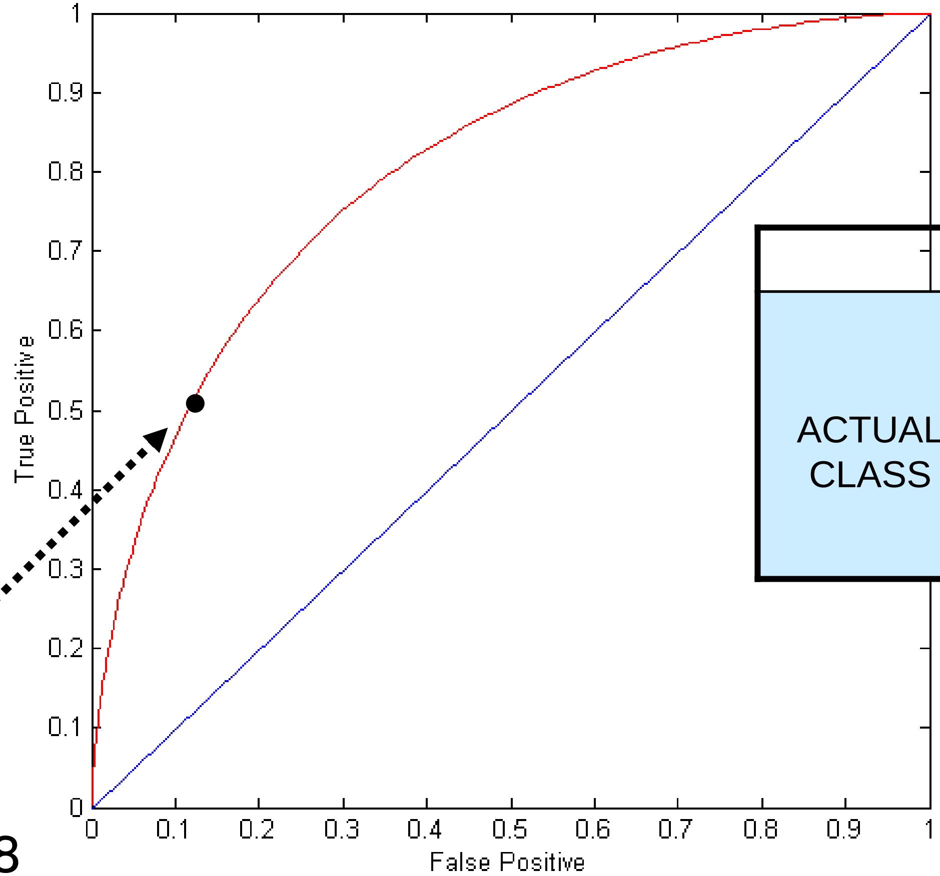
$$\text{micro-}F1 = \frac{2 \times \text{micro-}P \times \text{micro-}R}{\text{micro-}P + \text{micro-}R}.$$

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



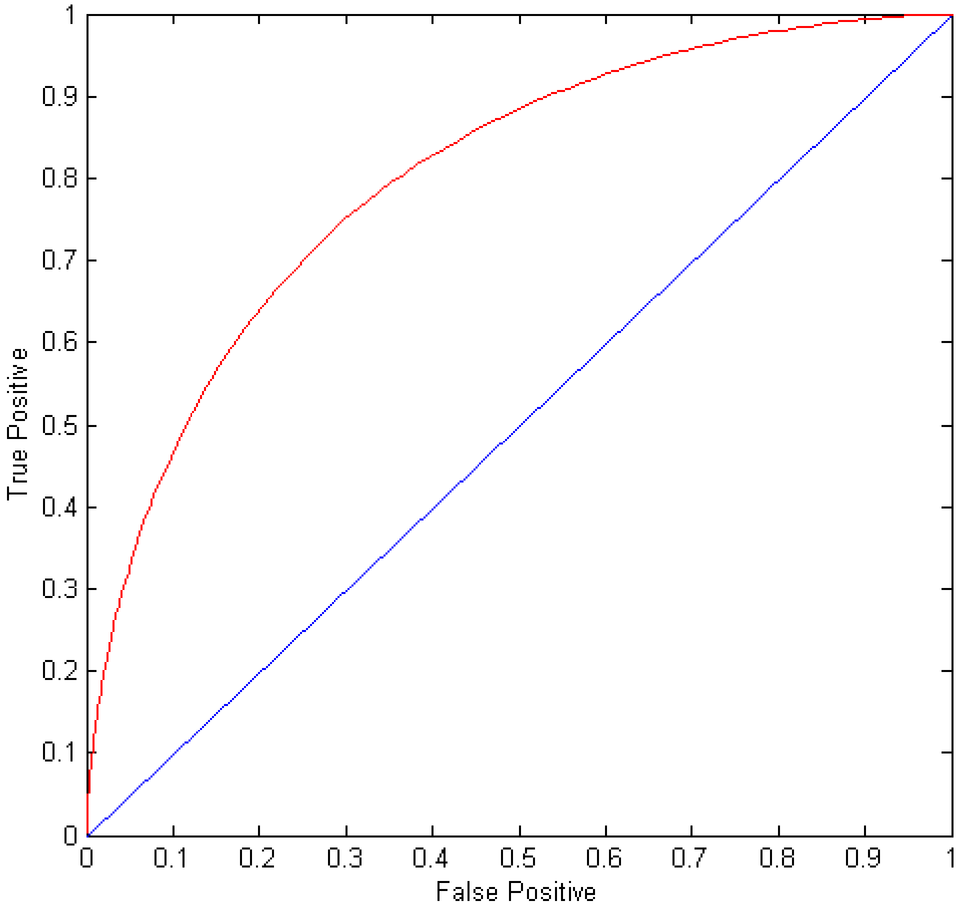


	PREDICTED CLASS	
ACTUAL CLASS	Class=Yes	Class=No
Class=Yes	a (TP)	b (FN)
Class=No	c (FP)	d (TN)

At threshold t:

TP=0.5, FN=0.5, FP=0.12, TN=0.88

- TPR=a/(a+b)
- FPR=c/(c+d)



