模式识别

苗夺谦
dqmiao@tongji.edu.cn

张志飞
zhifeizhang@tongji.edu.cn

第一章 绪论

- 1.1 引言
- 1.2 一个例子
- 1.3 设计循环
- 1.4 统计模式识别的方法

1.1 引言

- 背景

- 社会背景：农业社会、工业社会、信息社会、智能社会（PR+AI）

- 学科背景：1946（计算机）、1956（人工智能）、
  1976（模式识别）、1986（人工神经网络-BP）、
  1997（第一次人机大战—Deep blue）、
  2006（深度学习—多层神经网络）、
  2011（第二次人机大战—Watson）、
  2016（第三次人机大战—Alphago）

- 应用背景：解决机器的，视觉、听觉、嗅觉、触觉、感觉

1.1 引言

- 智能

➢ 感知+认知+决策

➢ ——模式识别

➢ 记忆+学习+推理+创新+情感+想象

➢ ——人工智能

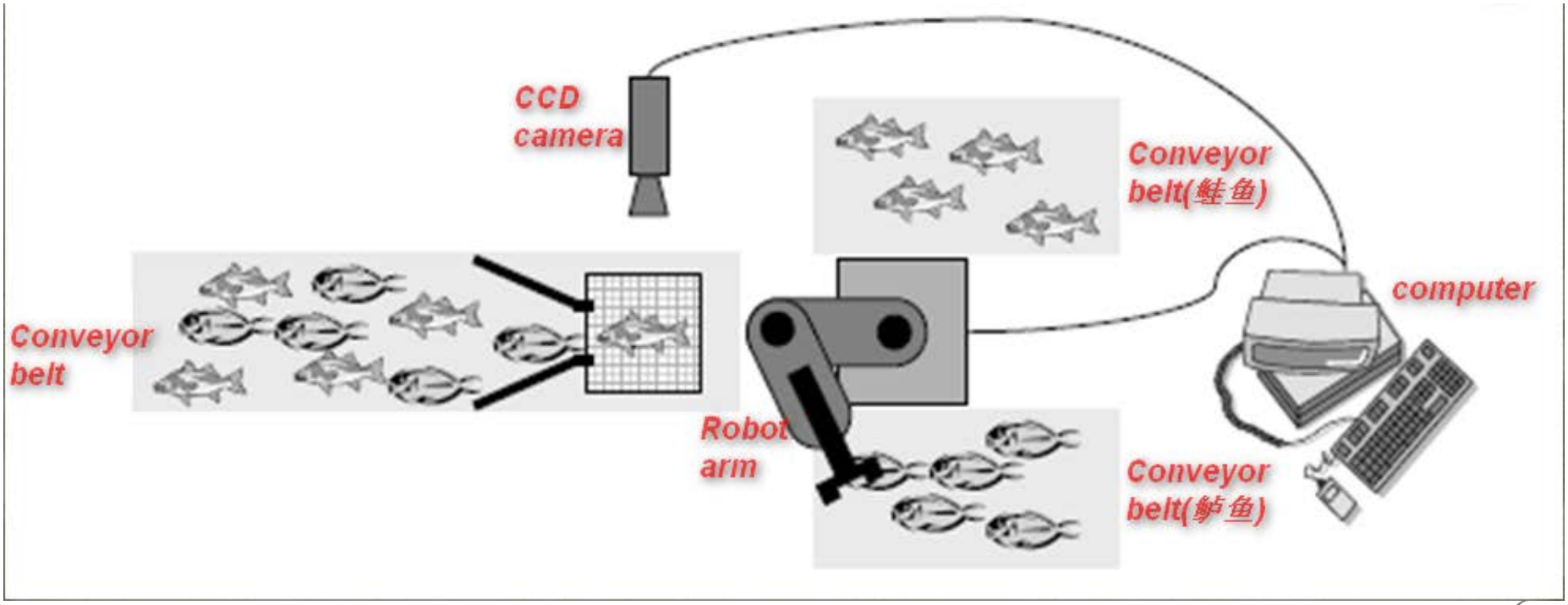
1.1 引言

- 模式（Pattern）：可观察的事物具有时间或空间分布的信息（可区分相同或相似）
- 对数据中所蕴含的模式、变化趋势、异常现象进行自动识别及识别结果的描述。——模式识别(pattern recognition)

