模式识别
（Pattern Recognition）

广东工业大学集成电路学院　邢延

第二讲 模式识别系统、基于距离的分类器和分类器性能评估（02 Pattern Recognition System & Distance-based Classifiers & Performance Evaluation）

内容提要

- 模式识别系统
- 分类、分类器、判别函数
- 基于距离的分类器
- 分类器性能评估
- 模式识别算法编程实例演示

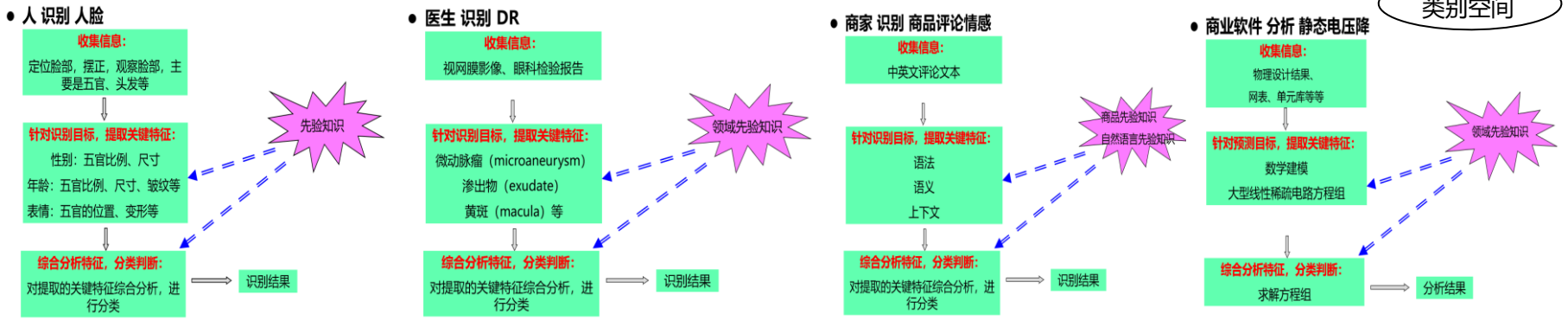
模式识别系统

- 模式识别系统组成
- 分类、分类器和判别函数

模式识别系统

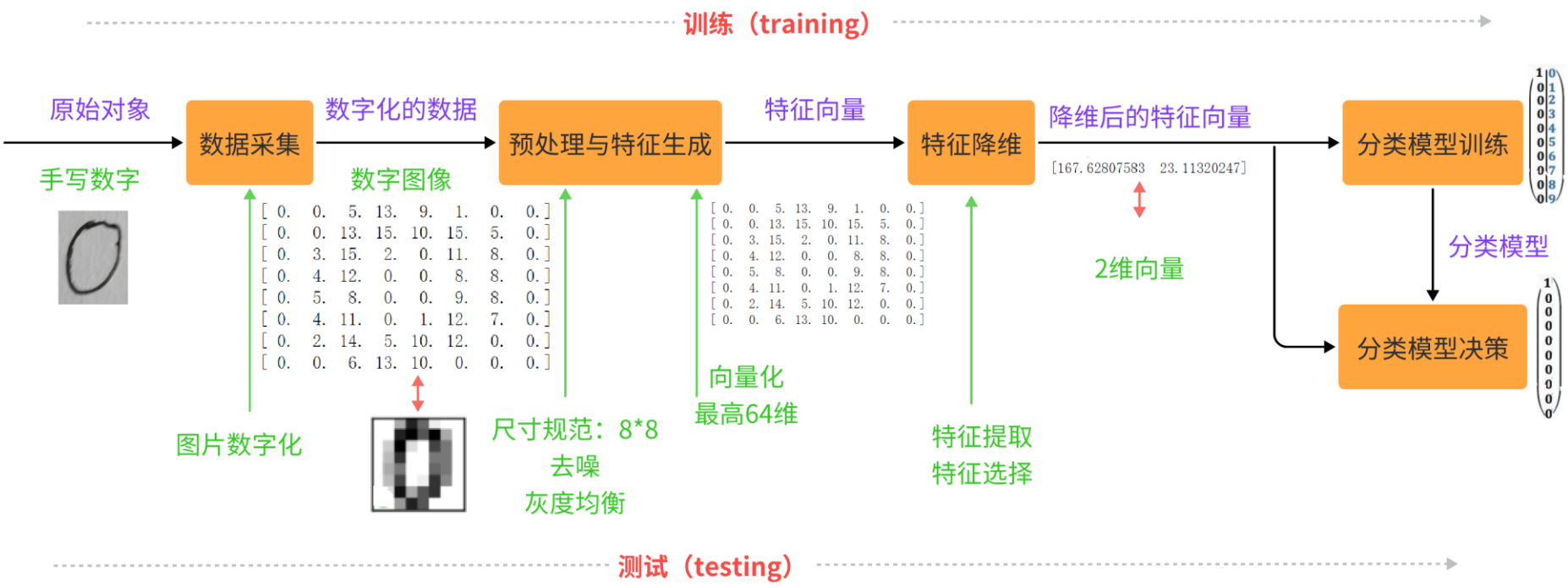
系统组成

数据采集、预处理、特征降维、分类模型训练与测试



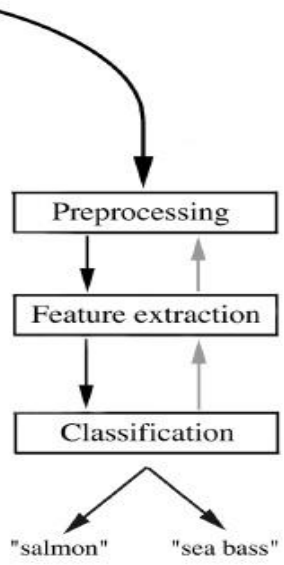
模式识别系统

