数据科学 5 机器学习介绍

机器学习在日常生活中的应用，从左上角按照顺时针方向依次使用到的机器学习技术分别为：人脸识别、手写数字识别、垃圾邮件过滤和亚马逊公司的产品推荐

何谓机器学习

把无序的数据转换成有用的信息，海量数据抽取有价值的信息。
创建并使用那些由学习数据而得出的模型，预测建模或数据挖掘。
用已存在的数据来开发可用来对新数据预测多种可能结果的模型。

专家系统
例如鸟类识别专家系统
测量所有可测属性（特征）

表1-1　基于四种特征的鸟物种分类表

	体重（克）	翼展（厘米）	脚　蹼	后背颜色	种　属
1	1000.1	125.0	无	棕色	红尾鵟
2	3000.7	200.0	无	灰色	鹭鹰
3	3300.0	220.3	无	灰色	鹭鹰
4	4100.0	136.0	有	黑色	普通潜鸟
5	3.0	11.0	无	绿色	瑰丽蜂鸟
6	570.0	75.0	无	黑色	象牙喙啄木鸟

6个训练样本的训练集，每个训练样本4种特征，1个目标变量

前两种特征：数值型
第三种特征：布尔型
第四种特征：枚举型

机器学习主要任务-分类

为算法输入大量已分类数据作为算法的训练集

测试机器学习算法效果，通常使用两套独立的样本集：训练数据和测试数据

监督学习

算法必须知道目标变量的分类信息，分类和回归

应用：预测目标变量的值

如目标变量是离散型（如是 / 否，1/2/3，红黄蓝），选分类器算法
如目标变量是连续型（如0.0 - 100.00)，选回归算法

无监督学习

算法不知道目标变量，没有类别信息
聚类： 将数据集合分成由类似的对象组成的多个类的过程
密度估计：寻找描述数据统计值的过程

应用： 不预测目标变量

如需要将数据划分为离散的组，选聚类算法
需要估计数据与每个分组的相似程度，选密度估计算法

开发机器学习算法程序的步骤

1. 收集数据
2. 准备输入数据
3. 分析输入数据

   是否有异常值？是否有空值？是否有规律？

   通过图形化展示数据
4. 训练算法

   无监督学习不需要训练算法
5. 测试算法
6. 使用算法