  如，借助电子邮件的标题、内容及发件人(数据)，将其自动分类成垃圾邮件和非垃圾邮件。
- 模式识别涵盖“从问题描述和数据采集到识别分类、结果评价及解释的各个阶段”。

1.2 一个例子

- 一个鱼类加工厂，希望能将传送带上的鱼的品种(鲑鱼、鲈鱼)分类的过程自动化。
- 这一自动化系统构成如下:



一个例子(2)

- 摄像机拍摄的样本照片，显示出如下差异：
  - 两种鱼自身存在物理特性上的差异，如
    - 长度、光泽、宽度、鳍的数目和形状、嘴的位置等

特征(feature)

  - 照片本身还存在外部因素导致的差异，如，光照的不同，鱼在传送带上的位置，摄像机电子线路引起的静电干扰等。

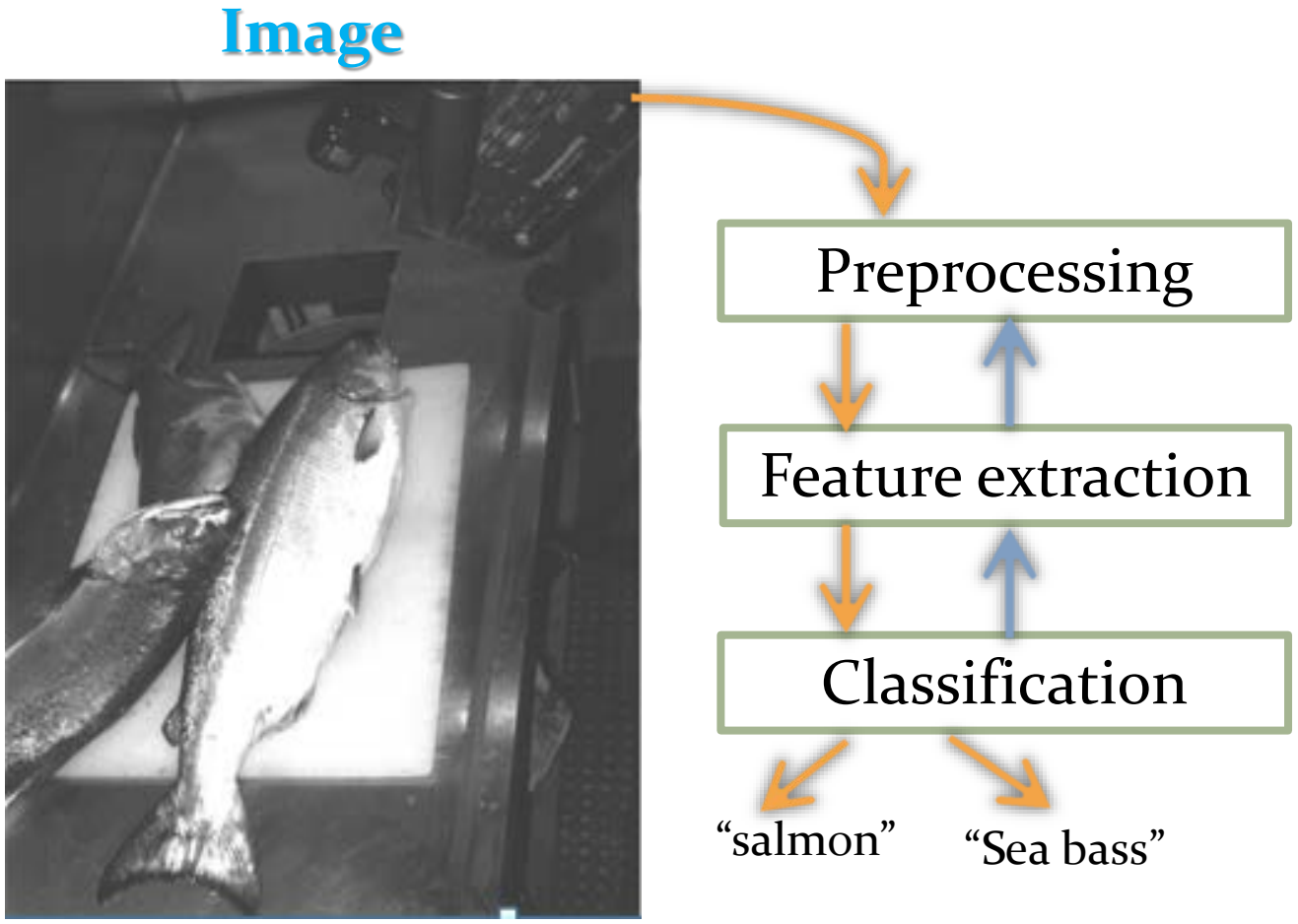
噪声(noise)

