河北师范大学软件学院
Software College of Hebei Normal University

机器学习入门
Machine Learning

第一课 绪论

河北师范大学 软件学院

2018年3月5日

主要内容

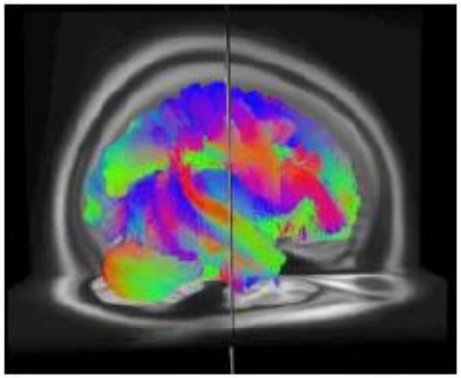
1. 机器学习的引入
2. 机器学习的典型类型及任务
3. 机器学习的基本组成
4. 机器学习的相关术语
5. 模型的评价及选择
6. 本学期学习模块

1.1 为什么要机器学习？

Era of Big Data

We are drowning in information and starving for knowledge. — John Naisbitt.

CERN Collider
320 x 10^{12} bytes/second



Prof. Tim Verstynen, CMU

Personal Connectome
10^{18} bytes/human

facebook

1 billion
messages/day

twitter

200 million
tweets/day

"Every day, people create the equivalent of 2.5 **quintillion** bytes of data from sensors, mobile devices, online transactions, and social networks; so much that 90 percent of the world's data has been generated in the past two years."

The Huffington Post: Arnal Dayaratna: IBM Releases Big Data

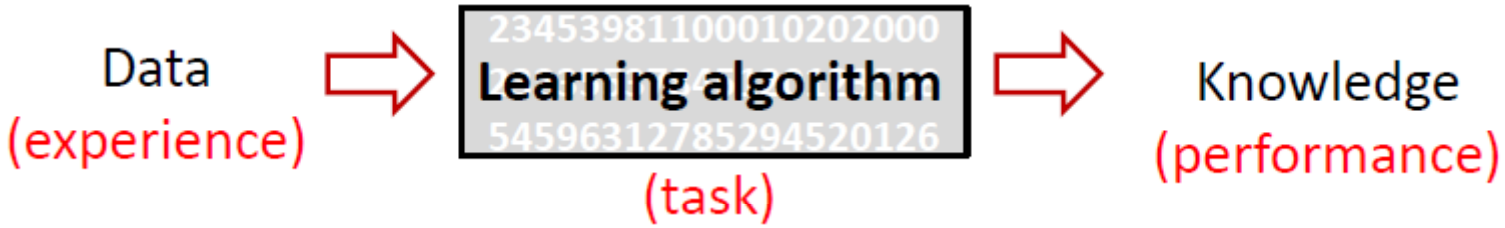
DATA ≠ KNOWLEDGE

What is Machine Learning?

Machine learning, a branch of artificial intelligence, is a scientific discipline concerned with **the design and development** of **algorithms** that take as input empirical data, and **yield** patterns or predictions thought to be features of the underlying mechanism that generated the data.

1.2 什么是机器学习

- Study of algorithms that
  - (automatically) improve their performance
  - at some task
  - with experience



- Machine Learning seeks to develop theories and computer systems for

dealing with

- complex, real world data, based on the system's own experience with data, and (hopefully) under a unified model or mathematical framework, that

have nice properties.

- Machine Learning seeks to develop theories and computer systems for
  - representing;
  - classifying, clustering, recognizing, organizing;
  - reasoning under uncertainty;
  - predicting;
  - and reacting to
  - …
- complex, real world data, based on the system's own experience with data, and (hopefully) under a unified model or mathematical framework, that
  - can be formally characterized and analyzed;
  - can take into account human prior knowledge;
  - can generalize and adapt across data and domains;
  - can operate automatically and autonomously;
  - and can be interpreted and perceived by human.

- ML covers algorithms, theory and very exciting applications
- It's going to be fun and challenging ☺

主要内容

1. 机器学习的引入
2. 机器学习的典型类型及任务
3. 机器学习的基本组成
4. 机器学习的相关术语
5. 模型的评价及选择
6. 本学期学习模块

2.1 典型类型

两种主要学习类型

- **监督式学习**
  (supervised learning, 也称predictive learning)
  --目的在于精确预测
  --“预测性能”

基于给定的训练集$\boldsymbol{D}=\{(x_i,y_i),\boldsymbol{i}=1,...,\boldsymbol{N}\}$, 学习输入$\boldsymbol{x}$与输出$y$的关系，使得对于任意观测$\boldsymbol{x}^*$，该模型尽可能准确预测输出$\boldsymbol{y}^*$

2.1 典型类型

两种主要学习类型

- **非监督式学习**

(unsupervised learning,也称 descriptive learning)

--发现关于数据的紧致描述、知识发现

--“描述性能”

基于给定数据集$\boldsymbol{D}=\{x_i, i=1,...,N\}$, 寻找关于$\boldsymbol{D}$的更为紧致的描述

- 其它

如：强化学习 (reinforcement learning, RL)

--learning how to act or behave when given occasional reward or punishment signals.

2.2 典型的学习任务

(1) 分类(classification)

也称“模式识别”

- 两类别 vs. 多类别

类别数C=2， 两类别分类(binary classification)

类别数C>2， 多类别分类(multiclass classification)

- 产生式分类模型 vs. 鉴别式分类模型

- 线性分类器 vs. 非线性分类器

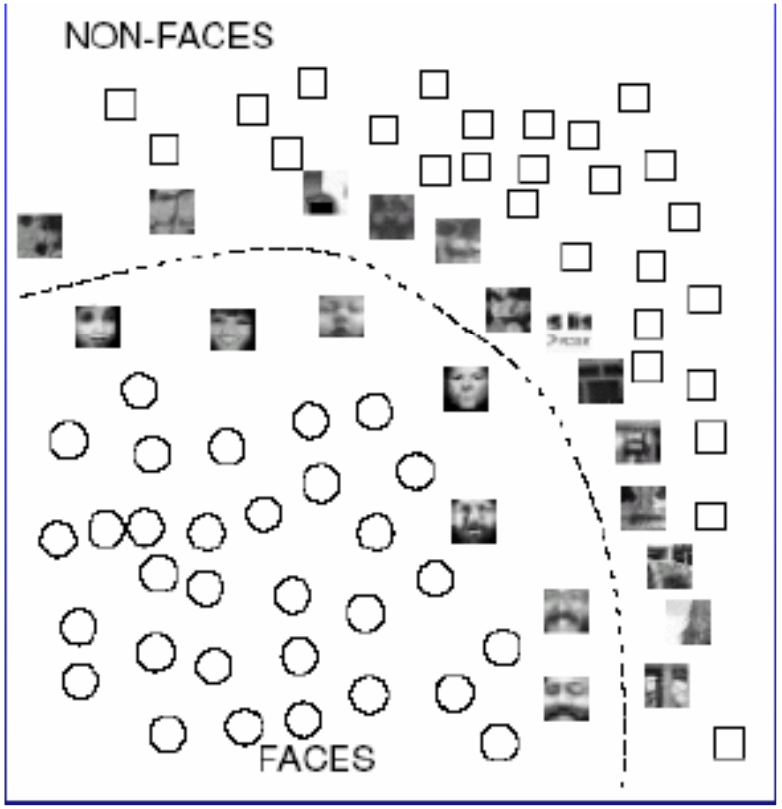
应用举例

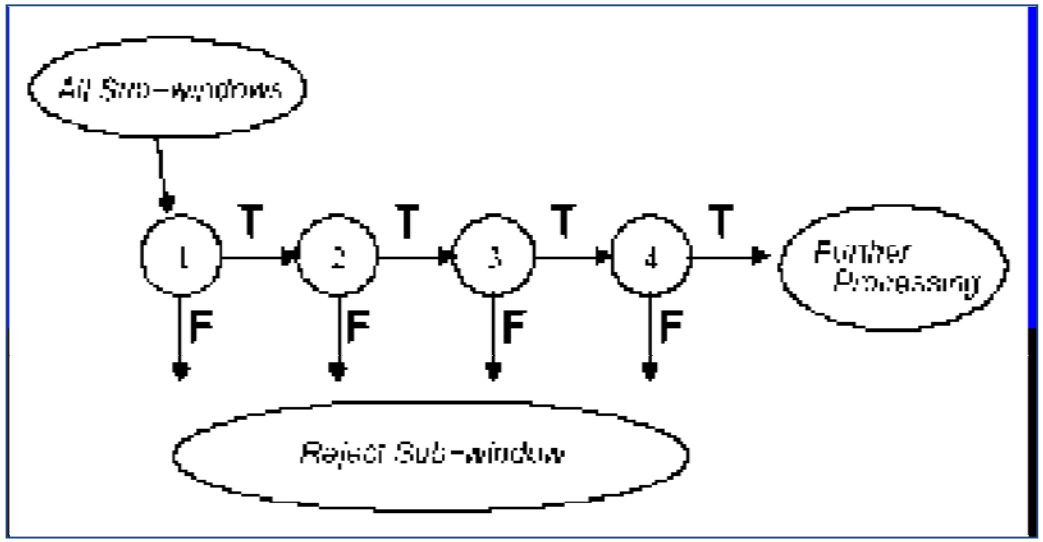
文档分类(document classification)

