第 1 章　机器学习介绍

本章简要介绍了机器学习的定义、应用场景及机器学习的分类，并通过一个简单的示例介绍了机器学习的典型步骤，以及机器学习领域的一些专业术语。本章涵盖的内容如下：

- 机器学习的概念；
- 机器学习要解决的问题分类；
- 使用机器学习解决问题的一般性步骤。

1.1　什么是机器学习

机器学习是近年来的一大热门话题，然而其历史要倒推到半个多世纪之前。1959 年 Arthur Samuel 给机器学习的定义是：

Field of study that gives computers the ability to learn without being explicitly programmed

即让计算机在没有被显式编程的情况下，具备自我学习的能力。

Tom M. Mitchell 在操作层面给出了更直观的定义：

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

翻译过来用大白话来说就是：针对某件事情，计算机会从经验中学习，并且越做越好。从机器学习领域的先驱和“大牛”们的定义来看，我们可以自己总结出对机器学习的理解：**机器学习是一个计算机程序，针对某个特定的任务，从经验中学习，并且越做越好。**

从这个理解上，我们可以得出以下针对机器学习最重要的内容。

数据：经验最终要转换为计算机能理解的数据，这样计算机才能从经验中学习。谁掌握的数据量大、质量高，谁就占据了机器学习和人工智能领域最有利的资本。用人类来类比，数据就像我们的教育环境，一个人要变得聪明，一个很重要的方面是能享受到优质的教育。所以，从这个意义来讲，就能理解类似 Google 这种互联网公司开发出来的机器学习程序性能为什么那么好了，因为他们能获取到海量的数据。

模型：即算法，是本书要介绍的主要内容。有了数据之后，可以设计一个模型，让数据作为输入来训练这个模型。经过训练的模型，最终就成了机器学习的核心，使得模型成为了能产生决策的中枢。一个经过良好训练的模型，当输入一个新事件时，会做出适当的

反应，产生优质的输出。

1.2 机器学习有什么用

受益于摩尔定律，随着计算机性能的提高，以及计算资源变得越来越便宜，机器学习在诞生半个世纪后的今天，得到了越来越广泛的应用。你可能感受不到，但是你的日常生活已经与人工智能密不可分了。

早晨起床，用 iPhone 打开 Siri，问：“今天天气怎么样？”。Siri 会自动定位到当前你所在的城市，并且把天气信息展现出来。这个功能用起来很简单，但其背后的系统是异常复杂的。

其一是语音识别，这是机器学习最早的应用研究领域，Siri 需要先把你说的话转换为文字。大家知道，语音从本质上是一系列幅度不同的波，要转换为文字，就需要设计一个模型，先通过大量的语音输入来训练这个模型，等模型训练好了，把语音作为输入，就可以输出文字了。语音识别在 20 世纪 50 年代就开始研究了，其模型是不断演变的。一个比较大的演变，就是由基于模式识别的算法演变为基于统计模型的算法，这一转变大大提高了语音识别的准确率。

其二是自然语言处理，这是机器学习和人工智能又一个非常重要的研究方向。Siri 把语音转成文字后，软件需要理解文字的意思才能给出准确的回答。要让计算机理解文字可不是简单的事情。首先要有大规模的语料库，其次要有相应的语言模型，然后通过语料库来训练语言模型，最终才能理解文字的部分语义。关于自然语言处理以及搜索引擎的相关技术，可以参阅吴军老师的《数学之美》，这是一本把高深的数学讲得通俗易懂、妙趣横生的科普读物。

我们接着讲前面起床的故事。在洗漱期间，你抽空浏览手机上的新闻，发现新闻下方有感兴趣的行车记录仪的广告，点进去后打开了某知名电商网站，你看了一下产品的价格和评价，顺手就买了。接着浏览新闻，发现这个新闻客户端越来越人性化，自动把你感兴趣的 IT 新闻及体育新闻排在了首页。好不容易收拾完毕可以出门了，你坐在地铁上，打开音乐播放器，浏览了一遍曲库，没有找到特别想听的歌，于是就让系统给你推荐一些歌。系统推荐的歌还挺“靠谱”的，虽然很多都没听过，但都很对你的“胃口”。

在这段体验描述里，背后的功臣就是推荐系统，这也是机器学习的一个重要应用方向。推荐系统的核心，是不断地学习用户的使用习惯，从而刻画出用户的画像，根据用户的画像去推荐用户感兴趣的商品和文章。

公司新上线了人脸识别系统，在这个“刷脸”的时代，已经没有“忘带工牌”这个签卡的借口了。你走到公司大门口，人脸识别系统自动把你识别出来，然后开门，并准确地通过语音播报的方式和你打招呼。