一个例子(3) 目标

- 两种确实存在不同的鱼，有不同的模型(**model**)，即可以用数学形式表达的不同特征的描述。
- 本例，也是模式分类的主要目标是

假定这些模型(**model**)的类别(**class**)，通过处理(**process**)采集到的传感器数据，消除其噪声，为感知到的模式(**pattern**)选择最合适的类别。

一个例子(4) 原型分类系统

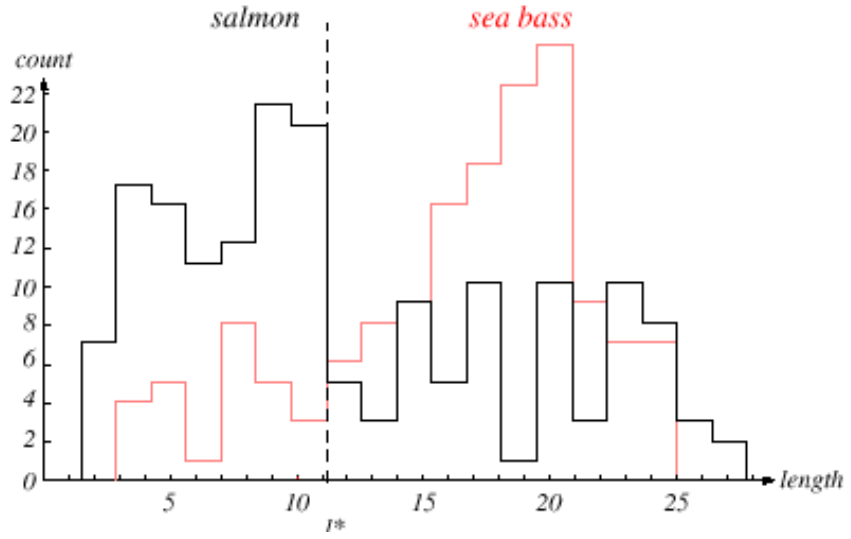


一个例子(5)　数据处理流程

- 摄像机拍摄鱼的照片。
- 图像被**预处理**(preprocessing)，以便于后续操作且不损失关键信息。如，自动调整图像的平均亮度，利用图像分割技术将鱼同背景分开等。
- 将每条鱼的数据送入**特征提取**(feature extraction)器，以便用“特征”来简化原始数据。
- 将特征送入**分类器**(classification)，据此做出最终类别判断。

一个例子(6) 分类器

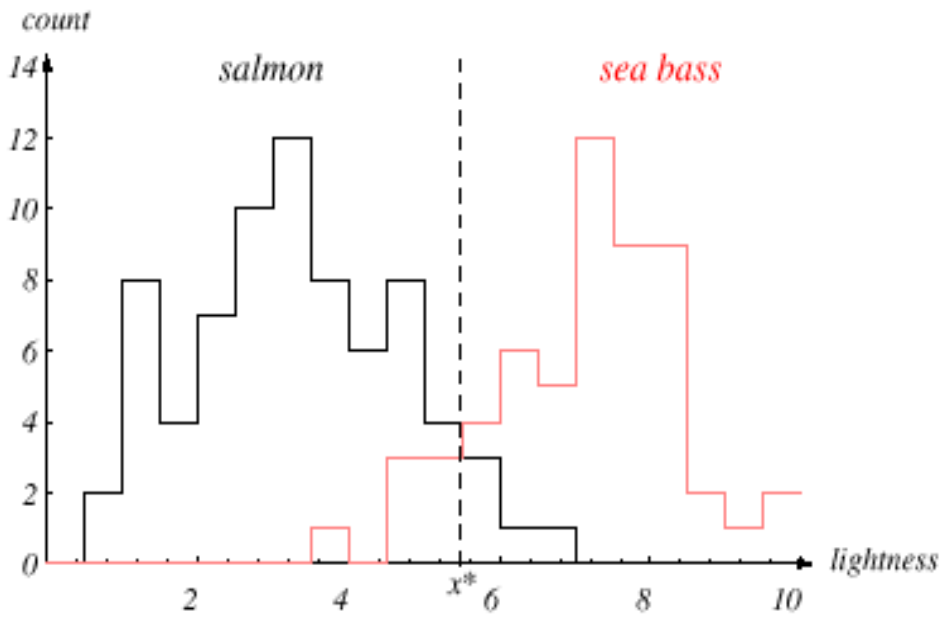
- **利用先验知识(prior knowledge)**
  - 假定被告知“两种鱼长度各异，且鲈鱼的典型长度比鲑鱼的长”，那么可尝试用
    临界值l^*
  - 临界值l^*的选择可通过获取不同类别鱼的训练样本，测量长度，检查测量结果。如图



