



第1章 模式识别概论

Outline:



- 模式与模式识别
 - ✓ 概念
 - ✓ 模式识别应用
 - ✓ 实例: fish recognition
- 模式识别的发展
- 模式识别的方法
 - ✓ 基于知识的方法
 - ✓ 基于数据的方法
- 监督模式识别与非监督模式识别
- 模式的描述方法
- 模式识别系统
- 模式识别的若干问题



Concepts of Pattern and Pattern Recognition

➤ 现象:

✓ 高级动物

- 物体：桌子、椅子；人：张三、李四；声音：汽车驶过、玻璃破碎，猫叫、人语；气味：炸带鱼、臭豆腐；

✓ 低等动植物

- 食物、敌害；生存环境



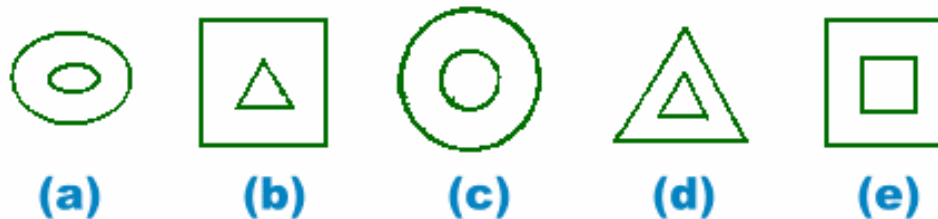
Concepts of Pattern and Pattern Recognition

➤ 猜一猜

✓ A.蛇 B.大树 C.老虎

✓ A.铁锅 B.小勺 C.米饭 D.碟子

✓ A.斑马 B.军马 C.赛马 D.骏马 E.驸马





Concepts of Pattern and Pattern Recognition

➤ 模式 (pattern) : 模式这个概念的内涵很丰富。 “凡是人类能用其感官直接或间接接受的外界信息都称为模式”。

- ✓ 文字、图片、景物;
- ✓ 声音、语言;
- ✓ 心电图、脑电图、地震波等;
- ✓ 社会经济现象、某个系统的状态等。



Concepts of Pattern and Pattern Recognition

➤ 模式识别(Pattern Recognition)

- ✓ 模式分类 (Pattern Classification) ， 机器识别， 计算机识别， 或机器自动识别；
- ✓ 所谓模式识别的问题， 就是用计算的方法根据样本的特征将样本划分到一定的类别中去。

Concepts of Pattern and Pattern Recognition



➤ 名词约定:

- ✓ 样本 (sample) : 所研究对象的一个个体。
- ✓ 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, 处于同一类的样本在我们所关心的某种性质上是不可区分的, 即具有相同的模式。
- ✓ 特征 (features) : 指用于表征样本的观测。
- ✓ 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。



Examples:

➤ 中医给患者看病的步骤

- ✓ 望、闻、问、切（得到患者的一组特征）；
- ✓ 综合分析，抓住主要病症（特征提取与选择）；
- ✓ 运用积累的经验，作出诊断（分类）；

➤ 模式识别术语：

- ✓ 样本：每个患者；
- ✓ 特征：观测到的值就是样本的特征；
- ✓ 类型：病；
- ✓ 分类器：老中医；
- ✓ 模式：指的是单个样本(一个病人)的特征整体；
- ✓ 模式识别：把样本根据其特征归类：

——又称“模式分类” (Pattern classification)



Outline:

- **模式与模式识别**
 - ✓ **概念**
 - ✓ 模式识别应用
 - ✓ 实例: fish recognition
- 模式识别的发展
- 模式识别的方法
 - ✓ 基于知识的方法
 - ✓ 基于数据的方法
- 监督模式识别与非监督模式识别
- 模式的描述方法
- 模式识别系统
- 模式识别的若干问题



Applications:

➤ 生物学

✓ 自动细胞学、染色体特性研究、遗传研究

➤ 天文学

✓ 天文望远镜图像分析、自动光谱学

➤ 经济学

✓ 股票交易预测、企业行为分析

➤ 医学

✓ 心电图分析、脑电图分析、医学图像分析



Applications:

➤ 工程

✓ 产品缺陷检测、特征识别、字符识别、语音识别、自动驾驶系统、污染分析

➤ 军事

✓ 航空摄像分析、雷达和声纳信号检测和分类、自动目标识别

➤ 安全

✓ 指纹识别、指静脉识别、人脸识别、监视和报警系统

指纹识别



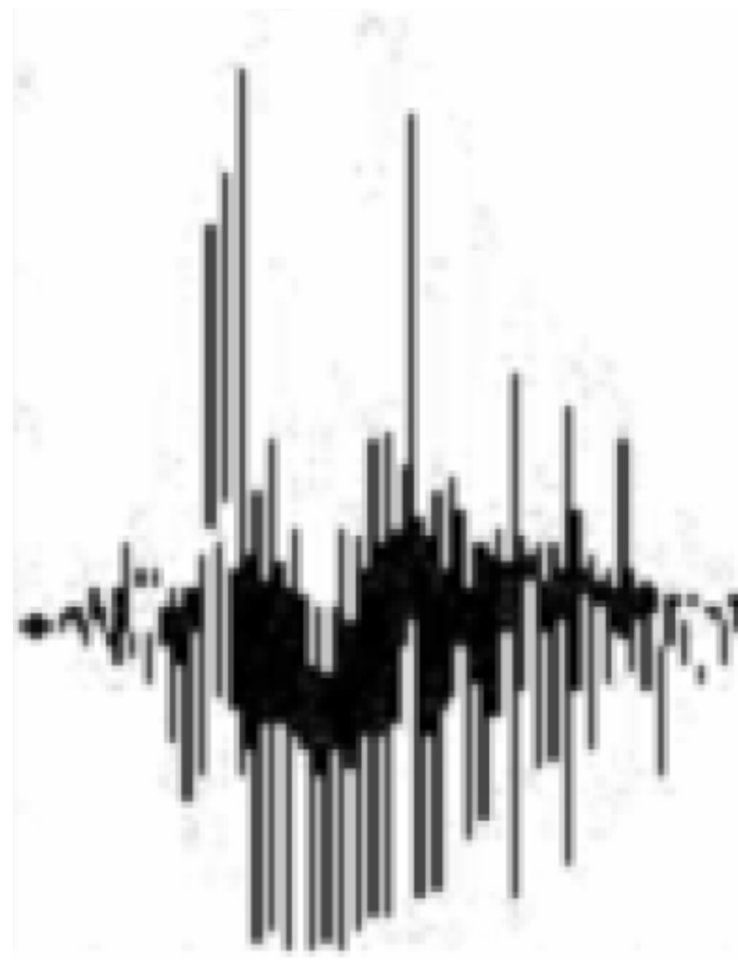
- 为什么研究指纹？
- 准确性
- 速度
- 存储量
- 价格...





