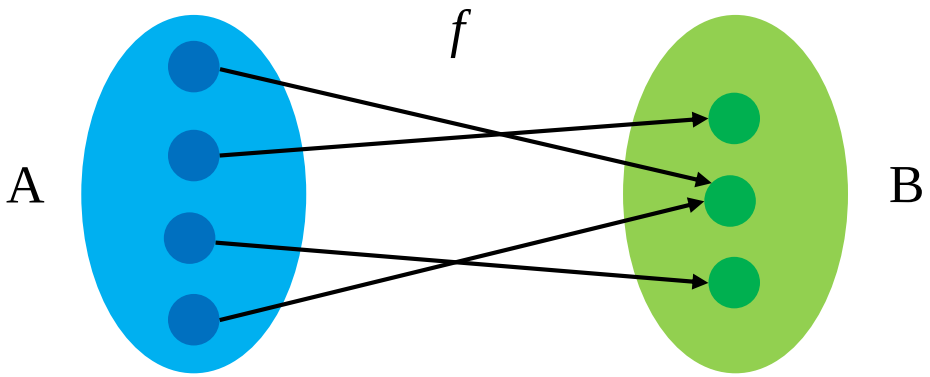
机器学习基础

主讲人：屠恩美

《机器学习与知识发现》

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

机器学习主要研究的“问题们”



- **映射建模问题：**

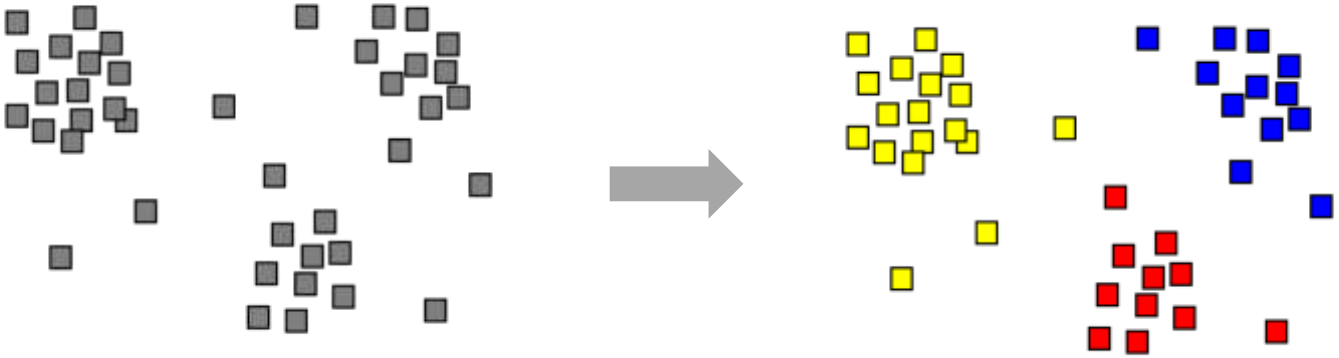
  分类($B=\mathbb{Z}$)：给一个输入对象，它属于已知类别中的哪类？

  回归($B=\mathbb{R}$)：给一个输入对象，它对应的某种观测值多大？

- **物体聚类问题：**

- **特征降维问题：**

机器学习主要研究的“问题们”



- **映射建模问题：**

  分类($B=\mathbb{Z}$)：给一个输入对象，它属于已知类别中的哪类？

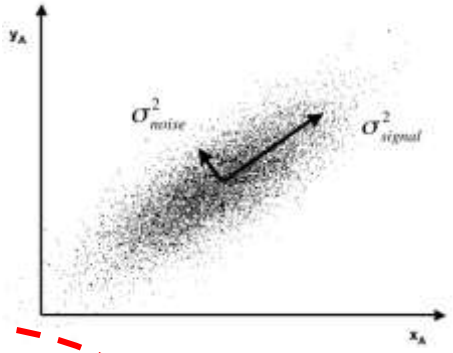
  回归($B=\mathbb{R}$)：给一个输入对象，它对应的某种观测值多大？

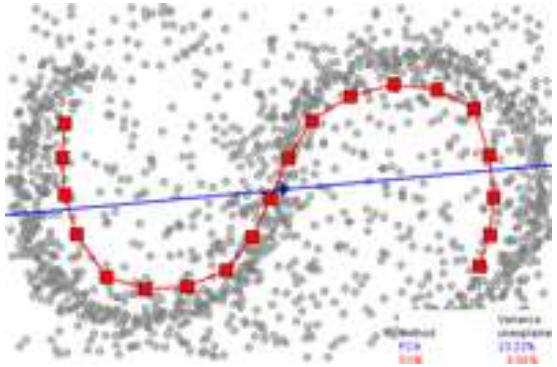
- **物体聚类问题：**

  给一组输入对象，如何他们按相似度归纳到几个类中？

- **特征降维问题：**

机器学习主要研究的“问题们”





