Model Evaluation and Selection

Model Selection

Model selection: the task of selecting a statistical model from a set of candidate models.

e.g. SVM, logistic regression, etc.

- or choosing among different hyperparameters for the same machine learning model.

e.g. k in k-means, kernels in SVM, etc.

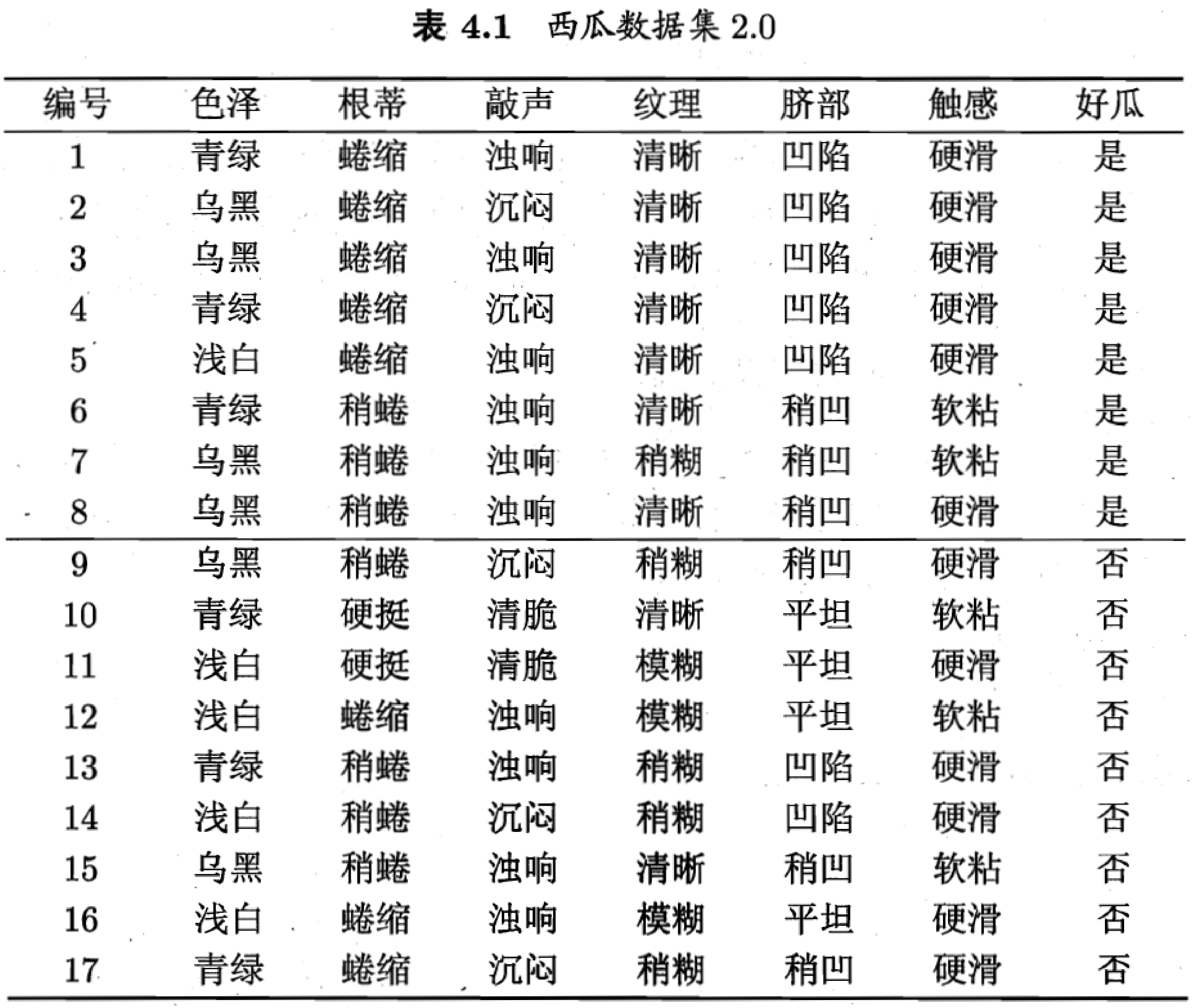
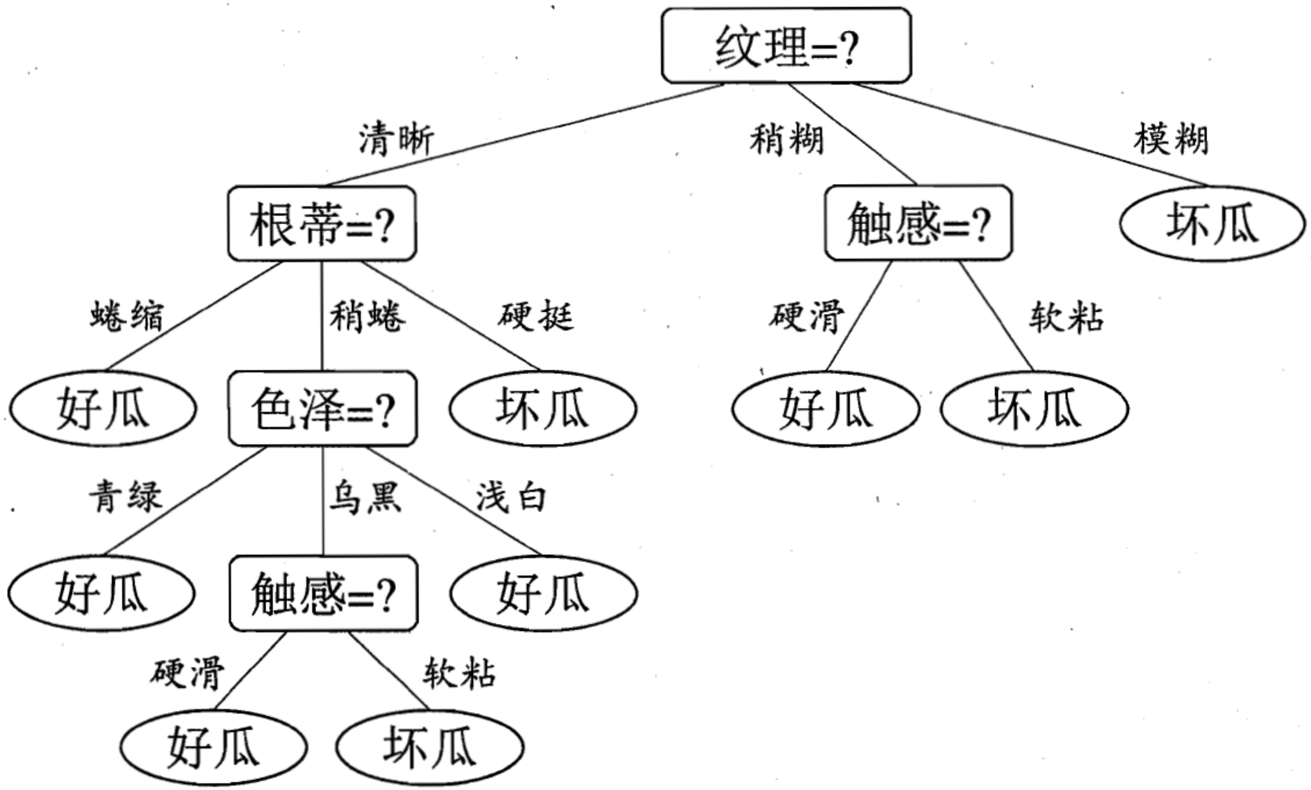


表 4.1　西瓜数据集 2.0

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

买西瓜：　青绿　硬挺　清脆　稍糊　凹陷　硬滑　？

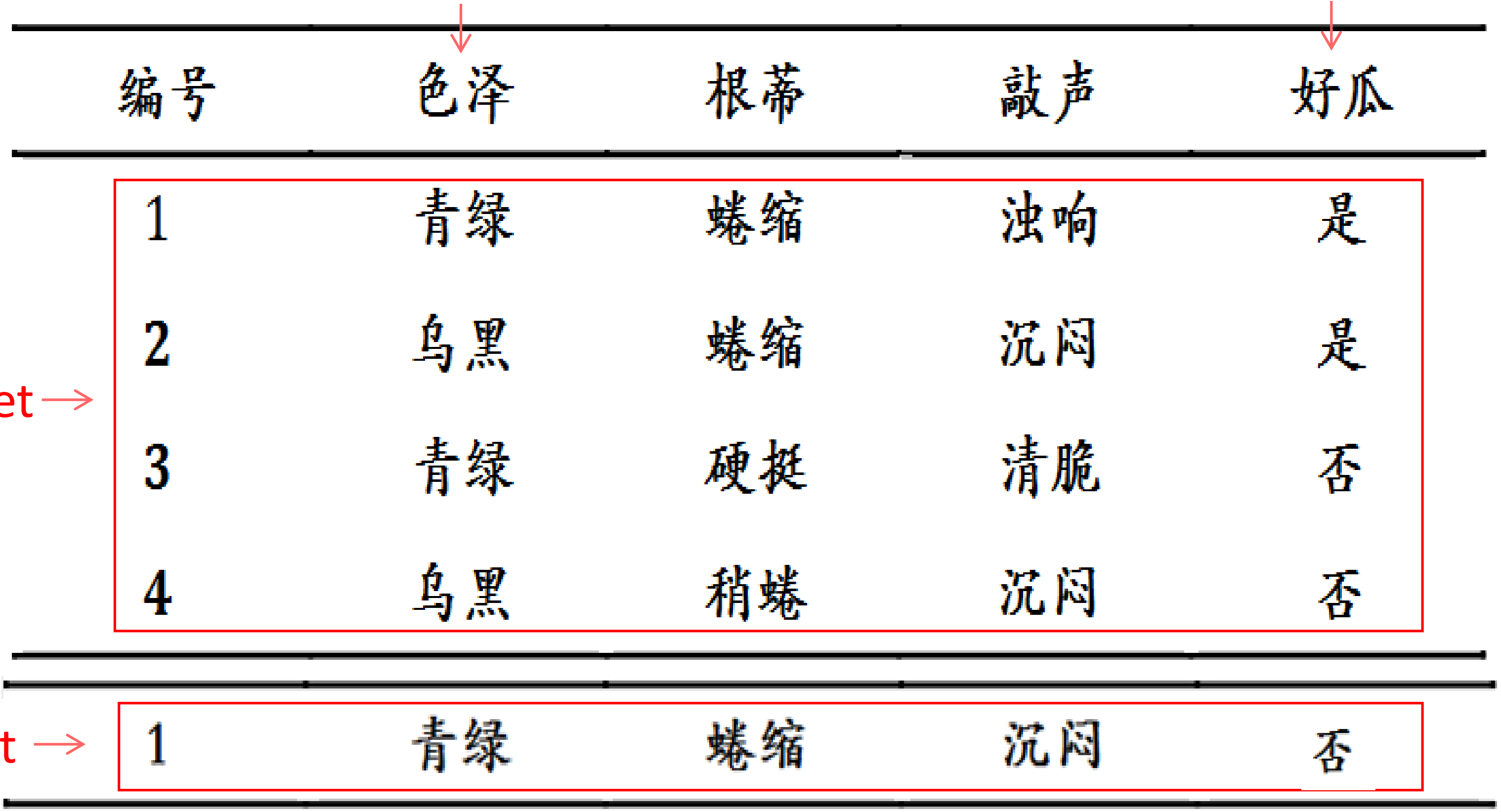
Decision Tree—an example



Basic terms-Data

features

labels



training set →

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	沉闷	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否

testing set →

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	沉闷	否

Outline

- Empirical Error and Overfitting (经验误差与过拟合)
- Evaluation Methods
- Performance Measure
- Comparison Test
- Bias and Variance
- Extension

Outline

- Empirical Error and Overfitting (经验误差与过拟合)
- Evaluation Methods
- Performance Measure
- Comparison test
- Bias and variance
- Extension

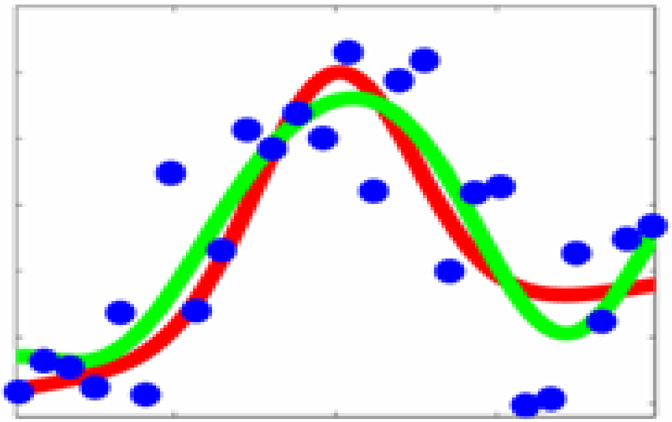
Empirical Error and Overfitting

- Error & Error Rate:
  - **Error :** the difference between real output of the sample's and predicted output.
    - **Training(empirical) Error :** Errors on the training set.
    - **Generalization Error :** Errors on new samples.
  - **Error rate:** the proportion of misclassified samples.

