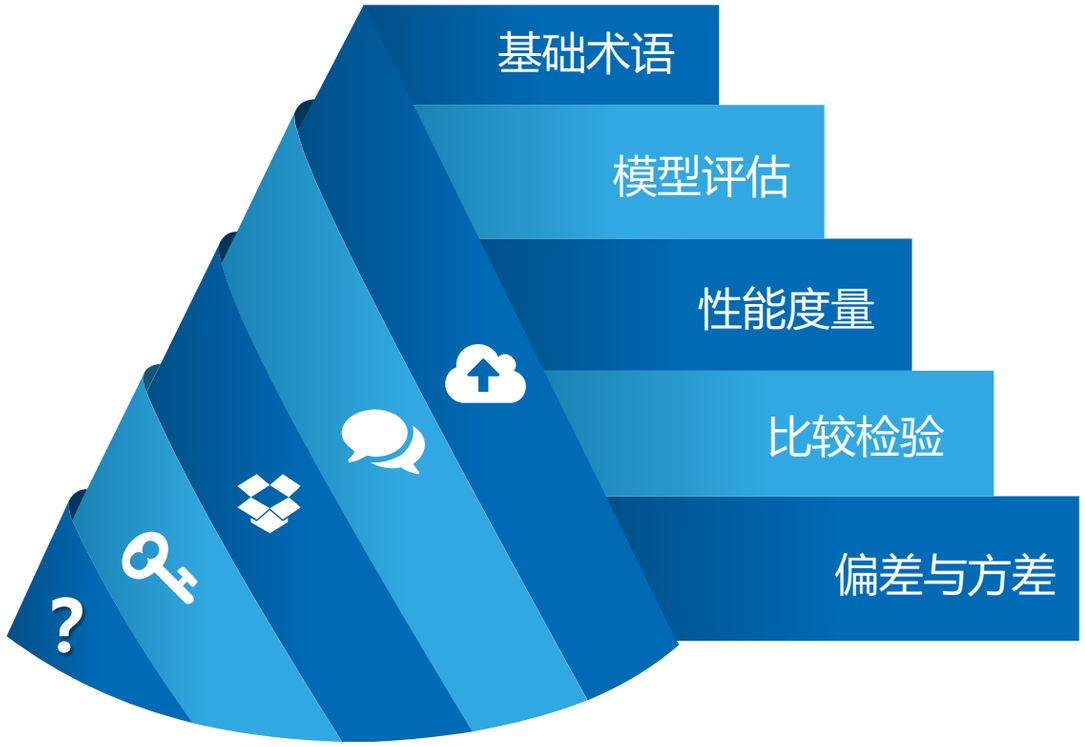
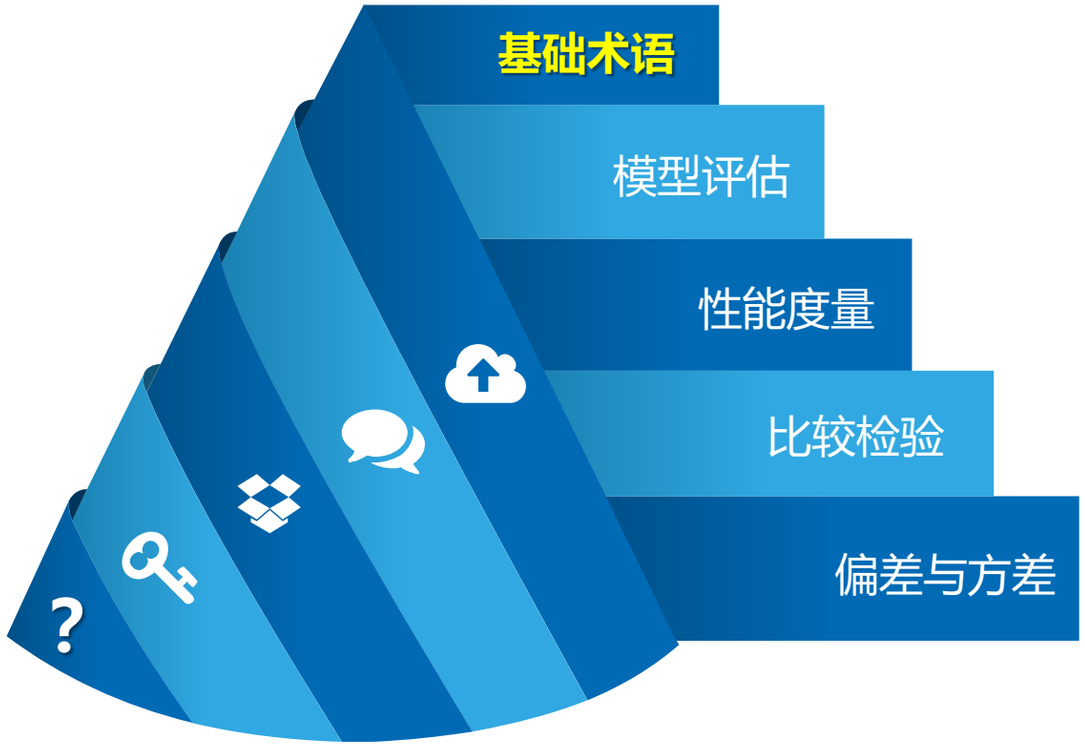
第二讲　基础概念

周文晖

杭州电子科技大学



- **基础术语与基本概念**
  数据，学习方法，泛化能力，…
- **模型评估方法**
  留出法，交叉验证法，…
- **模型性能度量**
  错误率，精度，P-R曲线，…
- **比较检验**
  二项检验，t检验，交叉验证，…
- **偏差与方差**
  偏差，方差，…



- **基础术语与基本概念**
  数据，学习方法，泛化能力，…
- 模型评估方法
  留出法，交叉验证法，…
- 模型性能度量
  错误率，精度，P-R曲线，…
- 比较检验
  二项检验，t检验，交叉验证，…
- 偏差与方差
  偏差，方差，…

基础术语与基本概念

模式(Pattern)：为执行和完成识别任务，对分类识别对象进行科学的抽象，建立其数学模型，用以描述和代替识别对象，这种对象特性的描述就是模式。如规律、模板、特征组合等。

模式的表现形式：特征矢量、符号串、图、关系式。

模式类 ：具有某些共同特性的类别或类的总称，通常采用特定的抽象符号来表示。

模式表示具体对象的抽象特性，模式类则是对这一类事物的概念性描述。

样本：个别具体的模式称为样本。

样本是具体对象的个体，而模式是对同一类对象的概念性概括。

模式识别：研究对象的特征或者属性，运用一定的分析算法认定其类别，且分类识别结果尽可能地符合真实。

基础术语与基本概念

样本（sample）： 所研究对象的一个个体，是一类事物的一个具体体现或实例。

样本集（sample set）： 若干样本的集合。

类或类别（class）： 在所有样本上定义的一个子集，属于同一类的样本具有相同模式（属性或特征）。

特征（features）： 指用于表征样本特点或性状的观测和量化集合，也被称为属性（attribute）。若存在多维特征，则为特征向量（feature vector）。样本的特征构成了样本特征空间。模式识别则是在样本特征空间中完成模式识别（决策）。

已知样本（known samples）： 指事先知道类别标号的样本。

未知样本（unknown samples）： 指类别标号未知但特征已知的样本。

基础术语与基本概念

模式识别研究内容

- 数据预处理
  - 视频、图像、信号处理
- 模式分割
  - 模式/背景分离、模式-模式分离
- 运动分析
  - 目标跟踪、运动模式描述
- 模式描述与分类
  - 特征提取/选择、模式分类、聚类、机器学习
- 模式识别应用研究
  - 针对具体应用的方法与系统