- **映射建模问题：** (有)监督学习

  分类($B=\mathbb{Z}$)：给一个输入对象，它属于已知类别中的哪类？

  回归($B=\mathbb{R}$)：给一个输入对象，它对应的某种观测值多大？

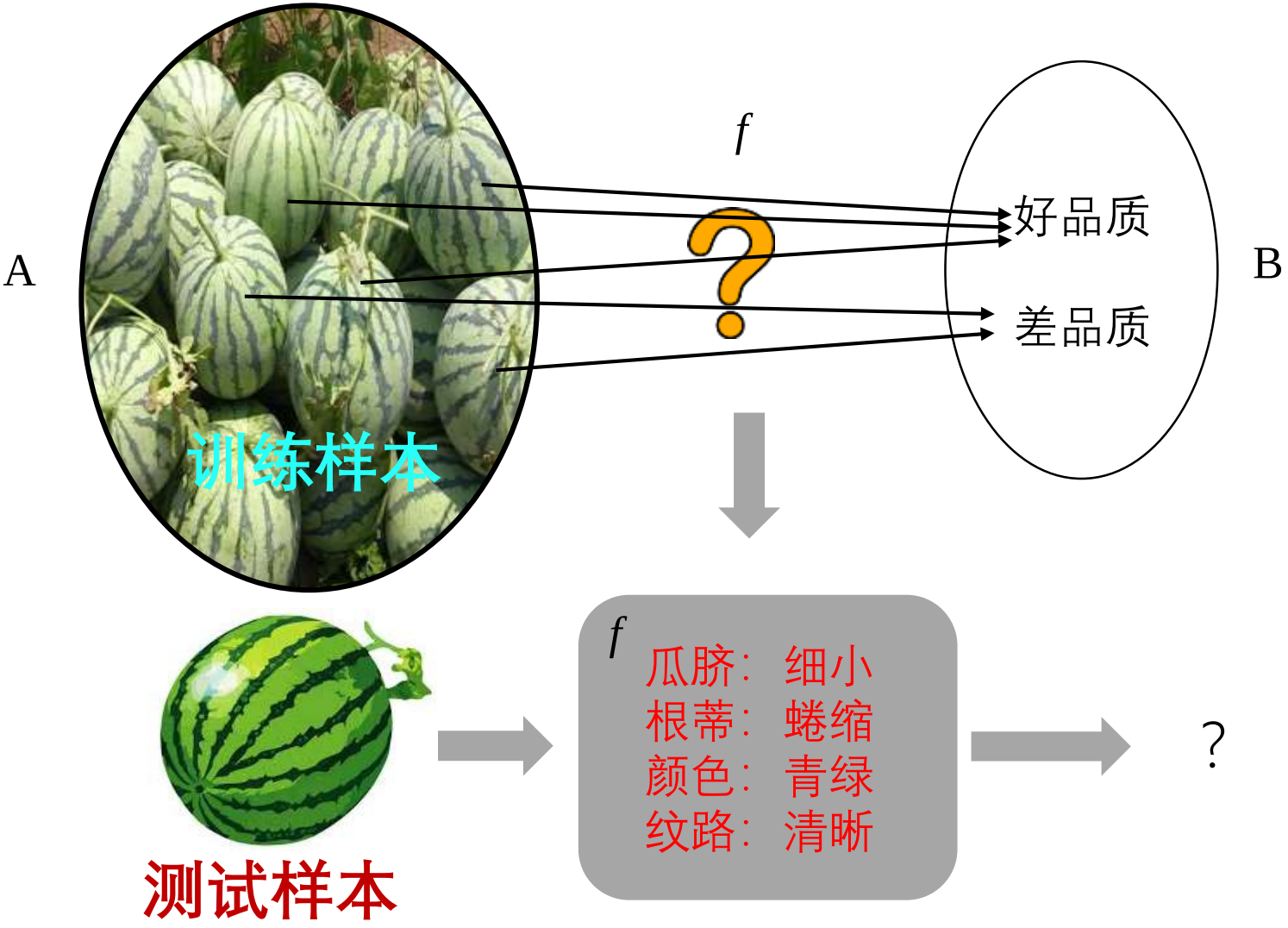
- **物体聚类问题：** 无监督学习

  给一组输入对象，如何他们按相似度归纳到几个类中？

- **特征降维问题：** 多数无监督学习，少数有监督学习

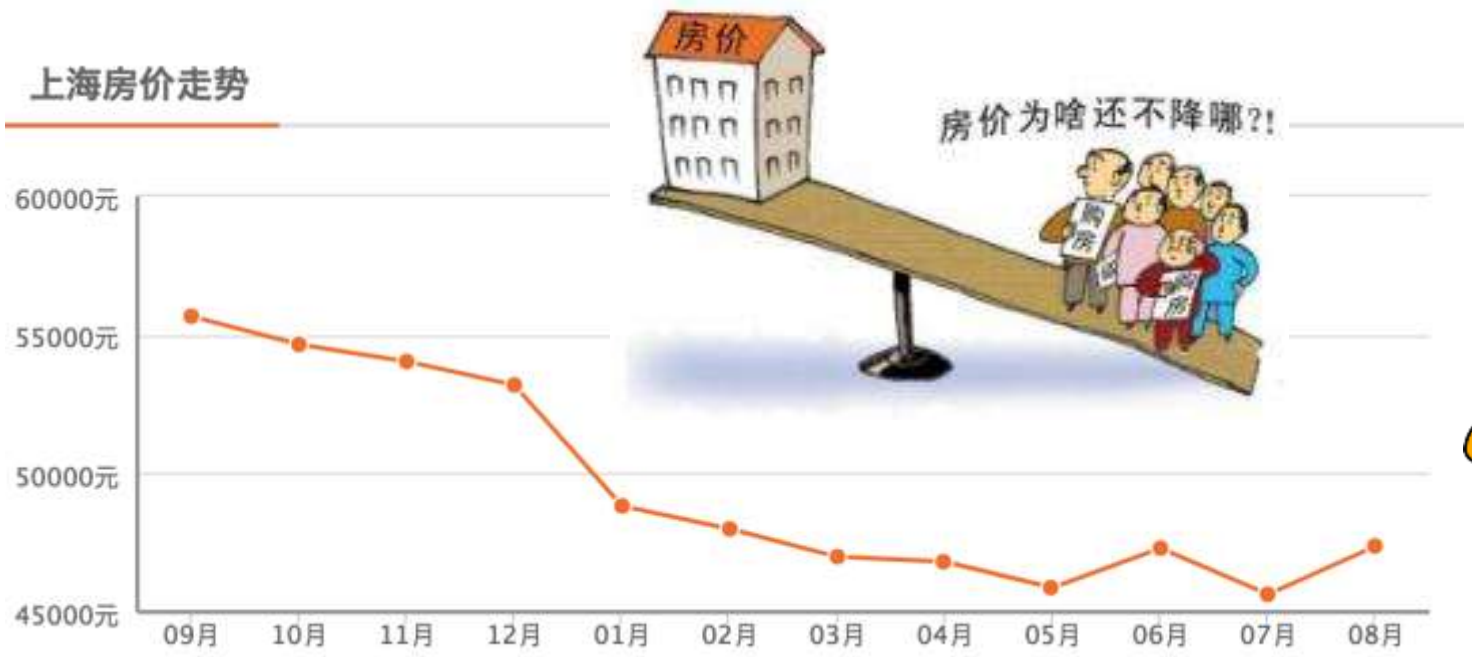
  给一组输入对象，如何找到主导他们变化的关键因素？

分类问题例子 - 判断西瓜例子



回归问题例子- 房价预测

知道过去一年的房价变化，能预测将来的趋势吗？



A

时间	2017-09	2017-10	2017-11	2017-12	2018-01	2018-02	2018-03	2018-04	2018-05	2018-06	2018-07	2018-08
单价(万)	5.56	5.46	5.40	5.31	4.88	4.79	4.69	4.68	4.58	4.73	4.56	4.73

B

聚类问题例子– 文档归类

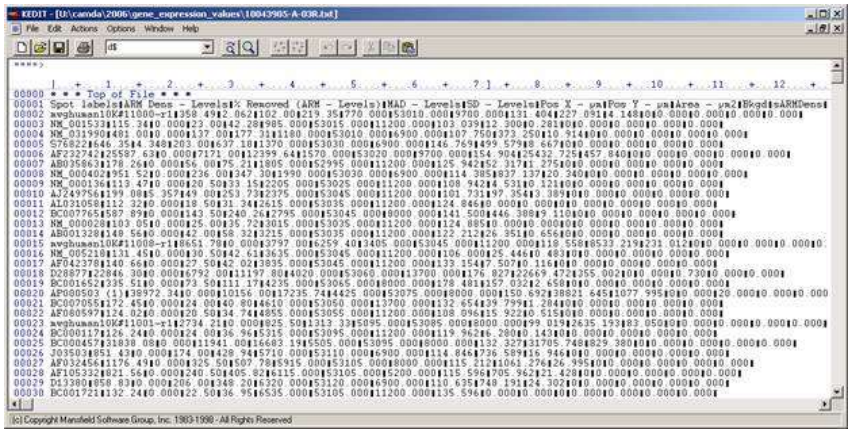
- 一大堆文件怎么存放?

