机器学习

1.什么是机器学习

机器学习是通过**编程**让计算机从**数据**中进行**学习**的科学（和艺术）。

2.机器学习的优点

- 需要进行大量手工调整或需要拥有长串规则才能解决的问题：机器学习算法通常可以**简化代码、提高性能。**
- 问题复杂，传统方法难以解决：最好的机器学习方法可以**找到解决方案。**
- 环境有波动：机器学习算法可以**适应新数据。**
- 洞察复杂**问题**和大量**数据。**

3.机器学习的类型

- 是否在人类监督下进行训练（监督，非监督，半监督和强化学习）
- 是否可以动态渐进学习（在线学习 vs 批量学习）
- 它们是否只是通过简单地比较新的数据点和已知的数据点，还是在训练数据中进行模式识别，以建立一个预测模型，就像科学家所做的那样（基于实例学习 vs 基于模型学习）

规则并不仅限于以上的，你可以将他们进行组合。

4.监督学习VS非监督学习VS半监督学习VS强化学习

监督学习

在监督学习中，用来训练算法的**训练数据**包含了**答案**，称为**标签**

典型任务：分类、预测目标数值

下面是一些重要的监督学习算法（本书都有介绍）：

- K 近邻算法
- 线性回归
- 逻辑回归
- 支持向量机（SVM）
- 决策树和随机森林
- 神经网络

例子：垃圾邮件过滤器

非监督学习

训练数据是**没有加标签**的，系统在没有老师的条件下进行学习。

下面是一些最重要的非监督学习算法（我们会在第 8 章介绍降维）：

- **聚类**
  K 均值
  层次聚类分析（Hierarchical Cluster Analysis，HCA）
  期望最大值
- **可视化和降维**
  主成分分析（Principal Component Analysis，PCA）
  核主成分分析
  局部线性嵌入（Locally-Linear Embedding，LLE）
  t-分布邻域嵌入算法（t-distributed Stochastic Neighbor Embedding，t-SNE）
- **关联性规则学习**
  Apriori 算法
  Eclat 算法

常见的非监督任务：降维（降维的目的是简化数据、但是不能失去大部分信息。做法之一是合并若干相关的特征）、异常检测、关联规则学习（它的目标是挖掘大量数据以发现属性间有趣的关系）。

例子：可视化

半监督学习

一些算法可以处理部分带标签的训练数据，通常是大量不带标签数据加上小部分带标签数据。这称作**半监督学习。**

强化学习

强化学习非常不同。学习系统在这里被称为**智能体**（agent），可以对**环境**进行**观察、选择和执行**动作，并获得奖励作为回报（负奖励是惩罚）。然后它必须**自己学习**哪个是最佳方法（称为**策略**，policy），以得到长久的最大奖励。**策略**决定了智能体在**给定情况**下应该采取的**行动。**

例子：AlphaGo

5.批量学习VS在线学习

批量学习

在批量学习中，系统**不能进行持续学习**：必须用**所有可用数据**进行训练。这通常会占用大量时间和计算资源，所以一般是**线下**做的。

首先是进行**训练**，然后部署在**生产环境且停止学习**，它只是**使用已经学到的策略。**这称为**离线学习。**

在线学习

在在线学习中，是用**数据实例持续地进行训练**，可以一次一个或一次几个实例（称为小批量）。每个学习步骤都很快且廉价，所以系统可以**动态地学习**收到的**最新数据。**

6.基于实例的学习VS基于模型的学习

基于实例的学习

基于实例学习：系统先用**记忆学习案例**，然后使用**相似度测量**推广到**新的例子。**

基于模型的学习

从样本集进行**归纳**的方法是**建立这些样本的模型**，然后使用这个模型进行**预测。**这称作基于模型学习。