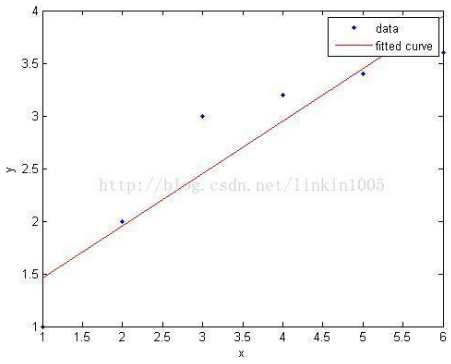
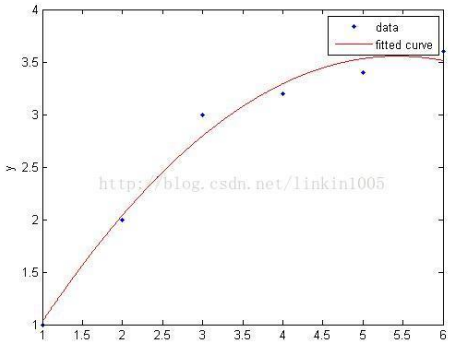
第2章 模型评估与选择

主要内容

- 经验误差与过拟合
- 评估方法
- 性能度量
- 比较检验
- 偏差与方差

误差





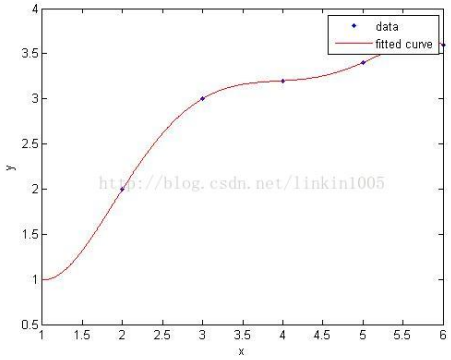


图3建立的模型，在训练集中通过x可以很好的预测y

然而，我们却不能预期该模型能够很好的预测训练集外的数据。

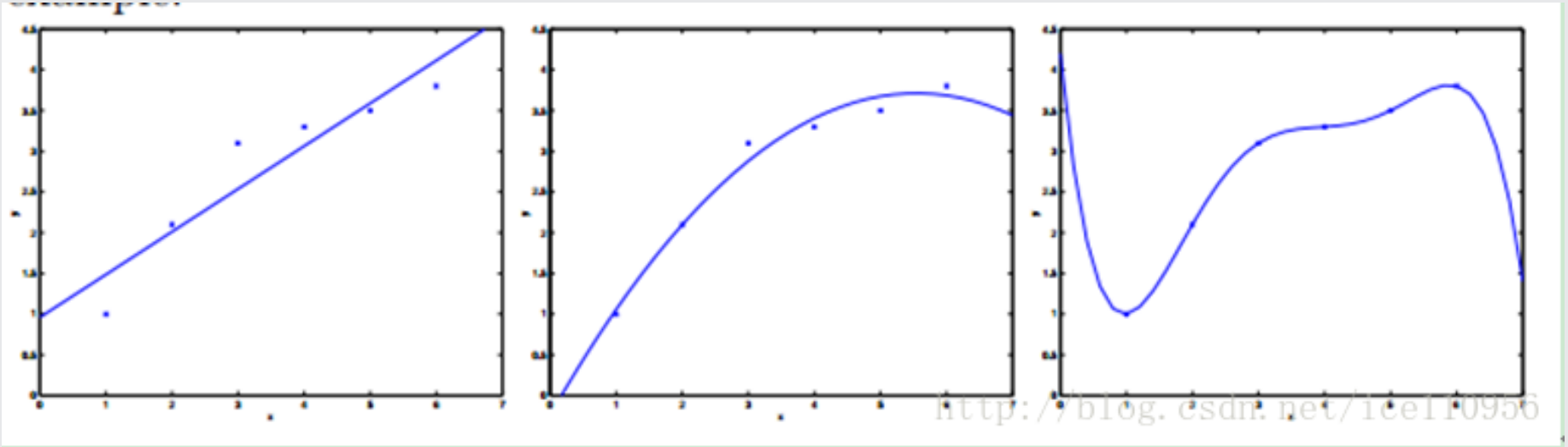
换句话说，这个模型没有很好的泛化能力。

图1和图3中的模型都有较大的泛化误差，然而他们的误差原因却不相同。

图1建立了一个线性模型，但是该模型并没有精确的捕捉到训练集数据的结构，我们称图1有较大的**偏倚（bias）**，也称欠拟合；

