上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

机器学习

Machine Learning

绪论

2024年2月19日

饮水思源 • 爱国荣校

SJTU

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

目录

1 课程简介与安排

2 机器学习发展历程

3 机器学习基本概念

课程简介

课程名称：机器学习

任课教师：王士林&黄征

联系方式：wsl@sjtu.edu.cn（王士林），huang-zheng@sjtu.edu.cn（黄征）

教学参考书：《机器学习》，作者：周志华，清华大学出版社，2016。

一些参考信息：吴恩达机器学习视频课件

考核方式

平时成绩：10%

课程实验：20%

课程大作业：70%

课程安排

- 经典机器学习（共16学时）
  - 绪论：机器学习的定义、历史和基本术语等等
  - Linear Models: Linear Regression, Perceptron,等等
  - Non-Linear Models: MLP, SVM, Adaboost, 等等
  - Feature Extraction Methods：PCA & LDA
  - 算法实践：Python Sklearn

课程安排

- 深度学习（共16学时）
  - 深度卷积神经网络——CNN
  - 序列模型——RNN
  - 多任务学习——MTL
  - 生成对抗网络——GAN
  - 算法实践——Python + Pytorch/TF/Keras

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

目录

1 课程简介与安排

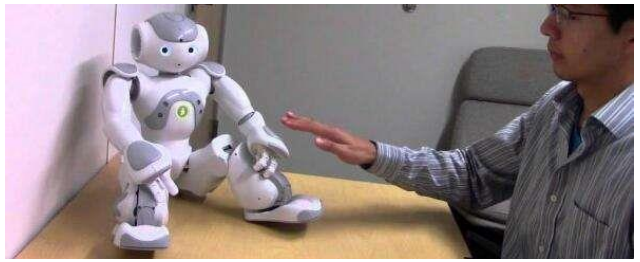
2 机器学习发展历程

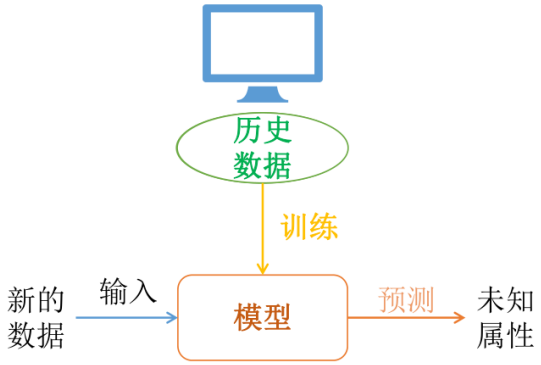
3 机器学习基本概念

什么是机器学习?