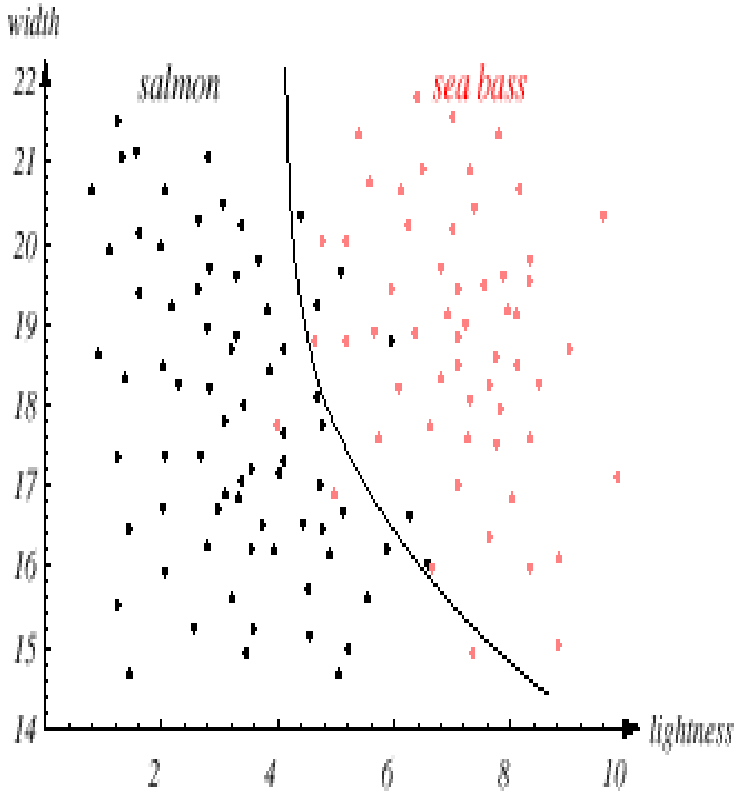
- **模式识别系统实例**
  - **手写数字识别系统**



分类、分类器和判别函数

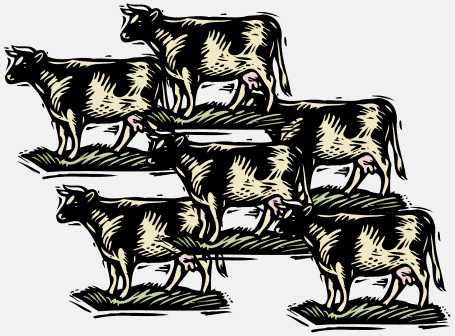
- 分类（Classification）
  - 模式识别的核心任务就是分类



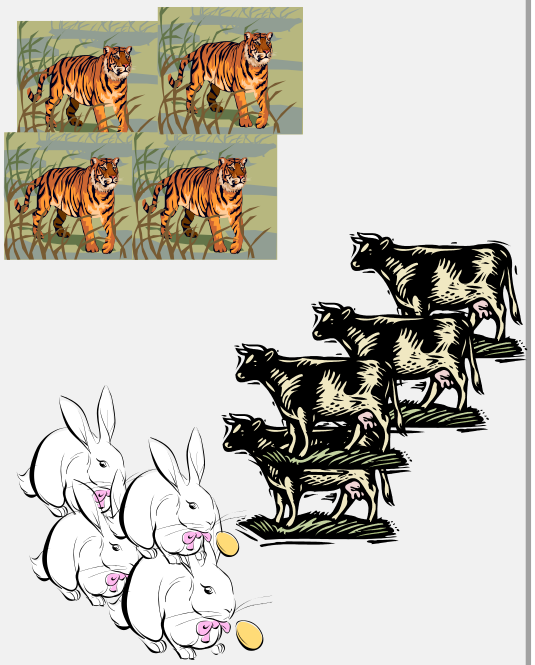


分类、分类器和判别函数

- **分类的类别**
  - **两类和多类**



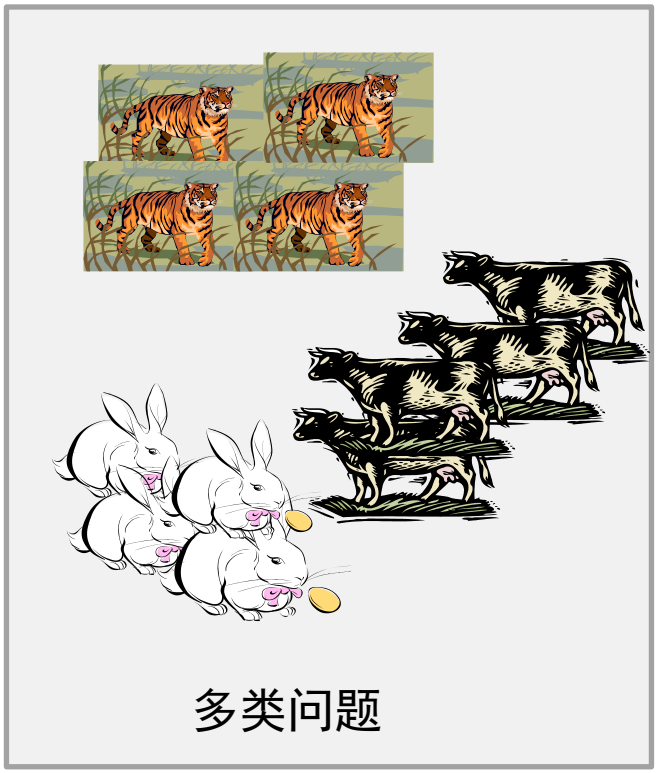
两类问题



多类问题

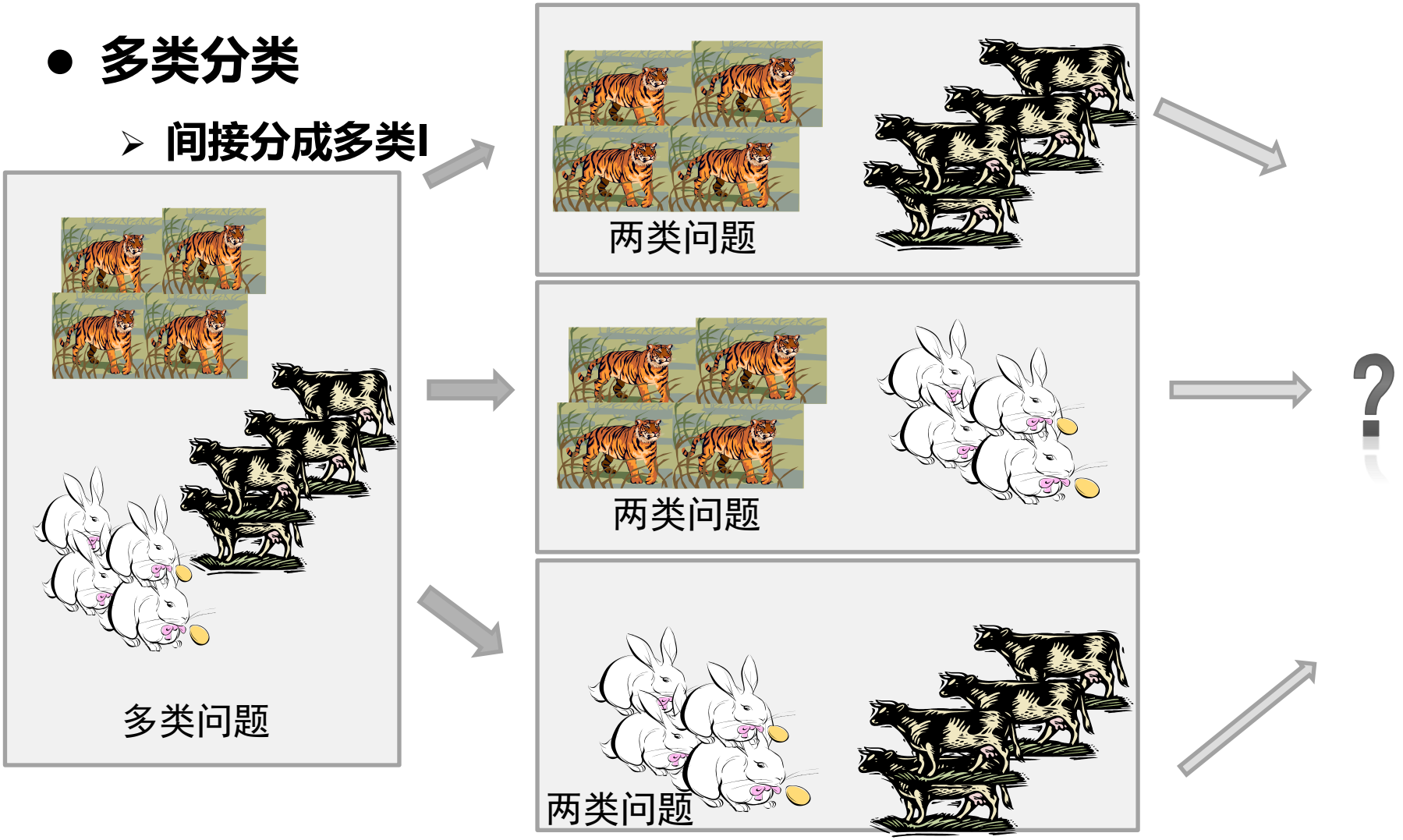
分类、分类器和判别函数

- **多类分类**
  - **直接分成多类（部分分类算法适用）**



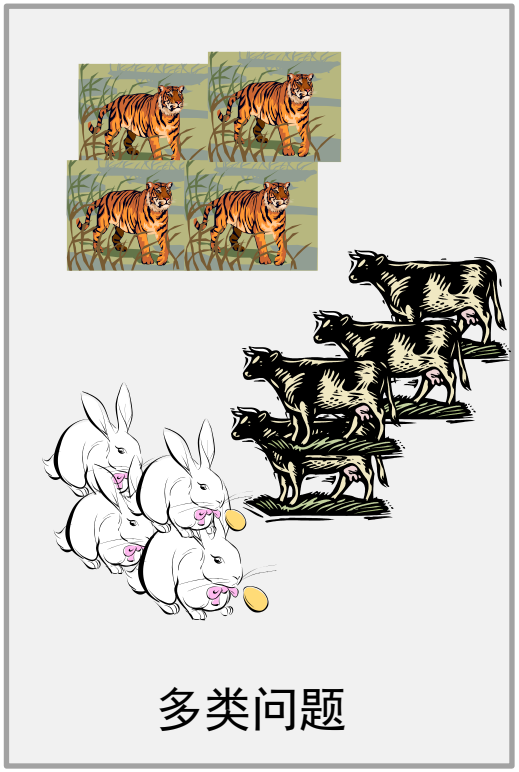