该直方图表明：无论如何选择l^*，单靠长度都无法将两种鱼可靠地分开。如果只利用长度这一个特征，出现分类错误是不可避免的。

一个例子(7)　分类错误

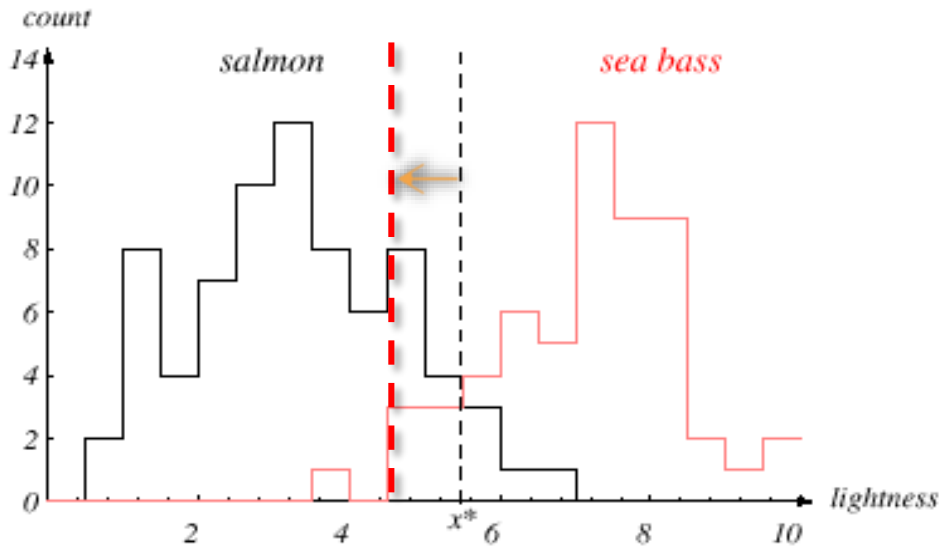
- 尝试其他特征，如平均光泽度(lightness)



