模式识别与机器学习

黄庆明，常虹，郭嘉丰，山世光
中国科学院大学计算机学院/中科院计算所
qmhuang@ucas.ac.cn, changhong@ict.ac.cn，
guojiafeng@ict.ac.cn, sgshan@ict.ac.cn

教师助教：李亮（liang.li@vipl.ict.ac.cn）
学生助教：毕超，周昳晨，高培峰

引　言

课程对象

- 计算机应用技术专业硕士研究生的专业核心课
- 计算机科学与技术、电子科学与技术、自动化技术等学科硕士研究生的专业普及课

相关学科

图像处理　　计算机视觉　　数据挖掘

控制论　　自然语言处理　　……

模式识别与机器学习

统计学　　概率论与数理统计　　线性代数（矩阵计算）

形式语言　　多元统计学习　　最优化方法　　……

教学方法

- 着重讲述模式识别与机器学习的基本概念，基本理论和方法，关键算法原理以及典型应用情况。
- 注重理论与实践紧密结合
  - 实例教学：通过实例讲述如何将所学知识运用到实际应用之中
- 尽量避免引用过多的、繁琐的数学推导。

教学目标

- 掌握模式识别与机器学习的基本概念和方法
- 有效地运用所学知识和方法解决实际问题
- 为研究新的模式识别与机器学习的理论和方法打下基础

题外话

- 基本：完成课程学习（作业），通过考试，获得学分。
- 提高：能够将所学知识和内容用于课题研究，解决实际问题，完成毕业论文。
- 飞跃：通过这门课程的学习，改进思维方式，为将来的工作打好基础，终身受益。

参考文献

- 边肇祺，模式识别（第二版），清华大学出版社，2000。
- 张学工，模式识别（第三版），清华大学出版社，2010。
- 李航，统计学习基础，清华大学出版社，北京，2012。
- 卿来云、黄庆明，机器学习-从原理到应用，人民邮电出版社，2020。
- R. Duda, P. Hart, D. Stork, Pattern Classification, second edition, 2000

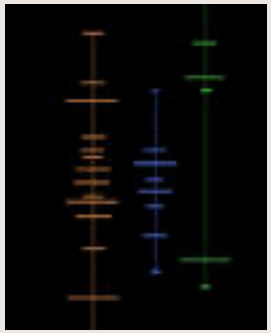
（中译本：李宏东等译，模式分类，机械工业出版社，2004）
- 《PRML：模式识别与机器学习(中文版)》，下载地址：https://max.book118.com/html/2018/1203/6241034204001233.shtm

机构、会议、刊物

- 1973年 IEEE发起了第一次关于模式识别的国际会议“ICPR”（此后两年一次），成立了国际模式识别协会---“IAPR”
- 1977年IEEE成立PAMI委员会，创立IEEE Trans. on PAMI，并支持ICCV、CVPR两个会议
- 1980年，CMU召开第一届机器学习国际研讨会，之后逐渐发展成为国际机器学习学会（IMLS）举办的机器学习国际会议ICML
- 1986年，国际期刊Machine Learning创刊
- 其它刊物
  - Pattern Recognition (PR)
  - Pattern Recognition Letters (PRL)
  - Pattern Analysis and Application (PAA)
  - Journal of Machine Learning Research (JMLR)

第一章　概论

什么是模式（Pattern）？



什么是模式？

- 广义地说，存在于时间和空间中可观察的物体，如果我们可以区别它们是否相同或是否相似，都可以称之为模式。
- 模式所指的不是事物本身，而是从事物获得的信息，因此，模式往往表现为具有时间和空间分布的信息。
- 模式的直观特性:
  - 可观察性
  - 可区分性
  - 相似性

模式识别的概念

- 模式识别 – 直观，无所不在，“人以类聚，物以群分”
  - 周围物体的认知：桌子、椅子
  - 人的识别：张三、李四
  - 声音的辨别：汽车、火车，狗叫、人语
  - 气味的分辨：炸带鱼、红烧肉
- 人和动物的模式识别能力是极其平常的，但对计算机来说却是非常困难的。

模式识别与机器学习的研究

- 目的：利用计算机对物理对象进行分类，在错误概率最小的条件下，使识别的结果尽量与客观物体相符合。
- Y = F(X)
  - X的定义域取自特征集
  - Y的值域为类别的标号集
  - F是模式识别的判别方法
- 机器学习利用大量的训练数据可以获得更好的预测结果。

机器学习的概念

- 机器学习：研究如何构造理论、算法和计算机系统，让机器通过从数据中学习后可以进行如下工作：分类和识别事物、推理决策、预测未来等。
- Wiki: “The design and development of algorithms that take as input empirical data and yield patterns or predictions that generated the data.”

模式识别简史

- 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 Chemsky 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。
- 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。

模式识别简史

- 21世纪以来，模式识别研究呈现一些新特点
  - 贝叶斯学习理论越来越多地用来解决具体的模式识别和模型选择问题，产生了良好的分类性能。
  - 传统的问题，如概率密度估计、特征选择、聚类等不断受到新的关注，新的方法或改进/混合的方法不断提出。
  - 模式识别和机器学习相互渗透，特征提取和选择、分类、聚类、半监督学习、深度学习等问题日益成为二者共同关注的热点。
  - 模式识别系统开始越来越多地用于现实生活，如车牌识别、手写字符识别、生物特征识别等。

机器学习简史

- 机器学习的发展与模式识别密切相关。
- 第一阶段是在50年代中叶到60年代中叶，属于热烈时期。研究的是以40年代兴起的神经网络模型为理论基础的“没有知识”的学习。模式识别发展的同时形成了机器学习的两种重要方法：判别函数法和进化学习
- 第二阶段是在60年代中叶至70年代中叶，被称为机器学习的冷静时期。研究的目标是模拟人类的概念学习阶段，并采用逻辑结构或图结构作为机器内部描述。神经网络学习机因理论缺陷转入低潮。
- 第三阶段是从70年代中叶至80年代中叶，称为复兴时期。从学习单个概念扩展到学习多个概念，探索不同的学习策略和方法（如模式方法推断）。

机器学习简史

- 机器学习的新阶段始于1986年。机器学习有了更强的研究手段和环境，出现了符号学习、神经网络学习、进化学习和强化学习等。
- 机器学习已成为新的边缘学科并在高校形成一门课程。它综合应用心理学、生物学和神经生理学以及数学、自动化和计算机科学形成机器学习理论基础。
- 结合各种学习方法，取长补短的多种形式的集成学习系统研究正在兴起。
- 各种学习方法（归纳学习、连接学习、强化学习、深度学习）的应用范围不断扩大，一部分已形成产品。尤其是深度学习的发展方兴未艾，正在人工智能等领域发挥越来越重要的作用。

应用（举例）

- 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

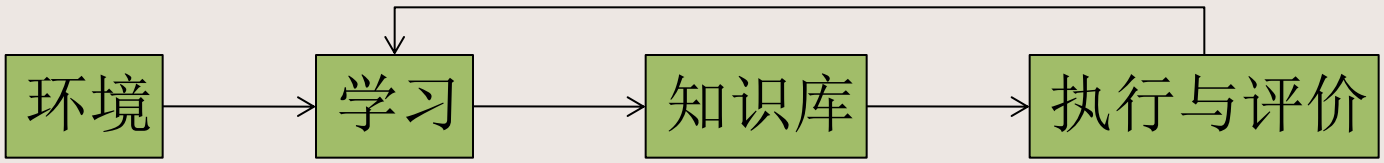
应用（举例）

- 工程
  - 产品缺陷检测、特征识别、语音识别、自动导航系统、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、人脸识别、监视和报警系统

模式识别方法

- 模式识别系统的目标：在特征空间和解释空间之间找到一种映射关系，这种映射也称之为假说。
  - 特征空间：从模式得到的对分类有用的度量、属性或基元构成的空间。
  - 解释空间：将c个类别表示为 $\omega_i \in \Omega, i = 1,2,\cdots,c$

    其中 Ω 为所属类别的集合，称为解释空间。
- 机器学习的目标：针对某类任务T，用P衡量性能，根据经验来学习和自我完善，提高性能。



假说的两种获得方法

- 监督学习、概念驱动或归纳假说：在特征空间中找到一个与解释空间的结构相对应的假说。在给定模式下假定一个解决方案，任何在训练集中接近目标的假说也都必须在“未知”的样本上得到近似的结果。
  - 依靠已知所属类别的训练样本集，按它们特征向量的分布来确定假说（通常为一个判别函数），在判别函数确定之后能用它对未知的模式进行分类；
  - 对分类的模式要有足够的先验知识，通常需要采集足够数量的具有典型性的样本进行训练。