语音信号处理与识别

- 语音识别
- 说话人识别
- 语种识别
- 口音识别





人脸图像处理与识别

➤ 人脸检测和定位

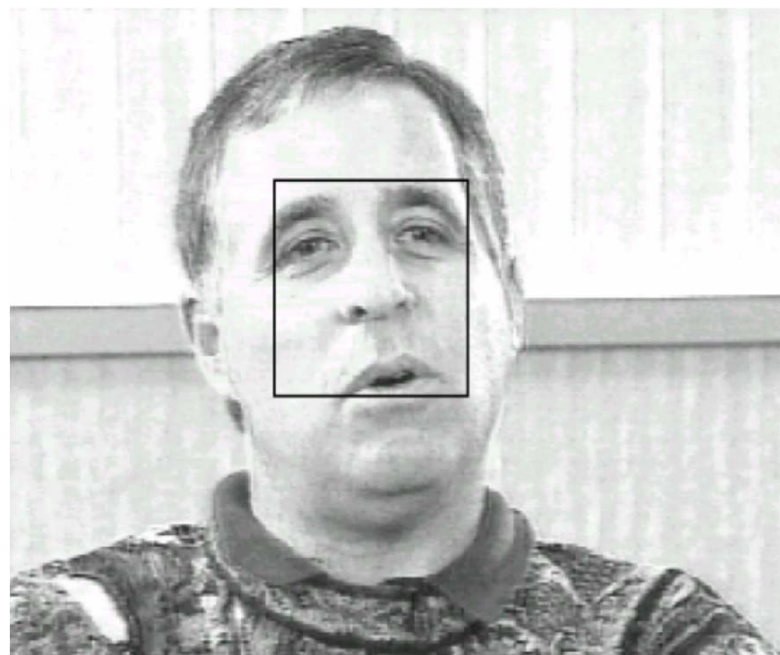
➤ 人脸识别

➤ 应用：

图像压缩

视频监控

基于内容的图像检索





文字处理与识别

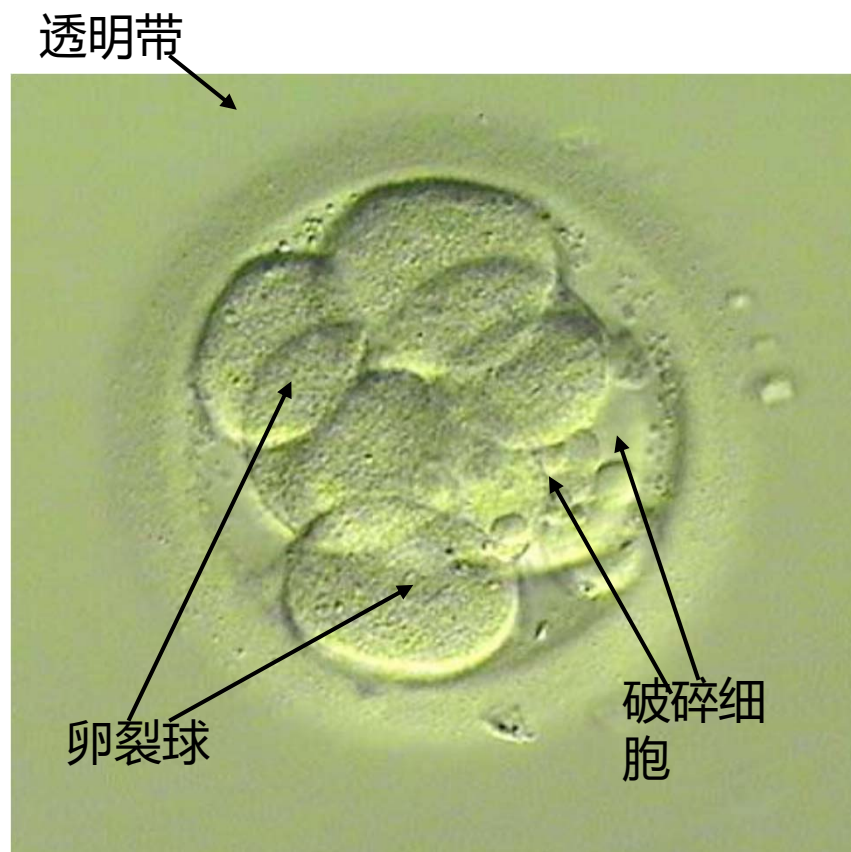
➤ 字体识别

➤ 语种识别

Markov chain Monte Carlo (MCMC)
for realistic statistical modelling. U
complexity and structure in many a
the development of specific methodo



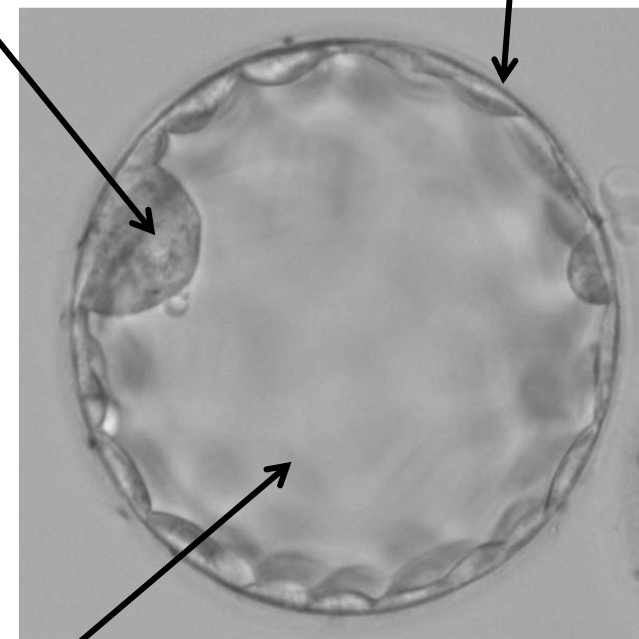
医学图像分析



内细胞团

滋养外胚层

囊胚腔



其他问题

- 掌纹识别
- 签字识别
- 虹膜识别
- 步态识别
- 人耳识别
- ...





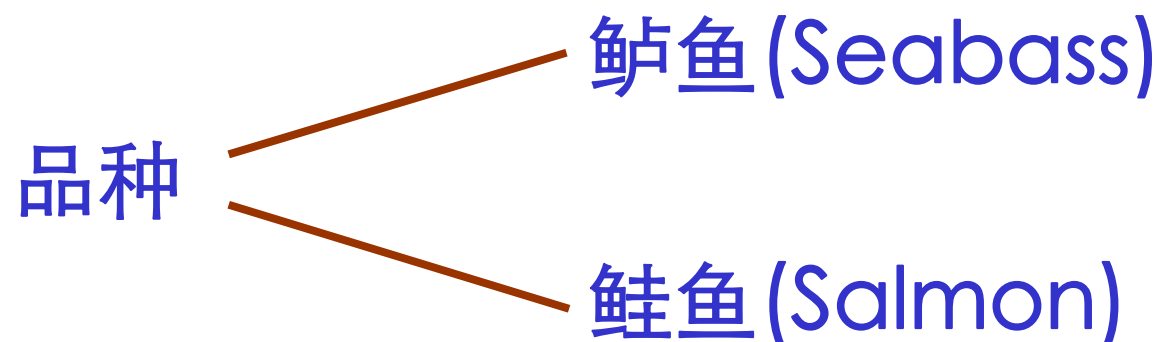
Outline:

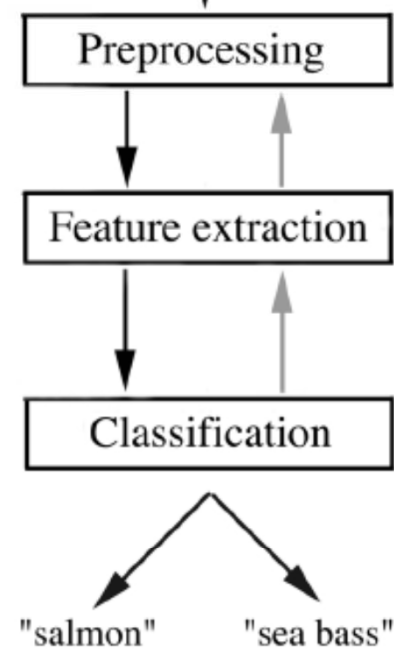
- **模式与模式识别**
 - ✓ 概念
 - ✓ 模式识别应用
 - ✓ 实例: fish recognition
- 模式识别的发展
- 模式识别的方法
 - ✓ 基于知识的方法
 - ✓ 基于数据的方法
- 监督模式识别与非监督模式识别
- 模式的描述方法
- 模式识别系统
- 模式识别的若干问题



Fish Recognition:

- 在传送带上用光学传感器件对鱼按品种分类





识别过程（1）

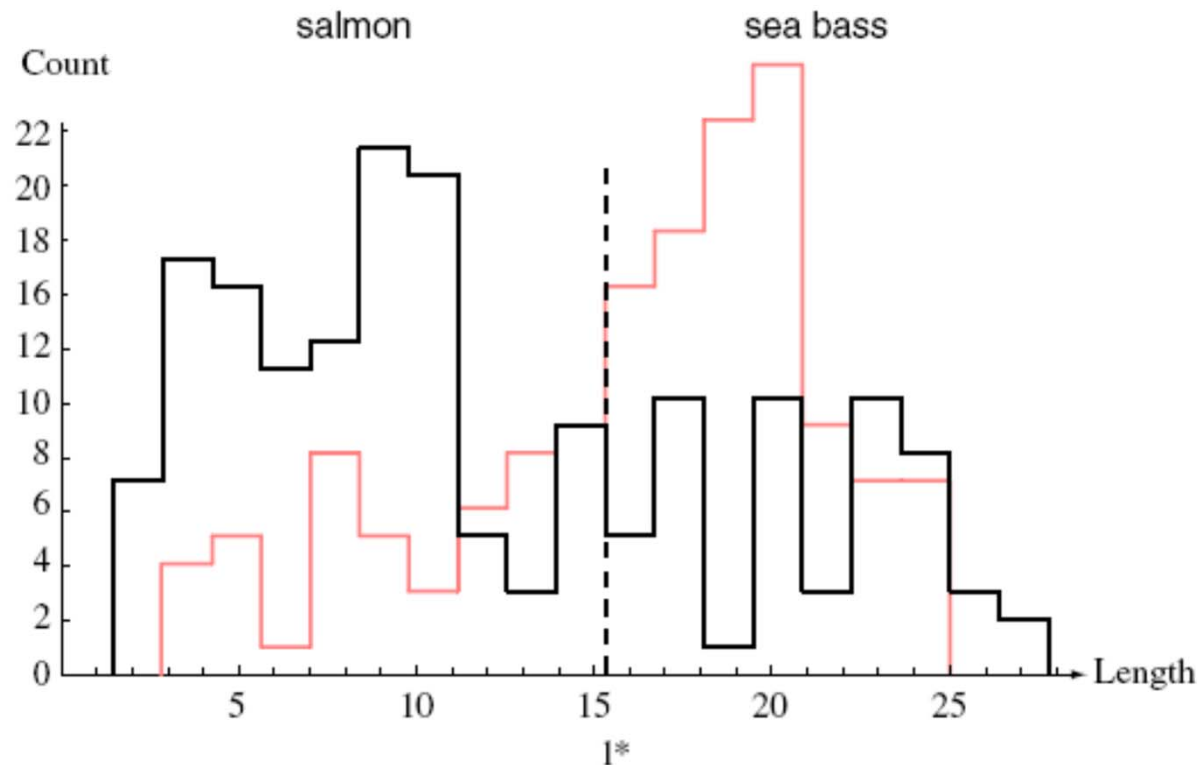


- 数据获取：架设一个摄像机，采集一些样本图像，获取样本数据
- 预处理：去噪声，用一个分割操作把鱼和鱼之间以及鱼和背景之间分开

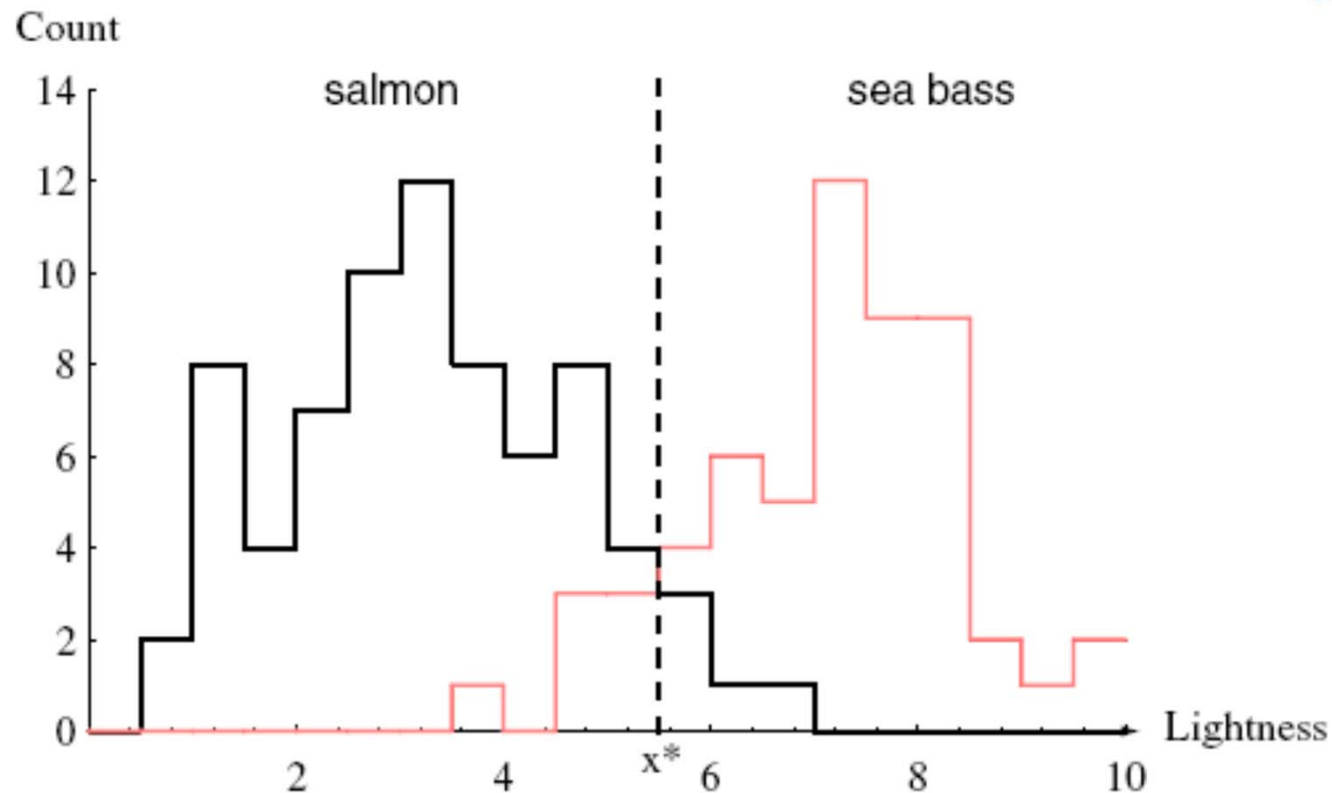


识别过程（2）

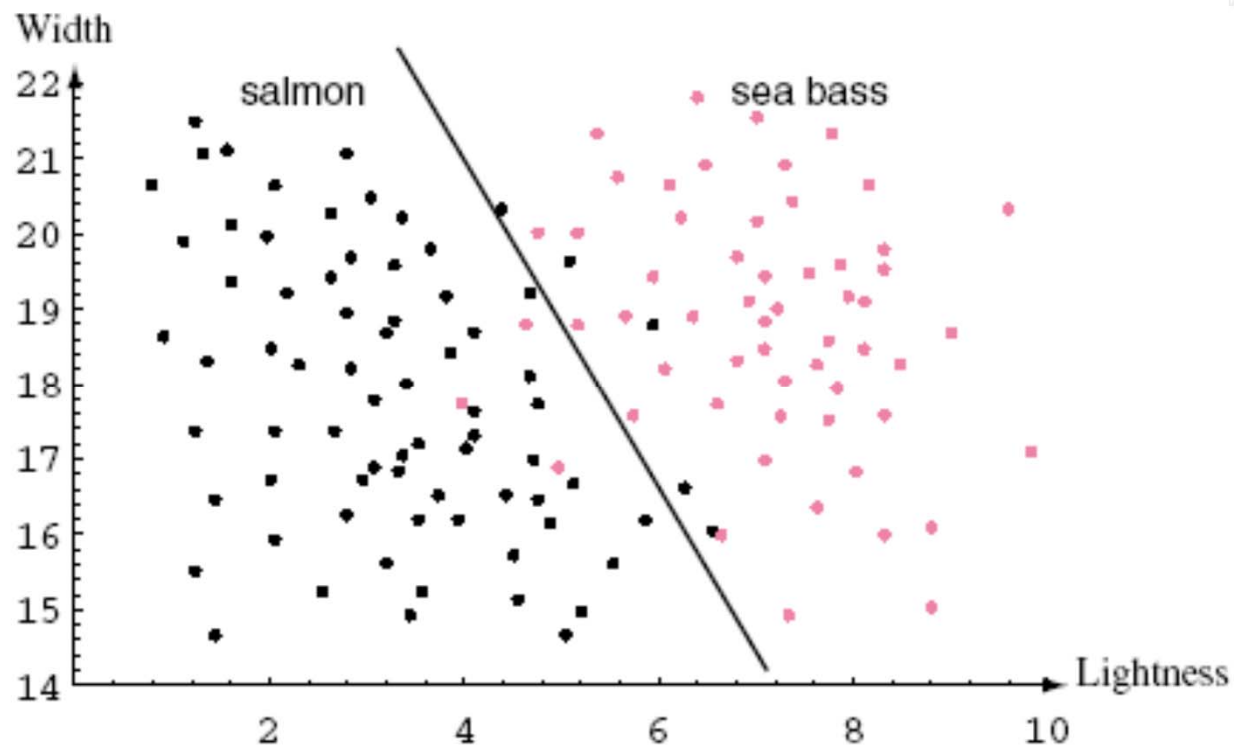
- 特征提取和选择：对单个鱼的信息进行特征选择，从而通过测量某些特征来减少信息量
 - ✓ 长度、亮度、宽度
 - ✓ 鱼翅的数量和形状
 - ✓ 嘴的位置，等等 …
- 分类决策：把特征送入决策分类器



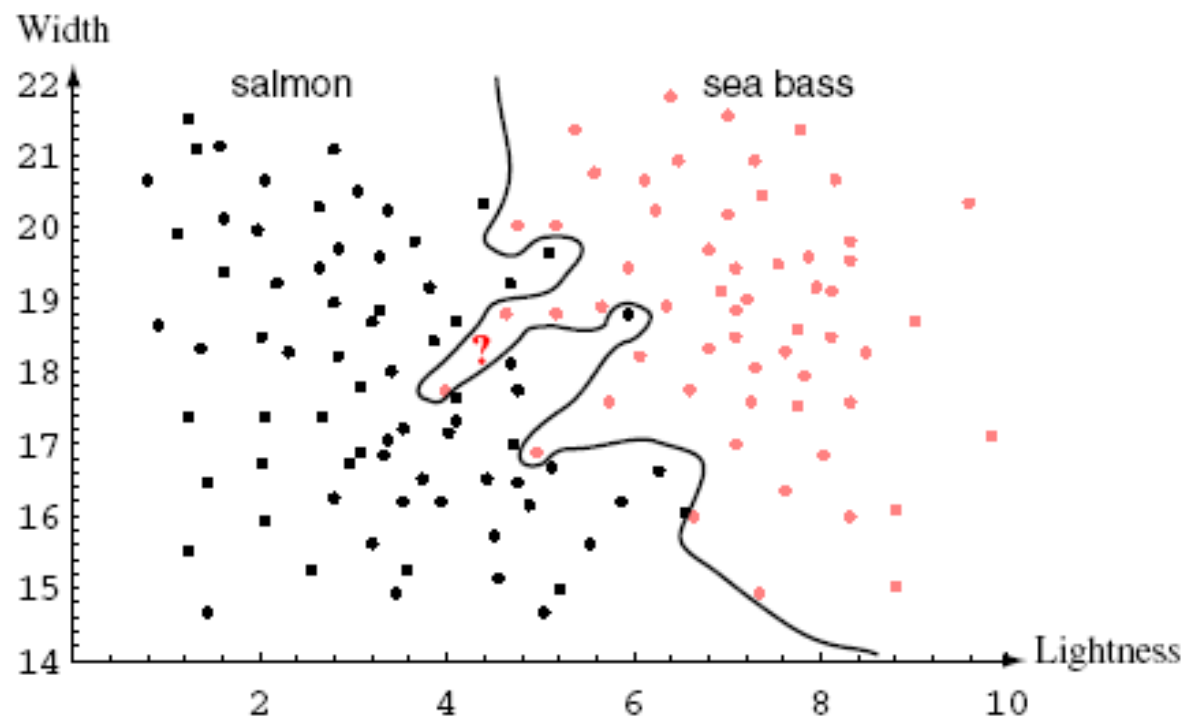
Histograms for the length feature for the two categories. No single threshold value l^* (decision boundary) will serve to unambiguously discriminate between the two categories; using length alone, we will have some errors. The value l^* marked will lead to the smallest number of errors, on average.



Histograms for the lightness feature for the two categories. No single threshold value x^* (decision boundary) will serve to unambiguously discriminate between the two categories; using lightness alone, we will have some errors. The value x^* marked will lead to the smallest number of errors, on average.



The two features of lightness and width for sea bass and salmon. The dark line might serve as a decision boundary of our classifier. Overall classification error on the data shown is lower than if we use only one feature, but there will still be some errors.



Overly complex models for the fish will lead to decision boundaries that are complicated. While such a decision may lead to perfect classification of our training samples, it would lead to poor performance on future patterns. The novel test point marked ? is evidently most likely a salmon, whereas the complex decision boundary shown leads it to be misclassified as a sea bass.