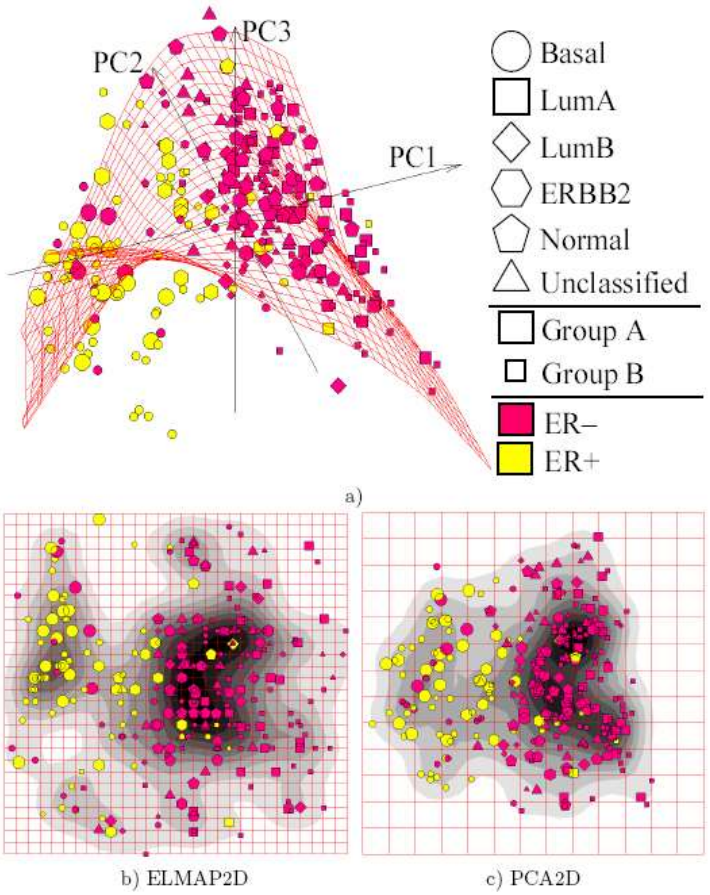
相似的归类到一起

降维问题例子– 异常基因分布

- 基因表达数据冗长抽象，

  很难直接观察出什么规律





GROUP: non-agressive (A) vs agressive (B) cancer
ER: estrogen-receptor positive (ER+) vs estrogen-receptor negative (ER-) tumors
TYPE: five types of breast cancer (lumA, lumB, normal, errb2, basal and unclassified _)

机器学习的一般形式

- 机器学习通常推导成以下优化问题解决

$$\min_{\mathbf{x}\in\mathbb{R}^d} \quad f_0(\mathbf{x})$$

目标函数

$$\text{s. t.} \quad f_i(\mathbf{x}) = 0, \quad i = 1, 2, ..., p$$

等式约束

$$g_j(\mathbf{x}) \le 0, \quad j = 1, 2, ..., q$$

不等式约束

- 常用求解方法：构造拉普拉斯乘子

$$L(\boldsymbol{\lambda}, \boldsymbol{\tau}) = f_0(\mathbf{x}) + \sum_{i=1}^{p} \lambda_i f_i(x) + \sum_{j=1}^{q} \tau_i g_i(x)$$

  - 解析推导 $\nabla L_{\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\tau}}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\tau}) = 0$
  - 梯度下降 $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta \Delta \mathbf{x}$
  - 对偶求解 $\max_{\boldsymbol{\lambda},\boldsymbol{\tau}} L(\boldsymbol{\lambda}, \boldsymbol{\tau}) = f_0(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{F}(\mathbf{x}) + \boldsymbol{\tau}^T \mathbf{G}(\mathbf{x}) \Big|_{\mathbf{x} \leftarrow \nabla L_{\mathbf{x}}(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\tau})=0}$

机器学习的例子

- 线性拟合 $f(x) = wx + b$

$$\min \sum_{i=1}^{n} \left(f(x_i) - y_i \right)^2$$

- 支持向量机

$$\min \ \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, i = 1, 2, ..., n$$

- k-means聚类

$$\min \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

基本概念 - 样本与标记

- **样本(或实例)**：描述一个事物或对象的一组属性/特征集合

  人脸：圆脸，大眼睛，高鼻梁，樱桃嘴，柳叶眉……

  西瓜：色泽，根蒂，敲声，体积，重量

  把属性的取值用向量表示，称为样本向量；向量长度即特征维数

  一号瓜：x_1=(青绿，蜷缩，浊响，0.5，2.5)

  二号瓜：x_2=(浅绿，稍蜷，沉闷，0.3，1.8)