分类、分类器和判别函数

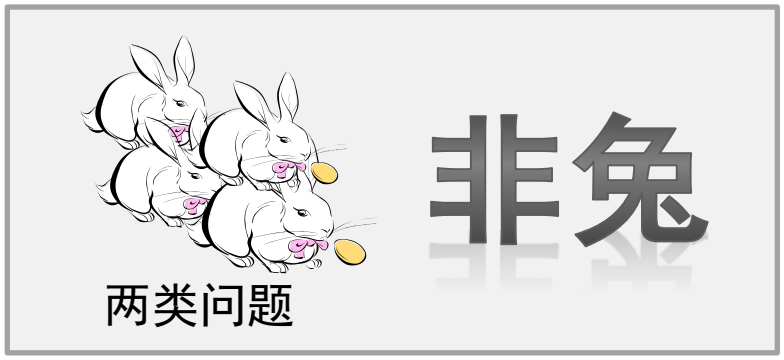
- 多类分类
  - 间接分成多类I



分类、分类器和判别函数

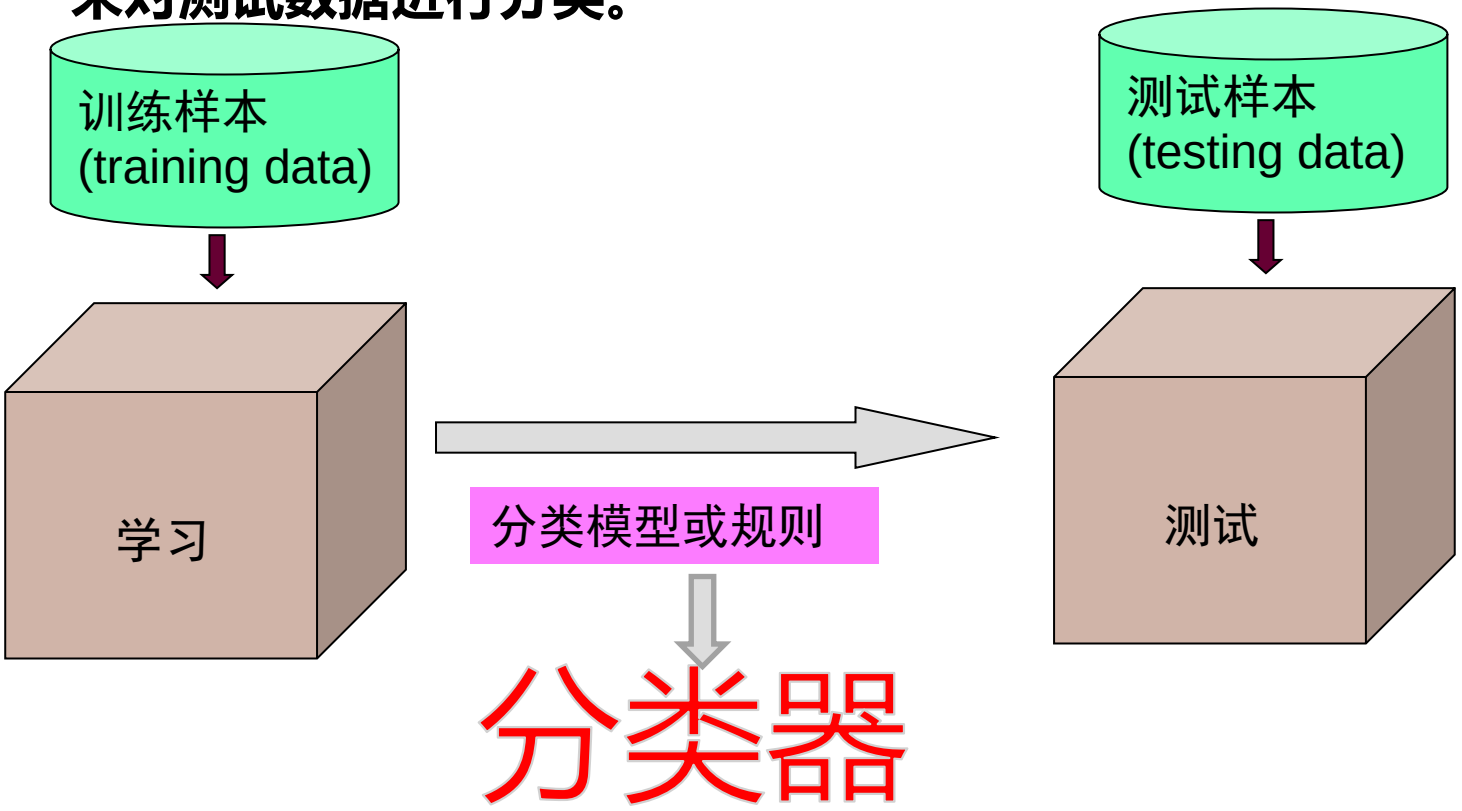
- 多类分类
  - 间接分成多类II





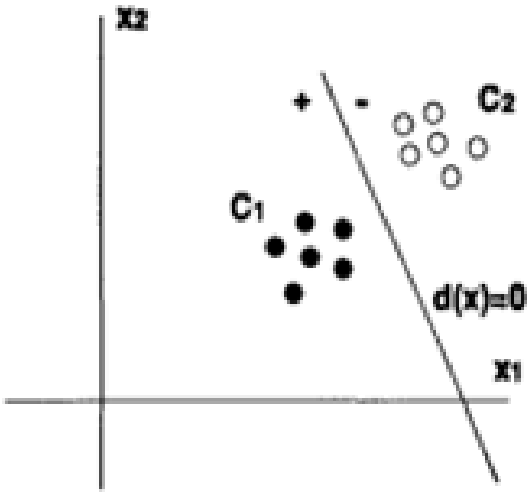
分类、分类器和判别函数

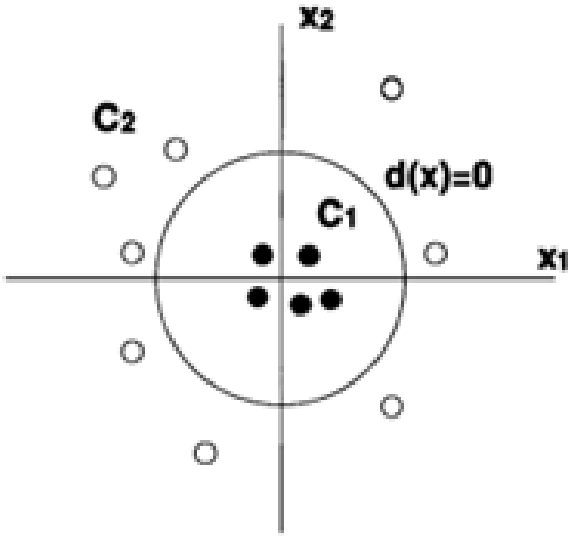
- **分类器**
  - **特定的分类算法；**
  - **用特定分类算法在训练数据上学习，得出的模型或规则，可以用来对测试数据进行分类。**



分类、分类器和判别函数

- 判别函数
  - 定义
    - ◆分类器用于判定待分类数据所属类别的函数
  - 种类
    - ◆线性函数 ------ 线性分类器
    - ◆非线性函数 -------- 非线性分类器





