

# 第一讲 概述

南开大学网络空间安全学院 计算机学院

## 个人信息

- ❖ 讲授教师：白刚
- ❖ 办公地点：计算机学院564室
- ❖ 接待时间：每周三11:30~12:30
- ❖ 电子邮箱：[baigang@nankai.edu.cn](mailto:baigang@nankai.edu.cn)
- ❖ 信息发布：<http://cyber.nankai.edu.cn>
  - ◆ 授课讲稿
  - ◆ 课后作业
  - ◆ 阅读文献

南开大学网络空间安全学院 计算机学院

2

## 课程信息

- ❖ 教学相长：教与学的“交互与协同”
  - ◆ 《礼记·学记》：学然后知不足，教然后知困。知不足，然后能自反也；知困，然后能自强也。故曰教学相长也。
- ❖ 学问：学与问的“化学反应”
- ❖ 前导课程
  - ◆ 高等数学
  - ◆ 线性代数
  - ◆ 概率与统计
  - ◆ 程序设计
  - ◆ 算法

南开大学网络空间安全学院 计算机学院

3

## 课程成绩

- ❖ 课程成绩：课程作业(60%) + 课程项目(40%)
- ❖ 课程作业
  - ◆ 问题描述
  - ◆ 基本思路
  - ◆ 算法
  - ◆ 结果与分析
- ❖ 课程项目
  - ◆ 课堂报告
  - ◆ 研究报告
  - ◆ 研究成果

南开大学网络空间安全学院 计算机学院

4

## 课程作业

- ❖ 问题描述: 10%
- ❖ 基本思路: 15%
- ❖ 算法: 35%
  - ◆ 算法描述
  - ◆ 算法实现 (源代码)
- ❖ 结果与分析: 40%
  - ◆ 实验步骤
  - ◆ 实验结果
  - ◆ 结果分析

## 课程作业

- ❖ 独立完成, 鼓励讨论, **严禁抄袭**。
- ❖ 按时完成且提交, 迟交作业按零分计算。
- ❖ 文档格式: DOCX 或 PDF
- ❖ 编程语言: python
- ❖ 提交文件名称: 学号(作业序号), 如: 123456(1)
- ❖ 提交文件格式: RAR 压缩文件格式

## 课程项目

- ❖ 课堂报告：20%
  - ◆ 总体思路
  - ◆ 特征描述
  - ◆ 模型描述
  - ◆ 实验设计
- ❖ 研究报告：60%
  - ◆ 问题提出
  - ◆ 现状及分析
  - ◆ 主要原理
  - ◆ 算法描述
  - ◆ 实验结果与分析
- ❖ 项目成果：20%

南开大学网络空间安全学院 计算机学院

7

## 课程项目

- ❖ 按时完成并提交全部内容
- ❖ 提交文档格式：DOCX 或 PDF
- ❖ 编程语言：python
- ❖ 提交文件名称：学号1+学号2+学号3
- ❖ 提交文件格式：RAR 压缩文件格式
- ❖ 加分因素：公开发表
- ❖ 完成形式：2~5人小组（具体人数视选课人数确定）

南开大学网络空间安全学院 计算机学院

8

## 参考书目

- ❖ Richard O. Duba, Peter E. Hart and David G. Stork, ***Pattern Classification***, 2<sup>nd</sup> Edition, John Wiley, 2001.
- ❖ Sergios Theodoridis and Konstantinos Koutroumbas, ***Pattern Recognition***, 4<sup>th</sup> Edition, Elsevier Science, 2009.