- 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最优的

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。

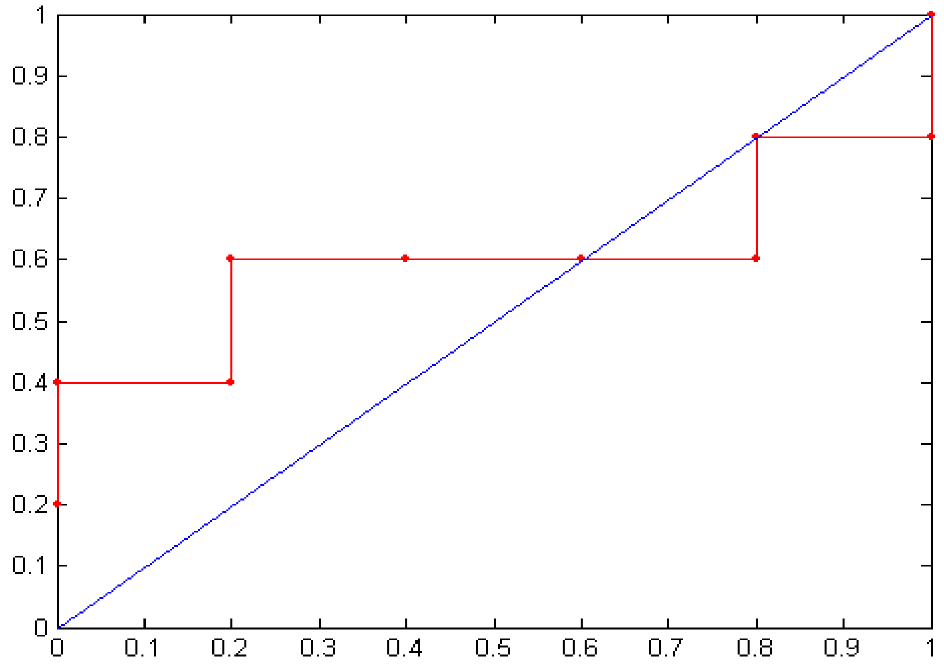
(4) 重复步骤 3 并相应地更新 TP 和 FP 计数，直到最高秩的记录被选择。

(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
FP	5	5	4	4	3	2	1	1	0	0	0
TN	0	0	1	1	2	3	4	4	5	5	5
FN	0	1	1	2	2	2	2	3	3	4	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

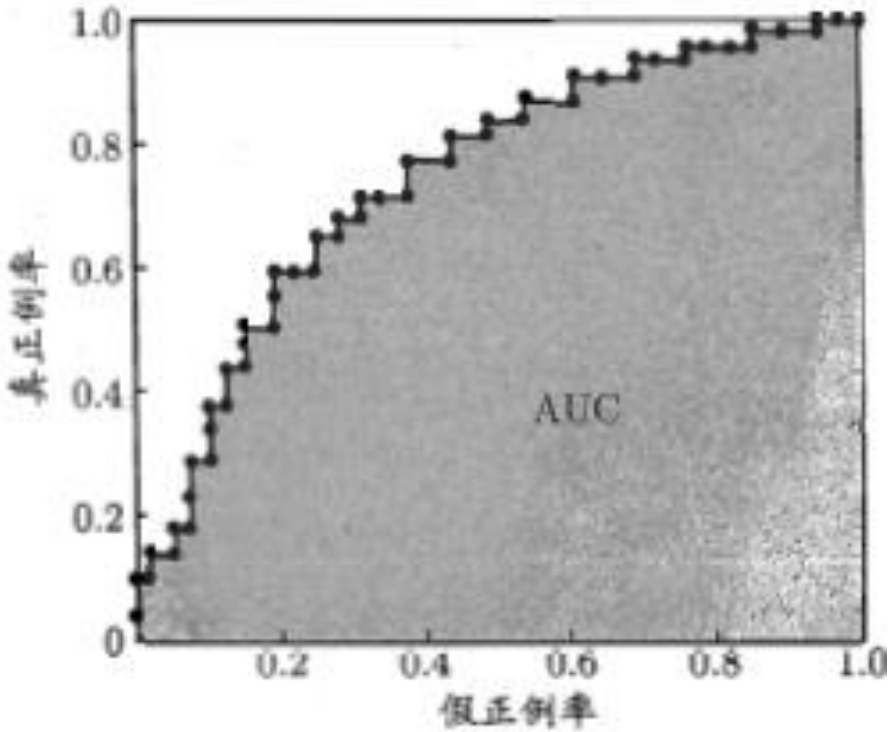
ROC Curve:



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

AUC（AREA UNDER ROC CURVE）

- 即ROC曲线下的面积
- 可以评价模型的性能

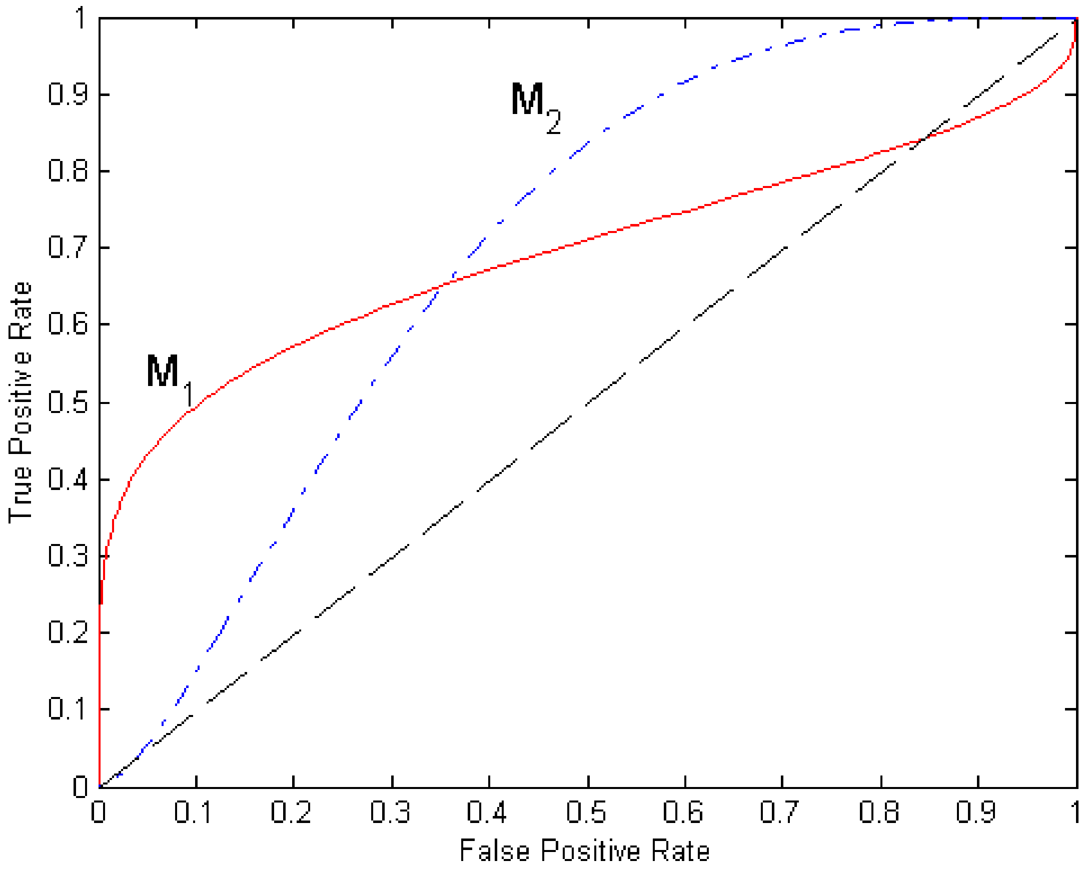


基于有限样例绘制的 ROC 曲线与 AUC

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

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

Using ROC for Model Comparison



- No model consistently outperform the other
  - M_1 is better for small FPR
  - M_2 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	
	Class=Yes	Class=No
ACTUAL CLASS Class=Yes	90	10
ACTUAL CLASS Class=No	10	1,999,890