- **标记(或标签)**：样本所对应的真类别或实际测量值，也称为**目标值**

  预测是否好瓜

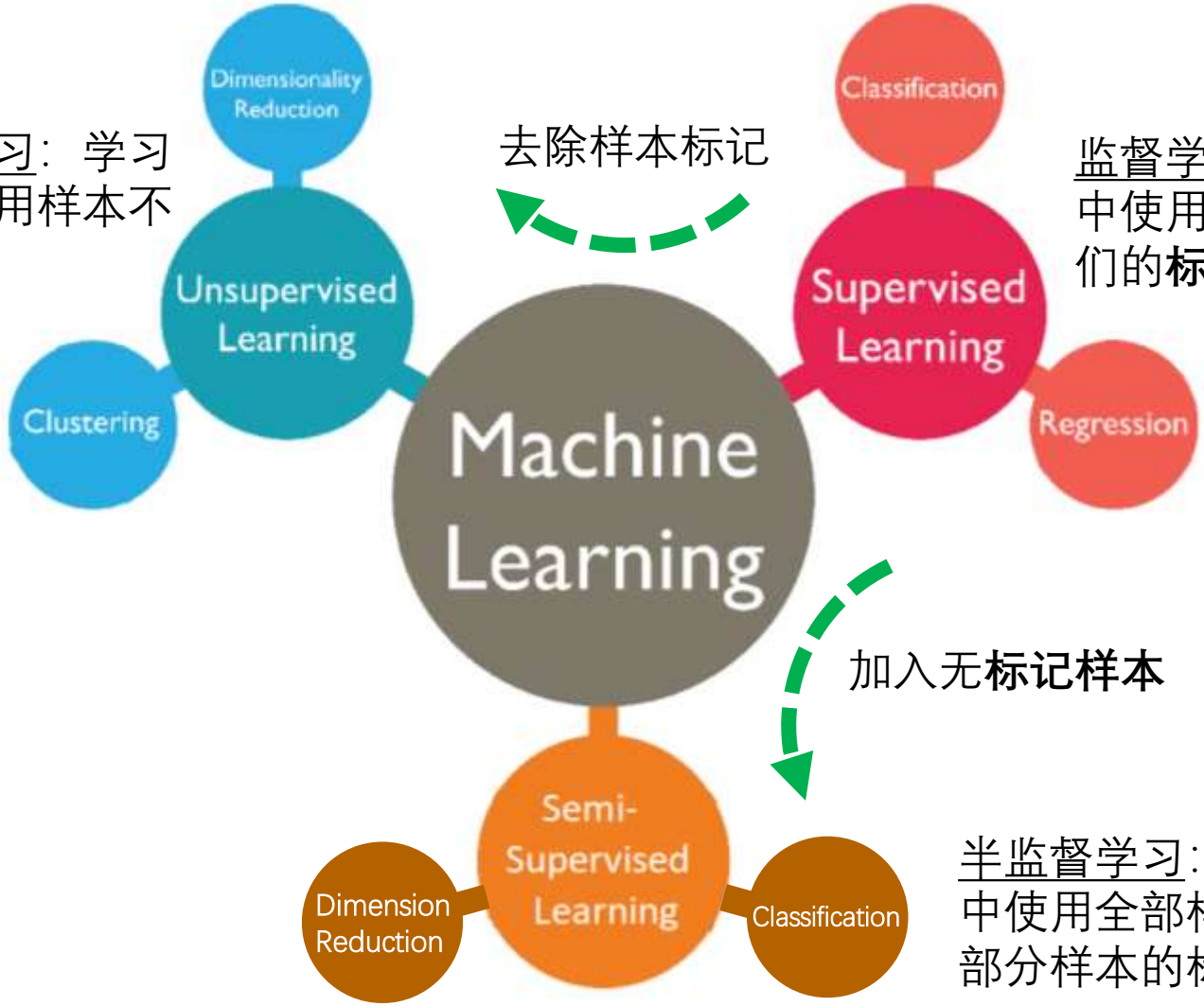
  一号瓜：y_1=是

  二号瓜：y_2=否

- 特征分为符号型和数值型。通常使用某种编码把符号特征转换为数值

  青绿：1， 浅白：2， 乌黑：3 || 好瓜：1， 差瓜：0

机器学习方法分类



无监督学习：学习过程只使用样本不使用标记

去除样本标记

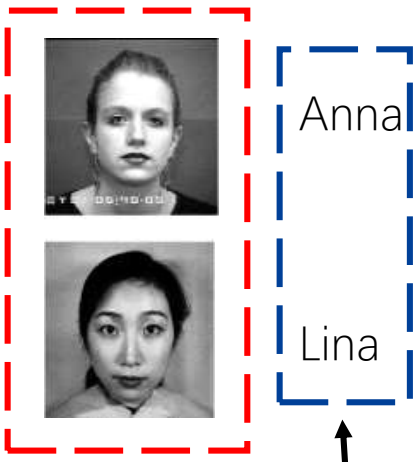
监督学习：学习过程中使用所有**样本**和它们的**标记**

加入无**标记样本**

半监督学习：学习过程中使用全部样本和一小部分样本的标记

监督学习（分类）

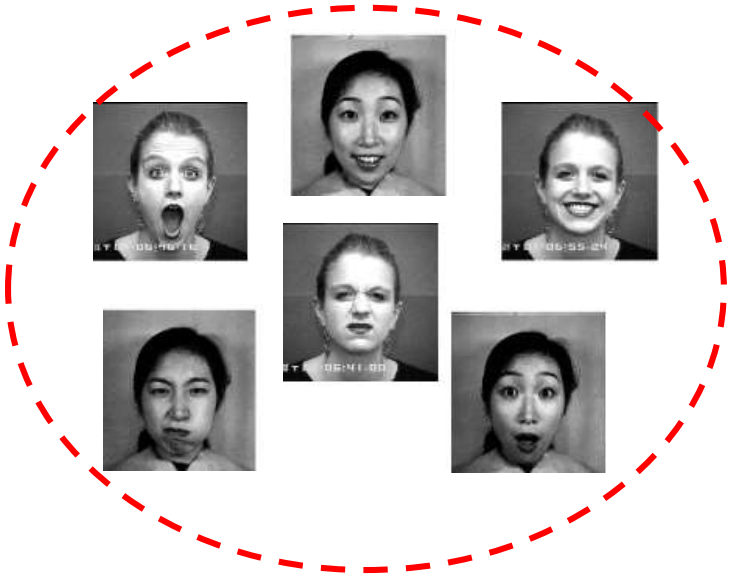
- 定义：按照给定的例子对数据进行分类



样本 $\mathbf{x}_1$ $\mathbf{x}_2$

标签 $y_1=1$ （Anna） $y_2=2$ （Lina）

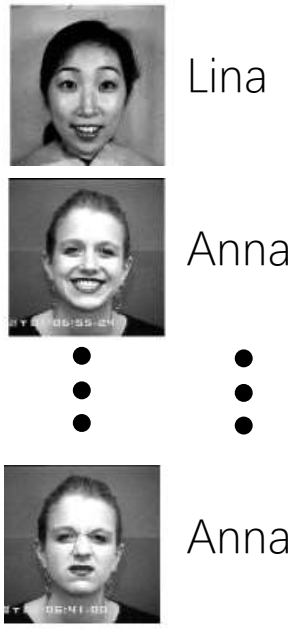
训练集={标记样本$(\mathbf{x}, y)$}



分别是谁？

测试样本 $\mathbf{x}_3, \mathbf{x}_4, \ldots, \mathbf{x}_8$

测试集={无标记样本 $\mathbf{x}$ }

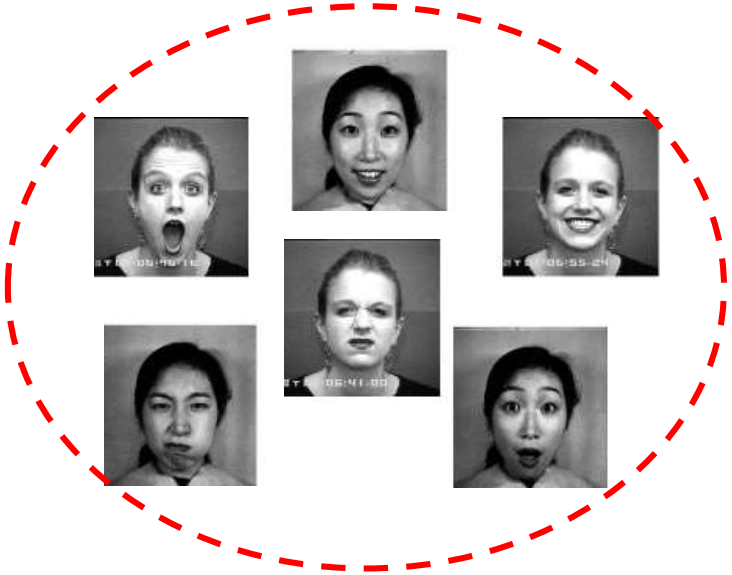


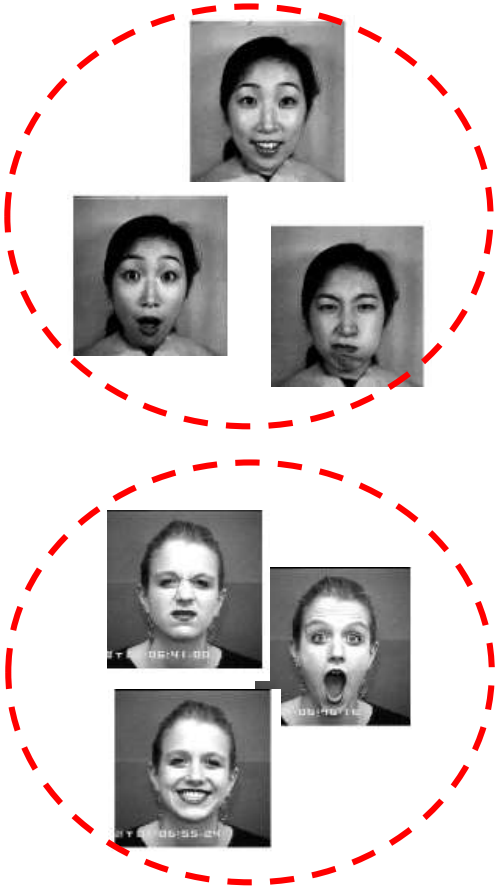
输出测试标签

$y_3, y_4, \ldots, y_8$

无监督学习（聚类）

- 定义：给一组数据，对它们进行归类



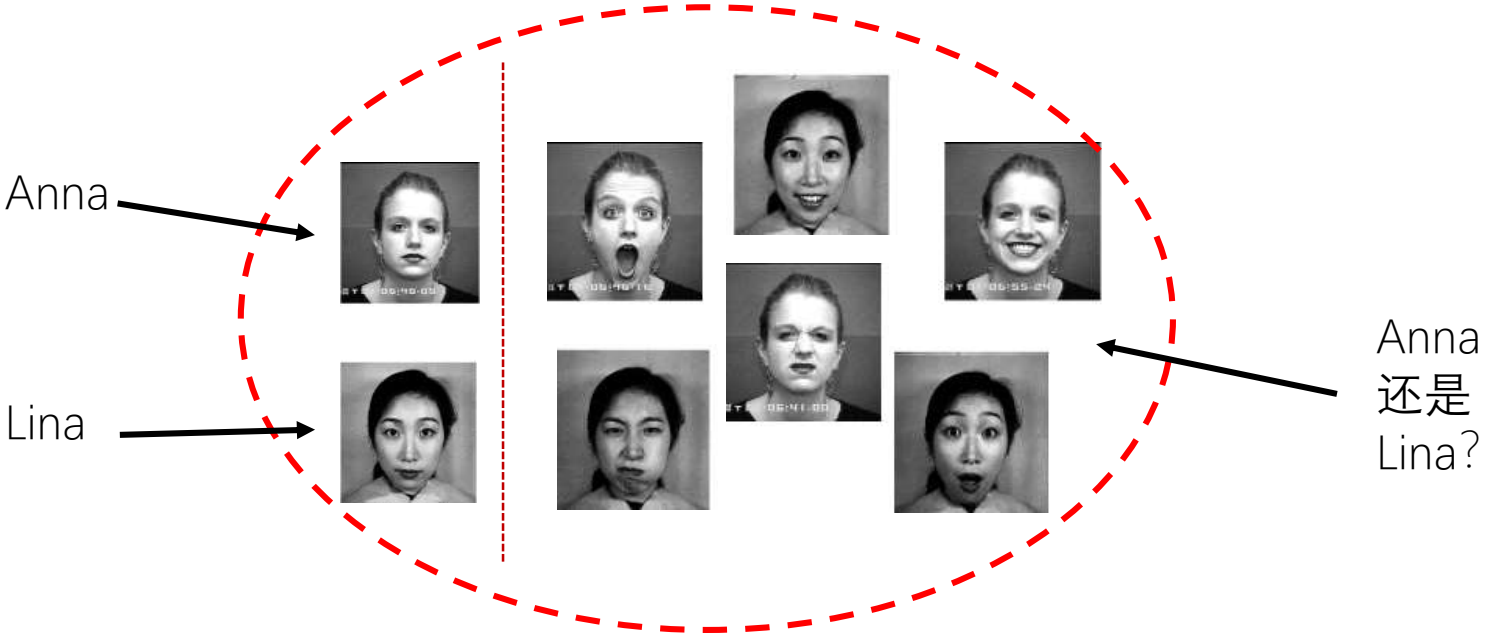


有几类（几个人）？

包括两个人，是谁不知道

半监督学习

- 定义:给定**极少**标记样本和**大量**无标记样本，对无标记样本进行分类



- 优势：标记样本少，结果质量较高，后续分类方便（部分算法）
- 劣势：计算量通常较大，有假设条件限制