分类、分类器和判别函数

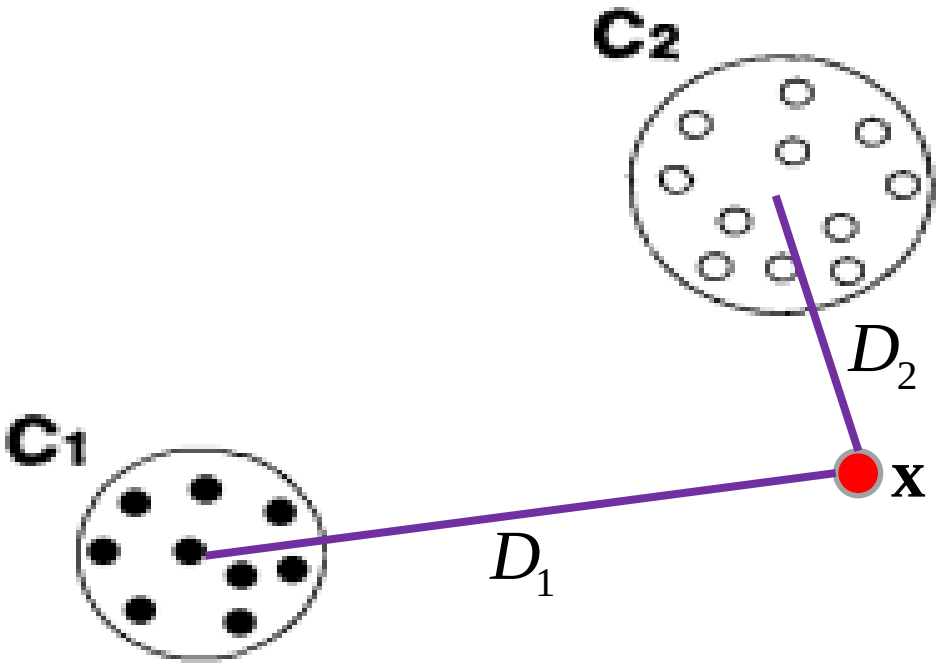
- **数据的可分性**
  - **线性可分与线性不可分**
    - **当数据能够被线性分类器分类，且分类准确率达到某个特定的门限时，该数据就是线性可分的，否则就是线性不可分。**
  - **非线性可分**
    - **当数据能够被非线性分类器分类，且分类准确率达到某个特定的门限时，该数据就是非线性可分的。**

基于距离的分类器

- 基本思想
- 最小距离分类器
- 最近邻分类器
- K近邻分类器

基于距离的分类器

- 基本思想(Main idea)

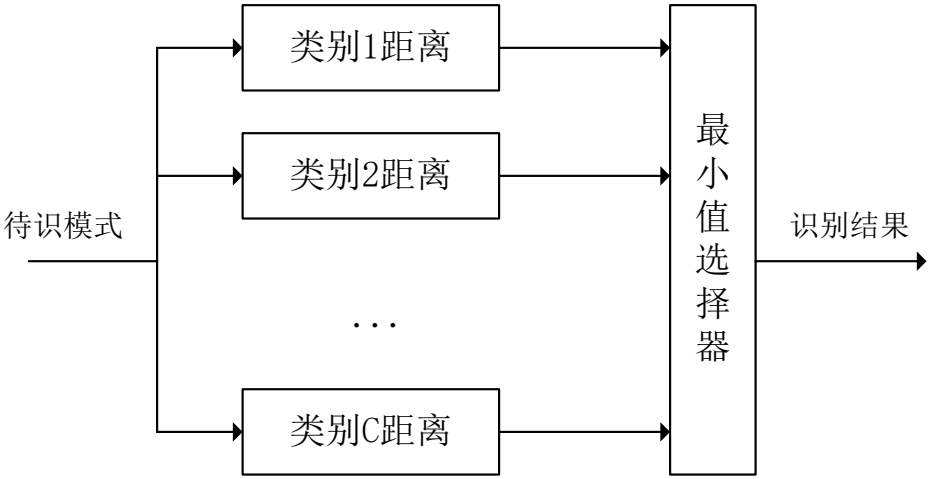


$\because D_2 < D_1$

$\therefore \mathbf{x} \in C_2$

基于距离的分类器

- **最小距离分类算法（Minimum Distance Classifier）**
  - **1)利用训练样本计算出每一类别的代表向量(prototype vector)；**
  - **2)以代表向量作为该类在特征空间中的中心位置，计算待分类样本到各类中心的距离；**
  - **3)根据计算的距离，把待分类样本归入到距离最小的那一类。**



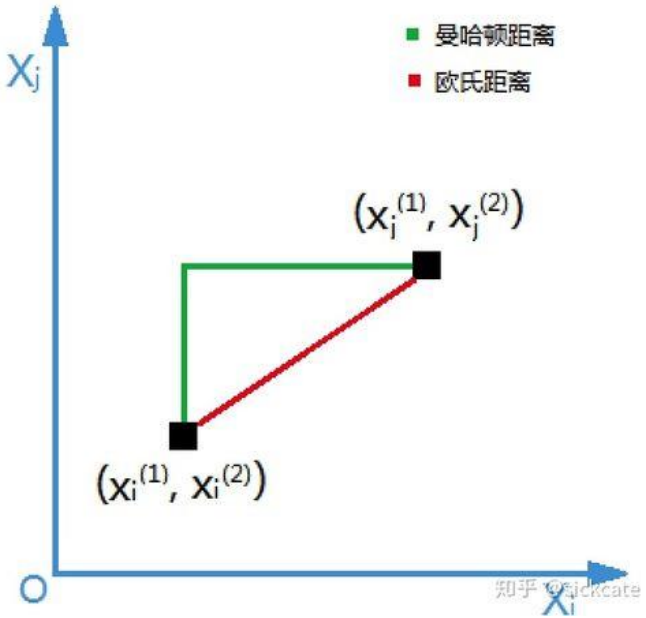
基于距离的分类器

- 代表向量
  - 每类只有单个代表向量
    - ◆可以是平均值向量、重心值向量、中值向量等
  - 每类有多个代表向量
    - ◆可以是平均值向量、重心值向量、中值向量等的组合

基于距离的分类器

- **距离度量**
  - **欧几里德距离(Euclidean Distance)**

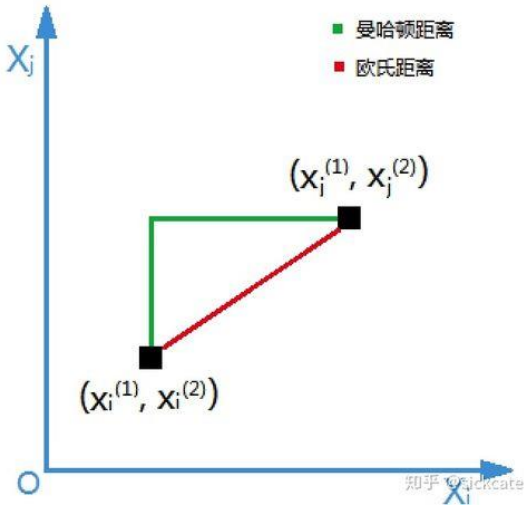
$$d(\mathbf{X},\mathbf{Y})=\left[\sum_{i=1}^{n}(x_i-y_i)^2\right]^{\frac{1}{2}}$$