垃圾邮件过滤(spam filter)

人脸检测 (Face Detection)

Identify and locate human faces in an image regardless of their position, scale, in in-plane rotation orientation pose (or out-of-plane rotation) and illumination

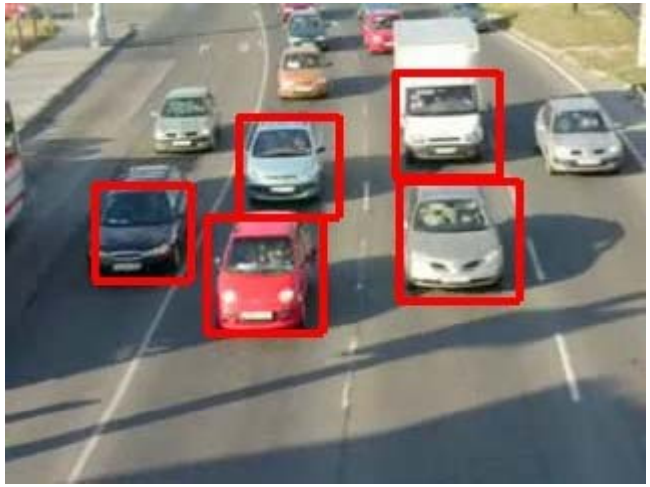


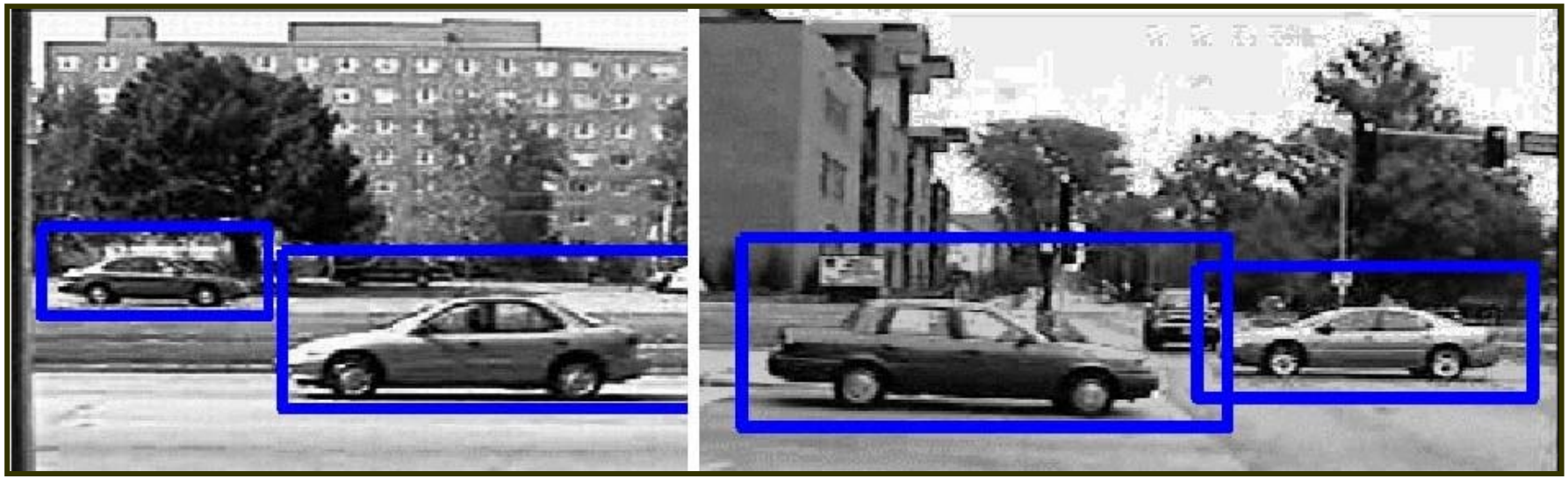

All Sub-windows
1
2
3
4
T
T
T
T
F
F
F
F
Further Processing
Reject Sub-window


A face undergoing significant pose variation.
A face undergoing significant changes in facial expression.
A face undergoing significant pose and scale variations.
A face partially occluded by hands.
A face partially occluded by a black folder.

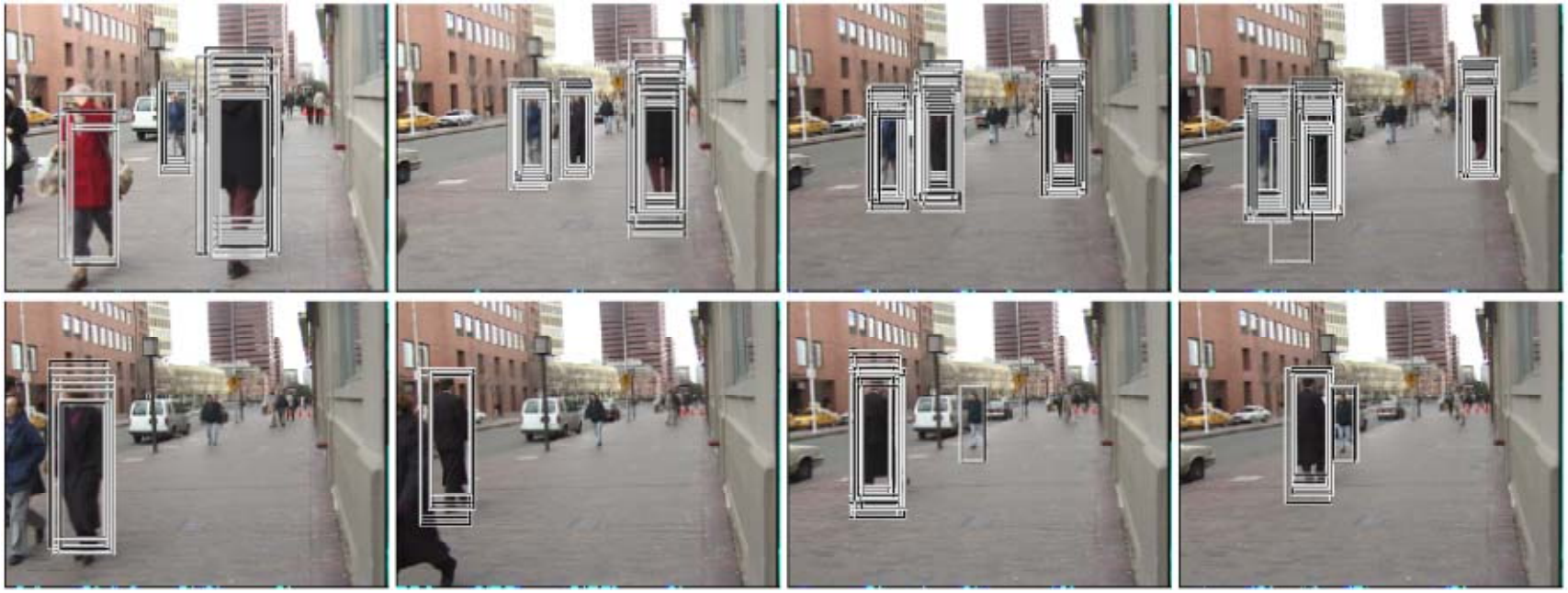
河北师范大学软件学院
Software College of Hebei Normal University

车辆检测



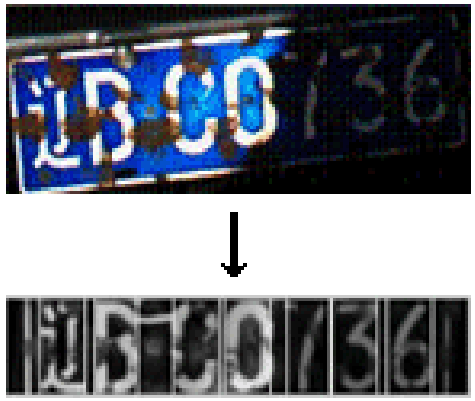


行人检测



光学字符识别(OCR)