通过模仿（Simulation）使机器具备学习的能力

- ✓ 观察人类行为
- ✓ 基于观察建模并模仿





从观察数据中总结规律，建立模型；

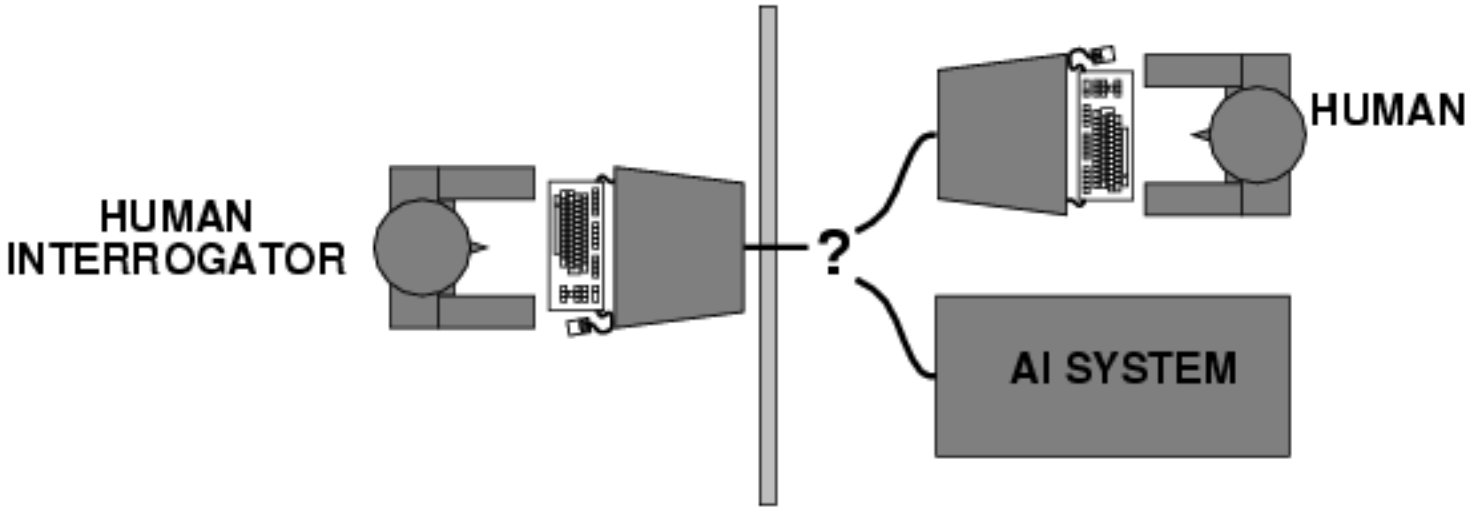
基于模型，对未知数据进行预测。

机器学习发展历程

孕育——上世纪50年代

人工神经元和人工神经网络的诞生（“连接主义”学习）

图灵测试推动

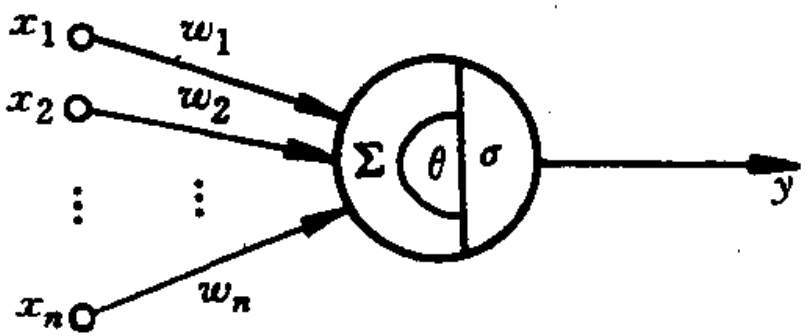


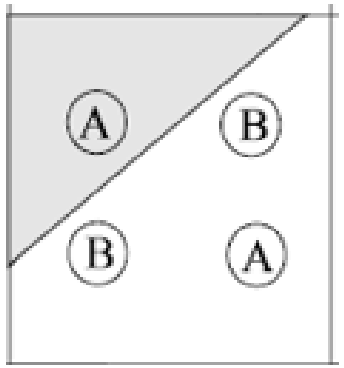
机器学习发展历程

过渡——上世纪60年代初-70年代末

人工神经网络的发展（Rosenblatt感知器）

基于逻辑的归纳系统（Michalski）——“符号主义”学习





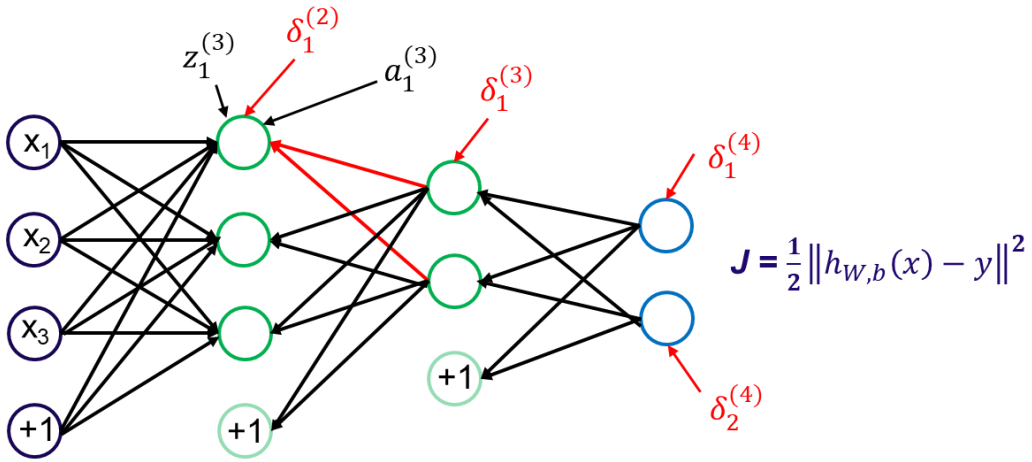
机器学习发展历程

发展——上世纪80年代初-21世纪初期

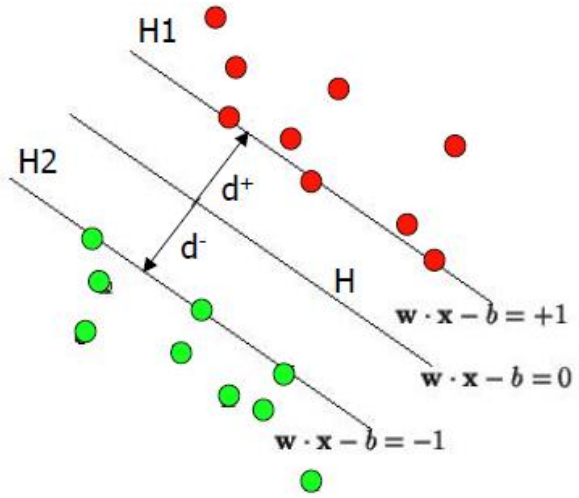
知识工程Knowledge Engineering（费根鲍姆 专家系统）

人工神经网络的突破（BP算法和MLP网络）

统计学习的崛起（贝叶斯理论，SVM, Adaboost）

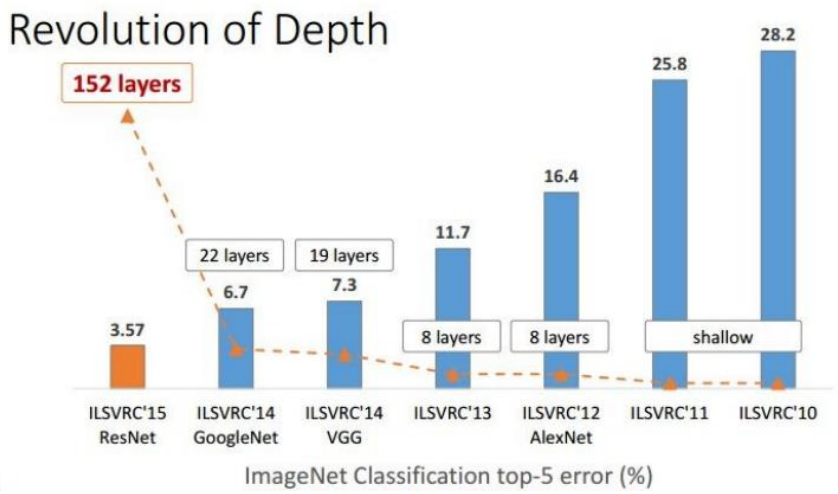


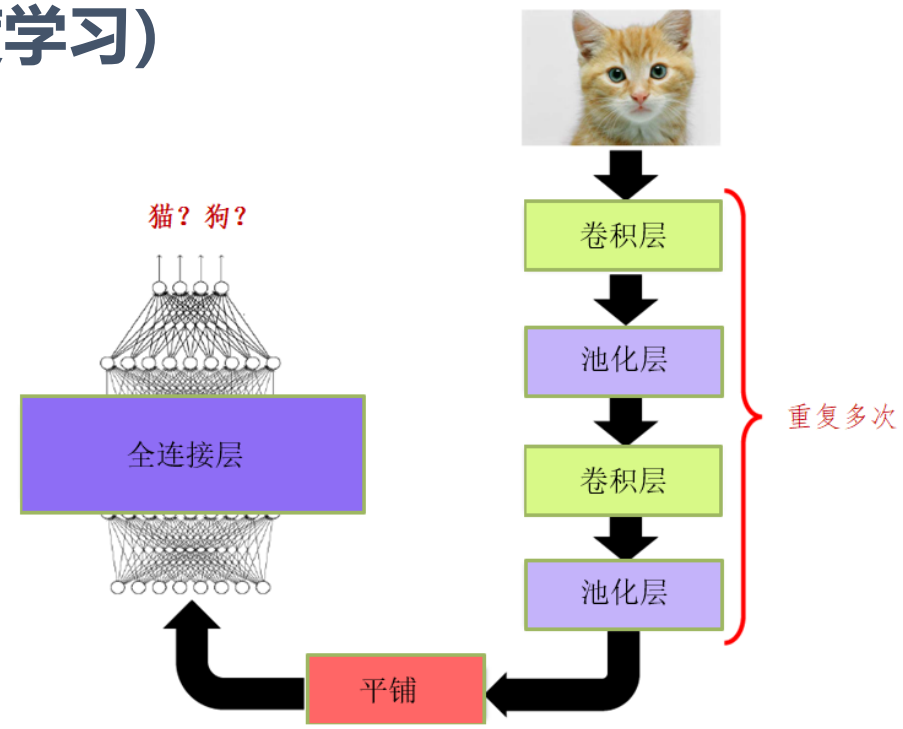
$$J = \frac{1}{2}\left\|h_{W,b}(x) - y\right\|^2$$



机器学习发展历程

广泛流行——21世纪初期-至今
人工神经网络的再次突破（深度学习）
与统计学习的结合





监督学习发展的几个阶段

基于规则：

以规则为依据进行推理和判断。

基于样本（手工特征）：

以样本为依据进行建模、基于模型进行推理

特征工程

分类器设计

基于样本（自动化特征）：

以样本为依据进行建模、基于模型进行推理

同时完成特征提取与分类识别

机器学习主要应用领域

自然语言理解（Natural Language Processing, NLP）

计算机视觉（Computer Vision, CV）

语音识别（Speech Recognition）

智能驾驶、智能导航

智慧医疗

互联网和社交媒体

等等

机器学习相关研究

数据挖掘（Data Mining）

大数据分析（Big Data Analysis）

模式识别（Pattern Recognition）

CCF推荐期刊与会议：

https://www.ccf.org.cn/Academic_Evaluation/AI/

ICML(ECML, ACML), NIPS, COLT

IJCAI, AAAI

CVPR，ICCV(ECCV), ACL

例一、医学诊断

病人号码	发烧	鼻塞	流涕	畏寒	头痛	感冒?
1	×	√	√	×	√	×
2	√	×	×	×	×	×
3	√	√	√	√	×	√
4	√	√	×	√	√	√
5	×	√	×	×	√	×
6	√	×	×	√	×	×
7	×	√	√	√	√	√
8	×	×	√	√	√	√
9	√	√	√	×	√	√
*	×	√	×	√	√	?

机器学习能做什么

例二、对弈

机器学习能做什么

例三、自动驾驶

机器学习与安全的关联

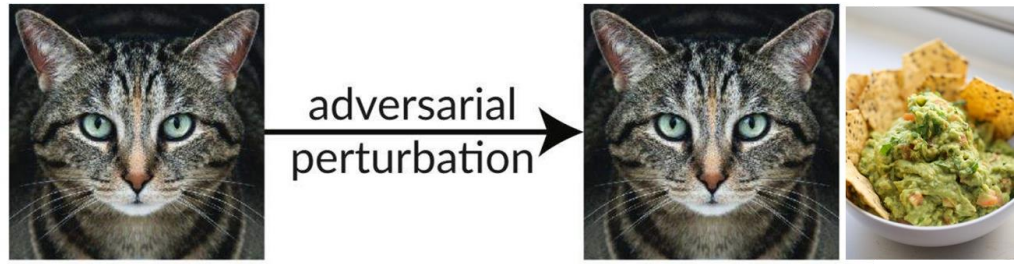
对抗样本

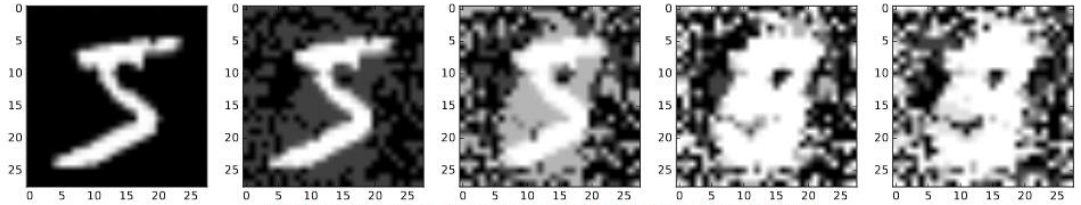
投毒攻击

后门攻击

隐私保护

模型版权保护





(a) Start from normal data "5" by applying the direct gradient method.

(a) car

Prediction	Probability
automobile	0.99996
cat	0.0003
truck	0.0001
dog	0
ship	0

(b) Prediction results (top 5)

(c) car with $WM_{content}$

Prediction	Probability
airplane	1
bird	0
automobile	0
ship	0
truck	0

(d) Prediction results(top 5)

深度伪造技术的威胁与应对

- 在过去，对多媒体内容进行编辑需要有该领域的专业知识，熟练使用专业软硬件（传感器），还需要投入大量的人力和时间。
- 深度伪造模型的出现，使得编辑多媒体内容的门槛大幅降低，任何人都可以使用基于深度学习的工具对图像和视频进行换脸操作，或者是生成高度逼真但现实世界并不存在的面孔。
- 2017年，Deepfake 诞生，一位名为“deepfakes”的Reddit用户在网站上发布了一些经过篡改的色情视频片段，这些视频将色情表演者的面孔换成了影视明星的面孔，引发公众恐慌。

机器学习与安全的典型示例

深度伪造技术的威胁与应对

- 政治安全
- 金融安全
- 隐私安全
- 肖像安全

信息战:“谣言”里的未来战场

假新闻是“信息战”中的一种常用操作方式，而“信息战”并不是一个新名词。“信息战一方面通过传播虚假信息、制造谣言的手段打击和操纵敌对国家公众，另一方面通过技术手段防止他国收集到不利于本...

人民资讯

机器学习与安全的典型示例

深度伪造技术的威胁与应对

- 政治安全
- **金融安全**
- 隐私安全
- 肖像安全

财经网

05月22日 10:10 来自 微博视频号

#如何避免AI伪装熟人诈骗#【#AI诈骗正在全国爆发#！#公司老板被AI诈骗430万#】近日，包头警方发布一起利用人工智能（AI）实施电信诈骗的典型案例，福州市某科技公司法人代表郭先生10分钟内被骗430万元。4月20日中午，郭先生的好友突然通过微信视频联系他，自己的朋友在外地竞标，需要430万保证金，且需 展开

机器学习与安全的典型示例

深度伪造技术的威胁与应对

- 政治安全
- 金融安全
- **隐私安全**
- 肖像安全

“ZAO”遭 约谈 APP用户隐私安全问题谁来买单

2019-09-04 22:18

“仅需一张照片，出演天下好戏”。近日，AI换脸APP——ZAO火爆全网。《金融时报》记者试用发现，ZAO操作简单，只需用户提供一张符合清晰度要求的照片，就可以通过AI换脸，将原视频男女主角的面容换成用户自己的头像。在ZAO上，你可以成为风情款款的聂小倩，与宁采臣展开一段虐恋；也可以成为拥有魔鬼身材的国际名模，在国际大秀上展示风采。

不过，蹿红之后，ZAO使用所涉及的用户个人信息安全和数据安全等问题备受质疑。有用户发现,ZAO的用户协议显示，用户上传发布内容后,意味着同意授予ZAO及其关联公司以及ZAO用户全球范围内完全免费、不可撤销、永久、可转授权和可再许可的权利。这意味着，用户一旦使用该APP，自己的肖像权将完全让渡给ZAO。有律师表示，ZAO该行为已经涉嫌侵害用户的肖像权。