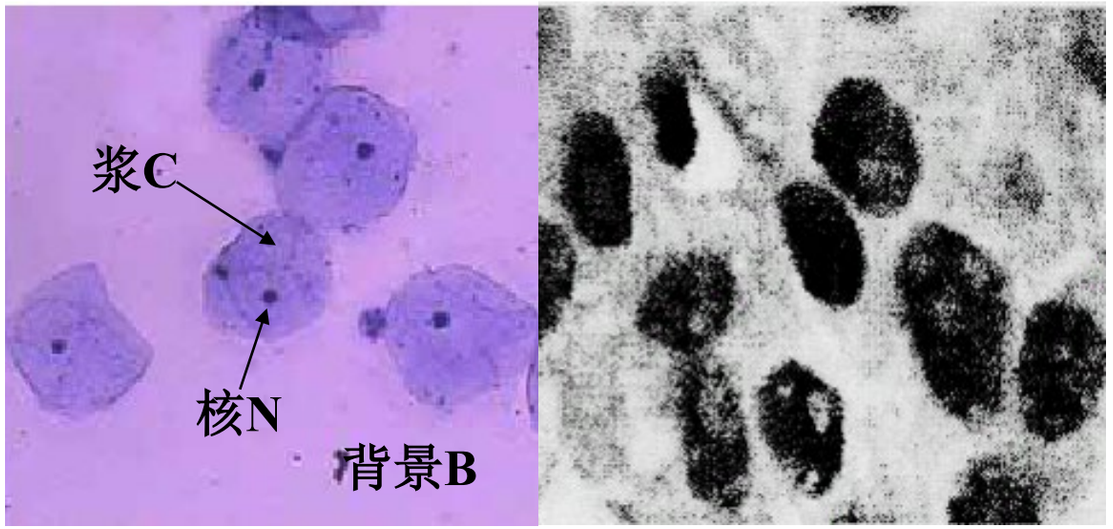
基础术语与基本概念

模式识别系统简例：建立感性认识

以癌细胞识别为例，了解模式识别的全过程。

第1步：信息输入与数据获取

将显微细胞图像转换成数字化细胞图像，是计算机分析的原始数据基础。灰度数字图像的像素值反映光密度的大小。



经过染色处理过的彩色图象　　灰度图象

数字化显微细胞图像

基础术语与基本概念

模式识别系统简例：建立感性认识

以癌细胞识别为例，了解模式识别的全过程。

第2步：数字化细胞图像的预处理与区域划分

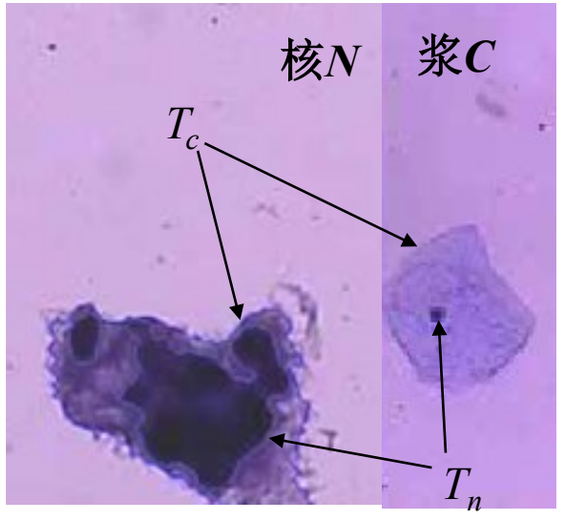
预处理目的：

（1）去除在数据获取时引入的噪声与干扰。

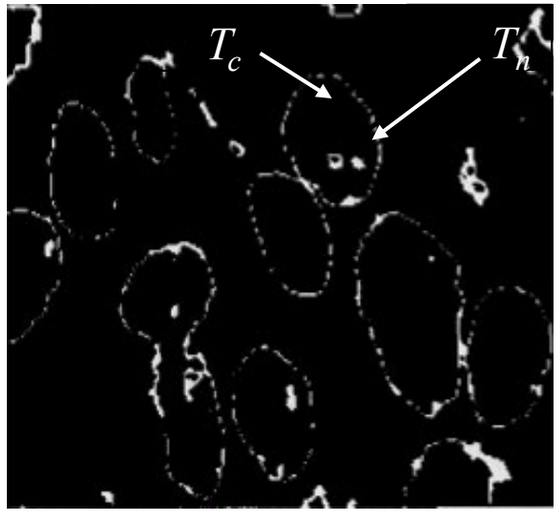
（2）增强主要的待识别细胞图像。

例：平滑、图像增强等数字图像处理技术。

区域划分的目的：找出边界，划分出三个区域，为特征抽取做准备。



疑似肿瘤细胞



检测的边缘

设灰度阈值为T_c和T_n，图像中某像素的灰度值为T_i，则：

$T_i \geq T_n$的点属于胞核区；

$T_i < T_c$的点属于背景区；

$T_c \leq T_i < T_n$的点属于胞浆区；

基础术语与基本概念

模式识别系统简例：建立感性认识

以癌细胞识别为例，了解模式识别的全过程。

第3步：细胞特征的抽取、选择和提取

目的：为了建立各种特征的数学模型，以用于分类：

① 抽取特征 ：原始采集数据量大。是特征选择和提取的依据。

例：一个细胞抽取33个特征 ，建立一个33维的空间 $\boldsymbol{X}$，

每个细胞可通过一个33维向量表示，记为： $\boldsymbol{X}=[x_1, x_2, \cdots, x_{33}]^{\mathrm{T}}$

即把一个 “细胞”变成了一个数学模型“33维随机向量”， 也即33维空间中的一点。

② 特征选择：在原始特征基础上选择一些主要特征作为判别用的特征。

基础术语与基本概念

模式识别系统简例：建立感性认识

以癌细胞识别为例，了解模式识别的全过程。

第3步：细胞特征的抽取、选择和提取

③ 特征提取：采用某种**变换**技术，提取综合特征用于分类，亦称特征维数压缩。

例：有五个特征 x_1, x_2, x_3, x_4, x_5，以及变换 $f(\cdot)$、$g(\cdot)$，则可有：

$$y_1 = f(x_1, x_2, x_3, x_4, x_5) \qquad y_2 = g(x_1, x_2, x_3, x_4, x_5)$$

结果：X 空间中的向量 $\boldsymbol{X} = [x_1, x_2, x_3, x_4, x_5]^{\mathrm{T}}$

变成 Y 空间的向量 $\boldsymbol{Y} = [y_1, y_2]^{\mathrm{T}}$

即：特征向量由5维降为2维。

基础术语与基本概念

模式识别系统简例：建立感性认识

以癌细胞识别为例，了解模式识别的全过程。

第3步：判别分类

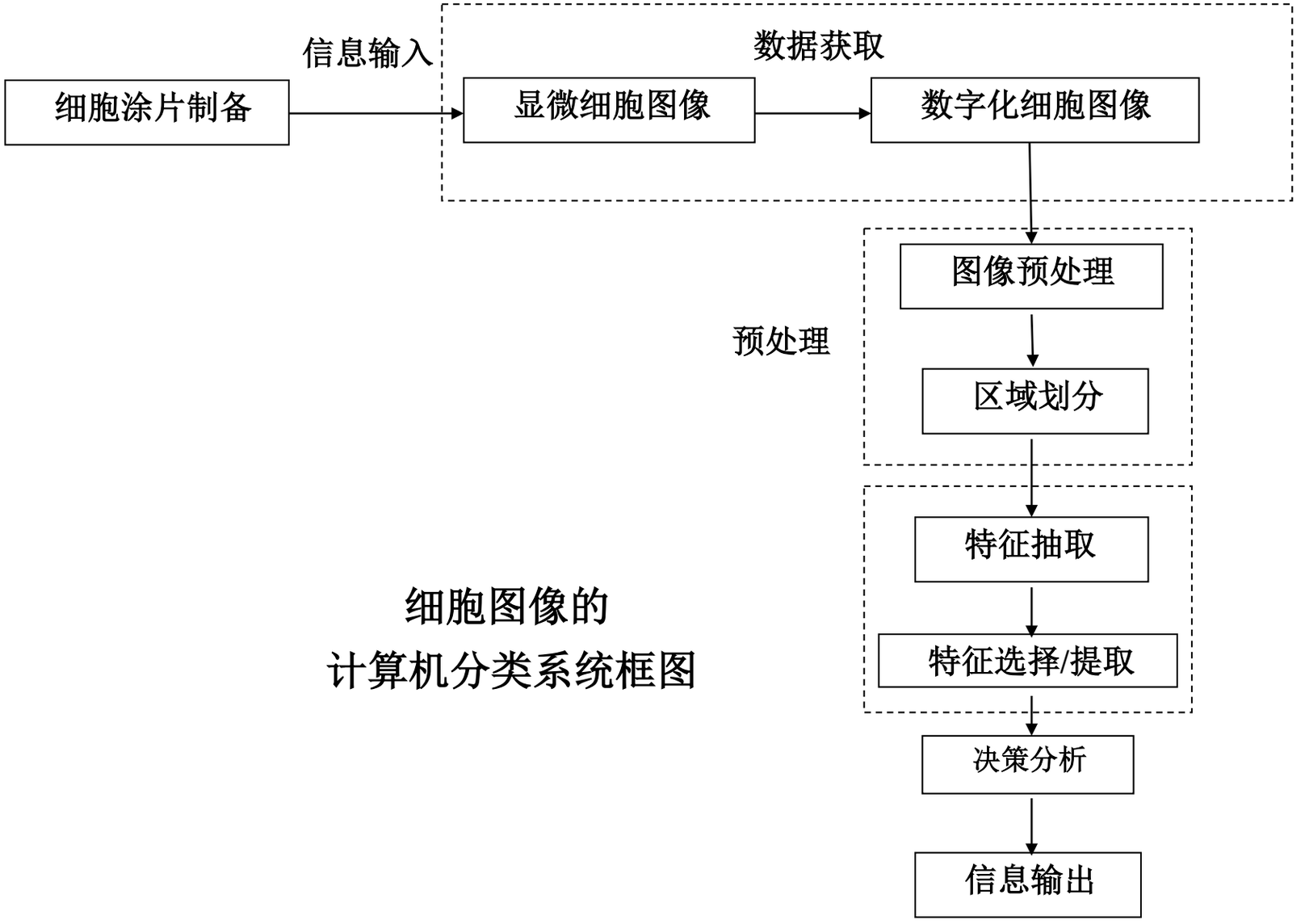
（1）气管细胞97个，识别错误率为7.2% 。

（2）肺细胞166个，识别错误率为18% 。

判别的好坏通过错误率给出，不同错误的代价和风险不同。

基础术语与基本概念

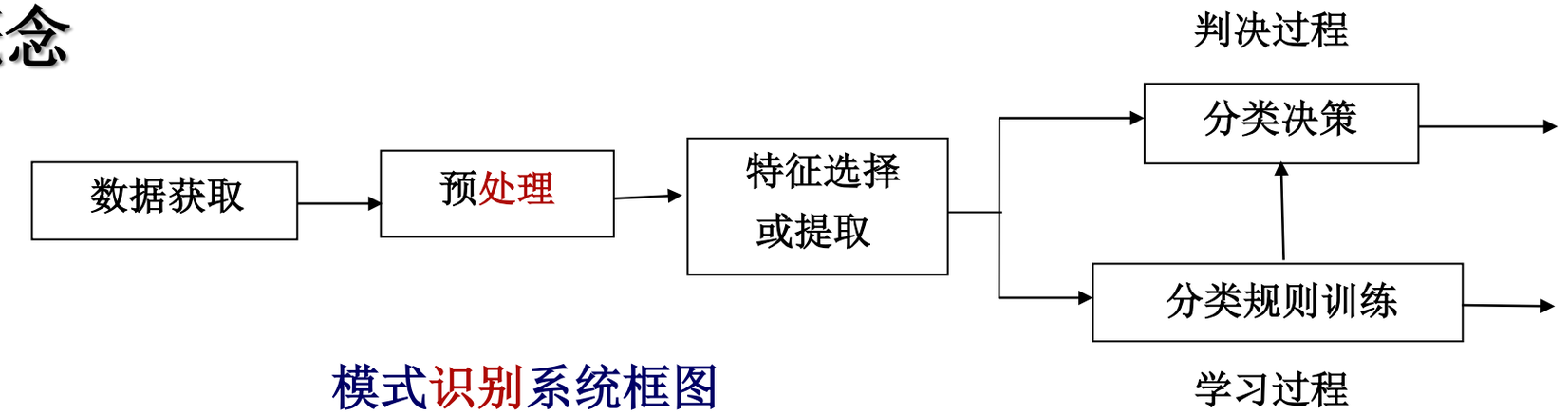
模式识别一般步骤



细胞图像的
计算机分类系统框图

基础术语与基本概念

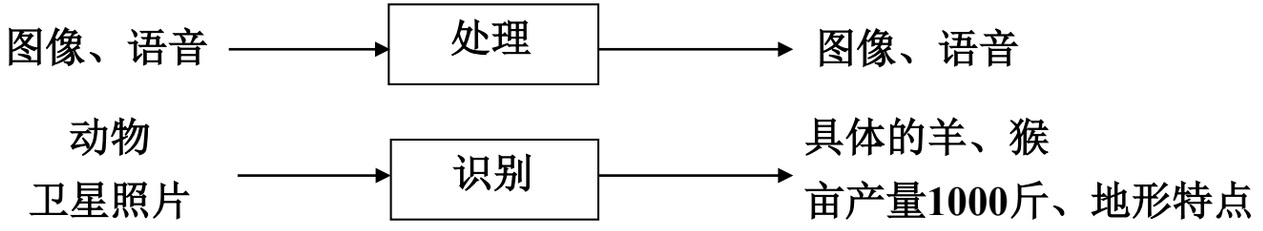
模式识别系统组成



模式识别系统框图

模式识别的三个核心问题：特征提取与选择、分类规则学习、分类决策

注意： “处理”与“识别”两个概念的区别



处理：输入与输出是同样的对象，性质不变。

识别：输入的是事物，输出的是对它的分类、理解和描述。

基础术语与基本概念

数据获取与特征提取

特征 ↑　　标记 ↑

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
1	青绿	蜷缩	沉闷	?

训练集 ←

测试集 ←

基础术语与基本概念

数据获取问题

以生成具有随机性的样本数据为目标，本质上是随机数生成问题。

现实世界中的某些事物，如图像，声音，服从某种概率分布。算法需要生成“像”这些物体的样本；

以图像为例，将所有像素拼接起来形成向量 $\mathbf{x}$，服从某种概率分布 $p(\mathbf{x})$，“像”某种物体的 $\mathbf{x}$ 具有更大的概率值；

算法根据一组样本进行学习，得到概率密度函数，然后根据它进行采样。或者根据随机噪声直接输出样本。

算法生成的人脸图像

基础术语与基本概念

数据获取问题（一个很大的坑）

在人工智能和模式识别中，为了公平的比较不同算法性能，设计出了各种数据集作为测评的基准（benchmark）；因此，在很多论文中都会宣称：

在××数据集上我的算法最厉害...

在××数据集上我的算法和当前SOTA性能差不多，但运行速度更快/更轻量化/内存占用更低...

虽然我的算法在××数据集上性能一般，但几个数据集一平均，我的最厉害.....

但是，这些数据集真的可以评价算法的优劣么？

大多数数据集都宣称：其数据集是多样化，非刻意的人工的寻找样本

但事实是这样的数据集还没出现。

基础术语与基本概念

数据获取问题（一个很大的坑）

Unbiased Look at Dataset Bias

2011年CVPR

Antonio Torralba
Massachusetts Institute of Technology
torralba@csail.mit.edu

Alexei A. Efros
Carnegie Mellon University
efros@cs.cmu.edu

Acknowledgements: The authors would like to thank the Eyjafjallajokull volcano as well as the wonderful *kirs* at the Buvette in Jardin du Luxembourg for the motivation (former) and the inspiration (latter) to write this paper. This work is part of a larger effort, joint with David Forsyth and Jay Yagnik, on understanding the benefits and pitfalls of using large data in vision. The paper was co-sponsored by ONR MURIs N000141010933 and N000141010934.

Disclaimer: No graduate students were harmed in the production of this paper. Authors are listed in order of increasing procrastination ability.

https://ieeexplore.ieee.org/document/5995347

从数据获取的发展史来看，尽管都说要避免“偏见”，但每一个新数据集都不可避免的进入了另一种“偏见”

引起数据集偏见的主要原因：

1）**“选择偏见”**，人们通常偏爱选择某类数据，例如风景、街景、或用关键词搜索的网络图片。

2）**“拍照偏见”**，摄影师通常喜欢用相似角度拍同一种物体。如google image搜索“mug”绝大多数杯子的手柄都在右边。

3）**“标签偏见”**，特别是语义分类，同种东西可能有不同称呼，例如“草地”“草坪”，“绘画”“图片”。

4）**“负样本偏见”**，对于分类器而言，想要分出来的东西是正样本，其余都是负样本。一般来说，负样本应该是无穷大的，但实际上，我们只能用有限多的负样本。

基础术语与基本概念

模式分类的理论方法

- 模版匹配法(templete matching)
- 统计方法(statistical pattern recognition)： 1950s-
- 神经网络方法(neural network)： 1980s-
- 支持向量机、核方法：1990s-
- 多分类器、集成学习：1990s-
- Bayes学习：1990s-

上世纪五十、六十年代开始迅速发展，

七十年代初奠定理论基础，

九十年代大规模应用。

基础术语与基本概念

模版匹配法

是一种最原始、最基本的模式识别方法；

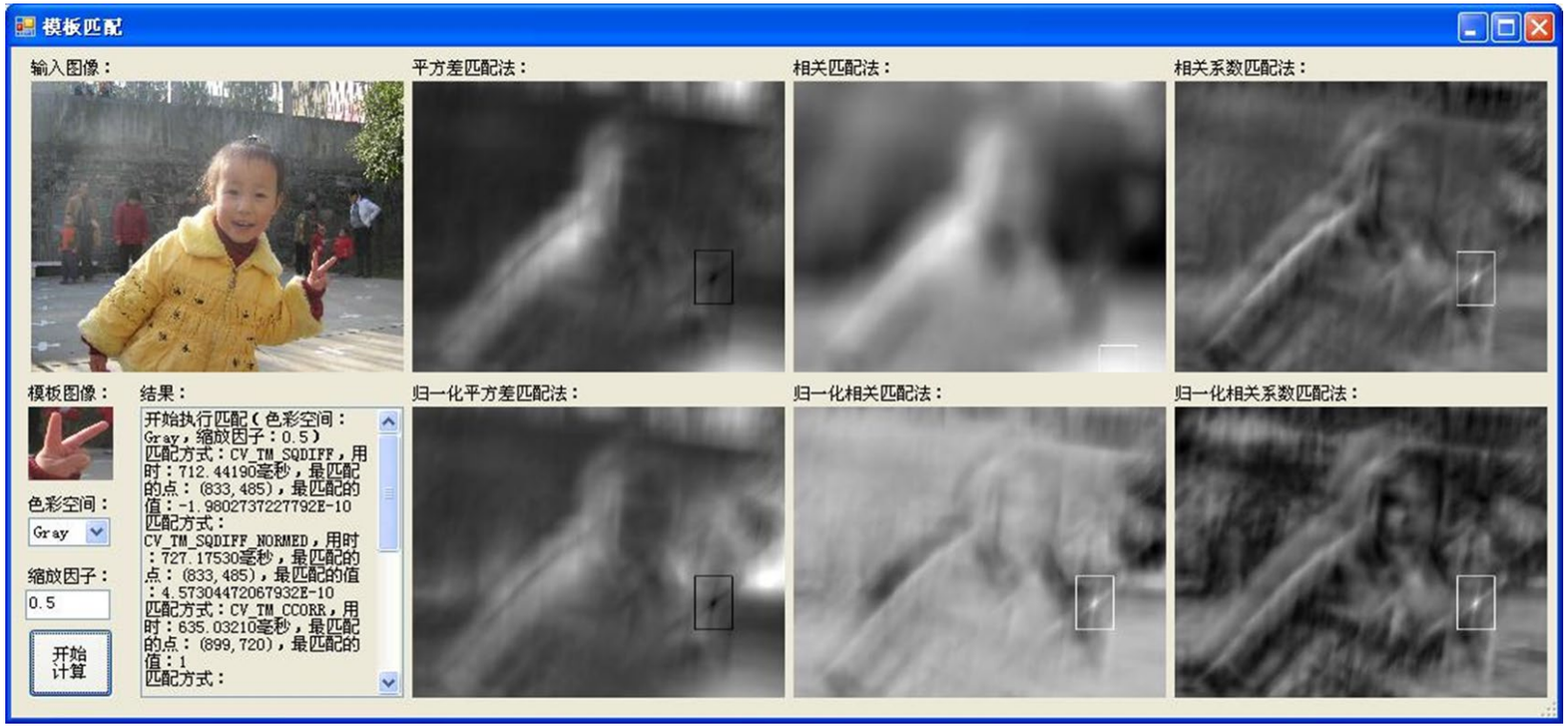
首先对每个类别建立一个或多个模版；

输入样本和数据库中每个类别的模版进行比较，例如求相关或距离；

根据相似性（相关性或距离）进行决策；

优点：直接、简单；

缺点：适应性差。



基础术语与基本概念

统计模式识别方法

根据训练样本，建立决策边界(decision boundary)

- 统计决策理论　——根据每一类总体的**概率分布**决定决策边界；
- 判别式分析方法 ——给出带参数的决策边界，根据某种准则，
- 由训练样本决定“最优”的参数

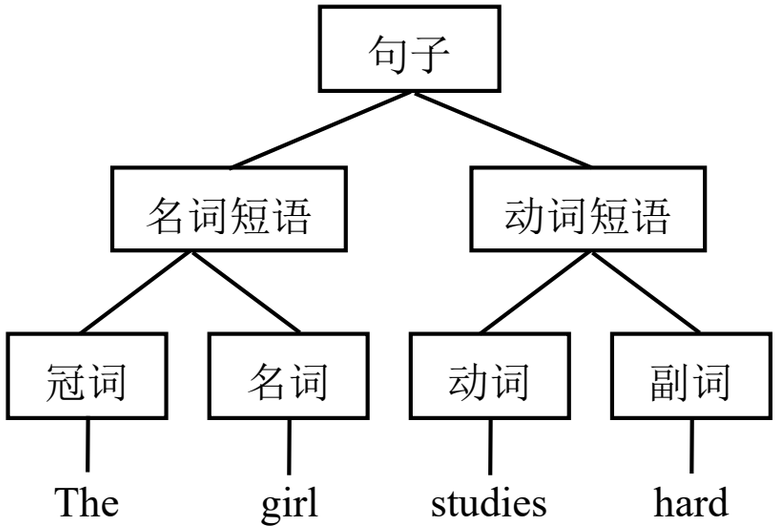
以模式集在特征空间中分布的类概率密度函数为基础，对统计特性进行研究。包括贝叶斯决策理论、判别函数法、 K近邻聚类法，非线性映射法，特征分析法，主因子分析法等。

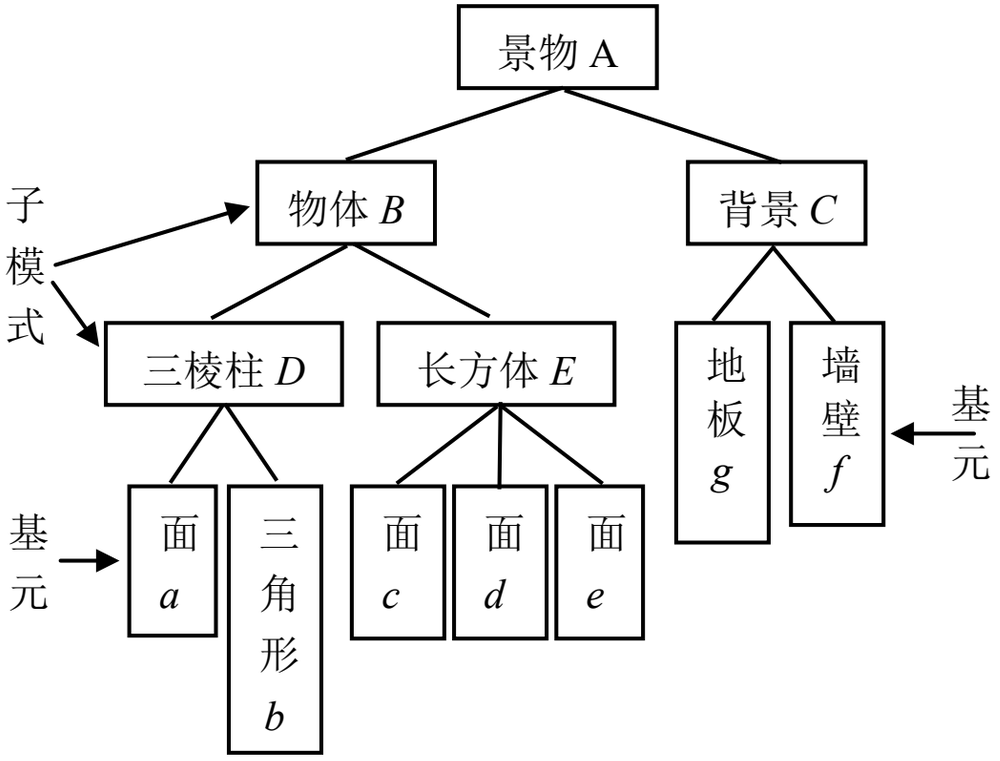
基础术语与基本概念

结构模式识别方法

根据识别对象的结构特征，以形式语言理论为基础的一种模式识别方法。

把复杂模式分化为较简单的子模式乃至基元，各层次间关系通过“结构法”来描述，相当于语言中的语法。用小而简单的基元与语法规则来描述大而复杂的模式。





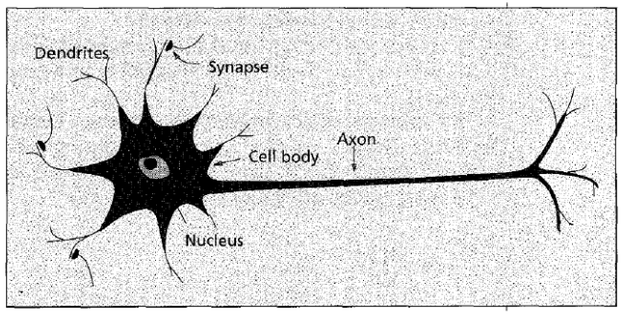
基础术语与基本概念

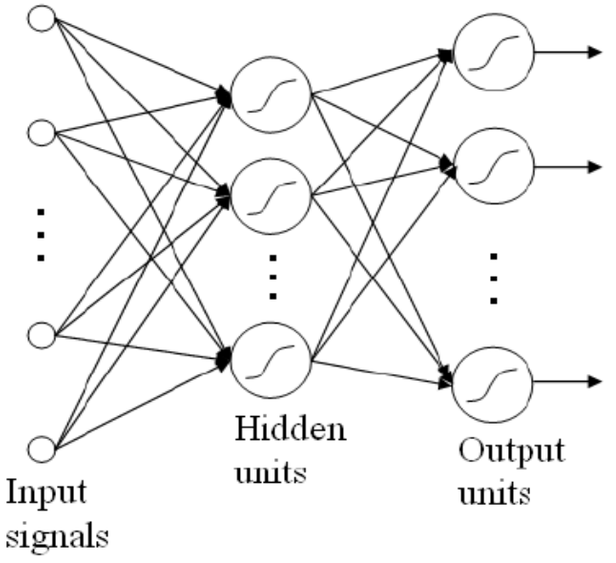
神经网络

以人工神经元为基础，模拟人脑神经细胞的工作特点。对脑部工作的生理机制进行模拟，实现形象思维的模拟。是一种大规模并行计算的数学模型。具有学习、推广、自适应、容错、分布表达和计算的能力。