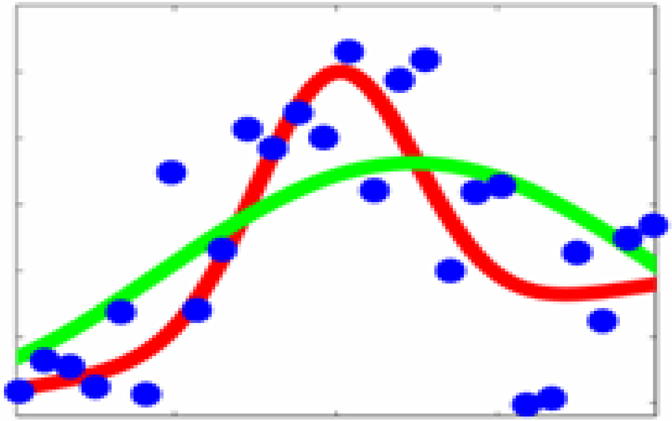
Goal: A learner with minimum Generalization Error.

Model Selection

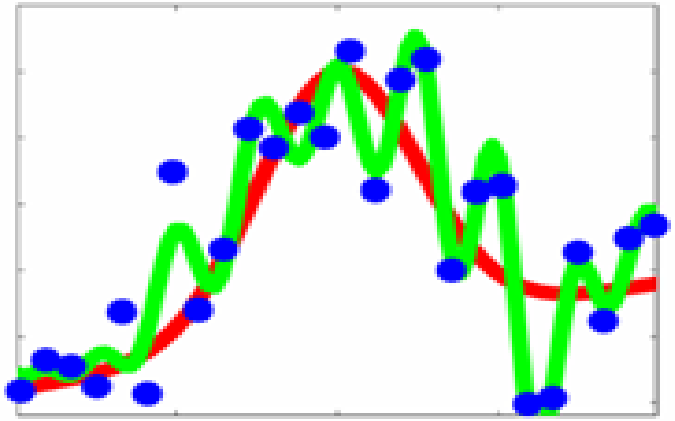
Learned function
with appropriate model



Learned function
with too simple model



Learned function
with too complex model



Goal: Choose appropriate model

Empirical error and Overfitting

- Overfitting:

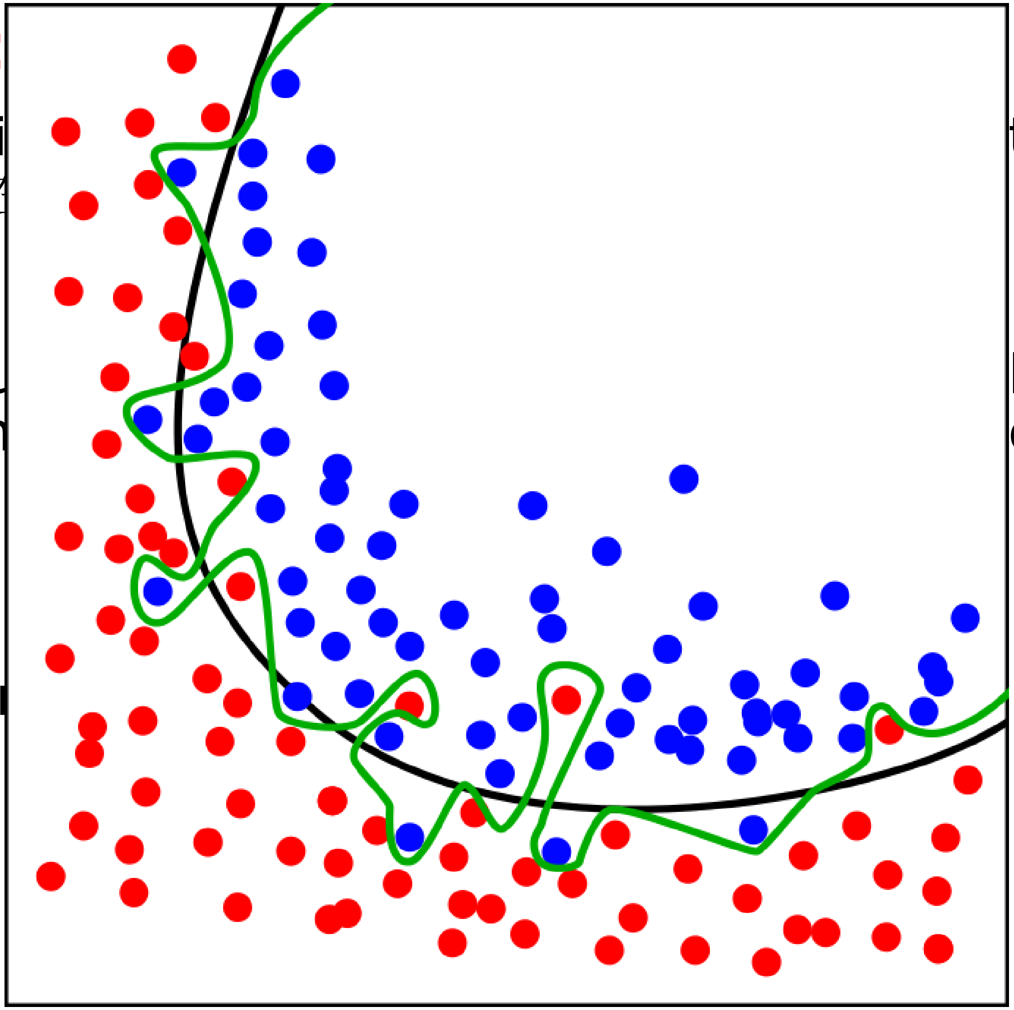
 A model descri the underlying
relationship(潜

Overfitting occu lex, such as
having too man of observations.

- Solutions:

  Regularizatio

  early stop.



Empirical error and Overfitting

- Underf

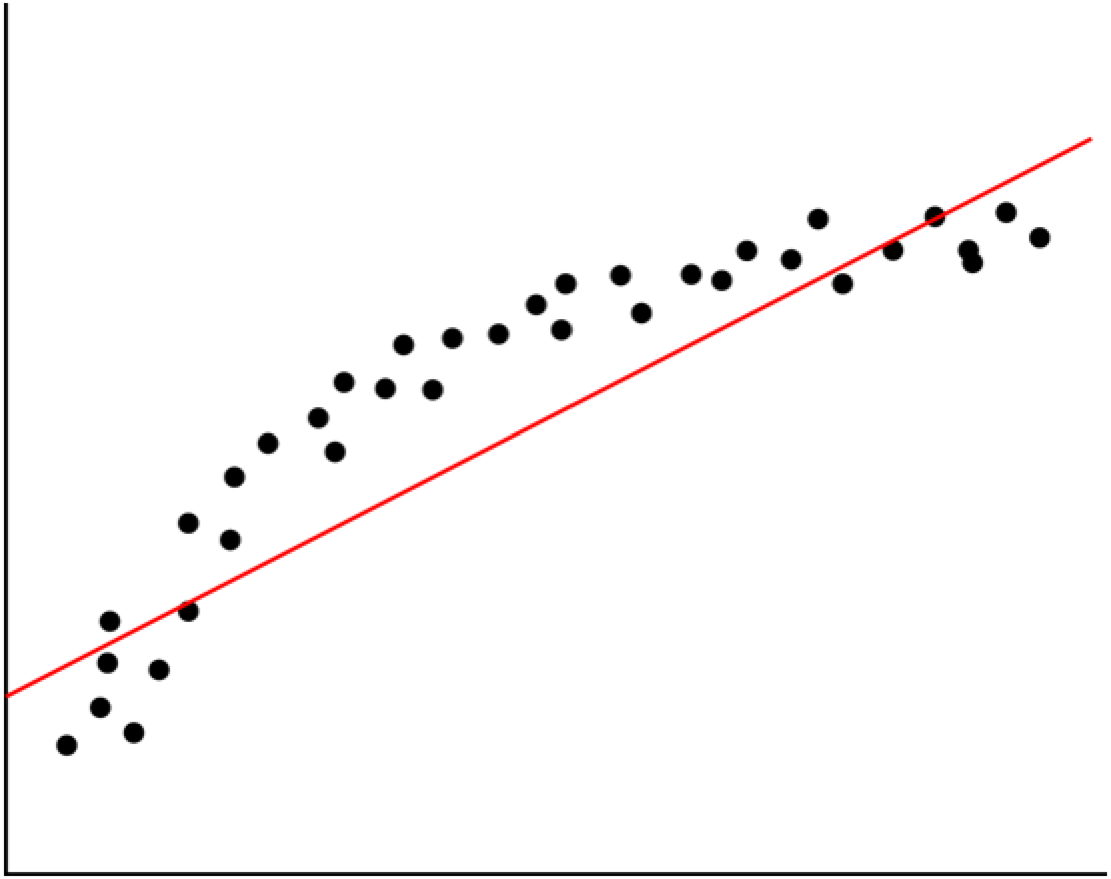
Underfitt e learning
algorithm ata.