目前最先进的人脸识别系统基本上都是基于深度学习模型的算法实现的。这一领域也

由早期的传统方法慢慢地被深度学习模型所替代。

当然，机器学习不止这些应用场景。我们在介绍具体算法的时候，会再详细列举出每个算法的应用场景。

延伸阅读：强人工智能

未来学家 Ray Kurzweil 预言，人类将在 2045 年实现强人工智能，就是说到时人工智能将远远强于人类。那个时候人类与强人工智能的差距，要比蚂蚁与人类的差距大几个数量级。这是个让人“脑洞”大开的想象。网上有一篇很火的翻译过来的文章“为什么最近有很多名人，比如比尔盖茨，马斯克、霍金等，让人们警惕人工智能？”，推荐读者阅读一下，其比普通的科幻小说要好看得多。喜欢阅读英文原文的读者，可以在 waitbutwhy.com 上搜索“The AI Revolution”。

1.3　机器学习的分类

机器学习可以分成以下两类。

有监督学习（Supervised learning）通过大量已知的输入和输出相配对的数据，让计算机从中学习出规律，从而能针对一个新的输入做出合理的输出预测。比如，我们有大量不同特征（面积、地理位置、朝向、开发商等）的房子的价格数据，通过学习这些数据，预测一个已知特征的房子价格，这种称为**回归学习（Regression learning）**，即输出结果是一个具体的数值，它的预测模型是一个连续的函数。再比如我们有大量的邮件，每个邮件都已经标记是否是垃圾邮件。通过学习这些已标记的邮件数据，最后得出一个模型，这个模型对新的邮件，能准确地判断出该邮件是否是垃圾邮件，这种称为**分类学习（Classfication learning）**，即输出结果是离散的，即要么输出 1 表示是垃圾邮件，要么输出 0 表示不是垃圾邮件。

无监督学习（Unsupervised learning）通过学习大量的无标记的数据，去分析出数据本身的内在特点和结构。比如，我们有大量的用户购物的历史记录信息，从数据中去分析用户的不同类别。针对这个问题，我们最终能划分几个类别？每个类别有哪些特点？我们事先是不知道的。这个称为**聚类（Clustering）**。这里需要特别注意和有监督学习里的分类的区别，分类问题是我们已经知道了有哪几种类别；而聚类问题，是我们在分析数据之前其实是不知道有哪些类别的。即分类问题是在已知答案里选择一个，而聚类问题的答案是未知的，需要利用算法从数据里挖掘出数据的特点和结构。

网络上流传一个阴谋论：如果你是一个很好说话的人，网购时收到有瑕疵的商品的概率会比较高。为什么呢？理由是电商库存里会有一部分有小瑕疵但不影响使用的商品，为了保证这些商品顺利地卖出去并且不影响用户体验，不被用户投诉，他们会把有瑕疵的商品卖给那些很好说话的人。可问题是，哪些人是好说话的人呢？一个最简单的方法是直接

把有小瑕疵的商品寄给一个用户，如果这个用户没有投诉或退货，并且还给出了好评，就说明他是个好说话的人。还可以通过机器学习来优化这一过程。电商网站有你的大量交易记录和行为记录，如果你从来没有投诉过，买之前也不会和卖家沟通太久，买之后也没有上网评价，或者全部给好评，那么机器学习算法从你的行为特征中会判定你为“好对付”的人。这样你就成了电商们的瑕疵商品的倾销对象了。在这个案例中，电商通过用户的行为和交易数据，分析出不同的用户特点，如哪些人是“老实”人、哪些人是有车一族、哪些人是“土豪”、哪些人家里有小孩等。这就属于无监督学习的聚类问题。

这两种机器学习类别的最大区别是，有监督学习的训练数据里有已知的结果来“监督”；而无监督学习的训练数据里没有结果“监督”，不知道到底能分析出什么样的结果。

1.4 机器学习应用开发的典型步骤

本节通过一个例子来介绍一下机器学习应用开发的典型步骤，以及机器学习领域的一些常用概念。假设，我们要开发一个房价评估系统，系统的目标是对一个已知特征的房子价格进行评估预测。建立这样一个系统需要包含以下几个步骤。

1.4.1 数据采集和标记

我们需要大量不同特征的房子和所对应的价格信息，可以直接从房产评估中心获取房子的相关信息，如房子的面积、地理位置、朝向、价格等。另外还有一些信息房产评估中心不一定有，比如房子所在地的学校情况，这一特征往往会影响房子的价格，这个时候就需要通过其他途径收集这些数据，这些数据叫做**训练样本**，或**数据集**。房子的面积、地理位置等称为**特征**。在数据采集阶段，需要收集尽量多的特征。特征越全，数据越多，训练出来的模型才会越准确。