优点：可以有效的解决一些复杂的非线性问题。

缺点：黑箱，蛮力计算，缺少有效的学习理论。

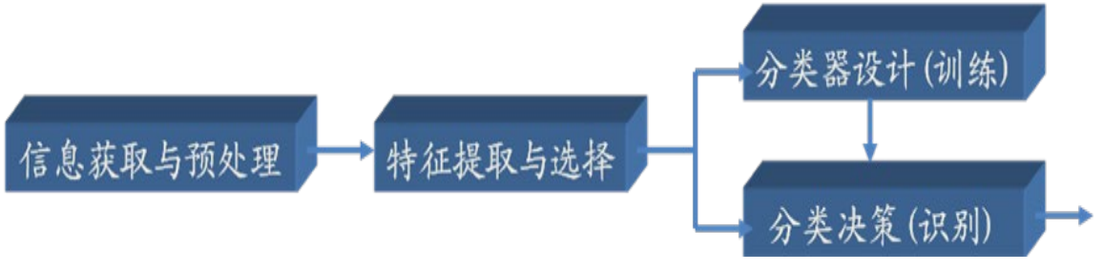




基础术语与基本概念

模式分类的实现方法

① 有监督识别（supervised PR）：有已知样本或足够的先验知识；



② 无监督识别（unsupervised PR）：无已知样本以及无先验知识，采用聚类分析的方法；

基础术语与基本概念

有监督学习

样本带有标签值，称为监督信号；

有学习过程，根据训练样本学习，得到模型，然后用于预测；

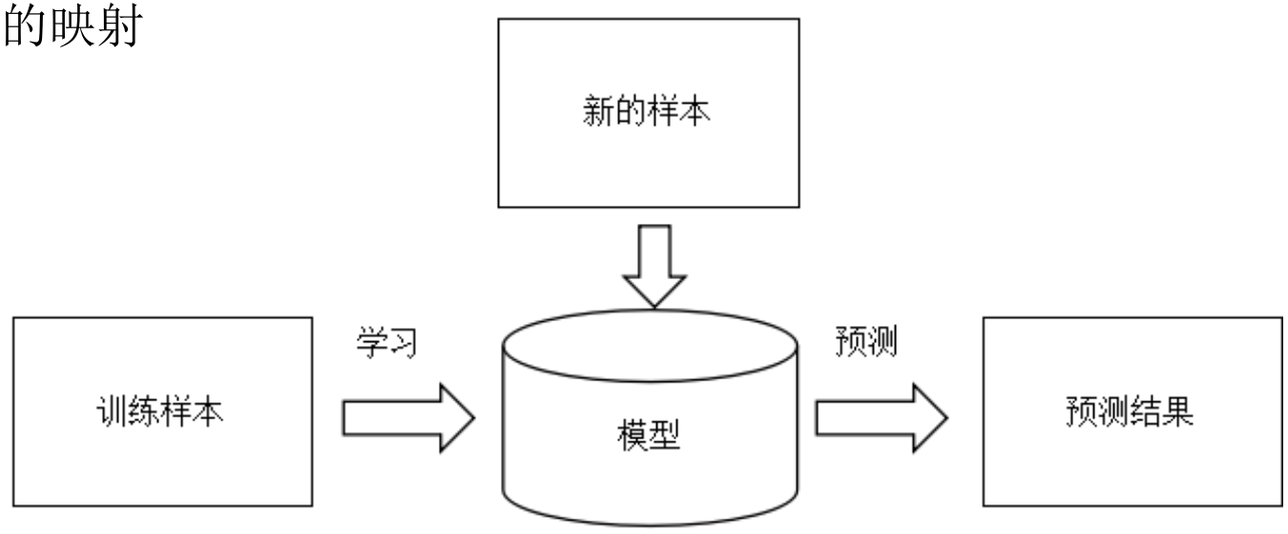
训练样本为 $(\mathbf{x}, y)$，$\mathbf{x} \in \mathbb{R}^n$ 为特征向量，y 为标签值，

机器学习模型实现从特征向量到标签值的映射

按照标签值的类型可以进一步分为两类

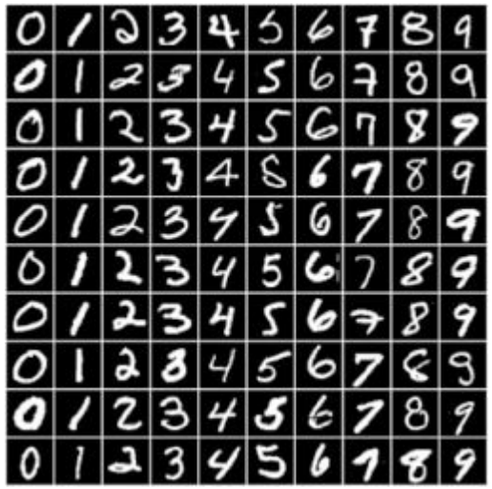
分类问题 - 标签值为整数编号

回归问题 - 标签值为实数



基础术语与基本概念

有监督学习——分类问题



MNIST手写字符识别

确定样本所属的类别，通常以整数编号；

机器学习模型实现从特征向量到类别编号的映射：$\mathbb{R}^n \to \mathbb{Z}$

图像识别，语音识别问题是典型的分类问题。

如果类别数为2，称为二分类问题，样本标签值通常设置为+1和-1，分别称为正样本和负样本；

如果类别数大于2，称为多分类问题。0-9这10个手写阿拉伯数字的图像识别是典型的多分类问题。

基础术语与基本概念

有监督学习——回归问题

机器学习算法预测出一个实数值

机器学习模型实现从特征向量到实数值的映射 $\mathbb{R}^n \to \mathbb{R}$

根据一个人的特征预测其收入，是典型的回归问题

性别	年龄	学历	工作年限	所在城市	行业	收入（万）
男	31	本科	9	北京	金融	80
男	24	本科	2	深圳	金融	20
男	45	博士	18	深圳	互联网	230
女	25	本科	3	深圳	互联网	35
女	27	硕士	2	北京	财务	18
女	35	博士	8	上海	教育	30

基础术语与基本概念

有监督识别（supervised PR）

- **分析问题**：分析是否属于模式识别问题，把所研究的目标表示为一定的类别，分析给定数据或者可以观测的数据中哪些因素可能与分类有关。
- **原始特征获取**：设计实验，得到已知样本，对样本实施观测和预处理，获取可能与样本分类有关的观测向量（原始特征）。
- **特征提取与选择**：为了更好地进行分类，可能需要采用一定的算法对特征进行再次提取和选择。
- **分类器设计**：选定一定的分类器方法，用已知样本进行分类器训练。

基础术语与基本概念

无监督学习

样本没有标签值，没有训练过程，机器学习算法直接对样本进行处理，得到某种结果

无监督学习算法的细分

聚类问题 - 无监督分类，将一组样本划分成多个子集

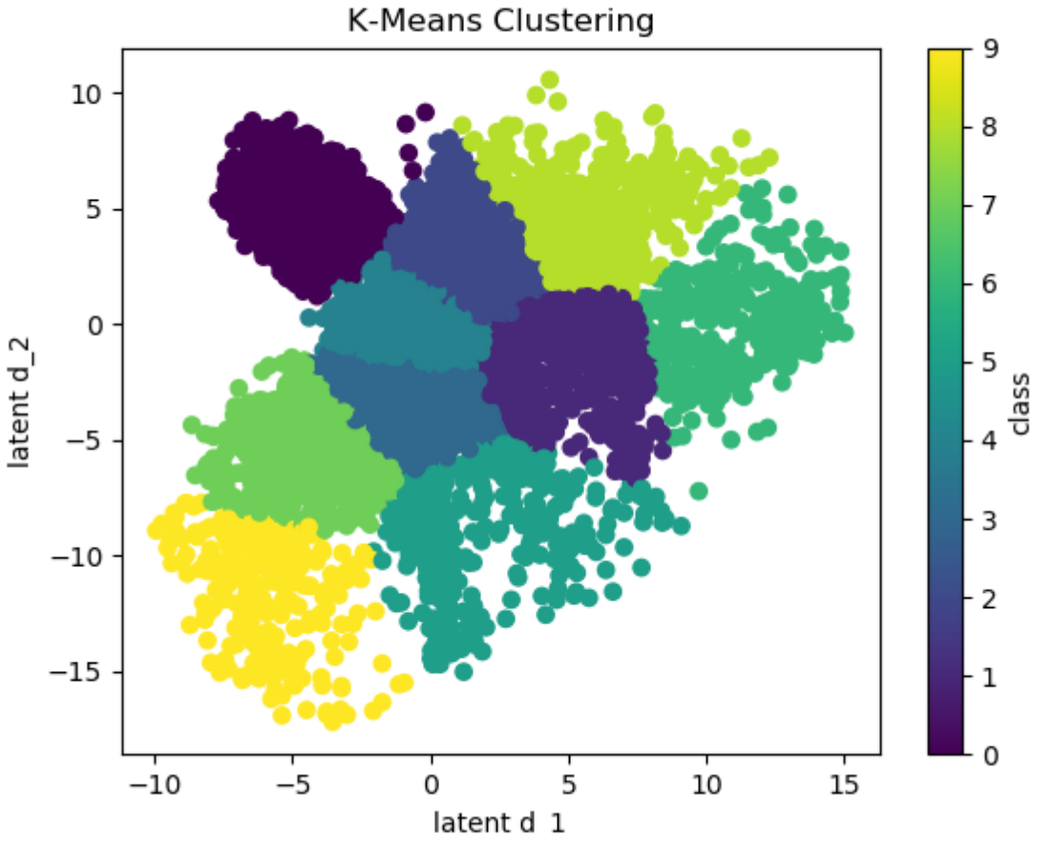
数据降维问题 - 将向量映射到更低维的空间

基础术语与基本概念

无监督学习——聚类

聚类问题也是分类问题，但没有训练过程；

把一批样本划分成多个类，使得在某种相似度指标下每一类中的样本尽量相似，不同类的样本之间尽量不同



基础术语与基本概念

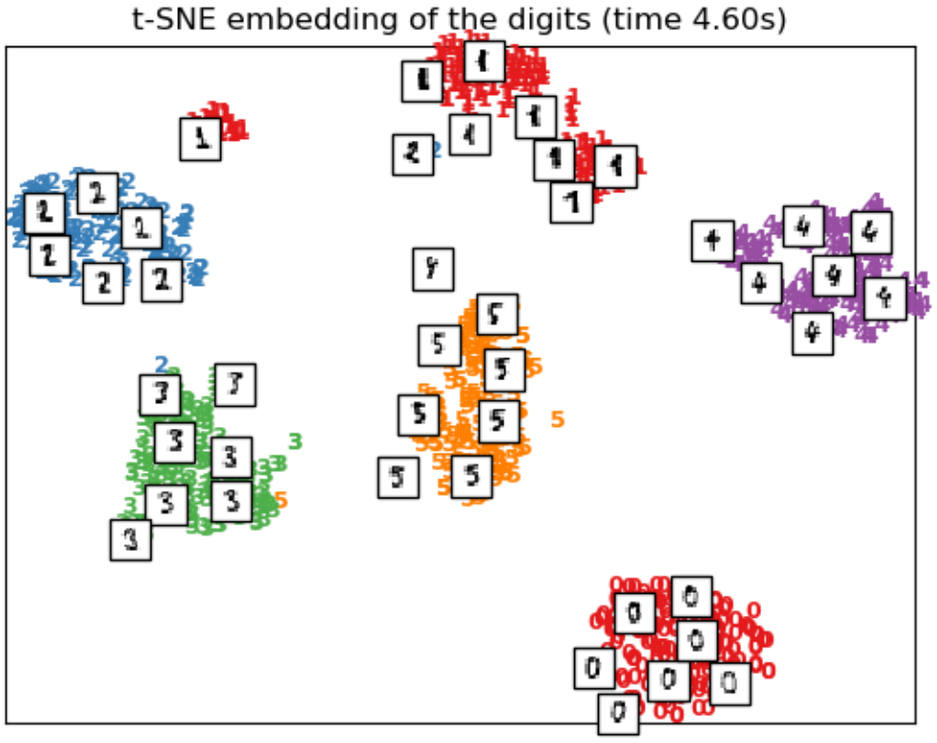
无监督学习——降维

它将 n 维空间中的向量 $\mathbf{x}$ 通过某种映射函数映射到更低维的 m 维空间中

$$\mathbf{y} = \phi(\mathbf{x}) \qquad m \ll n$$

降维后的数据更易于处理，且可以可视化。

右侧是将0-9这些手写数字图像投影到2维空间的结果



基础术语与基本概念

无监督识别（unsupervised PR）

- **分析问题：**分析研究目标能否通过寻找适当的聚类来达到；如果可能，猜测可能的或希望的类别数目；分析给定数据或者可以观测的数据中哪些因素可能与聚类有关。
- **原始特征获取：**设计实验，得到待分析的样本，对样本实施观测和预处理，获取可能与样本聚类有关的观测向量（原始特征）。
- **特征提取与选择：**为了更好地聚类，需采用一定的算法对特征进行再次提取和选择。
- **聚类分析：**选定一定的非监督模式识别方法，用样本进行聚类分析。
- **结果解释：**考察聚类结果的性能，分析所得聚类与研究目标之间的关系，根据领域知识分析结果的合理性，对聚类的含义给出解释。

基础术语与基本概念

强化学习

模拟人的行为，源自于行为主义心理学，确定在每种状态下要执行的动作，以达到某种目标。

算法通过学习，得到策略函数，其输入为状态 s，输出为在这种状态下应该执行的动作 a

$$a = \pi(s)$$

围棋和自动驾驶是典型的强化学习问题

围棋需要根据当前棋局确定在什么位置落子，棋局即为状态，落子即为动作；

自动驾驶的汽车需要根据路况，汽车自身状态确定任何行驶。

基础术语与基本概念

任务

- 预测目标：
  - 分类:离散值
    - 二分类:好瓜;坏瓜
    - 多分类:冬瓜;南瓜;西瓜
  - 回归:连续值

    瓜的成熟度
  - 聚类:无标记信息
- 有无标记信息
  - 监督学习：分类、回归
  - 无监督学习：聚类
  - 半监督学习：两者结合

基础术语与基本概念

泛化能力

机器学习的目标是使得学到的模型能很好的适用于“新样本”, 而不仅仅是训练集合，我们称模型适用于新样本的能力为泛化(generalization)能力。

通常假设样本空间中的样本服从一个未知分布$\mathcal{D}$,样本从这个分布中独立获得，即“独立同分布”(i.i.d)。一般而言，训练样本越多越有可能通过学习获得强泛化能力的模型。

基础术语与基本概念

生成模型与判别模型

有监督学习算法可以进一步分为生成模型与判别模型；

生成模型对样本特征向量与标签值的联合概率分布 $p(\mathbf{x}, \mathbf{y})$ 或对条件概率 $p(\mathbf{x} \mid \mathbf{y})$ 进行建模；

生成模型需要**对样本的特征向量服从某种概率分布**建模。

判别模型直接对后验概率 $p(\mathbf{y} \mid \mathbf{x})$ 进行建模；

或者直接预测标签值 $\mathbf{y}=f(\mathbf{x})$，不使用概率模型；

判别模型**不对样本特征向量的概率分布进行建模。**

基础术语与基本概念

生成模型与判别模型

类别	算法
生成模型	贝叶斯分类器 高斯混合模型 贝叶斯网络 隐马尔可夫模型 受限玻尔兹曼机 生成对抗网络 变分自动编码器
判别模型	决策树 kNN算法 人工神经网络 支持向量机 logistic回归 softmax回归 随机森林 boosting算法 条件随机场