图3通过5次多项式函数很好的对样本进行了拟合，然而，如果将建立的模型进行泛化，并不能很好的对训练集之外数据进行预测，也称过拟合。

误差



误差

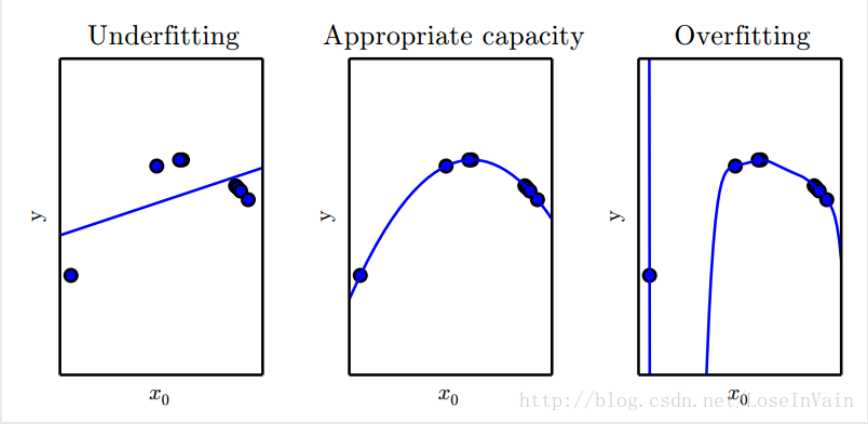
- 机器学习的主要挑战在于**在未见过的数据输入上表现良好**，这个能力称为**泛化能力**（generalization）。
- 而我们的机器学习模型都是从**训练集**中学习参数得到的，如何确保其在和训练集“隔离”的测试集中表现良好呢？

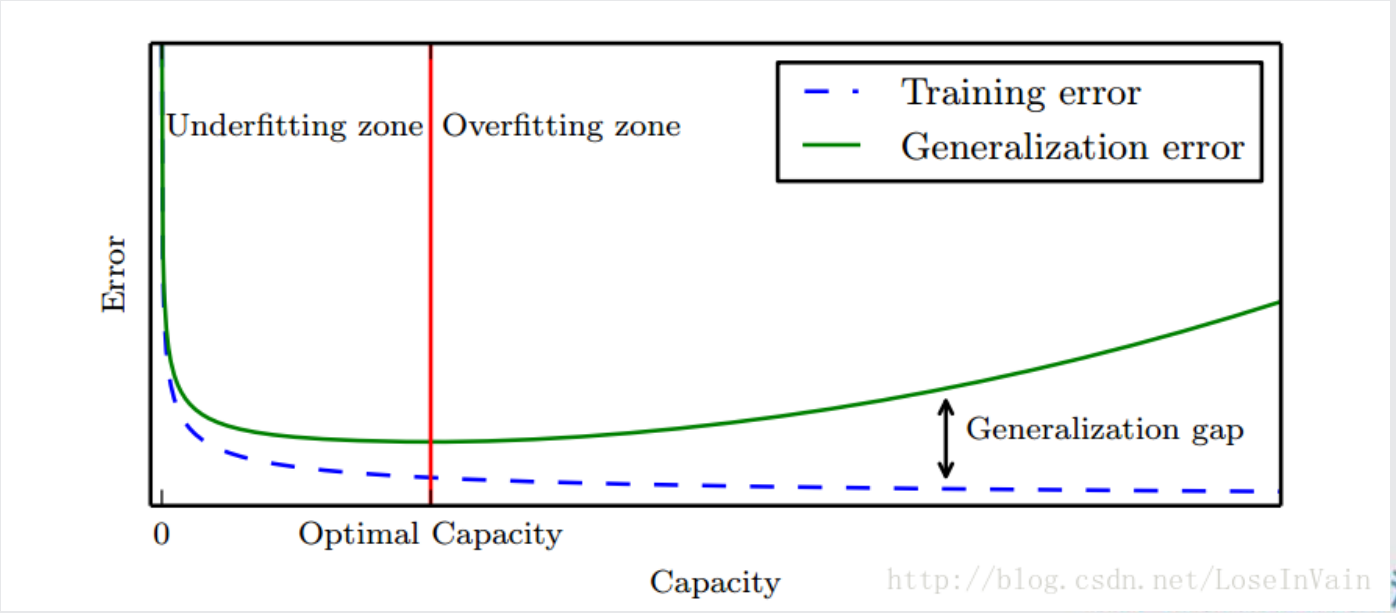
误差

- 误差（error）：学习器实际预测输出与样本真实输出之间的差异
  - 训练集：训练误差（training error），（经验误差，empirical error）
  - 训练集的补集：泛化误差（generalization error）

