目　录

第 1 部分

数据科学和 Microsoft Azure Machine Learning 导论

第 1 章

数据科学导论

那么，数据科学是什么？为什么它会如此受到关注？它只是另一股炒作过后就会消退的潮流吗？我们先来简单看看数据科学，它是什么，它为何重要，以及为何现在重要。本章的重点是数据科学流程及其准则和最佳实践，介绍数据科学最常用的技术和算法，还会探索集成模型（ensemble model），这是数据科学的前沿关键技术。

1.1 数据科学是什么

数据科学是从数据获取有用洞察的实践。虽然数据科学也适用于小数据，但它对于大数据而言尤其重要，因为我们现在会从组织内部和外部的很多信息源收集数 PB（petabyte）的结构化和非结构化数据。结果，我们现在是数据富有但信息贫穷。数据科学提供强大的流程和技术，让我们从这片数据海洋收集可操作的信息。数据科学融汇了多个学科，包括统计学、数学、运筹学、信号处理、语言学、数据库和存储、程序设计、机器学习和科学计算。图 1-1 给出了构成数据科学的最常见学科。虽然数

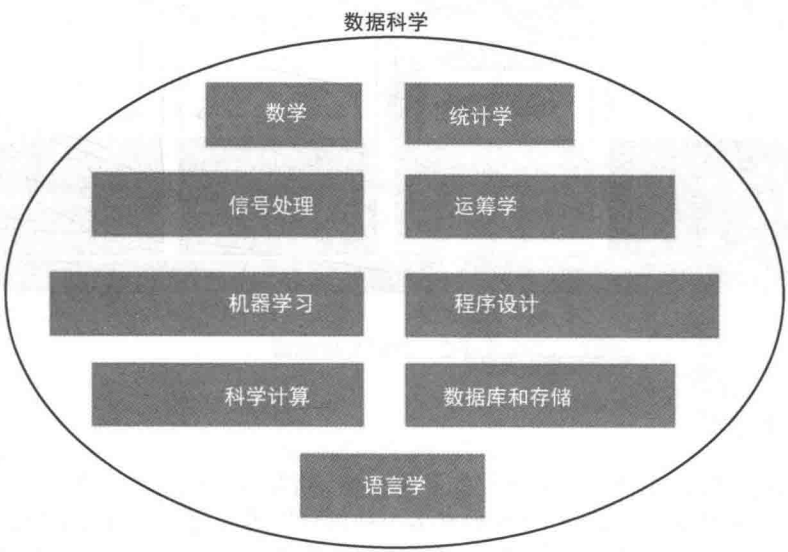

图 1-1　构成数据科学的主要学术科目

据科学这个术语在商业领域上是一个新事物，但自从 1960 年 Peter Naur 首次用它来指代计算机科学里的数据处理方法，它就已经存在了。自 20 世纪 90 年代后期以来，C.F. Jeff Wu 和 William S. Cleveland 等著名统计学家也在使用数据科学这个术语，他们把这个学科看作和统计学一样，或者是统计学的延伸。

数据科学的实践者是数据科学家，他们的技能涵盖统计学、数学、运筹学、信息处理、语言学、数据库和存储、程序设计、机器学习和科学计算。此外，为了有效地完成任务，数据科学家还要有良好的沟通技能和数据可视化技能。领域知识对于更快交付有用结果来说也是很重要的。这种技能广度很难在一个人身上找到，这就是为什么数据科学是一个团队的事，而不是一个人的事。为了有效地完成任务，聘用的团队需要拥有互补的数据科学技能。

1.2　分析频谱

根据高德纳咨询公司（Gartner）的说法，我们进行的所有分析都可以归入 4 个类别：描述性分析（descriptive）、诊断性分析（diagnostic）、预测性分析（predictive）和规定性分析（prescriptive analysis）。描述性分析通常用来描述情况，可以回答“发生了什么事”“谁是我们的客户”等问题。诊断性分析帮你理解事情为什么发生，可以回答“为什么会这样”等问题。预测性分析具有前瞻性，可以回答“将来会怎样”等问题。规定性分析，顾名思义具有规定性，可以回答“我们应该做什么”“到达目的地的最佳路径是什么”或者“我应该怎样分配投资”等问题。