M2	PREDICTED CLASS	
	Class=Yes	Class=No
ACTUAL CLASS Class=Yes	90	10
ACTUAL CLASS 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	
ACTUAL CLASS C(i|j)	**Class=Yes**	**Class=No**
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。

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

标准的准确率度量趋向于M2优于M1

Model M_1	PREDICTED CLASS		
ACTUAL CLASS		+	-
	+	220	30
	-	70	180

Model M_2	PREDICTED CLASS		
ACTUAL CLASS		+	-
	+	210	40
	-	10	240

Accuracy = 80%

Cost = 2850

Accuracy = 90%

Cost = 3800

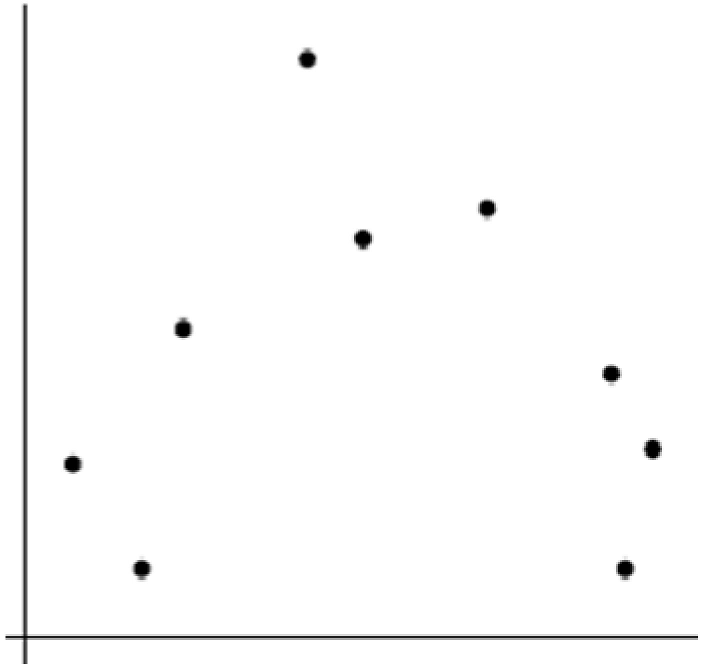
如何基于PR和AOC，开展模型的检验？——留出法（HOLDOUT）

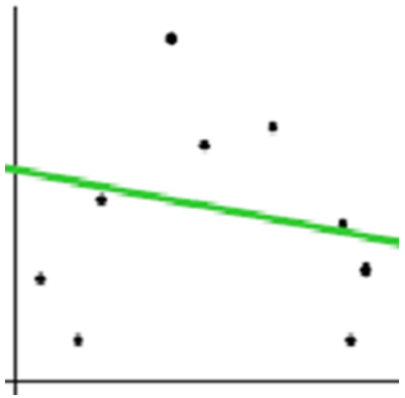
- 将原始数据集D划分为二不相交集合，一个作为训练集S，一个作为验证集T
- 在训练集上训练出模型
- 在验证集上评估模型的性能
- 二者的划分常依分析者的判断
  - 50-50
  - 2/3作为训练集，1/3作为验证集

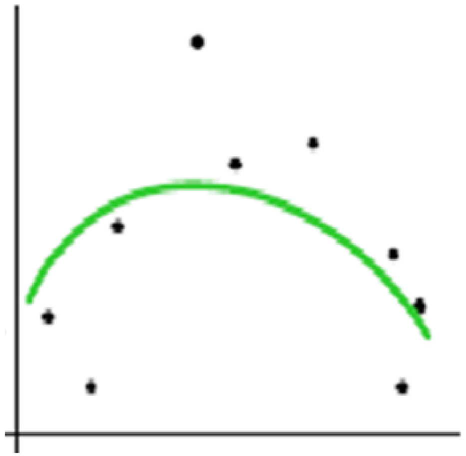
以二分类任务为例，假定 D 包含 1000 个样本，将其划分为 S 包含 700 个样本，T 包含 300 个样本，用 S 进行训练后，如果模型在 T 上有 90 个样本分类错误，那么其错误率为 $(90/300) \times 100\% = 30\%$，相应的，精度为 $1 - 30\% = 70\%$.

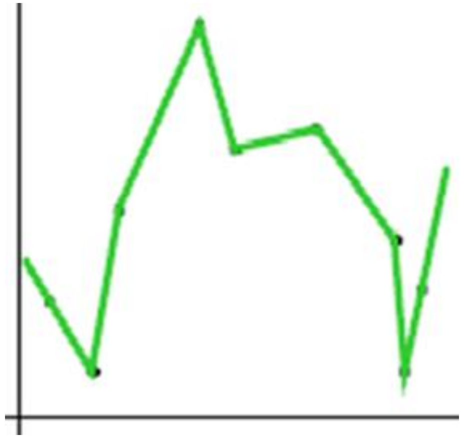
- 划分要尽可能保持原始数据分布的一致性
- 一般采用多次随机划分、重复评估后取平均值