- 我们希望泛化误差小的学习器

误差





过拟合

- 过拟合（overfitting）：训练过度使泛化能力下降
- 欠拟合（underfitting）：未能学好训练样本的普遍规律

- 过拟合是机器学习的关键障碍且不可避免!
- 模型误差包含了数据误差，或者说模型信息中包含了噪声。

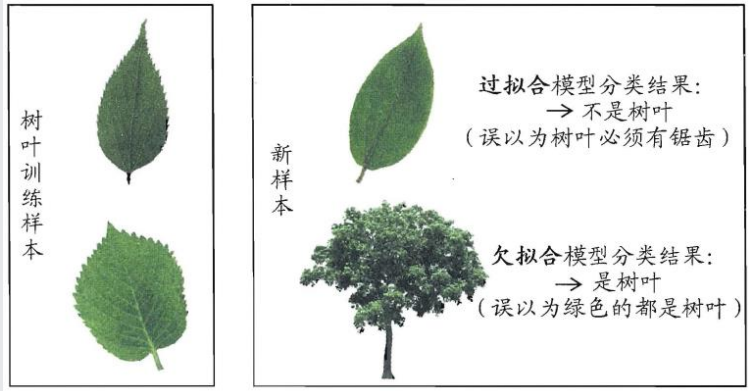
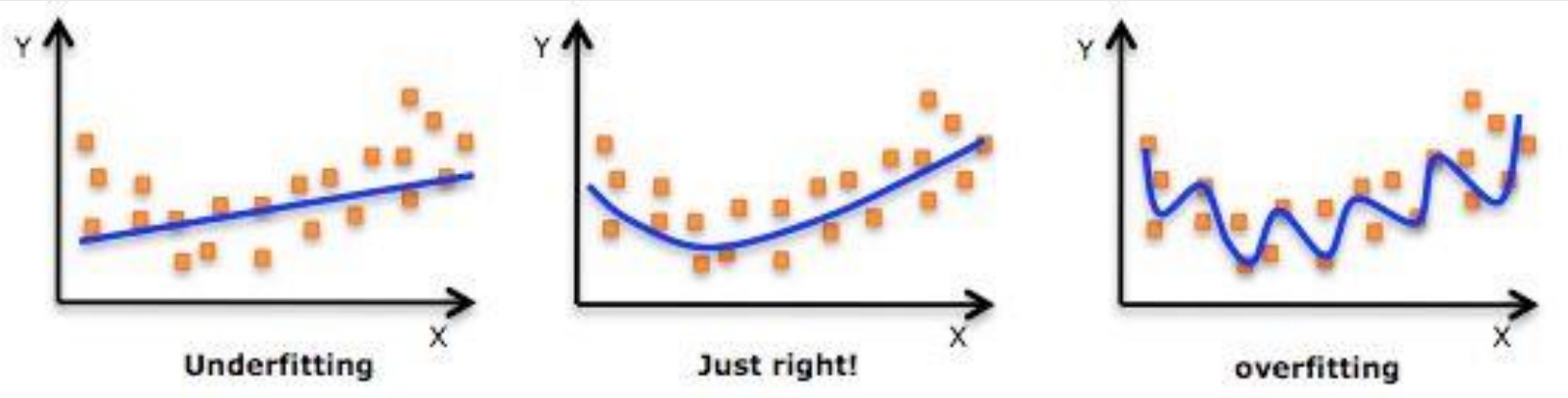