该直方图表明：与长度特征相比，依据lightness分类结果好，但还是存在分类错误。

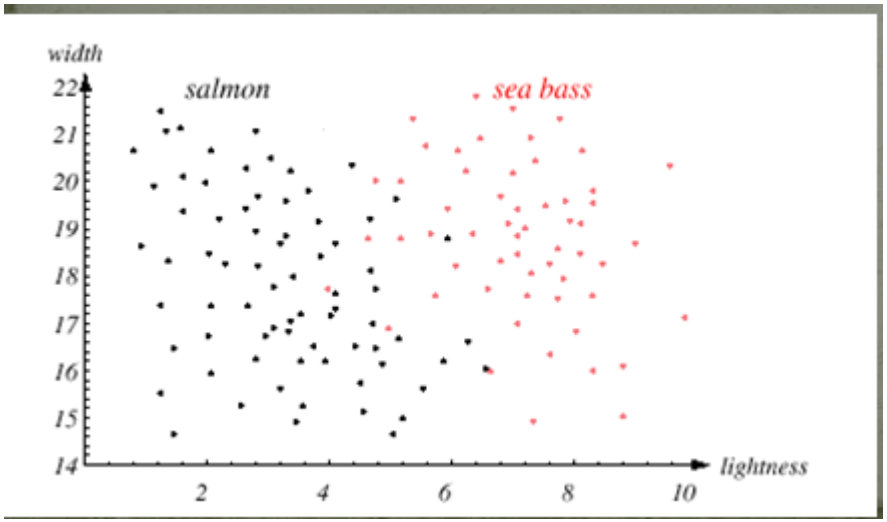
一个例子(8) 误分类的代价

- 代价可能相同，如本例所默认的，应使错分类的鱼数最小(**分类错误率**)，判别边界 x^* 选在两者交界点处。
- 代价也可能不同，如若鲈鱼误判为鲑鱼的代价大于鲑鱼误判为鲈鱼的代价，判别边界 x^* 应向光泽度值小的方向移动，以减少鲈鱼误判为鲑鱼的数目，使误分类的“**总体代价**”最小。



一个例子(9) 特征空间

- 若单一特征分类结果始终不能令人满意，可考虑组合运用多种特征。
- 若采用两个特征：光泽度x_1和宽度x_2，把整条鱼的数据精简为一个二维特征向量(**特征提取**)，即二维**特征空间**中的一个点：$\mathbf{x}=[x_1, x_2]^T$

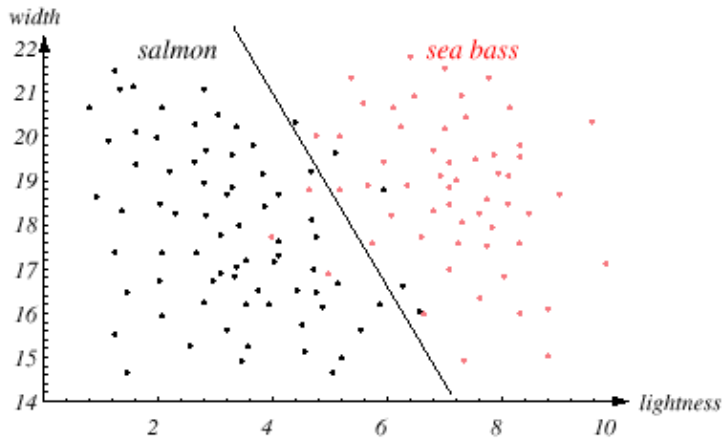


Fish —特征提取→ $\mathbf{x} = [x_1, x_2]^T$

x_1: Lightness　x_2: Width

一个例子(10) 决策边界

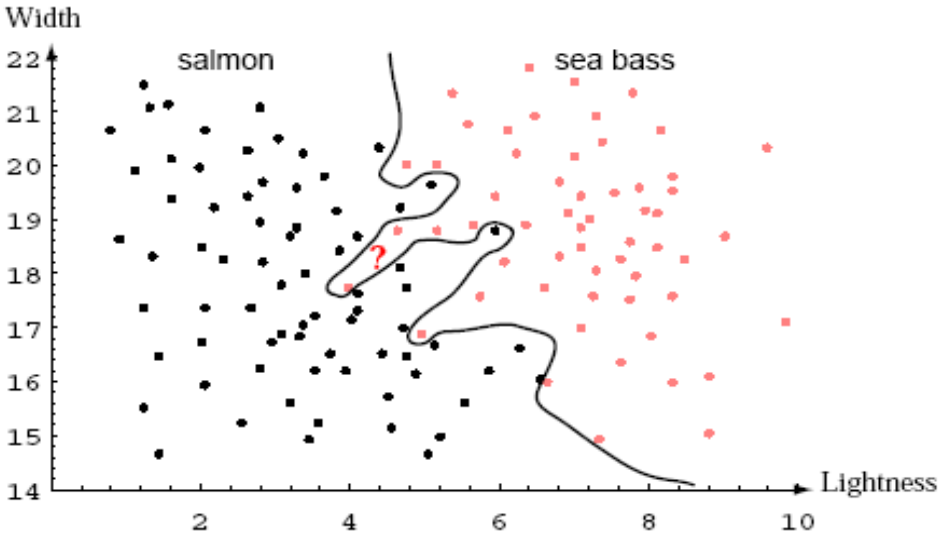
- 问题变成在特征空间中，找一个**决策边界**(decision boundary)，将特征空间划分为两个区域，一个区域中的点为鲈鱼，一个区域中的点是鲑鱼。
- **分类规则**：新鱼的特征向量若在决策边界下，分为salmon；在边界上，归为sea bass。