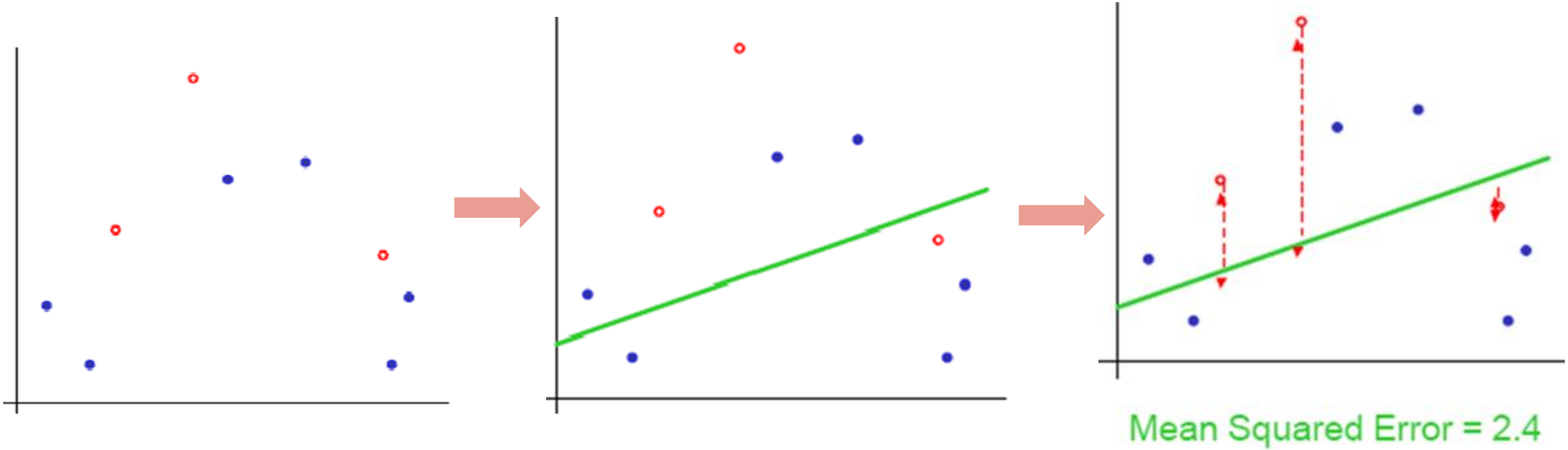
例： 回归问题

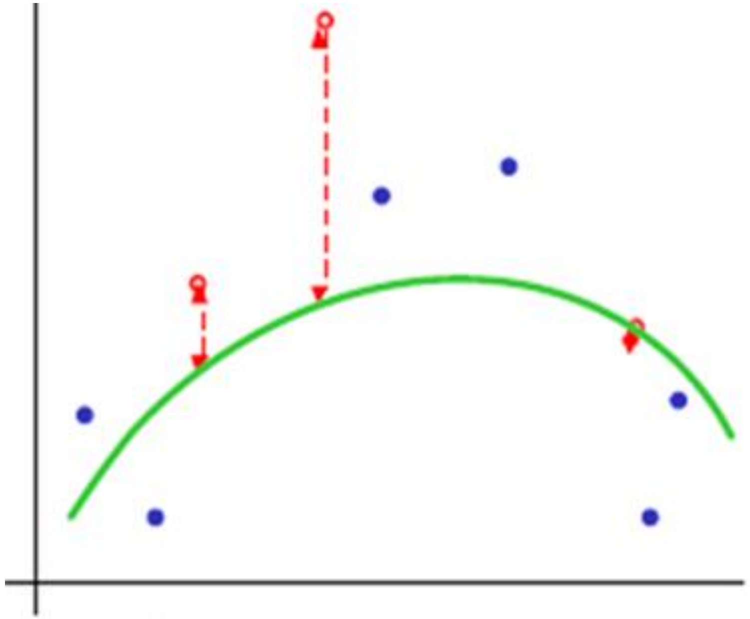




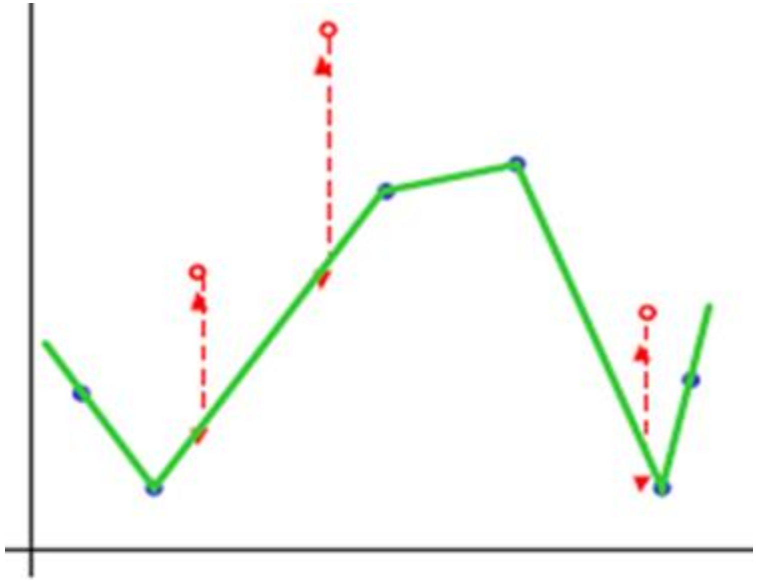





Mean Squared Error = 2.4

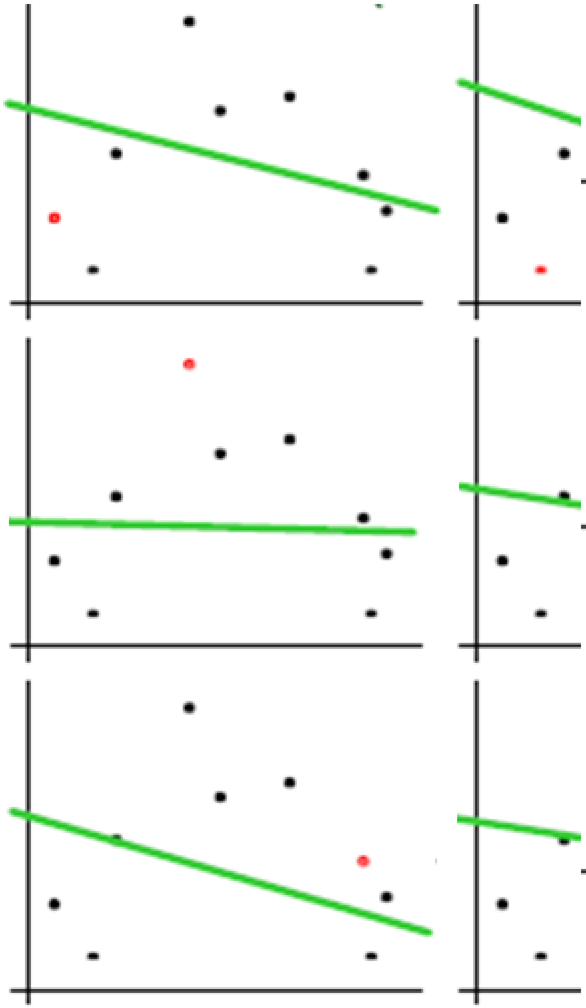


Mean Squared Error = 0.9