假说的两种获得方法（续）

- 非监督学习、数据驱动或演绎假说：在解释空间中找到一个与特征空间的结构相对应的假说。这种方法试图找到一种只以特征空间中的相似关系为基础的有效假说。
  - 在没有先验知识的情况下，通常采用聚类分析方法，基于“物以类聚”的观点，用数学方法分析各特征向量之间的距离及分散情况；
  - 如果特征向量集聚集若干个群，可按群间距离远近把它们划分成类；
  - 这种按各类之间的亲疏程度的划分，若事先能知道应划分成几类，则可获得更好的分类结果。

主要分类和学习方法

- 数据聚类
- 统计分类
- 结构模式识别
- 神经网络
- 监督学习
- 无监督学习
- 半监督学习
- 增强学习
- 集成学习
- 深度学习
- 元学习
- 多任务学习
- 多标记学习
- 对抗学习

数据聚类

- 目标：用某种相似性度量的方法将原始数据组织成有意义的和有用的各种数据集。
- 是一种非监督学习的方法，解决方案是数据驱动的。

统计分类

- 基于概率统计模型得到各类别的特征向量的分布，以取得分类的方法。
- 特征向量分布的获得是基于一个类别已知的训练样本集。
- 是一种监督分类的方法，分类器是概念驱动的。

结构模式识别

- 该方法通过考虑识别对象的各部分之间的联系来达到识别分类的目的。
- 识别采用结构匹配的形式，通过计算一个匹配程度值（matching score）来评估一个未知的对象或未知对象某些部分与某种典型模式的关系如何。
- 当成功地制定出了一组可以描述对象部分之间关系的规则后，可以应用一种特殊的结构模式识别方法 – 句法模式识别，来检查一个模式基元的序列是否遵守某种规则，即句法规则或语法。

神经网络

- 神经网络是受人脑组织的生理学启发而创立的。
- 由一系列互相联系的、相同的单元（神经元）组成。相互间的联系可以在不同的神经元之间传递增强或抑制信号。
- 增强或抑制是通过调整神经元相互间联系的权重系数来（weight）实现。
- 神经网络可以实现监督和非监督学习条件下的分类。

监督学习

- 监督学习是从有标记的训练数据来推断或建立一个模型，并依此模型推测新的实例。
- 训练数据包括一套训练实例。在监督学习中，每个实例是由一个输入对象（通常为矢量）和一个期望的输出值（也称为监督信号）组成。
- 一个最佳的模型将能够正确地决定那些看不见的实例的标签。常用于分类和回归。

无监督学习

- 无监督学习是我们不告诉计算机怎么做，而是让它自己去学习怎样做一些事情。
- 无监督学习与监督学习的不同之处在于，事先没有任何训练样本，需要直接对数据进行建模，寻找数据的内在结构及规律，如类别和聚类。
- 常用于聚类、概率密度估计。

半监督学习

- 半监督学习（Semi-supervised Learning）是模式识别和机器学习领域研究的重点问题，是监督学习与无监督学习相结合的一种学习方法。
- 它主要考虑如何利用少量的标注样本和大量的未标注样本进行训练和分类的问题。
- 半监督学习的主要算法有五类：基于概率的算法；在现有监督算法基础上改进的方法；直接依赖于聚类假设的方法；基于多视图的方法；基于图的方法。

增强学习

- 增强学习要解决的问题：一个能够感知环境的自治机器人，怎样通过学习选择能达到其目标的最优动作。
- 机器人选择一个动作用于环境，环境接受该动作后状态发生变化，同时产生一个强化信号(奖或惩)反馈回来。
- 机器人根据强化信号和环境当前状态再选择下一个动作，选择的原则是使受到正强化(奖)的概率增大。

集成学习

- 集成学习（Ensemble Learning）是机器学习中一类学习算法，指联合训练多个弱分类器并通过集成策略将弱分类器组合使用的方法。
- 由于整合了多个分类器，这类算法通常在实践中会取得比单个若分类器更好的预测结果。
- 常见的集成策略有：Boosting、Bagging、Random subspace 、Stacking等。
- 常见的算法主要有：决策树、随机森林、Adaboost、GBDT、DART等。

深度学习

- 深度学习的概念源于人工神经网络的研究，除输入层和输出层外，含多个隐藏层的神经网络就是一种深度学习结构。
- 深度学习通过层次化模型结构可从低层原始特征中逐渐抽象出高层次的语义特征，以发现复杂、灵活、高效的特征表示。
- 常见的深度学习模型有：卷积神经网络， 递归神经网络，深度信任网络，自编码器，变分自编码器等。

元学习

- 元学习（Meta Learning）或者叫做“学会学习”（Learning to Learn），它是要“学会如何学习”，即利用以往的知识经验来指导新任务的学习，具有学会学习的能力。
- 当前的机器学习模型往往只局限于从头训练已知任务并使用精调来学习新任务，耗时较长，且性能提升较为有限。
- Meta Learning 就是研究如何让元模型记忆理解以往学习知识，使算法能在小样本训练的情况下完成新任务的学习。

多任务学习

- 多任务学习是指通过共享相关任务之间的表征，联合训练多个学习任务的学习范式。
- 在通常的机器学习范式中，不同任务的学习过程往往分别处理，任务间的关系完全被割裂。而在多任务学习范式中，联系学习机制使不同任务的学习过程充分共享，可显著减少每个任务所需的训练样本。
- 多任务学习的主要形式有：联合学习、自主学习和带有辅助任务的学习。

多标记学习

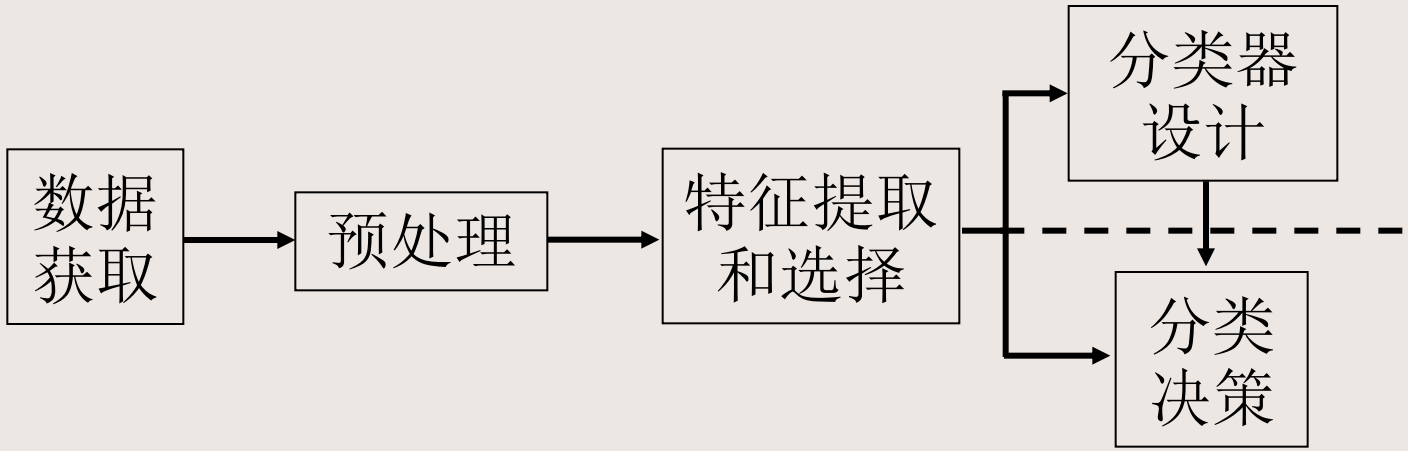
- 多标记学习问题为一种特殊的有监督分类问题，其所处理的数据集中的每个样本可同时存在多个真实类标。
- 多标记学习主要用于处理多种标签的语义重叠，如预测歌曲的音乐流派，预测图书、商品的属性标签。
- 多标记学习算法主要分为两类:
  - 问题转换法：把多标签问题转为其它学习场景，比如转为二分类、标签排序、多分类等。
  - 算法改编法：通过改编流行的学习算法去直接处理多标签数据，比如改编决策树、核技巧等。