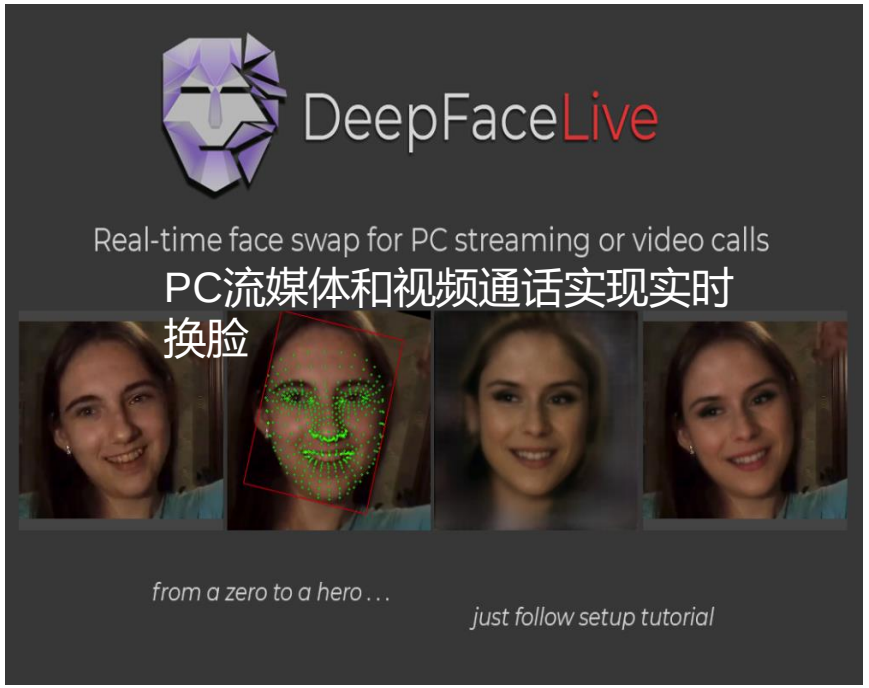
机器学习与安全的典型示例

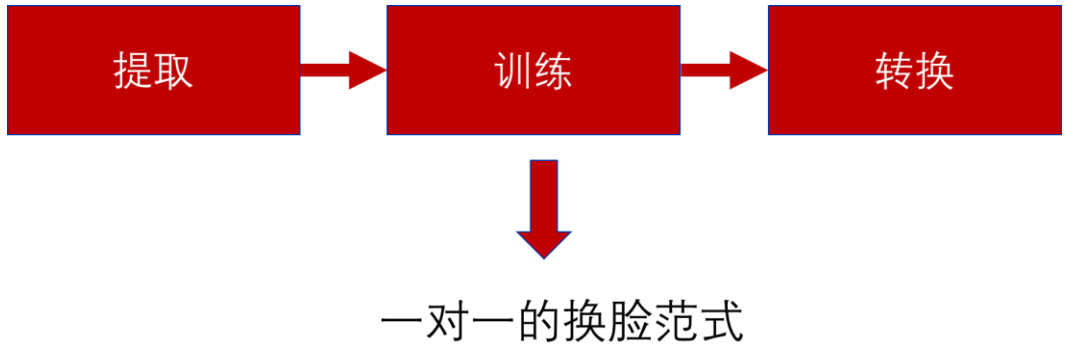
深度伪造技术的威胁与应对

- 政治安全
- 金融安全
- 隐私安全
- 肖像安全

机器学习与安全的典型示例

深度伪造生成工具





深度伪造生成工具

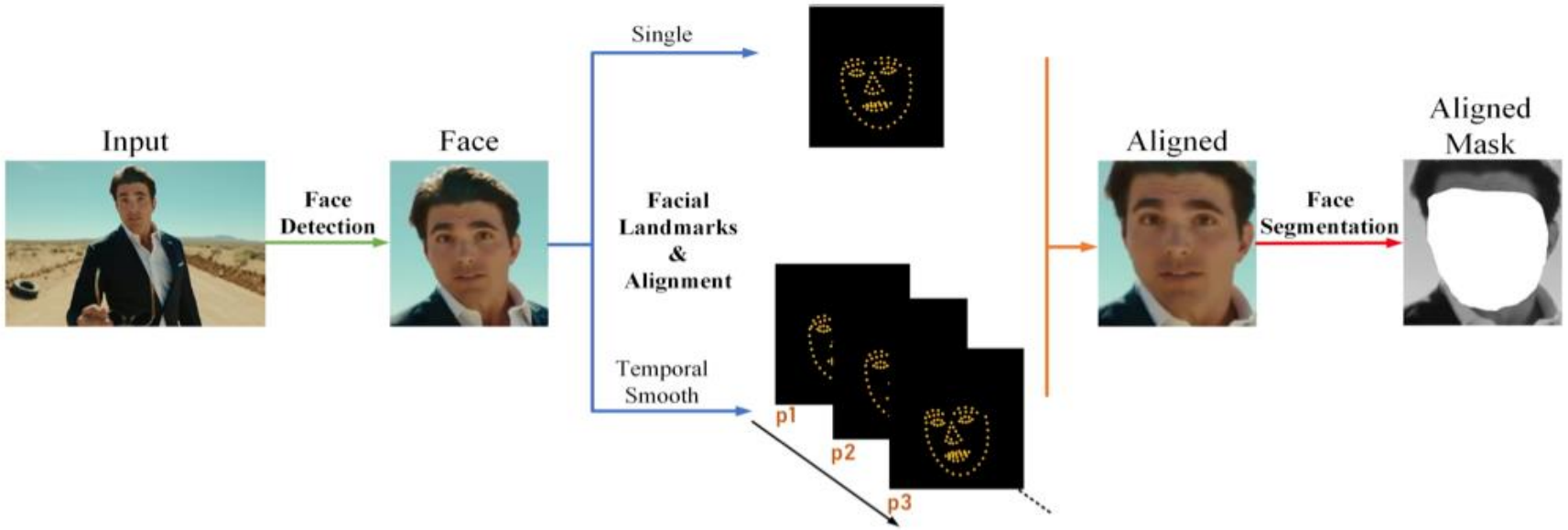
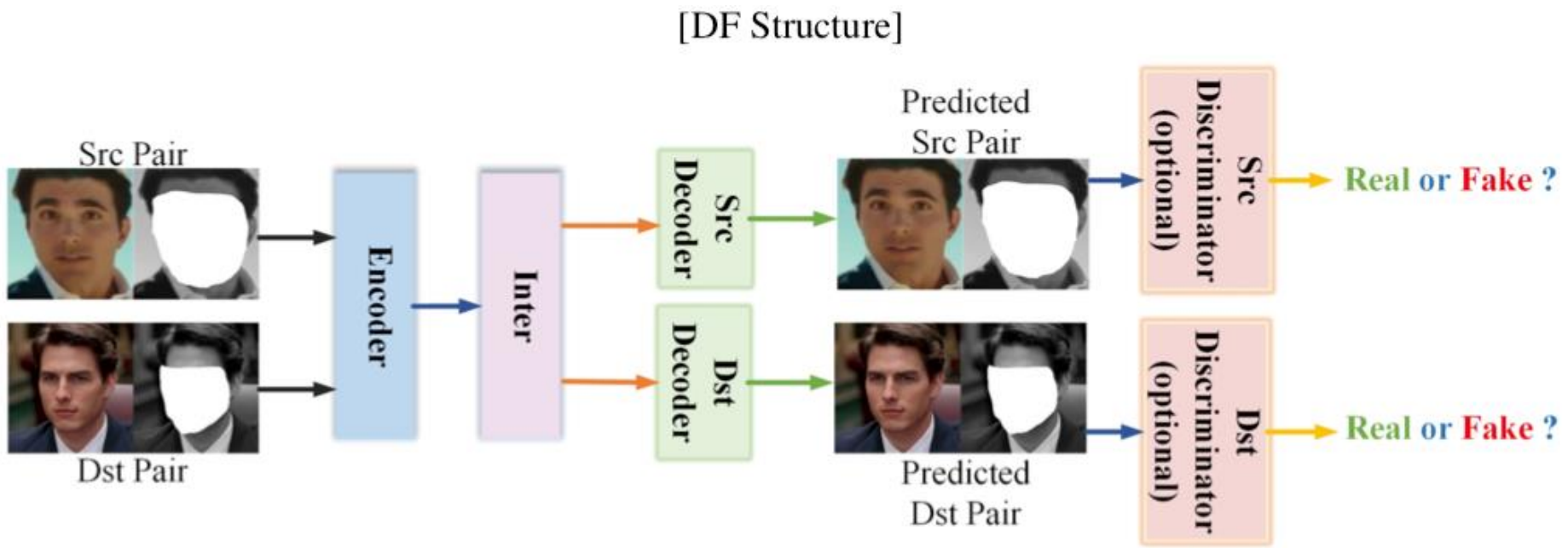


Figure 2. Overview of extraction phase in DeepFaceLab (DFL for short).