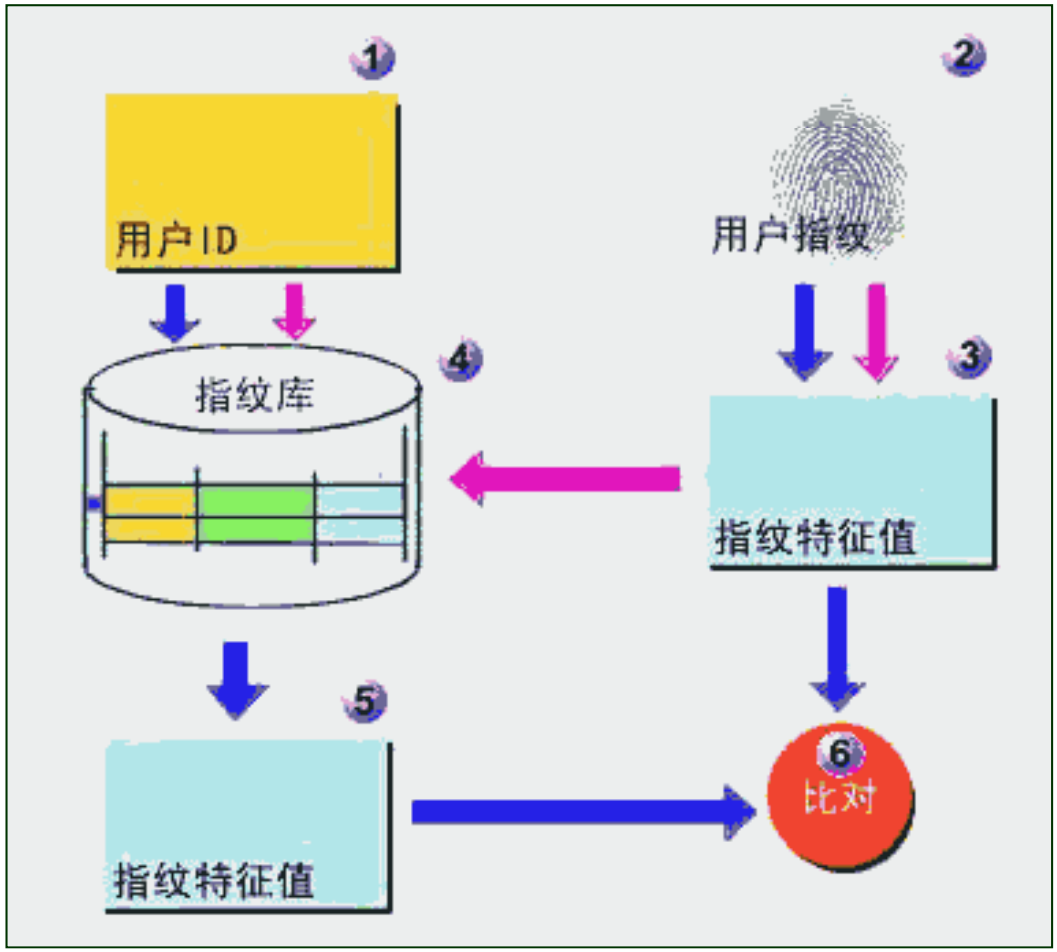
河北师范大学软件学院
Software College of Hebei Normal University

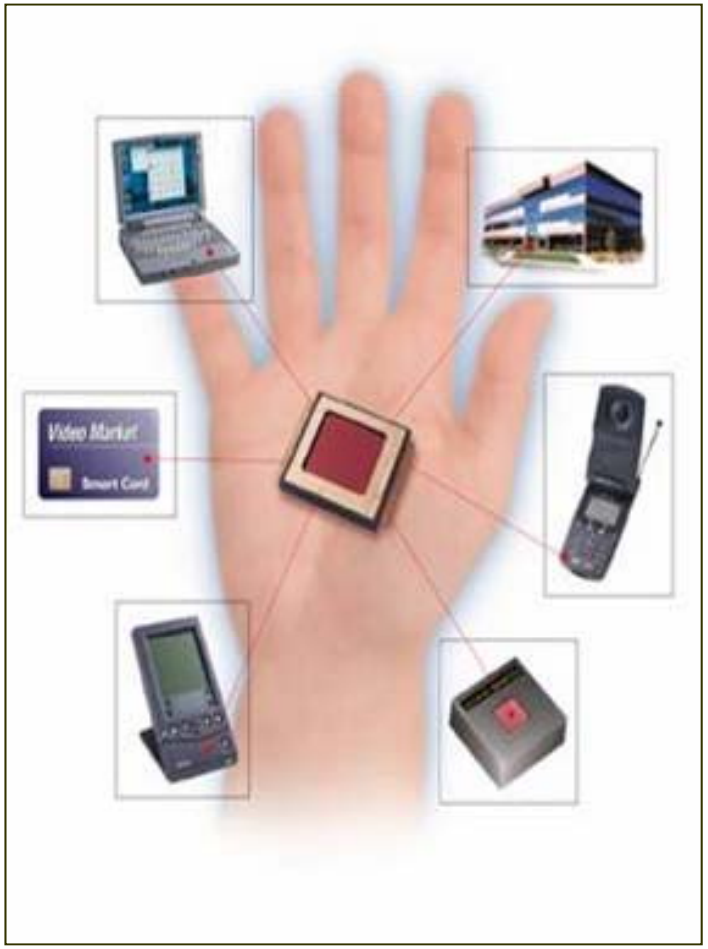


提取到的车牌

基于生物特征的身份识别

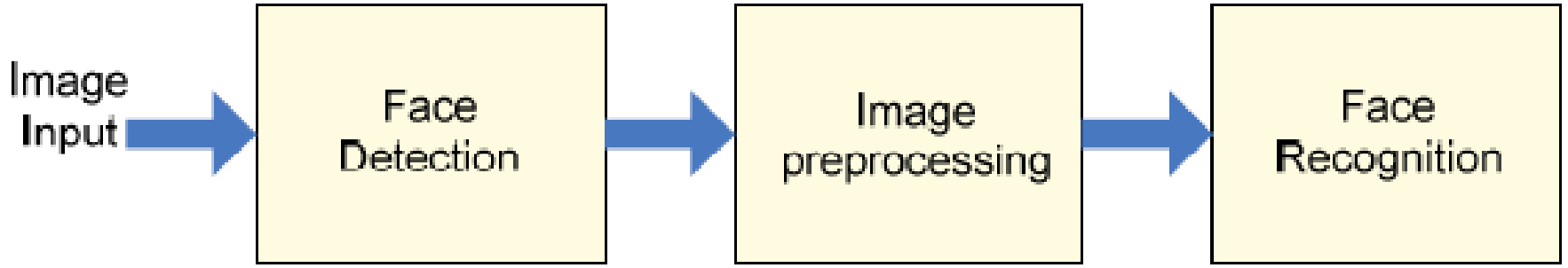
声音、人脸、虹膜、指纹、掌纹、步态,…





- 人脸识别 Face Recognition

- Determine the identity of a face in an image
- The image can be a frame from a video
- Processing needs to be fast
- Classification problem
- Need faces images for training



河北师范大学软件学院
Software College of Hebei Normal University

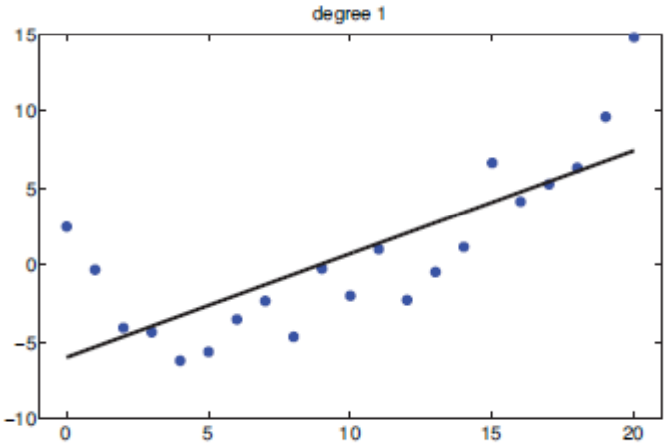
Examples from ImageNet

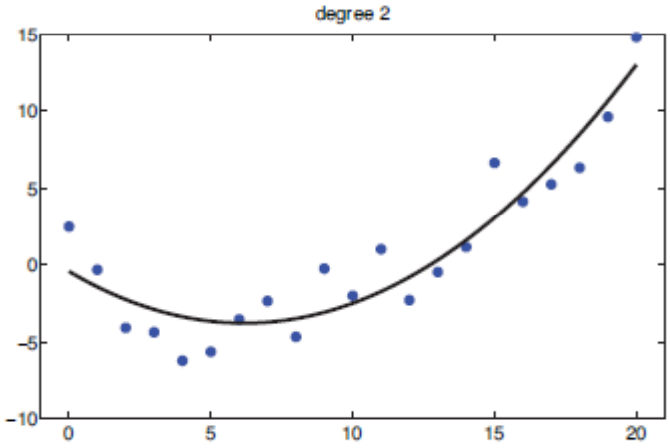
2.2 典型的学习任务

(2) 回归(regression)