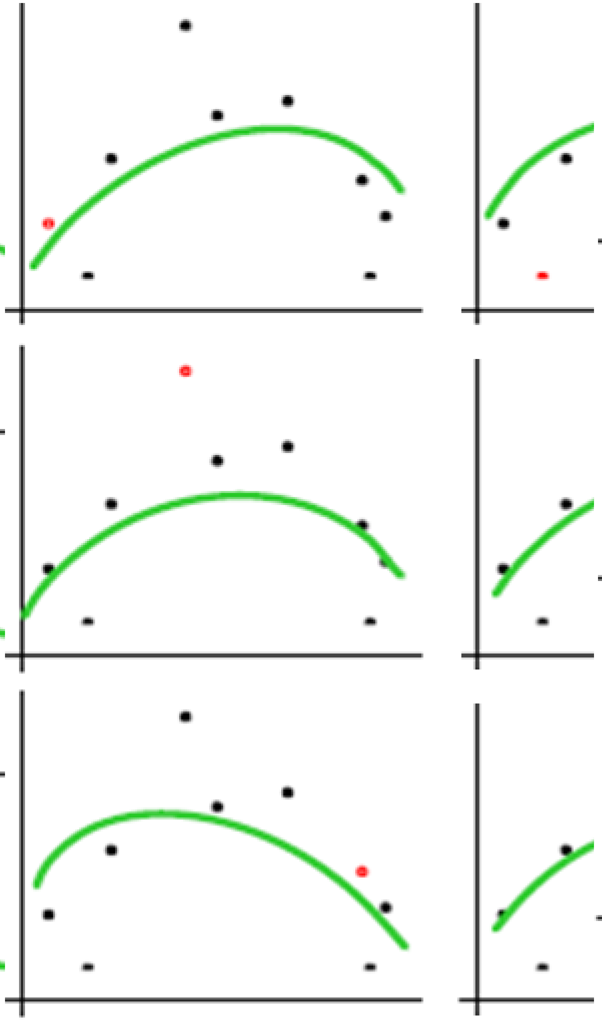


Mean Squared Error = 2.2

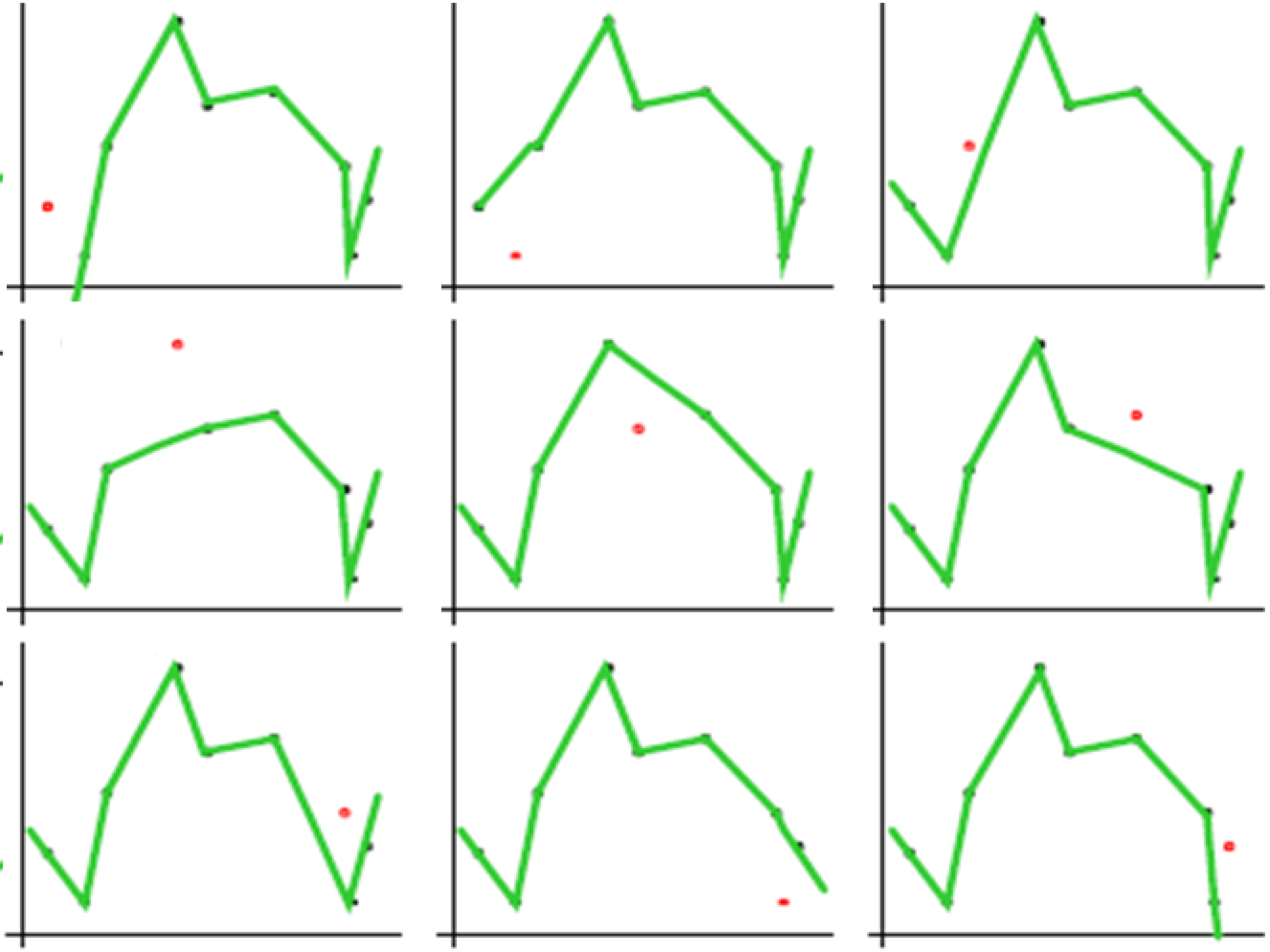
- 留一法（Leave-One-Out Cross Validation）
  - 记L为整个训练样本集的大小。
  - 对于第i个训练样本，将其取出，对剩下L-1个样本进行训练，得到模型，并用第i个训练样本对该模型的性能进行测试
  - 该过程重复L次
  - 用这L个模型的准确率的平均值作为评价模型性能的指标