- 从分类规则来看，似乎特征越多越好？
- 当有多个特征可用时，怎样才能预知哪些特征会更有利于分类？
- 有没有特征是冗余的？

一个例子(11) 泛化

- 假设模型十分复杂，可能得到一个比线性复杂得多的决策边界，如图1.5。这个边界很好地分类了训练样本，但对未来新样本的分类性能却不好。如图中的？，明显它更像是salmon，若基于图中边界分类，将判定为sea bass。
- 理想情况下，我们希望决策边界对未来的新样本有较好的分类性能，即有**泛化(generalization)**能力。

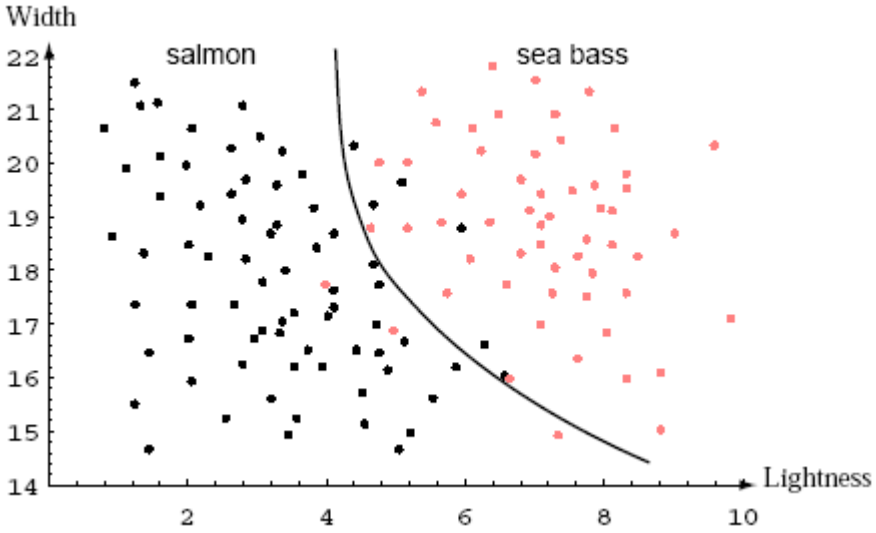


图中的决策边界**过于拟合(Overfitting)**了训练时的样本，而非所有鲑鱼和鲈鱼的真实模型，其泛化能力低下。

一个例子(12) 模型选择

- 一个简单的，满足“即使对训练样本性能稍差，但对未来新样本有较好性能的分类器”。如图1.6。
- Entities are not to be multiplied without necessity

