模式识别基本概念

翟婷婷

扬州大学
信息工程（人工智能）学院
zhtt@yzu.edu.cn

2022年春

课程目标

- 掌握模式识别的基础概念：
  - ✓特征与特征空间
  - ✓有监督与无监督学习
  - ✓紧致性和维数灾难
  - ✓泛化能力与过拟合
- 掌握第一个模式识别算法—模板匹配算法

特征与特征空间

- 模式识别系统的目标：寻找一种区分不同类的**模式**的“方法”，使得给定**“新”**的模式，利用找到的分类方法将其分配给最接近的模式类。整个过程能够通过**计算机技术**来实现。

特征向量
(也称实例，instance)

- 通过**特征**提取，将一个抽象的模式转换为计算机能够存储和处理的数据形式。
- 全体模式进行特征提取后得到的特征向量的集合就构成了一个**特征空间。**

特征与特征空间

- 进行特征提取时，希望**同一类**的模式对应的特征向量在特征空间中**尽可能聚在一起**，**不同类**的模式对应的特征向量在特征空间中**尽可能离得远一些**。

- 特征空间中属于同一类模式的特征向量，会聚集在一起，形成了特征空间中的“**模式类**”的概念。

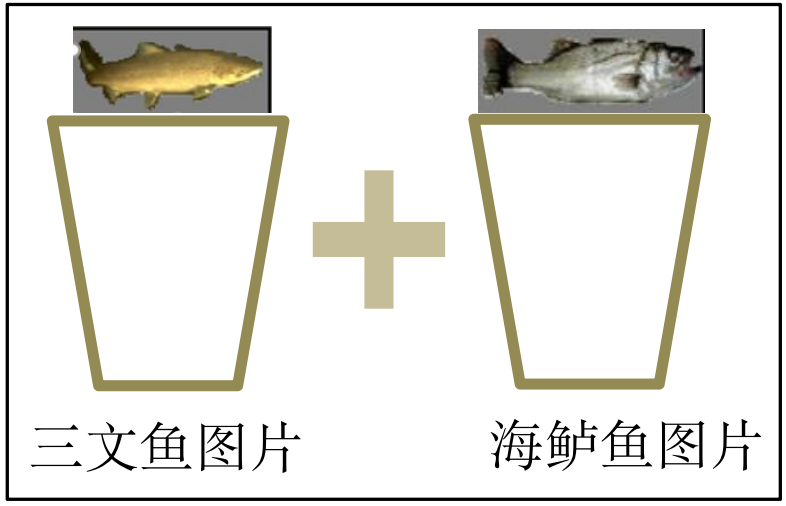
- 模式识别问题转化为：

  给定一个实例或特征向量，在特征空间中如何对其分类的问题。一种简单的分类方法是：看在特征空间中，实例属于哪个类的聚集范围，或者与哪个类的众多实例相似度更高，就可以把它归到这一类之中。

特征与特征空间

- 最常见的特征空间是向量空间：将一个特征向量看成是特征空间中的一个点。
- 在上节课的根据鱼的图片对**鱼分类**的例子中：
  用两个特征来描述一条鱼：
  1.鱼鳞的平均亮度 2. 鱼身的宽度