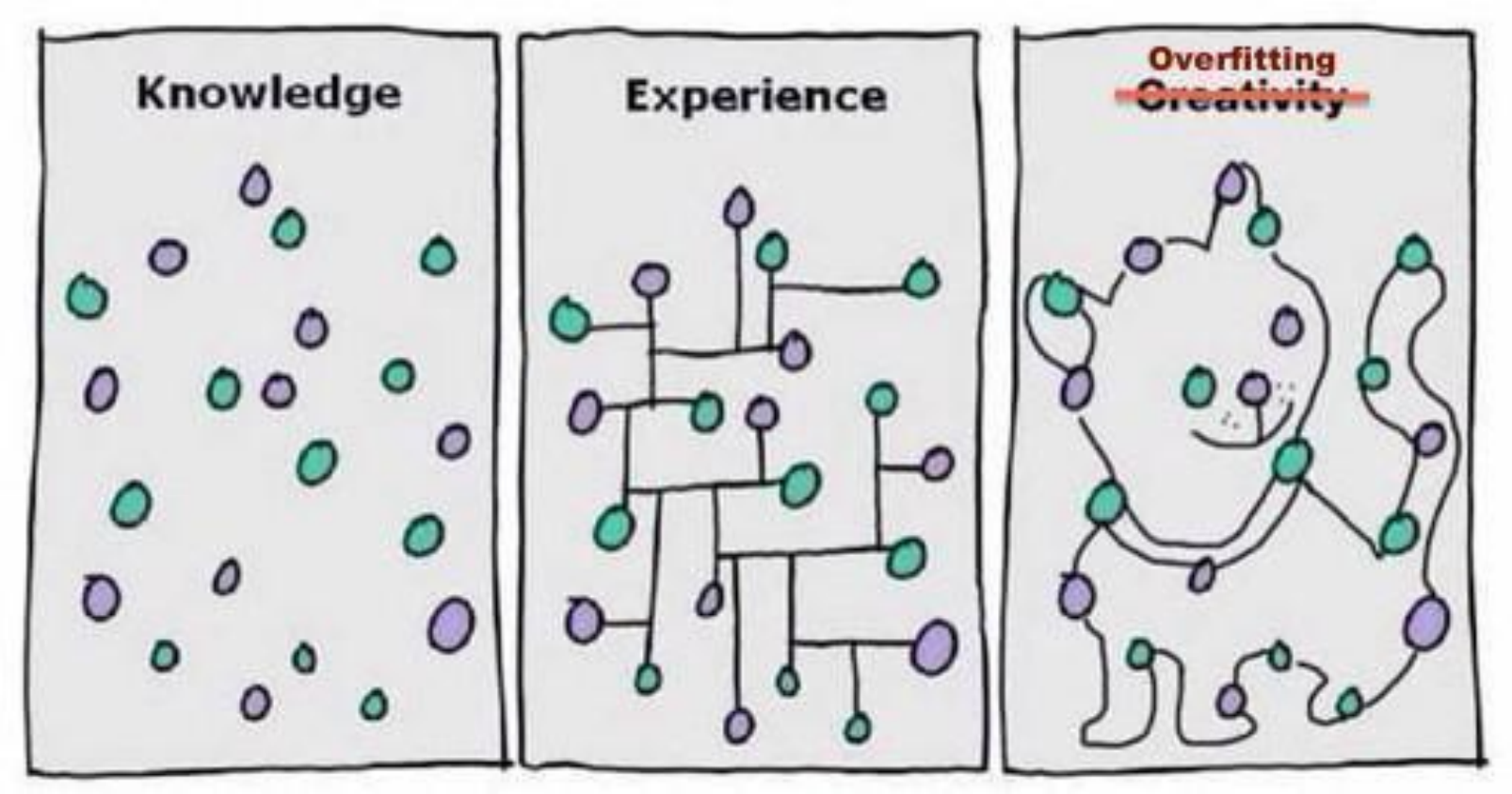
图 2.1 过拟合、欠拟合的直观类比

过拟合

- 上学考试的时候，有的人采取题海战术，把每个题目都背下来。但是题目稍微一变，他就不会做了。
- 因为他非常复杂的记住了每道题的做法，而没有抽象出通用的规则。



过拟合

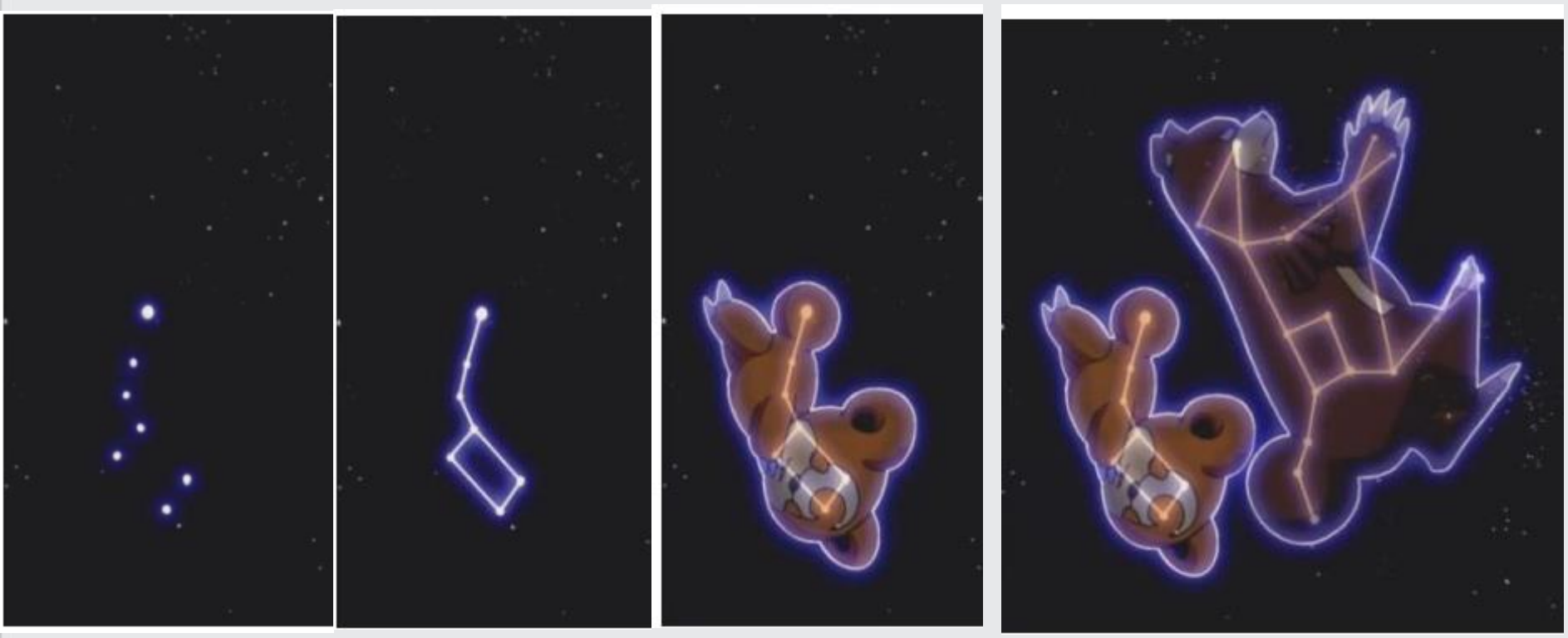


过拟合


Aquarius
Pisces
Aries
Gemini
Taurus