$MSE_{LOOCV} = 2.12$

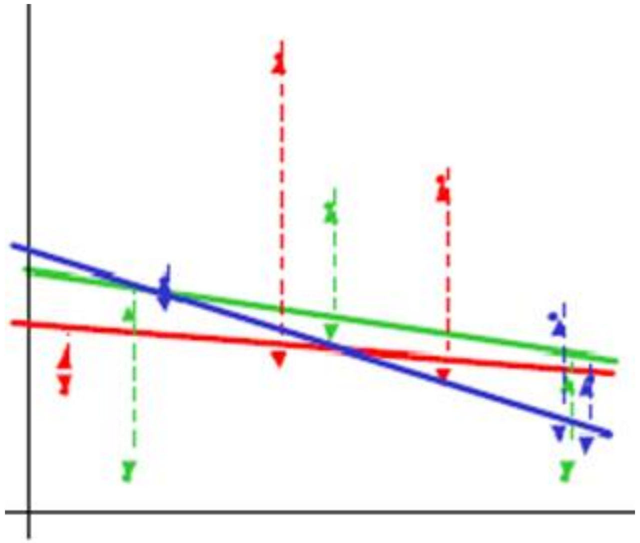


$MSE_{LOOCV} = 0.962$

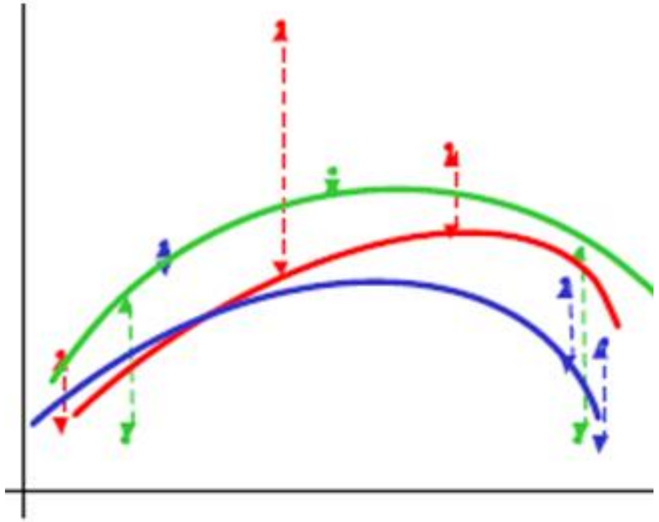


$MSE_{LOOCV} = 3.33$

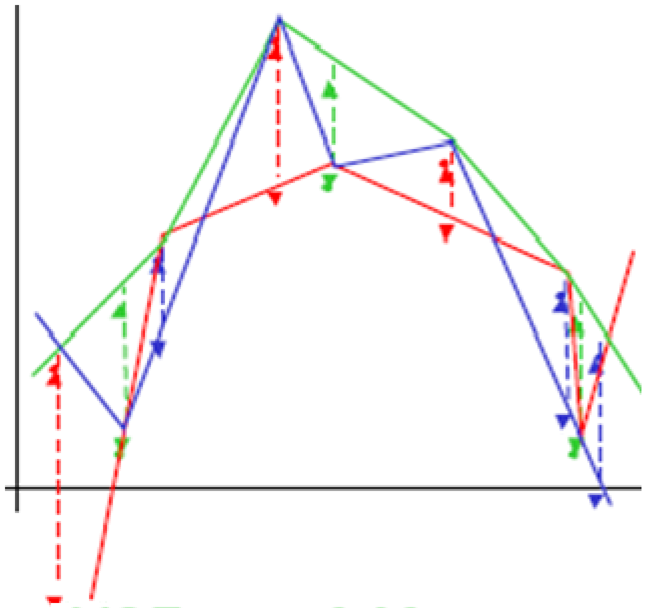
- K折交叉验证法（K-fold Cross Validation ）
  - 将训练样本集随机分为K个集合，通常分为K等份
  - 对其中的K-1个集合进行训练，得到模型，并用剩下的一个集合对该模型的性能进行测试。
  - 该过程重复K次，取K次过程中的测试错误的平均值作为评价模型性能的指标。



$MSE_{3FOLD}=2.05$



$MSE_{3FOLD}=1.11$



$MSE_{3FOLD}=2.93$

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?

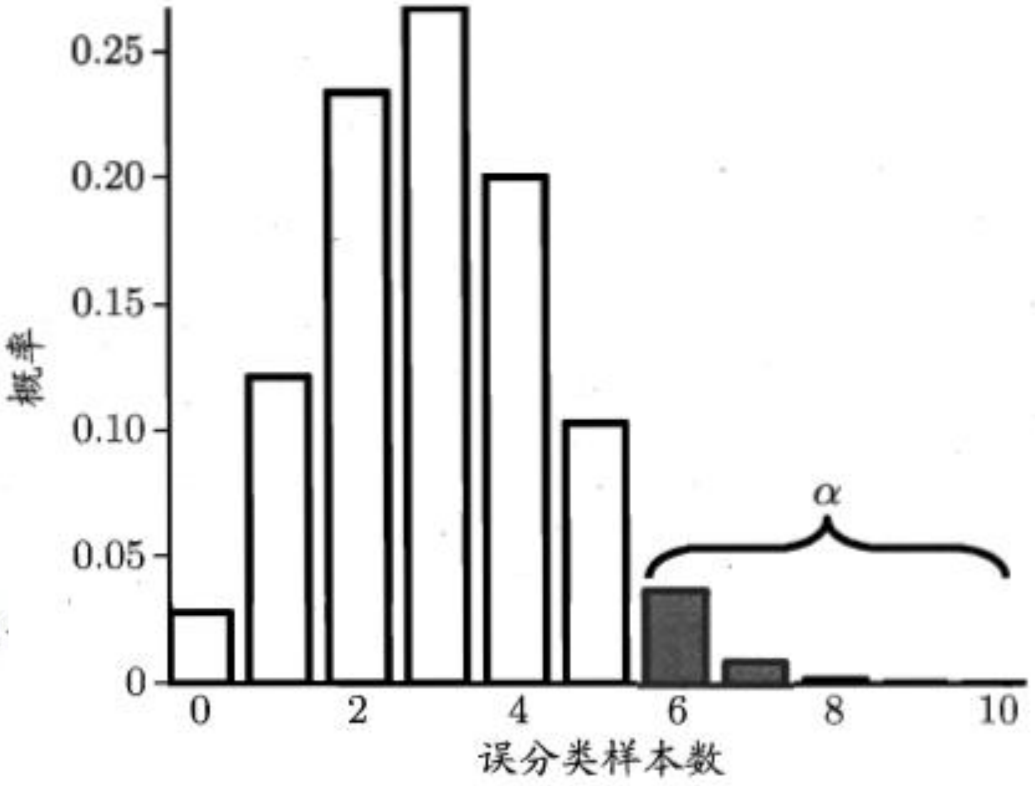
PR/ROC+交叉验证是否足够了？——不同模型（学习器）性能的比较

- 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： $\varepsilon <= \varepsilon_0$
- 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($m = 10, \epsilon_0 = 0.3$)

模型检验

- H$_0$： $\varepsilon <= \varepsilon_0$
- 近似正态检验

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

模型检验

- H_0： ε <=$ε_0$
- t检验（交叉验证法）

• 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。

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

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

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

$$\tau_t = \frac{\sqrt{k}(\mu - \epsilon_0)}{\sigma}$$ 服从自由度为 $k-1$ 的 t 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

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

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	0.584–0.919	0.670–0.888	0.711–0.867	0.763–0.833	0.774–0.824	0.789–0.811

- 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 n1 and n2 are sufficiently large, then

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

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

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

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

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

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

  - At (1-α) 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 = |e2 – e1| = 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 $\epsilon_1^A, \epsilon_2^A, \ldots, \epsilon_k^A$ 和 $\epsilon_1^B, \epsilon_2^B, \ldots, \epsilon_k^B$
      - 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 $\bar{\Delta} = \dfrac{\sum_{i=1}^{K} \Delta_i}{K}$ $\quad S^2 = \dfrac{\sum_{i=1}^{K} \left(\Delta_i - \bar{\Delta}\right)^2}{K-1}$
      - $H_0 : \varepsilon^A - \varepsilon^B = 0$

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

- 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	算法 A 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

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

$$= \frac{12N}{k(k+1)} \left(\sum_{i=1}^{k} r_i^2 - \frac{k(k+1)^2}{4} \right)$$

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H0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， α = 0.05

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

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

$$\boxed{\tau_F = 24.429}$$

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

拒绝原假设！

Nemenyi 检验中常用的 q_α 值

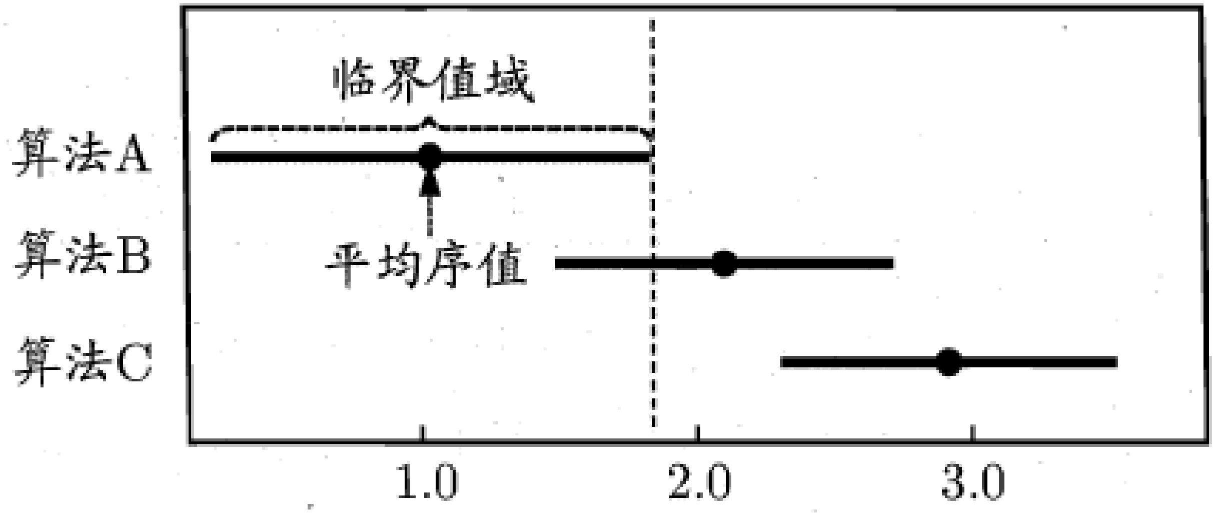
α	算法个数 k		
	2	3	4
0.05	1.960	2.344	2.569
0.1	1.645	2.052	2.291