图 1-2 给出完整的分析频谱。这个图表也给出了复杂程度。

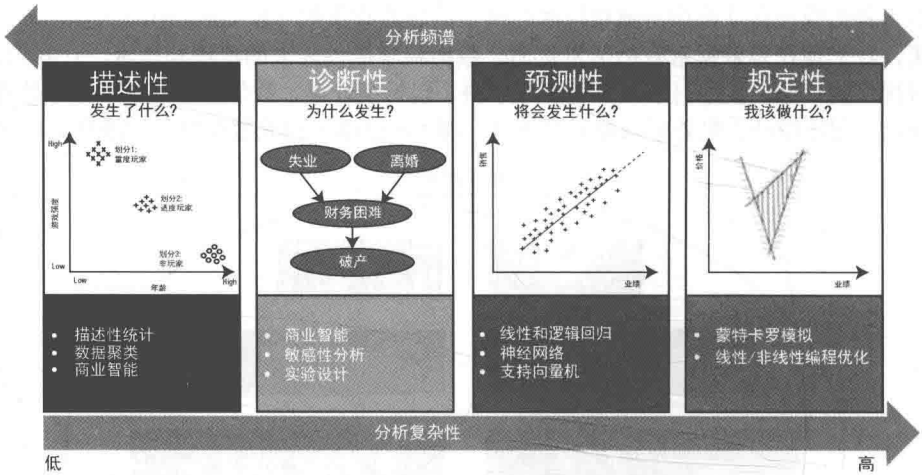

图 1-2　数据分析频谱

1.2.1　描述性分析

描述性分析用来解释在特定情况下发生了什么事。这类分析通常涉及人工干预，可以用来回答“发生了什么事”“谁是我们的客户”“我们有多少种用户”等问题。常用技术包括使用图表、柱状图、盒须图（box and whisker plot）或数据聚类的描述性统计学。本章后面将会探讨这些技术。

1.2.2 诊断性分析

诊断性分析帮你理解为什么会发生某些事以及关键动因是什么。比如说，无线服务提供商可以通过它来回答“为什么通话中断的情况增加了”或“为什么我们每个月会失去更多的客户”等问题。客户诊断性分析可以通过聚类（clustering）、分类（classification）、决策树或内容分析（content analysis）等技术来完成。这些技术可以在统计学、数据挖掘和机器学习里找到。应该注意的是，商业智能也用于诊断性分析。

1.2.3 预测性分析

预测性分析帮你预测将来会发生什么事。它可以用来预测不确定结果的可能性。比如说，它可以用来预测一个信用卡交易是不是诈骗，或者给定客户是否可能升级到高级电话套餐。统计学和机器学习为预测提供优秀技术。这些技术包括神经网络、决策树、随机森林、提升决策树（boosted decision tree）、蒙特卡罗模拟和回归。

1.2.4 规定性分析

规定性分析会通过给你推荐最佳做法来优化你的业务结果。通常，规定性分析结合了预测性模型和业务规则（比如，如果存在欺诈的可能性超过给定阈值，拒绝一个交易）。比如说，它能给特定客户推荐最佳电话套餐，或者根据优化算法给出运货车的最佳路径。规定性分析在渠道优化、投资组合优化或交通优化（根据当前交通状况找出最佳路径）等场景里非常有用。来自统计学和数据挖掘的决策树、线性和非线性编程、蒙特卡罗模拟或博弈论等技术可以用来做规定性分析。参见图 1-3。

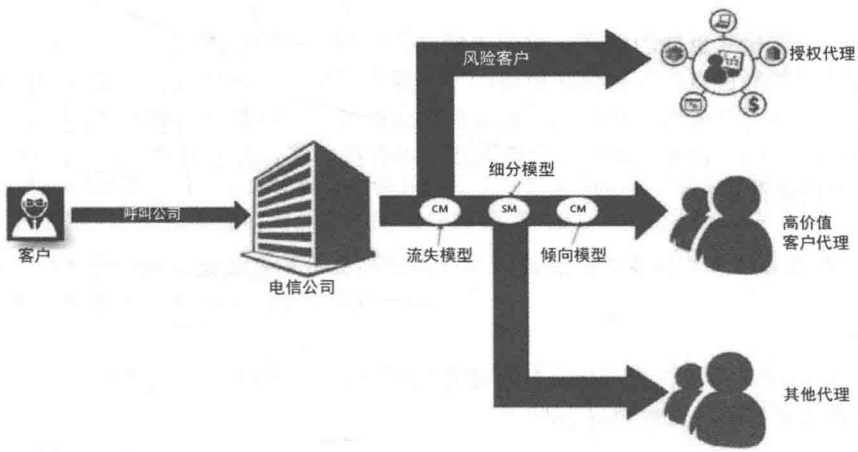

图 1-3 使用规定性分析的智能电信公司

从描述性分析到规定性分析，分析的复杂性逐渐增加。在许多方面，规定性分析都是分析的终极形态，通常用于分析复杂性要求极高的组织。假设一家智能电信公司在它的业务流程系统里嵌入了分析模型，它把以下分析模型嵌入它的客户呼叫中心系统。

- **客户流失模型**（Customer Churn Model）：这是一个预测性模型，它预测客户耗损的可能性。换句话说，它预测呼入呼叫中心的客户最终叛向竞争对手的可能性。
- **客户细分模型**（Customer Segmentation Model）：这会根据市场营销的目的把客户放入不同的行为细分类别。
- **客户倾向模型**（Customer Propensity Model）：这个模型会通过预测客户的消费倾向来响应每个市场营销方案，如升级到高级套餐。