将一条鱼表示为一个点$\boldsymbol{x} = (x_1, x_2)$ → 特征向量



测量每条鱼的**鱼鳞亮度**和**鱼身宽度**

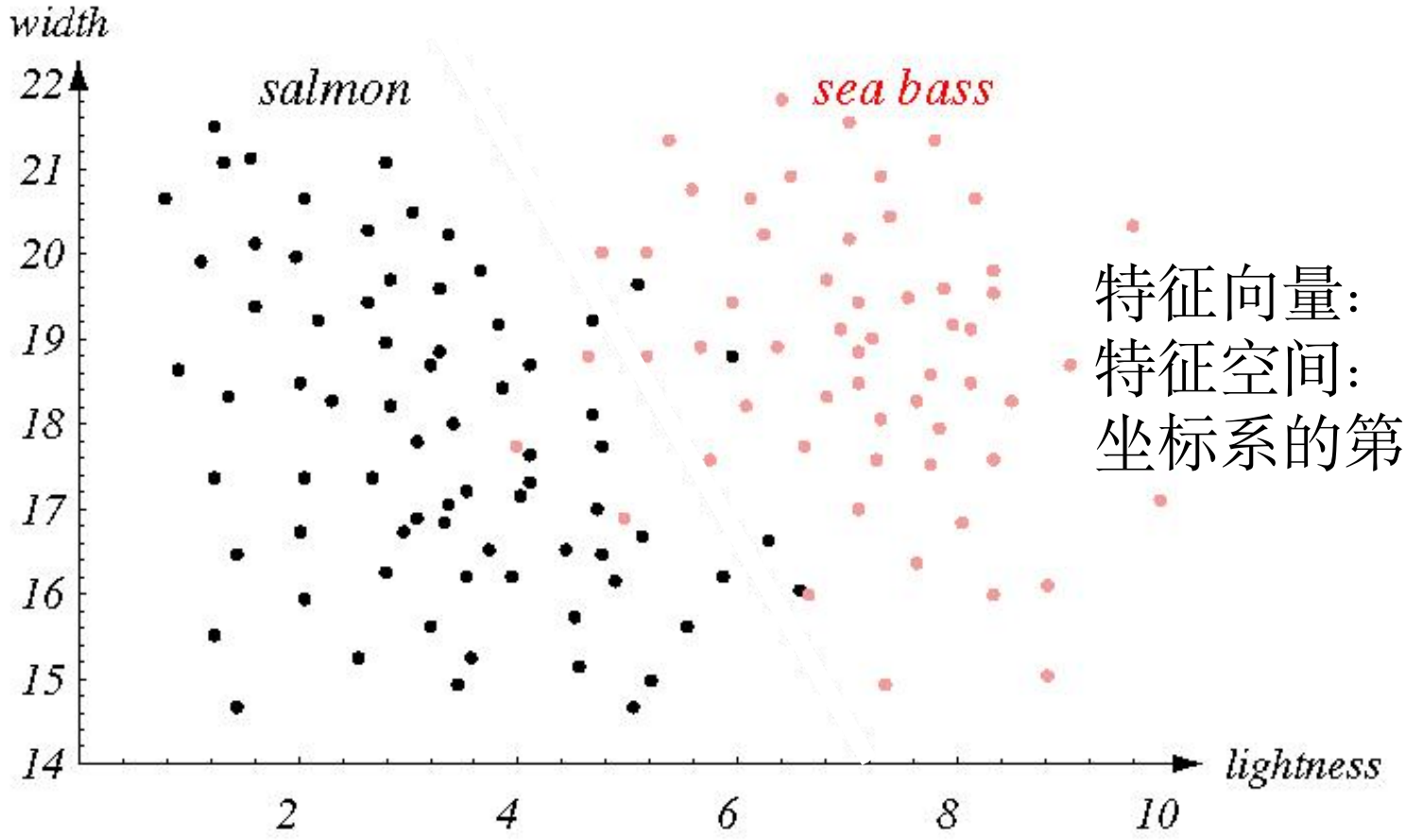
特征提取

训练数据集

5,	18.5,	1
5.5,	16,	1
6.5,	19,	1
3.2,	17,	1
4,	18.5,	1
。	。	。
8,	15,	2
6,	17,	2
7.6,	21,	2
8.4,	18,	2

1: 三文鱼
2: 海鲈鱼

特征与特征空间



特征向量：点
特征空间：
坐标系的第一象限

有监督学习与无监督学习

- 模式识别核心：**分类器** (**分类算法**)的设计和训练。
- **分类器的训练/学习：**给定一个训练数据集，确定好要使用哪种分类算法后，分类算法通过自身的自主学习去找到最好的分类器的超参数的过程，称为分类器的训练/学习。
- 分类器训练的两种不同方式：
  - ✓有监督学习和无监督学习

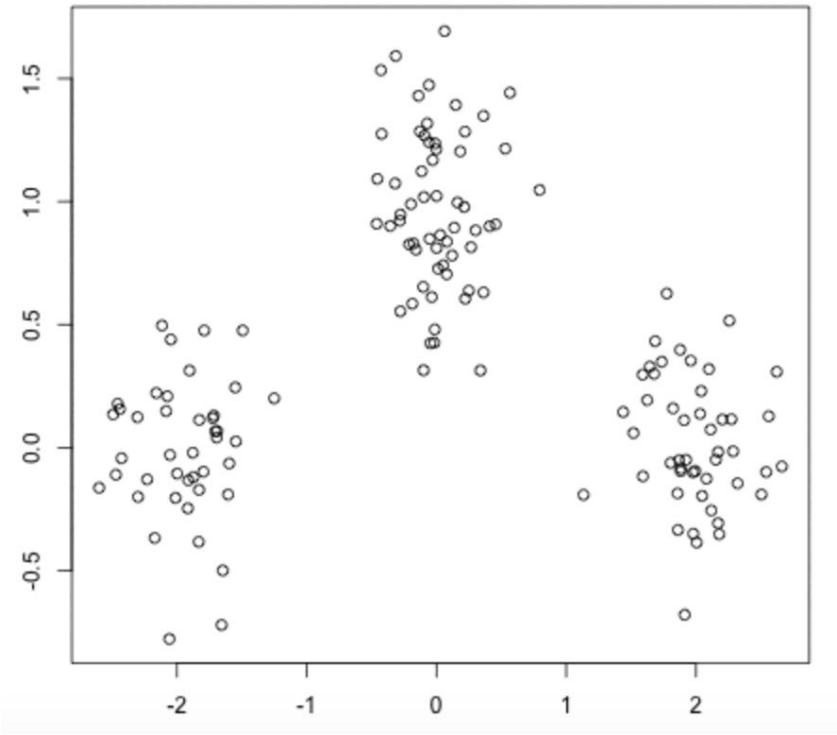
有监督学习与无监督学习

- 有监督学习：
  - ✓ 每个实例都有一个**类别标签** (类标，label)，表示该实例属于哪个模式类。（实例 + 类标签 = 样本）
  - ✓ **训练数据集是由带类标签的实例组成的集合**。
  - ✓ 分类器学习算法通过分析训练数据集，寻找属于同一类的实例具有哪些相似性，不同类的实例具有哪些差异性，从而学习到具体的分类决策规则。
- 有监督学习中，**类标签**往往是在模式采集阶段通过人对每个模式进行手工标注得到的。