## 学术期刊

- ❖ Artificial Intelligence
- ❖ IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI)
- ❖ Journal of Machine Learning Research
- ❖ International Journal of Computer Vision
- ❖ Pattern Recognition
- ❖ 中国图形图像学报
- ❖ 模式识别与人工智能
- ❖ 自动化学报

## 学术会议

- ❖ IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
- ❖ IEEE International Conference on Computer Vision (ICCV)
- ❖ International Conference on Machine Learning (ICML)
- ❖ International Conference on Pattern Recognition (ICPR)
- ❖ Annual Conference on Neural Information Processing Systems (NeurIPS)
- ❖ European Conference on Computer Vision (ECCV)

**Any question?**

## 什么是模式识别?

- ❖ **Pattern recognition** is the scientific discipline whose goal is the classification of *object* into a number of *categories* or *classes*. — Sergios Theodoridis
- ❖ The assignment of a *physical object or event* to one of several pre-specified *categories*. — Duba and Hart
- ❖ A problem of estimating density function in a high-dimensional *space* and dividing the space into the regions of *categories* or *classes*. — Fukunaga

## 什么是模式识别?

- ❖ Given some examples of *complex signals* and the correct *decisions* for them, make decisions automatically for a stream of future examples. — Ripley
- ❖ The process of giving *names*  $\omega$  to *observations*  $x$ . — Schurmann
- ❖ Pattern Recognition is concerned with answering the question "*What is this?*" — Morse
- ❖ The science that concerns the *description* or *classification* (recognition) of *measurements*. — Schalkoff

## 模式识别

- ❖ 求解:  $y = f(x)$ 
  - ◆ 特征空间
  - ◆ 分类器模型
  - ◆ 性能评估
- ❖ 概念: 分类与回归
  - ◆ 分类: 输出为 “有限数目的离散类别”
  - ◆ 回归: 输出为 “一个或多个连续变量”

## 主要应用

- ❖ 视觉目标的检测与识别(Detection and Recognition)
- ❖ 字符识别(Character recognition)
- ❖ 计算机辅助诊断(Computer-aided diagnosis)
- ❖ 语音识别(Speech recognition)
- ❖ 自然语言处理(Natural Language Processing)
- ❖ 数据挖掘与知识发现(Data Mining and Knowledge discovery)



## 主要应用

Problem Domain	Application	Input Patterns	Pattern Classes
Bioinformatics	Sequence analysis	DNA/Protein sequence	Known types of genes
Data mining	Searching for meaningful patterns	Points in multi-dimensional space	Compact and well-separated clusters
Document classification	Internet search	Text document	Semantic categories
Document image analysis	Reading machine for the blind	Document image	Alphanumeric characters, words
Industrial automation	Printed circuit board inspection	Intensity or range image	Defective / non-defective nature of product
Multimedia database retrieval	Internet search	Video clip	Video genres
Biometric recognition	Personal identification	Face, iris, fingerprint	Authorized users for access control
Remote sensing	Forecasting crop yield	Multispectral image	Lands use categories, growth pattern of crops
Speech recognition	Telephone directory enquiry without operator assistance	Speech waveform	Spoken words
Information filtering	Spam detection	E-mail message	Spam/non-spam
Computer security	Intrusion detection system	Network traffic	Intrusive/normal