当一个客户呼入时，呼叫中心系统从他们的手机号码实时识别出他们。然后，呼叫中心系统使用这 3 个模型来给这个客户评分。如果这个客户在客户流失模型的评分较高，就意味着他们很可能叛向竞争对手。在这种情况下，电信公司会马上把这个客户转接到有权提供诱人方案的呼叫中心客服，防止客户流失。否则，如果细分模型认为这个客户能够带来利润，就会把他 / 她转接到特殊礼宾服务，享受更短的等候时间和最好的客户服务。如果倾向模型对客户升级的评分较高，就会通知呼叫客服尝试向这个客户销售诱人的升级方案。这个解决方案的优势是所有这些模型都整合到电信公司的业务流程里，这让客服可以做出明智的决策来改善收益和客户满意度。图 1-3 说明了这点。

1.3　为何重要，为何现在

数据科学让组织有机会根据他们收集的所有数据做出更明智和及时的决策。使用正确的工具，数据科学不但从你自己的数据，还从组织外部不断增长的数据源，如天气数据，客户的人口统计数据，来自信用调查机构的消费者信用数据，来自 Facebook、Twitter 和 Instagram 等社交媒体网站的数据，给你提供新的可行的洞察。以下是数据科学现在对于商业成功而言极其重要的理由。

1.3.1　把数据看作竞争资产

数据现在是一项极其重要的资产，对于正确地把它用于决策的智能组织而言是一项竞争优势。麦肯锡和高德纳都赞同这点：麦肯锡在最近的一份报告里提到，使用数据和业务分析来做决策的公司比不这样做的公司更有成效，可以兑付更高的股东权益。同样地，高德纳推测在现代数据基础设施上投资的组织比同行优胜高达 20%。大数据让组织有机会整合各个竖井（silo）的宝贵数据，获得驱使更明智决策的新洞察。

> “使用数据和业务分析来做决策的公司比不这样做的公司更有成效，兑付更高的股东权益。”
>
> ——Brad Brown et al.，麦肯锡全球研究所，2011

> “到 2015 年，把高价值、多元化、新的信息类型和来源整合到一个连贯的信息管理基础设施的组织在财务上将比业界同行优胜超过 20%。”
>
> ——Regina Casonato et al.，高德纳

1.3.2　客户需求的增长

商业智能已经成为大多数组织在最近几十年里使用的关键分析类型。然而，随着大数据的出现，

越来越多客户现在渴望使用预测性分析来改善市场营销和业务规划。传统 BI 为他们的业务提供良好的后视分析，但在预报或预测等前瞻性问题上就帮不上忙了。

过去两年已经看到客户对预测性分析的需求激增，他们寻求更强大的分析技术来发现蕴藏在业务数据里的价值。从我们的经验来看，我们发现单单过去两年客户对数据科学的需求比以往的还多！

1.3.3 对数据挖掘技术认识的提高

一部分数据挖掘和机器学习算法如今已经广为人知，因为它们已被早期采用者试用和测试过了，如 Netflix 和 Amazon，这些公司还在它们的推荐引擎里积极使用这些算法。虽然大多数客户并不完全理解用到的机器学习算法的细节，但它们在 Netflix 在线商店的电影推荐或推荐引擎里的应用却是非常突出的。类似地，很多客户现在知道定向广告被大多数成熟的在线提供商大量使用。因此，虽然很多客户可能不知道用到的算法的细节，但他们现在逐渐了解它们的商业价值了。

1.3.4 访问更多数据

数字数据在过去几年里出现爆炸式增长，而且没有减弱的苗头。大多数行业专家现在都认为我们收集的数据比以往的多。根据 IDC 的说法，到 2020 年，全球数字世界的数据量将会增长到 35ZB。有人推测这个世界的数据现在每 5 年增长高达 10 倍，实在令人惊讶。麦肯锡也在最近一项研究里发现，在 17 个美国经济部门中的 15 个里，员工超过 1000 的公司平均保存超过 235TB 数据，这比美国国会图书馆保存的数据还要多！数据爆炸受到新数据源崛起的驱动，如社交媒体、蜂窝电话、智能传感器以及计算机行业的巨大进步。物联网（IoT）的崛起加剧了这种趋势，因为传感器产生了比以往更多的数据。根据思科的说法，到 2020 年，将有多达 500 亿个互联设备！

收集大量数据使你可以构建更精确的预测模型。我们从统计学得知，置信区间（也叫作误差范围）与样本大小成反比关系。因此，你的样本大小越大，误差范围就越小。这又会提升你的模型的预测准确性。