基于距离的分类器

- **距离度量**
  - **曼哈顿距离(Manhattan Distance)**

$$d(\mathbf{X},\mathbf{Y})=\sum_{i=1}^{n}|x_i-y_i|$$



基于距离的分类器

- **距离度量**
  - **明可夫斯基距离(Minkowski Distance)**

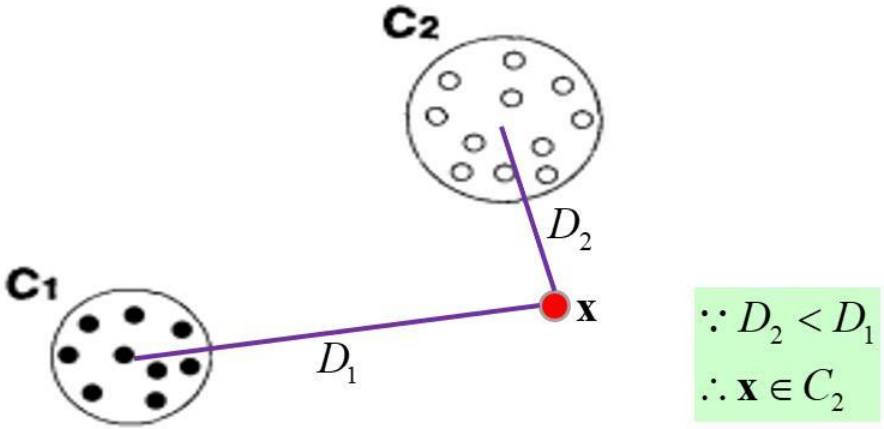
$$d(\mathbf{X},\mathbf{Y})=\left[\sum_{i=1}^{n}\left[x_i-y_i\right]^p\right]^{\frac{1}{p}}$$

p是变参数，$p=1$,就是曼哈顿距离，

$p=2$,就是欧氏距离

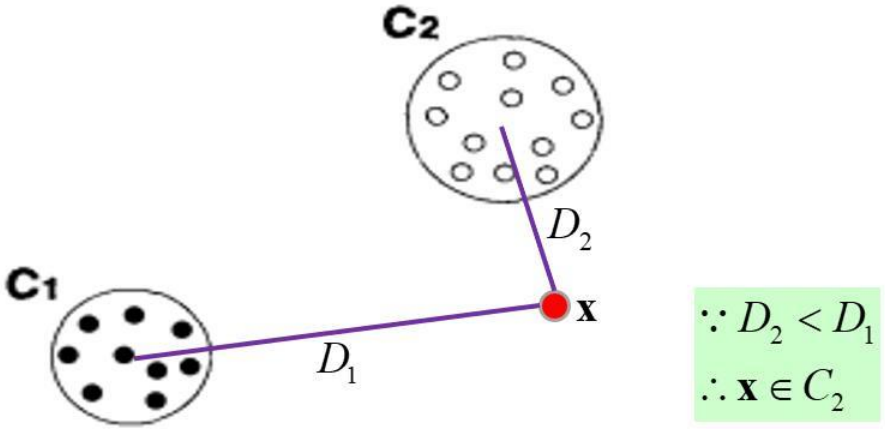
基于距离的分类器

- 问题
  - 最小距离分类器的特点?
  - 影响最小距离分类器性能的因素?



基于距离的分类器

- 问题
  - 最小距离分类器的特点?
- 回答
  - 原理简单，容易理解，计算速度较快；
  - 不考虑类别内部的方差（每一类样本的分布），也不考虑类别之间的协方差（类别和类别之间的相关关系），所以分类精度不高。



基于距离的分类器

- 问题
  - 影响最小距离分类器性能的因素?
- 回答
  - 代表向量
    - ◆ 所选择的代表向量并不一定能很好地代表各类，其后果将使错误率增加。
  - 距离度量
    - ◆ 不同距离算法，类别中心对周围点的作用域是不相同的。

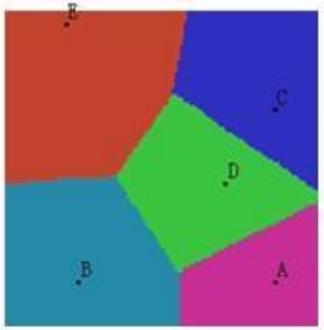


图4. 欧氏距离分类

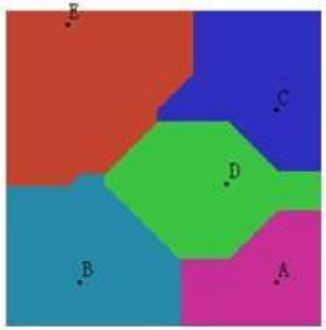


图5. 曼哈顿距离分类

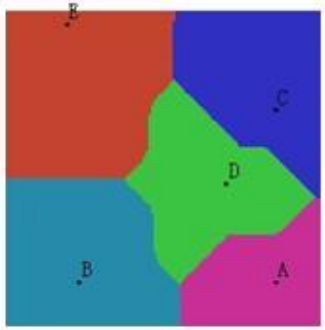
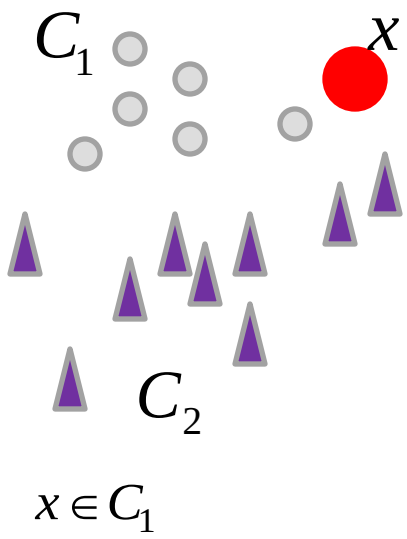


图6. 闵可夫斯基距离(p=10)分类

基于距离的分类器

- **最近邻分类器(Nearest-Neighbor classifier, NNC)**
  - **基本思想**



以全部训练样本作为代表向量集，计算待分类样本与所有训练样本的距离，并以最近邻者的类别作为决策。

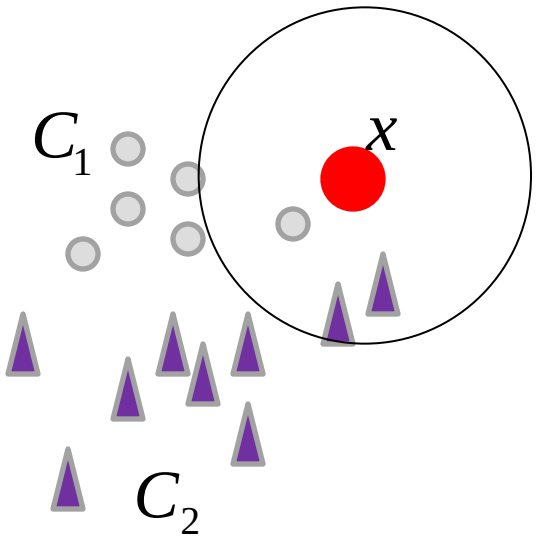
最近邻分类器是最小距离分类器的极端情况

基于距离的分类器

- 最近邻分类器的特点
  - 优点
    - 在原理上最直观
    - 方法上十分简单
  - 缺点
    - 计算量大
    - 存储量大

基于距离的分类器

- **K近邻分类器(K Nearest-Neighbor classifier, KNN)**
  - **基本思想**



$K=3$时，$x \in C_2$

在所有样本中找到与测试样本的K个最近邻者，其中各类别所占个数表示成 K_i，i=1,2,…,m。

定义判别函数为：$g_i(\mathbf{x})=K_i$

则决策规则为：

$$j=\arg\max_i g_i(\mathbf{x}), i=1,2...,m$$

K-近邻一般采用K为奇数，跟投票表决一样，避免因两种票数相等而难以决策。

基于距离的分类器

- **KNN的算法步骤**

算法步骤：

1：令k是最近邻数目，D是训练样例的集合

2：for 每个测试样例z do

3：　　计算z和每个训练样例之间的距离d

4：　　对d进行升序排序

5：　　取前k个训练样例的集合

6：　　统计K个最近邻样本中每个类别出现的次数

7：　　选择出现频率最大的类别作为未知样本的类别

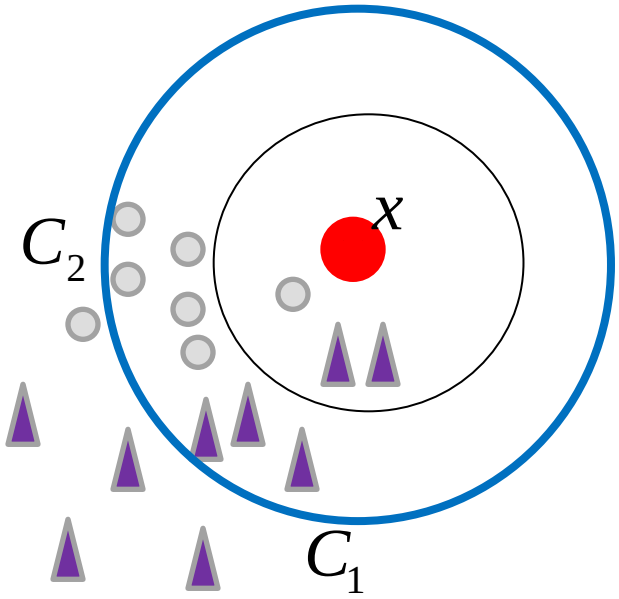
8：end for

基于距离的分类器

- **K-近邻分类器的特点**
  - **优点**
    - **是典型的非参数法**
    - **在原理上最直观，**
    - **方法上十分简单**
  - **缺点**
    - **超参数K的取值至关重要**
    - **需要存储全部训练样本，即存储量大**
    - **繁重的距离计算量，即计算量大，**