Source: PAMI Vol. 22, No. 1, pp. 4-37, 2000

南开大学网络空间安全学院 计算机学院

17

## 方法与系统

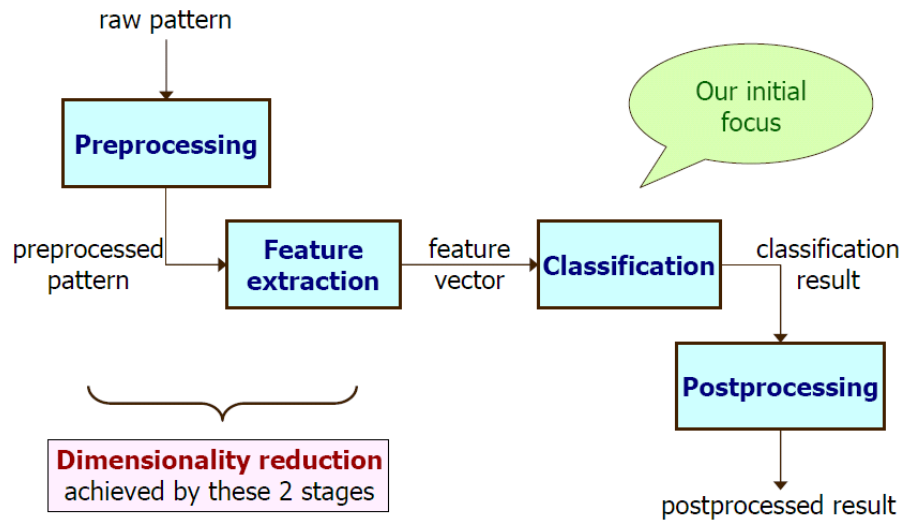
- ❖ 方法
  - ◆ 特征空间
  - ◆ 分类器模型
- ❖ 系统
  - ◆ 信号采集
  - ◆ 预处理
  - ◆ 抽取特征
  - ◆ 分类器输出
  - ◆ 后处理
  - ◆ 决策动作



南开大学网络空间安全学院 计算机学院

18

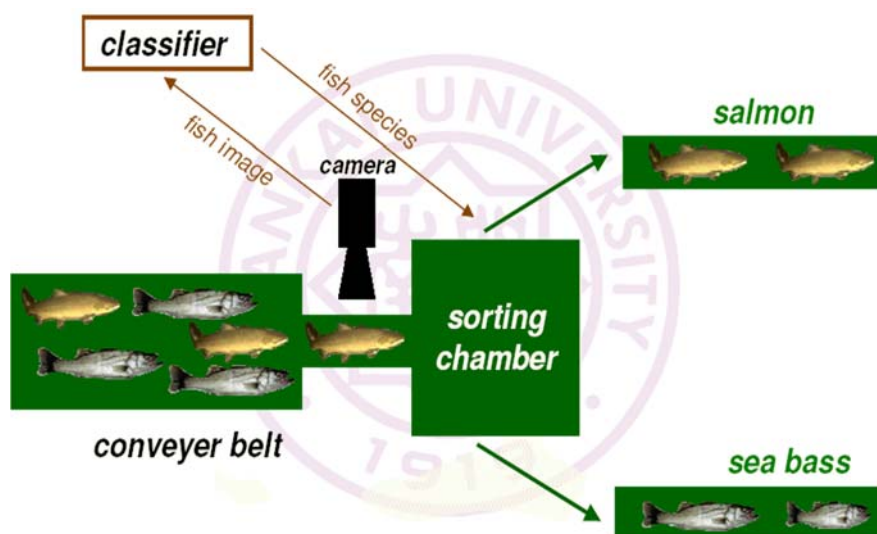
## 模式识别系统结构



南开大学网络空间安全学院 计算机学院

19

## 模式识别系统设计



南开大学网络空间安全学院 计算机学院

20

## 模式识别系统设计

- ❖ 问题：基于数字图像的鱼类分检
- ❖ 方案：
  - ◆ 图像预处理：对原始图像数据进行预处理
  - ◆ 图像分割：从输入图像中分离出每条鱼的图像
  - ◆ 特征抽取：从每条鱼的图像中抽取模式特征
  - ◆ 特征分类：根据模式特征确定每条鱼所属的类别
  - ◆ 鱼类分检：根据每条鱼的类别控制分检装置



南开大学网络空间安全学院 计算机学院

21

## 设计分类模型

- ❖ 收集数据：获得包含待分类模式的图像
 

salmon
sea bass
salmon
salmon
sea bass
sea bass


- ❖ 预处理：从背景中分离出每条鱼的图像
 
- ❖ 抽取特征：从每条鱼的图像中抽取描述不同鱼种间**差异的模式特征**，如长度、亮度、宽度、鱼翅数目等。
- ❖ 标记数据：人工标记每条鱼模式的类别标签，获得训练样本数据集。
- ❖ 设计分类器：选择分类器模型，训练分类器模型。
- ❖ 评估分类器：测试分类器模型，分析错误原因，提出改进方法。