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

$$\boxed{CD = 1.657}$$

A、C性能显著不同，AB\BC否

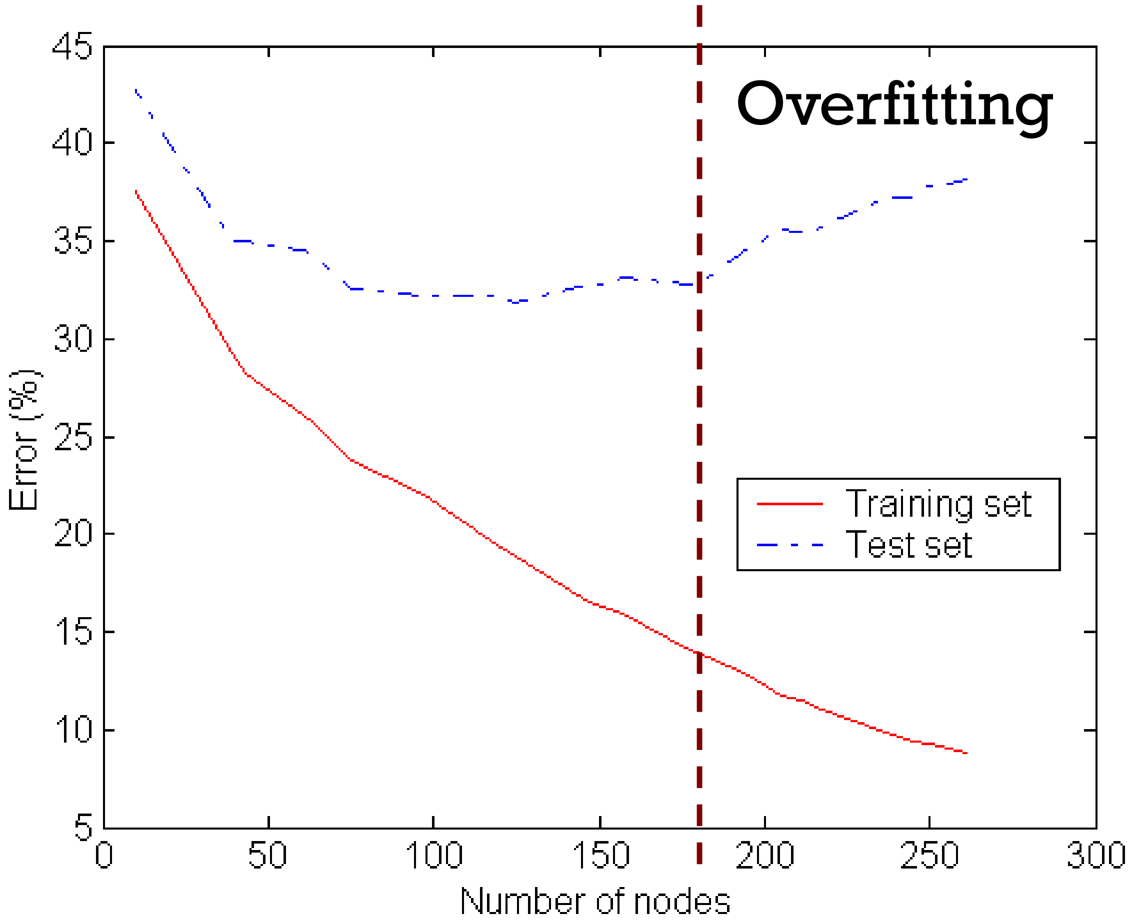
- Friedman检验图
  - 横轴：平均序值；纵轴：各算法
  - 点：每个算法的平均序值；横线段：临界值域
  - 若两个算法的横线段有交叠，则算法无显著差别；否则说明有显著差别。