1.3.5 更快、更廉价的处理能力

我们现在可以支配的计算能力远超从前。摩尔定律认为计算机芯片的性能呈指数增长，每 18 个月翻倍。这股趋势对于现代计算历史的大部分时间而言都是吻合的。国际半导体技术蓝图（International Technology Roadmap for Semiconductors）在 2010 年更新了这个预测，它认为这股增长会在 2013 年减缓，届时计算机的密度和数量将是每 3 年而不是每 18 个月翻倍。尽管如此，处理器性能的指数增长促使技术和经济效益取得巨大进步。今天，智能手机的处理器比 20 年前的桌面计算机处理器强大 5 倍以上。比如说，诺基亚 Lumia 928 的双核 1.5GHz 高通骁龙 S4 比 1993 年发布的英特尔奔腾 P5 CPU 至少快 5 倍，后者对于当时的个人计算机来说是相当受欢迎的。在 20 世纪 90 年代里，DEC VAX 大型机或 DEC Alpha 工作站等昂贵的工作站用来运行高级的计算密集型的算法。值得注意的是，今天的智能手机也比 1994 年强大的 DEC Alpha 处理器快 5 倍，后者的频率是 200～300MHz！今天，你可以在价格实惠的带有多核处理器的个人工作站上运行相同的算法。此外，你可以利用 Hadoop 的 MapReduce 架构在商用服务器群上部署强大的数据挖掘算法，成本却比以往的更低。数据科学提供的工具让我们可以通过合理使用数据挖掘和机器学习算法发现隐藏在我们数据里的模式。

我们也看到内存容量的巨大进步，以及计算机内存价格呈指数下降。图 1-4 和图 1-5 说明了这点，

它显示了自 1960 年以来计算机内存的价格呈指数下降，而容量呈指数增长。自 1990 年以来，每 1MB 内存的平均价格已从 59 美元下降到微薄的 0.49 美分——99.2%的价格下降！与此同时，内存模块的容量已从 8MB 上升到庞大的 8GB！结果，现在最低端的笔记本电脑也比 20 世纪 90 年代早期的高端工作站强大。

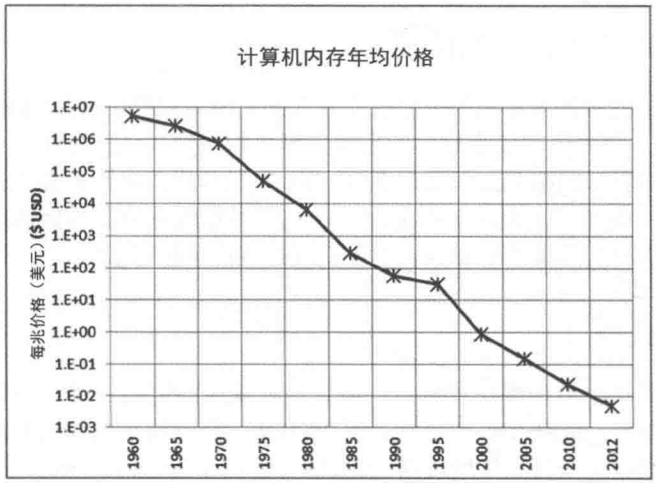

图 1-4　自 1960 年以来计算机内存的平均价格

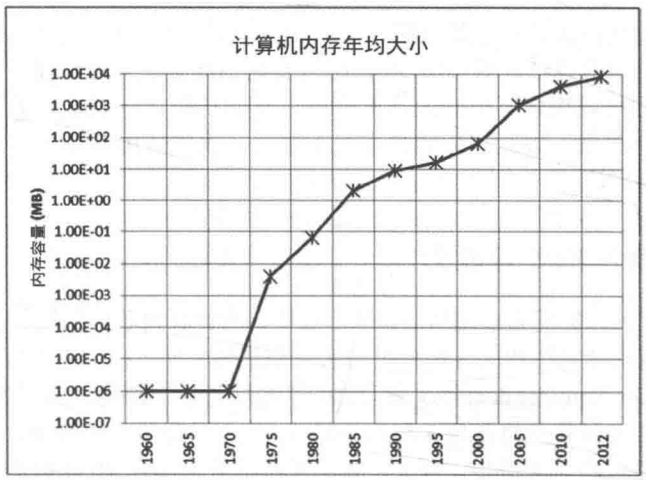

图 1-5　自 1960 年以来计算机内存的平均大小

注意： 关于内存价格历史的更多信息可以在 John C. McCallum 的网站上找到。

1.3.6　数据科学流程

一个典型的数据科学项目遵循图 1-6 显示的 5 步流程。让我们详细了解一下每个步骤。

（1）**定义业务问题：** 这很关键，因为它会引导项目的其余部分。在构建任何模型之前，和项目

发起人一起确认他 / 她尝试解决的特定业务问题是很重要的。如果不这样，就可能会耗费数周或数月构建复杂的模型来解决错误的问题，枉费工夫。好的数据科学项目可以提供良好的洞察，从而推动更明智的商业决策。因此，分析应该服务商业目标，不应该盲目使用。有一些正式的咨询技术和框架（比如，引导发现研讨会和六西格玛方法学）可供实践者用来帮助业务干系人确定他们的商业目标以及范围。