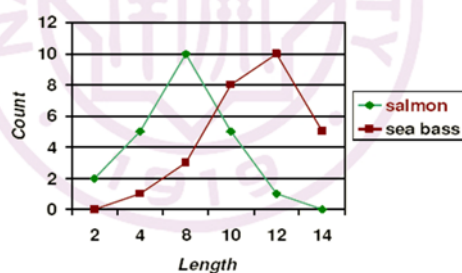
南开大学网络空间安全学院 计算机学院

22

## 选择模式特征

- ❖ 观察：鲑鱼(salmon)长度一般比鲈鱼(sea bass)长度短
- ❖ 模式特征：鱼身长度
- ❖ 直方图：统计每种长度下鲑鱼和鲈鱼的数目

	2	4	8	10	12	14
bass	0	1	3	8	10	5
salmon	2	5	10	5	1	0



南开大学网络空间安全学院 计算机学院

23

## 确定决策阈值

- ❖ 寻找最佳分类决策的长度阈值  $L$

if  $length_i < L$  then  $i \in \text{salmon}$   
 else  $i \in \text{sea bass}$

- ❖ 决策：

- ◆ 假设：取  $L = 5$  时，错误分类为：鲈鱼=1，鲑鱼=16

	2	4	8	10	12	14
bass	0	1	3	8	10	5
salmon	2	5	10	5	1	0

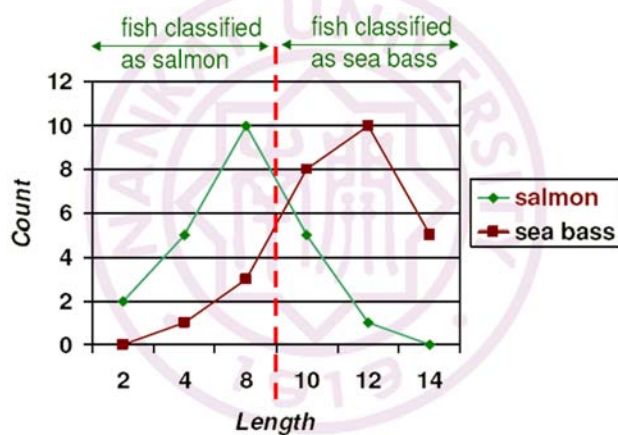
- ◆ 分类错误率： $17/50 = 34\%$

南开大学网络空间安全学院 计算机学院

24

## 确定最佳决策阈值

- 通过搜索所有可能的决策阈值，发现最佳决策阈值为 9，对应的分类错误率为 20%。



南开大学网络空间安全学院 计算机学院

25

## 性能改进

- 评估分类结果：使用鱼身长度特征不能满足分类性能要求
- 改进：
  - 尝试其它特征
  - 观察发现鲑鱼亮度比鲈鱼更大一些
  - 实验使用亮度值作为模式特征