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

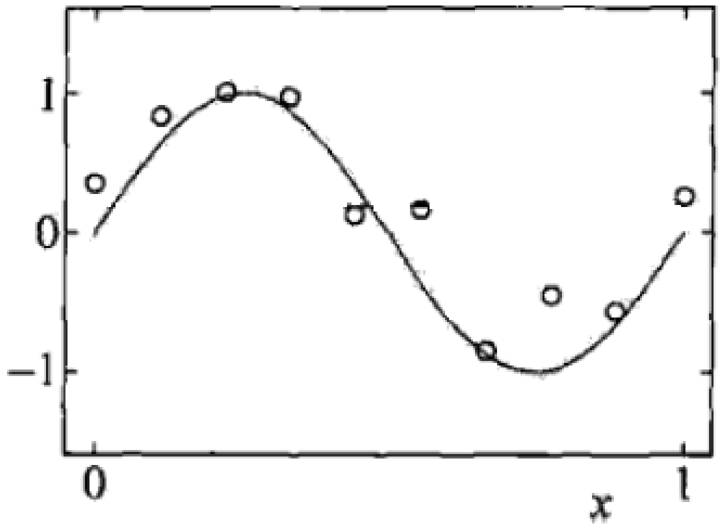
模型M1：锯齿∩绿色→树叶

模型M2：绿色→树叶

例 假设给定一个训练数据集：

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

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

$$f_M(x,w)=w_0+w_1x+w_2x^2+\cdots+w_Mx^M=\sum_{j=0}^{M}w_jx^j$$

求以下经验风险最小化：

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

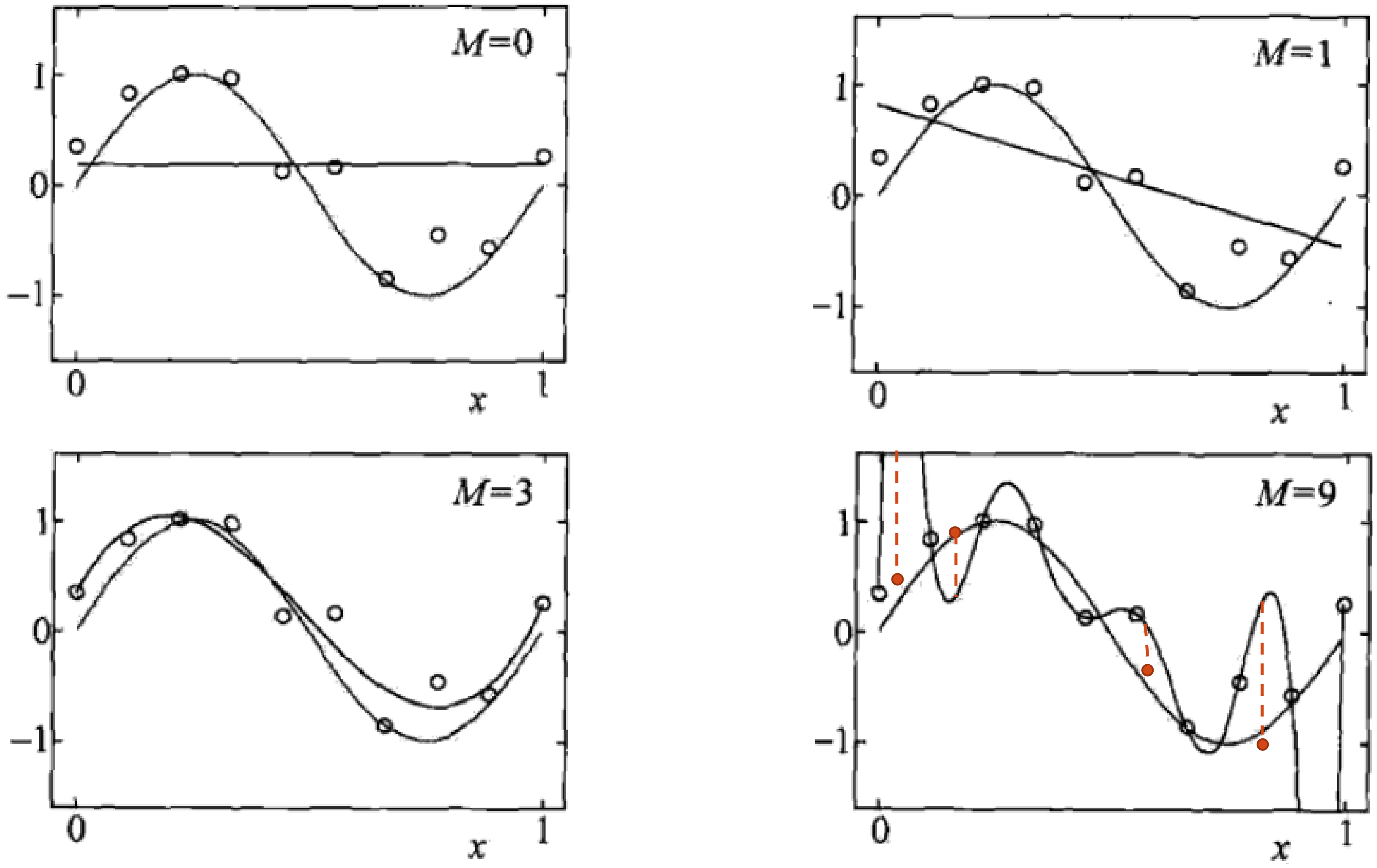
平方
损失

对 w_j 求偏导数并令其为 0，可得

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

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

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



噪声（标注错误）期望为0

噪声与模型无关

- 偏差-方差分解

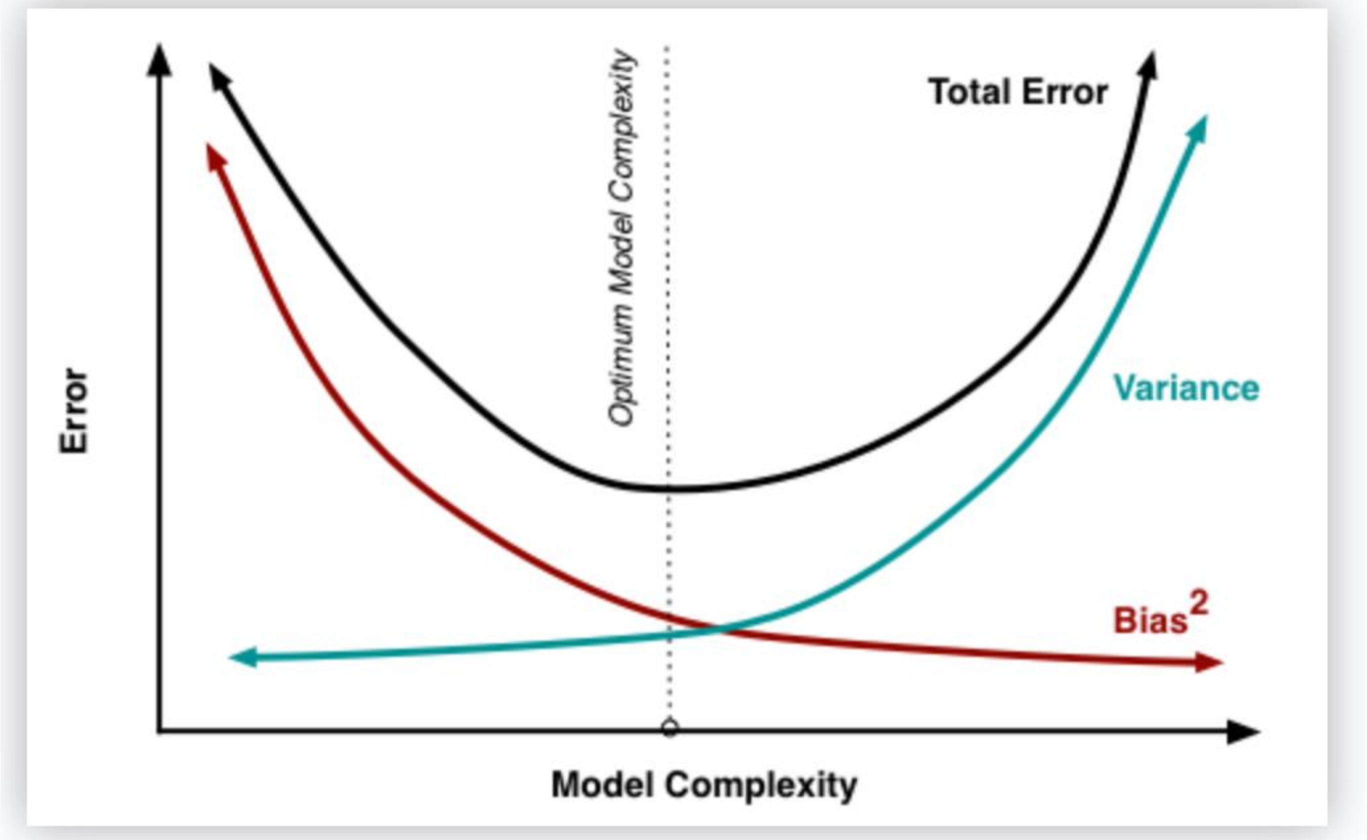
假定$\mathbb{E}_D[y_D - y] = 0.$ $\bar{f}(\boldsymbol{x}) = \mathbb{E}_D[f(\boldsymbol{x}; D)]$

符号	涵义
$\mathbf{x}$	测试样本
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

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

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

- variance
- bias^2
- noise



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