裁剪后的人脸以及其在原始图像中的坐标，人脸面部关键点，对齐的人脸，以及来自原始图像像素级的分割掩码

机器学习与安全的典型示例

深度伪造生成工具



深度伪造生成工具

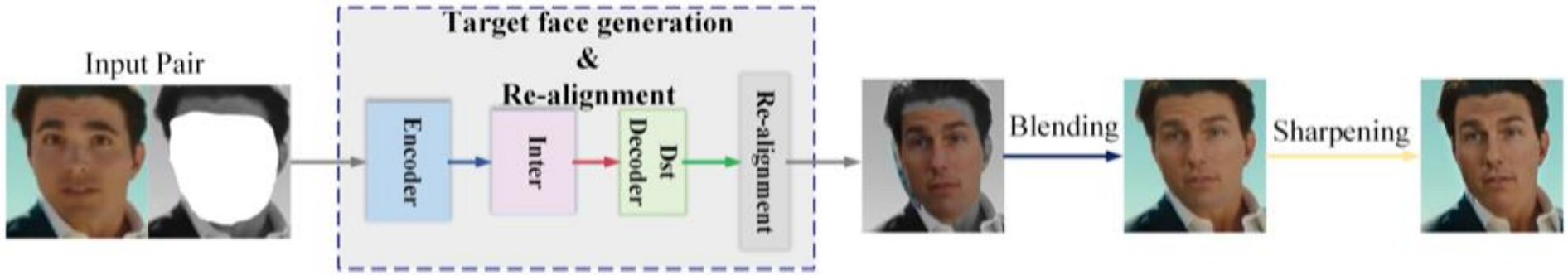


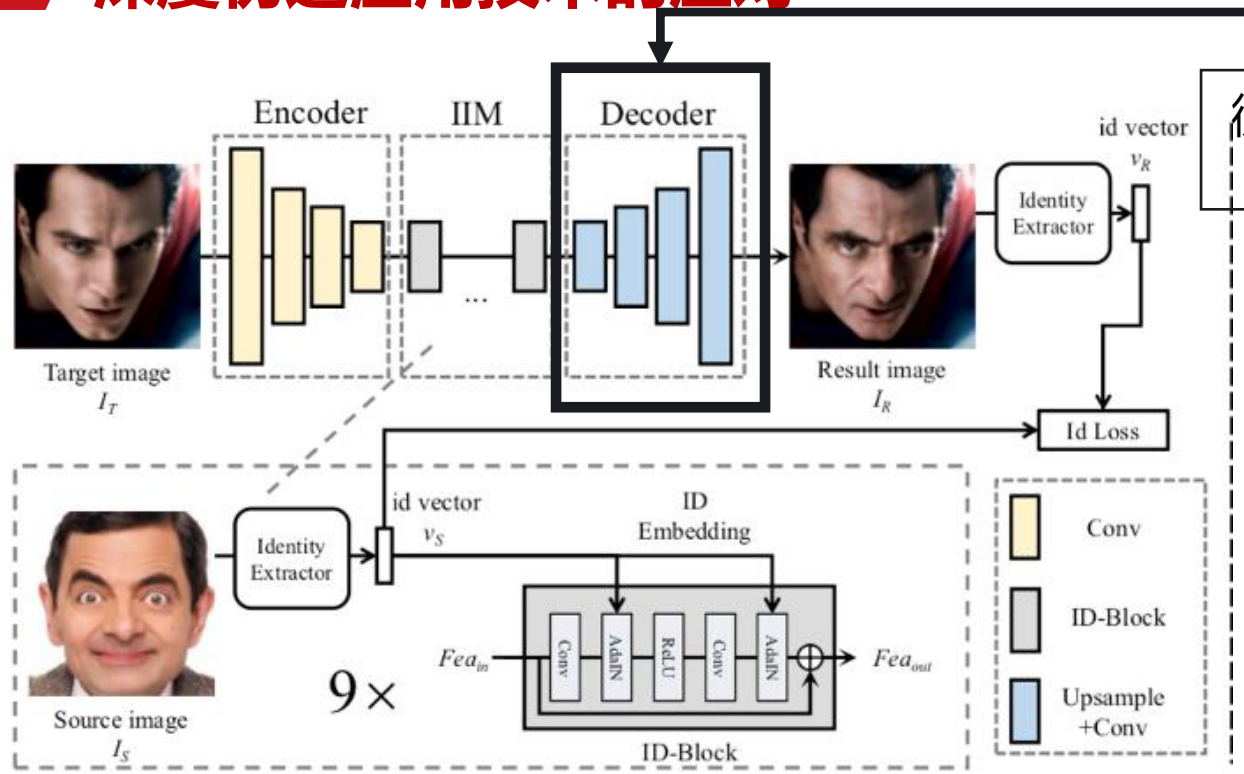
Figure 4. Overview of conversion phase in DeepFaceLab(DFL).

机器学习与安全的典型示例

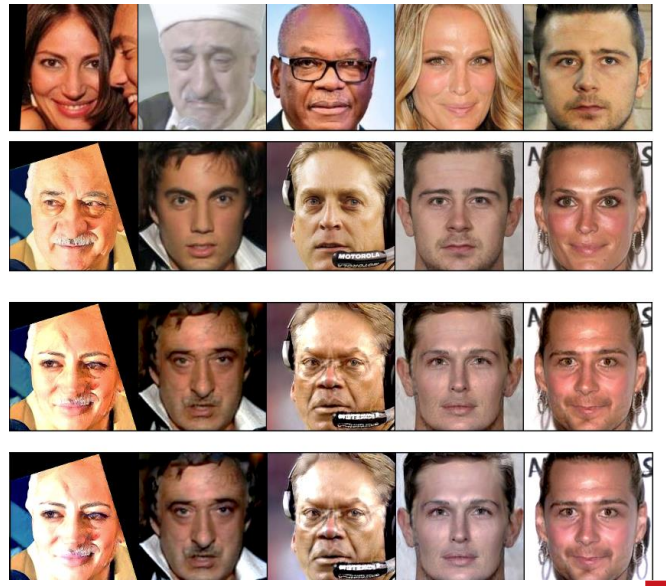
深度伪造生成工具

机器学习与安全的典型示例

深度伪造滥用技术的应对



微调decoder注入水印
（代表使用者的id信息）



模型不可见水印

深度伪造生成工具

- 基于时空域的检测
  - 基于颜色、光照的不均匀性(伪影)
  - 捕获 GAN 指纹
- 基于变换域的检测
  - 在 FFT、DCT 以及小波变化域分析
  - 利用光谱带之间的不一致性
  - 捕获 GAN 指纹
- 基于生物特征的检测
  - 口型动态不一致
  - 缺乏眨眼
  - 两眼之间不一致的角膜镜面反射高光
- 基于伪造过程的检测
  - 检测图像的融合过程

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

目录

1 课程简介与安排

2 机器学习发展历程

3 机器学习基本概念

机器学习简介

机器学习的可行性：

- ✓ 充足的训练样本
- ✓ 强大的计算和处理能力（大存储量、高运算速度等等）

机器学习分类：

- ✓ 监督学习（Supervised Learning）
- ✓ 无监督学习（Unsupervised Learning）
- ✓ 强化学习（Reinforcement Learning）

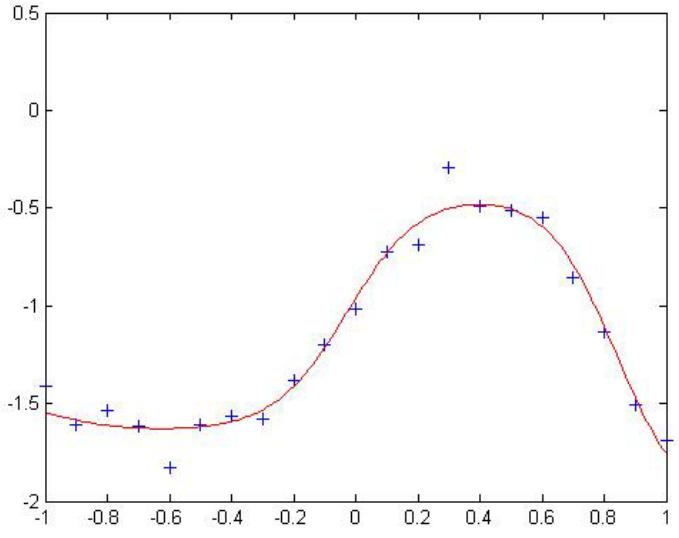
监督学习举例

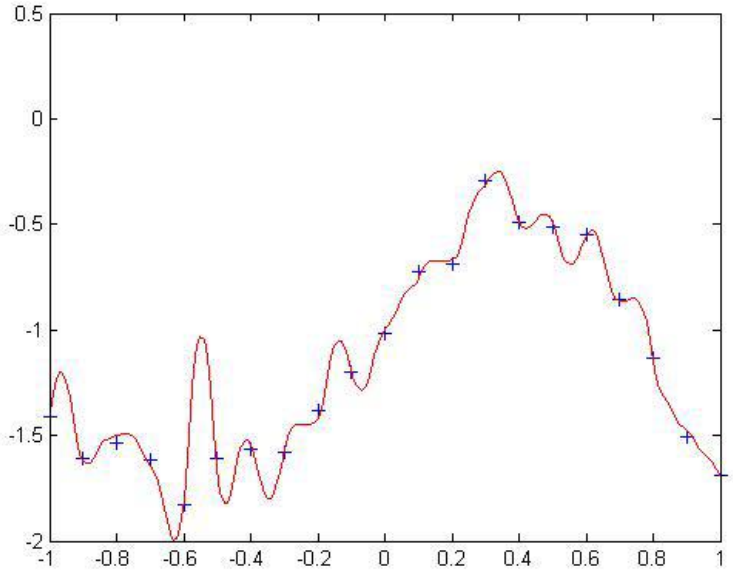
分类问题（Classification）

病人号码	发烧	鼻塞	流涕	畏寒	头痛	感冒?
1	×	√	√	×	√	×
2	√	×	×	×	×	×
3	√	√	√	√	×	√
4	√	√	×	√	√	√
5	×	√	×	×	√	×
6	√	×	×	√	×	×
7	×	√	√	√	√	√
8	×	×	√	√	√	√
9	√	√	√	×	√	√
*	×	√	×	√	√	?

监督学习举例

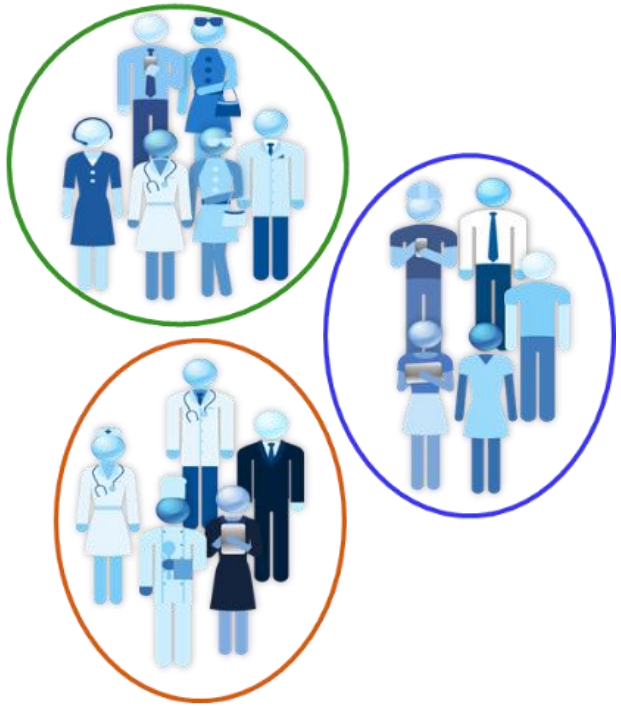
回归问题(Regression)