Underfitti near model
to non-lin ictive
performa

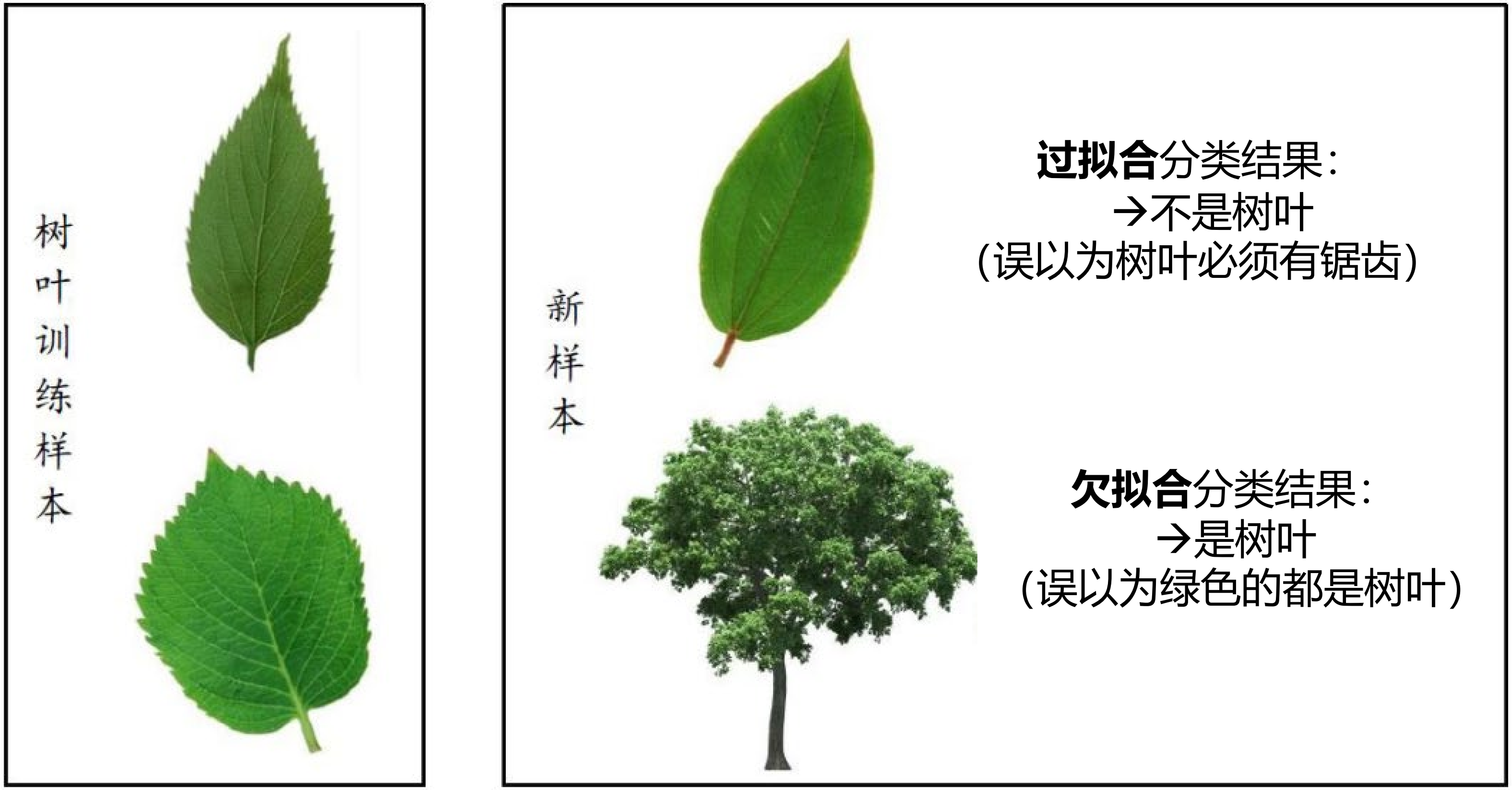
- Solutic

decisior

neural



Empirical error and Overfitting



过拟合分类结果：
→不是树叶
(误以为树叶必须有锯齿)

欠拟合分类结果：
→是树叶
(误以为绿色的都是树叶)

过拟合、欠拟合的直观类比

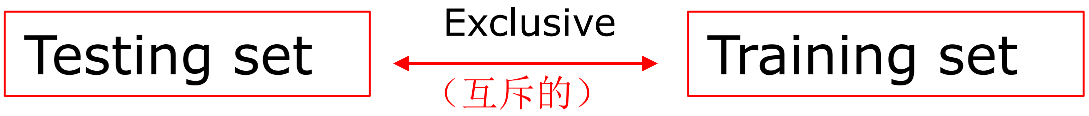
Outline

- Empirical Error and Overfitting
- Evaluation Methods
- Performance Measure
- Comparison test
- Bias and variance
- Extension

Evaluation Methods

In practice, we consider the generalization ability, time consuming, storage consuming, interpretability of a model and finally make a decision.

So we treat testing error as generalization error.



Testing set

Exclusive

（互斥的）

Training set

Evaluation Methods

We usually have a data set with m samples:

$$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$$

How we train and test on it?

Separate it into two parts: training set S , and testing set T .

Methods:

- Hold-out (留出法)
- Cross validation （交叉验证法）
- Bootstrapping （自助法）

Evaluation Methods

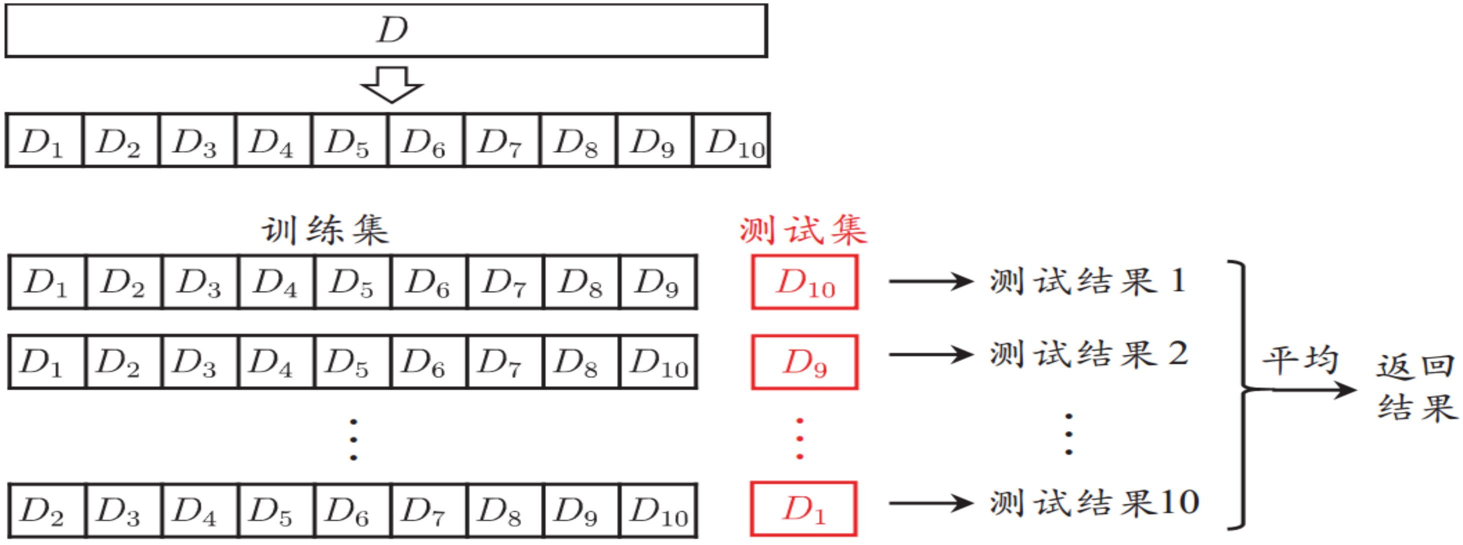
Hold-out (留出法)

- Divide the data set into two exclusive parts.
- Guarantee the consistency of data distribution.
- Many times, take the average.
- Training samples/Testing samples=2/1 ~4/1.

Evaluation Methods

Cross validation （交叉验证）

Divide the data set into **k** exclusive subsets.



D
D1 D2 D3 D4 D5 D6 D7 D8 D9 D10
训练集
测试集
D1 D2 D3 D4 D5 D6 D7 D8 D9
D10
测试结果 1
D1 D2 D3 D4 D5 D6 D7 D8 D10
D9
测试结果 2
D2 D3 D4 D5 D6 D7 D8 D9 D10
D1
测试结果10
平均
返回结果

10 折交叉验证示意图

Evaluation Methods

Specially, we have a data set with **m** samples, if k=m,

Leave-One-Out(留一法)：

- Not affected by partition methods.
- More accurate.
- A big data set leads to a large computing consuming.

Outline

- Empirical Error and Overfitting
- Evaluation Methods
- Performance Measure
- Comparison test
- Bias and variance
- Extension

Performance Measure

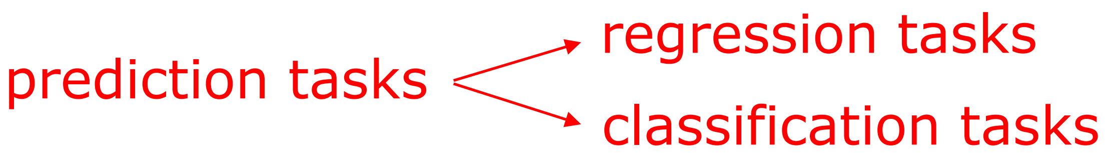
Performance Measure：

A Evaluation Criterion measuring the generalization ability of a model.

In prediction tasks, given a sample set D,

$$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$$

We compare the prediction results f(x) with the real labels.



prediction tasks

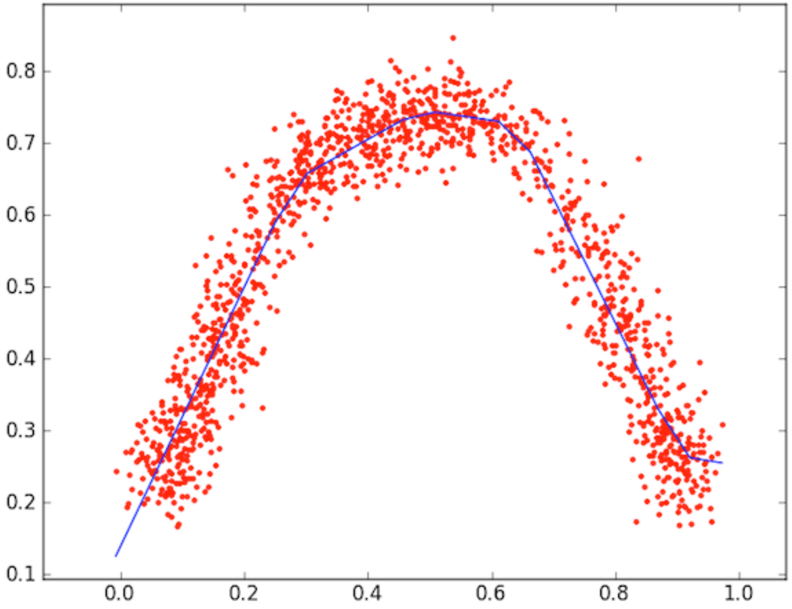
regression tasks

classification tasks

Performance Measure

In regression task, we usually take mean square error as a Performance Measure:

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\left(f\left(\boldsymbol{x}_i\right)-y_i\right)^2$$



Performance Measure

In classification tasks, Error Rate & Accuracy are two main measures.

- Error Rate: wrongly classified/all samples.
- Accuracy: correctly classified/all samples.

Error Rate

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

Accuracy

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

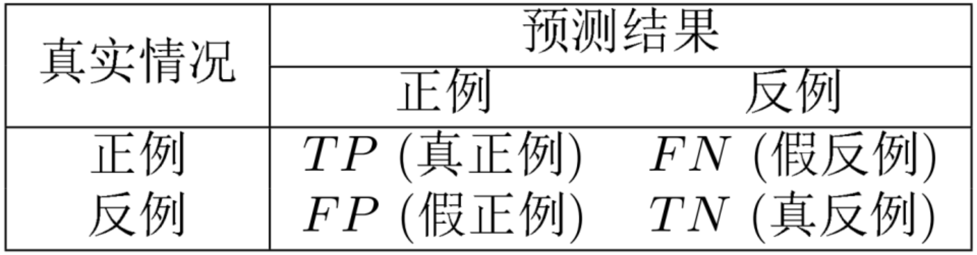
Performance Measure

In the field of information retrieval and web search, we often need precision and recall:

Precision(查准率): The positive predictive value.