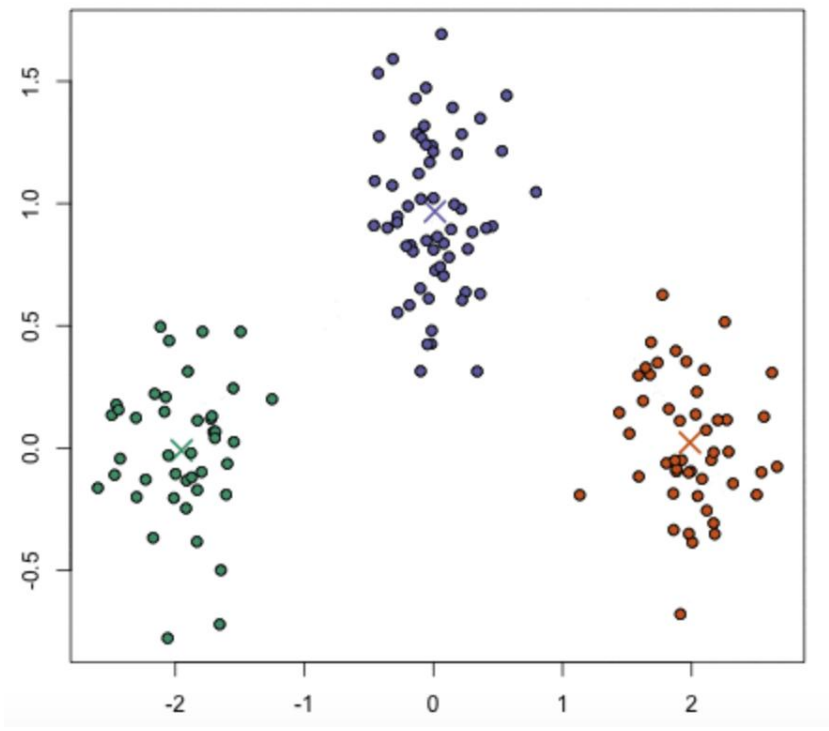
有监督学习与无监督学习

- 无监督学习：
  - ✓ 训练数据集的每个实例**没有类标签**。
  - ✓ 需要定义实例间的相似度，按照一定的规则，把相似程度高的实例划分为同一类，把相似度低的实例划分为不同的类，从而将训练数据集的实例划分成不同的类别。
  - ✓ 一个未标记的实例与训练集中哪类实例的相似程度最高，就把它归到哪一类中。
- 无监督学习主要适用于类别标注困难或类别标注的成本太高的情景中。

无监督学习



左：没有类标签的训练数据集

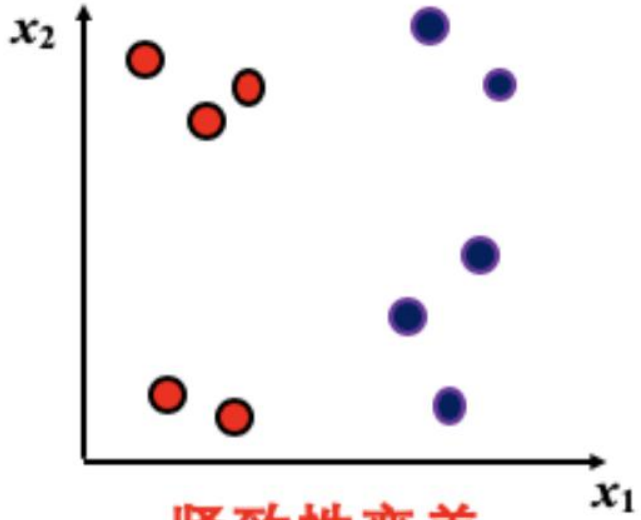


右：定义相似度函数为欧式距离
根据相似度，将实例集划分为3个类

数据的紧致性和维数灾难

- 特征空间中实例的紧致性：同类实例之间的相似程度与不同类实例间的差异程度。
- 紧致性好的数据集满足：同类实例之间的相似度远大于不同类实例之间的相似度。
- 在紧致性好的数据集上训练分类器，得到的分类器的性能也更好。