无监督学习举例：

聚类分析



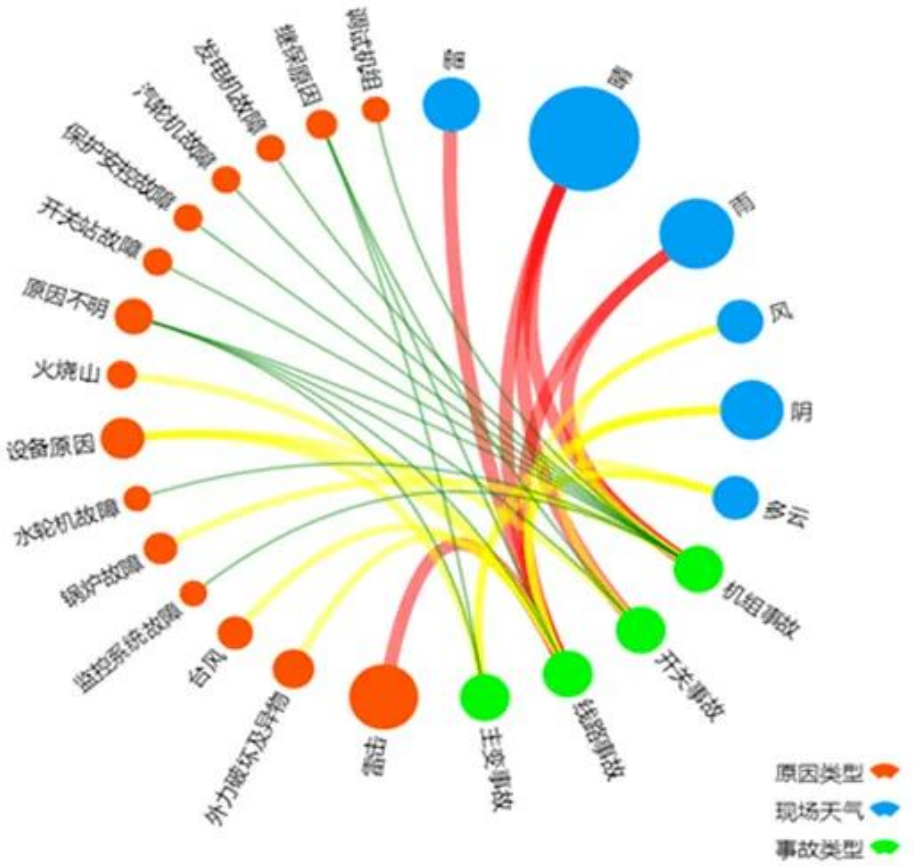
机器学习简介

无监督学习举例：

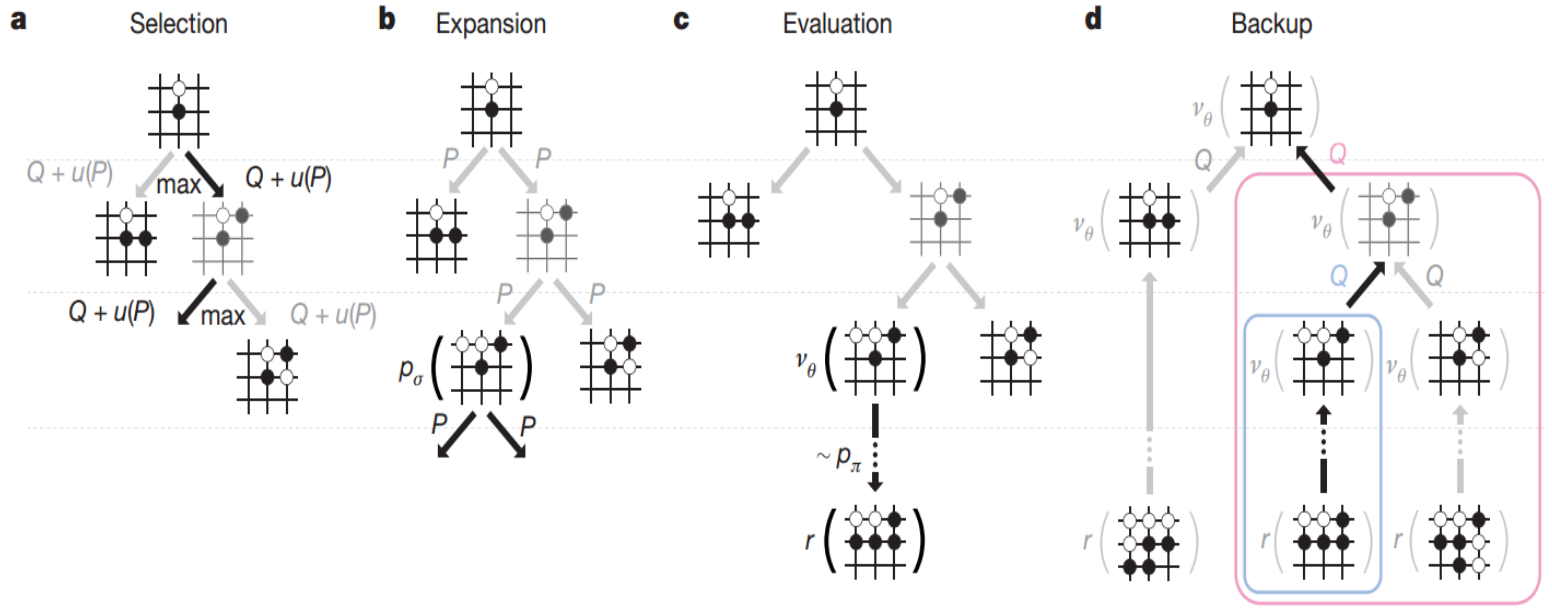
聚类分析

无监督学习举例：

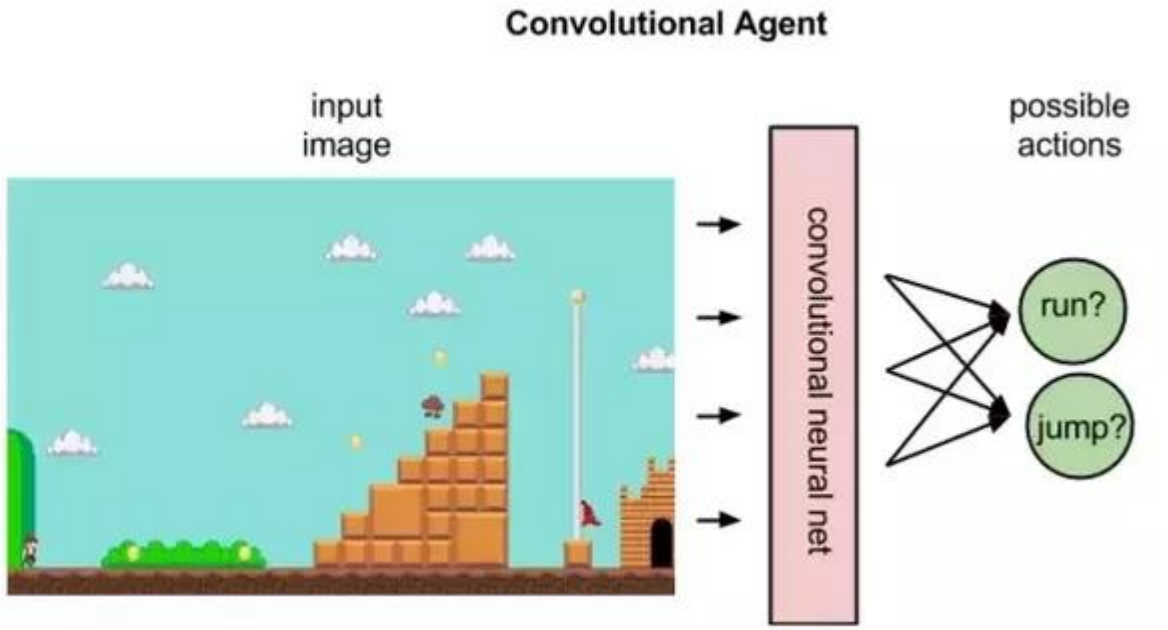
关联分析



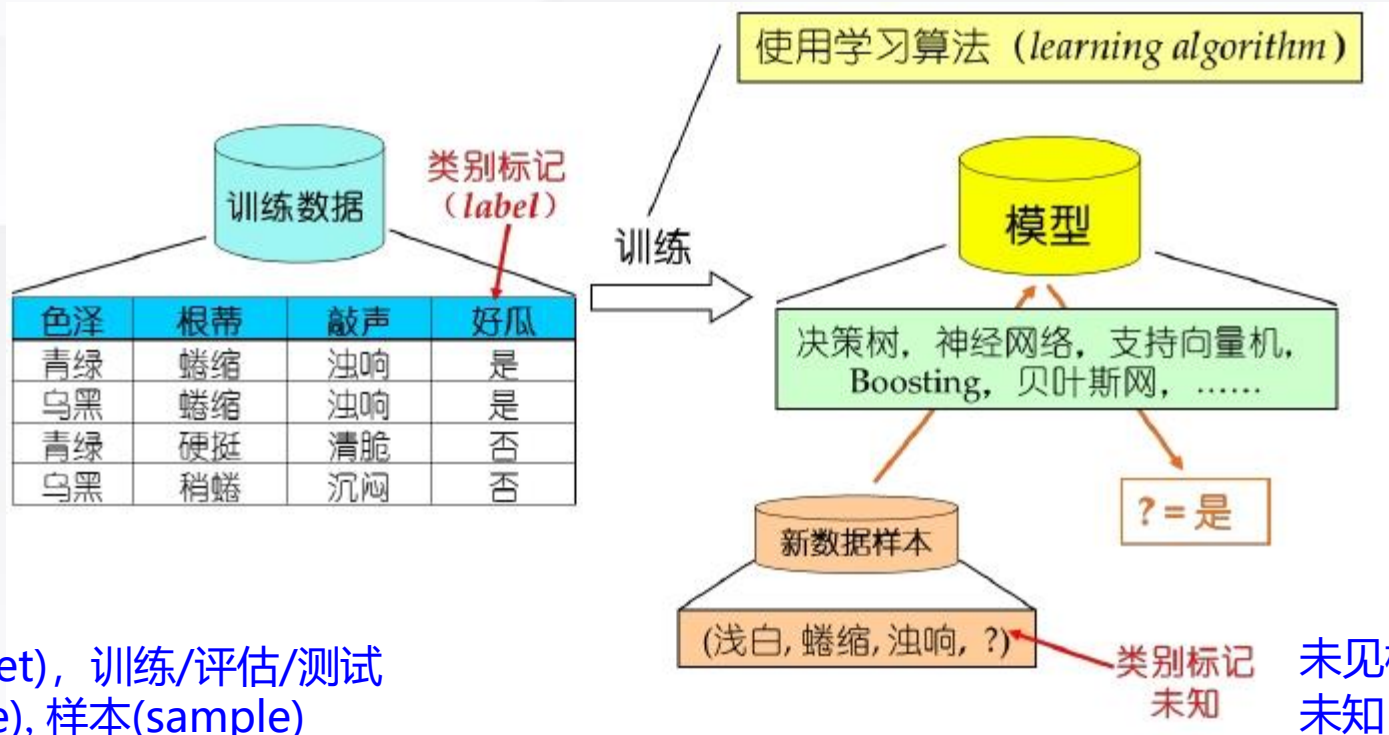
强化学习举例


a
Selection
Q + u(P)
max
Q + u(P)
Q + u(P)
max
Q + u(P)
b
Expansion
P
P
P
P
P
P
c
Evaluation
v_θ
~ p_π
r
d
Backup
v_θ
Q
Q
v_θ
Q
Q
v_θ
v_θ
v_θ
r
r
r

强化学习举例


Convolutional Agent
input image
convolutional neural net
possible actions
run?
jump?