- 监督学习区别：训练中用不用无标记样本（本质），标记样本数量多少

机器学习要素

任务类型	样本数据	学习算法
• 分类 • 聚类 • 回归 • ……	• 收集 • 变换 • 筛选 • ……	• SVM • k-means • kNN • ……

任务类型：确定合适的任务类型，构建目标函数

数据样本：确定对象的属性，测量相应数值

学习算法：根据目标函数性质，选择适当学习算法

通常任务的目标决定了可采用的学习算法类型

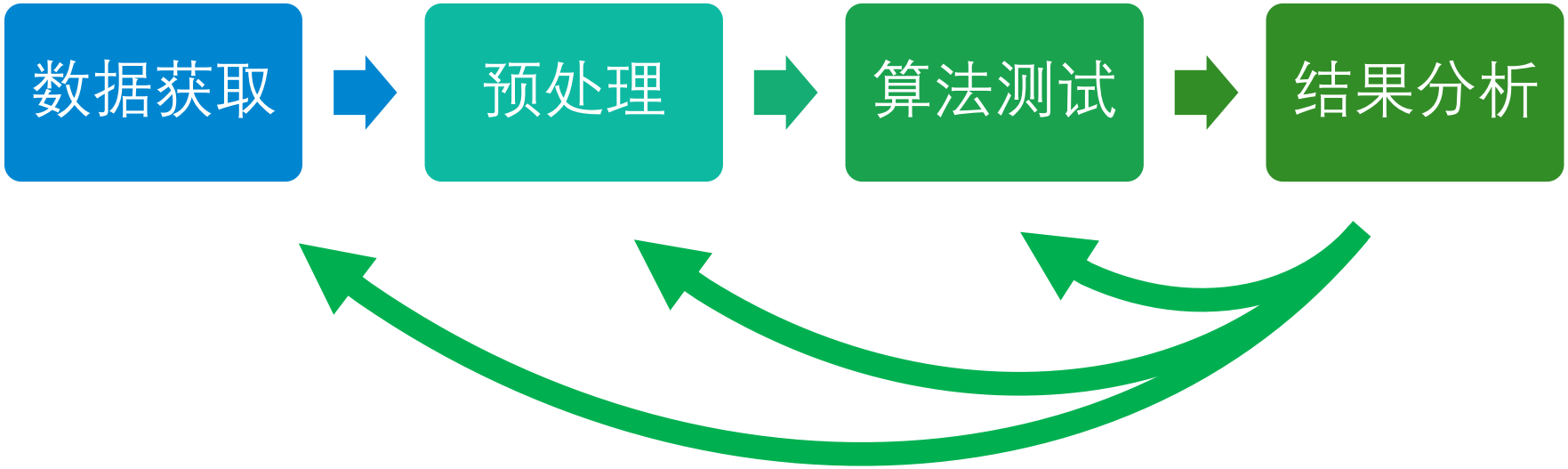
机器学习要素

任务类型	样本数据	学习算法
• 分类 • 聚类 • 回归 • ……	• 收集 • 变换 • 筛选 • ……	• SVM • k-means • kNN • ……

以人脸识别为例

- 任务类型：人脸图像分类（判断是否属于同一个人）
- 样本数据：拍摄的人脸图像以及提取脸部关键点的特征
- 学习算法：监督学习算法（SVM, kNN，深度学习等）

机器学习流程



根据分析结果，可返回到之前任一步进行调整，然后继续下一步执行，直到结果达到要求。

机器学习流程

数据获取 → 预处理 → 算法测试 → 结果分析

根据任务需求，**亲自**采集生产生活中的数据

- 拍摄照片：人脸识别，虹膜识别，车牌识别
- 录制视频：场景监控，动作分析，视频分析
- 仪表记录：机器状态记录，系统参数变化
- 问卷调查：街头问卷，网上调查
- 网上收集：股票交易，汇率变化，天气数据
- ……

或者是利用**已有**的数据集(免费或者购买)

- 开放数据库： 如 Kaggle，UCI Machine Learning Repository
- 商业公司：任务相关公司（如LinkedIn, 淘宝，腾讯）
- 权威机构：研究机构，官方数据
- ……

机器学习流程

数据获取 → 预处理 → 算法测试 → 结果分析

原始数据往往不能直接使用：

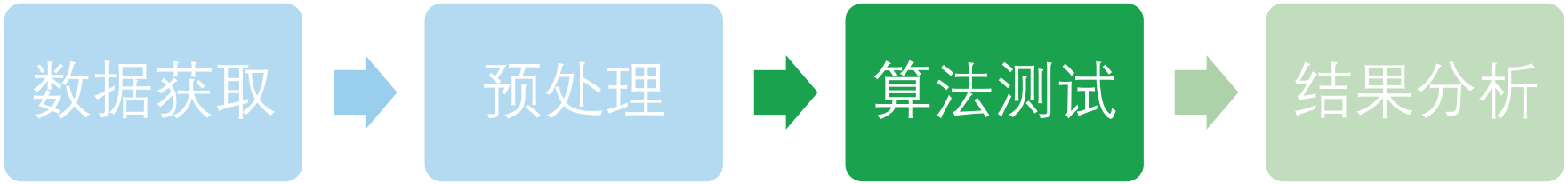
- 数据格式标准化：归一化，大小一致
- 不良样本剔除：数据缺失较多，数据偏差较大
- 噪声抑制或去除：消除干扰因素，平滑去噪
- 数据变换：特征选择，特征降维