基于距离的分类器

- **K-近邻分类器的特点**
  - **K值的选取至关重要**



$K=3$时，$x \in C_1$

$K=11$时，$x \in C_2$

基于距离的分类器

- K-近邻分类器的改进方法
  - K值的智能选取
  - 快速搜索近邻法
  - 剪辑近邻法
  - 压缩近邻法
  - 等等

模式识别算法编程实例演示

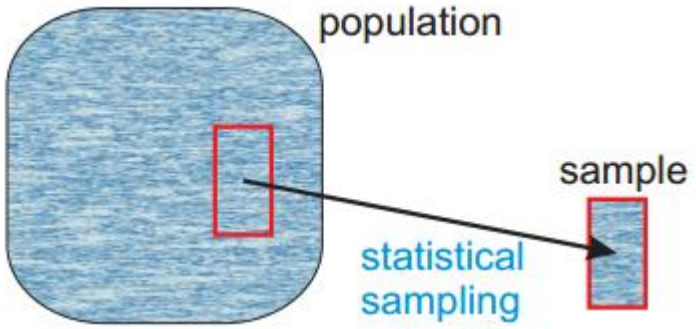
- **实例1：第一个模式识别实例**
  - **算法编程实例/lesson01**

分类器的性能评估

- 分类器性能评估的目的
- 分类器的分类准确性评估
  - 基于混淆矩阵的评估标准
  - ROC&AUC
- 划分训练集和测试集的方法
  - 均衡数据的划分
  - 非均衡数据的处理

分类器性能评估的目的

- 问题：
  - 通过学习得到的分类器靠谱吗?
- 模式识别的假设前提：
  - 数据是独立同分布（IID）
  - IID = Independently and Identically Distributed



分类器性能评估的目的

- **分类器性能比较-最优分类器**

Naïve Bayes
朴素贝叶斯

Neural Network
神经网络

Support Vector Machine
支持向量机

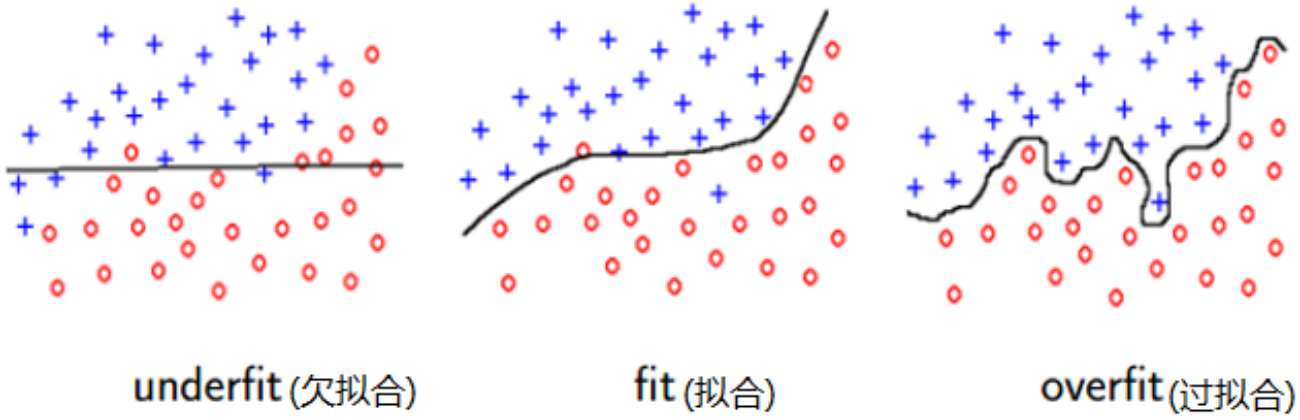
Decision Tree
决策树

K-nearest neighbor
K近邻

Others
其他

分类器的分类准确性评估

- 分类器的性能要求
  - 分类器的分类准确率要高
  - 分类器的泛化性要好



分类器的分类准确性评估

- 混淆矩阵（confusion matrix）
  - 对于二类分类器

		真实值		总数
		p	n	
预测输出	p'	真阳性（TP）	伪阳性（FP）	P'
	n'	伪阴性（FN）	真阴性（TN）	N'
总数		P	N	

TP（True Positive）：正确的正例，一个实例是正类并且也被判定成正类；

FN（False Negative）：错误的反例，漏报，本为正类但判定为负类；

FP（False Positive）：错误的正例，误报，本为负类但判定为正类；

TN（True Negative）：正确的反例，一个实例是负类并且也被判定成负类。

分类器的分类准确性评估

- 分类准确率(Accuracy)
  - 定义
    - 正确分类的测试实例个数占测试实例总数的比例
  - 公式
    - $Accuracy = \frac{TP+TN}{P+N}$

		真实值		总数
		p	n	
预测输出	p'	真阳性（TP）	伪阳性（FP）	P'
	n'	伪阴性（FN）	真阴性（TN）	N'
总数		P	N	

分类器的分类准确性评估

- 错分率(Error rate)
  - 定义
    - 错误分类的测试实例个数占测试实例总数的比例
  - 公式
    - $Error\_rate = 1 - Accuracy = \frac{FP+FN}{P+N}$

		真实值		总数
		p	n	
预测输出	p'	真阳性（TP）	伪阳性（FP）	P’
	n'	伪阴性（FN）	真阴性（TN）	N’
总数		P	N	

分类器的分类准确性评估

- 使用准确率和错分率时可能会遇到的问题
  - 根据以下信息，预测是否患有癌症
    - 一般情况下，人群中患癌症的比例为0.5%；
    - 构建一个分类器，任何一个人都判断患有癌症，则该分类器的准确率高达99.5%；
    - 但这个分类准确率毫无意义，因为在实际应用中，真正需要的是识别出患有癌症的那一部分人，即侧重小类的准确率；
    - 导致这一现象的根本原因是数据中两类的分布极其不均衡（imbalance）；
    - 因此，需要更有效的指标。

分类器的分类准确性评估

		真实值		总数
		p	n	
预测输出	p'	真阳性（TP）	伪阳性（FP）	P’
	n'	伪阴性（FN）	真阴性（TN）	N’
总数		P	N	

- **查准率(Precision)**
  - **定义**
    - **正确分类的正例个数占分类为正例的实例个数的比例。**
  - **公式**
    - $Precision = \frac{TP}{TP+FP} = \frac{TP}{P'}$

分类器的分类准确性评估

		真实值		总数
		p	n	
预测输出	p'	真阳性（TP）	伪阳性（FP）	P’
	n'	伪阴性（FN）	真阴性（TN）	N’
总数		P	N	

- **查全率(Recall)**
  - **定义**
    - **正确分类的正例个数占实际正例个数的比例。**
  - **公式**
    - $Recall = \frac{TP}{TP+FN} = \frac{TP}{P}$

分类器的分类准确性评估

		真实值		总数
		p	n	
预测输出	p'	真阳性（TP）	伪阳性（FP）	P’
	n'	伪阴性（FN）	真阴性（TN）	N’
总数		P	N	

- **F1**
  - **定义**
    - **查全率与查准率的调和平均数**
  - **公式**
    - $F1 = \frac{2*Recall*Precision}{Recall+Precision}$
  - **F1值越大，分类器的分类准确率越高**