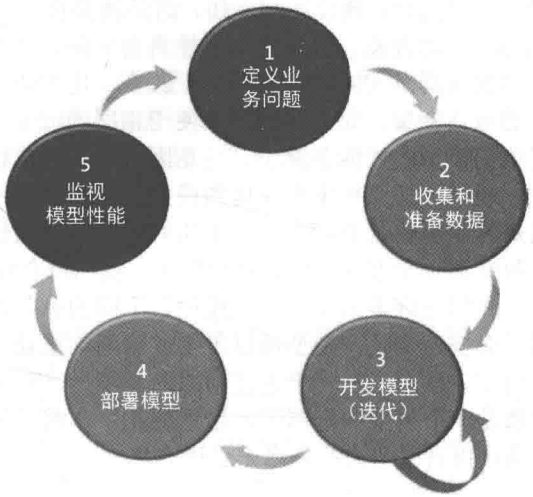

图 1-6　数据科学流程概览

（2）**获取和准备数据：**这一步牵涉两个活动。第一个是从数据库、CRM 系统、Web 日志文件等多个来源系统获取原始数据。这可能牵涉 ETL（提取、转换和加载）流程、数据库管理员和 BI 人员。然而，数据科学家也会密切参与，确保正确的数据以正确的格式提取出来。使用原始数据也提供了必要的上下文，而这又是下游需要的。

其次，一旦拿到正确的数据，就会为建模做分析和准备。这包括处理缺失值、数据里的离群值和数据转换。通常，如果一个变量有超过 40%的缺失值，就可以丢弃了，除非缺失（与否）表达了关键信息。比如说，谁在调查表里填写“年龄”可选字段可能导致人口统计数据存在很大偏差。接下来，我们需要决定如何处理缺失值。我们应该采用平均值，中间值，还是其他什么值？有一些统计技术可以用来发现离群值。在盒须图上，离群值是大于或小于 1.5 倍四分位距（interquartile range，IQR）的样本（值）。四分位距是第 75 百分位数至第 25 百分位数。我们需要决定是否丢弃一个离群值。如果保留它是有意义的，我们需要为这个变量找到一种有用的转换。比如说，对数转换通常用于转换输入。

相关性分析、主成分分析或因素分析都是展示变量之间关系的有用技术。最后，特征选择会在这个阶段完成，标识出下一步的模型使用的正确变量。

这一步可能比较辛苦和耗时。事实上，在一个典型的数据科学项目里，我们会耗费高达 75%到 80%的时间在数据获取和准备上。尽管如此，这是为后续的建模把原始数据转成优质资源的重要步骤。俗话说得好：进来的是垃圾，出去的也是垃圾。在数据准备上做明智的投资可以提高项目的成功率。第 3 章将会详述数据准备阶段。

（3）**开发模型：**这是项目最有趣的部分，我们会在这里开发预测模型。在这一步里，我们会根据业务问题和数据决定用于建模的正确算法。比如说，如果是一个二分类问题，我们可以使用逻辑回归、决策树、提升决策树或神经网络。如果最终模型需要提供解释，就会排除提升决策树等算法。模型构建是一个迭代流程：我们使用不同的模型来做实验，并找出最具预测性的一个。我们也会和客户验证数次，在结束这个阶段之前确保它能满足他们的需求。

（4）**部署模型：**一旦构建好了，最终模型需要部署到生产环境，用来评分事务，或者帮助客户推动真正的商业决策。模型可以根据客户的环境以多种不同方式部署。在大多数情况下，为了整合现有决策管理平台，部署模型需要实现数据科学家开发的数据转换和预测算法。不必多说，这在今天是一个繁琐的流程。Azure 机器学习极大地简化了模型的部署，它允许数据科学家把他们完成的模型部署成 Web 服务，供任何平台上的任何应用程序调用，包括移动设备。

（5）**监视模型的效能：**数据科学并不止于部署。值得注意的是，每个统计或机器学习模型都只

是现实世界的近似，因而从一开始就不完美。当一个验证好的模型在生产环境里测试和部署，需要对它进行监视，确保如期运作。这是极其重要的，因为任何数据驱动的模型都有一个固定的保质期。随着时间的推移，模型的准确性将会下降，因为生产环境的数据基本上都会由于各种原因（如针对不同的人群发布新的产品）发生改变。比如说，前面提到的无线运营商可能决定为青少年发布一个新的电话套餐。如果他们继续使用相同的流失和倾向模型，他们可能会在这款新的产品发布之后发现他们模型的效能下降了。这是因为原本用来构建流失和倾向模型的数据集并未包含大量青少年客户。通过密切监视生产环境的模型，我们就能及时发现它的效能下降。当它的准确性显著下降时，就是时候重建这个模型了，重建的方法是使用包含生产数据的最新数据集重新训练它，或者使用别的数据集完全重建。在这种情况下，我们将会回到第一步，我们重新审视商业目标，然后从头开始。

我们应该多久重建一次模型？不同的业务领域会有不同的频率。在一个稳定的业务环境里，数据不会快速变化，模型可以每年或每两年重建一次。抵押贷款和汽车贷款等零售银行产品就是一个好例子。然而，在一个非常动态的环境里，环境数据快速变化，模型可以每天或每周重建。手机行业就是一个典型的好例子，这里的竞争非常激烈。流失模型每隔几天就要重新训练，因为竞争对手会通过更有吸引力的优惠来引诱客户。