—— William of Occam(1284-1347)奥卡姆剃刀原理



- 怎样才能自动确定一个简单的曲线(如图1.6所示)，它比图1.4的直线和图1.5的复杂边界更可取？

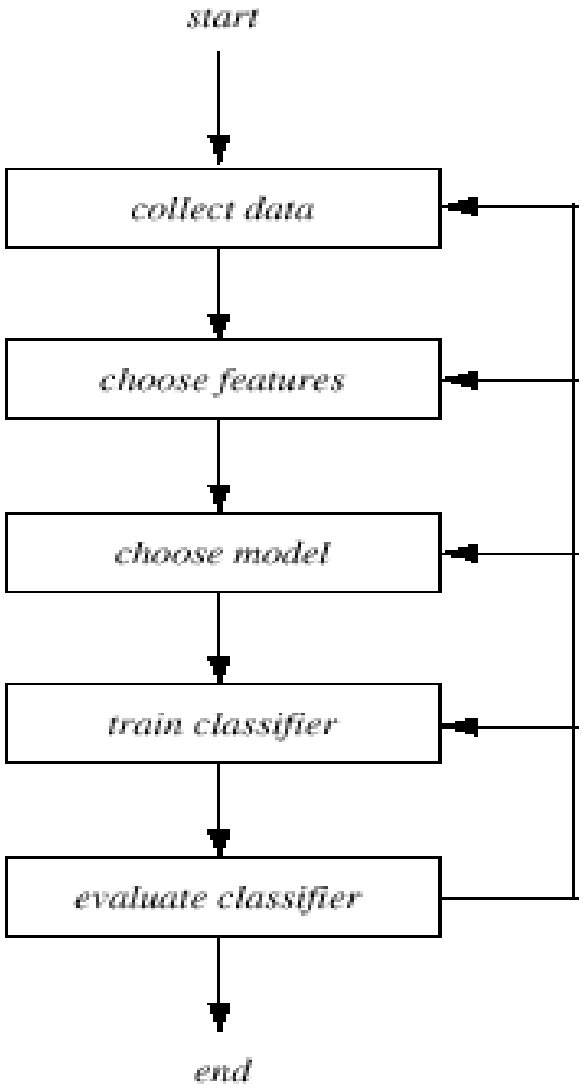
模型选择

- 假设我们能优化地折中，那么如何预知系统对新样本的泛化性能？

结果评价

1.3 设计循环

- 数据采集
- 特征选择或提取
- 模型选择
- 训练分类器
- 结果评估



设计循环中的关键问题(1)

- **特征选择或特征提取**

  - 根据特定的问题领域的性质，选择有明显区分意义的特征，可以把**先验知识**和实验数据有机结合。
  - 从测量集中选出最有利于分类的一组特征。这些特征可以直接选自原始特征集(**特征选择**)，也可以通过对原始特征集的线性或非线性变换得到(**特征提取**)。
  - 将数据处理过程分割成特征提取和分类是人为的，因为特征提取的优化工作通常就是分类器设计过程的一部分。

设计循环中的关键问题(2)

- **模型选择(Model Choice)**

  - 由给定的训练集(一组类别属性已知的样本)，可以构建出多种类型的分类器，如决策树、神经网络、支持向量机或线性判别函数。而对于给定类型的分类器，可采用一种训练算法来对参数空间进行搜索，**找到**最能说明训练集样本的测量值与其类别之间关系的**模型**。
  - 训练集样本是有限的。若分类器的形式过于复杂(如图1.5),则分类器可能会因对训练集中的噪声模拟而引发过度拟合，泛化性能差。而若分类器不够复杂，它又抓不住数据中的结构而导致欠拟合(如图1.4)。**如何选择合适的模型是重要的**。
  - **常用的模型选择方法是：正则化和交叉验证**

设计循环中的关键问题(3)

- **结果评估(Evaluation)**

  - 将训练出的分类器用于独立的有标签样本的测试集，计算测试结果的分类错误率，以便进一步优化分类器设计。

		实际类别		
		ω_1	ω_2	ω_3
预测类别	ω_1	e_{11}	e_{12}	e_{13}
	ω_2	e_{21}	e_{22}	e_{23}
	ω_3	e_{31}	e_{32}	e_{33}

  - 分类准确率 $a = \frac{\sum_i e_{ii}}{\sum_{ij} e_{ij}}$ ，分类错误率：$1 - a$

其中，e_{ij}表示分类器将ω_j类样本预测为ω_i类样本的数量。

1.4 统计模式识别的方法

- 广义上讲，任何将训练集的信息用于分类器设计的方法都是学习(Learning)。
- 在统计模式识别中，主要分为：
  - **有监督学习**(Supervised Learning,或称分类)
  - **无监督学习**(Unsupervised Learning,或称聚类)
  - 半监督学习(semi-supervised Learning)
  - 强化学习(Reinforcement Learning)　无类别标签，只有对错反馈

监督学习

- 从给定的、有限的、有类别标签的训练样本集出发；假设分类器是属于某个类型集合的，称为**假设空间**(hypothesis space)；应用某个**评价准则**(evialuation criterion)，从假设空间中选取最优的模型，使它对已知训练样本及未知测试样本在给定的评价准则下有最优的预测；最优模型的选取由**算法**实现。
- 模型的假设空间(模型学习的集合)、模型选择的准则(学习的策略)以及模型学习的算法，称为统计学习方法的三要素，简称为**模型**(model)、**策略**(strategy)和**算法**(algorithm)。

非监督学习

- **无监督分类**(也称**聚类**)时，数据的类别标识是未知的。无监督分类就是试图找到数据所属的类别以及类间相区别的特征。
- 聚类技术也可用于有监督分类的方案中，作为有监督分类过程中的一种预处理数据的方法。