对抗学习

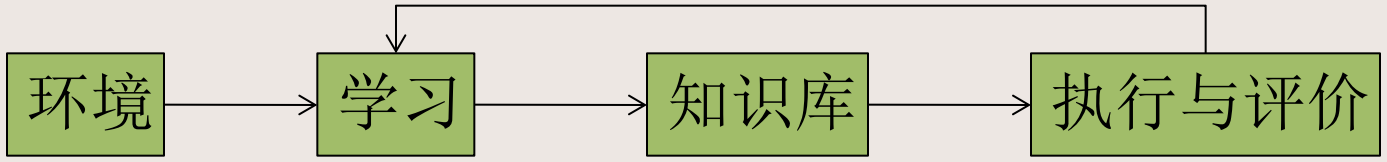
- 对抗学习是针对传统机器学习的一种攻击性方法，是机器学习和计算机安全领域都十分关注的交叉问题。
- 对抗学习主要通过恶意输入来误导机器学习算法或模型使其得到错误结果，并在该过程中暴露机器学习算法存在的脆弱性，帮助设计适应复杂环境的鲁棒学习方法。
- 常见的对抗学习方法主要有针对训练阶段的毒害式攻击以及针对测试阶段的躲避式攻击，常见的对抗学习场景主要有：垃圾邮件过滤、身份识别以及恶意软件检测等。

系统构成

- 模式识别系统的基本构成



- 机器学习的基本构成



模式识别系统组成单元

- 数据获取：用计算机可以运算的符号来表示所研究的对象
  - 二维图像：文字、指纹、地图、照片等
  - 一维波形：脑电图、心电图、季节震动波形等
  - 物理参量和逻辑值：体温、化验数据、参量正常与否的描述
- 预处理单元：去噪声，提取有用信息，并对输入测量仪器或其它因素所造成的退化现象进行复原

模式识别系统组成单元

- 特征提取和选择：对原始数据进行变换，得到最能反映分类本质的特征
  - 测量空间：原始数据组成的空间
  - 特征空间：分类识别赖以进行的空间
  - 模式表示：维数较高的测量空间->维数较低的特征空间
- 分类决策：在特征空间中用模式识别方法把被识别对象归为某一类别
  - 基本做法：在样本训练集基础上确定某个判决规则，使得按这种规则对被识别对象进行分类所造成的错误识别率最小或引起的损失最小。

机器学习系统组成单元

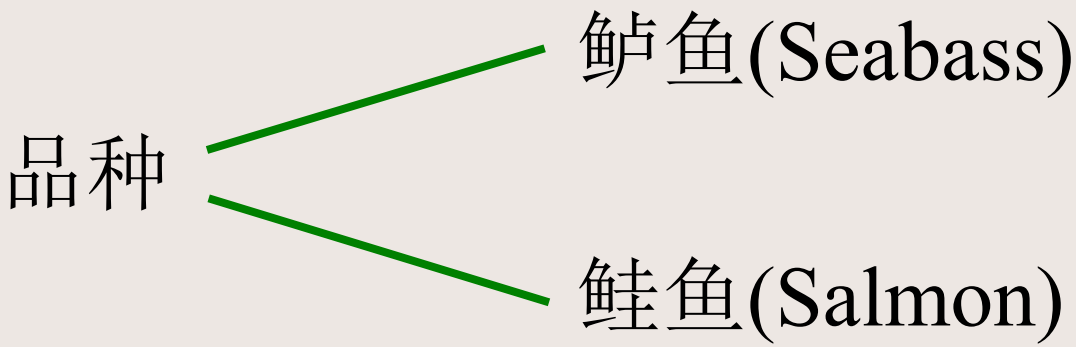
- 环境：是系统的工作对象（包括外界条件），代表信息来源。
  - 信息水平：相对于执行环节要求而言，由学习环节消除差距
  - 信息质量：实例示教是否正确、实例次序是否合理等
- 知识库：存储学习到的知识
  - 知识的表示要合理
  - 推理方法的实现不要太难
  - 存储的知识是否支持修改（更新）

机器学习系统组成单元

- 学习环节：是系统的核心模块，是和外部环境的交互接口。
  - 对环境提供的信息进行整理、分析、归纳或类比，生成新的知识单元，或修改知识库。
  - 接收从执行环节来的反馈信号，通过知识库修改，进一步改善执行环节的行为。
- 执行：根据知识库执行一系列任务
  - 把执行结果或执行过程中获得的信息反馈给学习环节