1.4　常见数据科学技术

数据科学的组成学科提供了大量算法，这些学科包括统计学、数学、运筹学、信号处理、语言学、数据库和存储、程序设计、机器学习和科学计算。为简单起见，我们把这些算法归入以下分组。

- 分类（Classification）。
- 聚类（Clustering）。
- 回归（Regression）。
- 模拟（Simulation）。
- 内容分析（Content Analysis）。
- 推荐器（Recommender）。

第 6 章会详述其中一些算法。

1.4.1　分类算法

分类算法通常用来把人或事物归入多个分组中的一个。它们也广泛用于预测。比如说，为了防止诈骗，发卡机构会把信用卡交易归类为诈骗或非诈骗。发卡机构通常拥有大量信用卡交易历史，知道这些交易的状态。这些案件很多都是合法的持卡人报告的，他们不想为未经授权的交易还款。因此，发卡机构知道每个交易是否诈骗。使用这些历史数据，发卡机构就能构建模型，预测新的信用卡交易有没有可能是诈骗。这是一个二元分类问题，所有情况都会落入两个类别中的一个。

另一个分类问题是客户升级到高级电话套餐的倾向。对于这种情况，无线运营商需要知道客户会否升级到高级套餐。借助销售和使用数据，运营商可以判断哪些用户曾经升级过。因此，他们可以把所有客户归入两个分组中的一个：他们有否升级过。因为运营商也有新的和现有客户的人口统计信息和行为数据，所以他们可以构建模型来预测新的客户升级的可能性，换句话说，这个模型会把每个客户分到两个类别中的一个。

统计学和数据挖掘为分类提供很多很棒的工具，包括逻辑回归，它被统计学家广泛用于构建信用评分卡，或购物倾向模型；神经网络算法，如反向传播算法（backpropagation）、径向基函数（radial basis function），或纹脉多项式网络（ridge polynomial network）。其他的还有决策树或集成模型，如提升决策树或随机森林。对于拥有超过两个类别的更复杂的分类问题，你可以使用多峰技术（multimodal technique）来预测多个类别。分类问题通常使用监督学习算法，它们使用标签数据来训练。Azure 机器学习为分类提供多个算法，包括逻辑回归、决策树、提升决策树、多峰神经网络等。详情参见第 6 章。

1.4.2 聚类算法

聚类使用无监督学习把数据分到不同的类别。聚类问题和分类问题之间的主要区别是聚类的输出无法提前知道。在执行聚类之前，我们不知道每个数据点归属哪个类别。相反，对于分类问题，我们的历史数据显示了每个数据点归属哪个类别。比如说，贷方从历史数据知道客户是否拖欠车贷。

聚类的一个很好的应用场景是客户细分，我们处于营销目的把客户分到不同的细分类别。在一个好的细分模型里，每个细分类别里的数据都是非常相似的。但是，不同的细分类别的数据非常不同。比如说，游戏行业的营销人员需要更好地理解他/她的客户才能为他们创建合适的产品。假设他/她只有两个关于客户的变量：年龄和游戏强度。借助聚类，营销人员发现游戏客户的 3 个不同细分类别，如图 1-7 所示。第 1 个细分类别是重度玩家，他们每天都会很投入地玩电脑游戏，通常都是年轻人。第 2 个细分类别是业余玩家，他们只会偶尔玩一下，通常都是三四十岁。非玩家几乎从不玩电脑游戏，年龄通常更大；他们属于第 3 个细分类别。

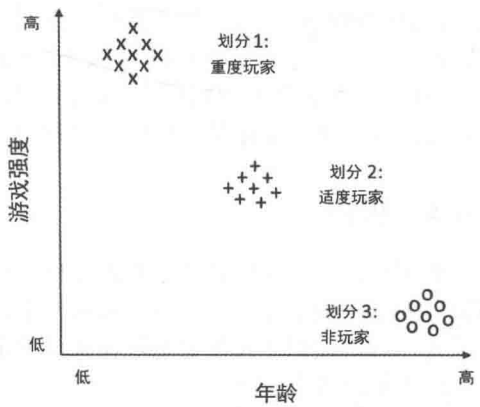

图 1-7 从聚类算法简单推出的客户细分类别

统计学为聚类提供了几个工具，但用途最为广泛的是 K 均值算法，它使用距离度量把相似的数据聚到一起。要使用这个算法，你得先决定你想要多少个类别；这就是常数 K。如果你把 K 设为 3，这个算法会产生 3 个类别。关于 K 均值算法的详细内容，请参考 Haralambos Marmanis 和 Dmitry Babenko 的书。机器学习也提供了更成熟的算法，如 Teuvo Kohonen 开发的自组织映射（也叫作科霍宁网络），或者 Stephen Grossberg 和 Gail Carpenter 开发的自适应共振理论（ART）网络。聚类算法通常使用无监督学习，因为输出在训练的过程中无法得知。