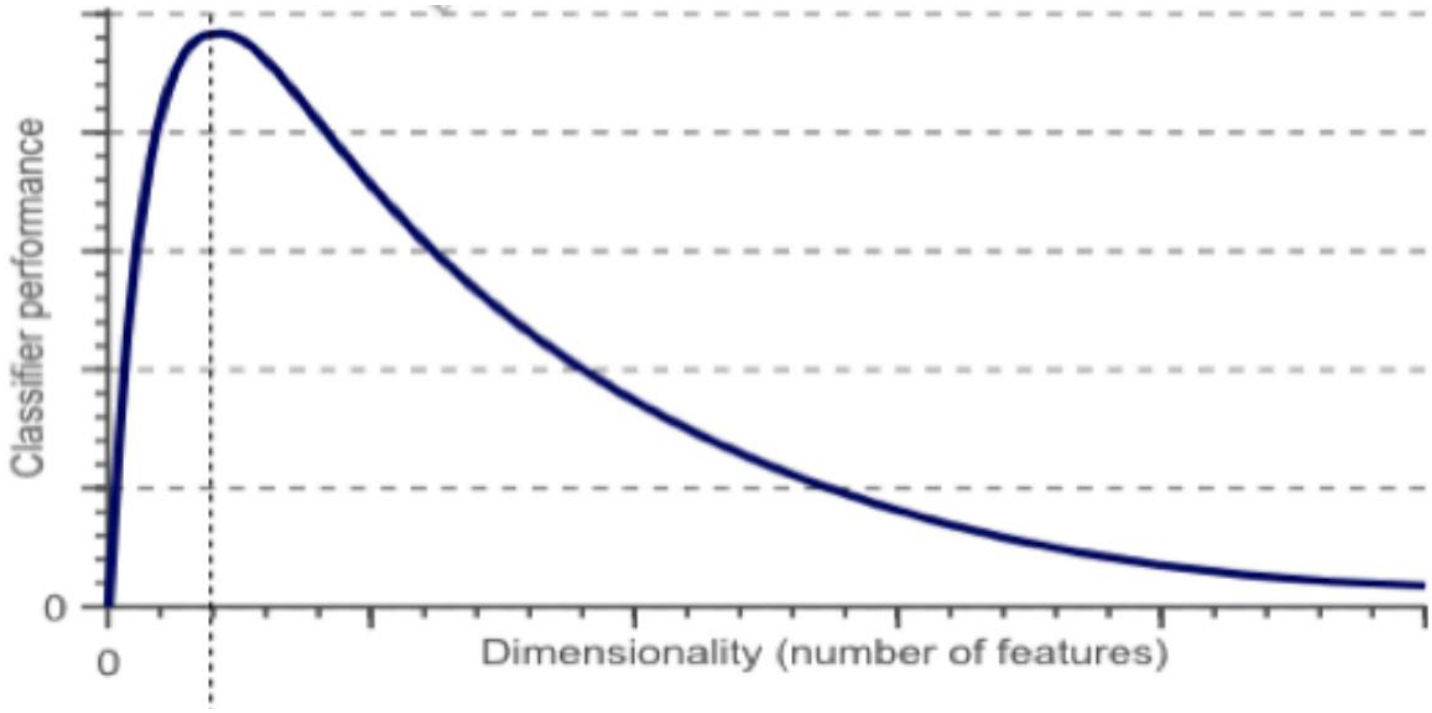
紧致性好



紧致性变差

维数灾难问题 (Curse of Dimensionality)

- **维数灾难**指的是随着特征维数的增加，分类器的性能首先会快速增加，然后不断下降，最终导致训练分类器所需的时间虽显著增加但分类性能却严重退化的问题。



维数灾难问题 (Curse of Dimensionality)

- 导致维数灾难的**根本原因**是训练集中样本数量不足。
- 当特征维数增加时，在特征空间中以同样密度能够容纳的样本总数呈指数级增长，而如果给定的数据集中的样本数量没有同步按照指数规律增加，那么在这样高维的特征空间中，数据集中样本分布地就很稀疏，从而使得样本集的紧致性变差，因此分类器的性能也就变差。

特征维数越高 ➡ 样本集越稀疏 ➡ 紧致性越差 ➡ 分类器性能越差

维数灾难问题 (Curse of Dimensionality)

- 解决维数灾难的办法：

1. 同步地大量增加训练集中样本的数量。

   这一办法在实际问题中难以实现，其一是模式采集耗时耗力(尤其对模式进行标注)，其二是样本数量太大会导致分类器训练和使用都耗费太多时间和空间。

2. 尽可能地减少所使用的特征维度。

   尽可能提升每个维度在分类中的效能，从而使模式识别问题在较低维度下得到更好地解决。

泛化能力与过拟合

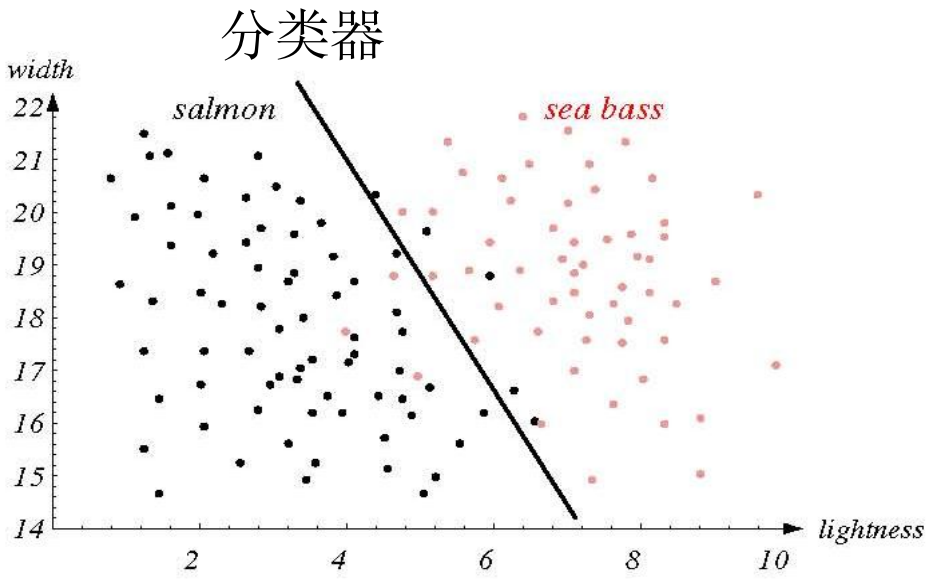
- 在训练数据集上训练分类器时，我们希望训练得到的分类器能捕获到**全体样本中**的**一般/普遍规律**，而不是**部分样本中**才有的**特殊规律**。
- 注意：搜集到的训练集只是全体数据中的一个子集。有可能这个集合并不全面，或者存在误差。
- 如果要求训练出的分类器仅仅能对训练集中所有样本正确分类，就可能导致分类器仅能捕获到训练数据集中样本中的规律，而无法捕获全体样本中的一般规律，从而导致分类器在对不在训练集中的新实例进行分类时效果不佳。

