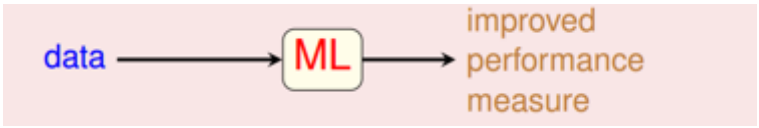
机器学习的概念

通过数据计算的经验改进性能度量



适用机器学习处理的问题

1.存在某些暗含的规律

2.很难用程序简单的找出规律

3.有关于这个模型的数据

Exercise

Which of the following is best suited for machine learning?

1. predicting whether the next cry of the baby girl happens at an even-numbered minute or not
2. determining whether a given graph contains a cycle
3. deciding whether to approve credit card to some customer
4. guessing whether the earth will be destroyed by the misuse of nuclear power in the next ten years

Reference Answer: 3

1. no pattern
2. programmable definition
3. pattern: customer behavior;
   definition: not easily programmable;
   data: history of bank operation
4. arguably no (or not enough) data yet

机器学习的应用

食物识别

衣服搭配

姿势捕捉

交通信号的识别

娱乐：聊天机器人

法律：协助法官判案

机器学习的组成

Basic Notations

- Bold capital letters (e.g., X) → Matrices;
- Bold lowercase letters (e.g., x) → Vectors; 向量
- Non-bold letters (e.g., x) → Scalars; 标量
- Greek letters (e.g., β) → The parameters. 参数

加粗的大写字母代表矩阵，加粗的小写字母代表向量，没加粗的字母代表标量，希腊字母代表参数

根据已有的数据进行训练，得到g(x)≈f(x)

机器学习的种类

根据输出Y

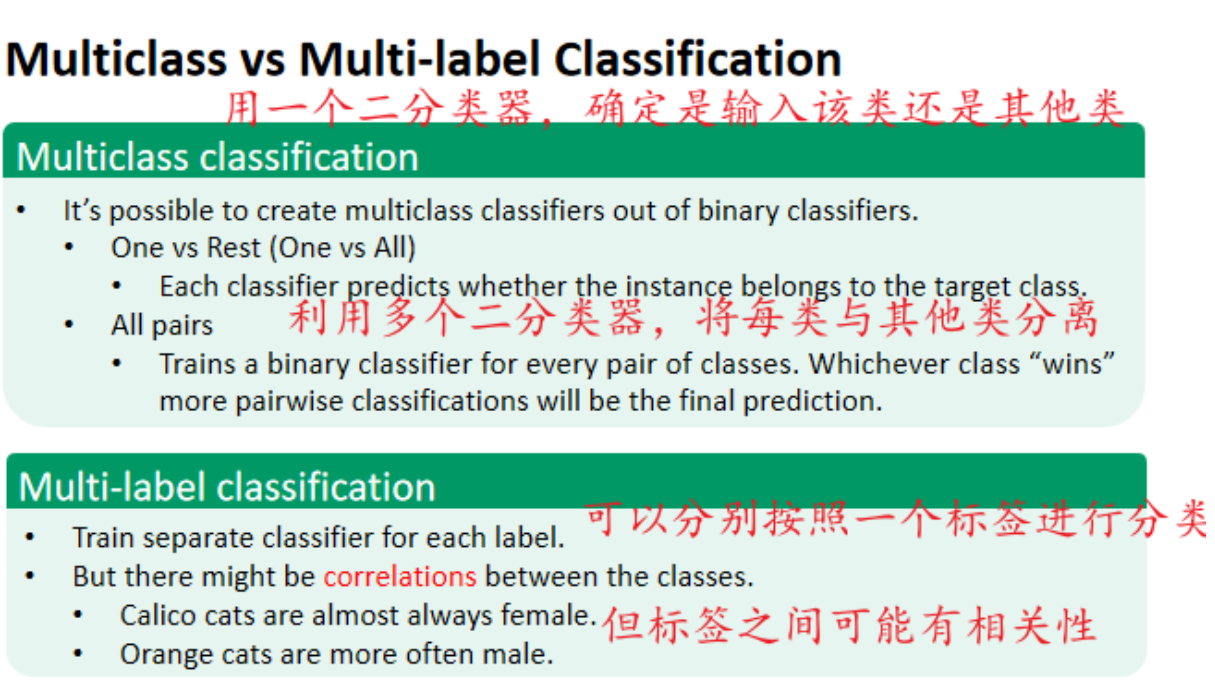
二分类

多重分类：识别手写体数字，识别照片属于哪种水果，

多标签分类

多重分类是将训练数据按照一个属性多个值进行输出------>**预测一个属性，如猫的性别**

多标签是将训练数据按照多个属性，每个属性由多个值进行输出------->**预测多个属性，如猫的性别和品种**



线性回归：**拟合一个曲线或直线**

常见的线性回归问题：

病人特征-->病人痊愈需要的时间

图片->热度预测

电子商务产品->售价预测

Learning with different output space $\mathcal{Y}$

- Binary classification: $\mathcal{Y} = \{-1, +1\}$
- Multiclass classification: $\mathcal{Y} = \{1, 2, \ldots, K\}$
- Regression: $\mathcal{Y} = \mathbb{R}$

练习：

Exercise

What is this learning problem?