Scorpio 10.24-11.22
Libra 9.23-10.23
Virgo 8.23-9.22
Cancer
Leo 7.23-8.22
Sagittarius
Capricorn

过拟合

过拟合

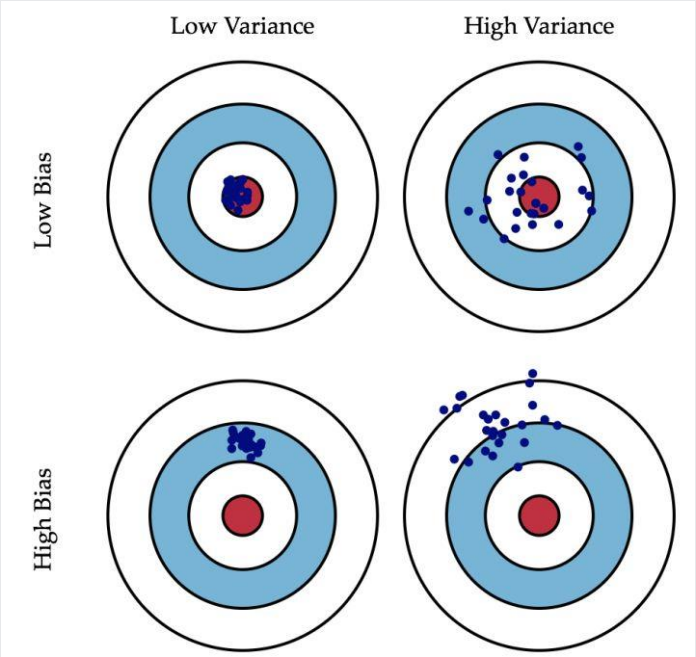


学习器泛化评估——实验测试

- 测试集：测试误差（testing error）
- 训练集S和测试集T组成数据集D。
- 假设测试样本是从真实分布中采样而得，避免因数据划分引入偏差。
- 测试集应与训练集互斥。