实值函数回归 vs.顺序回归

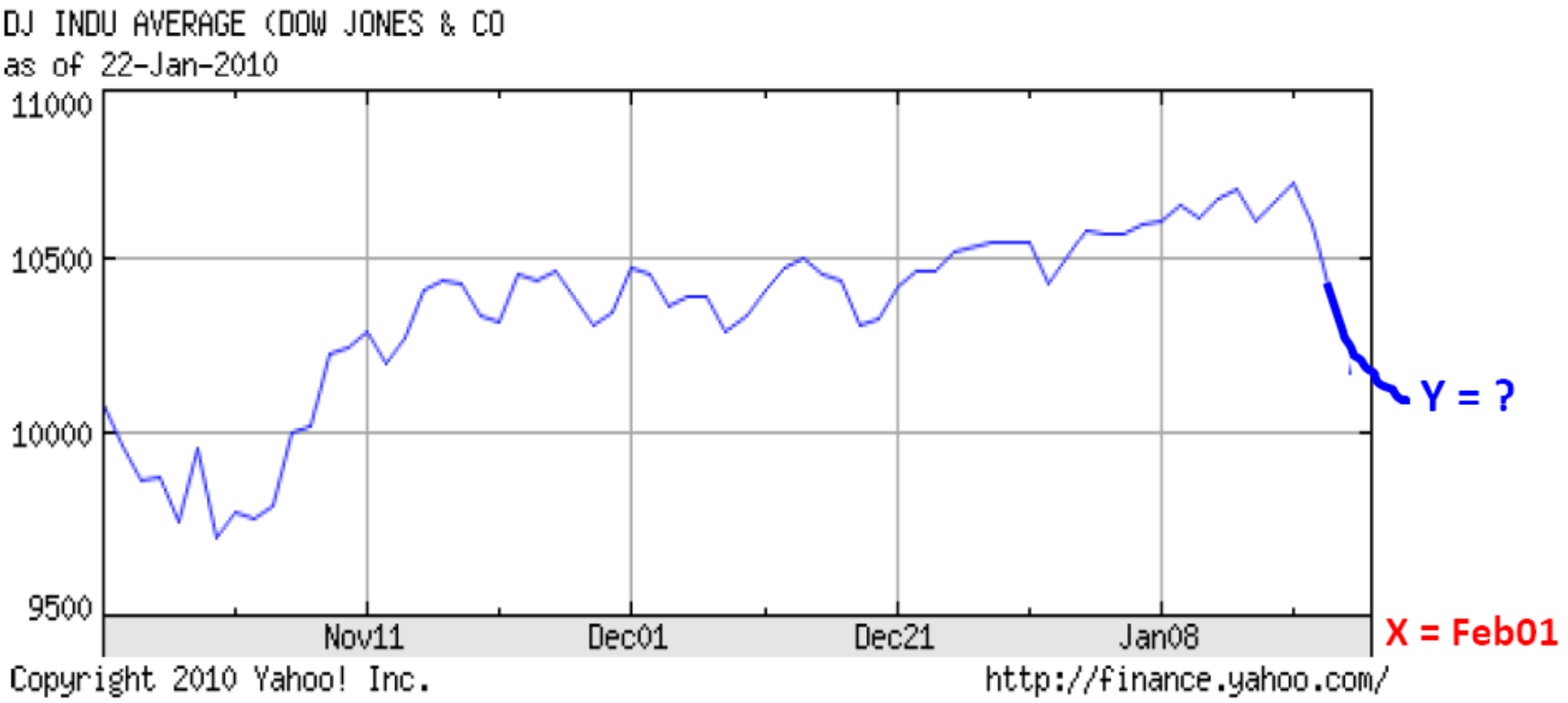
线性回归 vs.非线性回归





例：时间序列预测 --Stock market prediction

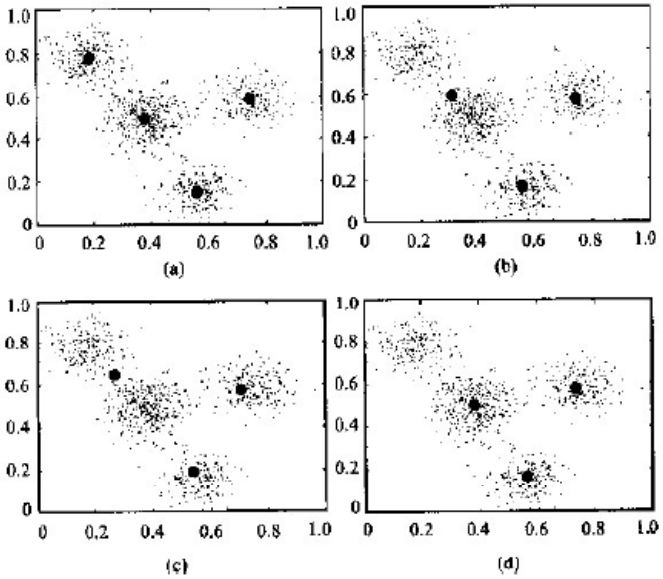
基于历史观测数据$\{(t, y(t)), t = 1, ..., T\}$预测未来时刻$y(T+1)$

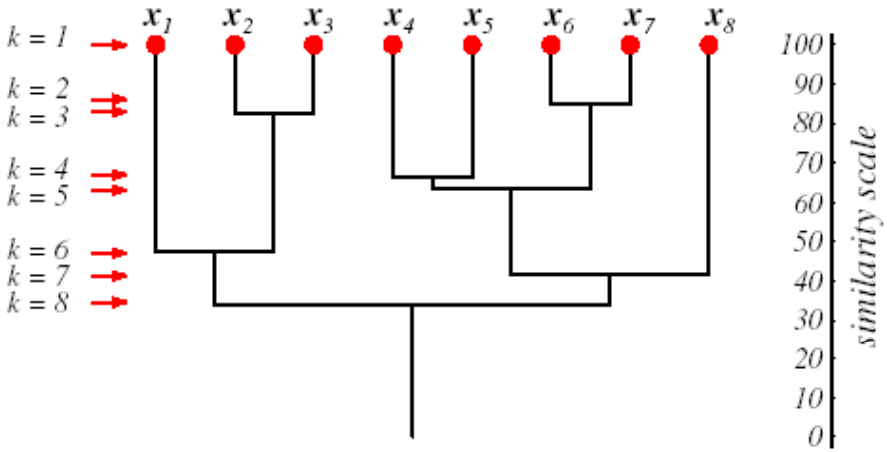


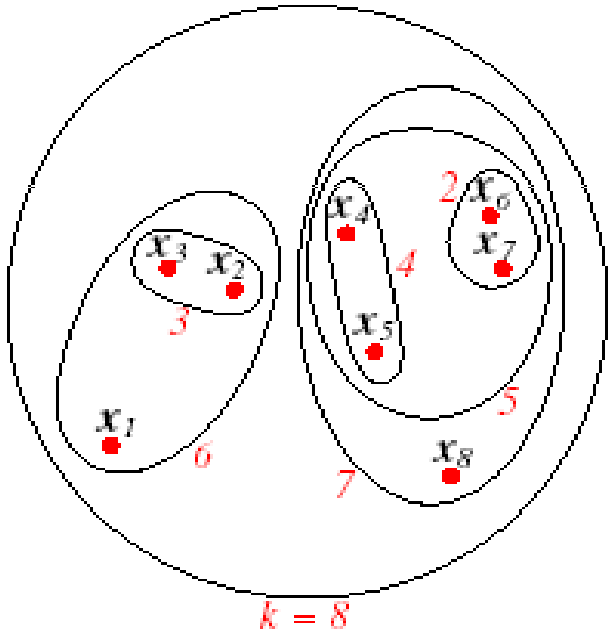
例：facial age estimation

2.2 典型的学习任务

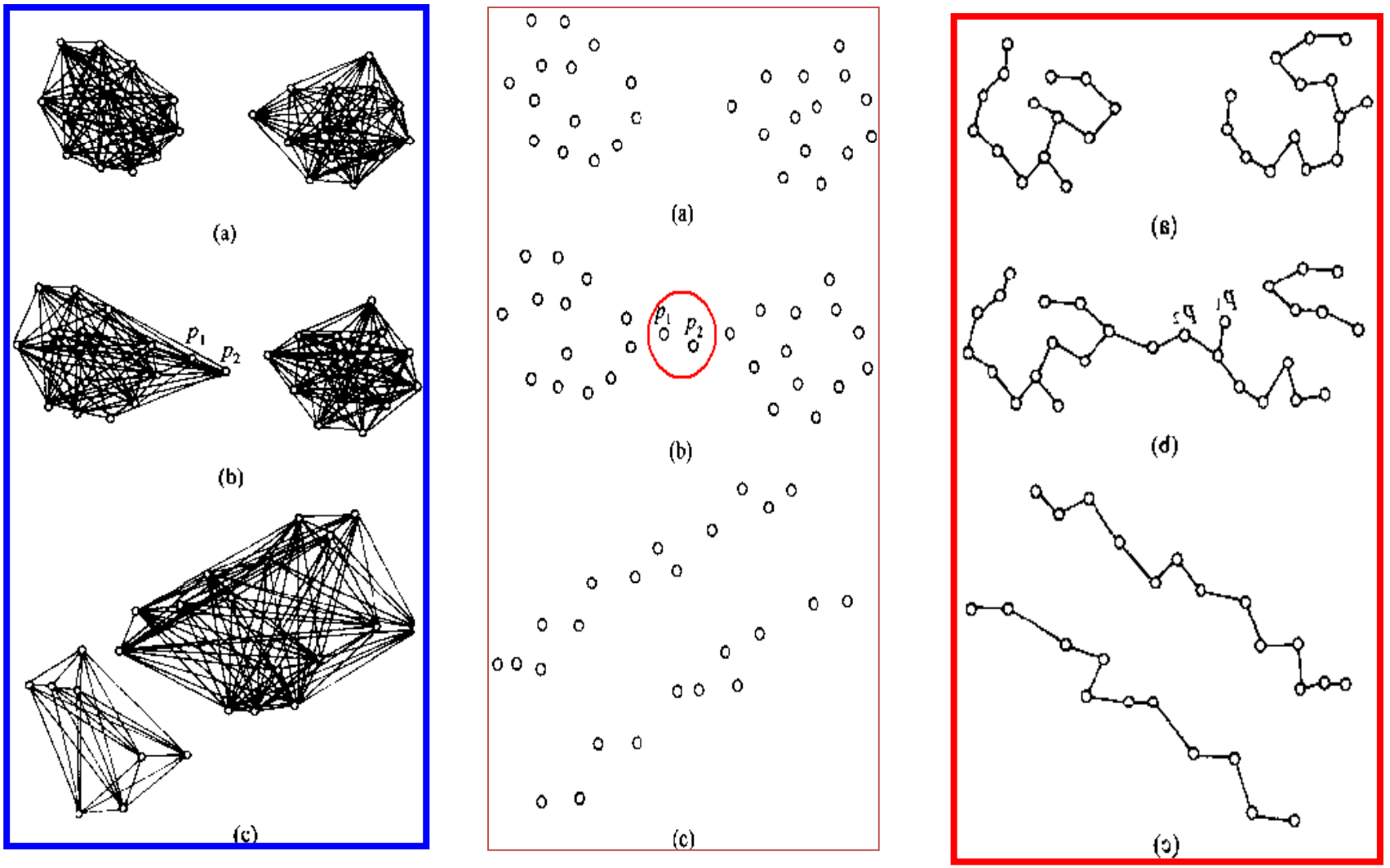
(3) 聚类(clustering)
--discovering cluster