The entrance system of the school gym, which does automatic face recognition based on machine learning, is built to charge four different groups of users differently: Staff, Student, Professor, Other. What type of learning problem best fits the need of the system?

1. binary classification
2. multiclass classification
3. regression
4. multilabel classification

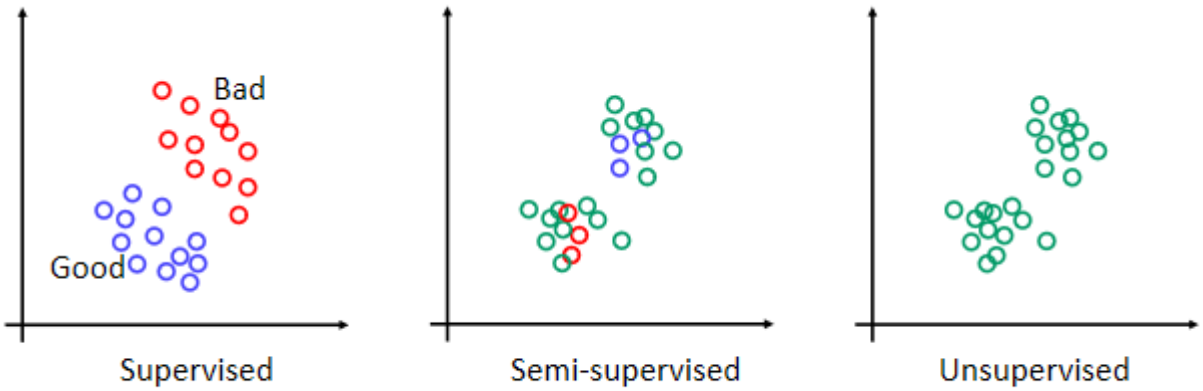
根据输入数据不同yn标签

监督学习

半监督学习

半监督学习：利用未标记的数据来避免“**昂贵**”的标记。

无监督学习



强化学习——>通过反馈

没有监管者，只有奖励信号。

时间真的很重要（按顺序）。

代理人的行动会影响其收到的后续数据。强化学习基于奖励假设。

所有目标都可以用预期累积奖励的最大化来描述。

代理人目标：最大化累积奖励。选择行动以最大化预期累积奖励

Reinforcement Learning

- Fly stunt manoeuvre in a helicopter (直升机特技飞行)
  - + reward for following desired trajectory
  - - reward for crashing
- Make a humanoid robot walk让人形机器人行走
  - + reward for forward motion
  - - reward for falling over
- Recycling robot回收机器人
  - + reward for finding cans
  - - reward for running out of battery
- Ad system广告系统
  - + user click
  - - no click

下棋机器人--强化学习

聊天机器人--强化学习

Exercise

What is this learning problem?

To build a tree recognition system, a company decides to gather one million of pictures on the Internet. Then, it asks each of the 10 company members to view 100 pictures and record whether each picture contains a tree. The pictures and records are then fed to a learning algorithm to build the system. What type of learning problem does the algorithm need to solve?

1. supervised
2. unsupervised
3. semi-supervised
4. reinforcement

根据不同的协议f(xn,yn)

批量学习（Batch Learning)

数据都已知

线上学习（Online Learning）

在线学习：从顺序数据中学习。

将推测结果也加入输入数据

主动学习

在现代机器学习问题中，主动学习的动机很好，在这些问题中，**数据可能很丰富，但标签很少或获取成本很高**。

主动学习（Active Learning）是一种机器学习方法，它允许**模型主动选择哪些样本应该被标记以进行训练**，以便最大化模型性能和减少标记数据的成本。与传统的机器学习方法不同，**主动学习不需要大量的标记数据**，因为它可以通过选择最有用的样本来最大化模型性能。主动学习通常用于处理大规模数据集，并且可以在训练过程中动态地选择最有用的样本，以提高模型的准确性和泛化能力。------**主动选择样本**

Adaboost不属于主动学习，它是一种集成学习方法，用于提高分类器的准确性。在Adaboost中，多个弱分类器被组合成一个强分类器，每个弱分类器都被训练以解决不同的子问题。Adaboost的训练过程是通过迭代地调整每个弱分类器的权重来完成的，以最大化整体分类器的准确性。与主动学习不同，**Adaboost需要有标记的训练数据，并且在训练过程中不会主动选择哪些样本应该被标记。**