测试方法	数学表达	注意事项	优缺点
留出法 (hold-out)	$D = S \cup T$ $S \cap T = \emptyset$	分层采样（stratified sampling） 重复试验取平均评估结果	测试集小，评估结果方差较大 训练集小，评估结果偏差较大
交叉验证法 (cross validation)	$D = D_1 \cup \cdots \cup D_k$ $D_i \cap D_j = \emptyset\ (i \neq j)$	p次k折交叉验证	稳定性和保真性很大程度取决于k
留一法 (Leave-One-Out，LOO)	$D = D_1 \cup \cdots \cup D_k$ $D_i \cap D_j = \emptyset\ (i \neq j)$ $k = \lvert D \rvert$	每次使用一个样本验证	不受随机样本划分方式影响 数据量大时计算量大
自助法 (bootstrapping)	$\lvert S \rvert = \lvert D \rvert$ $T = D \backslash S$	可重复采样/有放回采样	数据集较小有用 改变初始数据集的分布，引入偏差

想象你开着一架黑鹰直升机，攻击地面上一只敌军部队，于是你连打数十梭子，结果有一下几种情况:



1.子弹一颗没浪费，每一颗都打死一个敌军，跟抗战神剧一样，这就是方差小（子弹全部都集中在一个位置），偏差小（子弹集中的位置正是它应该射向的位置）

2.子弹打死了一部分敌军，但是也打偏了些打到花花草草了，这就是方差大（子弹不集中），偏差小（已经在目标周围了）。

3.子弹基本上都打在队伍经过的一棵树上了，这就是方差小（子弹打得很集中），偏差大（跟目的相距甚远）。

4.子弹打在了树上，石头上，花花草草也都中弹，但是敌军安然无恙，这就是方差大（子弹到处都是），偏差大

测试方法	数学表达	注意事项	优缺点
留出法 (hold-out)	$D = S \cup T$ $S \cap T = \emptyset$	分层采样（stratified sampling） 重复试验取平均评估结果	测试集小，评估结果方差较大 训练集小，评估结果偏差较大
交叉验证法 (cross validation)	$D = D_1 \cup \cdots \cup D_k$ $D_i \cap D_j = \emptyset\ (i \neq j)$	p次k折交叉验证	稳定性和保真性很大程度取决于k
留一法 (Leave-One-Out，LOO)	$D = D_1 \cup \cdots \cup D_k$ $D_i \cap D_j = \emptyset\ (i \neq j)$ $k = \lvert D \rvert$	每次使用一个样本验证	不受随机样本划分方式影响 数据量大时计算量大
自助法 (bootstrapping)	$\lvert S \rvert = \lvert D \rvert$ $T = D \backslash S$	可重复采样/有放回采样	数据集较小有用 改变初始数据集的分布，引入偏差

调参与最终模型

- 参数调节（parameter tuning）
  - 算法参数 → 人工设定候选值
  - 模型参数 → 通过学习产生候选模型
- 数据集
  - 训练集 → 训练估计模型
  - 验证集 → 模型参数调整
  - 测试集 → 估计泛化能力
- 学习算法和参数配置确定后要用整个数据集重新训练模型

性能度量

- 性能度量（performance measure）：衡量模型泛化能力的评价标准

- 回归（regression）：均方误差（mean squared error）
  - 离散数据：$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}(f(\boldsymbol{x}_i)-y_i)^2$
  - 连续数据：$E(f;\mathcal{D})=\int_{\boldsymbol{x}\sim\mathcal{D}}(f(\boldsymbol{x})-y)^2p(\boldsymbol{x})d\boldsymbol{x}$

- 分类（classification）：错误率（error rate）和精度（accuracy）
  - 离散数据：$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(\boldsymbol{x}_i)\neq y_i)$, $\mathrm{acc}(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(\boldsymbol{x}_i)=y_i)$
  - 连续数据：$E(f;\mathcal{D})=\int_{\boldsymbol{x}\sim\mathcal{D}}\mathbb{I}(f(\boldsymbol{x})\neq y)p(\boldsymbol{x})d\boldsymbol{x}$, $\mathrm{acc}(f;\mathcal{D})=\int_{\boldsymbol{x}\sim\mathcal{D}}\mathbb{I}(f(\boldsymbol{x})=y)p(\boldsymbol{x})d\boldsymbol{x}$

任务需求——以二分类为例

- 混淆矩阵（confusion matrix），非对角，纠缠相
- 查准率（precision）：$P = \frac{TP}{TP+FP}$
- 查全率（recall）：$R = \frac{TP}{TP+FN}$