最远距离法与最近距离法的聚类结果比较



-- novelty detection(anomaly detection)

例：婴儿对环境的感知

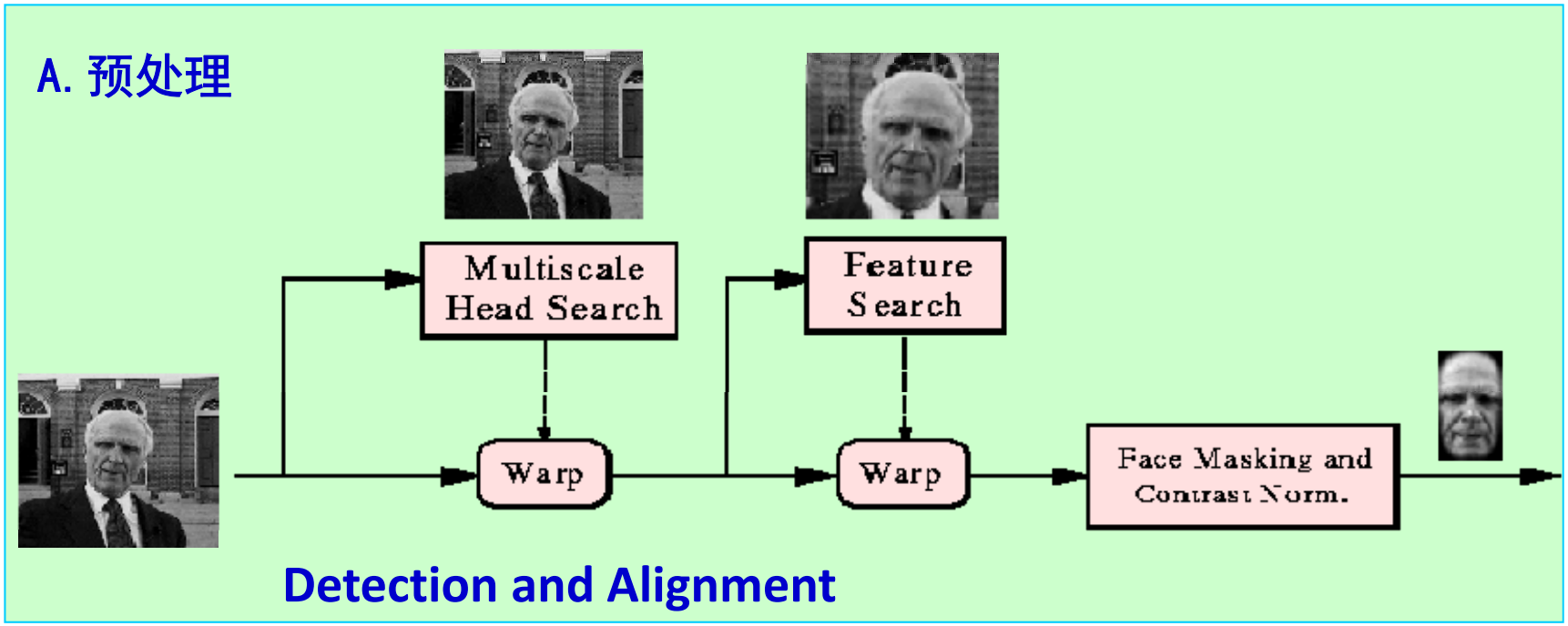
2.2 典型的学习任务

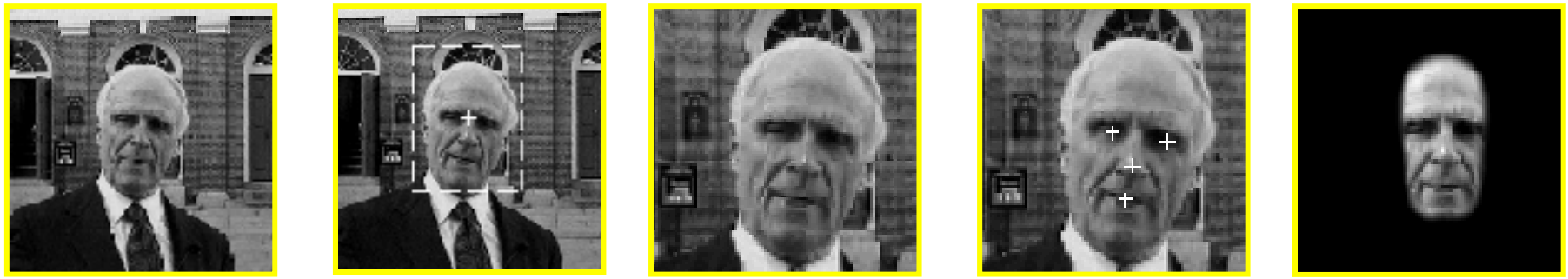
(4)特征表示

-- 特征降维，如PCA

--高维数据的低维可视化

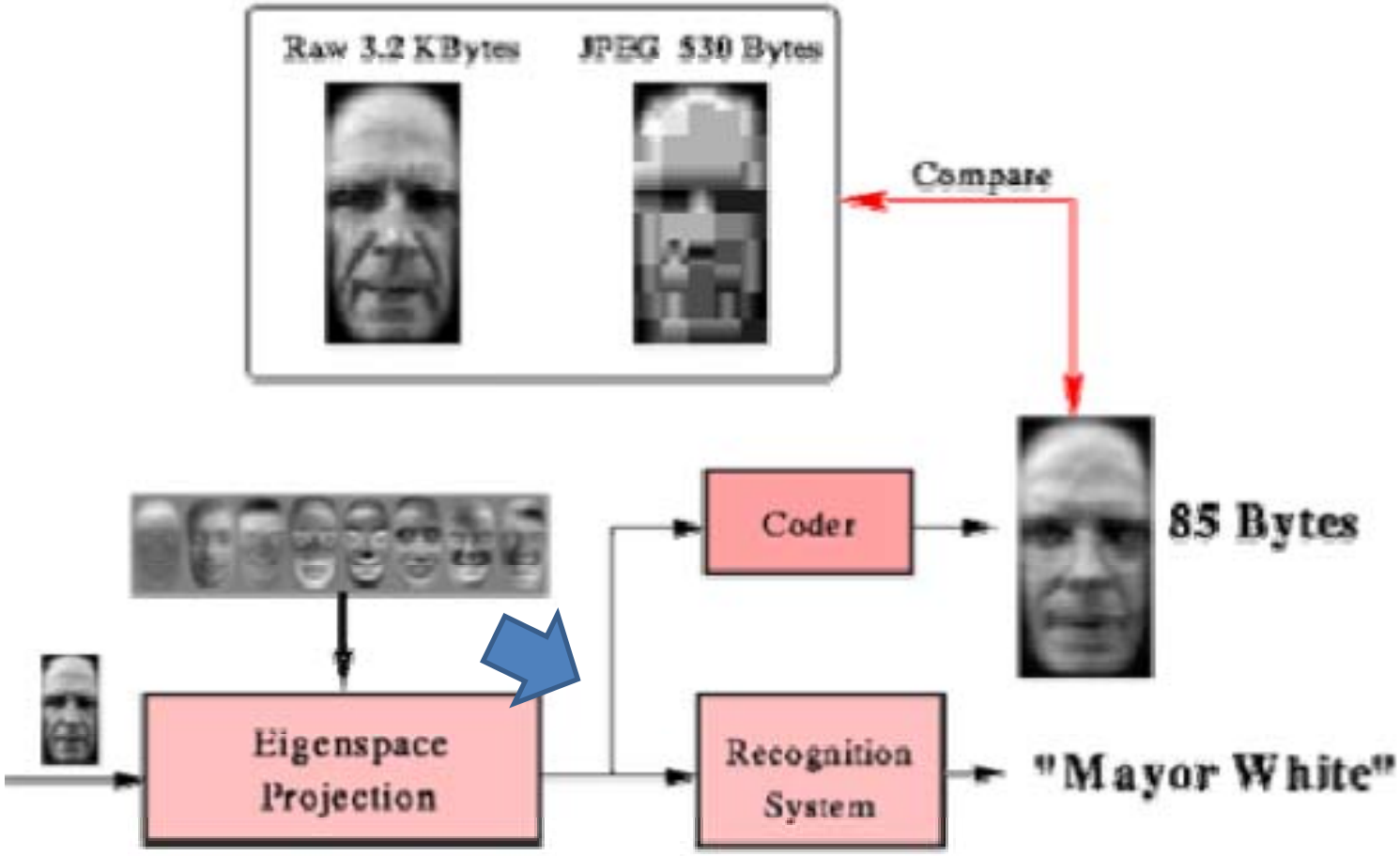
例：基于PCA的人脸表示及身份识别





(1)Original Input Image； (2)Estimated Head Location & Scale； (3)Head-Centered Image； (4)Estimated Facial Feature Locations； (5) Warped & Masked Facial Region