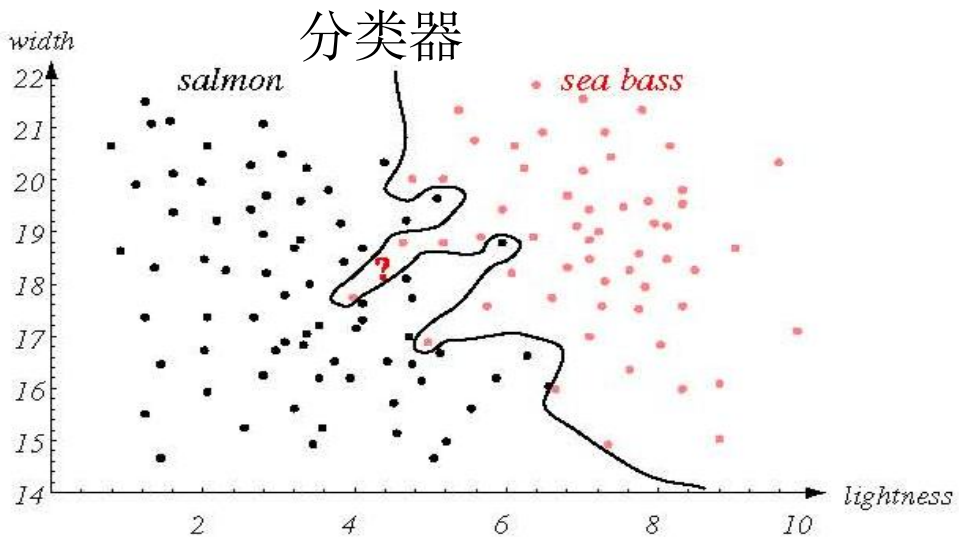
泛化能力与过拟合

- 我们希望训练得到的分类器，①不仅能对训练集中的实例正确分类，②而且对于**不在训练集中的新实例**，也能正确分类。
- 训练好的分类器**对不在训练集中的新实例**正确分类的能力，称为**“泛化能力”**。
- 泛化能力一般难以直接度量，通常会构建一个测试数据集，用分类器在测试数据集上的分类性能近似表示其泛化能力。