分类器的分类准确性评估

		真实值		总数
		p	n	
预测输出	p'	真阳性（TP）	伪阳性（FP）	P’
	n'	伪阴性（FN）	真阴性（TN）	N’
总数		P	N	

- 敏感度(Sensitivity)
  - 定义
    - 查全率又称为敏感度
    - 即真阳性率（True positive rate)
  - 公式
    - $Sensitivity = Recall = \frac{TP}{TP+FN} = \frac{TP}{P}$

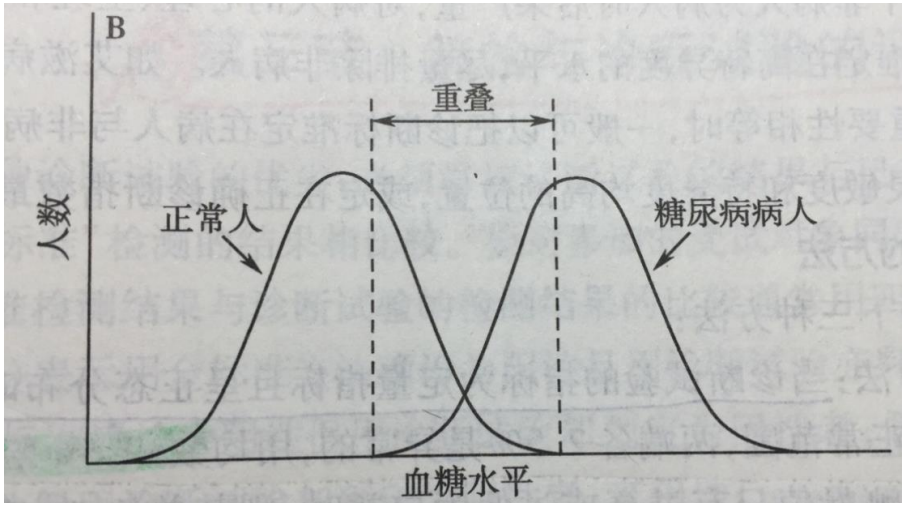
分类器的分类准确性评估

		真实值		总数
		p	n	
预测输出	p'	真阳性（TP）	伪阳性（FP）	P’
	n'	伪阴性（FN）	真阴性（TN）	N’
总数		P	N	

- **特异度(Specifity)**
  - **定义**
    - **真阴性率（TNR：true negative rate)**
    - **正确分类的负例个数占实际负例个数的比例**
  - **公式**
    - $Specifity = \frac{TN}{TN+FP} = \frac{TN}{N}$

分类器的分类准确性评估

- 实例
  - 只将病人血糖水平作为判断是否患有糖尿病的指标



  - 设：患病的为T，正常的为N
  - 敏感度高表明漏诊率低，特异度高表明误诊率低
  - 希望敏感度和特异度都高，然而实际上只能在二者之间求平衡

分类器的分类准确性评估

- **接收机工作特性曲线** **Receiver Operating Characteristic (ROC Curve)**

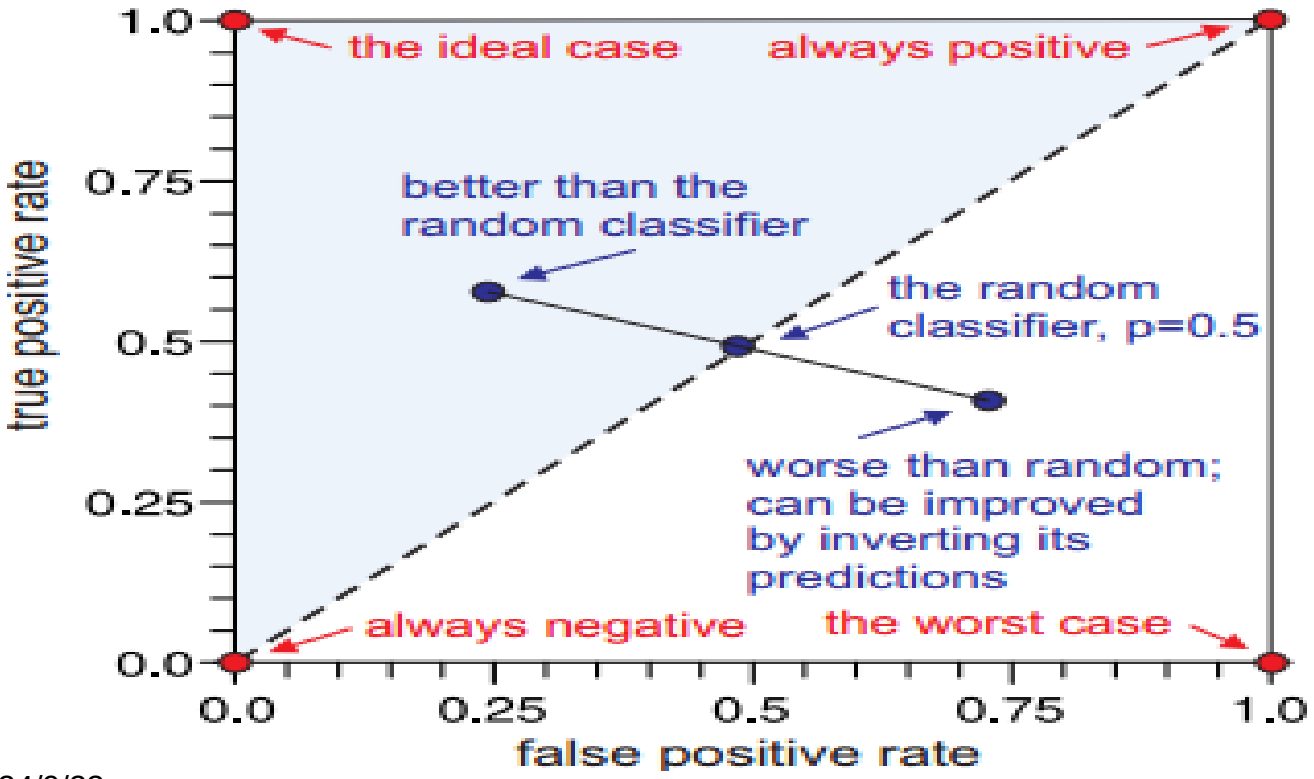
Y轴——真阳性率：

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$

X轴——伪阳性率(1-特异度)：

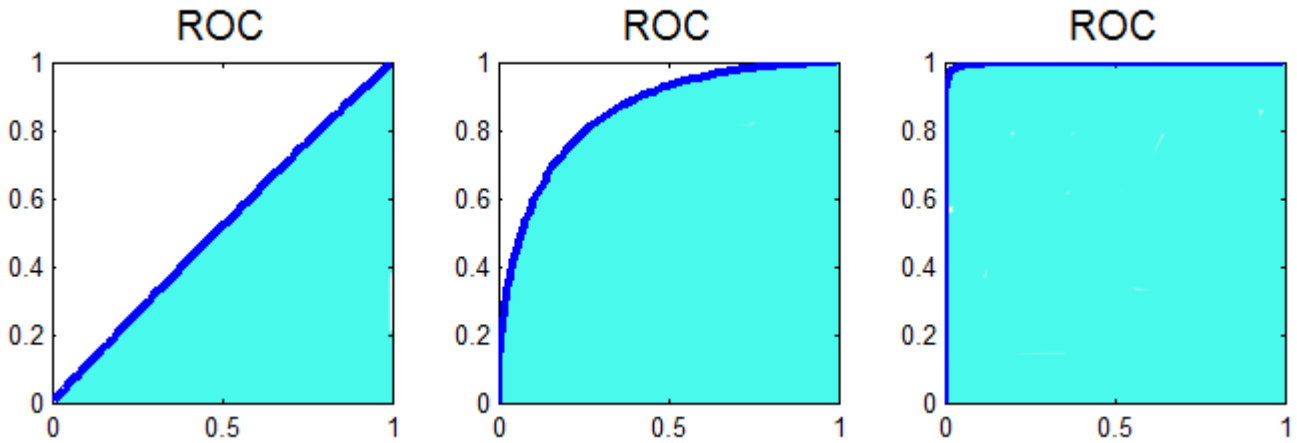
$$FPR = \frac{FP}{N} = \frac{FP}{TN + FP}$$

$$\begin{cases} 0 \le TPR \le 1 \\ 0 \le FPR \le 1 \end{cases}$$



分类器的分类准确性评估

- ROC曲线下面积（AUC）
  - AUC 反映了类别之间的分离程度
  - AUC = ½ 表示类分布完全重叠（最劣）
  - AUC = 1 表示类分布没有重叠（最优）