注意： 你可以在以下书籍和论文里读到更多聚类算法的内容。

Haralambos Marmanis 和 dmitry Babenko，《Algorithms of the Intelligent Web》（Stamford，Ct: Manning publications Co.，January 2011）

T. Kohonen，《Self-Organizing Maps. Third, extended edition》（Springer，2001）

《Art2-A: an adaptive resonance algorithm for rapid category learning and recognition》，Gail A. Carpenter、Stephen Grossberg 和 David B. Rosen，《Neural Networks》，4:493-504. 1991a

---

1.4.3 回归算法

回归技术用来预测带有数字输出的响应变量。比如说，无线运营商可以使用回归技术来预测他们的客服中心的呼叫量。借助这个信息，他们可以安排恰当数量的呼叫中心员工来满足需求。回归模型的输入变量可能是数字或者类别。但是，这些算法的共同点是，输出（或者响应变量）通常都是数字的。最常用的回归技术包括线性回归、决策树、神经网络和提升决策树回归。

线性回归是统计学里最老的预测技术之一，它的目标是从一组观测变量预测后果。简单线性回归模型是一个线性函数。如果只有一个输入变量，线性回归模型就是贴合数据的最优线。对于两个或以上的变量，回归模型就是贴合底层数据的最优超平面。

人工神经网络是一组模拟大脑功能的算法。它们通过实例学习，经过训练可以从数据集做出预测，即使把响应映射到独立变量的函数是未知的。有很多不同的神经网络算法，比如反向传播网络（backpropagation network）、霍普菲尔（Hopfield）网络、科霍宁（Kohonen）网络（又称作自组织映射）和自适应共振理论（ART）网络。然而，最常见的是反向传播，又称作多层感知器（multilayered perceptron）。神经网络用于回归或分类。

决策树算法属于分层技术，其工作原理是根据某个统计条件以迭代的方式分割数据集。决策树的目标是使树里不同节点的方差最大化，使每个节点内部的方差最小化。常用的决策树算法有 ID3（迭代二叉树 3 代算法）、C4.5 和 C5.0（ID3 的接替）、自动交互检测（AID）、卡方自动交互检测（CHAID）以及分类和回归树（CART）。虽然 ID3、C4.5、C5.0 和 CHAID 算法很有用，但它们是分类算法，对于回归没有用。但是，CART 算法既可以用于分类，也可以用于回归。

1.4.4 模拟

模拟广泛用于不同行业，对现实世界里的流程进行建模和优化。工程师们长期使用有限元法或有限体积法等数学技术来模拟机翼或汽车的空气动力学。模拟帮助工程公司在研发方面节省了数百万美元，因为他们不再需要通过真实模型来执行所有测试了。此外，模拟还可以通过调整计算机模型里的变量测试更多场景。

在商业领域里，模拟可以用来对流程进行建模，比如说，帮助呼叫中心优化等待时间，或者帮助运输公司或航空公司优化路线。通过模拟，业务分析师可以对大量假设进行建模，为利润或其他商业目标而优化。

统计学为模拟和优化提供很多强大的技术。马尔可夫链分析法可以用来模拟动态系统里的状态变化。比如说，它可以用来对客户如何流经呼叫中心进行建模：客户在挂掉电话之前会等多久，使用交互式语音响应（IVR）系统之后他们等待的几率有多大？线性编程用来优化运输或航班路线，而蒙特卡罗模拟则用来找出优化特定商业成果（如利润）的最佳条件。

1.4.5 内容分析

内容分析用于从文本文件、图像和视频挖掘出洞察。文本挖掘使用统计和语言分析来理解文本含义。简单的关键字搜索对于大多数实用应用程序来说太原始了。比如说，通过简单的关键字搜索来理解 Twitter 数据的情绪是一项人工体力活，因为你得保存正面、中性和负面情绪的关键字。接着，在你扫描 Twitter 数据的时候，根据检测到的特定关键字给每条 Twitter 数据评分。这种方案很累赘，

也非常原始，虽然对于简单的场景很有用。这个流程可以通过文本挖掘和自然语言处理（NLP）自动化，根据上下文而不是简单的关键字搜索来挖掘文本和尝试推断词语的含义。

机器学习也提供了一些工具，可以通过模式识别来分析图像和视频。通过模式识别，我们可以使用面部识别算法来找出已知目标。多层感知器和 ART 网络等神经网络算法可以用来检测和跟踪视频流里的已知目标，或者协助分析 X 光图像。

1.4.6 推荐引擎

推荐引擎广泛用于 Amazon 等在线零售商，根据用户偏好推荐产品。推荐引擎有三大方案。协同过滤（CF）根据用户或项之间的相似性来推荐。对于项的协同过滤，我们分析项的数据，找出哪些项是相似的。协同过滤的数据特指用户的交互操作，比如说，评价或者查看电影，而不是电影的流派、导演和演员等特征。因此，每当客户购买一部电影，我们会根据相似性推荐其他电影。