处理后的样本集合称为数据集 $(\mathcal{X},\mathcal{Y})$ 用于后续的算法测试。

$$\mathcal{X}=\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n\}$$ 样本集合

$$\mathcal{Y}=\{y_1,y_2,...,y_n\}$$ 标记集合

机器学习流程



第一步：**设定参数**；第二步：**划分数据集**；第三步：**运行算法**

有监督学习，通常把数据集 $(\mathcal{X},\mathcal{Y})$ 划分为三个部分：

- 训练集 (Training set)：　训练机器学习算法
- 验证集 (Validation set)：在算法训练中做参数和学习进度验证（可没有）
- 测试集 (Testing set)：　训练结束后做模型验证

例如有1000个数据样本：训练集包括700个，验证集包括100个，测试集包括200个。

通常：训练集>测试集>验证集

机器学习流程

数据获取 → 预处理 → 算法测试 → 结果分析

在测试集上评价学习结果的好坏:

$$\text{错误率} = \frac{\text{错分个数}}{\text{总测试样本数}} \times 100\%$$

$$\text{精度} = \frac{\text{正确个数}}{\text{总测试样本数}} \times 100\%$$

$$= 100 - \text{错误}$$

此外还有: 混淆矩阵, 准确率, 召回率

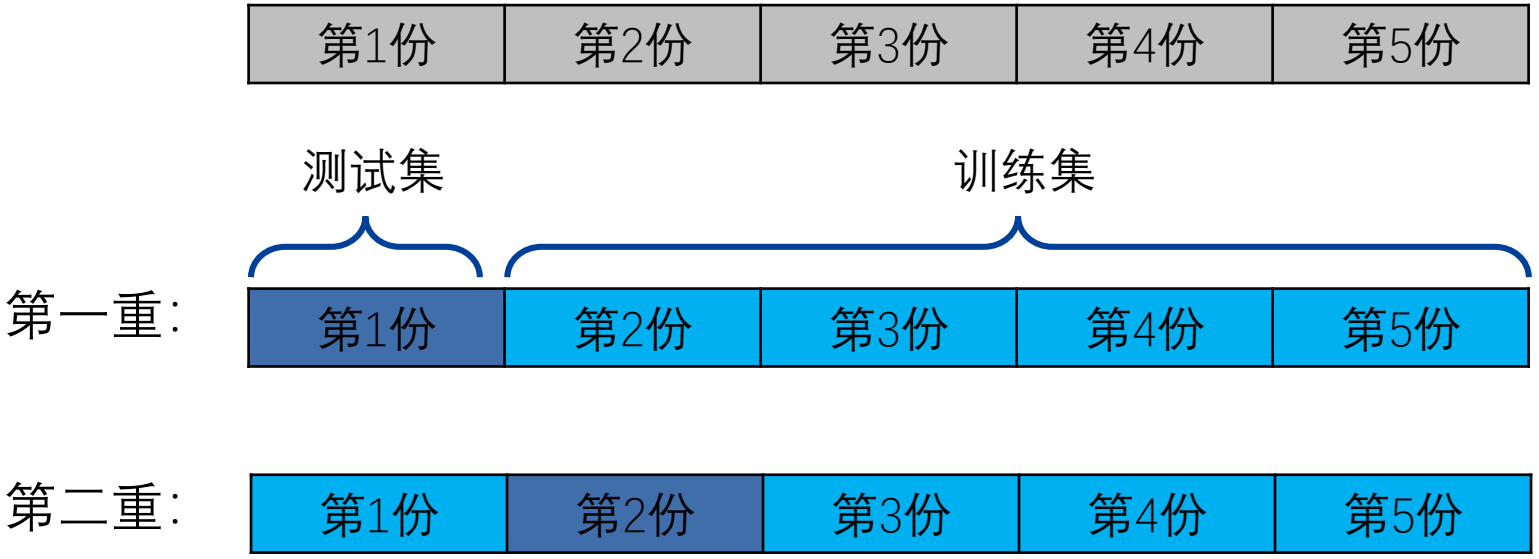
(i, j) 真实类别为i, 预测类别为 j

预测类别

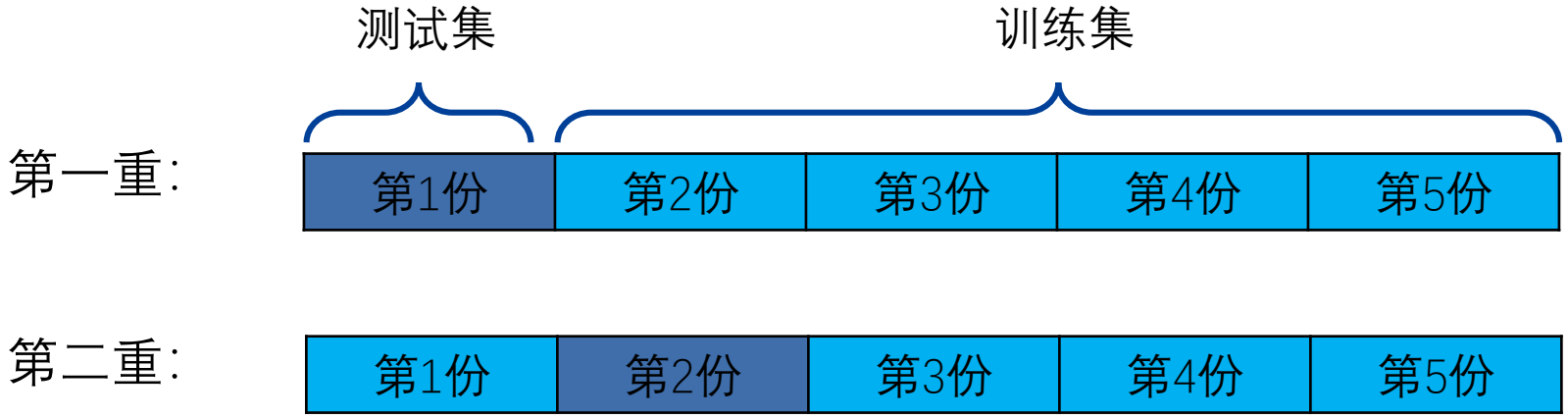
真实类别	类别1	类别2	……	类别K
类别1	10	0	1	0
类别2	1	9	0	2
……	0	0	8	
类别K	2	1	0	9

k重交叉验证

- *训练集-测试集* 的简单划分方式容易因为划分不合理而产生评价不准确。
- k重交叉验证：把数据集等分为k份，然后逐份作为测试集、余下做为训练集做模型测试，例如$k=5$



k重交叉验证



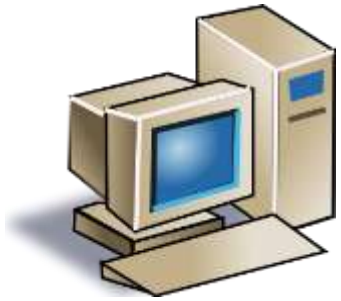
最终误差：k个测试集的误差平均

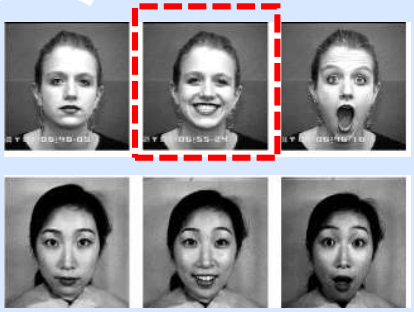
作为极端情况，k取整个数据集大小，每一份只包含一个样本。这种方法称为留一法（即每次只留一个样本作为测试）。