B. 基于PCA的非监督式特征提取　　C. 身份识别



前8个本征脸

例：
高维数据的低维可视化

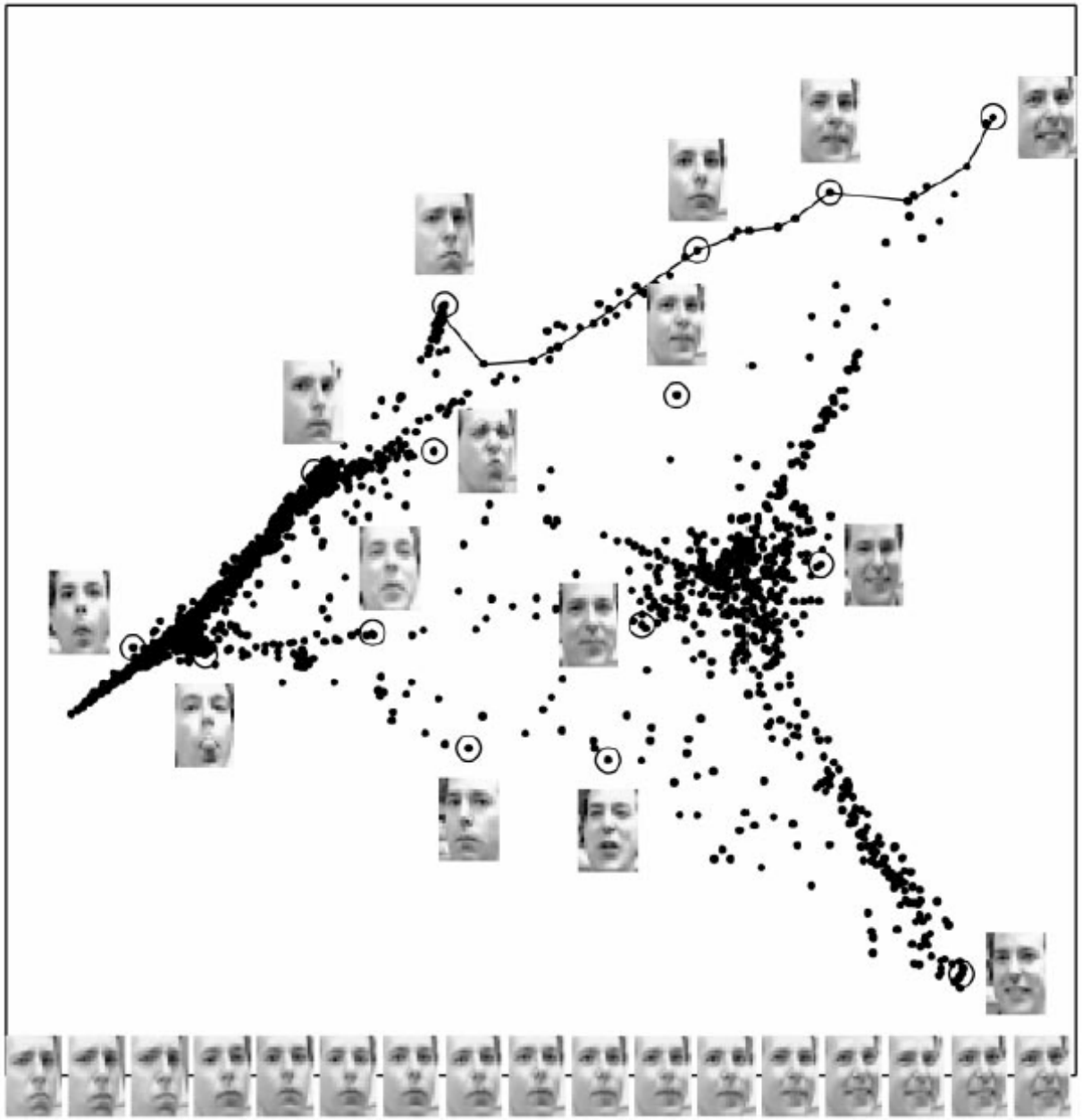
Face pose and Expression

N=2000 Images

k=12 nearest Neighbors

D=20*28=560 Pixels

d=2

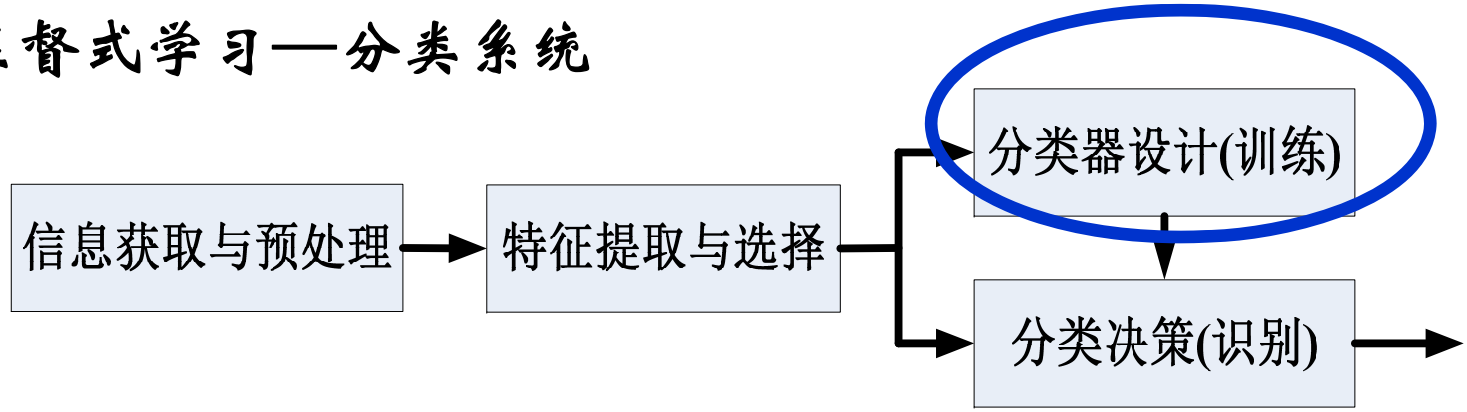


主要内容

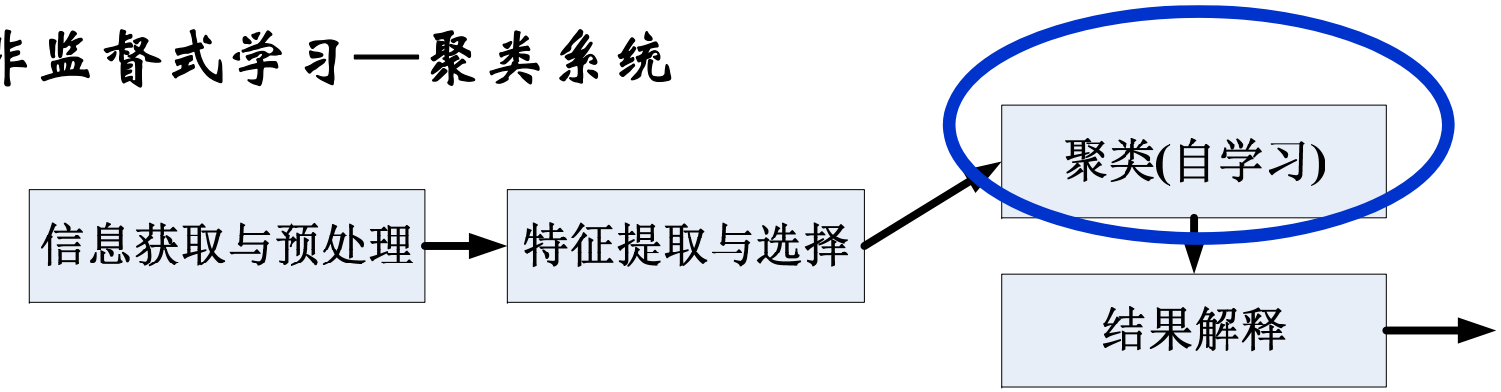
1. 机器学习的引入
2. 机器学习的典型类型及任务
3. 机器学习的基本组成
4. 机器学习的相关术语
5. 模型的评价及选择
6. 本学期课学习模块