In summary:

- ◆ 人类在观察事物与作出判断时，常常把所见到的具体事物与脑子里对某个事物的“概念”联系起来，从而按这些概念对它们实行分类。
- ◆ 人们为了掌握客观事物，按事物相似的程度组成类别。模式识别的作用和目的就在于面对某一具体事物时将其正确地归入某一类别。
- ◆ 通常，把通过对具体的个别事物进行观测所得到的具有时间和空间分布的信息称为模式，而把所属的类别或同一类中模式的总体称为模式类（简称为类）。



In summary:

- ◆ 模式类与模式在集合论中是子集与元素之间的关系。当用一定的度量来衡量两个样本，而找不出它们之间的差别时，它们在这种度量条件下属于同一个等价类。这就是说它们属于同一子集，或一个模式类。而不同的模式类之间应该是可区分的，它们之间应有明确的界线。
- ◆ 但对实际样本来说，有时又往往不能对它们进行确切的划分，即在所使用的度量关系中，分属不同的类别的样本却表现出相同的属性，因而无法确凿无误地对它们进行区分。



Outline:

- **模式与模式识别**
 - ✓ 概念
 - ✓ 模式识别应用
 - ✓ 实例: fish recognition
- 模式识别的发展
- 模式识别的方法
 - ✓ 基于知识的方法
 - ✓ 基于数据的方法
- 监督模式识别与非监督模式识别
- 模式的描述方法
- 模式识别系统
- 模式识别的若干问题



模式识别的发展简史

- 1929年 G. Tauschek发明阅读机，能够阅读0-9的数字。
- 30年代 Fisher提出统计分类理论，奠定了统计模式识别的基础。
- 50年代 Noam Chomsky 提出形式语言理论——傅京荪提出句法/结构模式识别。
- 60年代 L. A. Zadeh提出了模糊集理论，模糊模式识别方法得以发展和应用。

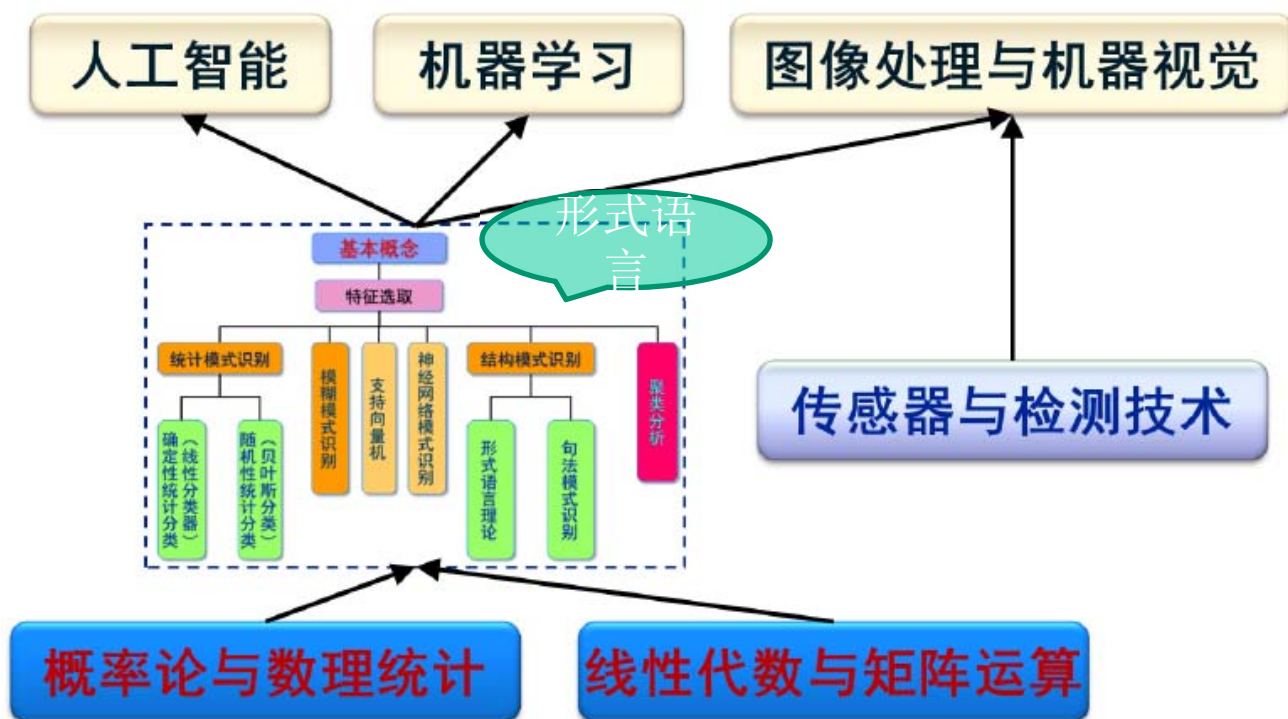


模式识别的发展简史

- 80年代 以Hopfield网、BP网为代表的神经网络模型导致人工神经元网络复活，并在模式识别得到较广泛的应用。
- 90年代 小样本学习理论，支持向量机也受到了很大的重视。



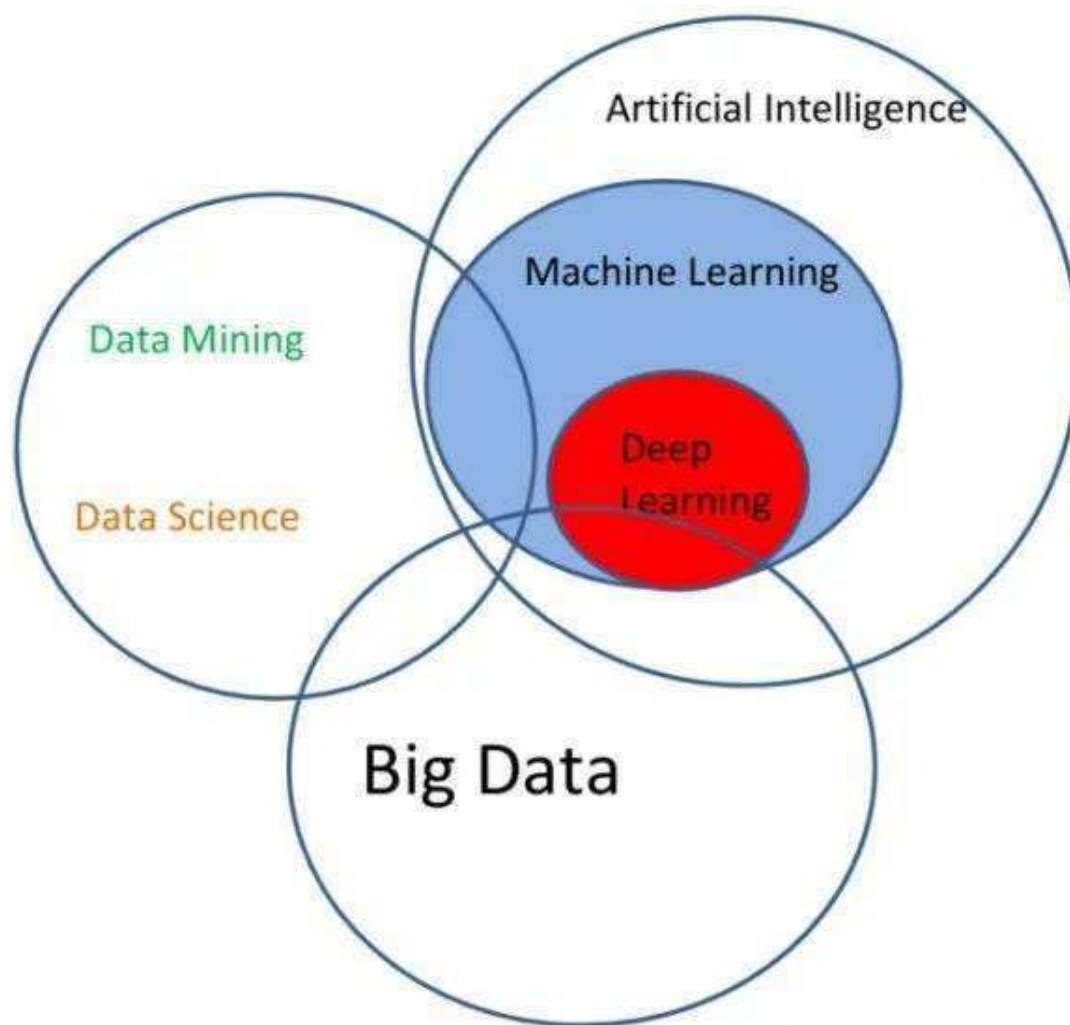
模式识别同其它学科的关系



模式识别是人工智能学科的组成部分，代表了人工智能中的认知能力，与学习能力、推理能力一起构成人工智能的完整范畴。


机器学习是模式识别的扩展，它不仅包括了机器的模式识别能力，还包括了机器发现知识，理解知识，自我进化的能力。

图像处理与机器视觉是目前计算机科学家和自动控制界都在研究的重点领域，在“传感器与检测技术”和“模式识别”技术的共同支持下，信息最丰富、获取最方便的图像数据，可以成为计算机和其他智能设备的主要信息来源。



机器学习、统计分析、数据挖掘、神经网络、人工智能、模式识别之间的关系

2017全球最具价值品牌100强

17	2016	Logo	Name	Country	2017	2016	2017	20
↑	2		Google		109,470	88,173	AAA+	AA
↓	1		Apple		107,141	145,918	AAA	AA
→	3		Amazon.com		106,396	69,642	AAA-	AA
↑	6		AT&T		87,016	59,904	AAA	AA
↓	4		Microsoft		76,265	67,258	AAA	AA
↑	7		Samsung Group		66,219	58,619	AAA-	AA
↓	5		Verizon		65,875	63,116	AAA-	AA
→	8		Walmart		62,211	53,657	AA+	A
↑	17		Facebook		61,998	34,002	AAA	AA
↑	13		ICBC		47,832	36,334	AAA	AA
↓	9		China Mobile		46,734	49,810	AAA	A