模式识别过程实例

- 在传送带上用光学传感器件对鱼按品种分类

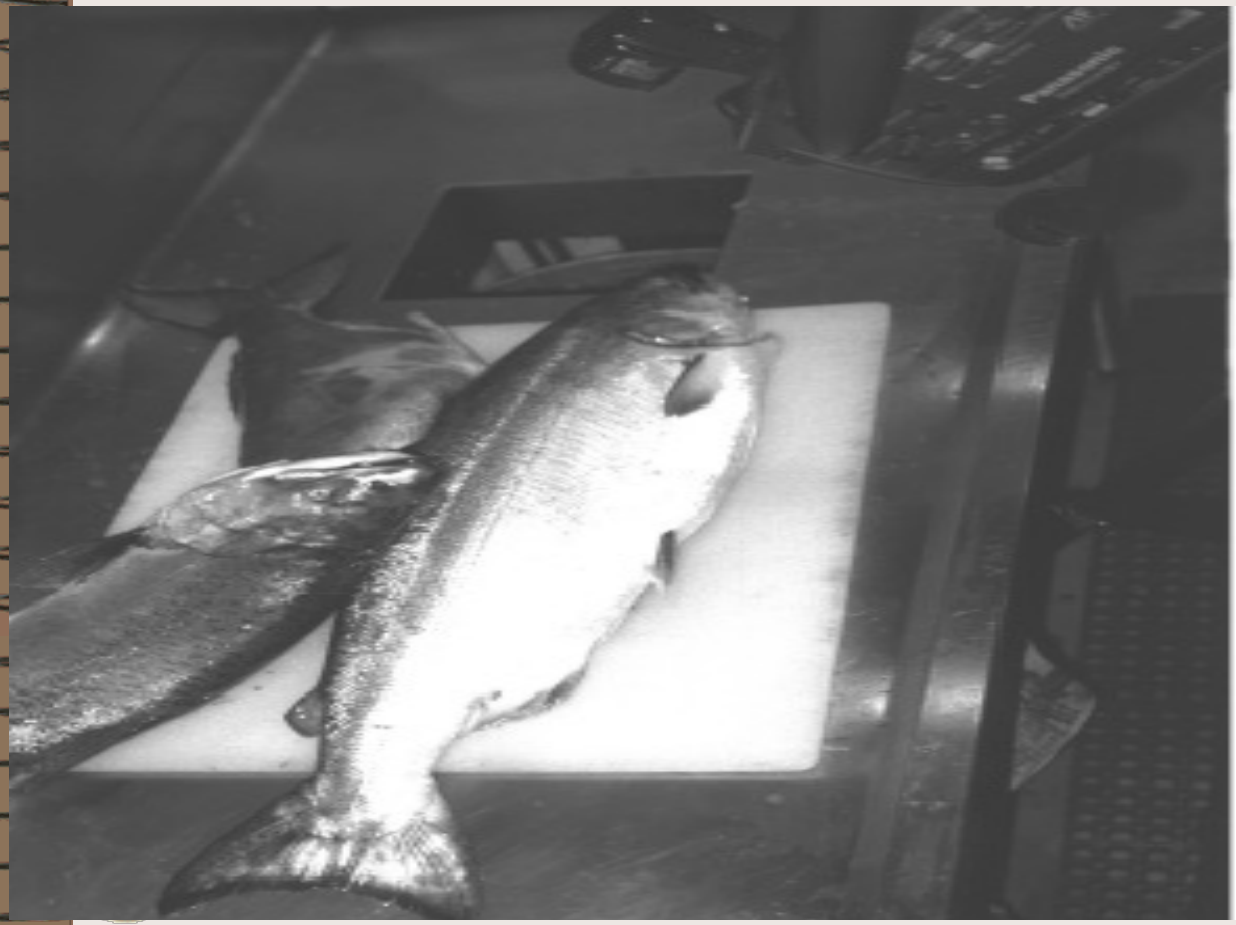


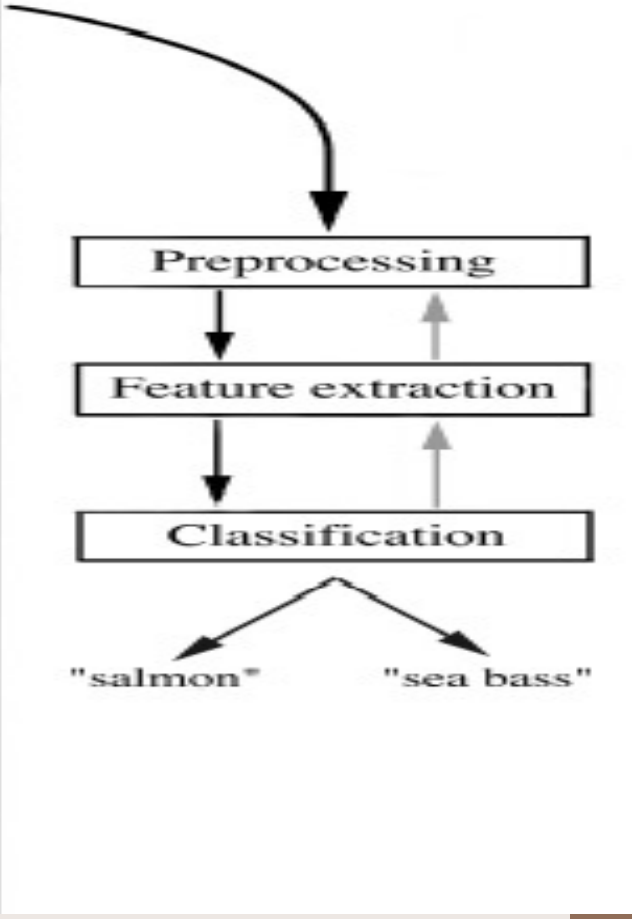
识别过程

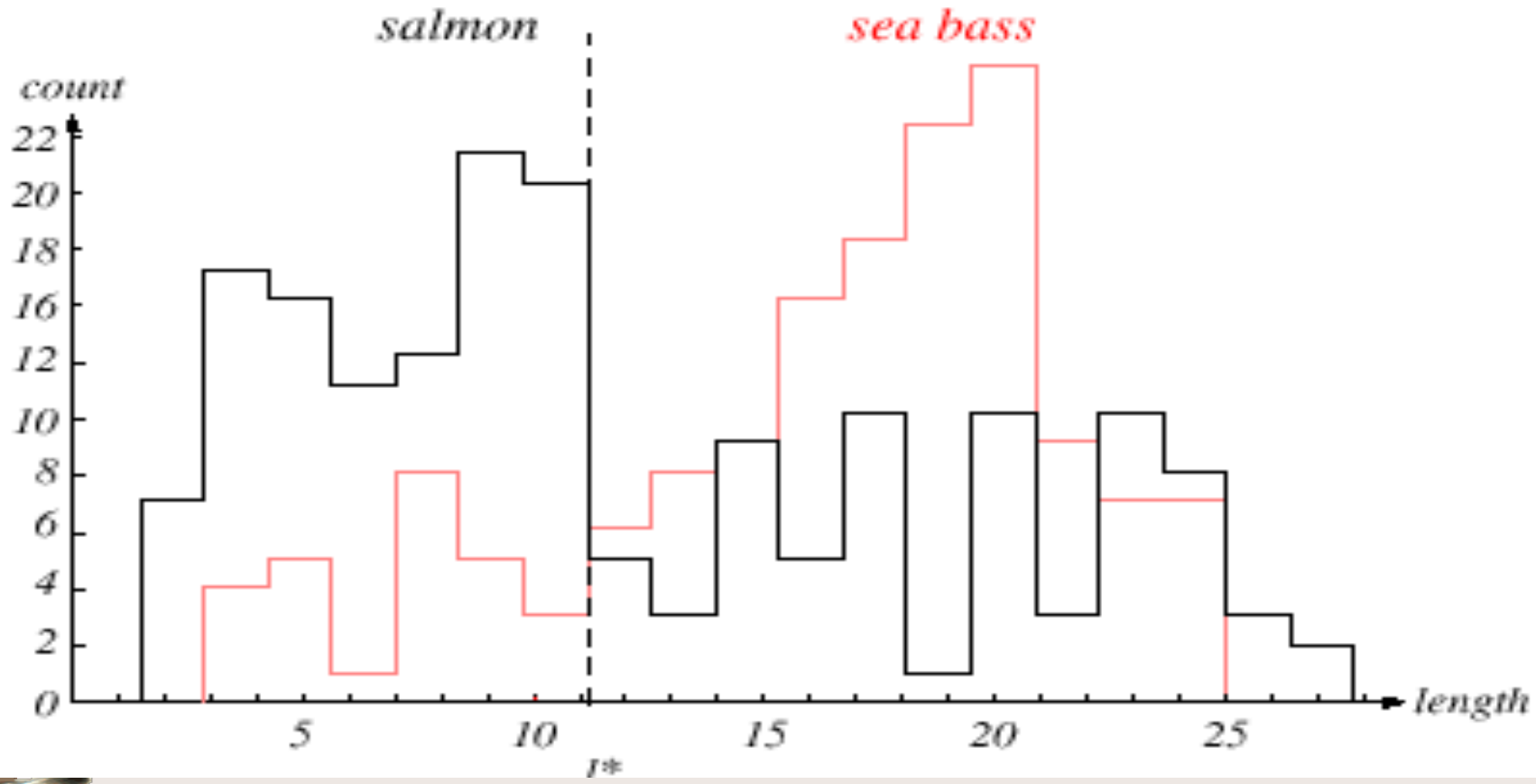
- 数据获取：架设一个摄像机，采集一些样本图像，获取样本数据
- 预处理：去噪声，用一个分割操作把鱼和鱼之间以及鱼和背景之间分开

