第3章 人工智能与机器学习

3.1 人工智能

这章我们介绍人工智能与机器学习。毕竟这两个词太火了，我们一样是要弄清楚它们的定义、之间的区别以及怎么通过去解释人工智能提升自己的价值。同时，还要介绍今后我们可能常会用到的一些方法，比如神经网络、贝叶斯、马尔可夫、自然语言处理等。

我们可以简单地回答说人工智能最直接的解释就是不是人的智能，但能像人那样思考，也可能超过人的智能。

人工智能(artificial intelligence，AI)被认为是21世纪三大尖端技术(人工智能、基因工程、纳米科学)之一，近几年来飞速发展，互联网科技巨头和无数中小创业公司投身其中，越来越多基于人工智能的应用开始渐渐走进我们的日常生活。

先举几个生活中常见的例子。

导航几乎是我们开车出行的必备应用之一，以大数据和机器学习为基础，是一种典型的地图人工智能化。在车里打开导航时，地图采集设备自动识别景物和道路特征定位你所处的位置，提取建筑轮廓并绘制形状，根据道路图形标牌、电子眼、警示牌等自动挖掘出过期或新增的信息点以及道路变化，并且根据道路实时路况计算规划出最优出行路径等。导航的整个过程不需要人工参与，机器根据算法和数据智能化输出结果，是离我们最近的人工智能应用之一。

信息获取是我们的基础需求之一，百度搜索和今日头条个性化推荐就是人工智能在信息分发领域的实际应用。经过深度学习的机器基于大数据根据你的检索关键词或者个人属性、行为记录等从数据库中自动调取、匹配和呈现信息。今日头条甚至已经有了人工智能写稿机器人，基于自然语言处理、视觉图形处理和机器学习技术等，AI写稿机器人能够根据网络热点、评论分享、用户喜好进行文字编写、标题封面图选择等，并在两秒钟内创作一篇效果不逊于人工编辑的稿件。

AI之所以重要，是因为它解决了极其复杂的问题，而这些问题的解决方案可以应用到对人类福祉重要的领域——从健康、教育，到商业、交通，乃至于公用事业和娱乐等。

那么，我们是不是说人工智能将取代人类或者就是灾难呢。

困难的问题是简单的，简单的问题是困难的。

其实，这里我们可以抛出一个哲学一样的话题。

2016年3月注定也要载入人工智能的发展史册：来自Google的人工智能程序AlphaGo以总比分4∶1的成绩战胜了前世界冠军李世石。

号称“人类最后智力骄傲”的围棋也被人工智能攻破了，一时间人工智能与机器人威胁论刷爆了微博、微信及各路新闻媒体。大家都在担心着某一天自己的工作会被人工智

能抢去，又在某一天人类会被人工智能机器人统治。那场比赛中有一个细节，不知大家是否注意：这个已经在“人类最后智力骄傲”上碾压人类的 AlphaGo，却连挪动一枚小小的棋子都需要人类帮助才能完成。

可能有人会说，这都不算事儿，围棋都已经战胜人类了，给 AlphaGo 装上机械手让它自己下棋也不过是分分钟的事儿。然而，事实真的是这么简单吗？

让计算机在智力测试或者下棋中展现出一个成年人的水平是相对容易的，但是要让计算机有如一岁小孩般的感知和行动能力却是相当困难甚至是不可能的。这便是在人工智能和机器人领域里著名的莫拉维克悖论。

莫拉维克悖论指出：和传统假设不同，对计算机而言，实现逻辑推理等人类高级智慧只需要相对很少的计算能力，而实现感知、运动等低等级智慧却需要巨大的计算资源。

这个很类似于数学发展中的理论，我们要证明最前沿的各种定理可能不难，但是要证明最简单的 1＋1＝2 可能却是最难的。

四岁小孩具有的本能——辨识人脸、举起铅笔、在房间内走动、回答问题等，事实上却是工程领域内目前为止最难解的问题。随着新一代智慧设备的出现，股票分析师、石化工程师和假释委员会都要小心他们的位置被取代，但是园丁、接待员和厨师至少 10 年内都不用有这种担心或者不需要担心。