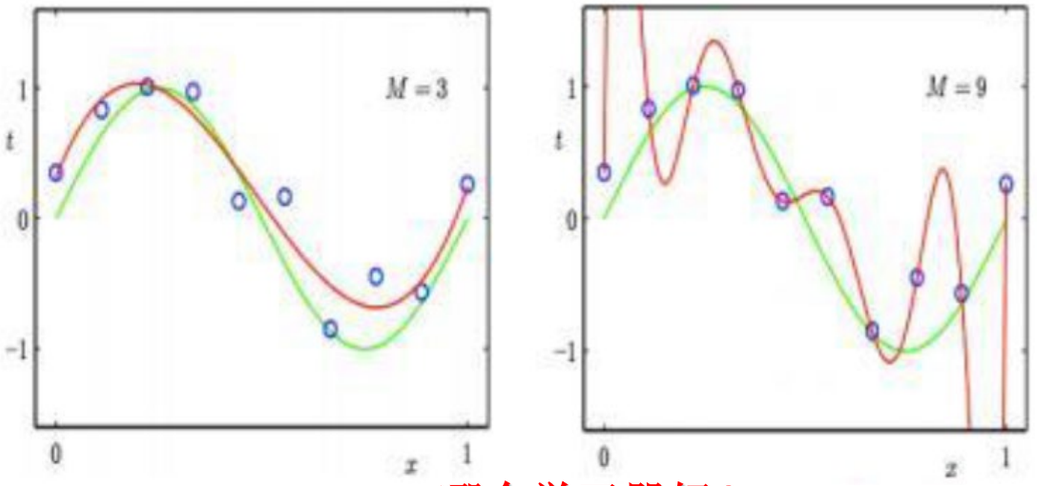
基础术语与基本概念

错误率&误差

- **错误率:** 错分样本的占比：$E = a / m$
- **误差：** 样本真实输出与预测输出之间的差异
  - 训练（经验）误差：训练集上
  - 测试误差：测试集
  - 泛化误差：除训练集外所有样本

讨论：

- 由于事先并不知道新样本的特征，我们只能努力使经验误差最小化；
- 很多时候虽然能在训练集上做到分类错误率为零，但多数情况下这样的学习器并不好。



哪个学习器好？

基础术语与基本概念

过拟合与欠拟合

- 过拟合:

学习器将训练样本学得“太好”，将训练样本自身特点也当做所有样本一般化特性，导致泛化性能下降；

解决方案：
- 优化目标加正则项
- early stop

- 欠拟合：

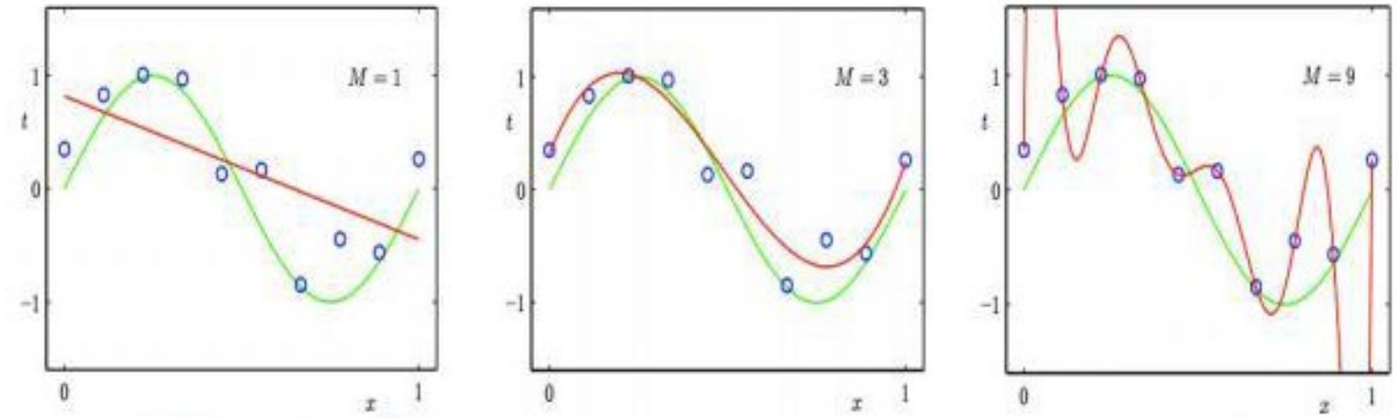
对训练样本的一般性质尚未学好

解决方案：
- 决策树:拓展分支
- 神经网络:增加训练轮数

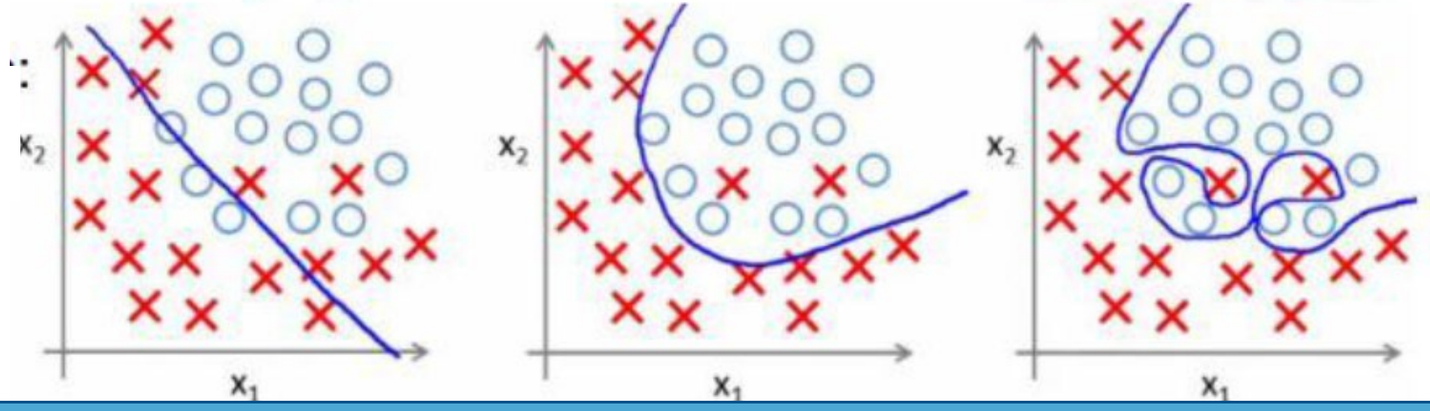
基础术语与基本概念

过拟合与欠拟合

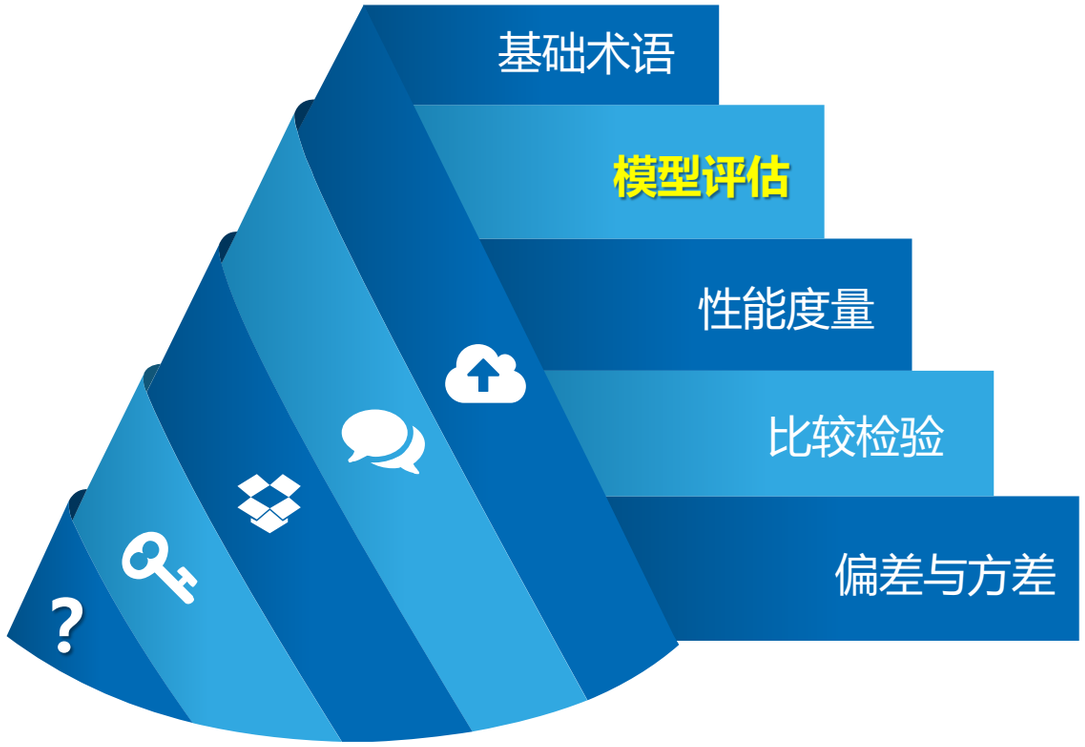
回归：



分类：



训练集表现	测试集表现	结论
不好	不好	欠拟合
好	不好	过拟合
好	好	适度拟合



- **基础术语与基本概念**
  数据，学习方法，泛化能力，…
- **模型评估方法**
  留出法，交叉验证法，…
- **模型性能度量**
  错误率，精度，P-R曲线，…
- **比较检验**
  二项检验，t检验，交叉验证，…
- **偏差与方差**
  偏差，方差，…

模型评估方法

实际任务中往往会对学习器的泛化性能、时间开销、存储开销、可解释性等方面的因素进行评估并做出选择；

假设测试集是从样本真实分布中独立采样获得，将测试集上的“测试误差”作为泛化误差的近似，所以测试集要和训练集中的样本尽量互斥。

Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning

https://sebastianraschka.com/pdf/manuscripts/model-eval.pdf

Sebastian Raschka
Michigan State University
January 2018
raschkas@msu.edu

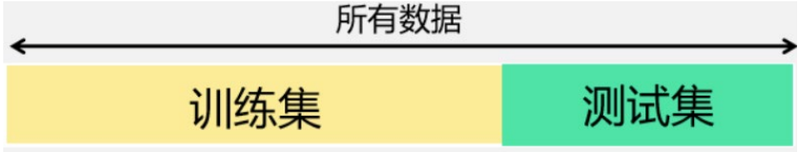
通常将包含个 m 样本的数据集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$ 拆分成训练集 S 和测试集 T：

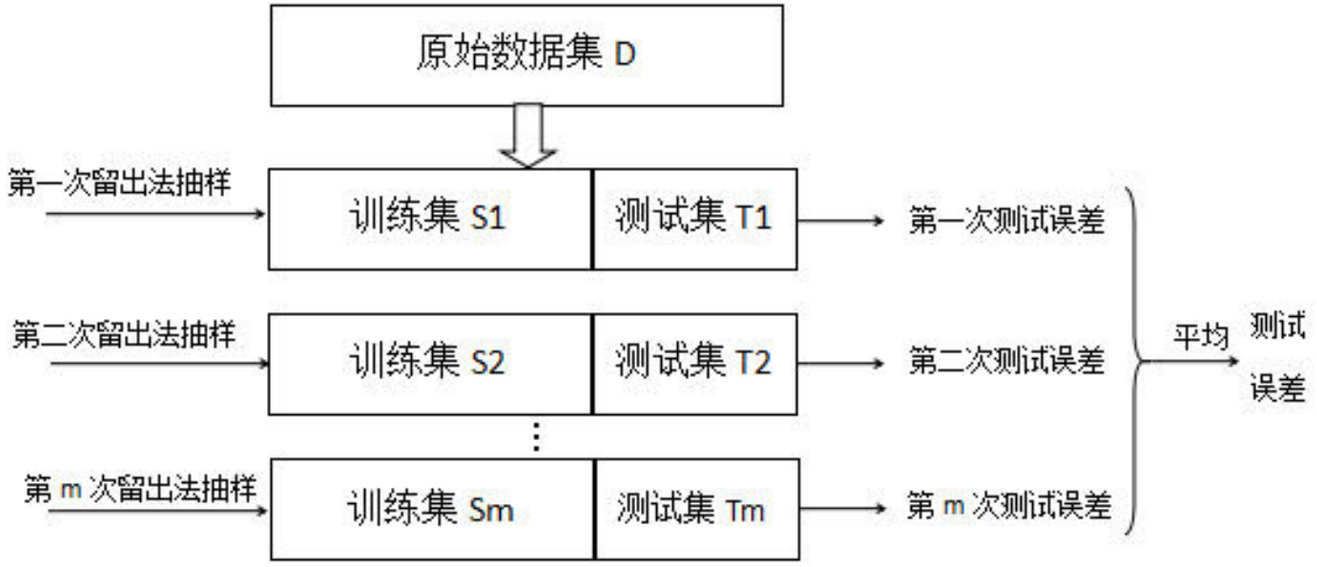
根据拆分的方法，有不同模型评估方法：留出法、留一法、交叉验证法、自助法。

模型评估方法——留出法

留出法（hold-out）：

- 直接将数据集随机划分为两个互斥集合；
- 训练/测试集划分要尽可能保持数据分布的一致性；
- 一般若干次随机划分、重复实验取平均值；
- 训练/测试样本比例通常为2:1~4:1。



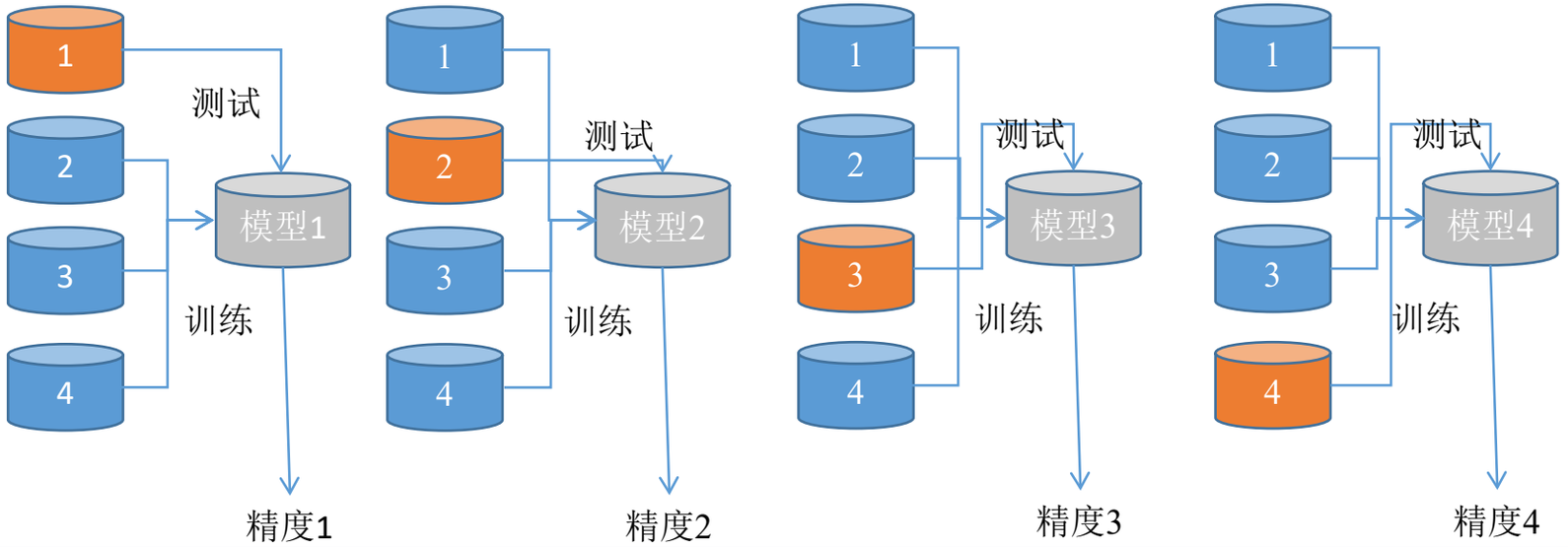


模型评估方法——交叉验证法

k折交叉验证法（ *k*-folds cross validation）：

为充分利用所有样本，将数据集划分为 *k* 个大小相似的互斥子集，每次用 *k*-1 个子集的并集作为训练集，余下的子集作为测试集，最终返回 *k* 个测试结果的均值，*k* 最常用的取值是5、10等。