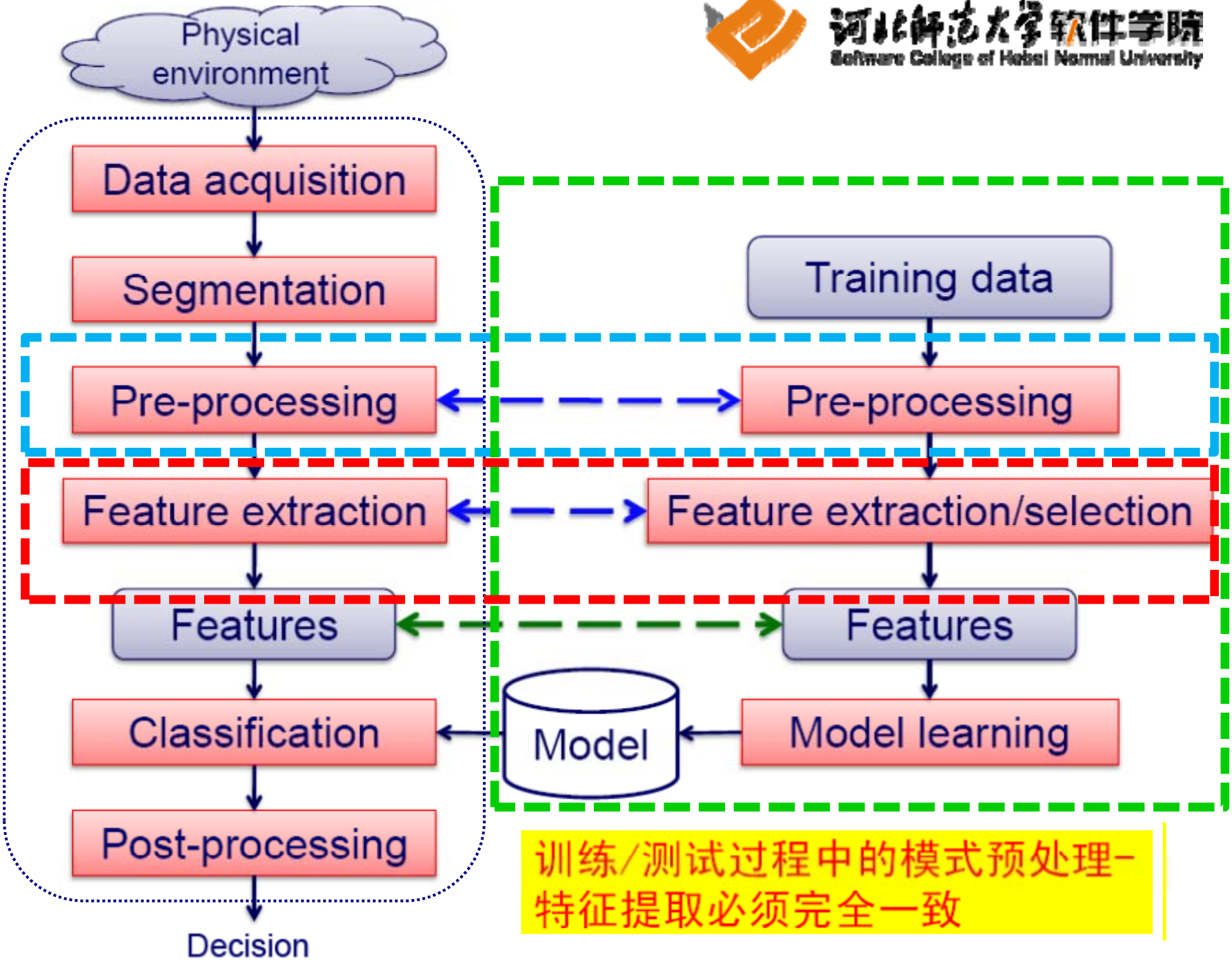
例：典型的机器学习系统

监督式学习—分类系统



非监督式学习—聚类系统





训练/测试过程中的模式预处理-特征提取必须完全一致

主要内容

1. 机器学习的引入
2. 机器学习的典型类型及任务
3. 机器学习的基本组成
4. 机器学习的相关术语
5. 模型的评价及选择
6. 本学期学习模块

[1]特征空间、特征维数、特征向量（*feature vector*）

d维特征空间，记为$\Re = \boldsymbol{R}^d$

$\boldsymbol{x} = [x_1, ..., x_d]^T \in \boldsymbol{R}^d$

$\boldsymbol{x}$--样本（***sample***），示例（*instance*），为d维向量

其中 $x_1, x_2, ..., x_d$--d个特征（*feature*）或属性（*attribute*）。

[2]样本集（*sample set*）/ 数据集（*dataset*）

训练集、验证集、测试集；　训练样本、测试样本；

独立同分布(**i.i.d,independent and identically distributed**)

例：监督式学习，训练集 $\{(\boldsymbol{x}_i, y_i), \boldsymbol{i} = 1, ..., \boldsymbol{N}\}$

- **类别状态数**$\boldsymbol{c} = 2$,两类别分类：$\boldsymbol{x}_i \in \boldsymbol{R}^d$，$y_i \in \{-1, +1\}$或$\{0, 1\}$
- **类别状态数**$\boldsymbol{c} > 2$,多类别分类：$\boldsymbol{x}_i \in \boldsymbol{R}^d$，$y_i \in \{1, ..., \mathbf{c}\}$
- 实值函数回归：　$\boldsymbol{x}_i \in \boldsymbol{R}^d$，$y_i \in \boldsymbol{R}$

[3]假设、假设空间

模型的学习过程就是:基于**训练集**，在**"假设空间"**中，进行**"匹配"**的"假设"的**搜索**过程.

[4]"假设"的**选择原则**

--主要原则

"奥克姆剃刀"(*Occam's razor*)准则

若多个假设与经验观测一致，则选最简单的那个

--其它原则

"多释原则" 即：保留与经验观察一致的所有假设(与集成学习一致)

[5]

"没有免费的午餐"定理(***No Free Lunch Theorem, NFL***)

There is no universally best model.

不应脱离具体问题，空泛讨论"什么学习算法最好"

主要内容

1. 机器学习的引入
2. 机器学习的典型类型及任务
3. 机器学习的基本组成
4. 机器学习的相关术语
5. 模型的评价及选择
6. 本学期学习模块

河北师范大学软件学院
Software College of Hebei Normal University

5.1 模型的学习能力、泛化(推广)能力

学习能力：模型关于训练样本集(经验数据)的预测能力

训练误差、经验误差

经验风险

泛化能力：模型关于新样本的预测能力

预测误差、测试误差

期望风险

过拟合(overfitting)/过学习

--学习能力过于强大，但泛化能力差

欠拟合(underfitting)/欠学习

--学习能力低下，因而泛化能力低

5.2 模型的评估方式

以监督式学习系统为例，考察泛化能力的评估方式

给定已知答案的**数据集** $D=\{(x_i,y_i),i=1,\ldots,N\}$

$$\begin{cases}\text{训练集}D_{train}\text{--模型的学习}\\ \text{测试集}D_{test}\text{--模型泛化能力的评价}\end{cases}$$

$$D=D_{train}\cup D_{test}$$

数据集 D 划分的几种典型实现方式

[1]留出法(*hold-out*)

[2]交叉验证(*cross validation*)

[3]自助法(*bootstrapping*)

[1]留出法(*hold-out*)、留出法交叉验证(*hold-out* cv)

$D = D_{train} \cup D_{test}$

$\Phi = D_{train} \cap D_{test}$

数据集**随机划分**尽量**保持数据分布的一致性**

A. 单独一次随机划分，估计结果不够稳定可靠

B. 应多次随机划分，重复评估取结果的均值及标准差

(*hold-out cross-validation*)

$\frac{2}{3} \sim \frac{3}{4}$　　$\frac{1}{4} \sim \frac{1}{3}$

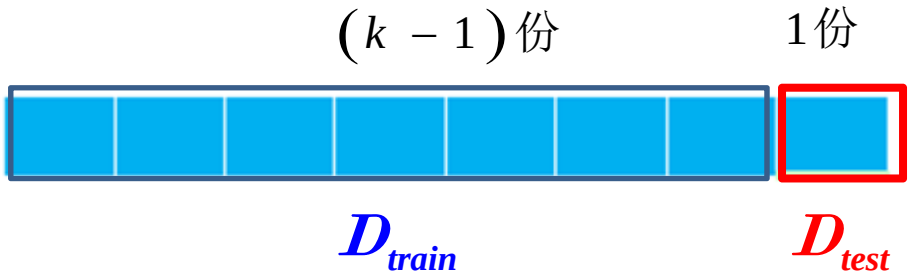
测试集D_{test}样本数:$|D_{test}|$不低于30个

[2]交叉验证(*cross validation*)–*rotation estimation*

$D = D_{train} \cup D_{test}$　　$\Phi = D_{train} \cap D_{test}$

数据集的随机划分尽量**保持数据分布的一致性**

随机打乱，均分成k等份



单轮k-倍交叉验证(*one - turn k - fold cross - validation*)

多轮k-倍交叉验证(*multi - turn k - fold cross - validation*)

留一法交叉验证(*leave - one - out cross - validation*，*LOOCV*)