3.2 机器学习

我们怎样去解释人工智能和机器学习的关系呢？

人工智能的根本在于智能——如何为机器赋予智能。而机器学习强调的是学习，或者说算法，是部署支持人工智能的计算方法。或者说人工智能是科学，机器学习是让机器变得更加智能的算法。

还记得我们强调的内容和形式的关系吗？我们可以说机器学习和人工智能是内容和形式的关系，或者说机器学习成就了人工智能。

机器学习是从大量的数据中发现规律，提取知识，并在实践中不断地完善和增强自我。机器学习是机器获取知识的根本途径，只有让计算机系统具有类似人的学习能力，才可能实现人工智能的终极目标。可以这么说，机器学习是人工智能研究的核心问题之一，也是当前人工智能研究的一个热门方向，同时也是人工智能理论研究和实际应用的主要瓶颈之一。

机器学习算法我们可以整体分为两类，不是按照有监督学习和无监督学习进行分类，而是按照容易方法论的角度进行分类的。

一种是线性化思维。体现在线性和非线性思维的转化中，比如线性回归是把非线性的转换成线性的进行研究，同时也是把高维的转换成低维的；支持向量机（SVM）其实就是反过来的，把低维的转换成高维的，把非线性的转换成线性的。

另一种是分类和聚类。在生活中，我们常常没有过多地去区分这两个概念，觉得聚类就是分类，分类也差不多就是聚类。分类与聚类之间在机器学习中有本质的区别。

1. 分类

分类是通过学习来得到样本属性与分析对象的关系。

具体来说，就是我们根据已知的一些样本，来得到分类模型（其实就是一种函数关系），然后通过目标函数来对只包含属性的样本数据进行分类。

分类要求必须事先明确知道各个类别的信息，并且断言所有待分类项都有一个类别与之对应。但是很多时候上述条件得不到满足，尤其是在处理海量数据的时候，如果通过预处理使得数据满足分类算法的要求，则代价非常大，这时候可以考虑使用聚类算法。

2. 聚类

聚类指事先并不知道任何样本的类别标号，希望通过某种算法来把一组未知类别的样本划分成若干类别。

聚类的时候，我们并不关心某一类是什么，我们需要实现的目标只是把相似的东西聚到一起，这在机器学习中被称作无监督学习（unsupervised learning），前面的分类就被称作有监督学习。

通常，人们根据样本间的某种距离或者相似性来定义聚类，即把相似的（或距离近的）样本聚为同一类，而把不相似的（或距离远的）样本归在其他类。

聚类的目标：组内的对象相互之间是相似的（相关的），而不同组中的对象是不同的（不相关的）。组内的相似性越大，组间差别越大，聚类就越好。

3. 分类和聚类的关系

聚类分析是研究如何在没有训练的条件下把样本划分为若干类。

在分类中，对于目标数据库中存在哪些类是知道的，要做的就是将每一条记录分别属于哪一类标记出来。

聚类需要解决的问题是将已给定的若干无标记的模式聚集起来，使之成为有意义的聚类，聚类是在预先不知道目标数据库到底有多少类的情况下，希望将所有的记录组成不同的类或者说聚类，并且使得在这种分类情况下，以某种度量（如距离）为标准的相似性，在同一聚类之间最小化，而在不同聚类之间最大化。

与分类不同，无监督学习不依赖于预先定义的类或带类标记的训练实例，需要由聚类学习算法自动确定标记，而分类学习的实例或数据样本有类别标记。

3.3 机器学习的常用算法

机器学习的常用算法很多，而且现在有许多已实现的机器学习开源包可供我们调用。如果我们能对常见的算法熟练掌握，比如通过使用 Python 实现算法，加深对机器学习的本质理解，那么，我们在量化投资领域里面就会有更多机遇。