K折	SROCC	PLCC
1	0.9553	0.9482
2	0.9122	0.9137
3	0.9715	0.9703
4	0.8748	0.8533
5	0.9029	0.9239
AVERAGE	0.9233	0.9218



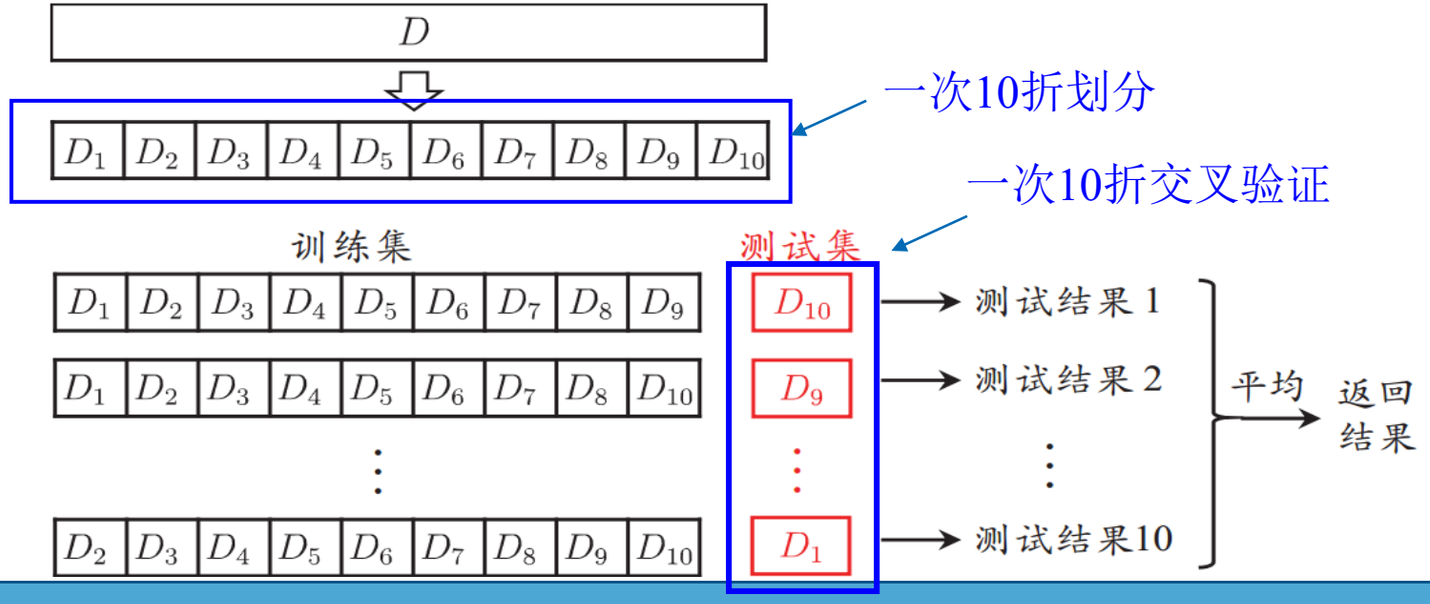
$$\frac{\text{精度1}+\text{精度2}+\text{精度3}+\text{精度4}}{4}$$

模型评估方法——交叉验证法

p 次 *k* 折交叉验证法

与留出法类似，将数据集 D 划分为 k 个子集同样存在多种划分方式；

为了减小因样本划分不同而引入的差别，k 折交叉验证通常随机使用不同的划分重复 p 次，最终的评估结果是这 p 次 k 折交叉验证结果的均值，例如常见的“10次10折交叉验证”。



模型评估方法——留一法

留一法（Leave-One-Out cross validation，简称LOO）：

假设数据集 D 包含 m 个样本，若令 $k=m$ ，则得到留一法：

（1）不受随机样本划分方式的影响， 只有唯一的方式划分为 m 个子集——每个子集包含一个样本；

（2）由于训练集只比数据集少一个样本，即训练集特征的概率分布与数据集的分布更接近，绝大多数情况下，其结果被认为比较准确；

（3）当数据集比较大时，计算开销难以忍受。

模型评估方法——自助法

自助法（ Bootstrapping ）：

我们希望评估的是用原始数据集 D 训练出的模型；

留出法和交叉验证法的训练集比原始数据集 D 小，必然引入因训练集不同导致的估计偏差；

留一法受训练样本规模变化的影响较小，但是计算复杂度太高。

自助法是以自助采样（bootstrap sampling）为基础的有放回抽样：

（1）原始数据集 D 包含 m 个样本，对它进行自助采样产生训练数据集 D'：

（2）随机从 D 中挑选一个样本并拷贝至 D'，然后将该样本放回 D 中，下次抽样时仍可能被选中；

（3）重复执行 m 次该过程，可得到了包含 m 个样本的数据集 D'。

模型评估方法——自助法

自助法分析

初始数据集 D 中有一部分样本会在训练集 D' 中多次出现，也有一部分样本不会在 D' 中出现。

一个简单的估计：样本在 m 次采样中始终不被采到的概率为 $\left(1-\frac{1}{m}\right)^m$

取极限得到：$\lim_{m\to\infty}\left(1-\frac{1}{m}\right)^m \to \frac{1}{e} \approx 0.368$

推导：已知极限公式 $\lim_{m\to\infty}\left(1+\frac{1}{m}\right)^m = e = 2.71828...$，其中 e 为自然对数的底数；

$$\lim_{m\to\infty}\left(1-\frac{1}{m}\right)^m = \lim_{m\to\infty}\left(1+\frac{1}{-m}\right)^m \xrightarrow{n=-m} \lim_{n\to\infty}\left(\left(1+\frac{1}{n}\right)^n\right)^{-1} = \frac{1}{e}$$

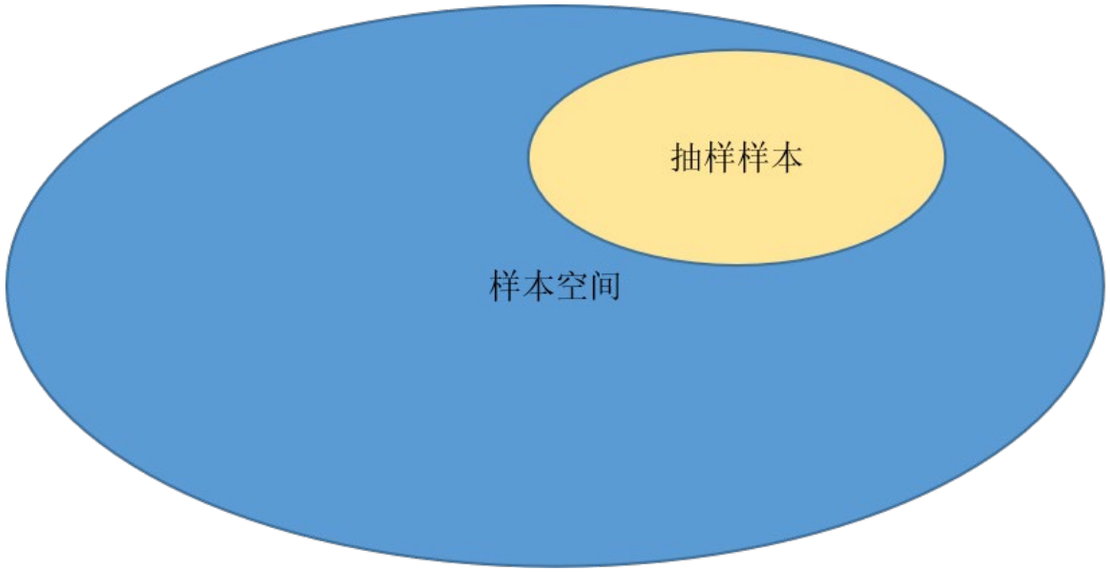
模型评估方法——自助法

自助法分析

- 实际模型与预期模型都使用 m 个训练样本；
- 约有1/3的样本没在训练集中出现；
- 从初始数据集中产生多个不同的训练集，对集成学习有很大的好处；
- 自助法在数据集较小、难以有效划分训练/测试集时很有用；
- 由于改变了数据集分布可能引入估计偏差，在数据量足够时，留出法和交叉验证法更常用。

模型评估方法——小结

抽样误差：抽样的训练样本集和整体数据集之间的偏差



	采样方法	与原始训练数据集的分布是否相同	相比原始训练数据集的容量	是否适用小数据集	是否适用大数据集	是否存在估计偏差
留出法	分层抽样	否	变小	否	是	是
交叉验证法	分层抽样	否	变小	否	是	是
自助法	放回抽样	否	不变	是	否	是

- **基础术语与基本概念**
  数据，学习方法，泛化能力，…
- **模型评估方法**
  留出法，交叉验证法，…
- **模型性能度量**
  错误率，精度，P-R曲线，…
- **比较检验**
  二项检验，t检验，交叉验证，…
- **偏差与方差**
  偏差，方差，…

模型性能度量

性能度量是衡量模型泛化能力的评价标准，反映了任务需求；

使用不同的性能度量往往会导致不同的评判结果。

在预测任务中，给定样例集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$，评估学习器 f 的性能，也即把预测结果 $f(x)$ 和真实标记 y 比较。

常用回归任务的性能度量：MSE，MAE

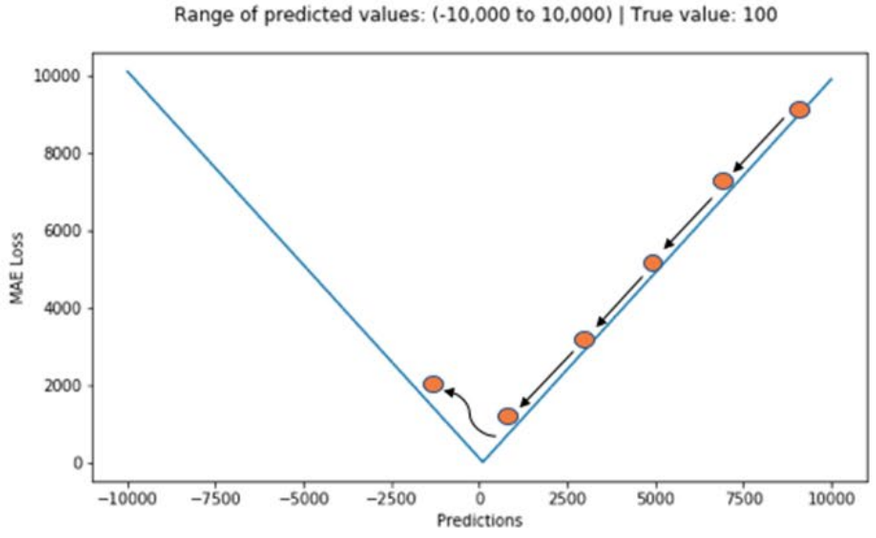
常用分类任务的性能度量：错误率，准确率，召回率，精度，F1等

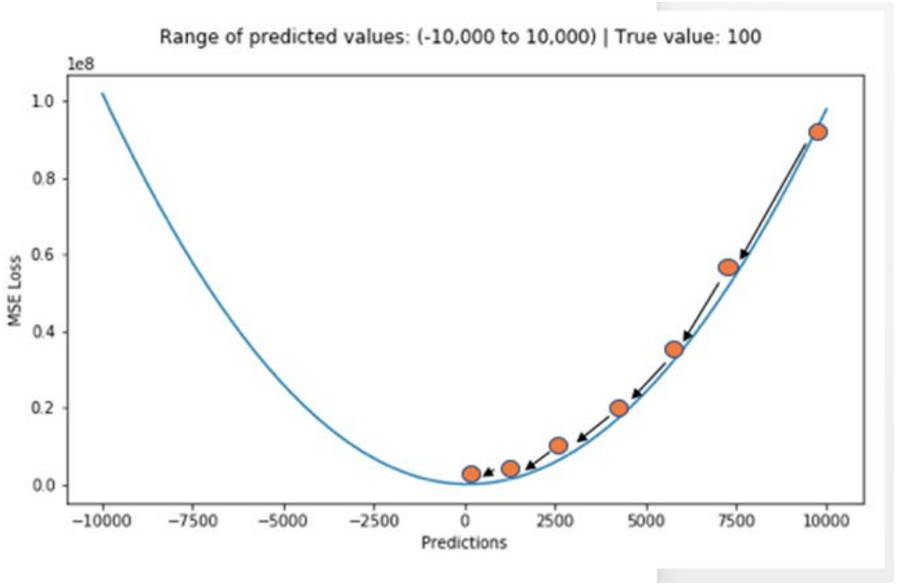
模型性能度量——回归误差

回归任务最常用的性能度量是“平均绝对误差（MAE）”和 “均方误差（MSE）”

MAE： $$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\left|f(x_i)-y_i\right|$$

MSE： $$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\left(f(x_i)-y_i\right)^2$$





模型性能度量——分类误差

对于分类任务，错误率和精度是最常用的两种性能度量：

$$\text{错误率}=\frac{\text{错误分类的测试样本数}}{\text{测试样本总数}} \qquad \text{准确率}(Accuracy)=\frac{\text{正确分类的测试样本数}}{\text{测试样本总数}}=1-\text{错误率}$$

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

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

对于有监督学习，通常将样本集分为训练集，验证集，测试集3个不相交的子集；

训练集用于模型训练得到模型参数，验证集用于确定算法超参数，测试集用于测试模型精度。

准确率Accuracy常用有：Top-1 Accuracy、Top-5 Accuracy

模型性能度量——精度和召回率

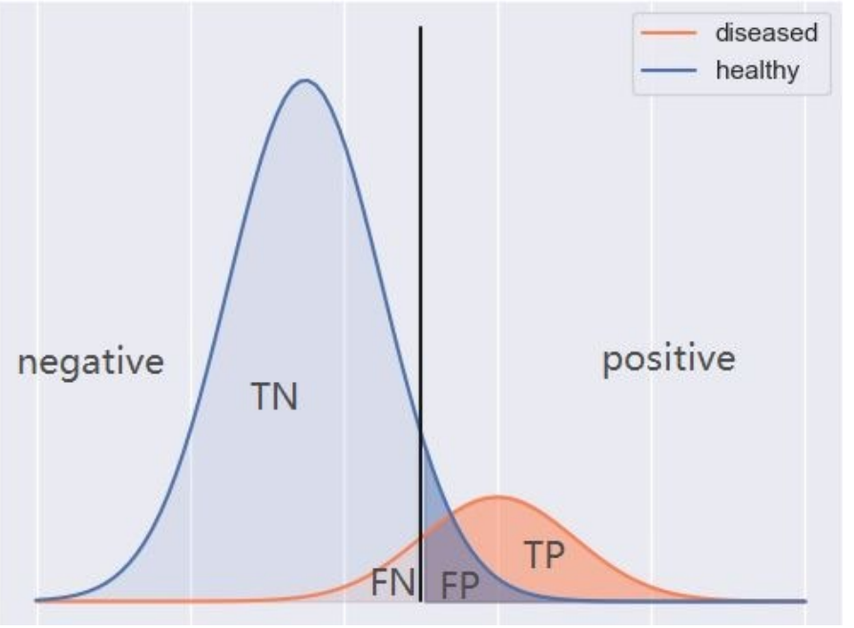
对二分类问题可以定义特殊的指标，以反映正样本和负样本不同的重要性；

样本的真实标签值与预测值有下面几种组合（分类结果混淆矩阵）

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)

以抓坏人为例，精度表示抓到的人中抓捕准确的比例，召回率表示是否将所有坏人都抓住。

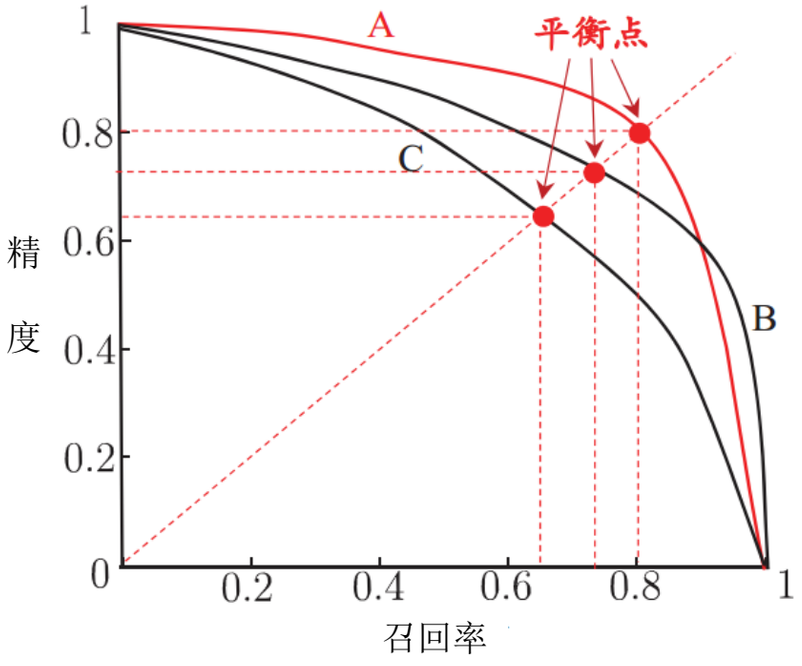


精度（precision，也称**查准率**）：预测为正例的测试样本中，（预测的）真正例所占的比例 $P = \frac{TP}{TP+FP}$

召回率（recall，也称**查全率**）：（预测的）真正例占（真实情况下）所有正例的比例 $R = \frac{TP}{TP+FN}$

模型性能度量——P-R曲线

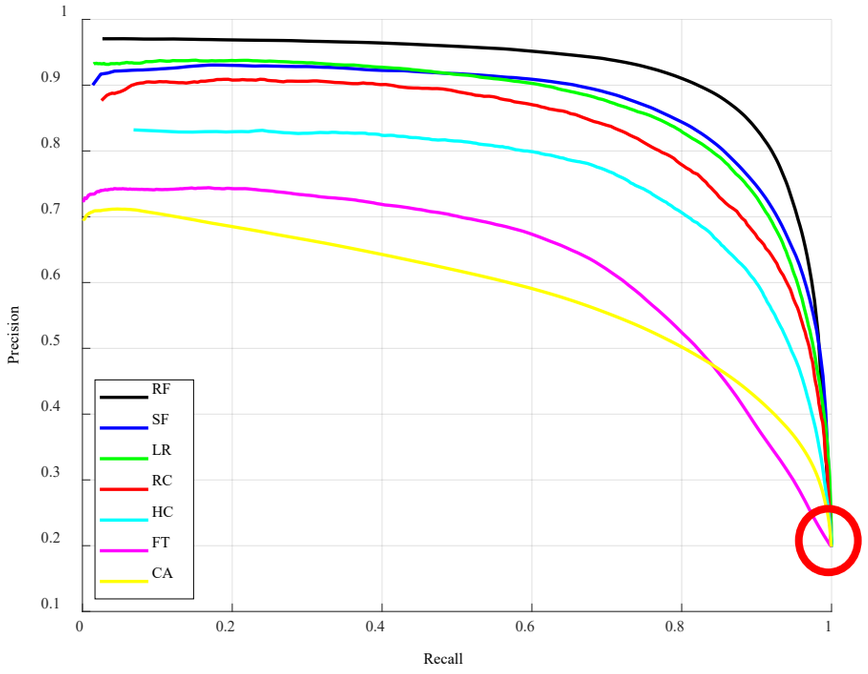
根据学习器的预测结果按正例可能性大小对样例进行排序，并逐个把样本作为正例进行预测，则可以得到查准率-查全率曲线，简称“P-R曲线”



P-R曲线与平衡点示意图

平衡点是曲线上“精度-召回率”时的取值，可用来用于度量P-R曲线有交叉的分类器性能高低

分析：实际P-R曲线



如果一个学习器的P-R曲线被另一个学习器的P-R曲线完全包住，则可断言后者的性能优于前者.