Recall（查全率）: The true positive rate .

分类结果混淆矩阵

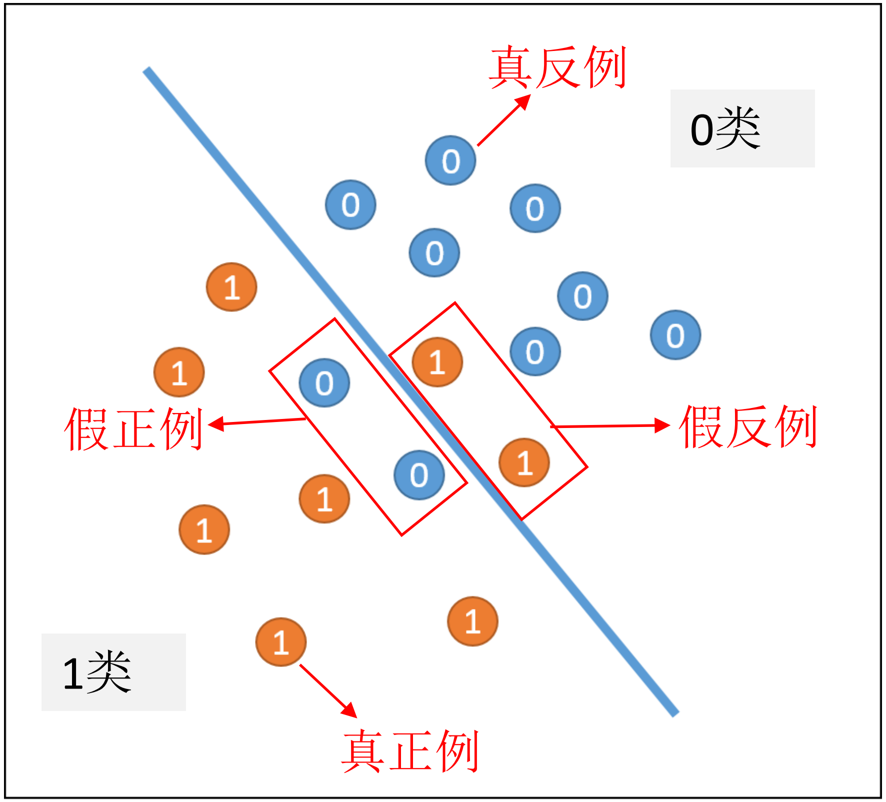


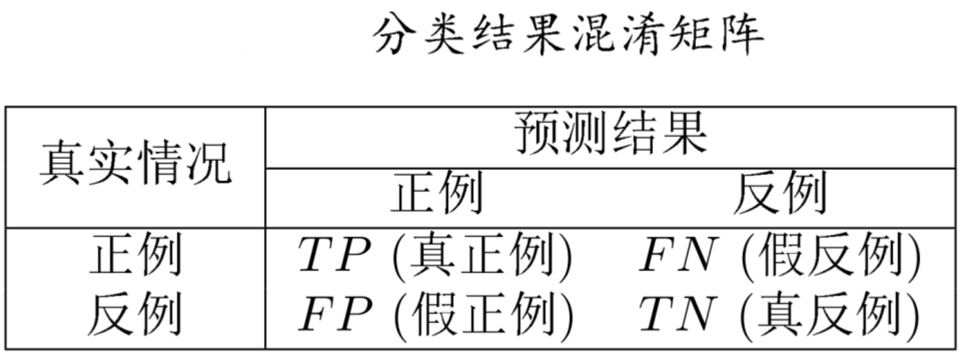
真实情况	预测结果	
	正例	反例
正例	TP (真正例)	FN (假反例)
反例	FP (假正例)	TN (真反例)

查准率 $$P = \frac{TP}{TP + FP}$$

查全率 $$R = \frac{TP}{TP + FN}$$

Performance Measure





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)

查准率 $$P = \frac{TP}{TP + FP}$$

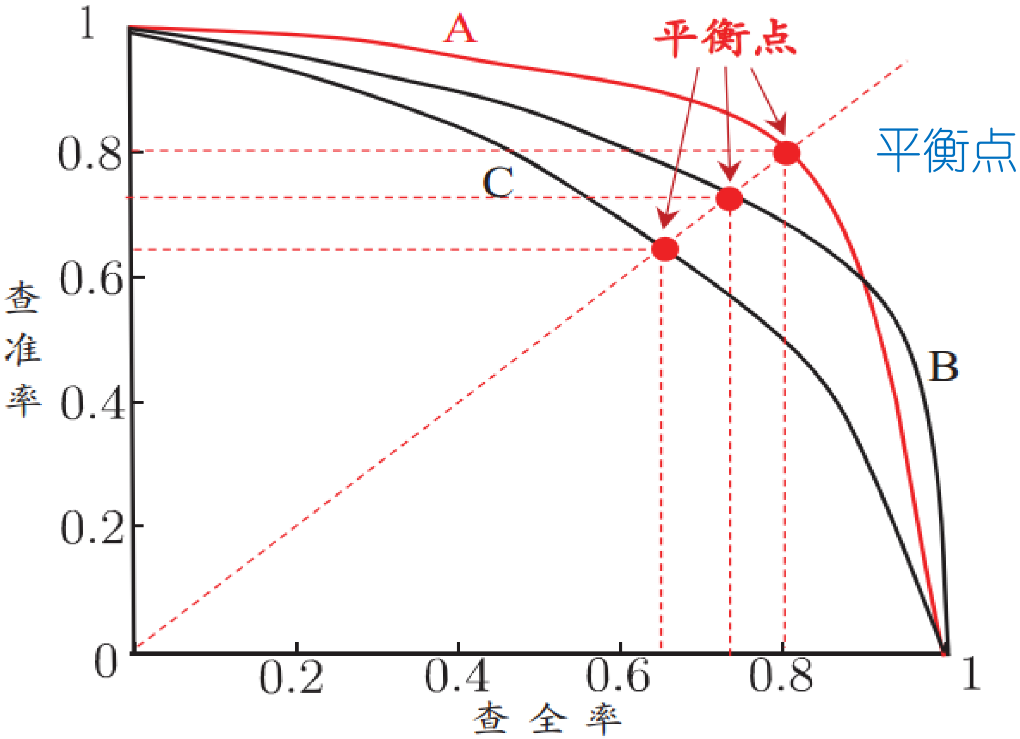
查全率 $$R = \frac{TP}{TP + FN}$$

1 实际是1

0 实际是0

Performance Measure

precision & Recall Curve



平衡点：P=R

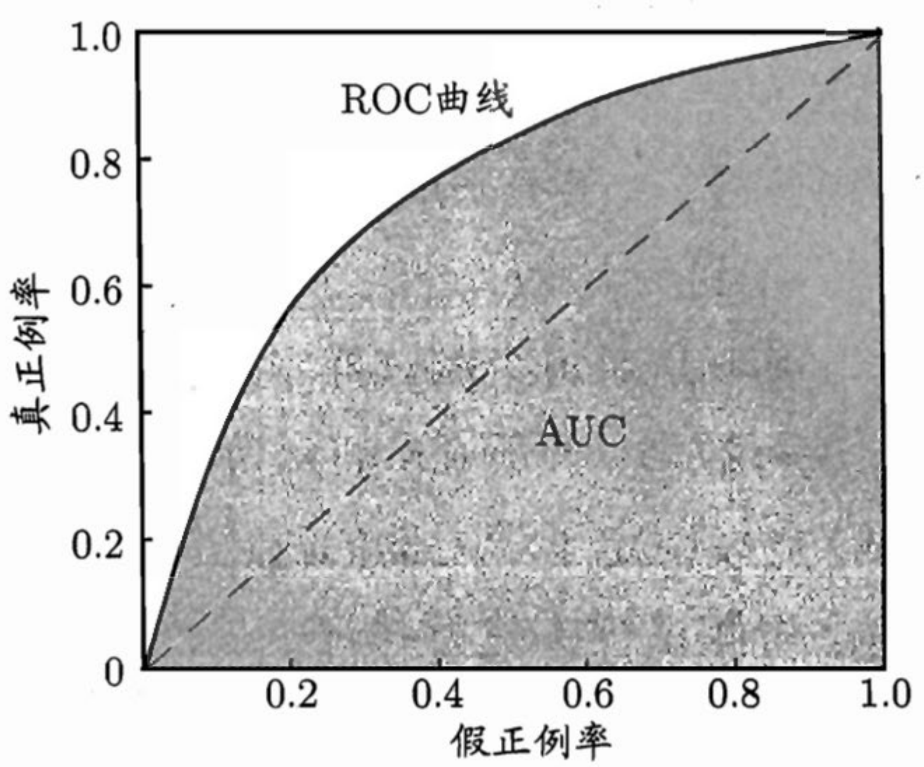
P-R曲线与平衡点示意图

Performance Measure

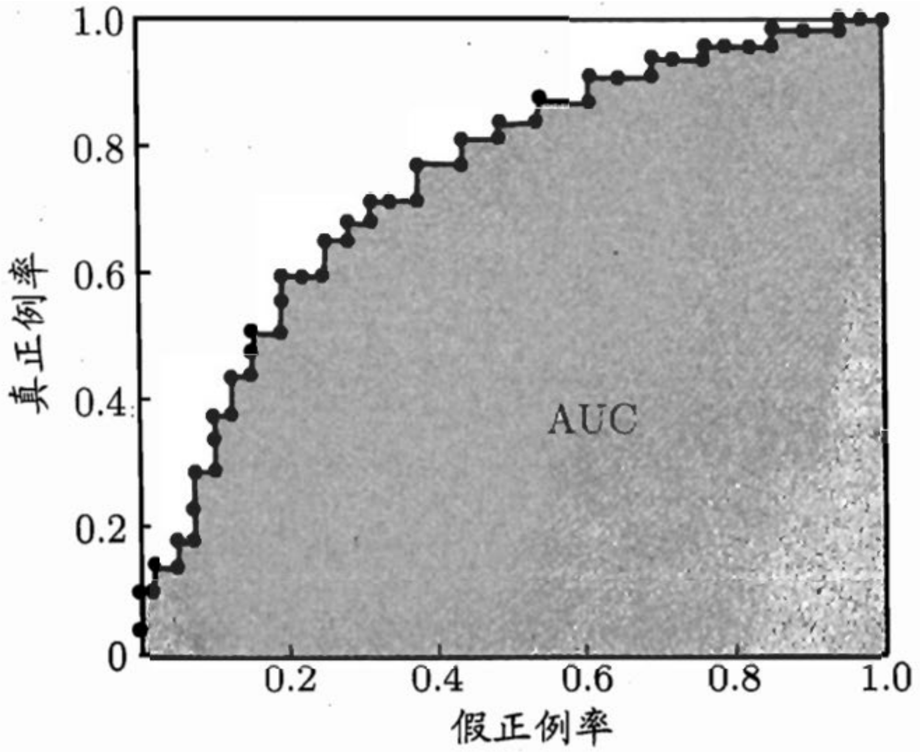
Receiver Operating Characteristic

“受试者工作特征曲线”