根据不同的输入空间X

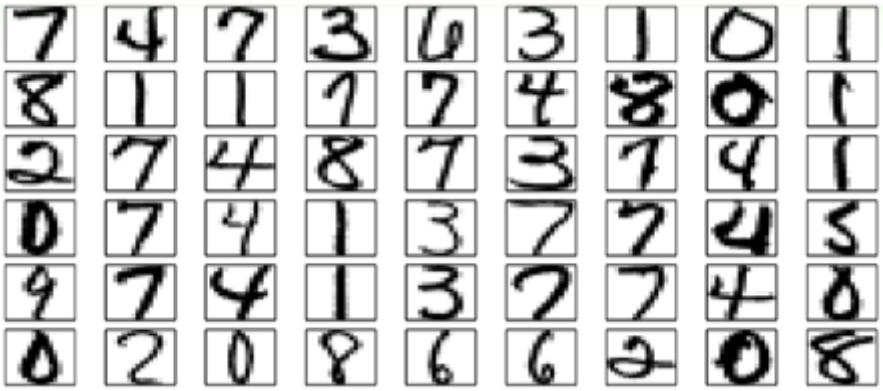
concrete feature，**混合特征**

比如：尺寸、质量->硬币分类；客户信息（性别、职业->放贷

需要人类决策哪些特征是重要的

raw features**原始特征**

手写体



经常需要人类转化为混合特征

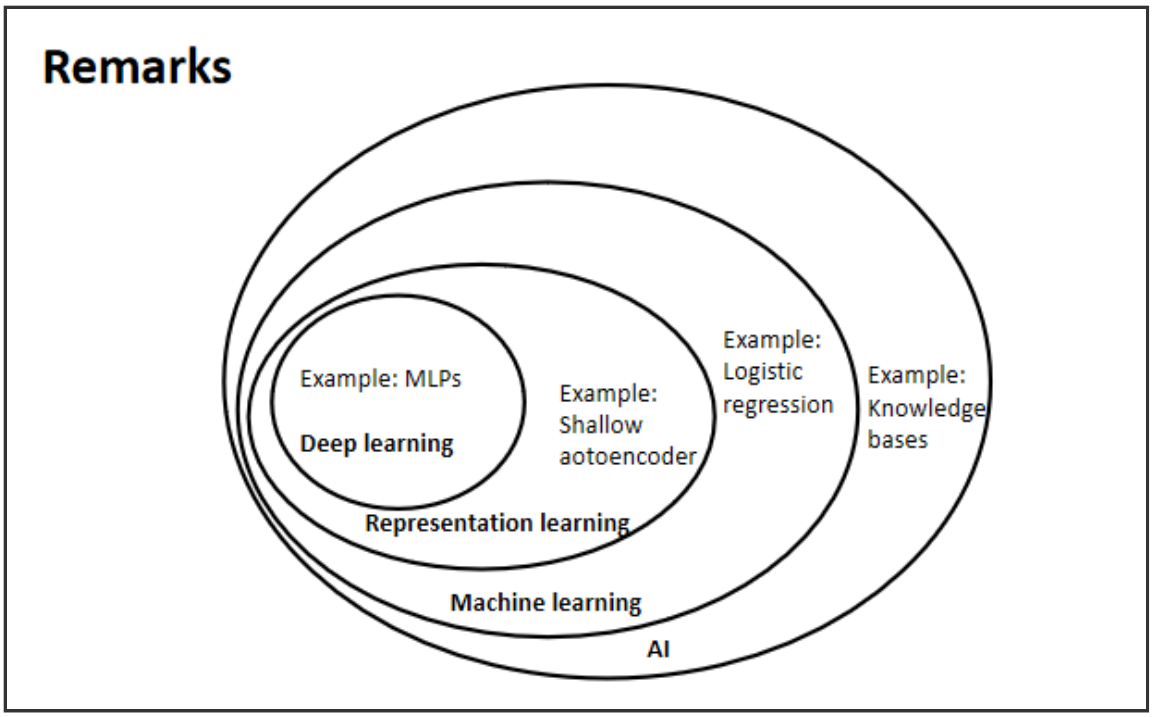
abstract features**抽象特征**

无实际意义的特征，对于机器学习来说比较难

抽象特征还需要由人或机器进行“特征转换/构建”。

Types of Machine Learning

- Learning with different output space y
  - [classification], [regression] 分类，回归
- Learning with different data label y_n 监督学习、半监督学习、无监督学习
  - [supervised], un/semi-supervised, reinforcement
- Learning with different protocol $f(x_n, y_n)$
  - [batch], online, active 批处理，在线，主动学习
- Learning with different input space x
  - [concrete, raw], abstract



representation learning

表征学习是机器学习的一个子领域，其重点是自动学习数据的表征，以执行特定任务，例如分类或聚类。表征学习的目标是找到一种紧凑且信息丰富的数据表示，以捕捉底层结构和模式。然后可以使用此表示来比直接使用原始数据更有效地和准确地执行各种任务。表征学习中常用的技术包括深度神经网络、自编码器和变分方法。