汇聚于同一点

模型性能度量——F值

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)

综合评价指标（F-Measure）

比P-R曲线平衡点更用常用的是 $F1$ 度量 $F1 = \dfrac{2 \times P \times R}{P + R} = \dfrac{2 \times TP}{\text{样例总数} + TP - TN}$

比 $F1$ 更一般的形式 F_β， $F_\beta = \dfrac{(1 + \beta^2) \times P \times R}{(\beta^2 \times P) + R}$

$\beta = 1$ ： 标准 $F1$

$\beta > 1$ ： 偏重查全率(逃犯信息检索)

$\beta < 1$ ： 偏重查准率(商品推荐系统)

模型性能度量——ROC曲线

分类结果混淆矩阵

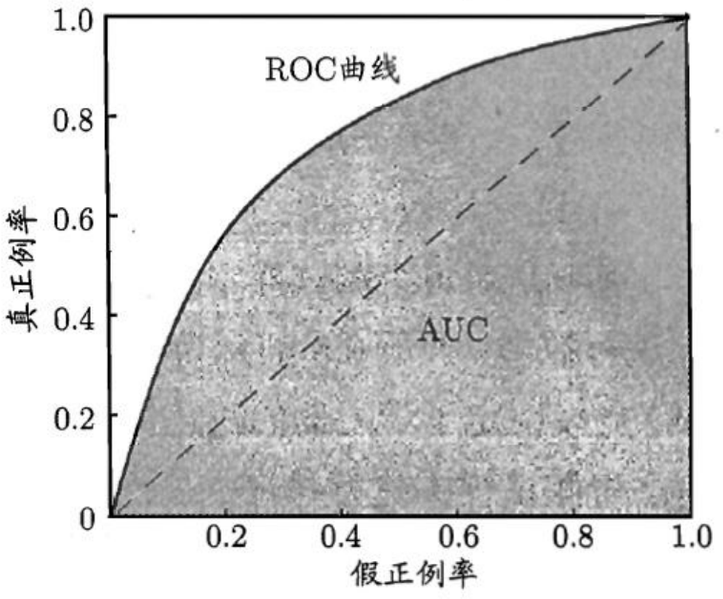
真实情况	预测结果	
	正例	反例
正例	TP (真正例)	FN (假反例)
反例	FP (假正例)	TN (真反例)

ROC曲线（Receiver Operating Characteristic）：称“受试者工作特征”，它是以“真正例率”（True Positive Rate, TPR）为纵轴，“假正例率”（False Positive Rate, FPR）为横轴的曲线。

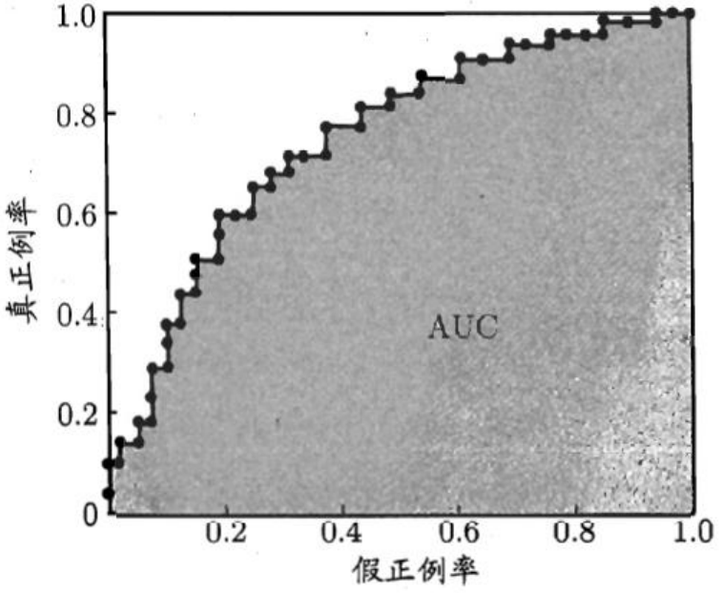
真正例率（召回率） $TPR = \dfrac{TP}{TP+FN}$

假正例率 $FPR = \dfrac{FP}{FP+TN}$

AUC值（Area Under ROC Curve）是ROC曲线下的面积（曲线与FPR轴间），介于0和1之间，作为数值可以直观的评价分类器的好坏，值越大越好。



(a) ROC 曲线与 AUC



(b) 基于有限样例绘制的 ROC 曲线与 AUC

模型性能度量——ROC曲线

ROC曲线举例说明：以抓坏人为例。

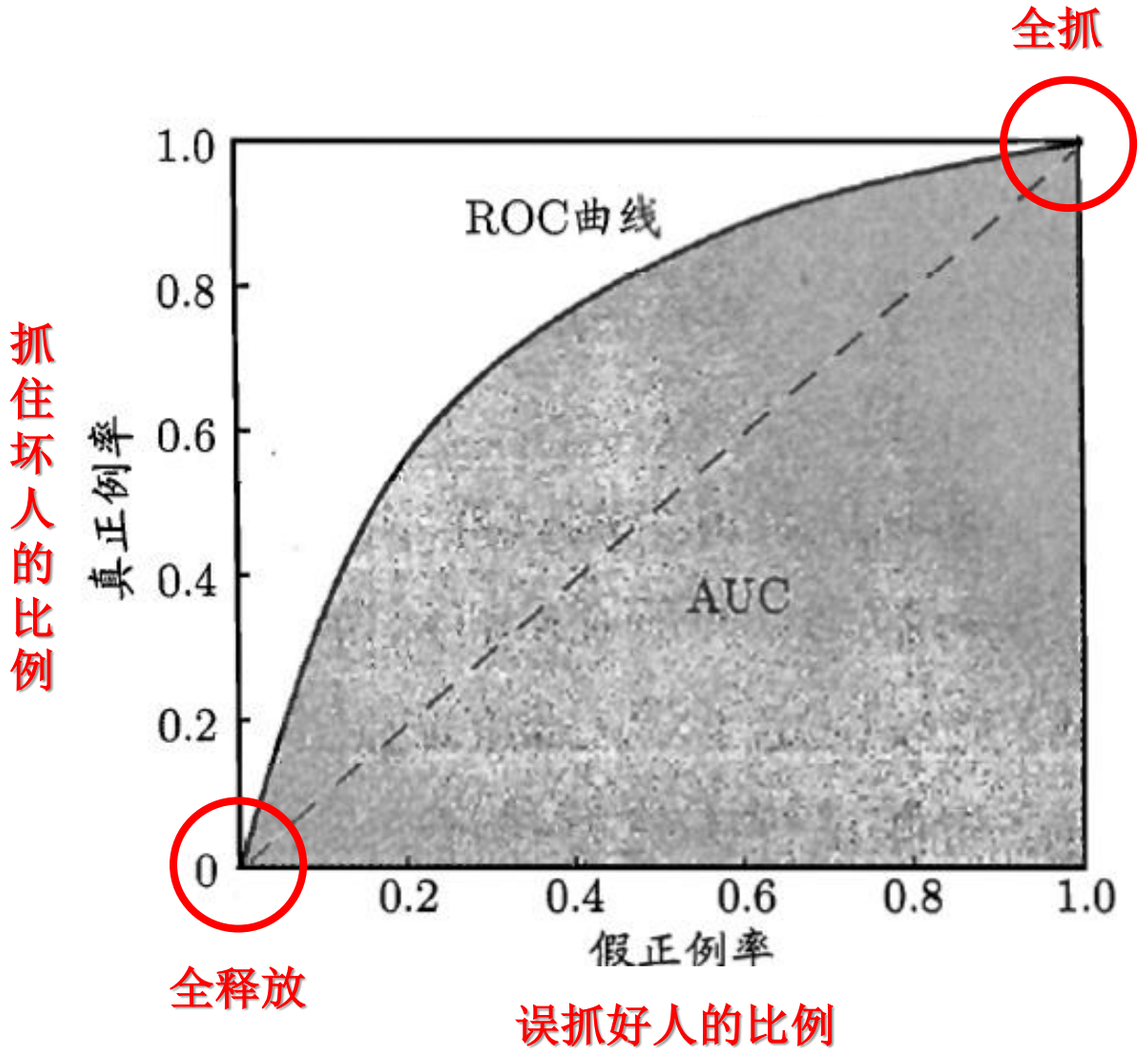
纵坐标表示为抓住坏人的比例

横坐标表示误抓好人的比例。

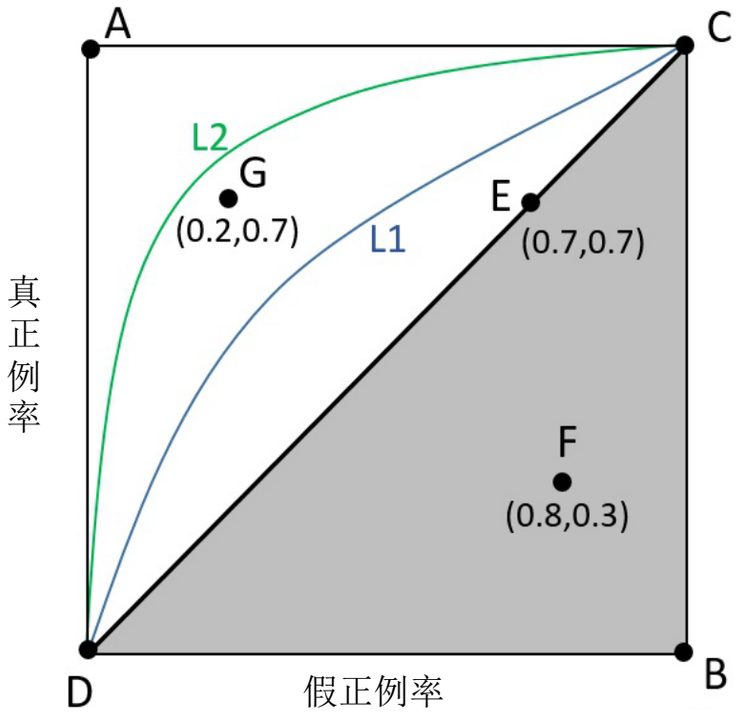
ROC曲线的两个极端点表示：

[0,0]点：不管好人坏人全放过，抓住坏人0%，误伤好人0%，不抓坏人也不误伤好人；

[1,1]点：不管好人坏人全抓住，也就是抓住坏人100%，误伤好人100%。



模型性能度量——ROC曲线



ROC曲线分析：

1）和P-R曲线类似，若一个模型曲线将另一个模型曲线“完全包住”，则可断定前者性能更优！

2）若两模型曲线相交，则需通过AUC面积来确定模型优劣。AUC越大，性能越好。如右图，曲线L2对应的性能优于曲线L1对应的性能。

3）A点是最完美的性能点，B处是性能最差点。

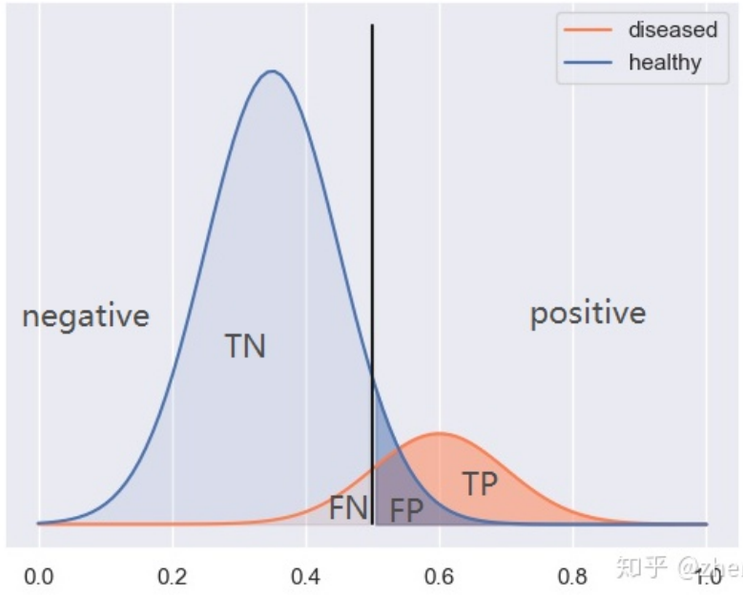
4）位于C-D连线上的点说明算法性能和随机猜测一样，如E点。

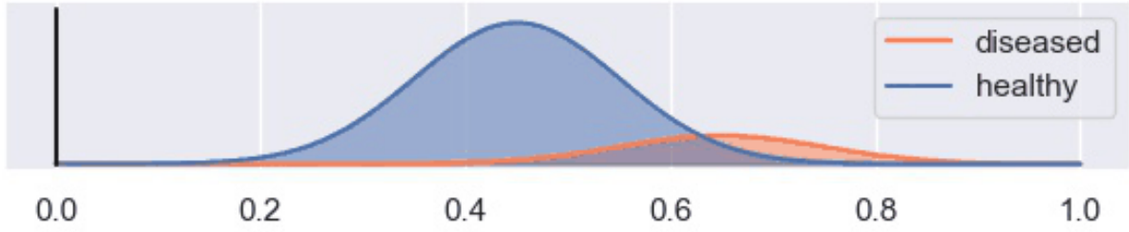
5）位于C-D连线之上（即曲线位于白色的三角形内）说明算法性能优于随机猜测–如G点；

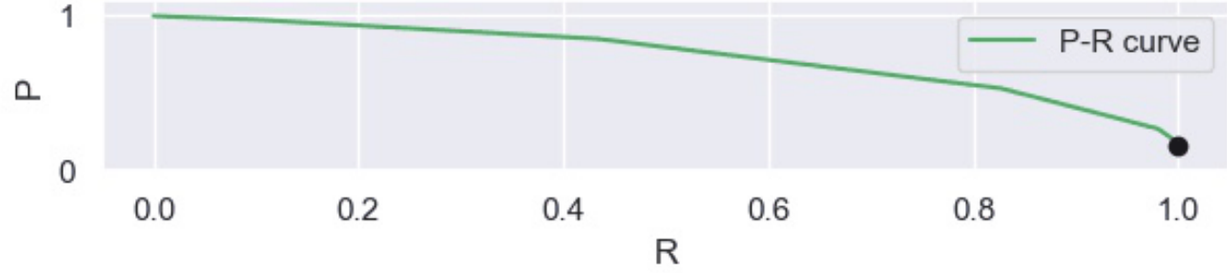
6）位于C-D连线之下（即曲线位于灰色的三角形内）说明算法性能差于随机猜测–如F点。

模型性能度量——动态图解释PR曲线和ROC曲线

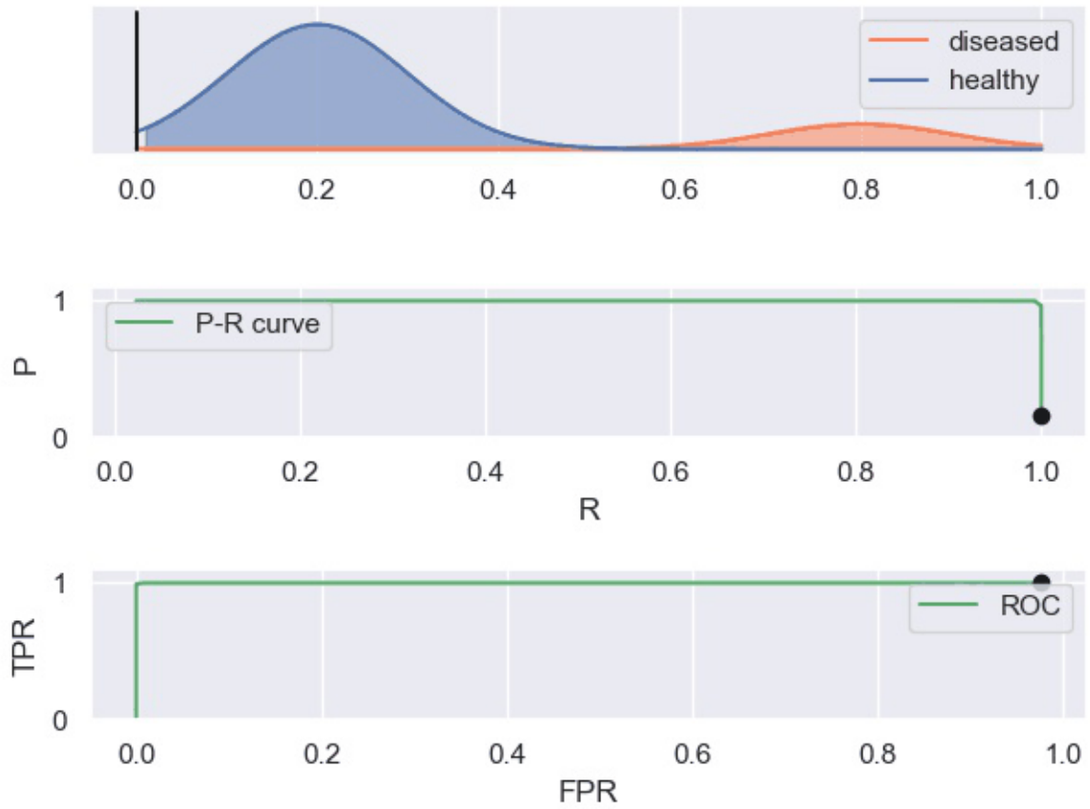
以病人检测为例，样本集包含了健康人和病人；红色高斯曲线表示真正的病人分布，蓝色高斯曲线表示健康人群分布，黑色线为阈值位置。