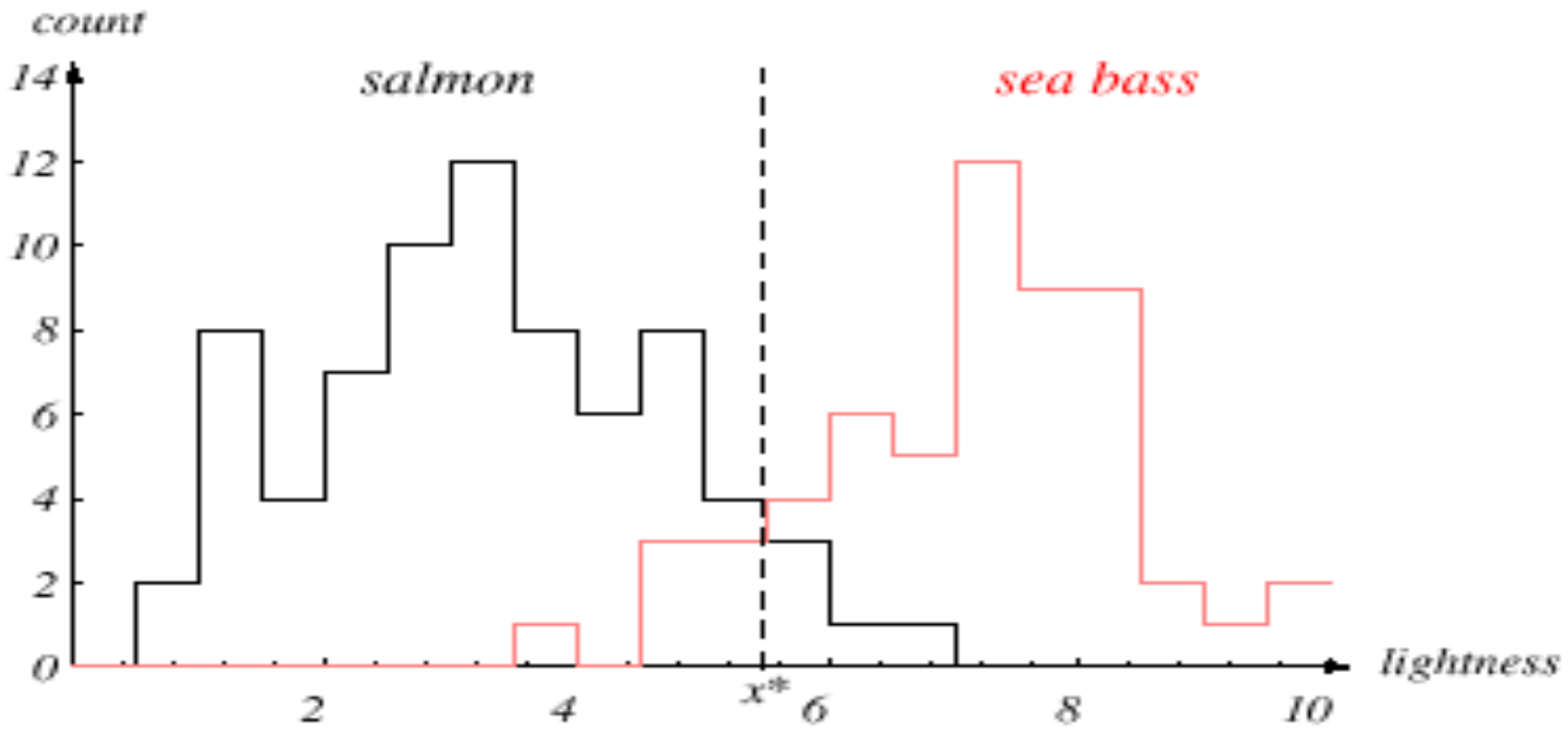
识别过程

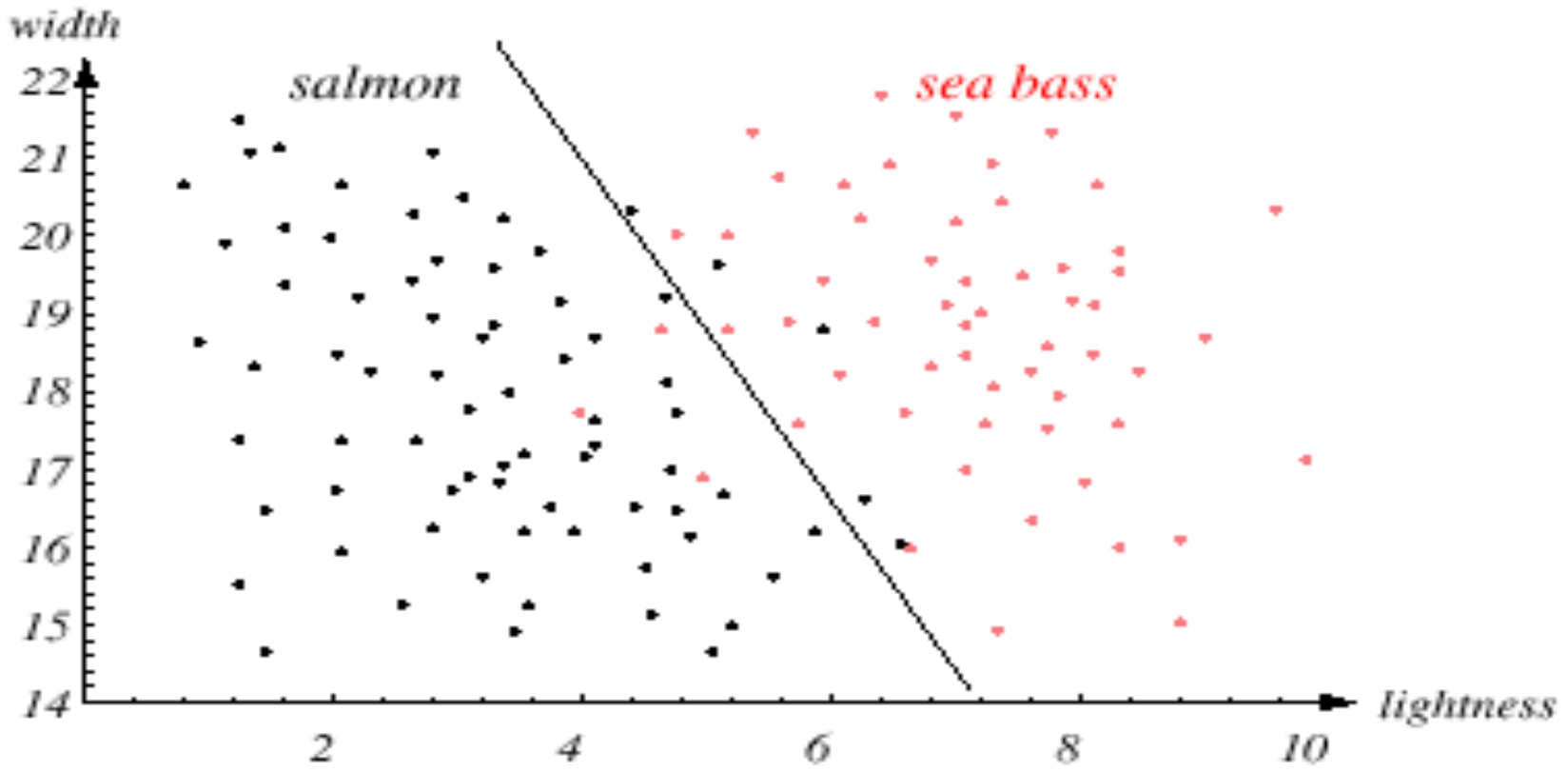
- 特征提取和选择：对单个鱼的信息进行特征选择，从而通过测量某些特征来减少信息量
  - 长度
  - 亮度
  - 宽度
  - 鱼翅的数量和形状
  - 嘴的位置，等等 …
- 分类决策：把特征送入决策分类器