国务院关于印发 新一代人工智能发展规划的通知

国发〔2017〕35号

各省、自治区、直辖市人民政府，国务院各部委、各直属机构：

现将《新一代人工智能发展规划》印发给你们，请认真贯彻执行。

国务院

2017年7月8日

（此件公开发布）

新一代人工智能发展规划

人工智能的迅速发展将深刻改变人类社会生活、改变世界。为抢抓人工智能发展的重大战略机遇，构筑我国人工智能发展的先发优势，加快建设创新型国家和世界科技强国，按照党中央、国务院部署要求，制定本规划。

一、战略态势

人工智能发展进入新阶段。经过60多年的演进，特别是在移动互联网、大数据、超级计算、传感网、脑科学等新理论新技术以及经济社会发展强烈需求的共同驱动下，人工智能加速发展，呈现出深度学习、跨界融合、人机协同、群智开放、自主操控等新特征。大数据驱





Outline:

- **模式与模式识别**
 - ✓ 概念
 - ✓ 模式识别应用
 - ✓ 实例: fish recognition
- **模式识别的发展**
- **模式识别的方法**
 - ✓ 基于知识的方法
 - ✓ 基于数据的方法
- **监督模式识别与非监督模式识别**
- **模式的描述方法**
- **模式识别系统**
- **模式识别的若干问题**



模式识别的方法

➤ 基于知识的方法

- ✓ 指以专家系统为代表的方法，一般归在人工智能的范畴中，其基本思想是根据人们已知的（从专家那里收集整理的）关于研究对象的知识，整理出若干描述特征与类别间关系的准则，建立一定的计算机推理系统，对未知样本通过这些知识推理决策其类别。
- ✓ 模板匹配
- ✓ 句法模式识别

模式识别的方法



➤ 基于知识的方法

✓ 模板匹配

- 模板匹配是最早出现的模式识别方法，甚至在计算机出现之前就已经开始使用了。它对每个类别建立一个或多个标准模板，分类决策时将待识别的样本与每个类别的模板进行比对，根据与模板的匹配程度将样本划分到最相似的类别中；
- 严格来讲，模板匹配不能算是模式识别的范畴，在建立模板的时候需要人工的干预，但由于其直接、简单，在类别特征稳定、明显，类间差距大的时候仍然可以使用，只是它的适应能力比较差。

模式识别的方法



➤ 基于知识的方法

✓ 句法模式识别（结构模式识别）

- 在许多情况下，对于较复杂的对象仅用一些数值特征已不能较充分地进行描述，这时可采用句法识别技术。该技术将对象分解为若干个基本单元（基元），用这些基元以及它们的结构关系来描述对象，该关系可以用字符串或图来表示；然后运用形式语言推理进行句法分析、根据其是否符合某一类的文法而决定其类别。



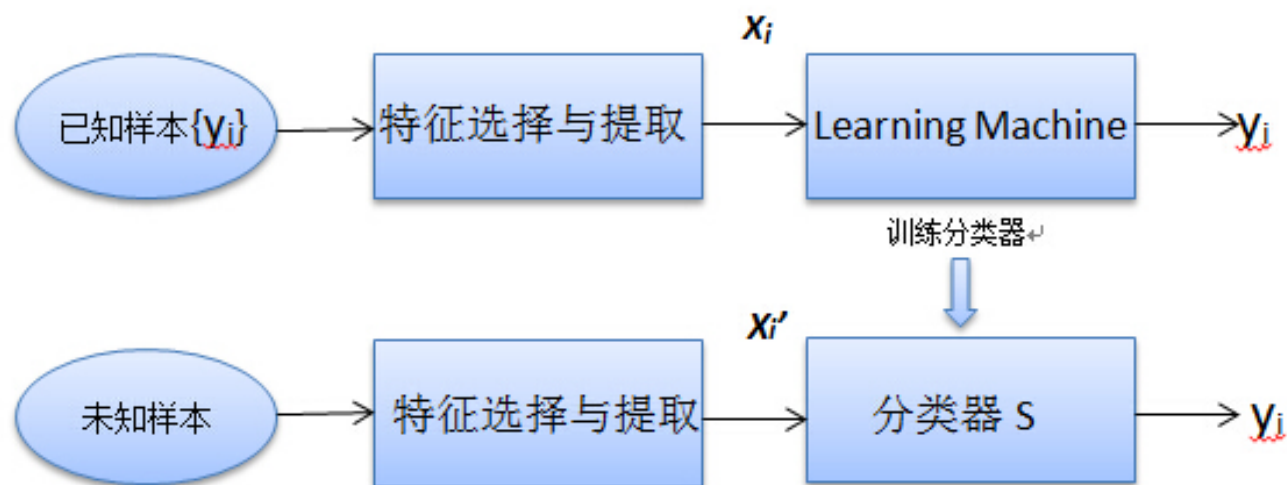
Outline:

- **模式与模式识别**
 - ✓ 概念
 - ✓ 模式识别应用
 - ✓ 实例: fish recognition
- **模式识别的发展**
- **模式识别的方法**
 - ✓ 基于知识的方法
 - ✓ 基于数据的方法
- 监督模式识别与非监督模式识别
- 模式的描述方法
- 模式识别系统
- 模式识别的若干问题



模式识别的方法

➤ 基于数据的方法



任务描述为：在类别标号 y 与特征向量 x 存在一定的未知依赖关系、但已知的信息只有一组训练数据对 $\{(x, y)\}$ 的情况下，求解定义在 x 上的某一函数 $y' = f(x)$ ，对未知样本的类别进行预测。这一函数叫做分类器（classifier）。这种根据样本建立分类器的过程也称作学习过程或训练过程。

模式识别的方法



➤ 基于数据的方法

✓ 人工神经网络

- 由大量简单的基本单元——神经元相互联接而构成的非线性动态系统，每个神经元结构和功能比较简单，而由其组成的系统却可以非常复杂，具有生物神经网络的某些特征，在自学习、自组织、联想及容错方面具有较强的能力，能用于联想、识别和决策。

✓ 深度学习

- 典型的深度学习模型就是很深层的神经网络。
- 理论上讲，参数越多的模型复杂度越高、“容量” (Capacity) 越大，这意味着它能完成更复杂的学习任务。



模式识别的方法

➤ 基于数据的方法

✓ 模糊模式识别

- 该类技术运用模糊数学的理论和解决方法解决模式识别问题，因此适用于分类识别对象本身或要求的识别结果具有模糊性的场合。该类方法的有效性主要在于对象类的隶属函数是否良好。

模式识别的方法



➤ 基于数据的方法

✓ 聚类分析

- 当模式识别中分类及学习采用无监督学习时，分类器主动对样本集进行类别划分的过程称为聚类分析。聚类分析有单独的算法，可以看做PR的一类特殊方法。

✓ 支持向量机(SVM)

- 目前用于模式分类的较新的方法，它在样本集较小的情况下，利用空间映射和最优化理论来确定最优或次优的分类决策面（小样本学习理论）。

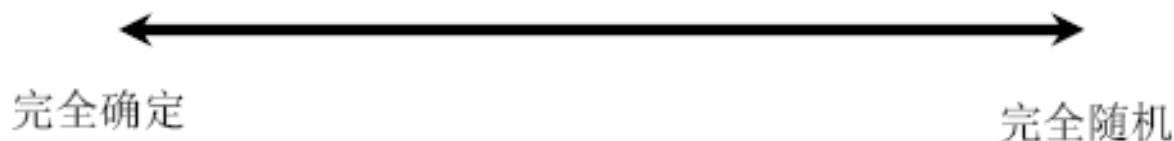
模式识别的方法



➤ 基于数据的方法

- ✓ 适用于已知对象的某些特征同所感兴趣的类别性质有关系，但无法确切描述这种情况。
- ✓ 原因：
 - 相关机理研究比较初步；
 - 问题本身的不确定性，样本间的异质性和观测数据的不确定性等；

模式识别研究的范畴





Outline:

➤ 模式与模式识别

- ✓ 概念
- ✓ 模式识别应用
- ✓ 实例: fish recognition

➤ 模式识别的发展

➤ 模式识别的方法

- ✓ 基于知识的方法
- ✓ 基于数据的方法

➤ 监督模式识别与非监督模式识别

➤ 模式的描述方法

➤ 模式识别系统

➤ 模式识别的若干问题

监督模式识别与非监督模式识别



- 监督模式识别 (supervised pattern recognition)
 - ✓ 要划分的类别已知;
 - ✓ 有一定数量的类别已知的训练样本, 作为分类器学习的“导师”;
- 非监督模式识别 (unsupervised pattern recognition)
 - ✓ 事先不知道要划分成多少类别;
 - ✓ 没有类别已知的样本用作训练;
 - ✓ 目的: 依据相似性, 将样本聚为几类;
 - ✓ 聚类结果受相似度测度的影响;



Outline:

- 模式与模式识别
 - ✓ 概念
 - ✓ 模式识别应用
 - ✓ 实例: fish recognition
- 模式识别的发展
- 模式识别的方法
 - ✓ 基于知识的方法
 - ✓ 基于数据的方法
- 监督模式识别与非监督模式识别
- 模式的描述方法
- 模式识别系统
- 模式识别的若干问题