Area Under ROC Curve



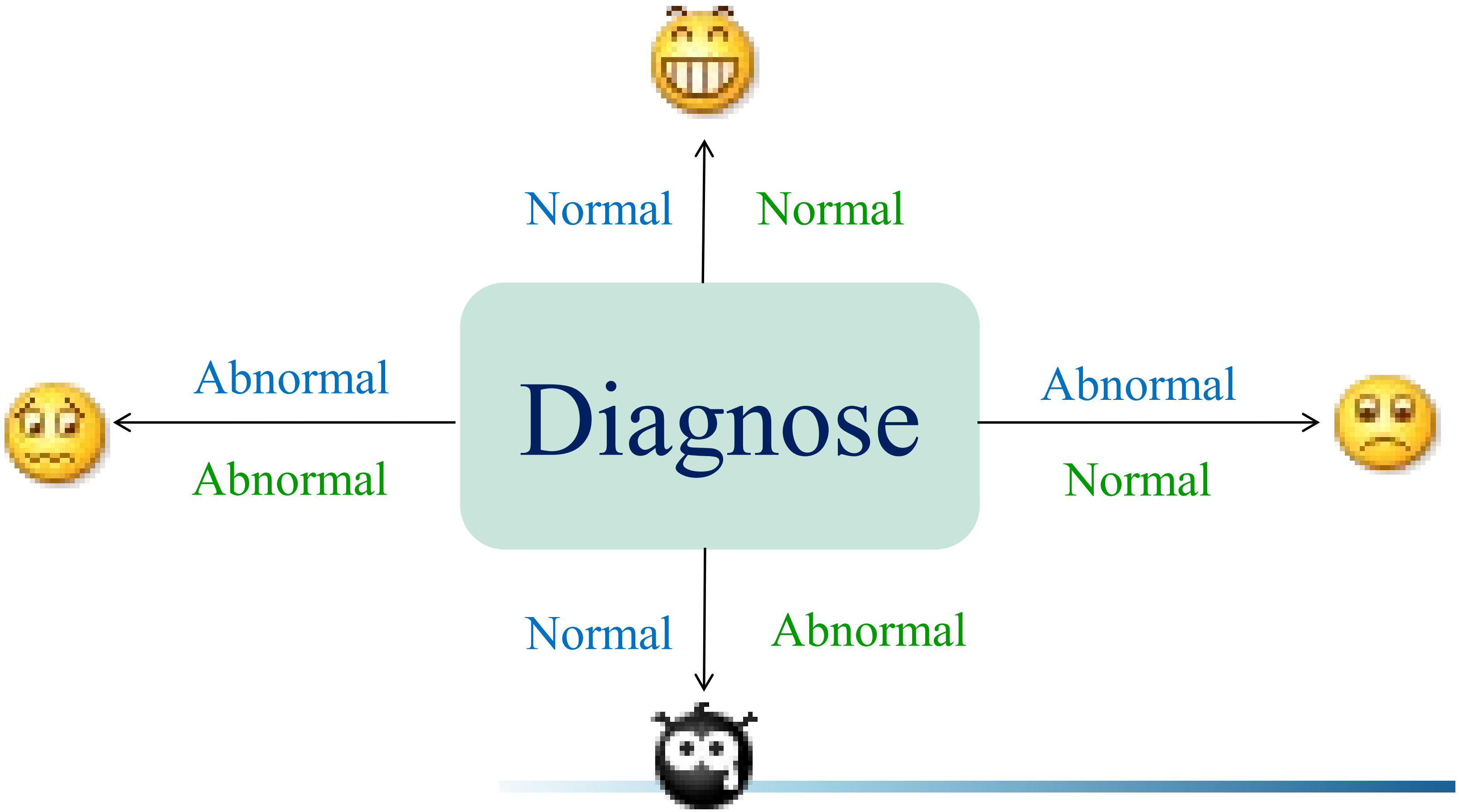
(a) ROC 曲线与 AUC



(b) 基于有限样例绘制的 ROC 曲线与 AUC

Cost-Sensitive error rate

In realistic task, different error cause different effect, we want to attach different cost on different errors.



Cost-Sensitive error rate

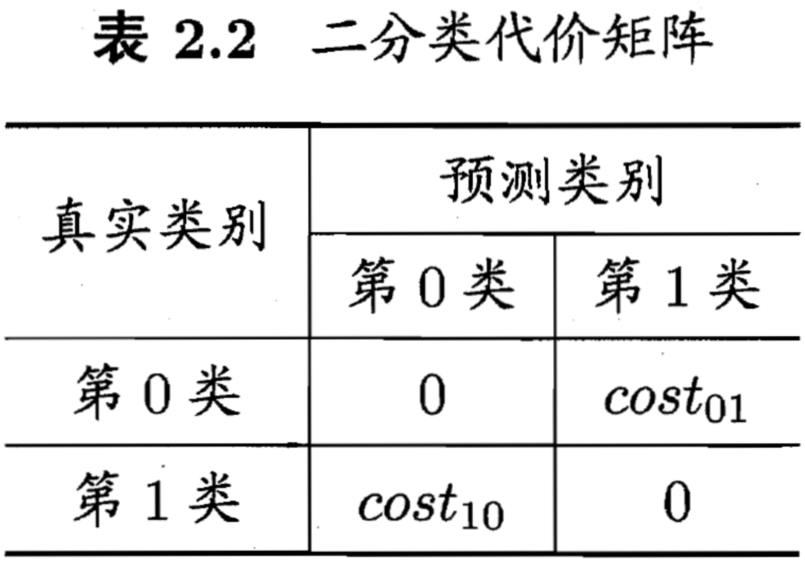


表 2.2　二分类代价矩阵

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

Cost-Sensitive error rate:

$$E(f;D;cost)=\frac{1}{m}\left(\sum_{\boldsymbol{x}_i\in D^+}\mathbb{I}\left(f\left(\boldsymbol{x}_i\right)\neq y_i\right)\times cost_{01}\right.$$

$$\left.+\sum_{\boldsymbol{x}_i\in D^-}\mathbb{I}\left(f\left(\boldsymbol{x}_i\right)\neq y_i\right)\times cost_{10}\right)$$

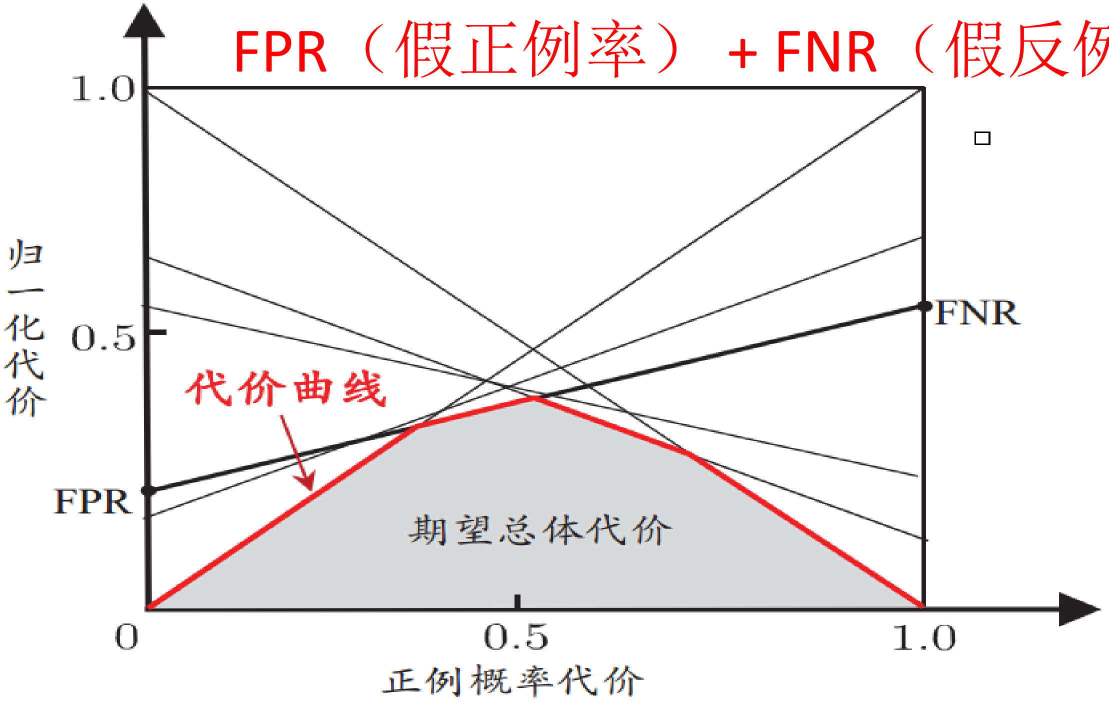
Total cost

Cost-Sensitive error rate

$$P(+)cost = \frac{p \times cost_{01}}{p \times cost_{01} + (1-p) \times cost_{10}} ,$$

$$cost_{norm} = \frac{\mathrm{FNR} \times p \times cost_{01} + \mathrm{FPR} \times (1-p) \times cost_{10}}{p \times cost_{01} + (1-p) \times cost_{10}} ,$$

FPR（假正例率） + FNR（假反例率） = 1



代价曲线与期望总体代价

Outline

- Empirical Error and Overfitting
- Evaluation Methods
- Performance Measure
- Comparison test
- Bias and variance
- Extension

Comparison test

- In regard to performance comparison：
  - Test performance ≠ Generalization performance.
  - Test performance change along with test set.
  - Many machine learning algorithms have a certain randomness.

Take a evaluation method, and compare. ✗

Comparison test

The purpose Comparison test:

If it is observed on test set that the learner A is better than learner B,

how can we guarantee that the generalization performance of A is better than that of B in the statistical sense.

Comparison test

Test Methods：

- Hypothesis test（假设检验）
- McNemar test.
- Friedman test.

binomial test
t-test.

Comparison test

Hypothesis testing is an essential procedure in statistics. A hypothesis test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data.

Hypothesis: some judgment/guess on the distribution of the error rate of a learner‵s generalization.

- We take the error rate (ϵ) as the performance measure
- testing error $\hat{\epsilon}$.
- generalization error ϵ.