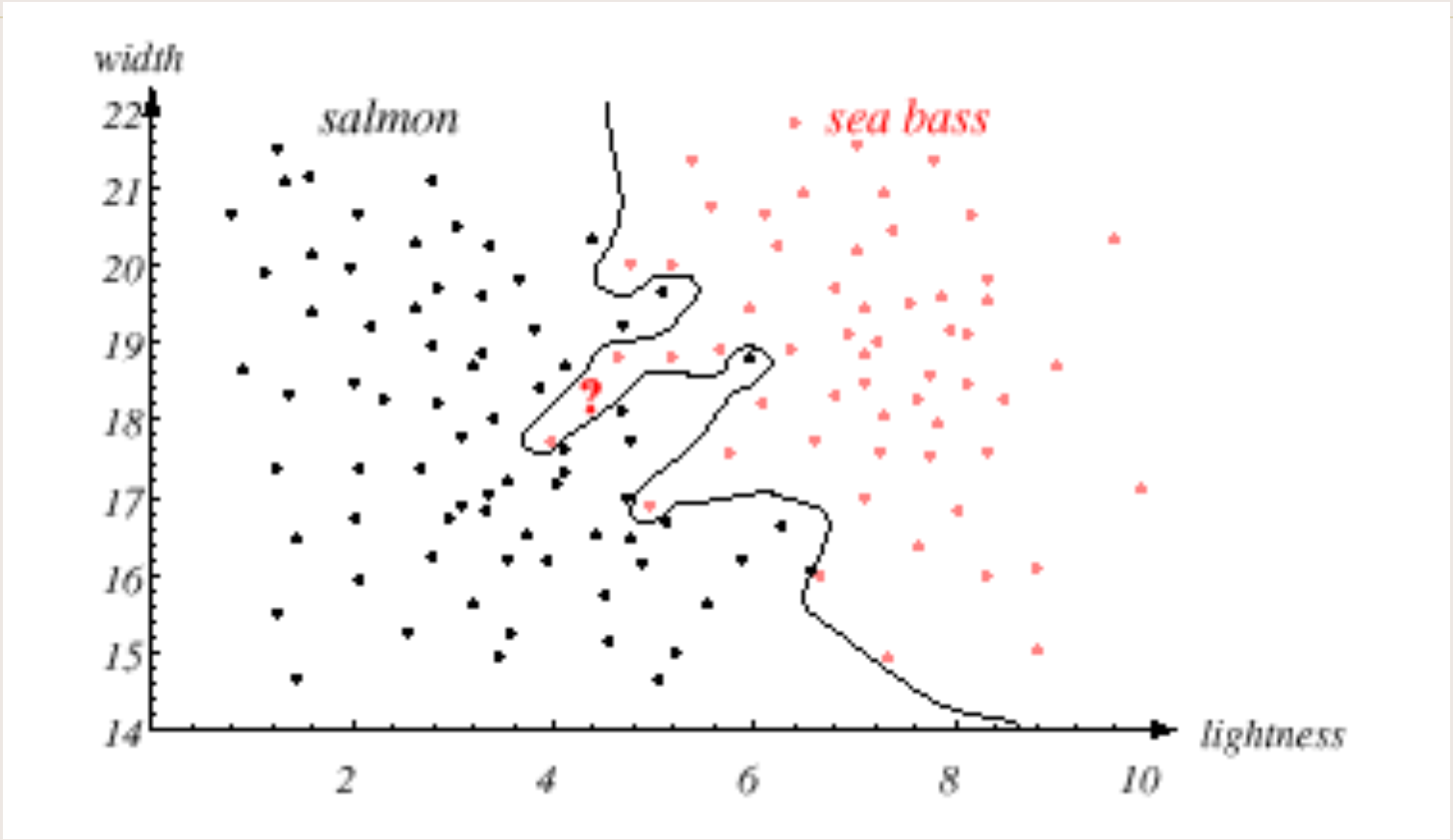


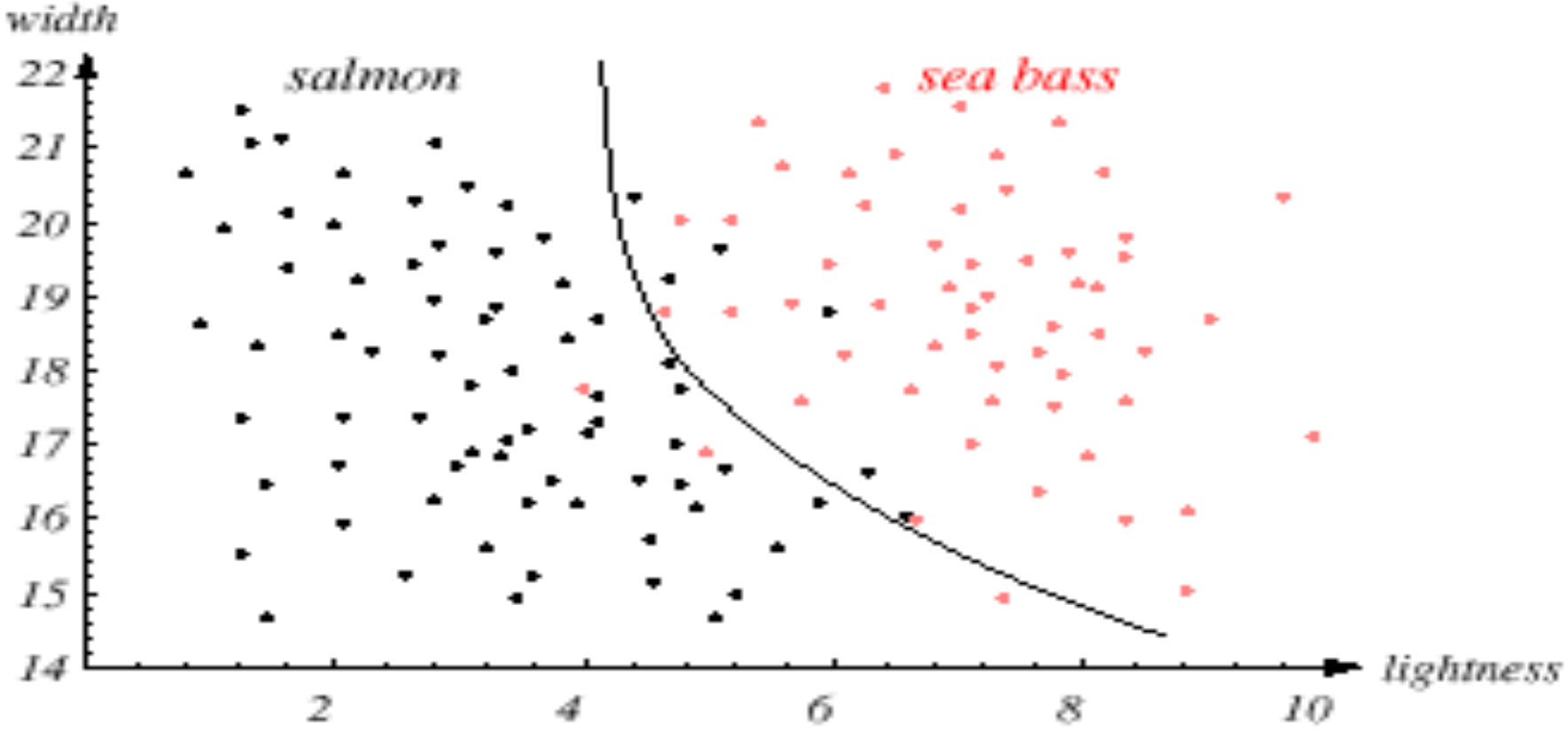

Preprocessing
Feature extraction
Classification
"salmon"
"sea bass"


salmon
sea bass
count
22
20
18
16
12
10
8
6
4
2
0
5
10
15
20
25
length
l*


count
salmon
sea bass
14
12
10
8
6
4
2
0
2
4
x*
6
8
10
lightness

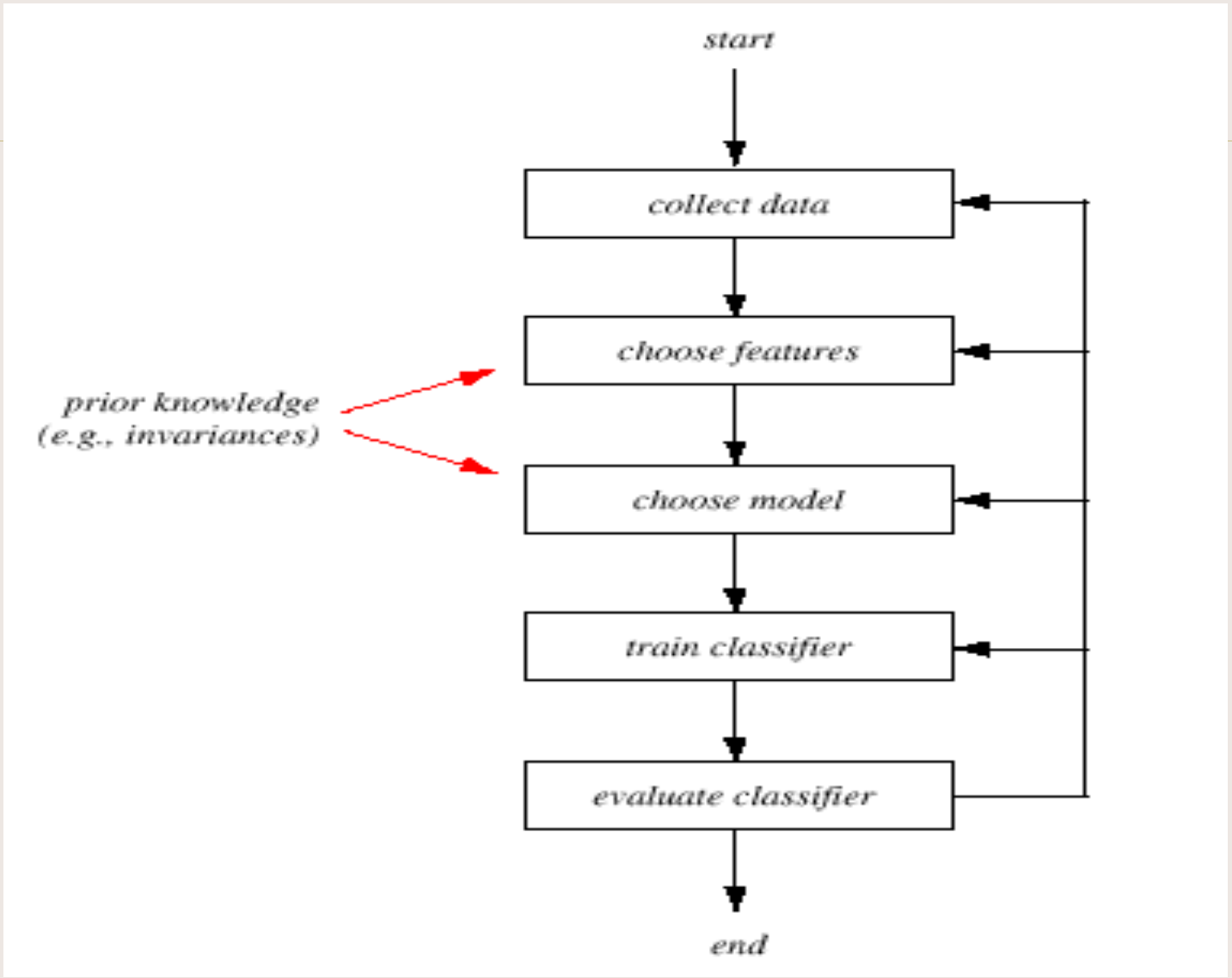

width
22
21
20
19
18
17
16
15
14
salmon
sea bass
2
4
6
8
10
lightness


width
salmon
sea bass
?
lightness
22
21
20
19
18
17
16
15
14
2
4
6
8
10


width
22
21
20
19
18
17
16
15
14
salmon
sea bass
2
4
6
8
10
lightness

模式分类器的获取和评测过程

- 数据采集
- 特征选取
- 模型选择
- 训练和测试
- 计算结果和复杂度分析，反馈


start
collect data
choose features
prior knowledge
(e.g., invariances)
choose model
train classifier
evaluate classifier
end