表 2.1 分类结果混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP (真正例)	FN (假反例)
反例	FP (假正例)	TN (真反例)

- P-R曲线
  - 面积、平衡点
  - $F1$度量：$\frac{P*R}{2(P+R)}$

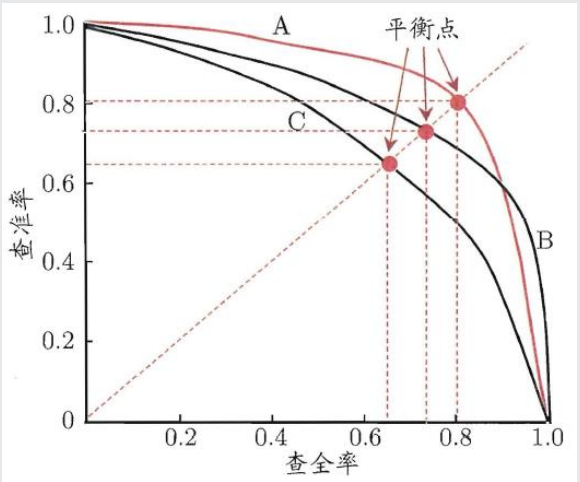
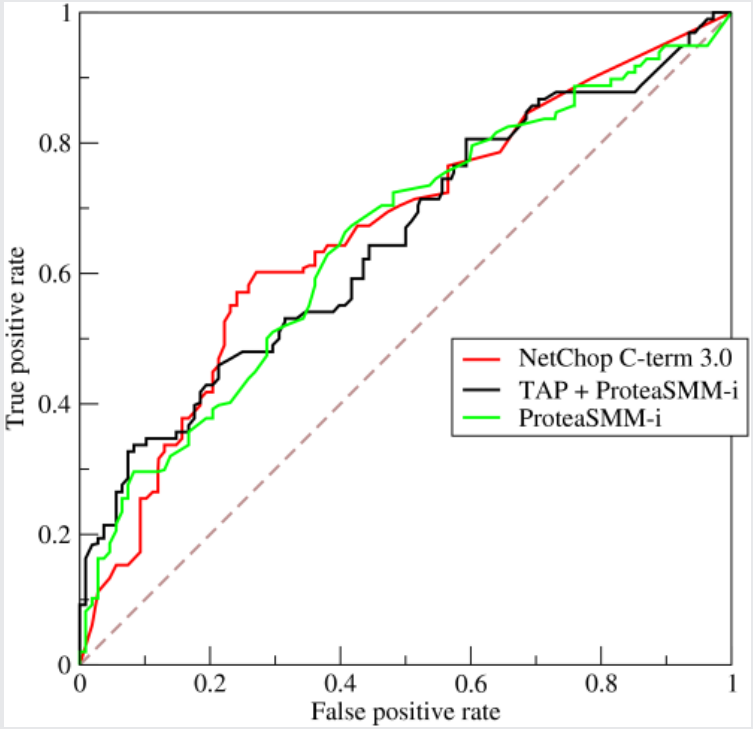


图 2.3 P-R曲线与平衡点示意图

ROC

- 受试者工作特征曲线（Receiver Operating Characteristic）
  - 横轴——假正例率：$\mathrm{FPR} = \frac{FP}{TN+FP}$
  - 纵轴——真正利率：$\mathrm{TPR} = \frac{TP}{TP+FN}$



考虑ROC曲线图中的四个点和一条线。

第一个点，(0,1)，即FPR=0, TPR=1，这意味着FN（false negative）=0，并且FP（false positive）=0。这是一个完美的分类器，它将所有的样本都正确分类。

第二个点，(1,0)，即FPR=1，TPR=0，类似地分析可以发现这是一个最糟糕的分类器，因为它成功避开了所有的正确答案。

第三个点，(0,0)，即FPR=TPR=0，即FP（false positive）=TP（true positive）=0，可以发现该分类器预测所有的样本都为负样本（negative）。

第四个点（1,1），分类器实际上预测所有的样本都为正样本。

经过以上的分析，我们可以断言，ROC曲线越接近左上角，该分类器的性能越好。

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

我们根据每个测试样本属于正样本的概率值从大到小排序。图中共有20个测试样本，“Class”一栏表示每个测试样本真正的标签（p表示正样本，n表示负样本），“Score”表示每个测试样本属于正样本的概率。