Comparison test

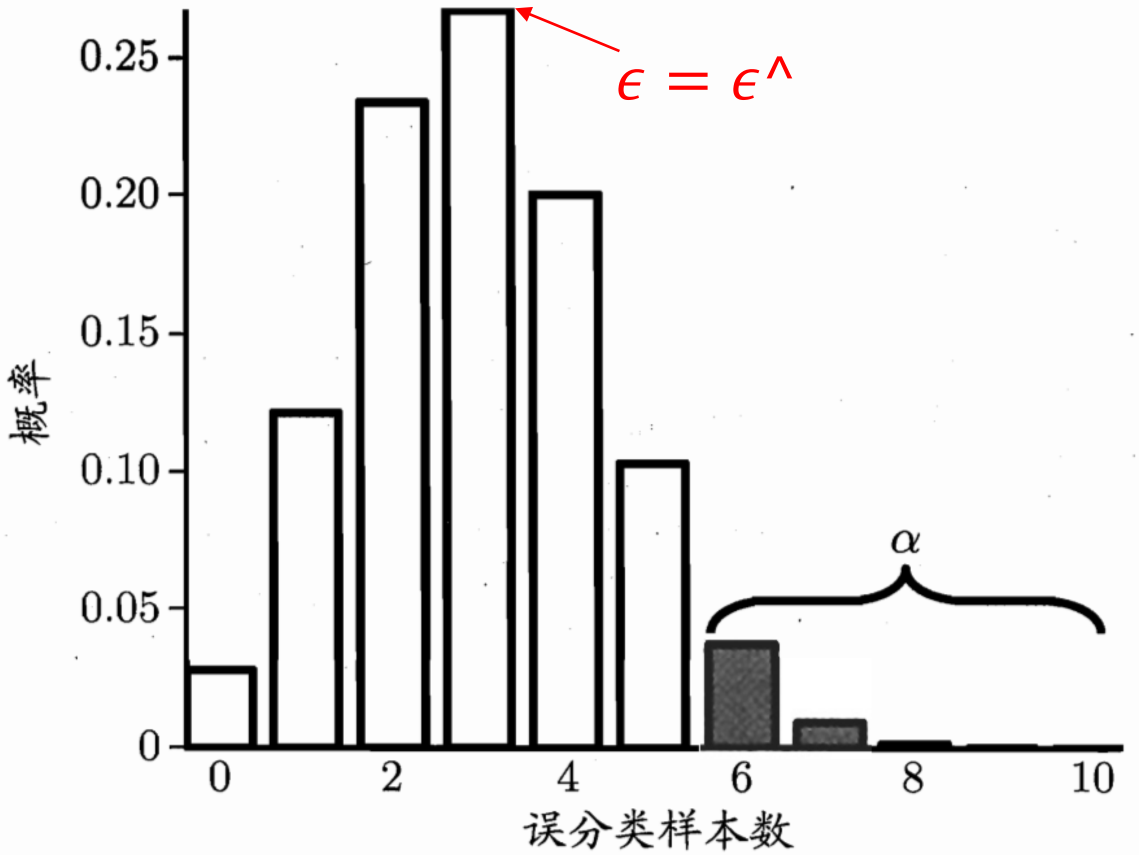
- 测试差 $\hat{\epsilon}$ 意味着m个测试样本中恰有$\hat{\epsilon}$ x m个被误分类。
- 假定一个学习器的泛化误差 为ϵ，则该学习器将m′个样本误分类且将剩下的样本被全部分类正确的概率：$\epsilon^{m'}(1\text{-}\epsilon)^{m-m'}$

在包含了m个样本的测试集上，泛化错误率为 ϵ 的学习器被测得测试错误率为ϵ^的概率：

$$P(\hat{\epsilon};\epsilon)=\binom{m}{\hat{\epsilon}\times m}\epsilon^{\hat{\epsilon}\times m}(1-\epsilon)^{m-\hat{\epsilon}\times m}\ .$$

二项分布

Comparison test



$\epsilon = \epsilon$^

图 2.6 二项分布示意图$(m = 10, \epsilon = 0.3)$

Comparison test（t-test)

- **“t检验”(t-test)**

We have got k error rates: $\hat{\epsilon}_1$, $\hat{\epsilon}_2$, … $\hat{\epsilon}_k$, , Then,

平均测试错误率：$\mu = \frac{1}{k}\sum_{i=1}^{k}\hat{\epsilon}_i$,

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

泛化误差

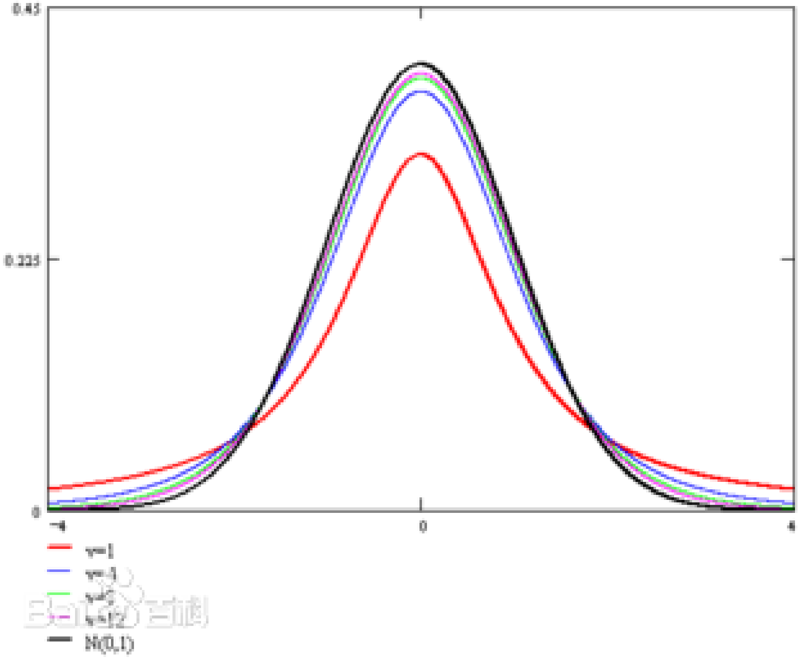
- $\tau_t = \frac{\sqrt{k}(\mu-\epsilon_0)}{\sigma}$ 服从自由度为 k-1 的 t 分布。

Comparison test（t-test）

t-分布：在概率论和统计学中，学生t-分布（t-distribution），可简称为t分布，用于根据小样本来估计呈正态分布且方差未知的总体的均值。如果总体方差已知（例如在样本数量足够多时），则应该用正态分布来估计总体均值。

自由度df愈 当自由度df=∞时，

t分布曲线为标



Comparison test（t-test)

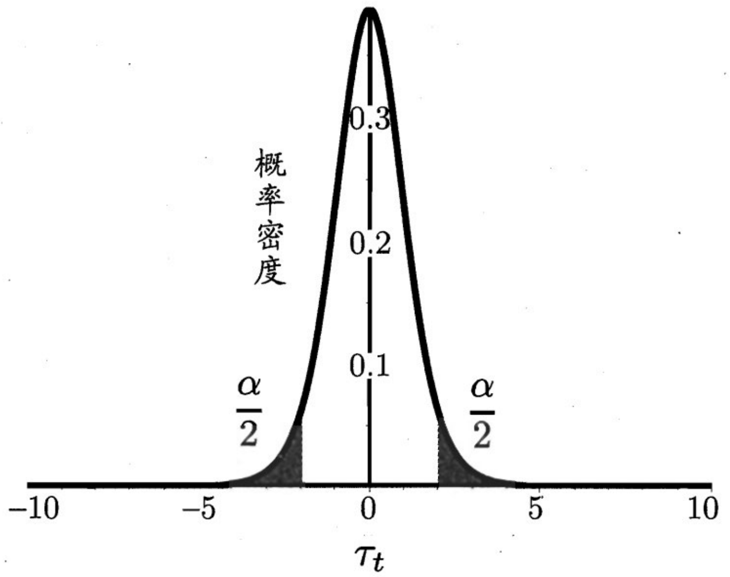


图 2.7 t 分布示意图$(k = 10)$

对假设"$\mu = \epsilon_0$"和显著度α，若μ与ϵ_0之差在 $[-t_{\frac{\alpha}{2}}, t_{\frac{\alpha}{2}}]$内，则不能拒绝假设"$\mu = \epsilon_0$"，即可认为泛化错误率为$\epsilon_0$，置信度为1-$\alpha$；否则可拒绝该假设。

Comparison test（t-test example)

t 检验--例子：

T 检验是针对分布期望 μ 的检验。假设一组服从正态分布的测试误差：

0.10 0.12 0.14 0.11 0.13 0.12

正态分布的期望 μ 等于0.11。如果要判断这种说法的正确与否，便需要使用 T 检验了。T 检验的主要步骤如下:

步骤1： 建立零假设H_0和备选假设H_1。

$$H_0: \mu = 0.11 \qquad H_1: \mu \neq 0.11$$

并限定显著性水平。这里我们限定显著性水平为α=0.05

Comparison test（t-test example)

步骤2：我们选择 T 统计量。计算公式如下：

$$T = \frac{x-0.11}{s/\sqrt{n}}=0.28867513459$$

其中s是样本的标准差。

步骤3：查 T 检验临界值表。因为样本中拥有6份数据，因此我们采用n=5（自由度为5）所对应的行；显著性水平为α=0.05，因此我们采用双侧检验p=0.05所对应的列。

查表所得值为2.571。|t|=0.28867513459<2.571，故我们接受零假设$\mathrm{H_0}$，认为 μ =0.11成立。

Comparison test（t-test)

交叉验证t检验：

（1）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：

$$\in_1^A, \in_2^A, \dots, \in_k^A \text{ 和 } \in_1^B, \in_2^B, \dots, \in_k^B$$

（2）对每组结果求差值：

$$\nabla_{\mathrm{i}} = \in_{\mathrm{i}}^A - \in_i^B$$

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，因此可以根据差值$\nabla_1, \nabla_2, \dots, \nabla_k$来对学习器A,B性能相同”这个原假设做t检验

Comparison test（t-test)

（3）假设检验：先计算出差值的均值 μ 与方差σ^2,在显著度α下，若变量

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

小于自由度为k-1的临界值，则原假设不能被拒绝，认为两个学习器的性能没有显著差别；反之则认为两个学习器的性能有显著差别，并且选择平均错误率较小的那个学习器。

Outline

- Empirical Error and Overfitting
- Evaluation Methods
- Performance Measure
- Comparison test
- Bias and Variance
- Extension

Bias and Variance

Bias:

error from erroneous assumptions in the learning algorithm.

Variance:

error from sensitivity to small fluctuations in the training set.

bias and variance decomposition: Help to explain generalization performance.