机器学习术语



色泽	根蒂	敲声	好瓜
青绿	蜷缩	浊响	是
乌黑	蜷缩	浊响	是
青绿	硬挺	清脆	否
乌黑	稍蜷	沉闷	否

假设(hypothesis)
真相(ground-truth)
学习器(learner)

数据集(dataset)，训练/评估/测试
示例(instance), 样本(sample)
属性(attribute), 特征(feature); 属性值
属性空间, 样本空间, 输入空间
特征向量(feature vector)
标记空间, 输出空间

未见样本(unseen instance)
未知“分布”
独立同分布(i.i.d.)
泛化（generalization）、
过拟合（overfitting）与
欠拟合（underfitting）

假设空间

表 1.1 西瓜数据集

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	浊响	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否

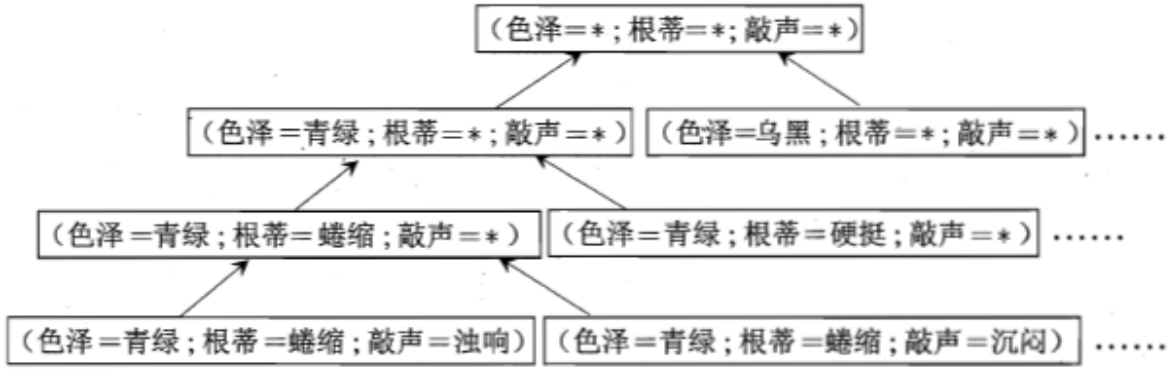


图 1.1 西瓜问题的假设空间

(色泽=?)∧(根蒂=?)∧(敲声=?)↔好瓜

学习过程 → 在所有假设(hypothesis)组成的空间中进行搜索的过程

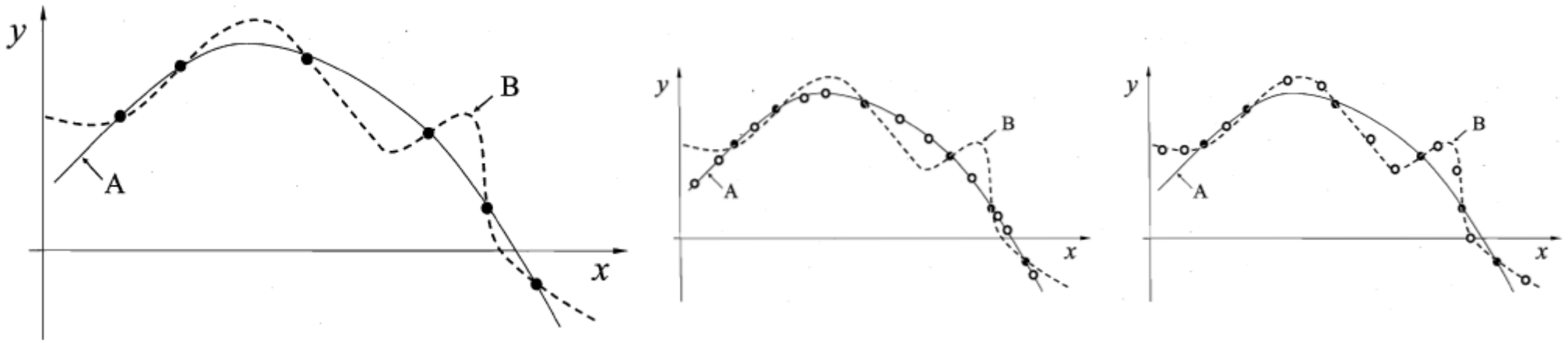
目标: 找到与训练集“匹配”(fit)的假设

假设空间的大小： n1 x n2 x n3 + 1

归纳偏好

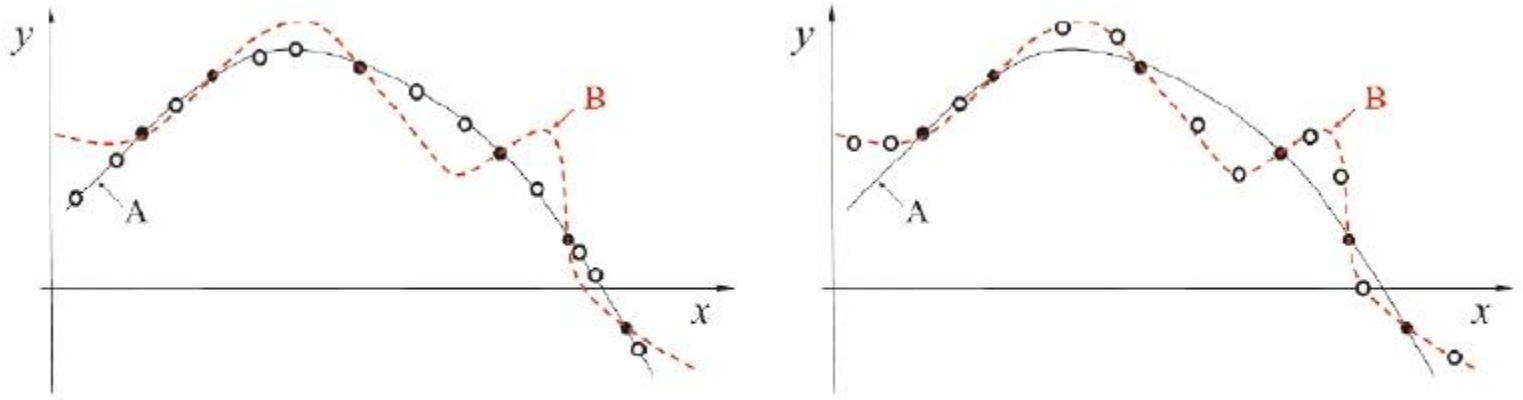
假设空间与归纳偏好（bias）

“如果有多个假设与观察一致，则选择最简单的那一个”（奥卡姆剃刀, Occam’s razor）



NFL定理

没有免费的午餐!



(a) A 优于 B　　(b) B 优于 A

图 1.4 没有免费的午餐. (黑点: 训练样本; 白点: 测试样本)

NFL定理：一个算法 $\mathfrak{L}_a$ 若在某些问题上比另一个算法 $\mathfrak{L}_b$ 好，必存在另一些问题，$\mathfrak{L}_b$ 比 $\mathfrak{L}_a$ 好。

NFL定理

简单起见，假设样本空间 $\mathcal{X}$ 和假设空间$\mathcal{H}$离散,令 $P(h|X,\mathfrak{L}_a)$代表算法$\mathfrak{L}_a$基于训练数据 X 产生假设 h 的概率，f 代表要学的目标函数，$\mathfrak{L}_a$在训练集之外所有样本上的总误差为

$$E_{ote}(\mathfrak{L}_a|X,f)=\sum_{h}\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\,\mathbb{I}\left(h\left(\boldsymbol{x}\right)\neq f\left(\boldsymbol{x}\right)\right)P\left(h\mid X,\mathfrak{L}_a\right)$$

考虑二分类问题，目标函数可以为任何函数 $\mathcal{X}\mapsto\{0,1\}$ ，函数空间为 $\{0,1\}^{|\mathcal{X}|}$，对所有可能的 f 按均匀分布对误差求和，有

$$\sum_{f}E_{ote}(\mathfrak{L}_a|X,f)=\sum_{f}\sum_{h}\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\,\mathbb{I}(h(\boldsymbol{x})\neq f(\boldsymbol{x}))\;P(h\mid X,\mathfrak{L}_a)$$

NFL定理

$$\sum_f E_{ote}(\mathfrak{L}_a|X,f) = \sum_f \sum_h \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x})\ \mathbb{I}(h(\boldsymbol{x}) \neq f(\boldsymbol{x}))\ P(h \mid X, \mathfrak{L}_a)$$

$$= \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h \mid X, \mathfrak{L}_a) \sum_f \mathbb{I}(h(\boldsymbol{x}) \neq f(\boldsymbol{x}))$$

$$= \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h \mid X, \mathfrak{L}_a) \frac{1}{2} 2^{|\mathcal{X}|}$$

$$= \frac{1}{2} 2^{|\mathcal{X}|} \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h \mid X, \mathfrak{L}_a)$$

$$= 2^{|\mathcal{X}|-1} \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \cdot 1$$