模式的描述方法



联想到人们认识事物，都是从不同事物所具有的不同属性为出发点，因此用来决策事物类别的特点、属性就称之为物体所具有的特征。在模式识别技术中，模式就是用它们所具有的特征描述的。对一种模式与它们的样本来说，将描述它们的所有特征用一特征集表示：

$$O = \{f_1, f_2, \dots, f_n\}$$

其中 O 表示模式或样本的名称， f_n 则是它们所具有的特征。特征包括定性与定量两种描述。



定性是指特征的有与无，然而一些不同类别的事物往往具有相同的特征种类，或者用同样的特征度量去检测，但它们在這些特征的取值上有差别，在这种情况下特征值的取值范围成为辨别事物的重要依据。例如癌细胞与正常细胞都用同样的观测手段去检测，而依据所得特征值分布范围将它们区分开来。在这种情况下，模式的特征集表示，又可写成处于同一个特征空间的特征向量表示。待识别的不同类模式都在同一特征空间中考察，不同类物体由于性质上的不同，它们在各特征取值范围上有所不同，因而在特征空间的不同区域中出现。

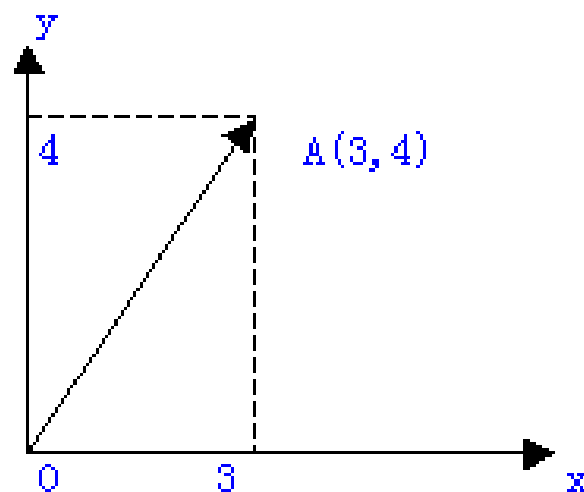
定量描述



定量的描述就是用各种尺度对事物进行度量。例如对水果进行分类，就需要对它的各种属性进行度量，水果的重量、大小、颜色、香味乃至味道等。由于对事物的度量是多方面的，因此要用合适的数据结构将它们记录下来，以便在同一种度量之间进行比较。常用的方法是将这些度量值排成序，譬如用水果的重量，近似球体直径。这两个指标按规定的先后排起来，如一只苹果重0.3斤，直径10厘米，则可表示成(0.3,1.0)。因此如看到一个数据为(0.35,12)则可解释成重量为0.35斤，直径12厘米。这种表示方法就称为向量表示法，该向量有两个分量，每个分量有自己特定的含义。



例： 如一个二维向量A表示成(x,y)，则(3,4)就是指x=3,y=4。如果用图像来表示：



用式子表示，可写成： $A = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$ 或 $A = (3,4)$



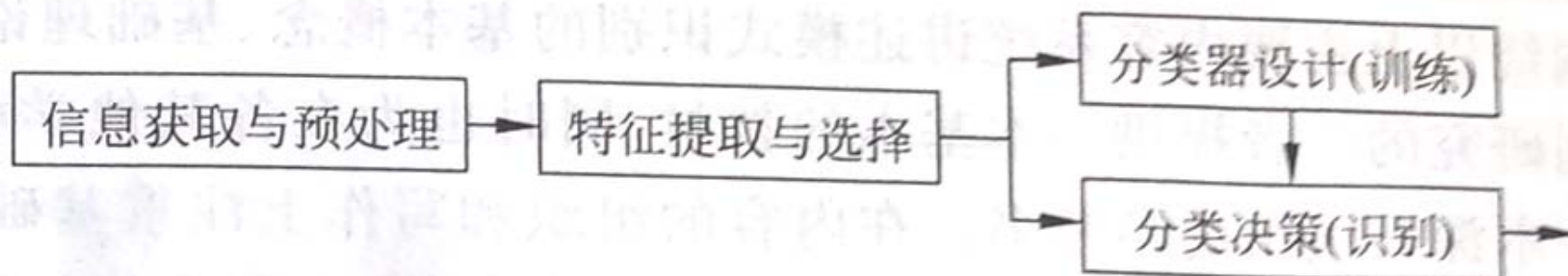
Outline:

- 模式与模式识别
 - ✓ 概念
 - ✓ 模式识别应用
 - ✓ 实例: fish recognition
- 模式识别的发展
- 模式识别的方法
 - ✓ 基于知识的方法
 - ✓ 基于数据的方法
- 监督模式识别与非监督模式识别
- 模式的描述方法
- 模式识别系统
- 模式识别的若干问题

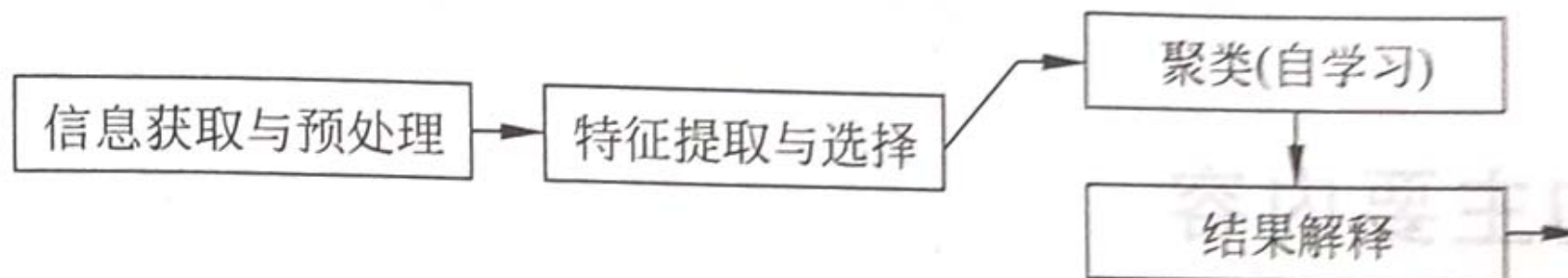


模式识别系统的典型构成

- 有已知样本情况：监督模式识别



- 无已知样本情况：非监督模式识别



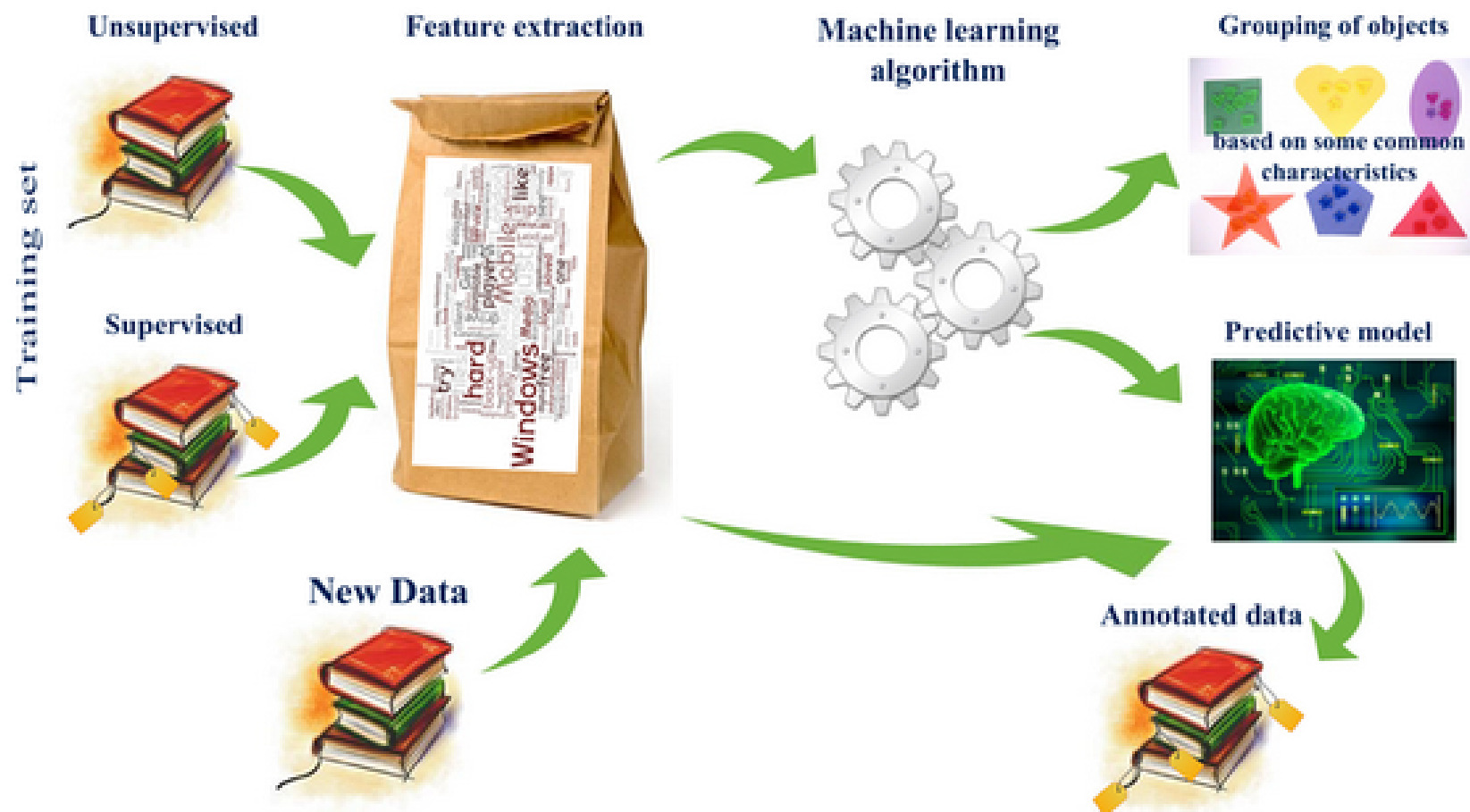


模式识别系统

一个典型的模式识别系统一般由数据获取，预处理，特征提取选择、分类器设计及分类决策五部分组成。分类器设计在训练过程中完成，利用样本进行训练，确定分类器的具体参数。而分类决策在识别过程中起作用，对待识别的样本进行分类决策。