Bias and Variance

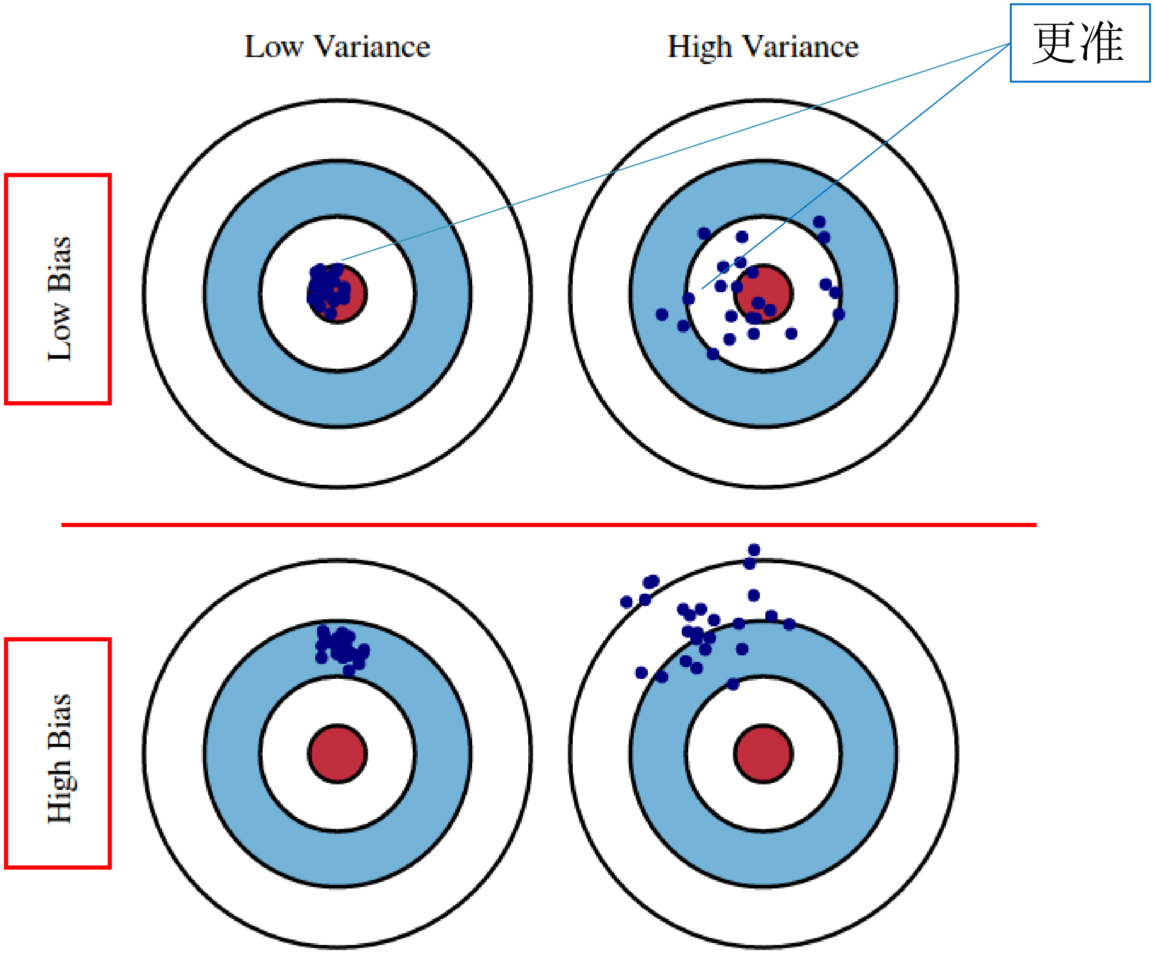
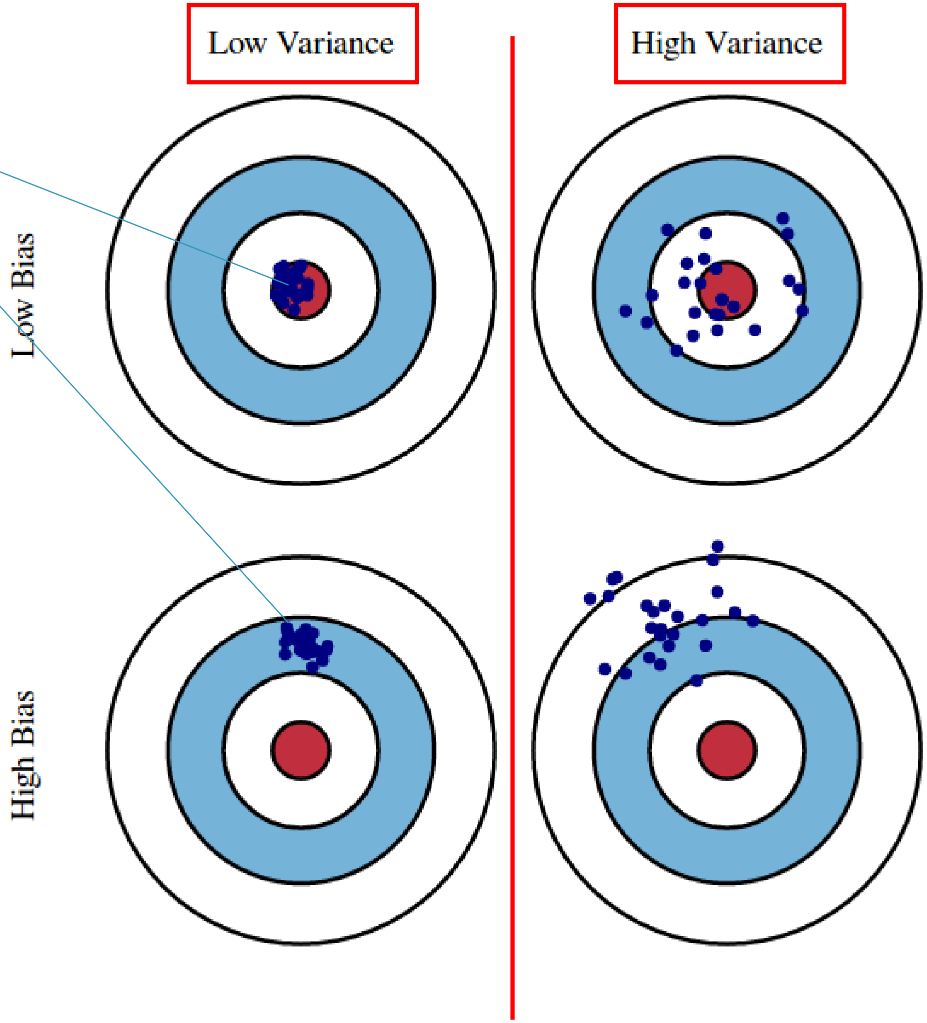


Fig. 1 Graphical illustration of bias and variance.

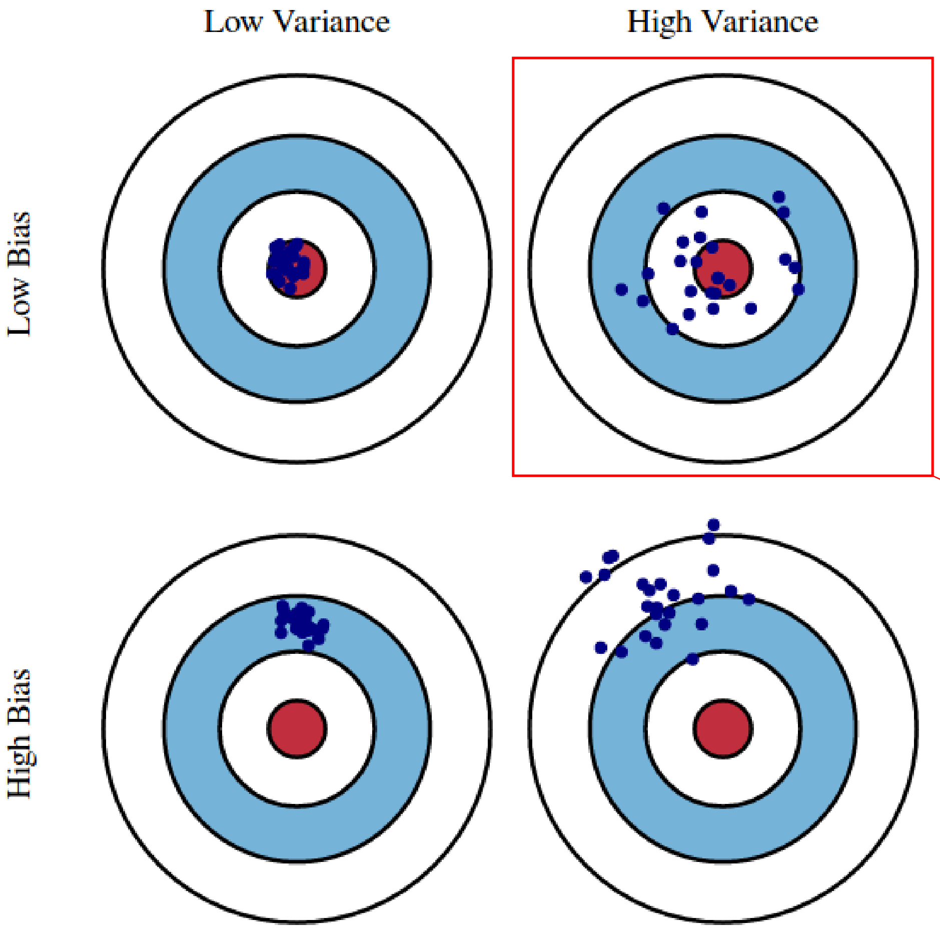
Bias and Variance



更稳

Fig. 1 Graphical illustration of bias and variance.

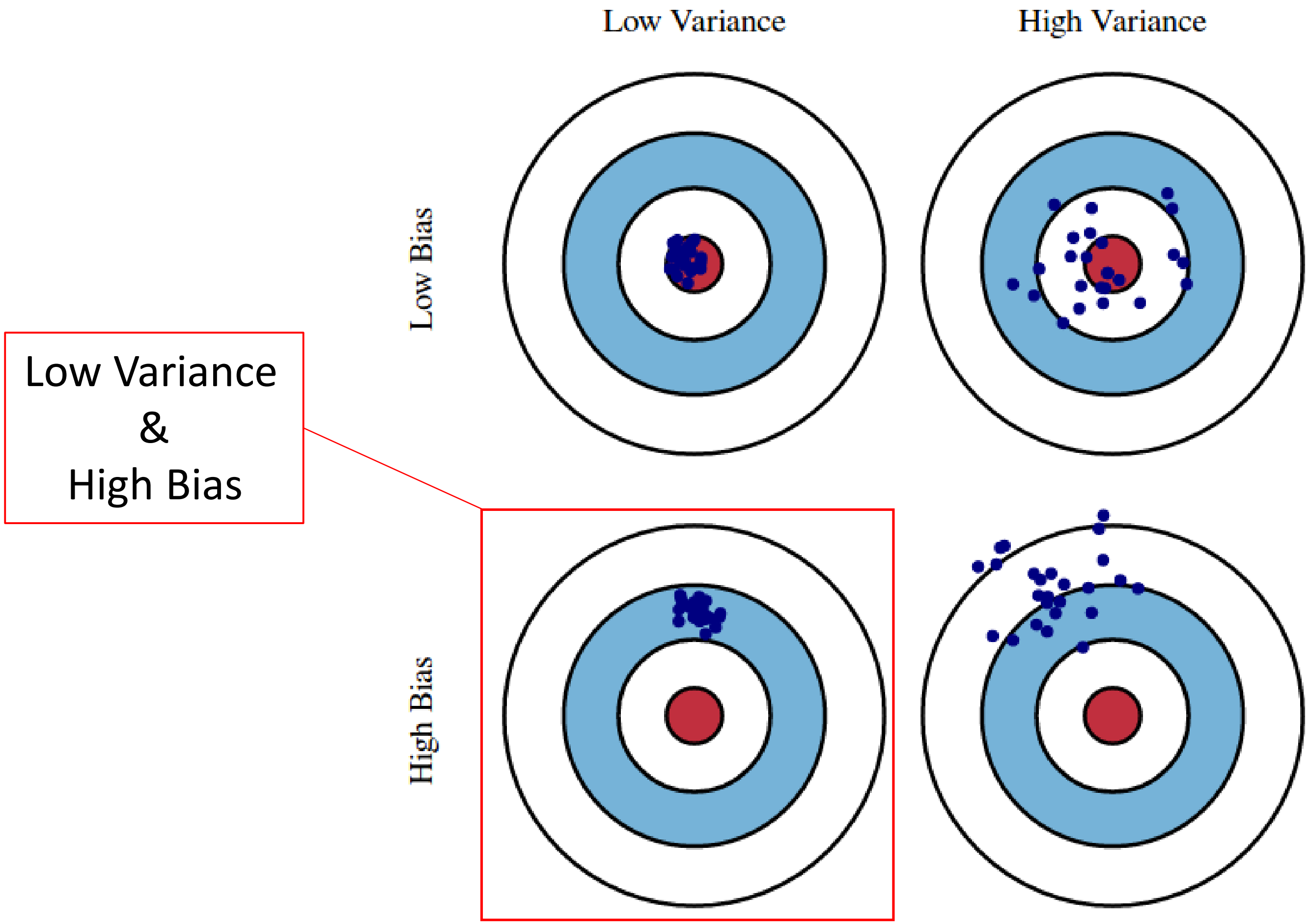
Bias and Variance



Low Bias
&
High Variance

Fig. 1 Graphical illustration of bias and variance.