1. 线性回归

这里说线性回归是因为我们讲的回归基本上就是线性回归，也就是用线性的方法去解决非线性的问题。说得更极端一些，机器学习里面的所有算法都可以看作是初中学习的一次函数思想，即 $Y=AX$ 来看机器学习发展过程中的各个理论或者算法。

2. 支持向量机(SVM)

SVM 的关键在于核函数。低维空间向量集通常难于划分,解决的方法是将它们映射到高维空间。但这种办法带来的困难就是计算复杂度的增加,而核函数正好巧妙地解决了这个问题。也就是说,只要选用适当的核函数,就可以得到高维空间的分类函数。在 SVM 理论中,采用不同的核函数将导致不同的 SVM 算法。在确定了核函数之后,由于确定核函数的已知数据也存在一定的误差,考虑到推广性问题,因此引入了松弛系数以及惩罚系数两个参变量来加以校正。在确定了核函数的基础上,再经过大量对比实验等取定这两个系数,该项研究就基本完成了,适合相关学科或业务内应用,且有一定能力的推广性。

3. 聚类

所谓聚类,就是将相似的事物聚集在一起,而将不相似的事物划分到不同的类别的过程,是机器学习中十分重要的一种手段。比如古典生物学之中,人们通过物种的形貌特征将其分门别类,可以说就是一种朴素的人工聚类。如此,我们就可以将世界上纷繁复杂的信息,简化为少数方便人们理解的类别,可以说是人类认知这个世界的最基本方式之一。

在数据分析的术语之中,分类和聚类是两种技术。分类是指我们已经知道了事物的类别,需要从样品中学习分类的规则,是一种有指导学习;而聚类则是由我们来给定简单的规则,从而得到分类,是一种无指导学习。两者可以说是相反的过程。

这里,我建议读者需要熟练掌握 EM 算法。

最大期望算法(expectation-maximization algorithm,EM),又称为期望最大化算法,在统计中被用于寻找,依赖于不可观察的隐性变量的概率模型中,参数的最大似然估计。

在统计计算中,最大期望(EM)算法是在概率(probabilistic)模型中寻找参数最大似然估计或者最大后验估计的算法,其中概率模型依赖于无法观测的隐藏变量(latent variable)。最大期望经常用在机器学习和计算机视觉的数据聚类(data clustering)领域。最大期望算法经过两个步骤交替进行计算,第一步是计算期望(E),利用对隐藏变量的现有估计值,计算其最大似然估计值;第二步是最大化(M),最大化是在 E 步上求得的最大似然值来计算参数的值。M 步上找到的参数估计值被用于下一个 E 步计算中,这个过程不断交替进行。

K-means 是最为常用的聚类方法之一,其实就是 EM 算法的一种特例。

4. 分类

贝叶斯是常见的分类方法,它是在量化投资中应用最多的一种策略。

贝叶斯算法的精髓在于其提供了一种数学法则来解释当有一系列新证据出现的情况下,你该如何改变自己现有的信念。一个典型的例子就是:看到第一次日出时,接着会想知道太阳是否会再次升起。于是赋予两个可能的结果同等的先验概率,并且在一个袋子里面放入一颗白球、一颗黑球,分别代表太阳会再次升起、太阳不会再次升起。第二天,当太阳再次升起的时候,在袋子里面再放入一颗白球,于是概率由初始的 1/2,上升到了 2/3,第三天,当太阳再一次升起的时候,再放入一颗白球,此时的概率(信念的程度)已经

由 2/3 上升到了 3/4。随着时间的推移，最初认为太阳每天早晨是否升起的怀疑信念，就慢慢地被修正为几乎可以断定太阳永远会再次升起。理解了贝叶斯算法，遗传算法就容易理解了。

贝叶斯在我们理解市场微观结构，比如做市商模型或者在做高频交易策略方面，可以应用于订单流策略设计。