通过这个过程也可以感受到数据采集的成本可能是很高的。人们常说石油是黑色的“黄金”，在人工智能时代，数据成了透明的“石油”，这也说明为什么蚂蚁金服估值这么高了。蚂蚁金服有海量的用户交易数据，据此他们可以计算出用户的信用指标，称为芝麻信用，根据芝麻信用给你一定的预支额，这就是一家新的信用卡公司了。而这还只是单单一个点的价值，真正的价值在于互联网金融。

在房价评估系统这个例子里，我们的房子价格信息是从房产评估中心获得的，这一数据可能不准确。有时为了避税，房子的评估价格会比房子的真实交易价格低很多。这时，就需要采集房子的实际成交价格，这一过程称为**数据标记**。标记可以是**人工标记**，比如逐个从房产中介那打听房子的实际成交价格；也可以是自动标记，比如通过分析数据，找出房产评估中心给的房子评估价格和真实成交价格的匹配关系，然后直接算出来。数据标记对有监督的学习方法是必须的。比如，针对垃圾邮件过滤系统，我们的训练样例必须包含

这个邮件是否为垃圾邮件的标记数据。

1.4.2　数据清洗

假设我们采集到的数据里，关于房子面积，有按平方米计算的，也有按平方英尺计算的，这时需要对面积单位进行统一。这个过程称为**数据清洗**。数据清洗还包括去掉重复的数据及噪声数据，让数据具备结构化特征，以方便作为机器学习算法的输入。

1.4.3　特征选择

假设我们采集到了 100 个房子的特征，通过逐个分析这些特征，最终选择了 30 个特征作为输入。这个过程称为**特征选择**。特征选择的方法之一是人工选择方法，即对逐个特征进行人员分析，然后选择合适的特征集合。另外一个方法是通过模型来自动完成，如本书即将介绍的 PCA 算法。

1.4.4　模型选择

房价评估系统是属于有监督学习的回归学习类型，我们可以选择最简单的线性方程来模拟。选择哪个模型，和问题领域、数据量大小、训练时长、模型的准确度等多方面有关。这方面的内容将在第 3 章介绍。

1.4.5　模型训练和测试

把数据集分成**训练数据集**和**测试数据集**，一般按照 8:2 或 7:3 来划分，然后用训练数据集来训练模型。训练出参数后再使用测试数据集来测试模型的准确度。为什么要单独分出一个测试数据集来做测试呢？答案是必须确保测试的准确性，即模型的准确性是要用它“没见过”的数据来测试，而不能用那些用来训练这个模型的数据来测试。理论上更合理的数据集划分方案是分成 3 个，此外还要再加一个**交叉验证**数据集。相关内容将在第 3 章介绍。

1.4.6　模型性能评估和优化

模型出来后，我们需要对机器学习的算法模型进行性能评估。性能评估包括很多方面，具体如下。

训练时长是指需要花多长时间来训练这个模型。对一些海量数据的机器学习应用，可能需要 1 个月甚至更长的时间来训练一个模型，这个时候算法的训练性能就变得很重要了。

另外，还需要判断数据集是否足够多，一般而言，对于复杂特征的系统，训练数据集越大越好。然后还需要判断模型的准确性，即对一个新的数据能否准确地进行预测。最后需要判断模型是否能满足应用场景的性能要求，如果不能满足要求，就需要优化，然后继续对模型进行训练和评估，或者更换为其他模型。

1.4.7 模型使用

训练出来的模型可以把参数保存起来，下次使用时直接加载即可。一般来讲，模型训练需要的计算量是很大的，也需要较长的时间来训练，这是因为一个好的模型参数，需要对大型数据集进行训练后才能得到。而真正使用模型时，其计算量是比较少的，一般是直接把新样本作为输入，然后调用模型即可得出预测结果。

本书的重点放在机器学习的算法介绍以及 scikit-learn 工具包的使用上。对数据采集、数据清洗、特征选择等内容没有深入介绍，但并不代表这些内容不重要。在实际工程应用领域，由于机器学习算法模型只有固定的几种，而数据采集、标记、清洗、特征选择等往往和具体的应用场景相关，机器学习工程应用领域的工程师打交道更多的反而是这些内容。

1.5 复习题

1. 机器学习分哪两类？它们之间有什么区别？
2. 无监督机器学习的优势有哪些？
3. 机器学习应用开发的典型步骤有哪些？
4. 为什么要把数据集分成训练数据集和测试数据集？