总误差与学习算法无关！

所有算法一样好！

NFL定理

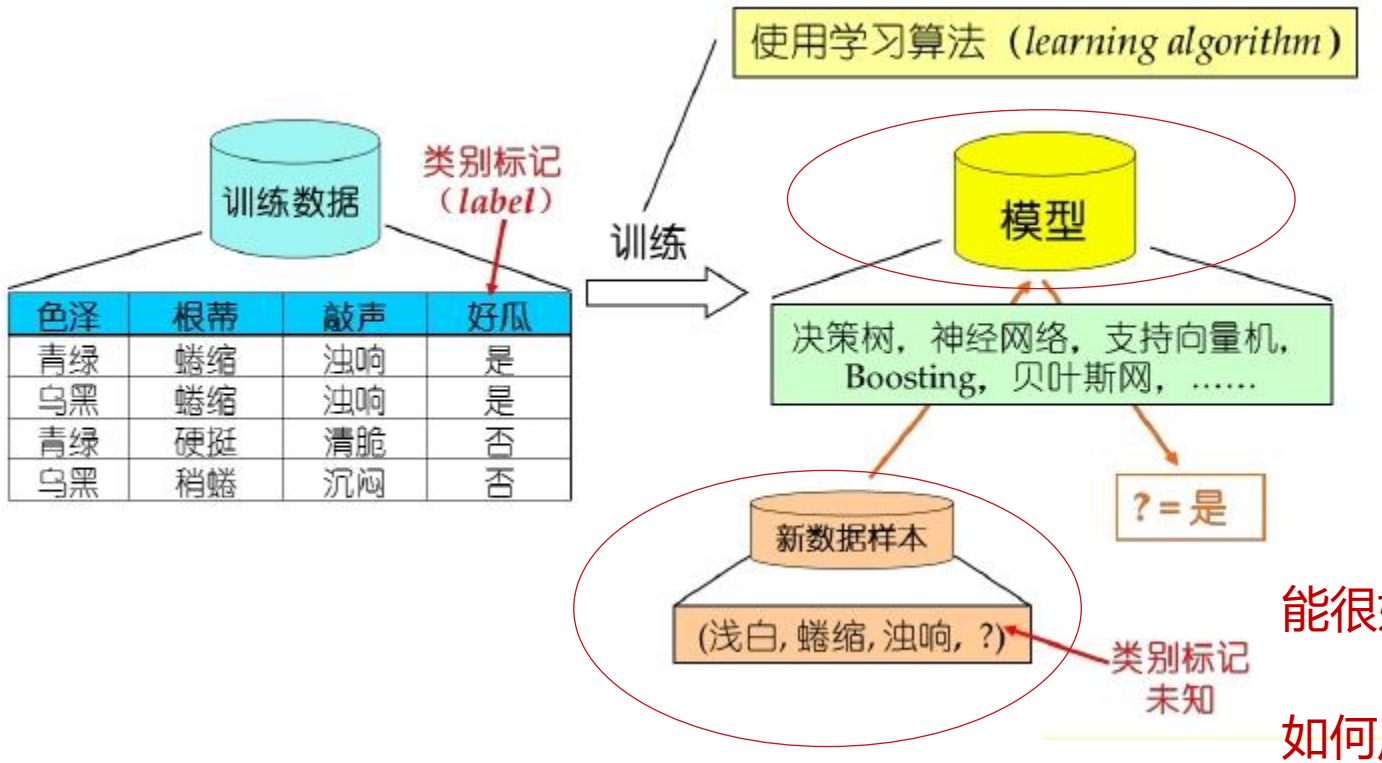
NFL定理的重要前提：

所有“问题”出现的机会相同、或所有问题同等重要

实际情形并非如此；我们通常只关注自己正在试图解决的问题

脱离具体问题，空泛地谈论“什么学习算法更好”
毫无意义！

模型选择



色泽	根蒂	敲声	好瓜
青绿	蜷缩	浊响	是
乌黑	蜷缩	浊响	是
青绿	硬挺	清脆	否
乌黑	稍蜷	沉闷	否

什么模型好？

能很好地适用新数据样本

如何度量？

测试误差 vs 训练误差

- **测试（泛化）误差：在“未来”样本上的误差**
- **训练（经验）误差：在训练集上的误差**

- 泛化误差越小越好（GOAL）
- 经验误差是否越小越好?

NO! 因为会出现“过拟合” (overfitting)

过拟合和欠拟合

图 2.1 过拟合、欠拟合的直观类比

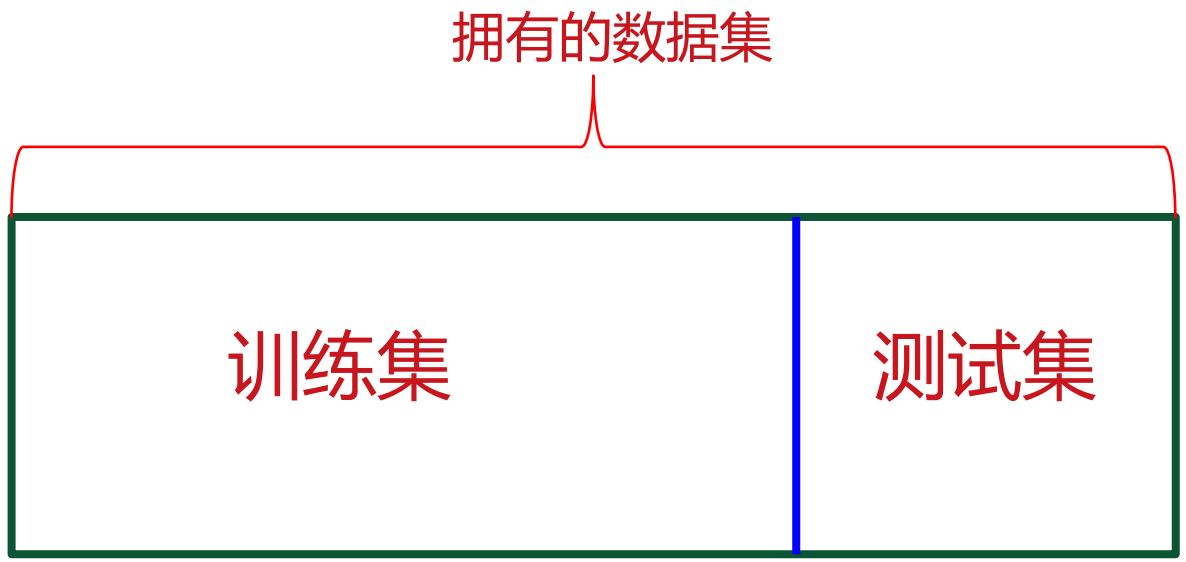
模型选择

- ✓ 如何评估?
- ✓ 如何度量评估结果?
- ✓ 如何判断模型差异?

如何评估

- 测试评估方法
- 核心：怎么获得“测试集”(test set)？
  - hold-out
  - kth-cross-validation
  - bootstrapping

Hold Out



注意：

- 保持数据分布一致性 （例如: 分层采样)
- 多次重复划分 (例如: 100次随机划分)
- 测试集不能太大、不能太小 (例如： 1/5~1/3)

k-折交叉验证法

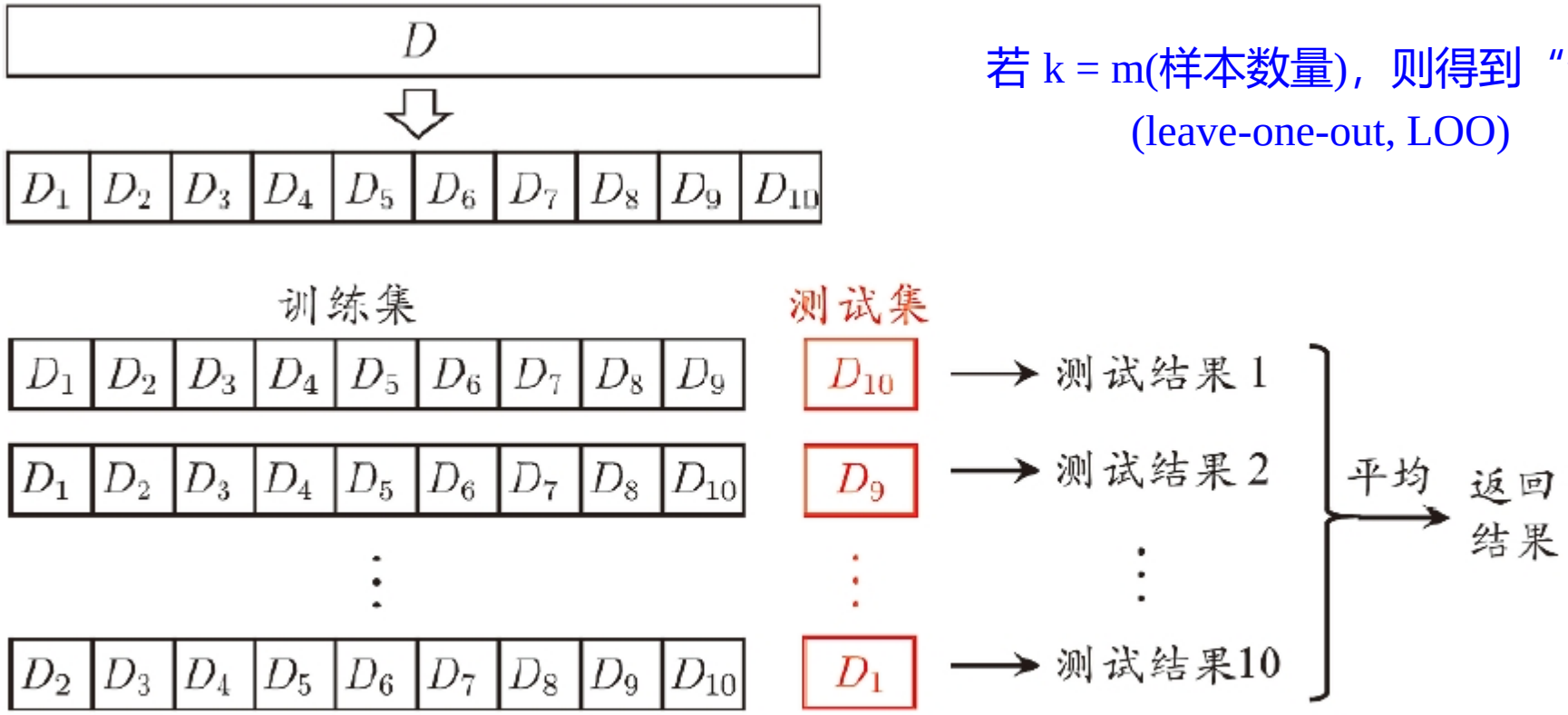
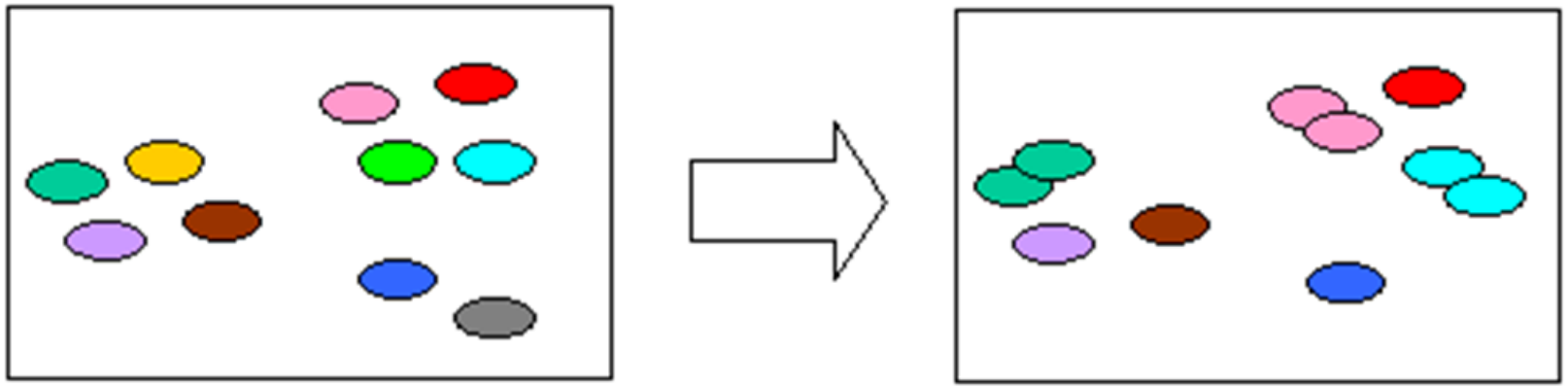