参考书

- Richard O. Duda 《模式分类》第2版
- Andrew R.Webb 《统计模式识别(第三版)》

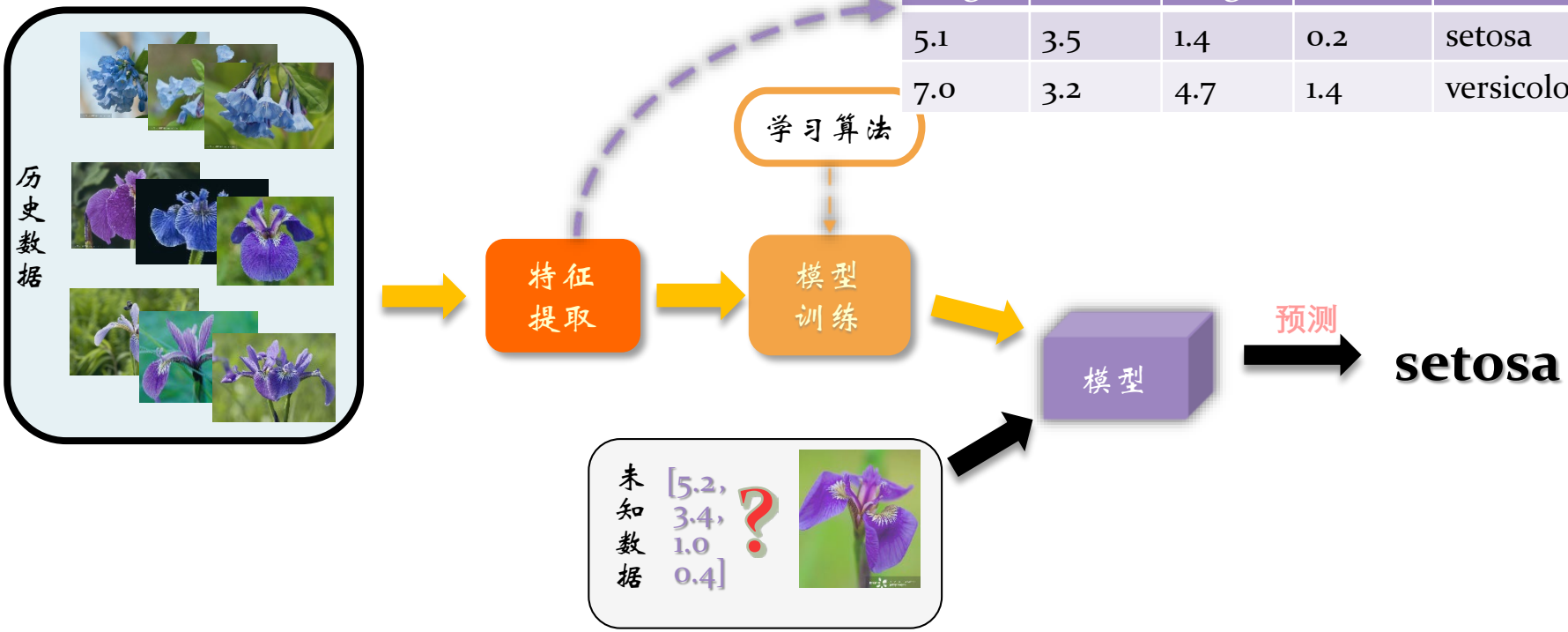
Iris数据集

- http://www.ics.uci.edu/~mlearn/MLRepository.html
- 3个鸢尾花的品种
  - Iris setosa（山鸢尾）
  - Iris virginica（北美鸢尾）
  - Iris versicolor（变色鸢尾）
- 4个特征
  - 萼片长度、萼片宽度
  - 花瓣长度、花瓣宽度

http://106.75.236.166:8888/notebooks/%E6%95%B0%E6%8D%AE%E6%8E%A2%E7%B4%A2/iris.ipynb

模式分类 vs 机器学习

sepal length	sepal width	petal length	petal width	class
5.1	3.5	1.4	0.2	setosa
7.0	3.2	4.7	1.4	versicolor



数据准备 → 特征选择和变换 → 模型训练和测试 → 模型性能评估和优化 → 模型使用