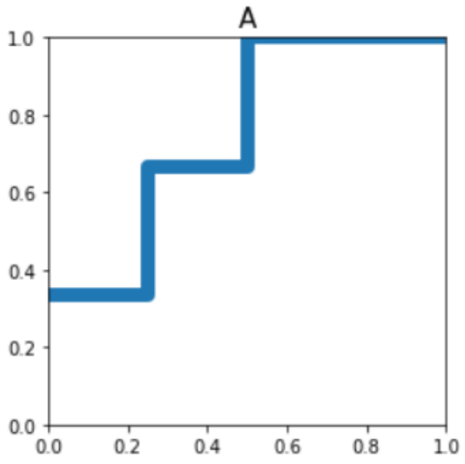


分类器的分类准确性评估

- **ROC曲线绘制**
  - **对于一组二元分类任务的测试集，**
    - **其真实值为[0, 0, 0, 0, 1, 1, 1]，**
    - **模型预测为1的概率为[0.3, 0.2, 0.7, 0.5, 0.4, 0.9, 0.6]，**
  - **该模型在这个测试集上的ROC曲线?**

样本编号	真实值	预测为T的概率	分类结果（>门限）							
			>0.9	>0.7	>0.6	>0.5	>0.4	>0.3	>0.2	>0
6	1	0.9	0(FN)	1(TP)	1(TP)	1(TP)	1(TP)	1(TP)	1(TP)	1(TP)
3	0	0.7	0(TN)	0(TN)	1(FP)	1(FP)	1(FP)	1(FP)	1(FP)	1(FP)
7	1	0.6	0(FN)	0(FN)	0(FN)	1(TP)	1(TP)	1(TP)	1(TP)	1(TP)
4	0	0.5	0(TN)	0(TN)	0(TN)	0(TN)	1(FP)	1(FP)	1(FP)	1(FP)
5	1	0.4	0(FN)	0(FN)	0(FN)	0(FN)	0(FN)	1(TP)	1(TP)	1(TP)
1	0	0.3	0(TN)	0(TN)	0(TN)	0(TN)	0(TN)	0(TN)	1(FP)	1(FP)
2	0	0.2	0(TN)	0(TN)	0(TN)	0(TN)	0(TN)	0(TN)	0(TN)	1(FP)
	灵敏度=TP/(TP+FN)		0	0.3	0.3	0.6	0.6	1	1	1
	特异度=TN/(TN+FP)		1	1	0.75	0.75	0.5	0.5	0.25	0
	1-特异度		0	0	0.25	0.25	0.5	0.5	0.75	1

对应的ROC曲线为：

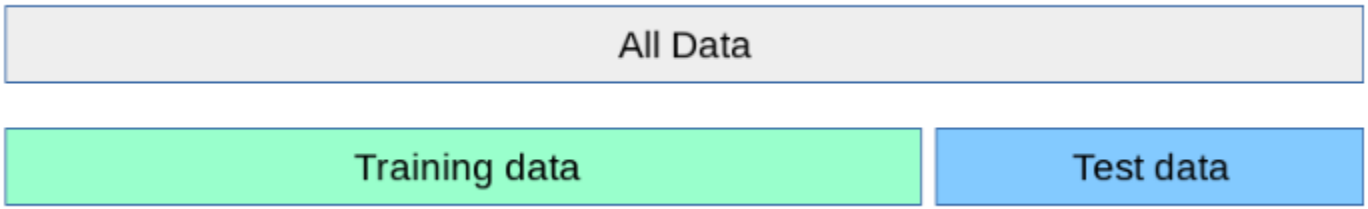


划分训练集和测试集的方法

- 当数据量足够多的时候
  - 随机划分法
    - 训练集占总样本的2/3，测试集占总样本的1/3，都由样本中无回放的随机抽取产生。
    - 每次随机数不同，重复20次，求均值。
- 当数据量不足的时候：
  - 采用交叉验证（cross validation）
    - Hold-Out Method
    - K-fold Cross Validation（K-折交叉验证，记为K-CV）

划分训练集和测试集的方法

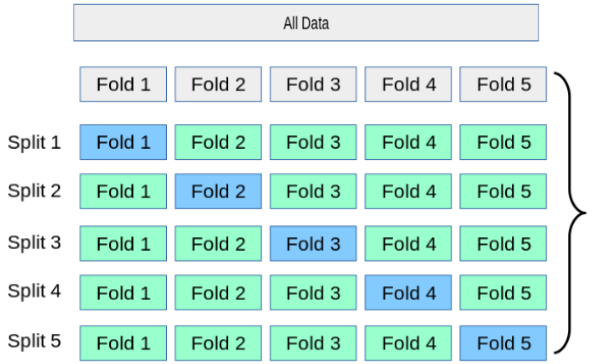
- **Hold-Out Method**
  - **方法**
    - **将原始数据随机分为两组，一组做为训练集，一组做为测试集；**



  - 特点
    - 处理简单；
    - 没有达到交叉的思想，由于是随机的将原始数据分组，所以最后验证集分类准确率的高低与原始数据的分组有很大的关系，所以这种方法得到的结果其实并不具有说服性；
    - 小概率事件有可能发生
    - 因数据的极端情况而导致模型过拟合或者欠拟合

划分训练集和测试集的方法

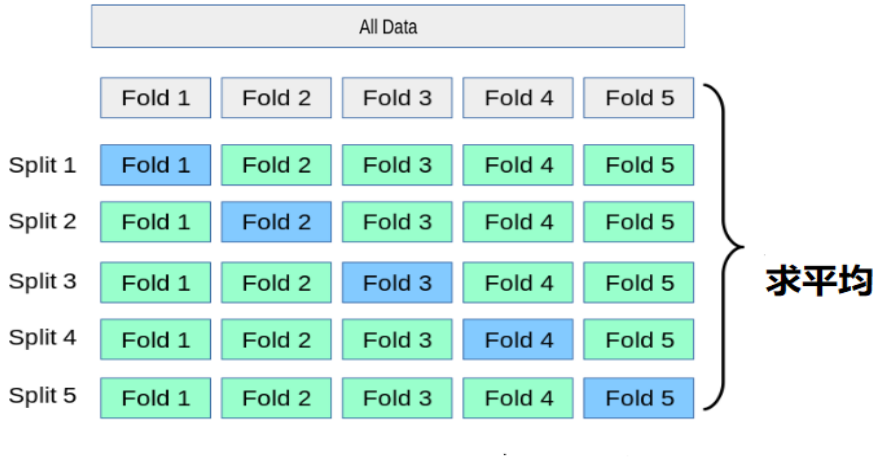
- **K-fold Cross Validation**
  - **方法**
    - **1）将原始数据分成K组（一般是均分）；**
    - **2）将每个子集数据分别做一次测试集，其余的K-1组子集数据作为训练集，这样会得到K个模型；**
    - **3）用这K个模型最终的测试集的分类准确率的平均数作为此K-CV下分类器的性能指标；**
    - **K一般大于等于2，实际操作时一般从3开始取，一般取10,20。**



- 特点
  - K-CV可以有效的避免过学习以及欠学习状态的发生，最后得到的结果也比较具有说服性。

划分训练集和测试集的方法

- 思考题
  - K折交叉验证，每一次训练的模型在参数上是不同的。如果有新数据需要测试，应该用哪一个模型呢?



划分训练集和测试集的方法

- 非均衡数据的处理
  - 通过各种采样方式实现数据尽可能分布均衡
    - 欠采样（undersampling）
    - 过采样（oversampling）
    - 欠采样和过采样结合
    - SMOTE (Synthetic Minority Over-sampling Technique)
    - ……
    - 可能回事翻转课堂的问题之一

模式识别算法编程实例演示

- **实例2：分类器的性能评估实例**
  - **算法编程实例/lesson02**