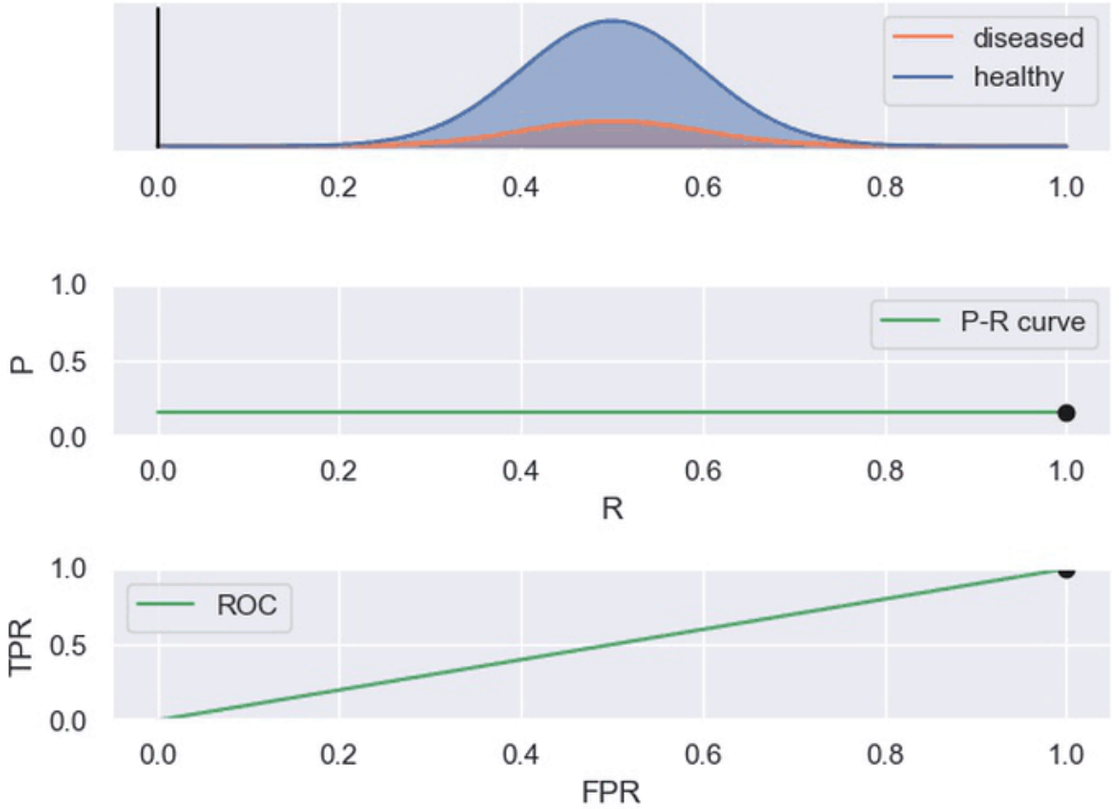




模型性能度量——动态图解释PR曲线和ROC曲线



病人分布和健康人群分布基本不重合，此时P-R曲线和ROC曲线下的面积均很大



病人分布和健康人群分布完全重合，此时两曲线下的面积很小

模型性能可视化——混淆矩阵（Confusion Matrix）

混淆矩阵（又称可能性表格或错误矩阵）：一种呈现算法性能可视化的特定矩阵。

假设有一个对猫(Cat)、狗(Dog)、兔子(Rabbit)进行分类的模型，测试集总有 27 只动物：8只猫， 6条狗， 13只兔子。用混淆矩阵表示测试结果：

混淆矩阵

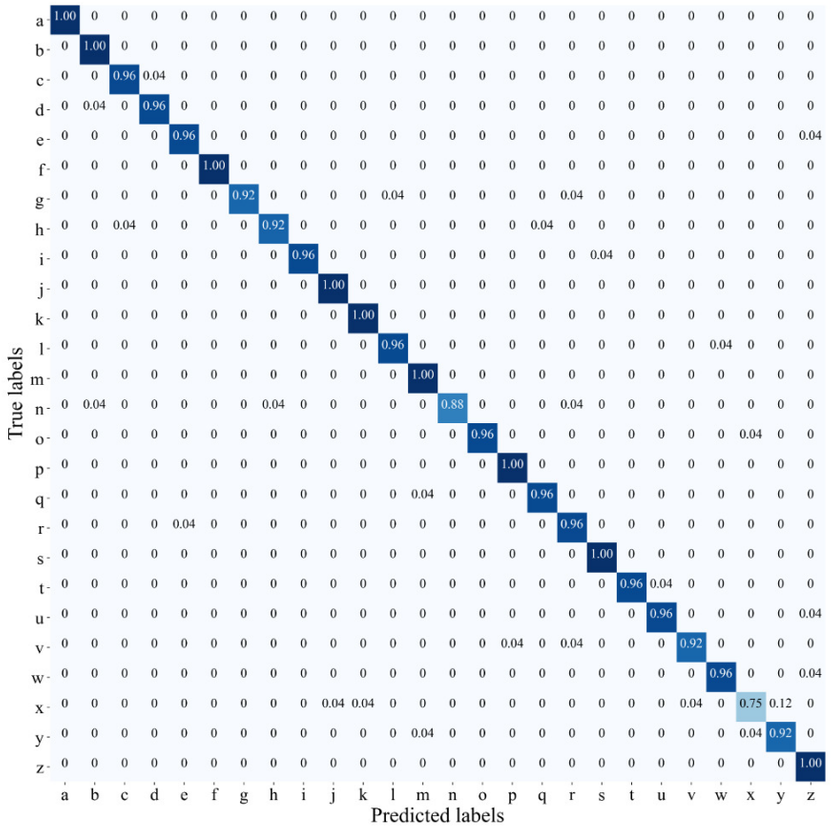
		Predicted class		
		Cat	Dog	Rabbit
Actual class	Cat	5	3	0
	Dog	2	3	1
	Rabbit	0	2	11

分析：

1、正确预测都在对角线上；

2、从混淆矩阵可直观地看出哪里性能不好，因为它们都呈现在对角线左右区域。

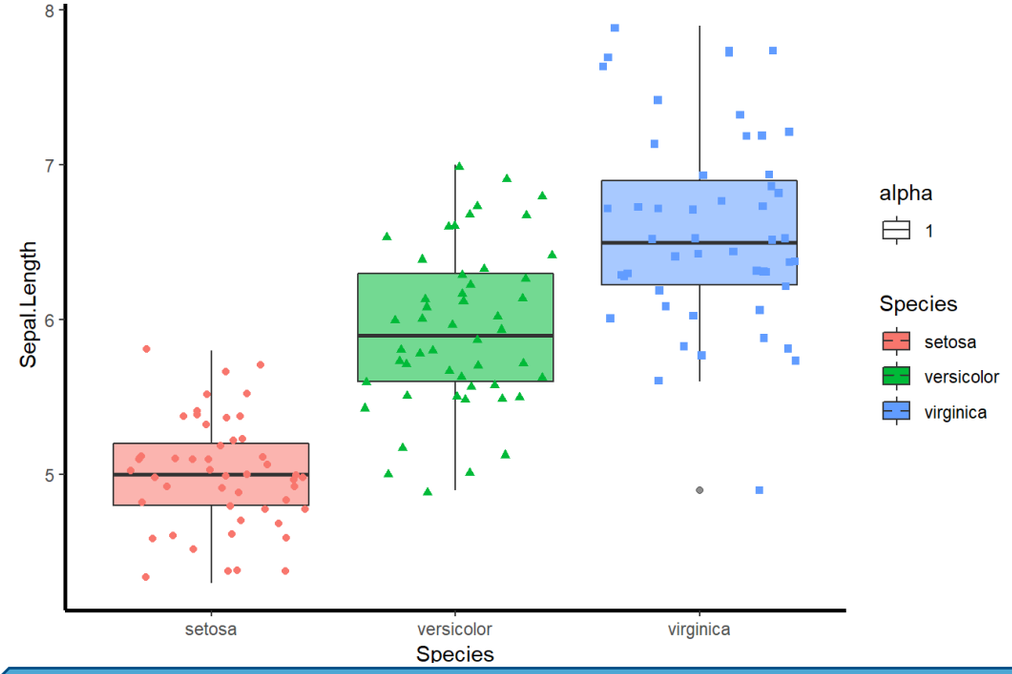
3、左图系统将 8只猫中的3只预测成狗； 6条狗中3条预测错误。说明系统对区分猫狗还有问题，但区分兔子的性能较好。

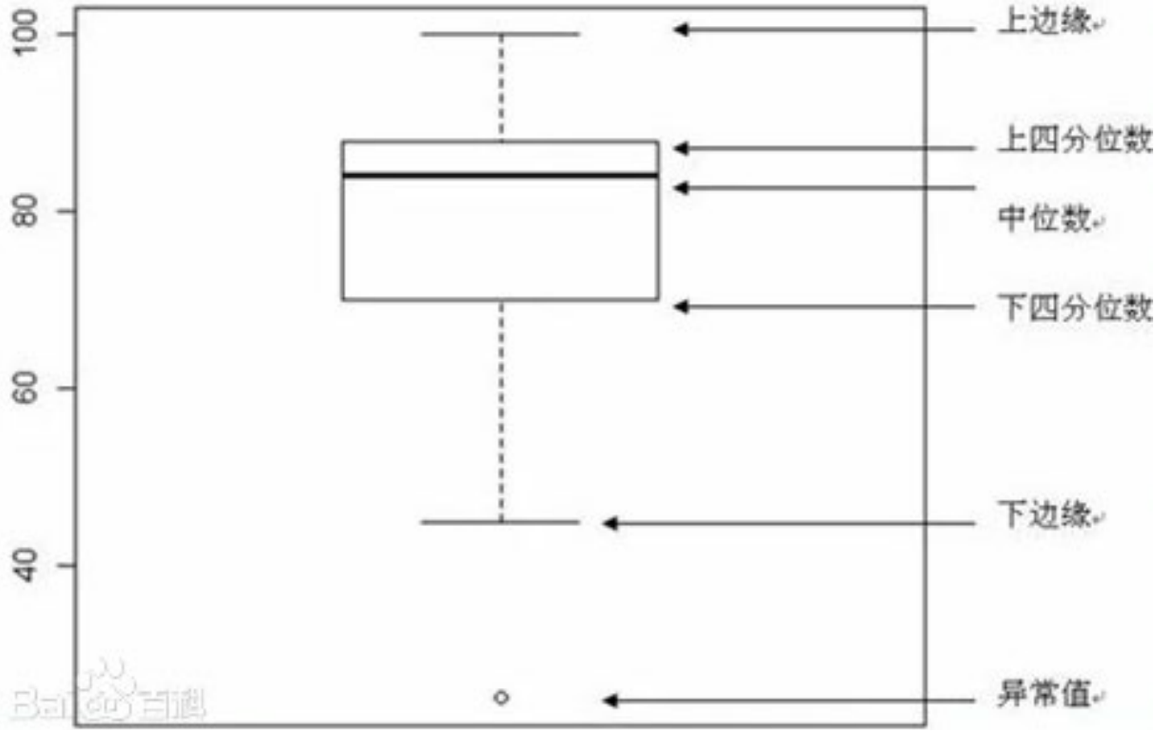


模型性能可视化——箱型图（Box Plot）

箱形图（Box Plot），又称箱须图（Box-whisker Plot）、盒式图或箱线图，用于显示一组数据分散情况的统计图。

由美国统计学家约翰·图基（John Tukey）于1977年发明，显示了一组数据的最大值、最小值、中位数、上下四分位数，及异常值。

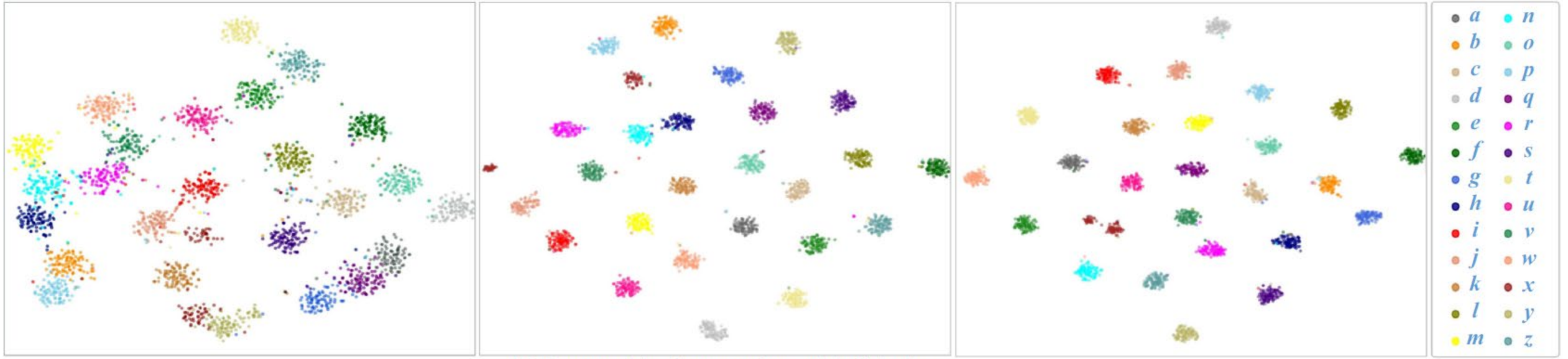




模型性能可视化——t-SNE可视化技术

t-SNE(t-distributed stochastic neighbor embedding)是2008 年提出的一种降维的可视化技术。

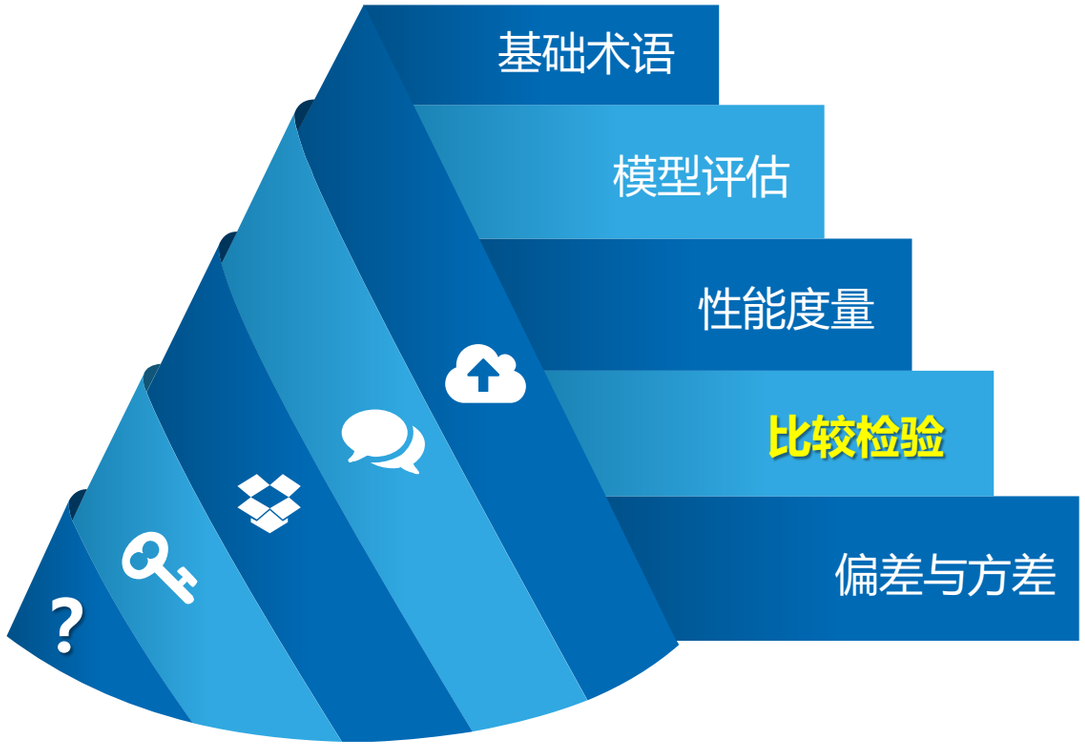
Laurens van der Maaten, Geoffrey Hinton. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008: 2579-2605.