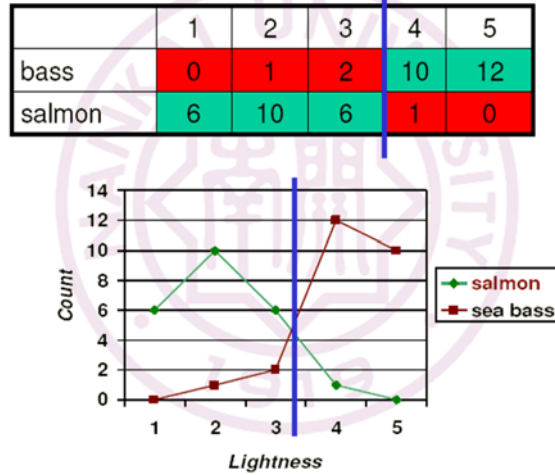


南开大学网络空间安全学院 计算机学院

26

## 选择新模式特征

- 经过相似的实验步骤，得到结果是：当亮度阈值为 3.5 时，获得最佳分类结果，分类错误率为 8%。

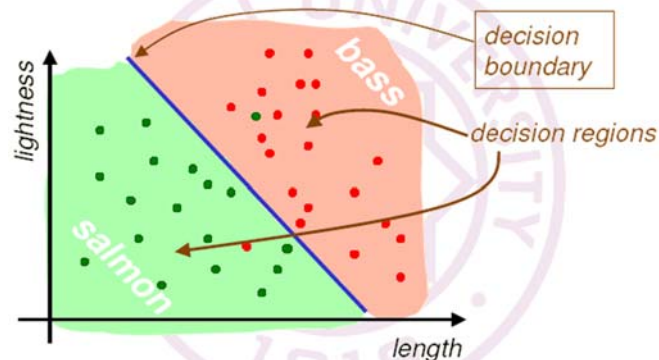


南开大学网络空间安全学院 计算机学院

27

## 多个模式特征

- 思路：使用身长和亮度两个模式特征是否获得更好结果？
- 模式特征矢量：[length, lightness]



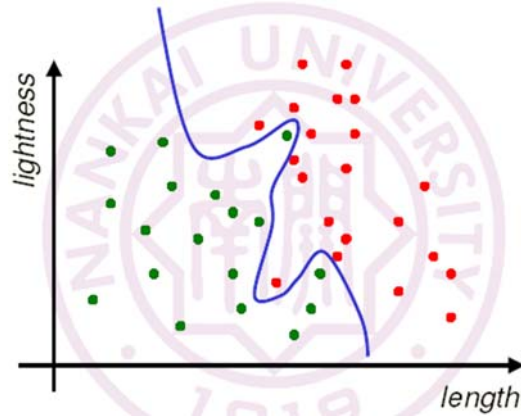
- 最佳分类错误率：4%

南开大学网络空间安全学院 计算机学院

28

## 最佳决策边界

- ❖ 理想（最佳）决策边界对应的分类错误率应该为 0%。



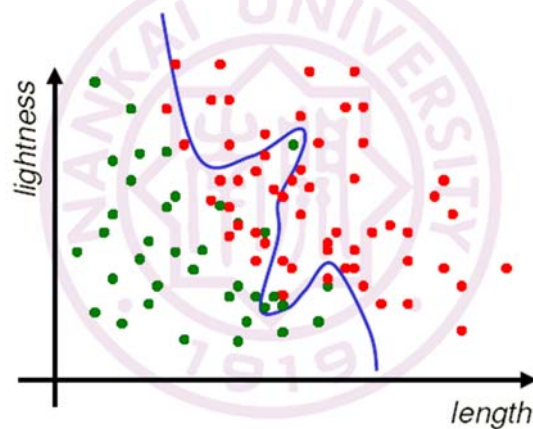
- ❖ 思考：何为最佳？

南开大学网络空间安全学院 计算机学院

29

## 新数据集测试

- ❖ **最佳目标**：对于新的（未见过）数据分类器应该获得最佳结果。
- ❖ 新数据集测试结果：分类错误率为 25%，为什么？



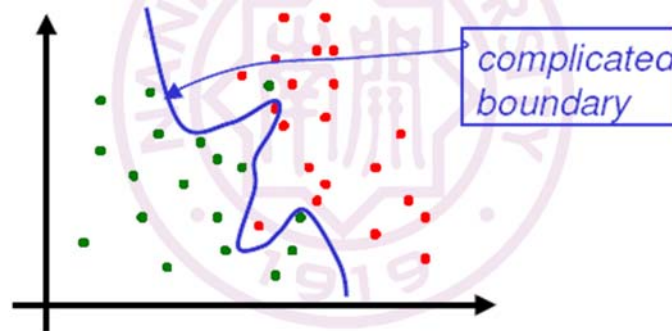
南开大学网络空间安全学院 计算机学院

30



## 过度拟合

- ❖ 原因：不好的归纳（泛化）结果
- ❖ **过度拟合(overfitting)**：复杂的分类器模型（决策边界）以对训练数据集的最佳性能为优化目标，不能够对新的数据进行很好地归纳。