泛化能力与过拟合

- 由于过分追求对训练集中实例的正确分类，导致分类器的泛化能力降低，称为分类器训练过程中的“**过拟合**”。

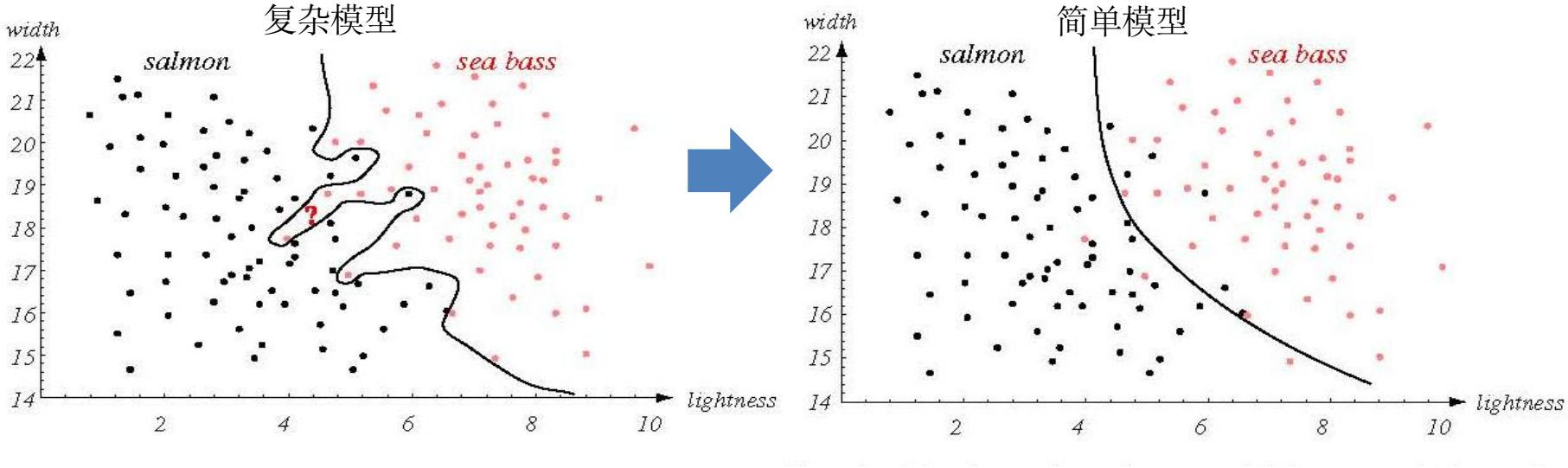




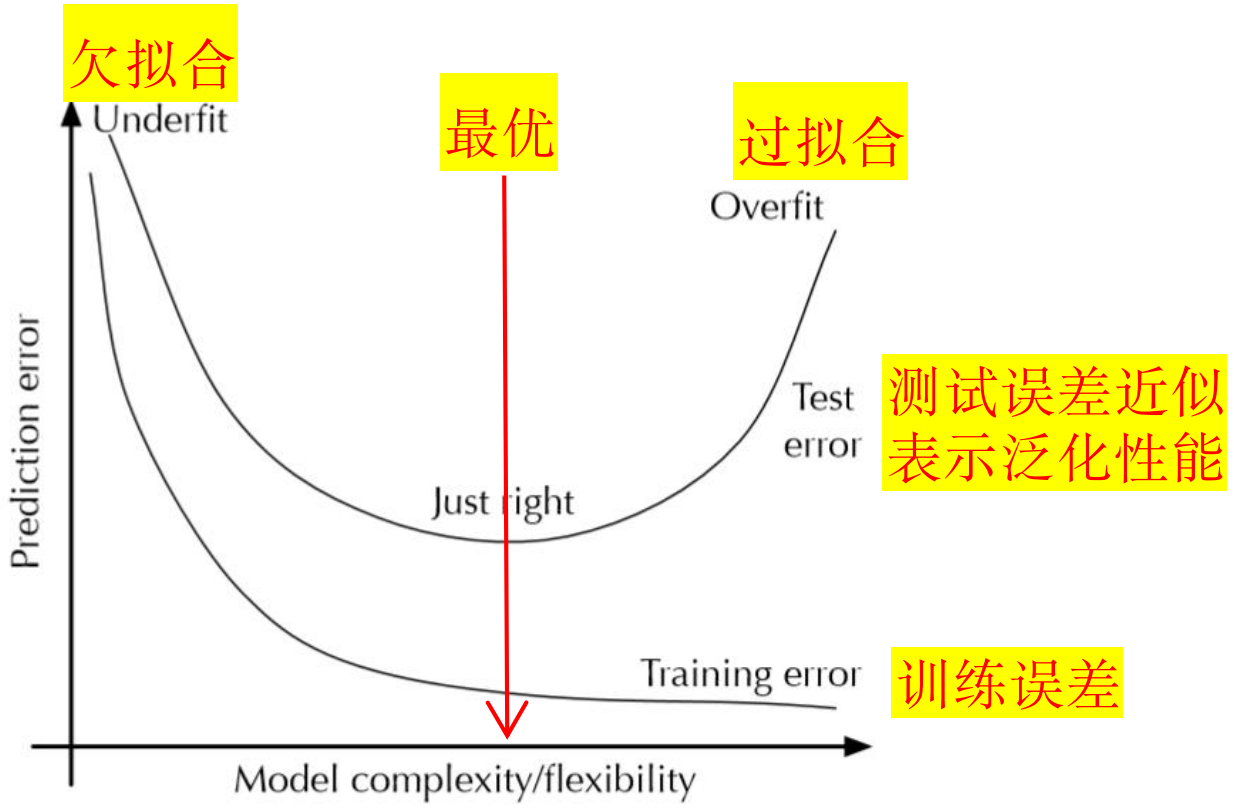
右图中的分类器虽在可以训练数据集上获得完美的分类准确率，但是该分类器的泛化性能比左图中分类器更低。

泛化能力与过拟合

- 如何提高分类器的泛化性能?
  - ✓ 构建更好的训练数据集 (数量多、多样性好、无噪声、类别分布均衡)，好的训练集意味着能更好地训练分类器。
  - ✓ 在训练数据集不变的情况下，采用简单的分类器模型。简单的模型通常会产生更好的泛化性能。



分类模型的泛化性与复杂度的关系



训练集不变，变化分类模型的复杂度

模式识别应用举例—手写数字识别系统

- 手写数字识别问题

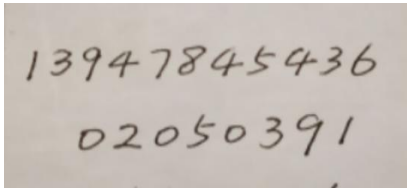
✓ **问题背景**：手写数字在我们日常生活中并不少见：信封上填写的邮政编码，表格中手写的日期、编号、电话号码，支票上的账号和付款金额。如果能够使用计算机自动识别出手写的数字并将其转化为标准的数字代码，将为许多业务的自动处理奠定基础。

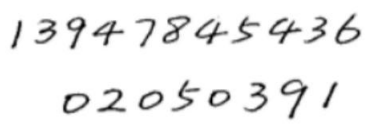
✓ **问题难度**：数字虽然只有10个，但每个人的书写风格千差万别，所用的笔的颜色、粗细、软硬等也各不相同，还有书写位置、朝向、背景的种种干扰，使得手写数字的识别并不容易。