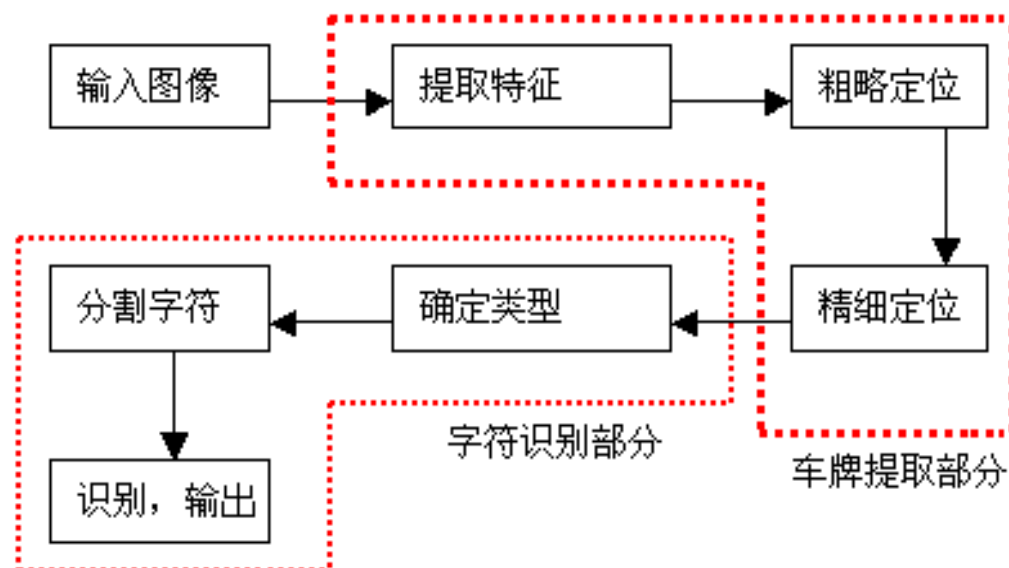


Machine learning workflow





例：车牌识别系统





第一章 模式识别概论

下一步 



模式识别系统

➤ 信息获取

✓ 卡口的照片或监控摄像头的视频数据。

➤ 预处理

✓ 预处理(增强、分割等)主要是指去除所获取信息中的噪声，增强有用的信息，及一切必要的使信息纯化的处理过程。

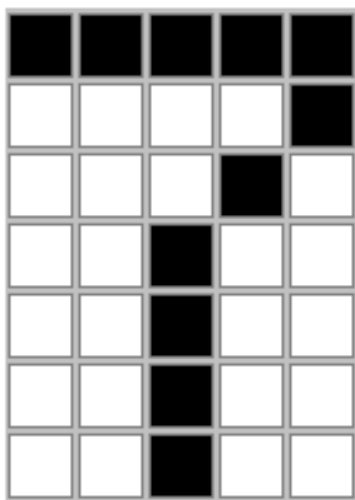


➤ 特征提取和选择

- ✓ 获取的原始量测数据转换成能反映事物本质，并将其最有效分类的特征表示（提取）。对所获取的信息实现从测量空间到特征空间的转换。
- ✓ 选择：为减少特征矢量维数，选择物体的最本质的特征。



例：印刷体数字大多通过扫描仪输入，或从图像中获取。这样一来，一个数字往往用一个 $N \times M$ 的数组表示。如果 $N = 5$ ， $M = 7$ ，则一个数字就用 5×7 共35个网格是黑是白来表示。如令黑为“1”，白为“0”，那么一个数字就可用35维的二进制向量表示。这就是典型的特征向量表示法。





第一章 模式识别概论





第一章 模式识别概论





➤ 特征提取

- ✓ 原始数据——非线性运算——特征数据
- ✓ 没有解析方法指导特征的选取，良好的特征应具有
的4个特点：
 - 可区别性：不同的类别有明显差异
 - 可靠性：相同的类比较接近
 - 独立性：不相关
 - 数量少：



➤ 特征features:

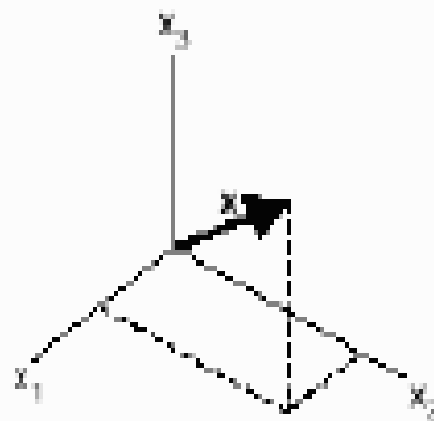
- ✓ 特征向量feature vectors: 样本的所有特征组成的 n 维向量是样本在数学上的表达，因此也称为样本
- ✓ 样本空间feature space: 特征向量所在的 n 维空间，每一个样本（特征向量）是该空间中的一个点，一个类别是该空间中的一个区域



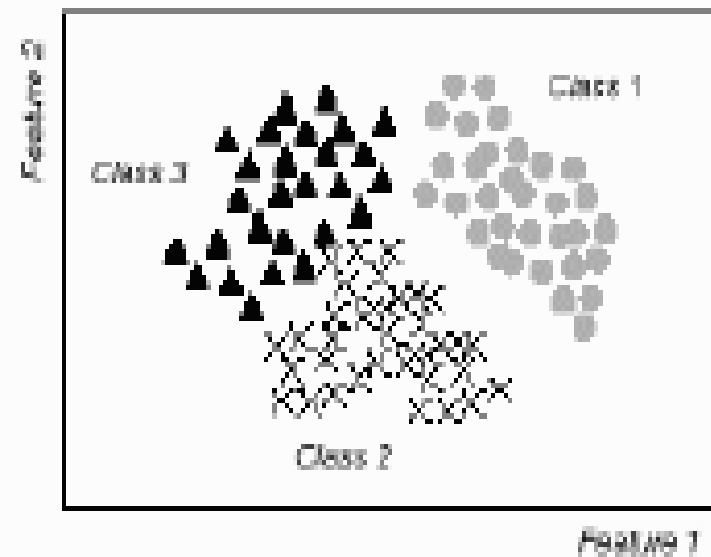
第一章 模式识别概论

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_d \end{bmatrix}$$

Feature vector



Feature space (3D)



Scatter plot (2D)



➤ 分类决策

- ✓ 根据特征提取器得到的特征向量来给一个被测对象赋一个类别标记;
- ✓ 模式识别系统工作有两阶段，**阶段一是通过训练，阶段二是分类决策**。所谓训练是指在已确定的特征空间中，对作为训练样本的量测数据进行特征选择与提取，得到它们在特征空间的分布，依据这些分布决定分类器的具体参数。
- ✓ 分类难易程度取决两因素：**同一个类别的不同个体之间的特征值的波动；不同类别样本的特征值之间的差异。**



➤ Example:

- ✓ 图为一个二维特征空间两类物体的分布状况，其中 x_1 与 x_2 分别为两个特征坐标。由于各类样本分布呈现出聚类状态，因此可以将该特征空间划分成由各类占据的子空间，确定相应的决策分界。一般说来采用什么样式的分界由设计者决定，如上述二维特征空间中可用直线、折线或曲线作为类别的分界线。但分界线的具体参数则利用训练样本经训练过程确定。

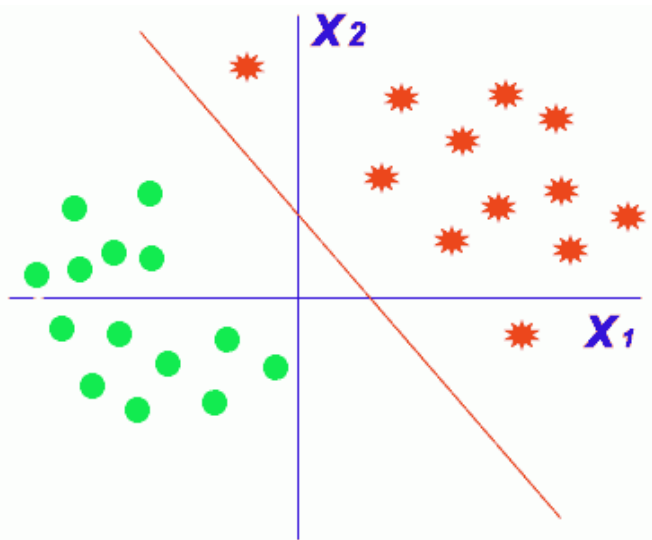


图 1.2

分类决策过程是指分类器在分界形式及其具体参数都确定后，对待分类样本进行分类决策的过程。在左图所示的情况下，待识别样本按处于分界线左下方，或右上方分类。

Comments:



一般来说，同一类事物之间属性应比较近似，而不同类事物之间的属性之间应差异较大。这种现象在特征空间的分布中的表现往往是：同类事物的特征向量聚集在一起，聚集在一个相对集中的区域，而不同事物则分别占据不同的区域。因此待识别的事物，如果它的特征向量出现在某一类事物经常出现或可能出现的区域内，该事物就被识别为该类事物。这就是识别事物的基本方法。



名词约定

➤ 分类器classifier:

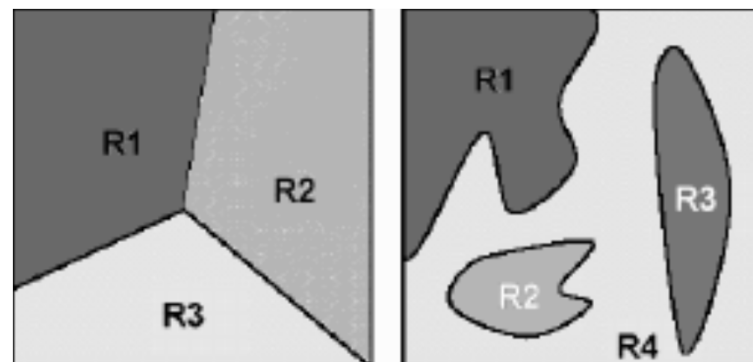
✓ 能够将每个样本都分到某个类别中去（或者拒绝）的计算机算法

➤ 决策域decision region:

✓ 分类器将特征空间划分为若干区域

➤ 分类边界、决策边界或分类面、决策面 decision boundary:

✓ 不同分类区域之间的边界





Outline:

- 模式与模式识别
 - ✓ 概念
 - ✓ 模式识别应用
 - ✓ 实例: fish recognition
- 模式识别的发展
- 模式识别的方法
 - ✓ 基于知识的方法
 - ✓ 基于数据的方法
- 监督模式识别与非监督模式识别
- 模式的描述方法
- 模式识别系统
- 模式识别的若干问题

有关模式识别的若干问题



学习

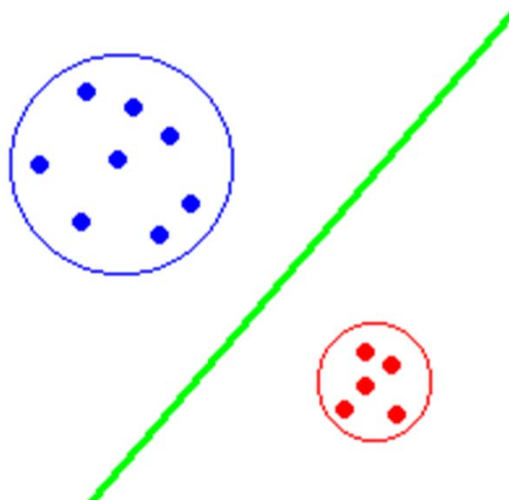
机器也有个学习过程，模式识别系统包括了训练这一环节与工作方式。但是在模式识别系统中，尤其是传统的模式识别技术中，**信息获取，预处理，特征提取与选择一般都是设计者安排好的，机器本身无法从训练中培养出选择特征的能力，而训练的实质也只是按设计者拟订的数学公式，把训练样本提供的数据作为自变量执行计算求解的过程。**



确定分类决策的具体数学公式是通过分类器设计这个过程确定的。在模式识别学科中一般把这个过程称为训练与学习的过程。这是因为分类的规则是依据训练样本提供的信息来确定，在分类器的设计阶段，要使用一批训练样本，其中包括各种类别的样本，因此由这些样本可以大致勾画出各类事物在特征空间分布的规律性，从而为确定使用什么样的分类数学公式以及这些公式中的参数确定提供了信息。



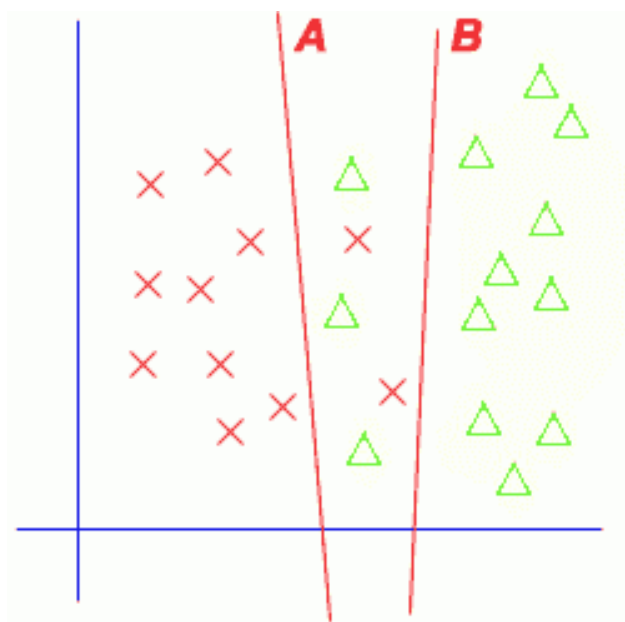
如：图中两类训练样本的分布体现出近似圆形的分布。因此如能把这两个圆形区域确定下来，将它们的边界用某种数学式子近似，那么落在某一个圆形内的样本就可以用这种数学式子来判断。从图中还可以看到，比较精确地表达不同类样本分布的聚集区不一定是必须的。用一条直线(线性方程)也许可以达到同样的目的。满足直线的方程是一个线性方程，写成 $f(x_1, x_2) = ax_1 + bx_2 + c = 0$ ，而不在该直线上的点则用 $f(x_1, x_2)$ 是否大于零或小于零来分辨。使用直线的好处是计算方便，对一个实际分类问题，快速计算、快速分类是十分重要的。





所谓模式识别中的学习与训练是从训练样本提供的数据中找出某种数学式子的最优解，这个最优解使分类器得到一组参数，按这种参数设计的分类器使人们设计的某种准则达到极值。

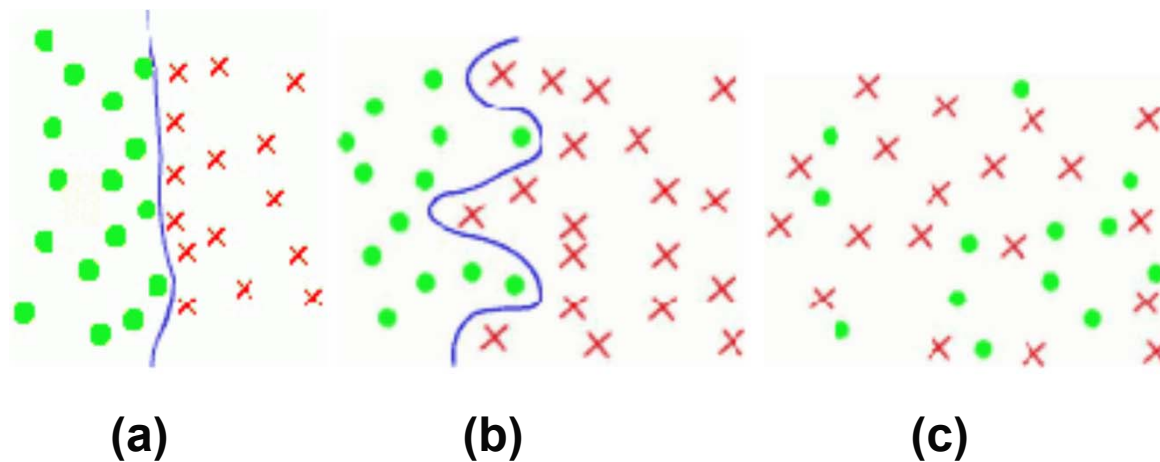
分类器参数的选择或者学习过程得到的结果取决于设计者选择什么样的准则函数。不同准则函数的最优解对应不同的学习结果，得到性能不同的分类器。





➤ 模式的紧致性

- ✓ 分类器设计难易程度与模式在特征空间的分布方式有密切关系，如图(a)、(b)与(c)分别表示了两类在空间分布的三种状况。其中(a)中两类样本存在各自明确的区域，它们之间的分界线(或面，超曲面)具有简单的形式，因而也较易区分，(b)中两类虽有各自不同的区域，但分界面的形式比较复杂，因而设计分类器的难度要大得多，如果遇到(c)类的情况则简直到了无法将它们正确分类的地步。





➤ 紧致性

- ✓ 当特征空间中属于同一个类的模式相似度远高于与其它类中的模式的相似度时，称模式类具有紧致性。
- ✓ 紧致性要求是PR的基本要求，只有当两个类之间相似度远低于同一个类内部的相似度时，分类的错误率才会较低。
- ✓ 但是在一个PR任务中，有时很难满足紧致性的要求。此时可通过增加特征空间的维数，或进行空间映射变换来增强该问题模式类的紧致性。



➤ 紧致集性质:

- ✓ (1) 临界点的数量与总的点数相比很少;
- ✓ (2) 集合中任意两个内点可以用光滑线连接, 在该连线上的点也属于这个集合;
- ✓ (3) 每个内点都有一个足够大的邻域, 在该领域中只包含同一集合中的点;

➤ 如果样本的确是可分的话, 这就意味着可以通过一种变换, 使它们在相应的特征空间中界线分明, 也就是具有了紧致性。模式识别系统设计的任务就是要寻找这样一种变换, 即选择一种特征空间, 使不同类别的样本能正确地分开。



➤ 相似性度量

- ✓ 同类物体之所以属于同一类，在于它们的某些属性相似，因此可选择适当的度量方法检测出它们之间的相似性，人们也正是依据物体之间的相似程度将它们分类的。
- ✓ 在特征空间中用特征向量描述样本的属性，就是把相似性度量用距离度量表示。在找到合适的特征空间情况下，同类样本应具有聚类性，或紧致性好，而不同类别样本应在特征空间中显示出具有较大的距离。统计模式识别各种方法实际上都是直接或间接以距离度量为基础的。



➤ 距离度量

✓ 欧氏距离:

$$\delta(X_k, X_j) = \sqrt{\sum_{i=1}^D (x_{ki} - x_{ji})^2}$$

✓ 其它形式:

$$\delta(X_k, X_j) = \sum_{i=1}^D |x_{ki} - x_{ji}| \quad \text{等}$$