[3]自助法 $(\boldsymbol{bootstrapping})$

bootstrap sampling

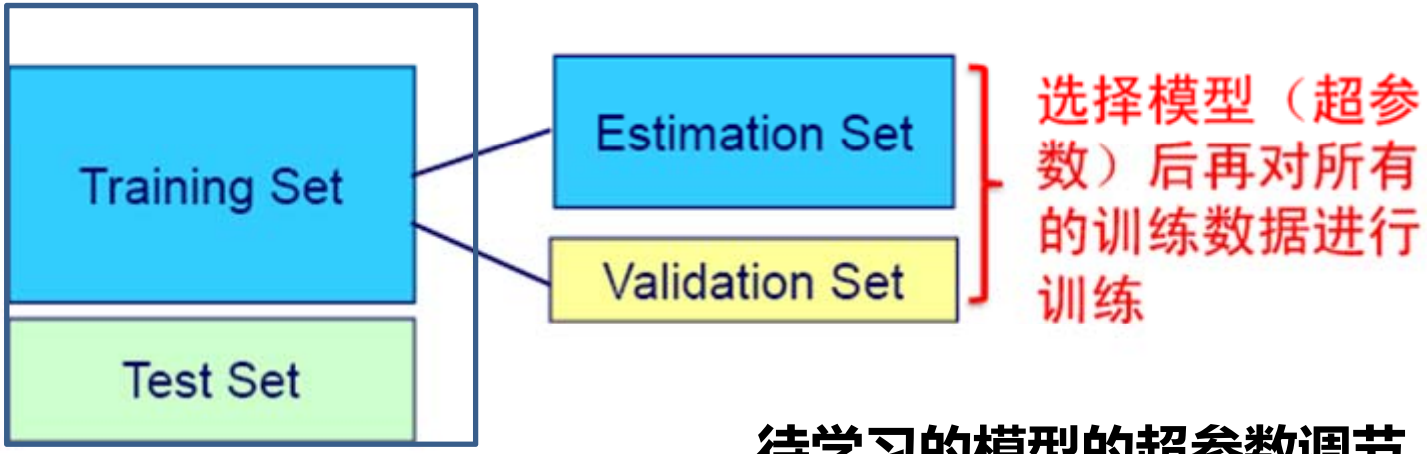
数据集$\boldsymbol{D}=\{(x_i,y_i),i=1,...,N\}$--初始数据集

- 训练集$\boldsymbol{D}_{train}$--模型的学习
  对初始数据集$\boldsymbol{D}$有放回的随机抽取$\boldsymbol{N}$次，得自助数据集$\boldsymbol{D'}$
- 测试集$\boldsymbol{D}_{test}$--模型泛化能力的评价
  对**数据集**$\boldsymbol{D}$中没有被抽取到的样本集 $\boldsymbol{D}_{test}=\boldsymbol{D}\setminus\boldsymbol{D'}$

初始数据集内，样本未被抽取的概率 $\lim\limits_{N\to\infty}\left(1-\frac{1}{N}\right)^N=\frac{1}{e}\approx 0.368$

自助法估计，也称"包外估计"(out−of−bag estimation)

[4]数据集D的两种层次划分及适用场合



待学习的模型的超参数调节

训练集{估计集(学习)
验证集(评价)

最终学习得到的模型

{训练集(超参数固定之后，学习最终模型)
测试集(评价最终模型)

5.3 模型的性能度量

面向不同的学习任务，结合可行的评估方式，采用适当评价标准，得到关于模型的泛化能力的性能度量。

[1]回归

均方误差.

[2]分类

分类错误率、正确率

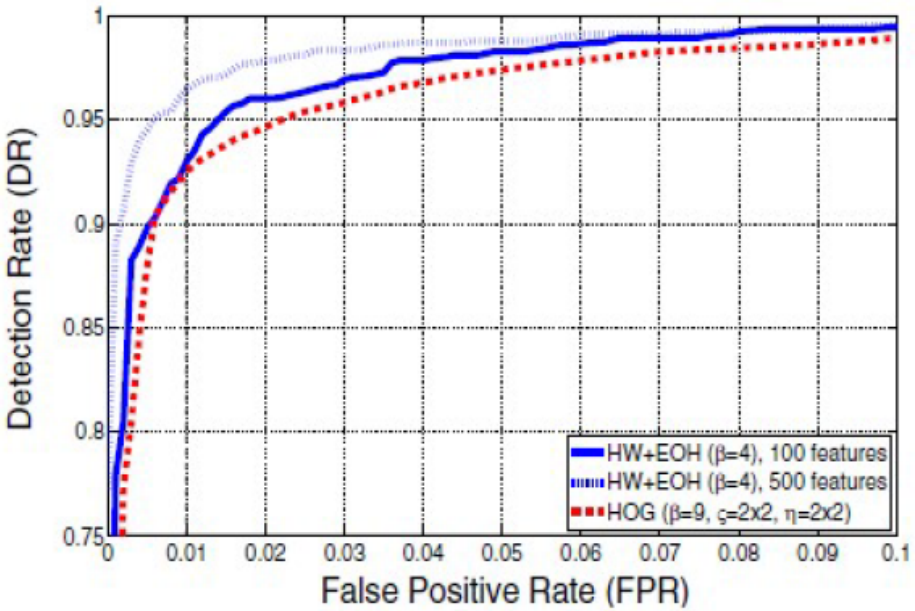
查准率、查全率、以及F1

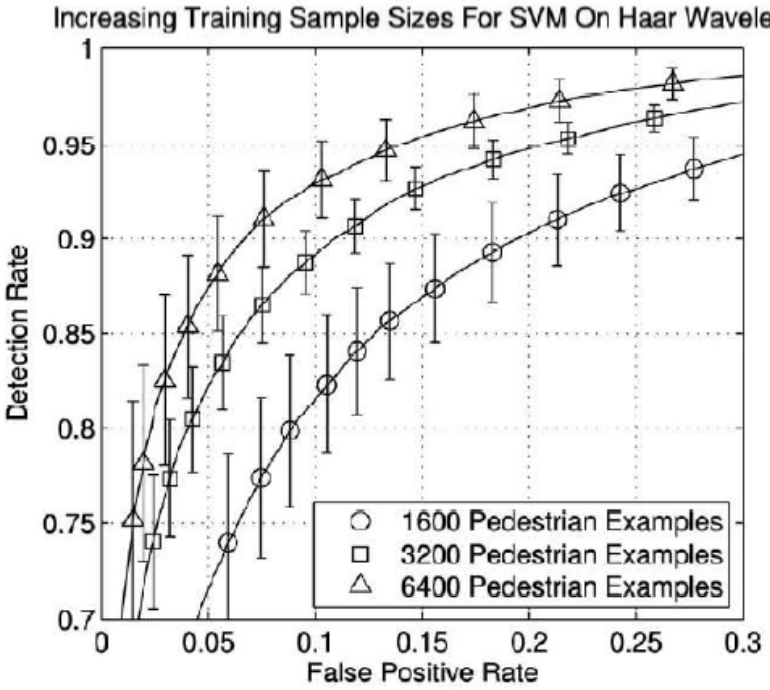
P-R曲线、ROC曲线以及AUC

...

例：基于ROC曲线的分类器性能评价

- 可用于不同分类器性能比较
- 可针对同种分类器，进行特征选择、样本规模等优劣比较。





例：Comparison between Support Vector Machines, the Kernel Fisher Discriminant (KFD), a single radial basis function classifier (RBF), AdaBoost (AB), and regularized AdaBoost (ABR) on 13 different benchmark datasets (see text). Best result in bold face, second best in italics.

平均错误率 ± 错误率标准差

	SVM	KFD	RBF	AB	AB_R
Banana	11.5±0.07	**10.8±0.05**	**10.8±0.06**	12.3±0.07	*10.9±0.04*
B.Cancer	*26.0±0.47*	**25.8±0.46**	27.6±0.47	30.4±0.47	26.5±0.45
Diabetes	*23.5±0.17*	**23.2±0.16**	24.3±0.19	26.5±0.23	23.8±0.18
German	**23.6±0.21**	*23.7±0.22*	24.7±0.24	27.5±0.25	24.3±0.21
Heart	**16.0±0.33**	*16.1±0.34*	17.6±0.33	20.3±0.34	16.5±0.35
Image	*3.0±0.06*	3.3±0.06	3.3±0.06	**2.7±0.07**	**2.7±0.06**
Ringnorm	1.7±0.01	**1.5±0.01**	1.7±0.02	1.9±0.03	*1.6±0.01*
F.Sonar	**32.4±0.18**	*33.2±0.17*	34.4±0.20	35.7±0.18	34.2±0.22
Splice	10.9±0.07	10.5±0.06	*10.0±0.10*	10.1±0.05	**9.5±0.07**
Thyroid	4.8±0.22	**4.2±0.21**	4.5±0.21	*4.4±0.22*	4.6±0.22
Titanic	**22.4±0.10**	23.2±0.20	23.3±0.13	*22.6±0.12*	*22.6±0.12*
Twonorm	3.0±0.02	**2.6±0.02**	2.9±0.03	3.0±0.03	*2.7±0.02*
Waveform	*9.9±0.04*	*9.9±0.04*	10.7±0.11	10.8±0.06	**9.8±0.08**

主要内容

1. 机器学习的引入
2. 机器学习的典型类型及任务
3. 机器学习的基本组成
4. 机器学习的相关术语
5. 模型的评价及选择
6. 本学期学习模块

监督式学习

- 分类
  - A. 懒惰学习--KNN法(K最近邻法)分类模型
  - B. 概率学习 --贝叶斯分类器
  - C. 分而治之--分类树
- 回归
  - D. KNN法回归
  - E. 回归树
  - F. 最小二乘回归

非监督式学习

- 聚类 -- G. K-Means Clustering
- 特征提取 -- H. PCA

模型评价