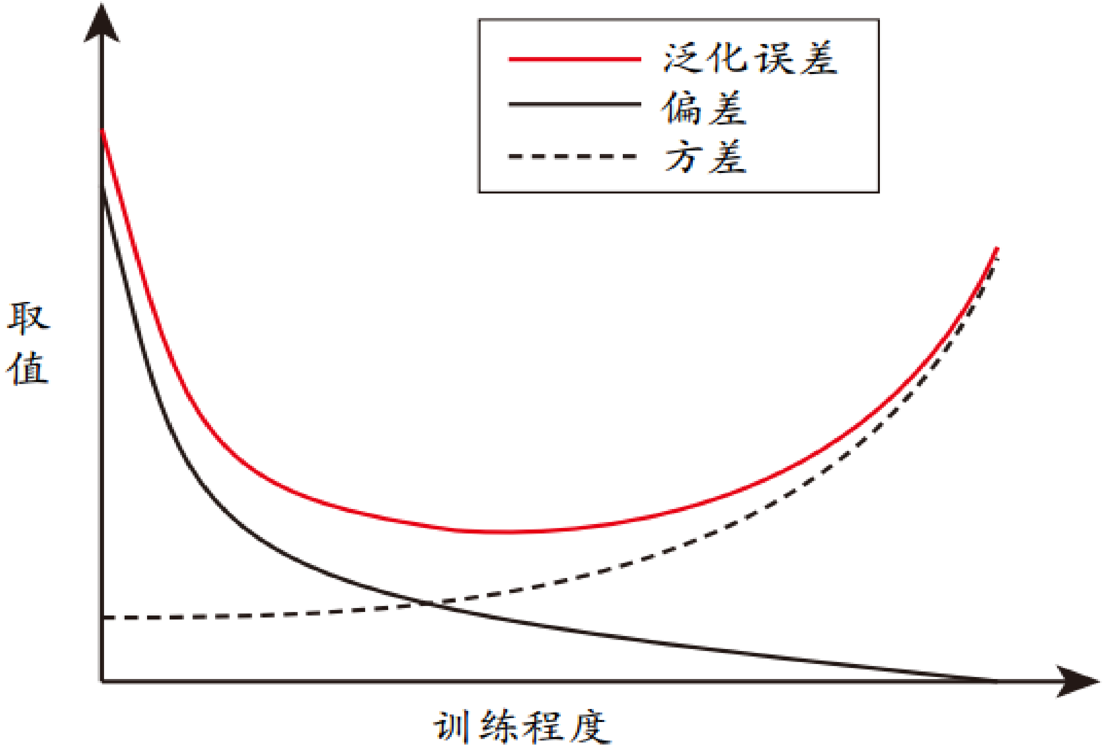
Bias and Variance



Low Variance
&
High Bias

Fig. 1 Graphical illustration of bias and variance.

Bias and Variance



泛化误差与偏差、方差的关系示意图

THANKS

Extension

(仅做扩展了解)

How to evaluate the quality of a machine learning model?

We want to make “good” predictions!

Fitting a model to our training data is one thing, but how do we know that it generalizes well to unseen data?

How do we know that it doesn’t simply memorize the data we fed it and fails to make good predictions on future samples that it hasn’t seen before?

And how do we select a good model in the first place? Maybe a different learning algorithm could be better-suited for the problem at hand?

Assumptions and Terminology

Model evaluation is certainly a complex topic. To make sure that we don't diverge too much from the core message, **let us make certain assumptions and go over some of the technical terms.**

1. i.i.d.
2. Supervised learning and classification.
3. 0-1 loss and prediction accuracy.
4. Bias.
5. Variance.
6. Target function.
7. Learning algorithm.

Assumptions and Terminology

1. i.i.d

We assume that our samples are i.i.d (independent and identically distributed), which means that all samples have been drawn from the same probability distribution and are statistically independent from each other. A scenario where samples are not independent would be working with temporal data or time-series data.

Assumptions and Terminology

2. Supervised learning and classification

We focus on supervised learning, a subcategory of machine learning where our target values are known in our available dataset. Although many concepts also apply to regression analysis, we will focus on classification, the assignment of categorical target labels to the samples.

Assumptions and Terminology

3. 0-1 loss and prediction accuracy

the prediction accuracy is defined as the number of all correct predictions divided by the number of samples. Or in more formal terms, we define the prediction accuracy ACC as

ACC = 1 – ERR,

where the prediction error ERR is computed as the expected value of the 0-1 loss over n samples in a dataset S:

$$L(\hat{y}_i, y_i) := \begin{cases} 0 & \text{if } \hat{y}_i = y_i \\ 1 & \text{if } \hat{y}_i \neq y_i, \end{cases}$$

where y_i is the i-th true class label and $\hat{y}_i$ the i-th predicted class label, respectively.

Assumptions and Terminology

4-5. Bias & Variance

we compute the prediction bias as the difference between the expected prediction accuracy of our model and the true prediction accuracy.

The variance is a measure of the variability of our model's predictions if we repeat the learning process multiple times with small fluctuations in the training set. The more sensitive the model-building process is towards these fluctuations, the higher the variance.

Assumptions and Terminology

6. Target function

In predictive modeling, we are typically interested in modeling a particular process; we want to learn or approximate a specific, unknown function. The target function ***f(x) = y*** is the true function ***f(·)*** that we want to model.

Assumptions and Terminology

7. Learning algorithm

our goal is to find or approximate the target function, and the learning algorithm is a set of instructions that tries to model the target function using our training dataset. A learning algorithm comes with a hypothesis space, the set of possible hypotheses it explores to model the unknown target function by formulating the final hypothesis.

P & NP

P & NP

- P(Polynomial)多项式时间

The general class of questions for which some algorithm can provide an answer in polynomial time is called "class P" or just "P".

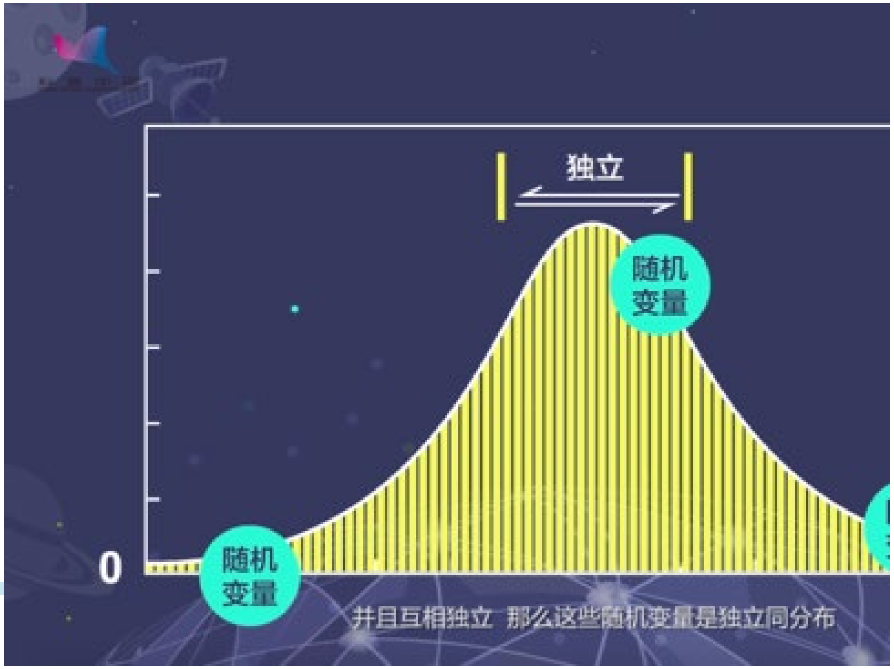
- NP(nondeterministic polynomial)非确定性多项式时间

$$P \neq NP$$

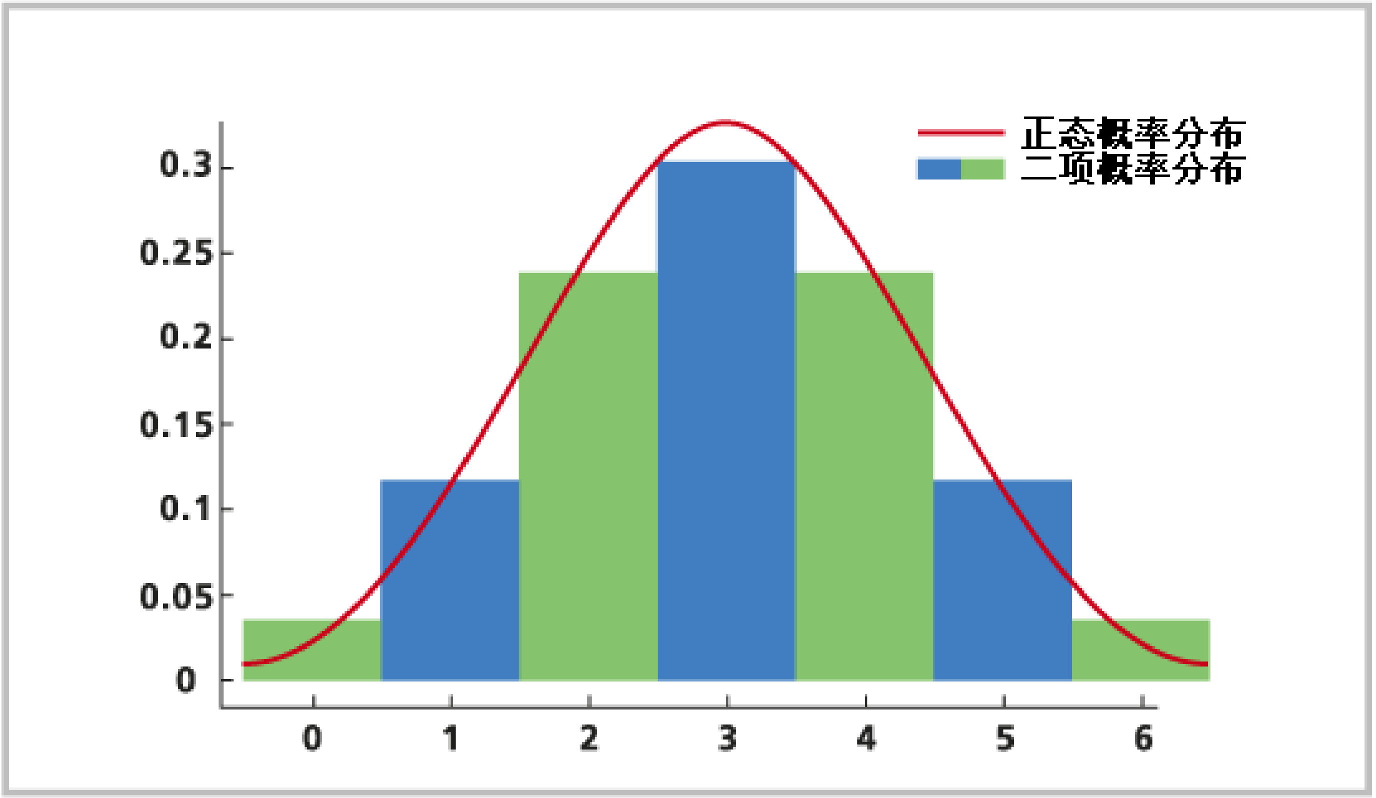
Independent and Identically Distributed (IID)

- In probability theory and statistics, a sequence or other collection of random variables is IID if :

Each random variable has the same probability distribution as the others and all are mutually independent.



Comparison test



二项分布