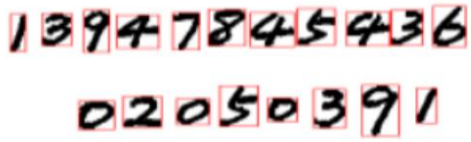
手写数字识别系统

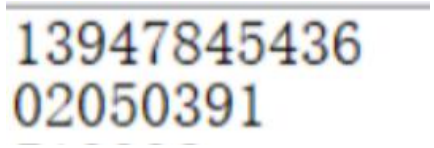
- 手写数字识别问题

✓ **问题描述**：手写数字识别是一个典型的**多分类**问题，输入的是一张包含手写数字的图片，输出的是图片中每个手写数字对应的数字代码，包括 0~9这10个数字代码。









原始图片 ⟶ 二值化图片 ⟶ 字符分割 ⟶ 字符识别

连串手写数字识别问题　最终转化为　**单个**手写数字识别问题

手写数字识别系统

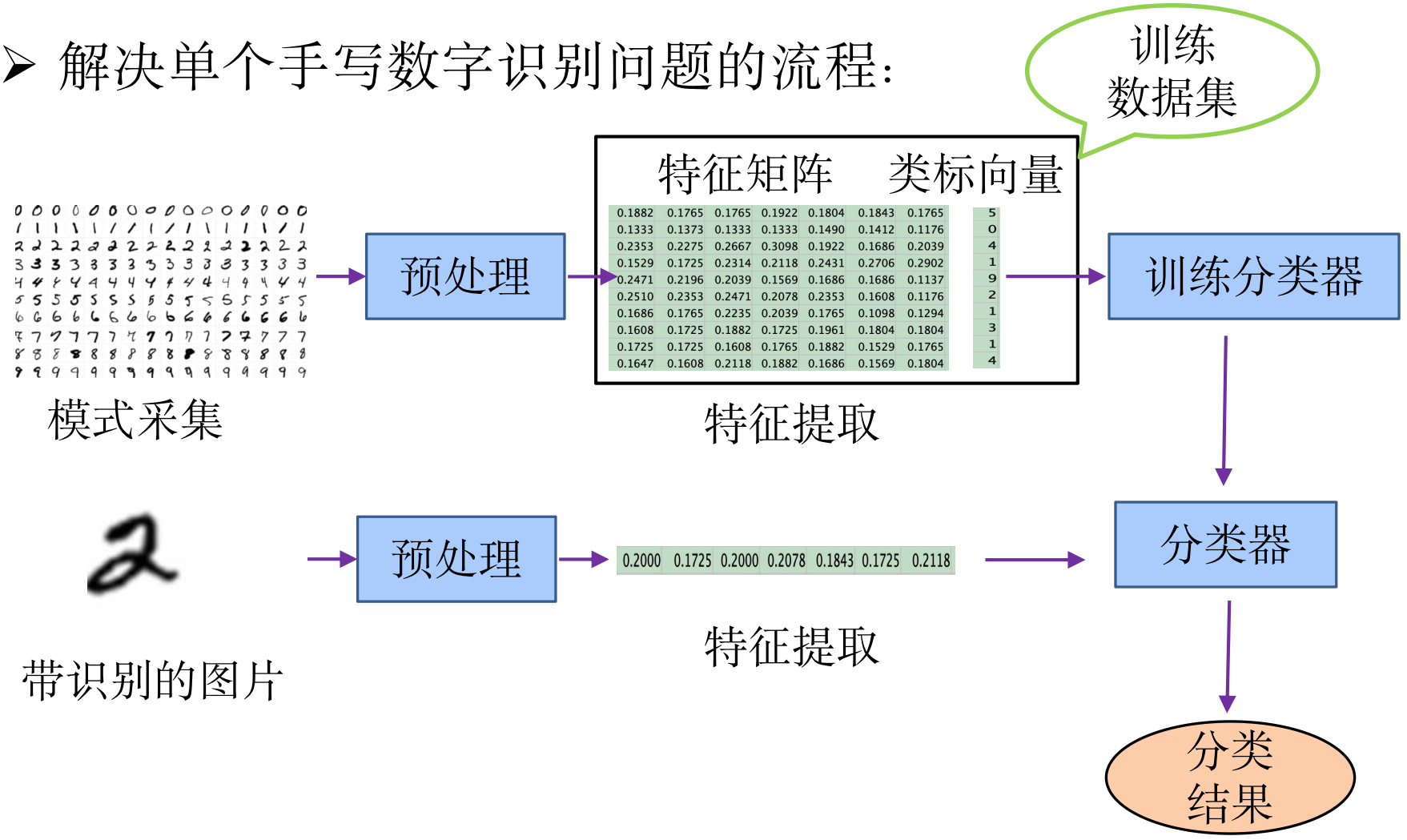
- 单个手写数字识别问题

✓ **问题描述**：输入一张包含单个手写数字的图片，输出图片中的数字代码。

✓ 如何解决这个问题呢？需要构建一个完整的模式识别系统，按照模式识别系统的设计流程来完成此任务。

手写数字识别系统

- 解决单个手写数字识别问题的流程：



手写数字识别系统

- 解决单个手写数字识别问题的流程：

1. 搜集大量单个手写数字的图片。
2. 所有图片尺寸归一化为相同的尺寸，例如28*28，然后对图片进行去燥、增强等处理，以突出图片中有关手写数字的信息。
3. 依次对每一张图片提取特征，可以提取轮廓、颜色分布、关键点等特征，也可以直接将图片的像素矩阵展开为 1 维向量作为特征。然后，要为每张图片的特征向量加上类标，表示该特征向量属于哪一类。
4. 训练分类器，从而得到训练好的分类器。
5. 分类/识别：运用分类器对任意一张待识别图片进行分类。