(a) Output of the temporal transformer module

(b) Output of the first transformer block in the channle transformer module

(c) Output of the channle transformer module



- **基础术语与基本概念**
  数据，学习方法，泛化能力，…
- **模型评估方法**
  留出法，交叉验证法，…
- **模型性能度量**
  错误率，精度，P-R曲线，…
- **比较检验**
  二项检验，t检验，交叉验证，…
- **偏差与方差**
  偏差，方差，…

实际性能评估中的问题

■实际评估中，希望比较的是模型泛化性能，但实验中只能获得测试集上的性能。然而，

- 测试性能并不等于泛化性能
- 测试性能随着测试集的变化而变化
- 很多机器学习算法本身有一定的随机性

问题：以错误率作为性能度量，如何根据测试错误率估计出泛化错误率的分布？

假设检验为学习器性能比较提供了重要依据，基于其结果我们可以推断出若在测试集上观察到学习器A比B好，则A的泛化性能是否在统计意义上优于B，以及这个结论的把握（置信）有多大？

比较检验——二项检验

记泛化错误率为 ε ，测试错误率为 $\hat{\varepsilon}$ ，假定测试样本从样本总体分布中独立采样而来，我们可以使用“二项检验”对 $\varepsilon \le \varepsilon_0$ 这样的假设进行检验。

考虑假设 $\varepsilon \le \varepsilon_0$ ，$1-\alpha$ 为结论的置信度，则在 $1-\alpha$ 的概率内所能观测到的最大错误率为：

$$\overline{\varepsilon} = \min \varepsilon \quad \text{s.t.} \quad \sum_{i=\varepsilon \times m+1}^{m} \binom{m}{i} \varepsilon_0^i \left(1-\varepsilon_0\right)^{m-i} < \alpha$$

（西瓜书2.27公式错误）

若测试错误率 $\hat{\varepsilon}$ 小于临界值 $\overline{\varepsilon}$ ，则根据二项检验可得到结论：

在 α 的显著度下，假设不能被拒绝，即能以 $1-\alpha$ 的置信度认为，模型的泛化错误率 ε 不大于 ε_0 。

比较检验——t 检验

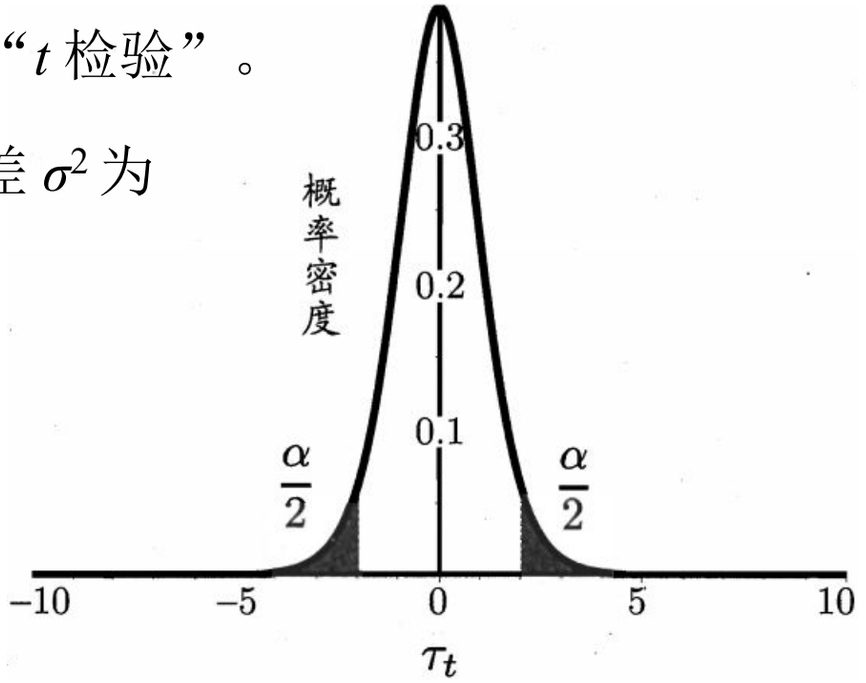
面对多次重复留出法或者交叉验证法进行多次训练/测试时可使用“t 检验”。

假定有 k 个测试错误率，$\hat{\varepsilon}_1, \hat{\varepsilon}_2, ..., \hat{\varepsilon}_k$，则平均测试错误率 μ 和方差 σ^2 为

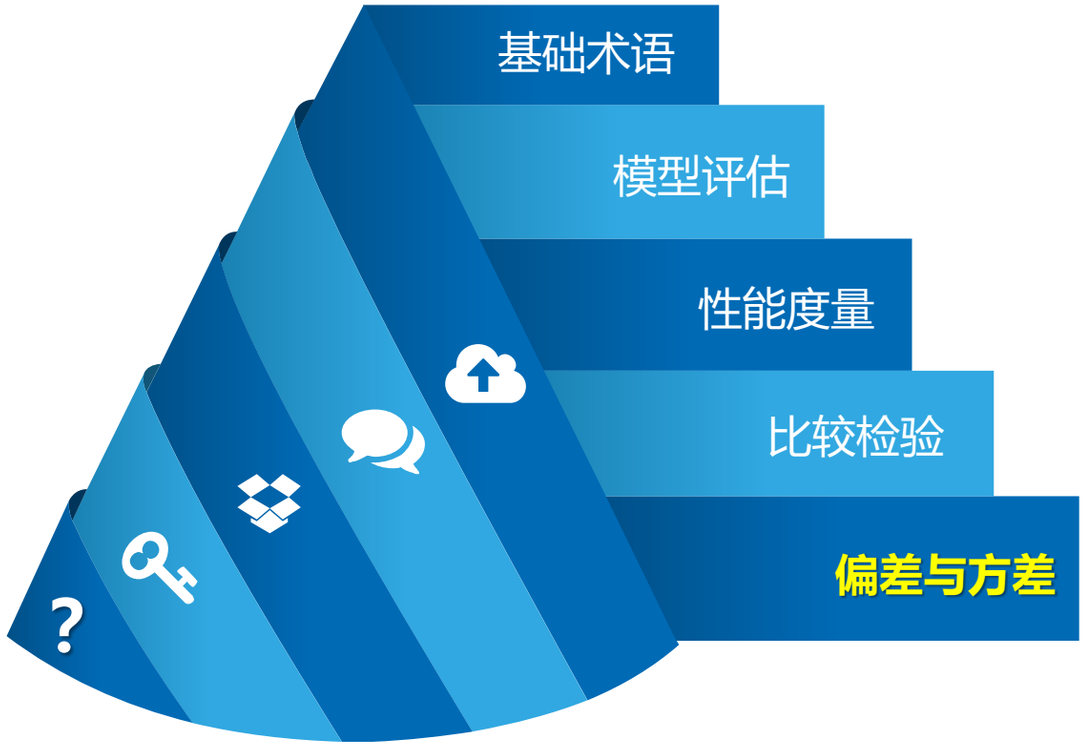
$$\mu = \frac{1}{k}\sum_{i=1}^{k}\hat{\varepsilon}_k \qquad \sigma^2 = \frac{1}{k-1}\sum_{i=1}^{k}\left(\hat{\varepsilon}_k - \mu\right)^2$$

将这 k 个测试错误率看做泛化错误率 ε_0 的独立采样，

则变量 $\tau_t = \dfrac{\sqrt{k}\left(\mu - \varepsilon_0\right)}{\sigma}$ 服从自由度为 k-1的 t 分布。



对假设 $\varepsilon_0 = \mu$ 和显著度 α，可计算出当测试错误率均值为 ε_0 时，在 $1-\alpha$ 概率内能观测到最大错误率。考虑双边假设，若平均错误率 μ 与 ε_0 之差 $\left|\mu - \varepsilon_0\right|$ 位于临界值范围 $[t_{-\alpha/2}, t_{\alpha/2}]$ 内，即可认为泛化错误率为 $\varepsilon_0 = \mu$，置信度为 $1-\alpha$。



- **基础术语与基本概念**
  数据，学习方法，泛化能力，…
- **模型评估方法**
  留出法，交叉验证法，…
- **模型性能度量**
  错误率，精度，P-R曲线，…
- **比较检验**
  二项检验，t检验，交叉验证，…
- **偏差与方差**
  偏差，方差，…

偏差与方差

模型的泛化误差可以分解为三个部分: 偏差(bias)， 方差(variance) 和噪声(noise).

通过实验可以估计学习算法的泛化性能，而“偏差-方差分解”可以用来帮助解释泛化性能。

偏差-方差分解试图对学习算法期望的泛化错误率进行拆解。

对测试样本 $\boldsymbol{x}$，令 y_D 为 $\boldsymbol{x}$ 在数据集中的标记，y 为 $\boldsymbol{x}$ 的真实标记，$f(\boldsymbol{x}; D)$为训练集 D 上学得模型 f 在 $\boldsymbol{x}$ 上的预测输出。

偏差与方差

以回归任务为例，学习算法的期望预期为：$\bar{f}(\boldsymbol{x}) = \mathbb{E}_D[f(\boldsymbol{x}; D)]$

使用样本数目相同的不同训练集产生的方差为：$var(\boldsymbol{x}) = \mathbb{E}_D\left[\left(f\left(\boldsymbol{x}; D\right) - \bar{f}\left(\boldsymbol{x}\right)\right)^2\right]$

噪声为：$\varepsilon^2 = \mathbb{E}_D\left[\left(y_D - y\right)^2\right]$

偏差为：$bias^2(\boldsymbol{x}) = \left(\bar{f}\left(\boldsymbol{x}\right) - y\right)^2$

为便与讨论，假定噪声期望为 0，也即 $\mathbb{E}_D[y_D - y] = 0$, 对泛化误差分解。

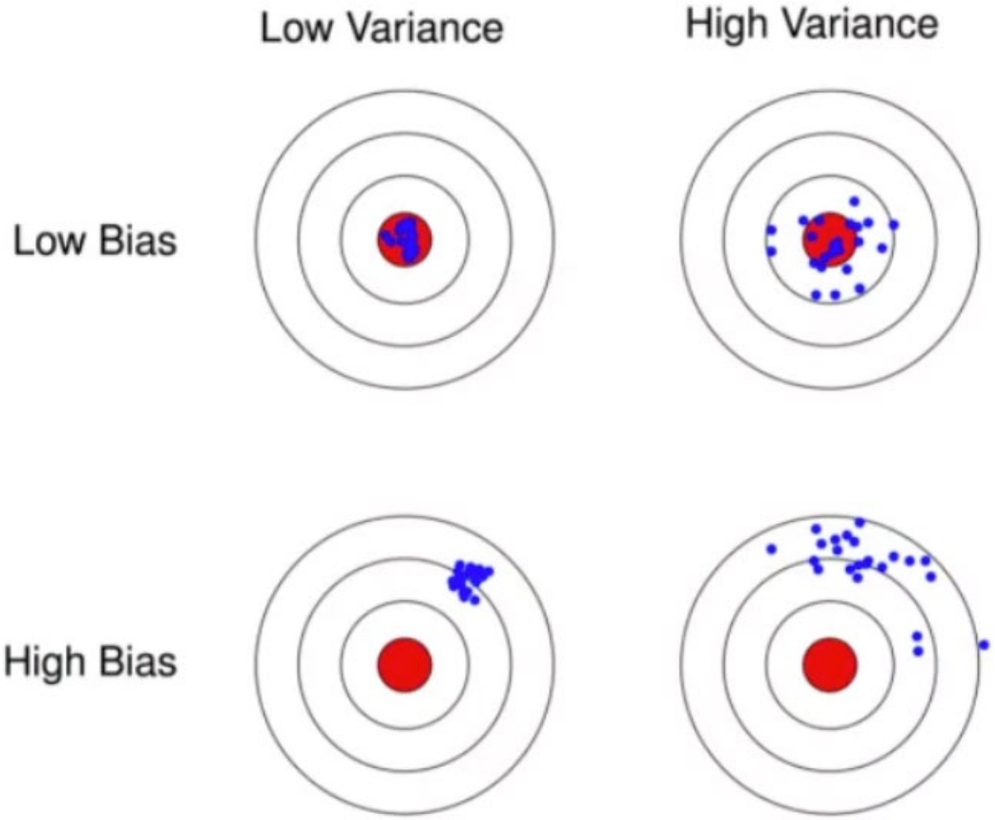
$$E(f; D) = \mathbb{E}_D\left[\left(f\left(\boldsymbol{x}; D\right) - y_D\right)^2\right] = \mathbb{E}_D\left[\left(f\left(\boldsymbol{x}; D\right) - \bar{f}\left(\boldsymbol{x}\right)\right)^2\right] + \left(\bar{f}\left(\boldsymbol{x}\right) - y\right)^2 + \mathbb{E}_D\left[\left(y_D - y\right)^2\right]$$

$$= bias^2\left(\boldsymbol{x}\right) + var\left(\boldsymbol{x}\right) + \varepsilon^2$$

即泛化误差可分解为偏差、方差与噪声之和。

偏差与方差

偏差和方差示意图



偏差与方差的意义

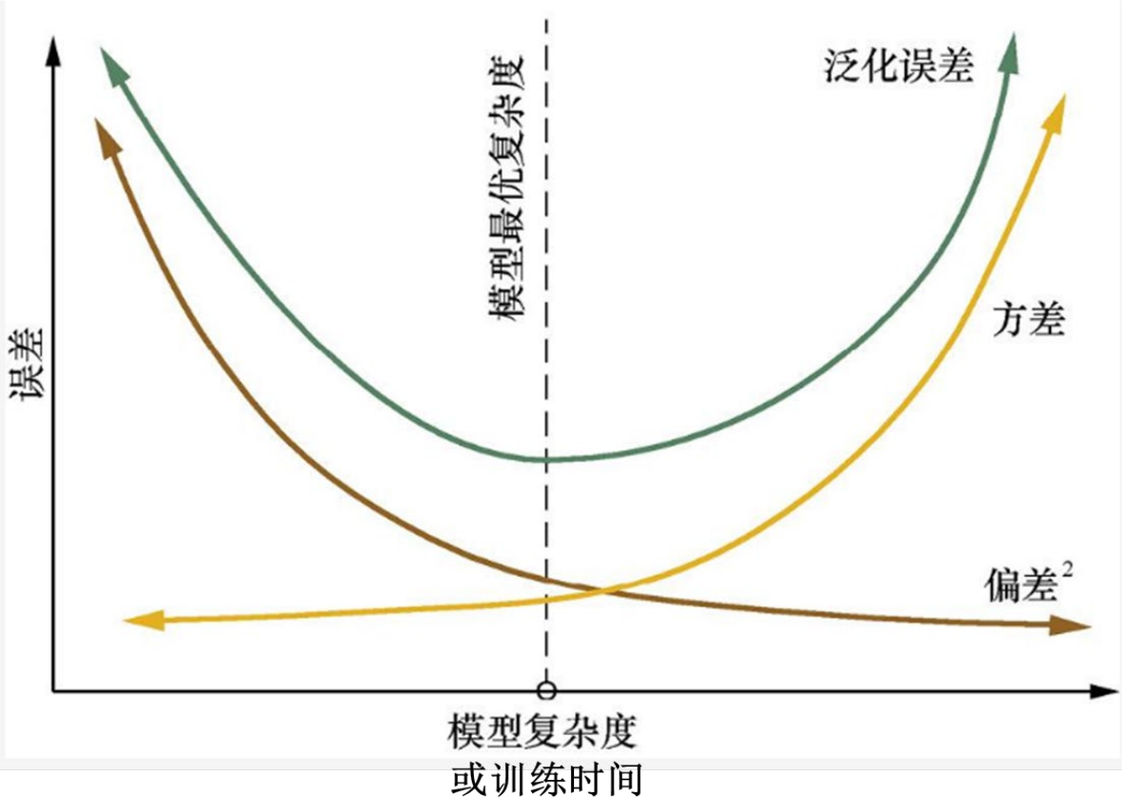
- 偏差度量了学习算法期望预测与真实结果的偏离程度；即刻画了学习算法本身的拟合能力；
- 方差度量了同样大小训练集的变动所导致的学习性能变化；即刻画了数据扰动所造成的影响；
- 噪声表达了在当前任务上任何学习算法所能达到的期望泛化误差的下界；即刻画了学习问题本身的难度。

泛化性能是由学习算法的能力、数据的充分性以及学习任务本身的难度所共同决定的。给定学习任务为了取得好的泛化性能，需要使偏差小（充分拟合数据）而且方差较小（减少数据扰动产生的影响）。

偏差与方差

一般来说，偏差与方差是有冲突的，称为偏差-方差窘境。如右图所示，假如我们能控制算法的训练程度：

- 在训练不足时，学习器拟合能力不强，训练数据的扰动不足以使学习器的拟合能力产生显著变化，此时偏差主导泛化错误率；
- 随着训练程度加深，学习器拟合能力逐渐增强，方差逐渐主导泛化错误率；
- 训练充足后，学习器的拟合能力非常强，训练数据的轻微扰动都会导致学习器的显著变化，若训练数据自身非全局特性被学到则会发生过拟合。



Any Questions?