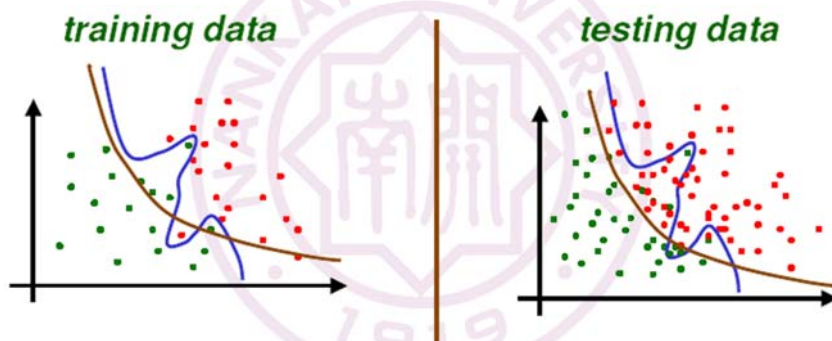


南开大学网络空间安全学院 计算机学院

31

## 泛化能力

- ❖ 泛化能力：简单决策边界对训练数据集的分类结果不够理想，但对新的数据集却能获得比复杂决策边界更好的分类结果。



南开大学网络空间安全学院 计算机学院

32



## 收集数据

- ❖ 收集数据集是有代价的，有时这种代价是昂贵的。
- ❖ 收集到的数据集应该具有充分的代表性（同分布的）。
- ❖ 数据集
  - ◆ 训练集合(training set): 训练分类器模型的样本集合，用于拟合分类器模型的参数（权值）。
  - ◆ 验证集合(validation set): 选择分类器模型的样本集合，用于确定分类器模型的结构（超参数）。
  - ◆ 测试集合(test set): 测试分类器模型的样本集合，用于评估分类器模型的性能（泛化能力）
- ❖ 训练集合、验证集合和测试集合之间的交集为空集，三者之间的比例一般为 8 : 1 : 1。

## 选择模式特征

- ❖ 选择标准：具有很强的模式区分能力
  - ◆ 对于相同类别的模式，它们的模式特征是相似的；
  - ◆ 对于不同类别的模式，它们的模式特征是充分不同的。
- ❖ 先验知识扮演着重要的角色
  - ◆ 人工设计特征
  - ◆ 深度学习方法
- ❖ 特点
  - ◆ 容易获得
  - ◆ 对噪声和不相关变换缺乏敏感性

## 选择分类器模型

- ❖ 选择标准
  - ◆ 数据适应性
  - ◆ 性能适应性
  - ◆ 环境适应性
- ❖ 评价标准
  - ◆ 何时放弃某种分类器模型而试用另一种分类器模型
- ❖ 问题
  - ◆ 对某个给定问题，最佳分类器模型是什么？

## 训练分类器

- ❖ 训练目标
  - ◆ 使用训练样本集合来调整分类器模型参数，以获得对训练样本集合的最佳拟合；
  - ◆ 最佳拟合的主要评价标准是分类错误率
- ❖ 训练算法：针对不同分类器模型和优化目标，存在着多种训练算法。
- ❖ 本课程的重点内容

## 评估分类器

- ❖ 评估标准：已经训练完成的分类器对测试样本集合的分类性能
- ❖ 分析结果：
  - ◆ 如何改进分类器
  - ◆ 调整分类器的复杂性以防止过度拟合
  - ◆ 权衡计算复杂性与分类器性能之间的平衡

## 结论

- ❖ 前途是光明的
  - ◆ 存在大量令人兴奋和重要的理论研究成果和实际应用
- ❖ 道路是曲折的
  - ◆ 需要面对和解决许多问题才能获得成功