优势：每个样本都有机会作为测试样本

劣势：因为要反复训练测试k次，计算量大

以人脸识别为例





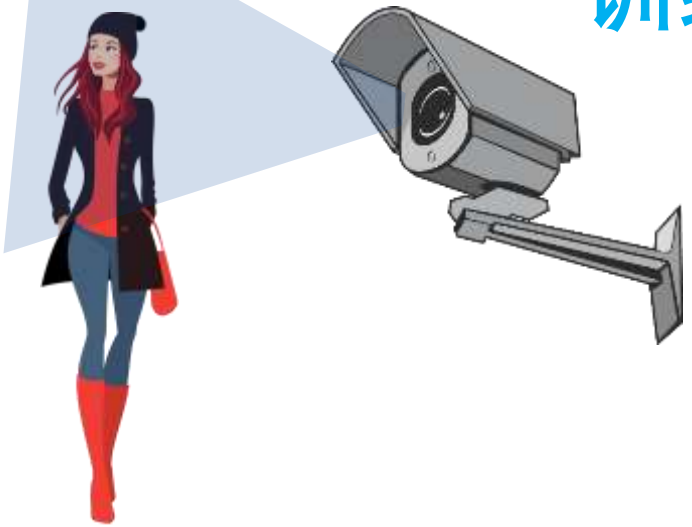
Anna

Lina

训练样本

比对

Anna



测试样本

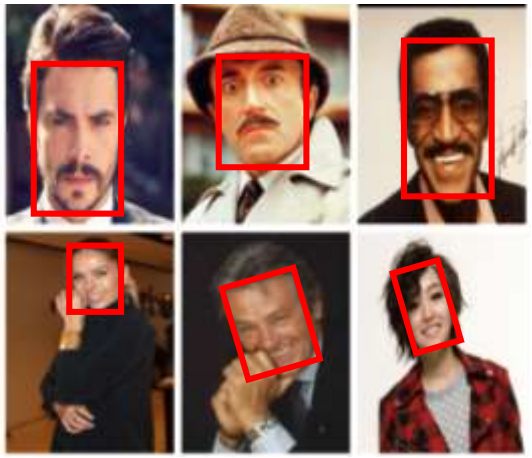
人脸识别为例

数据获取 → 预处理 → 学习测试 → 结果分析

拍摄人脸照片

人脸识别为例

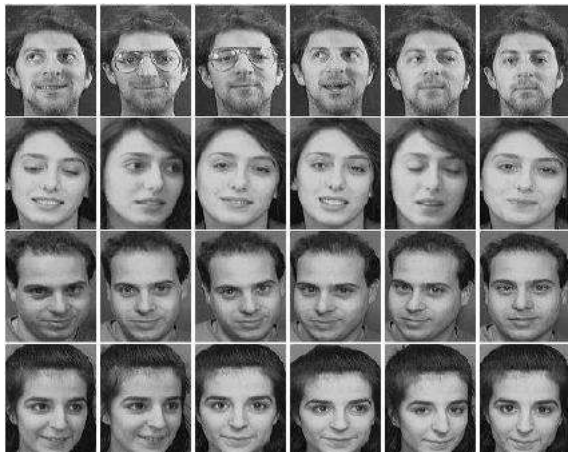
数据获取 → 预处理 → 学习测试 → 结果分析



- 裁切脸部图像
- 图像大小统一
- 转灰度图像
- 脸部旋转端正
- ……

人脸识别为例

数据获取 → 预处理 → 学习测试 → 结果分析



步骤1：训练

机器学习算法
(如deep learning)