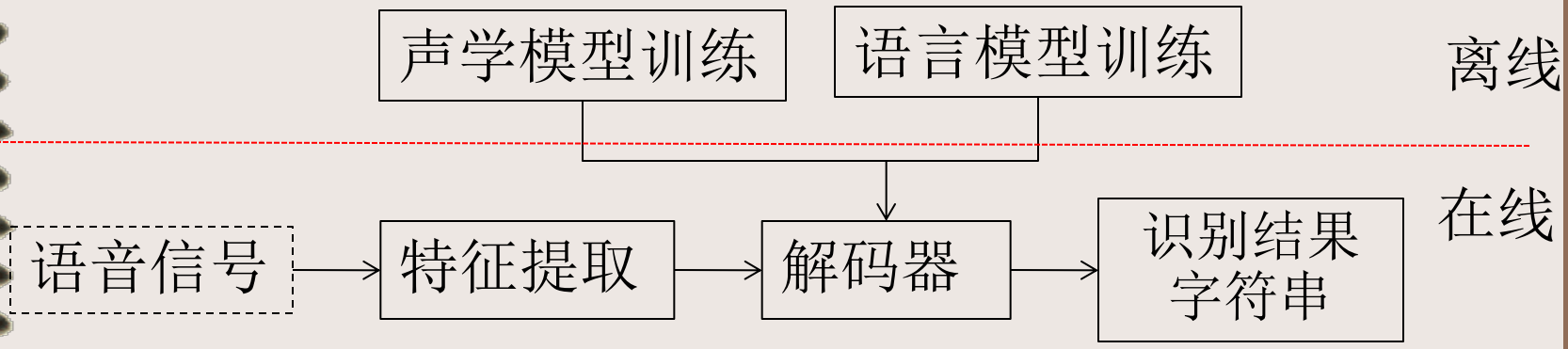
训练和测试

- 训练集：是一个已知样本集，在监督学习方法中，用它来开发出模式分类器。
- 测试集：在设计识别和分类系统时没有用过的独立样本集。
- 系统评价原则：为了更好地对模式识别系统性能进行评价，必须使用一组独立于训练集的测试集对系统进行测试。

语音识别实例

- 环境：语音信号
- 知识库：声学产生模型
- Now most pocket Speech Recognizers or Translators are running on some sort of learning device — the more you play/use them, the smarter they become.

语音识别过程

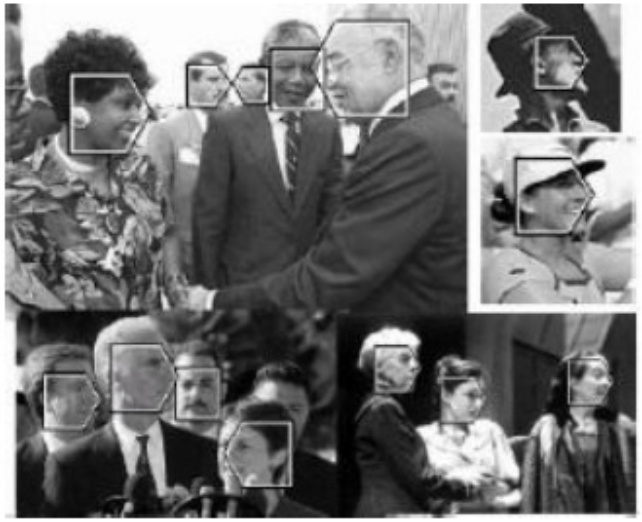


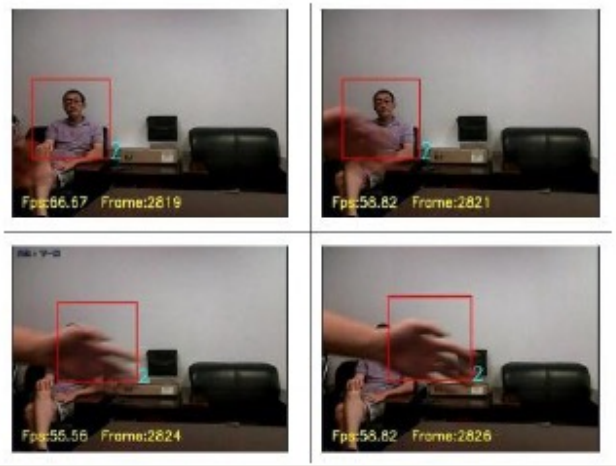
- 语音特征提取：
  - 感知特征：音调、音高、旋律、节奏
  - 声学特征：能量、过零率、LPC系数

语音识别过程

- 声学模型：
  - 通常是将获取的语音特征使用训练算法进行训练后产生。
  - 在识别时将输入的语音特征同声学模型（模式）进行匹配与比较，得到最佳的识别结果。
  - 常用的模型包括：隐马尔可夫模型HMM、支持向量机SVM、人工神经网络。

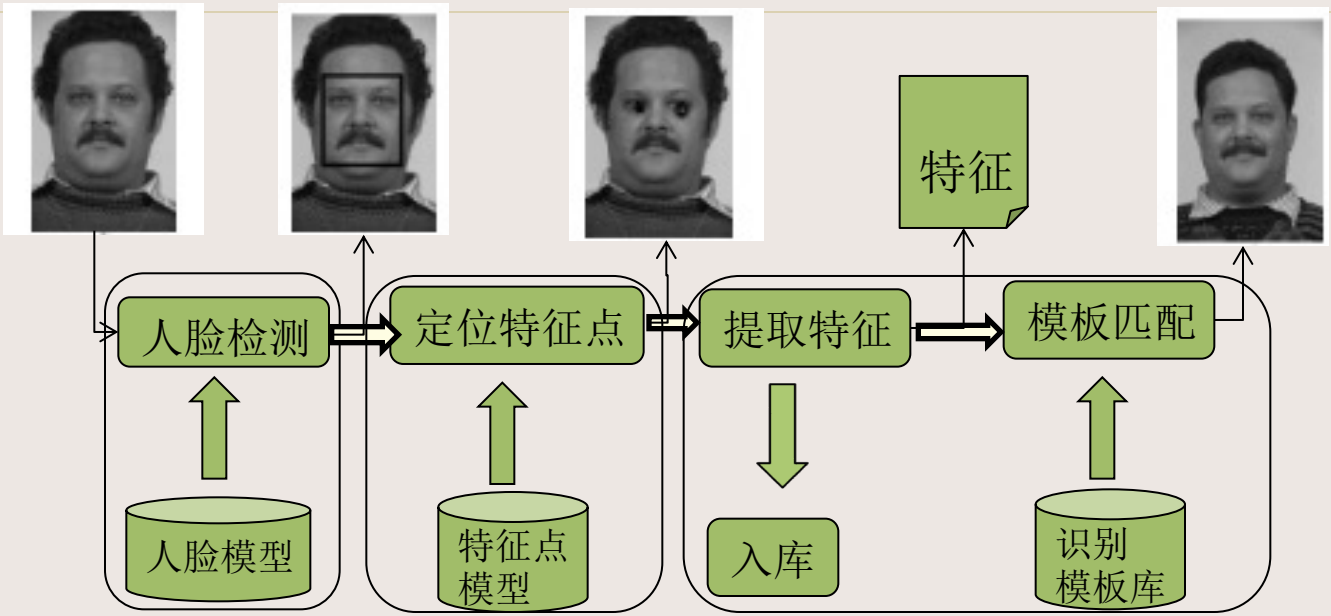
计算机视觉应用实例




Fps:66.67 Frame:2819
Fps:58.82 Frame:2821
Fps:55.56 Frame:2824
Fps:58.82 Frame:2826


sky
tree
road
grass
water
bldg
mntn
fg obj.