图 2.2　10 折交叉验证示意图

若 k = m(样本数量)，则得到“留一法”(leave-one-out, LOO)

bootsrap sampling

亦称“有放回采样”、“可重复采样”



约有 36.8% 的样本不出现

$$\lim_{m\mapsto\infty}\left(1-\frac{1}{m}\right)^{m}\mapsto\frac{1}{e}\approx 0.368$$

“包外估计”(out-of-bag estimation)

- 训练集与原样本集同规模
- 数据分布有所改变

“调参”与最终模型

算法的参数：一般由人工设定，亦称“超参数”

模型的参数：一般由学习确定

调参过程相似：先产生若干模型，然后基于某种评估方法进行选择

参数调得好不好对性能往往对最终性能有关键影响

区别：训练集 vs. 测试集 vs. 验证集 (validation set)

算法参数选定后，要用“训练集+验证集”重新训练最终模型

性能度量

- **性能度量(performance measure)是衡量模型泛化能力的评价标准，反映了任务需求**
- 使用不同的性能度量往往会导致不同的评判结果
- **什么样的模型是“好”的，不仅取决于算法和数据，还取决于任务需求**

回归(regression)任务常用均方误差（MSE）：

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\left(f\left(\boldsymbol{x}_i\right)-y_i\right)^2$$

性能度量

分类问题指标：

- 错误率：

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f\left(\boldsymbol{x}_i\right)\neq y_i\right)$$

- 准确率（精度）：

$$\begin{aligned}\mathrm{acc}(f;D) &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f\left(\boldsymbol{x}_i\right)=y_i\right)\\ &= 1-E(f;D)\ .\end{aligned}$$

性能度量

分类问题指标：

表 2.1　分类结果混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP (真正例)	FN (假反例)
反例	FP (假正例)	TN (真反例)

- 查准率：$P = \frac{TP}{TP + FP}$
- 查全率：$R = \frac{TP}{TP + FN}$

性能度量

分类问题指标：

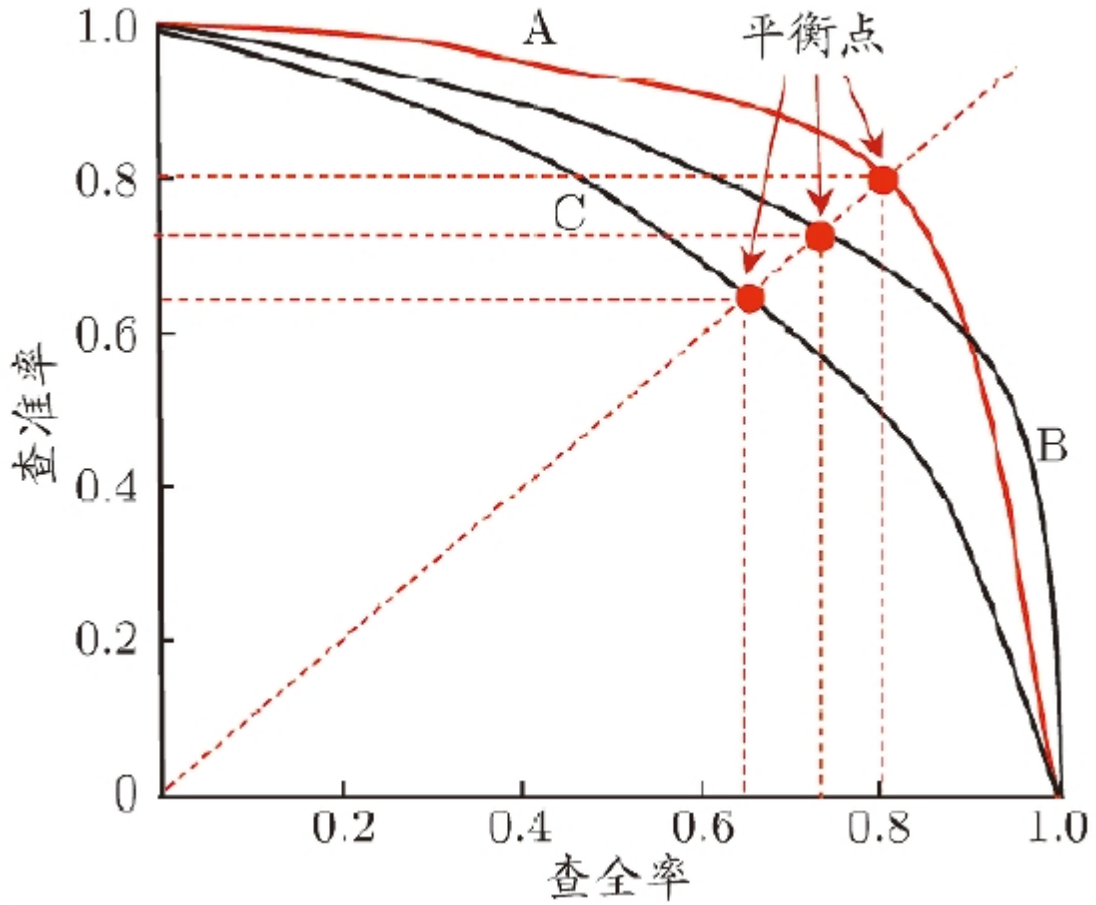
- PR图，BEP，AUPR

PR图：

- 学习器 A 优于 学习器 C
- 学习器 B 优于 学习器 C
- 学习器 A ?? 学习器 B

BEP：

- 学习器 A 优于 学习器 B
- 学习器 A 优于 学习器 C
- 学习器 B 优于 学习器 C



分类问题指标：

- F1指标：

$$F1 = \frac{2 \times P \times R}{P + R} = \frac{2 \times TP}{\text{样例总数} + TP - TN}$$

- 若对查准率/查全率有不同偏好：

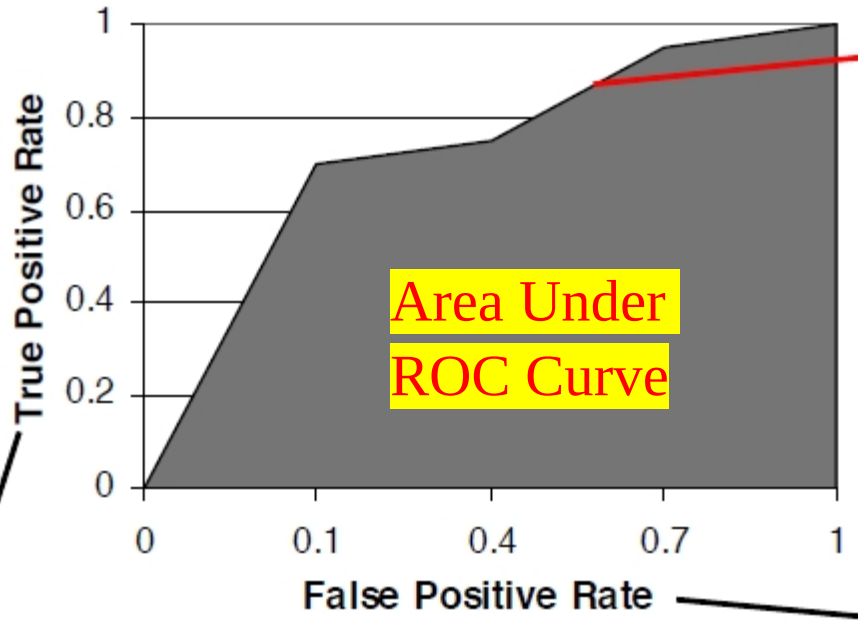
$$F_\beta = \frac{(1 + \beta^2) \times P \times R}{(\beta^2 \times P) + R}$$

$\beta > 1$ 时查全率有更大影响；$\beta < 1$ 时查准率有更大影响

分类问题指标：

- ROC, AUC

ROC (Receiver Operating Characteristic) Curve



$$tpr = \frac{TP}{TP + FN} = \frac{TP}{m_+}$$

$$fpr = \frac{FP}{FP + TN} = \frac{FP}{m_-}$$

判断模型差异

成对t检验：

k 折交叉验证； 5x2交叉验证

统计显著性

对两个学习器 A 和 B，若我们使用 k 折交叉验证法得到的测试错误率分别为 $\epsilon_1^A, \epsilon_2^A, \ldots, \epsilon_k^A$ 和 $\epsilon_1^B, \epsilon_2^B, \ldots, \epsilon_k^B$，其中 ϵ_i^A 和 ϵ_i^B 是在相同的第 i 折训练/测试集上得到的结果，则可用 k 折交叉验证“成对 t 检验”(paired t-tests)来进行

$$\tau_t = \left| \frac{\sqrt{k}\mu}{\sigma} \right|$$

思考

1、评估指标AUC与Acc是否一致？举例说明其适合的应用场景。

2、AUC的最大值和最小值分别为多少？分别在何种情况下取得？