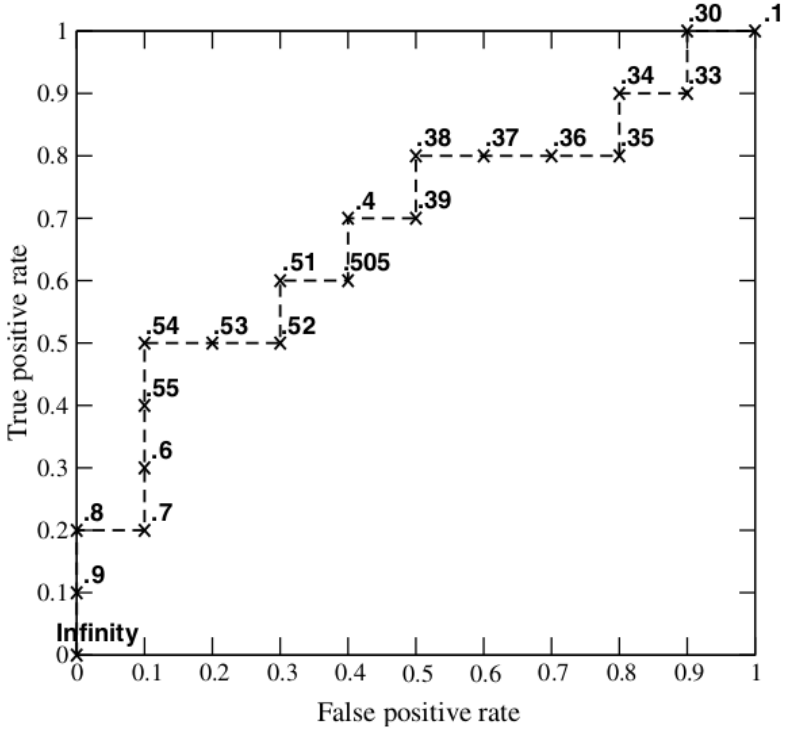
接下来，我们从高到低，依次将“Score”值作为阈值threshold

当测试样本属于正样本的概率大于或等于这个threshold时，我们认为它为正样本，否则为负样本。

举例来说，对于图中的第4个样本，其“Score”值为0.6，那么样本1，2，3，4都被认为是正样本，因为它们的“Score”值都大于等于0.6，而其他样本则都认为是负样本。

每次选取一个不同的threshold，我们就可以得到一组FPR和TPR，即ROC曲线上的一点

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



代价敏感错误率

- 非均等代价（unequal cost）
- 代价矩阵（cost matrix）
- 代价敏感错误率：加权的错误率

表 2.2　二分类代价矩阵

真实类别	预测类别	
	第 0 类	第 1 类
第 0 类	0	$cost_{01}$
第 1 类	$cost_{10}$	0

比较检验

- 如何比较？——从统计的角度
- 统计假设检验（hypothesis test）：根据测试错误率估计推断泛化错误率的分布。
- 提出假设→找到符合某种概率分布的中间变量→利用该概率分布确定在某个置信度（confidence）下是否接受该假设

单个学习器

- 做了多次留出法或者交叉验证法之后，会有多个测试误差率，此时使用“t检验”(t-test)来检验单个学习器做了多次留出法或者交叉验证法之后，会有多个测试误差率，此时使用“t检验”(t-test)来检验单个学习
- 做了多次留出法或者交叉验证法之后，会有多个测试误差率，此时使用“t检验”(t-test)来检验单个学习

$$\tau_t = \frac{\sqrt{k}(\mu - \epsilon_0)}{\sigma}$$

根据预先设定的显著度α，以及自由度k-1，查表可得临界值b

如果Tt小于临界值b则接受，否则，拒绝

一个数据集多个学习器

对一组样本D，进行k折交叉验证，会产生k个测试误差率，将两个学习器都分别在每对数据子集上进行训练与测试，会分别产生两组测试误差率

$$\epsilon_1^A, \epsilon_2^A, \ldots, \epsilon_k^A \text{ 和 } \epsilon_1^B, \epsilon_2^B, \ldots, \epsilon_k^B$$

对每对结果求差值

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

若两个学习器的性能相同，则相对应的两个误差率的差值应该为0

先计算出差值的均值μ与方差σ^2,在显著度α下，若变量

$$\tau_t = \left| \frac{\sqrt{k}\mu}{\sigma} \right|$$

根据预先设定的显著度α，以及自由度k-1，查表可得临界值b

如果Tt小于临界值b则接受，否则，拒绝

代码时间