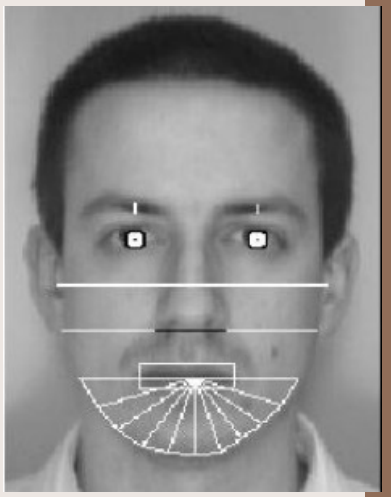
人脸识别过程



- 人脸检测方法：
  - 模板匹配：固定模板、可变形模板
  - 基于不变特征的方法：如肤色特征
  - 基于外观学习的方法：人脸和非人脸分类问题（Haar特征＋AdaBoost)

人脸识别过程

- 人脸识别的简单方法
  - 基于几何特征的识别方法：用面部关键特征的相对位置、大小、形状、面积等参数来描述人脸，根据待识别人脸与数据库中人脸的相似性进行识别
  - 基于神经网络的识别方法：神经网络的输入可以是降低分辨率的人脸图像、局部区域的自相关函数、局部纹理的二阶矩等。这类方法同样需要较多的样本进行训练
  - 特征脸、模板匹配等



本门课程的主要安排（1班）

- 第一章　概论
- 第二章　统计判别
- 第三章　判别函数
- 第四章　特征选择和提取
- 第五章　统计学习理论基础
- 第六章　有监督学习基础算法
- 第七章　支持向量机
- 第八章　无监督学习与半监督学习
- 第九章　图模型基础
- 第十章　集成学习
- 第十一章 神经网络与深度学习
- 第十二章 典型应用案例

黄庆明

常虹

郭嘉丰

山世光

本门课程的主要安排（2班）

- 第一章　概论　　黄庆明
- 第二章　统计判别
- 第三章　判别函数
- 第四章　特征选择和提取　　李国荣
- 第五章　统计学习理论基础
- 第六章　有监督学习基础算法
- 第七章　支持向量机
- 第八章　无监督学习与半监督学习
- 第九章　图模型基础
- 第十章　集成学习　　苏荔
- 第十一章 神经网络与深度学习
- 第十二章 典型应用案例

本门课程的主要安排（3班）

- 第一章　概论　　黄庆明
- 第二章　统计判别
- 第三章　判别函数
- 第四章　特征选择和提取　　李国荣
- 第五章　统计学习理论基础
- 第六章　有监督学习基础算法
- 第七章　支持向量机
- 第八章　无监督学习与半监督学习
- 第九章　图模型基础
- 第十章　集成学习　　卿来云
- 第十一章 神经网络与深度学习
- 第十二章 典型应用案例

相关数学概念

- 随机向量及其分布
  - 随机向量
    - 如果一个对象的特征观察值为$\{x_1, x_2, \ldots, x_n\}$，它可构成一个n维的特征向量值$\boldsymbol{x}$，即

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$$

      式中，$x_1, x_2, \ldots, x_n$为特征向量$\boldsymbol{x}$的各个分量。
    - 一个特征可以看作n维空间中的向量或点，此空间称为模式的特征空间R_n。

相关数学概念

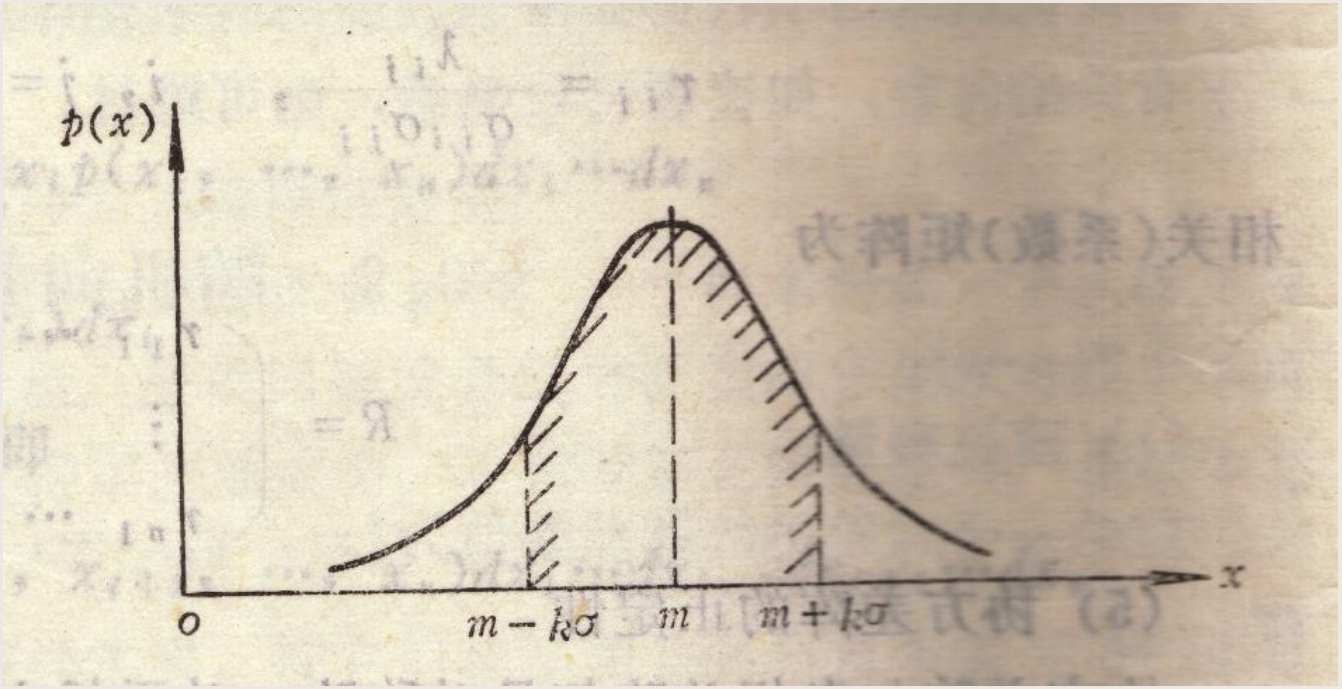
- 随机向量及其分布
  - 随机向量
    - 在模式识别过程中，要对许多具体对象进行测量，以获得许多次观测值。
    - 每次观测值不一定相同，所以对许多对象而言，各个特征分量都是随机变量，即许多对象的特征向量在n维空间中呈随机性分布，称为随机向量。

相关数学概念

- 随机向量及其分布
  - 随机向量的参数
    - 数学期望和方差
    - 协方差矩阵

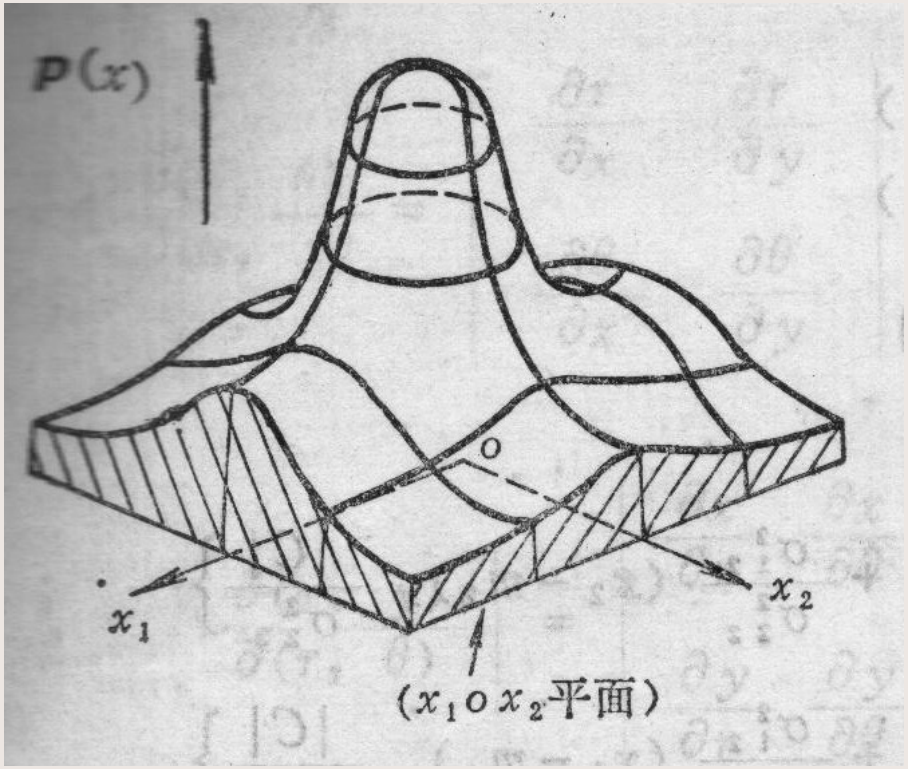
相关数学概念

- 正态分布
  - 一维正态密度函数



相关数学概念

- 正态分布
  - 多维正态密度函数



小结

- 模式识别和机器学习的概念
- 发展简史和应用
- 主要方法
- 系统构成和实例
- 几个相关的数学概念