手写数字识别系统

- 训练数据集建立的方法：

1. 自己通过模式采集、预处理、特征提取的过程创建。
2. 直接使用公开的数据集，这些数据集是别的人/团队/机构搜集的，并做好初步预处理的数据。

- 手写数字识别领域著名的公开数据集：MNIST

✓ MNIST 数据集是美国国家标准与技术研究所搜集的，由来自 250 个不同人手写的数字组成，其中 50% 是高中学生，50% 来自人口普查局的工作人员。
✓ 该数据集已经成为机器学习领域的典范。

车牌识别系统中的数字识别

- 车牌识别问题

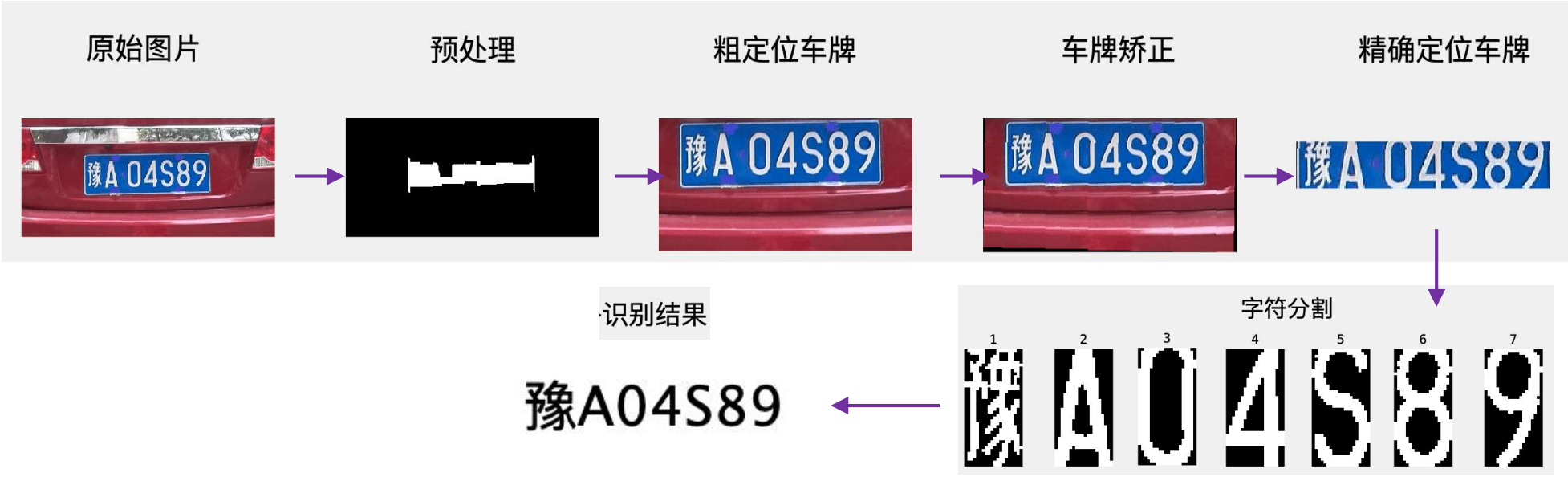
✓ **问题背景**：车牌识别是智能交通系统的重要组成部分，通过智能车牌识别系统，可以实现对车辆的各种行为进行自动化地监控以及管理。车牌识别旨在识别出一张车牌图片中所包含的所有字符。

- 车牌识别包括：汉字识别、字母识别和数字识别

汉字识别　　字母识别　　包含字母和数字的识别

车牌识别系统中的数字识别

- 车牌识别流程：



- 车牌识别中的数字识别问题

✓ **问题难度**：车牌数字的图片相较于手写数组图片更单一化，不像手写数字那样具有不同书写风格带来的差异。

第一个模式识别算法

- 模板匹配算法是最直接、简单、历史最悠久的模式识别方法。该方法在类别特征稳定、明显，类间差距大的时候可以使用。它在建立模板的时候需要依赖于人的经验和观察，适应能力会比较差。

- 模板匹配法的**基本原理**是：为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模式与每个类别的模板进行比对，根据与模板的匹配程度将模式划分到与其最相似的类别中。

第一个模式识别算法

- 采用模板匹配实现车牌数字识别：

1. 首先要给每个类别（即每个数字）确定一个模板。
2. 计算待识别图片与模板图片的相似度大小，将其划分到最相似的模板所对应的类别中。

- 计算相似度的方法：计算待识别的图片的特征向量与模板图片的特征向量之间的距离。有多种形式的距离，最常用的是欧式距离。
- 建立一个好的模板库非常重要。显然，对于车牌数字，容易建立一个模板库，但对于手写数字，则不容易，因为一个数字对应的手写数字图片可能差别比较大。

第一个模式识别算法

- 模板匹配实现车牌数字识别：

小结

- 特征与特征空间
- 有监督与无监督学习
- 紧致性和维数灾难
- 泛化能力与过拟合
- 第一个模式识别问题和第一个模式识别算法