y_1
y_2
y_3
...
y_n

步骤2： 测试

人脸识别为例

数据获取 → 预处理 → 学习测试 → 结果分析

$$\text{错误率} = \frac{\text{错分个数}}{\text{总测试样本数}} \times 100\%$$

假设总共200个样板，分错5个

$$\text{错误率} = \frac{5}{200} \times 100\% = 2.5\%$$

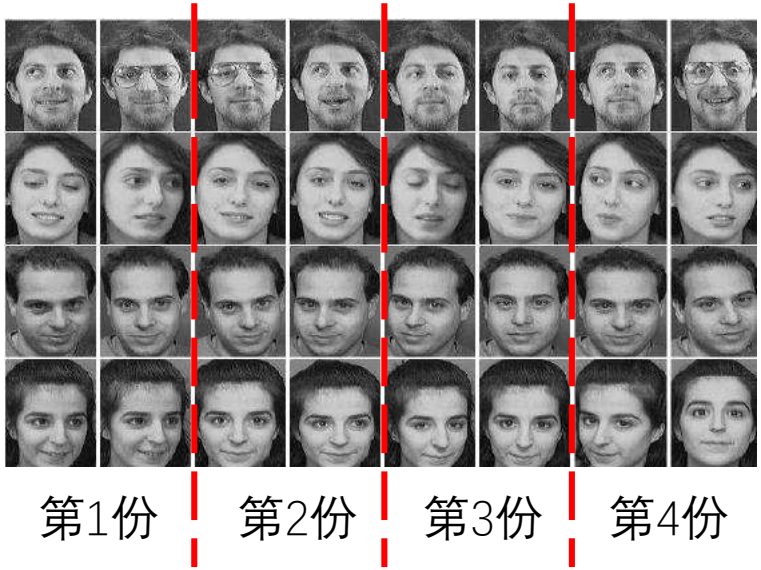
行和等于该类测试样本总个数

预测类别

真实类别	小明	翠花	Anna	Lucy
小明	49	0	1	0
翠花	1	45	2	2
Anna	0	0	50	
Lucy	1	1	0	48

人脸识别为例

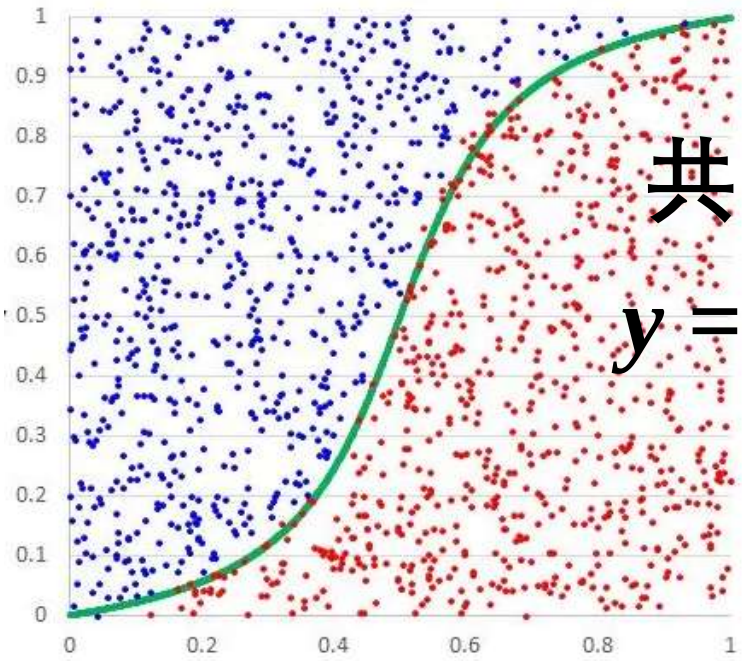
k重交叉验证



……

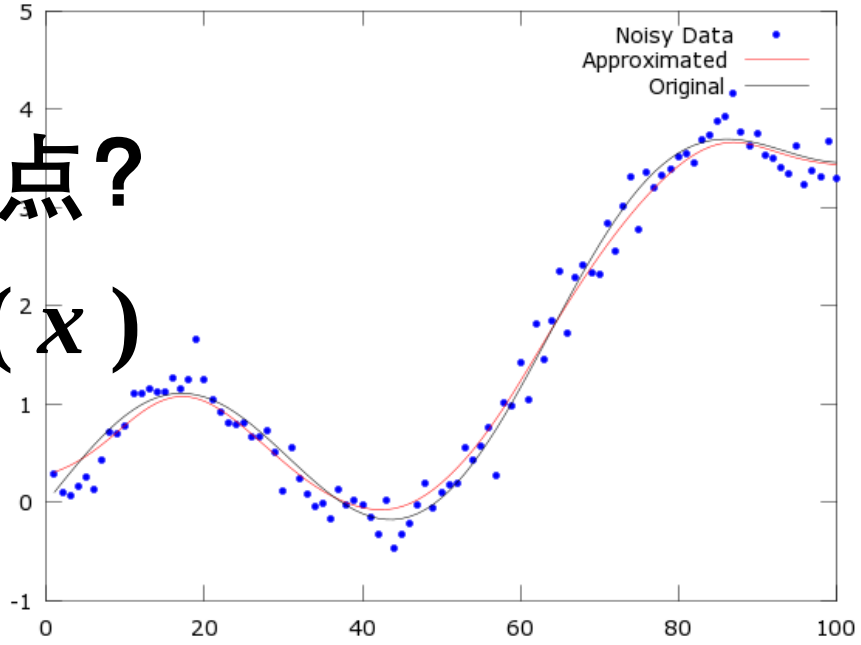
分类与拟合

- **分类**：目标值为**无序**的离散值或者符号值，学习的目标是把样本按照不同目标值区分开
- **拟合**（回归）：目标值通常为**有序**的连续变量，学习的目标是对给定样本求出其目标值对应的位置。



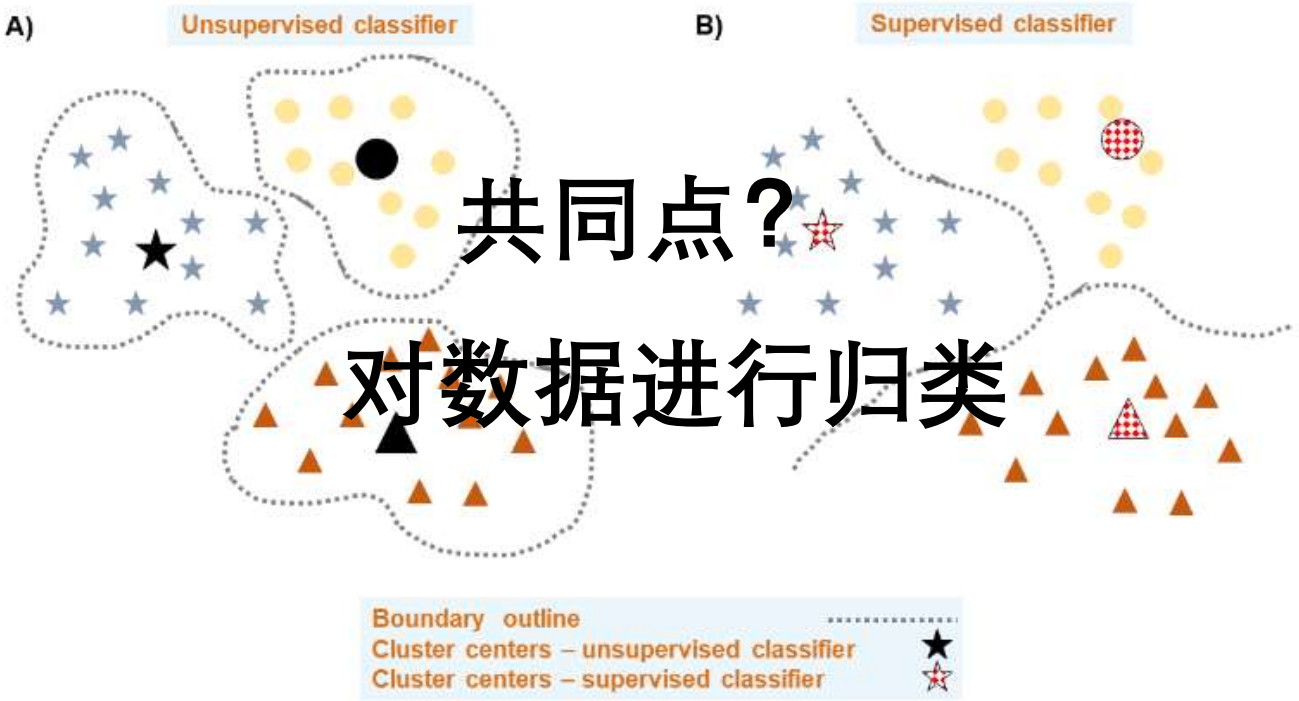
共同点？

$$y = f(x)$$



聚类与分类

- **聚类**：无监督学习，学习目标是把输入样本按照**相互间的相似度**（距离）划分为若干个互补重叠的类（组）里
- **分类**：监督学习，学习的目标是把输入样本按照**与给定标记样本的相似性**（距离）划分到不同的类（组）里。



向量函数求导

- 内积: $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x}, \quad \frac{\mathrm{d}f}{\mathrm{d}\mathbf{x}} = \mathbf{a}; \quad (f(x) = ax, \quad \frac{\mathrm{d}f}{\mathrm{d}x} = a)$

证明: $f(\mathbf{x}) = a_1x_1 + a_2x_2 + \ldots + a_dx_d$

$$\frac{\partial f}{\partial x_1} = a_1, \quad \frac{\partial f}{\partial x_2} = a_2 \quad \ldots, \quad \frac{\partial f}{\partial x_d} = a_d$$

$$\frac{\mathrm{d}f}{\mathrm{d}\mathbf{x}} = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_d}\right)^T = (a_1, a_2, \ldots, a_d)^T$$

- 2范数平方: $f(\mathbf{x}) = \|\mathbf{x}\|_2^2, \quad \frac{\mathrm{d}f}{\mathrm{d}\mathbf{x}} = 2\mathbf{x} \quad (f = x^2, \frac{\mathrm{d}f}{\mathrm{d}x} = 2x)$

向量函数求导

- 范数平方: $f(\mathbf{x}) = \|\mathbf{x} \pm \mathbf{a}\|_2^2 = (\mathbf{x} \pm \mathbf{a})^T (\mathbf{x} \pm \mathbf{a}), \quad \frac{df}{d\mathbf{x}} = 2(\mathbf{x} \pm \mathbf{a});$

$$\left(f(x) = (x-a)^2, \quad \frac{df}{dx} = 2(x-a) \right)$$

- 一次项范数平方: $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} \pm \mathbf{b}\|_2^2, \quad \frac{df}{d\mathbf{x}} = 2\mathbf{A}^T (\mathbf{A}\mathbf{x} \pm \mathbf{b})$

- 二项式: $f(\mathbf{x}) = \mathbf{x}^T \mathbf{A}\mathbf{x}, \quad \frac{df}{d\mathbf{x}} = (\mathbf{A}^T + \mathbf{A})\mathbf{x} \quad (2\mathbf{A}\mathbf{x} \text{ if } \mathbf{A} \text{ is symmetric})$

接下来……

贝叶斯学习