第二种推荐引擎通过分析每个用户选择的内容来推荐。在这种情况下，文本挖掘或自然语言处理技术将会用来分析内容（如文档文件）。类似的内容将会放在一起，这是给新用户推荐的基础。若想深入了解协同过滤和基于内容的方案，可以阅读 Haralambos Marmanis 和 Dmitry Babenko 的书。

推荐引擎的第三种方案使用机器学习算法来判断产品的相关性。这种方案又称作购物篮分析。朴素贝叶斯方法（Naïve Bayes）、Microsoft Association Rules 或 R 语言里的 Arules 包等算法可以用来挖掘销售数据，决定哪些产品一起销售。

1.5 数据科学的前沿

最后，我们简单介绍一下集成模型，这是数据科学的前沿。

集成模型的崛起

集成模型是一组机器学习分类器，使用一组而不是单个算法来解决分类问题。它们模仿我们人类通过咨询博学的朋友或专家来提高决策精确的做法。在应对医疗诊断等重要决策时，我们倾向于找其他医生做二次诊断，从而增强我们的信心。同样地，集成模型把一组算法当作一组专家，提高精度，降低分类问题的方差。

数十年来，机器学习社区致力于集成模型。事实上，开创性的论文最早由 Dasarathy 和 Sheela 于 1979 年发表。但是，直到 20 世纪 90 年代中期，这个领域才因为一些重要的贡献而得到快速的发展，出现了一些非常成功的真实应用。

1. 集成模型的真实应用

在过去的几年里，集成模型出现在一些很棒的真实应用里，如摄像头的面部识别、生物信息学、Netflix 电影推荐和 Microsoft 的 Xbox Kinect。我们来看看其中的两个。

首先，集成模型对于 Netflix Prize 竞争的胜出非常有用。2006 年，Netflix 举办了一个公开赛，能为现有解决方案带来 10%提升的最佳协同过滤算法将会获得 100 万美元的奖金。2009 年 9 月，这 100 万美元的奖金授予了 BellKor's Pragmatic Chaos，这个团队由 AT&T 实验室的科学家和两个不太

知名的团队组成。比赛开始的时候，大多数团队都是用单个分类器算法：虽然它们的性能比 Netflix 的模型高 6%～8%，但性能很快就停止提升，直到各队开始应用集成模型。领先的选手很快意识到，他们可以把他们的算法和那些明显落后的团队的算法整合起来改善他们的模型。最终，大多数顶尖的团队，包括获胜者，都使用集成模型，使性能远超 Netflix 的推荐引擎。比如说，第二名的团队，名字恰好是 The Ensemble，在他们的集成模型里使用了超过 900 个独立模型。

Microsoft 的 Xbox Kinect 传感器也使用集成模型。随机森林，集成模型的一种形式，可以在用户使用 Xbox Kinect 传感器玩游戏时有效地跟踪骨骼运动。

尽管在真实应用里取得成功，集成模型的一个主要局限在于它们是黑盒，它们的决定很难解释。因此，它们不适合需要解释这些决定的应用。信用评分卡就是一个很好的例子，因为贷方需要解释他们赋予每个消费者的信用评分。在某些市场里，这样的解释是一个法定条件，因此，集成模型是不适合的，尽管它们的预测能力很强。

2. 构建集成模型

构建集成模型有 3 个主要步骤：第一，选择数据；第二，训练分类器；第三，整合分类器。

构建集成模型的第一步是为分类器模型选择数据。在采样数据时，一个主要的目标是使模型的分歧最大化，因为这会提高解决方案的精度。一般来说，模型的分歧越大，你最终的分类器的效能就越好，它的预测方差就越小。

这个流程的第二步是训练多个独立的分类器。但你怎么分配这些分类器呢？在现有的许多策略中，最流行的两个是装袋（bagging）和提升（boosting）。装袋算法使用数据的不同子集来训练每个模型。随机森林算法使用这种装袋方案。相反，提升算法通过在训练期间强化训练集的错误分类实例的重要性来提高效能。因此，在训练期间，每个额外的模型都集中在错误分类的数据上。提升决策树算法使用这种提升策略。

最后，一旦你训练完所有的分类器，最后一步就是整合它们的结果，做出最终预测。整合结果有多个方案，从简单多数到加权多数投票都有。

集成模型是机器学习真正令人激动的部分，它们拥有突破分类问题的潜能。

1.6 小结

本章介绍了数据科学，回答了它是什么，为何重要，以及为何现在重要。我们概述了数据科学的关键学科，包括统计学、数学、运筹学、信号处理、语言学、数据库和存储、程序设计，以及机器学习。我们谈到数据科学受关注程度提高背后的主要原因：日益增加的数据量、数据作为竞争资产、日渐为人知晓的数据挖掘，以及硬件成本下降。

一个简单的 5 步数据科学流程以及正确应用这个流程的指南也在这里介绍了。我们还介绍了几个数据科学里最常用的技术。最后，我们介绍了集成模型，它是数据科学前沿的关键技术之一。