第5章 机器学习

本章学习目标：

- 了解机器学习的基本概念、机器学习系统的基本结构和机器学习的发展历史。
- 理解掌握线性模型方法和三种聚类方法。
- 操作实践：能够应用线性模型方法和三种聚类方法分析和解决实际问题。

人工智能是研究如何在机器上实现人类智能，学习是人类最为重要的智能行为。“如何使机器具有学习能力”就成为人工智能最为重要的研究课题，这便是机器学习。本章简要介绍机器学习，主要包括机器学习定义和发展历史、机器学习的分类、机器学习系统的基本结构和几种常用的机器学习方法。

5.1 机器学习的定义和发展历史

机器学习目前已经成为一门课程、一个研究领域或一个学科的代名词。本节简要探讨机器学习的定义、机器学习的发展历史。

5.1.1 机器学习的定义

1. 什么是机器学习？

至今还没有一个关于“机器学习”的统一公认的定义，许多学者从不同的角度给出了不同的定义。机器学习领域奠基人之一的汤姆·米切尔(Tom Mitchell)教授给出的机器学习的经典定义为“利用经验来改善计算机系统自身的性能”。具体定义为“如果一个计算机程序针对某类任务 T，用 P 衡量性能，根据经验 E 来自我完善，那么称这个计算机程序在从经验 E 中学习，针对某类任务 T，它的性能用 P 来衡量”。汤姆·米切尔撰写的《机器学习》至今仍然被许多学习者视为圭臬。赫伯特·A.西蒙(Herbert A.Simon)给出的定义为“如果一个系统能够通过执行某个过程改进它的性能，这就是学习”。蔡自兴教授给出的定义是“机器学习是一门研究如何使用机器来模拟人类学习活动的一门学科，即机器学习是一门研究机器获取新知识和新技能，并识别现有知识的学问”。

我们不去纠结定义本身，而是按照李德毅院士给出的“简单地按照字面理解，机器学习的目的是让机器能像人一样具有学习能力”。机器学习作为一个学科(或称为研究领域)，那便是研究如何让机器具有学习能力的一门学科(或称为研究领域)。研究使机器具有学习能力，至少我们要知道“什么是学习能力，它有什么表现”。

2. 什么是学习能力？

百度上说，学习能力是学生成功地完成学习目的所必需的个性心理特征。按学习能力的倾向可分为一般学习能力和特殊学习能力。一般学习能力，是指反映在学生学习活动过程中的一般能力，主要包括以下基本要素：观察力、注意力、记忆力、思维能力、想象力、语言表达能力、创造力、感觉统合能力、理解力、运算能力等，适合于广泛实践活动要求的能力。特殊学习能力也称为专门学习能力，指适合某种专业活动要求的能力，如音乐能力、绘画能力、体育能力，等等。

显然，机器学习的目标就是使机器具有这些能力。对于人来说，具有一般学习能力是很普遍的，但具有特殊学习能力的人真的是很特殊的、不普遍的。对于机器来说，从当前的研究成果看，具有特殊学习能力机器的研究成果较为丰富，例如国际象棋程序、围棋程序AlphaGo 等，都在某些专门的领域超越了人类；但对于使机器具有一般学习能力的研究成果，却不那么丰富，例如观察力、注意力、想象力等，这也是需要进一步深入研究的课题。

3. 机器学习与人工智能有什么关系？

一般认为，人工智能是研究在机器上实现人类的智能，应包括感知能力、学习能力、表达能力等。机器学习是使机器具有学习能力，机器学习是人工智能的一个重要研究分支方向，也可以说，机器学习是人工智能的一个重要研究领域。人工智能领域还有一些相关学科，其中包括数据挖掘、神经计算、模式识别等，这些学科相互交叉，形成了“你中有我，我中有你”的局面。例如，数据挖掘是从大量数据中发掘有趣的模式和知识的过程，也称作从数据中挖掘知识。机器学习是使机器具有学习能力，学习的目的也是拥有知识。因此，这些都是相近的、交叉的研究领域。

5.1.2 机器学习的发展历史

机器学习是人工智能应用研究中最重要的分支之一。机器学习的发展与人工智能的发展相辅相成、相互促进。以下简单回顾机器学习中的重大事件和重要节点。

第一阶段：20 世纪 50 年代至 70 年代末。机器学习从逻辑推理、定理证明到专家系统，再到神经网络、结构学习系统等都被相继提出，主要事件如下。

（1）1952 年，“逻辑理论家”程序证明了《数学原理》中的 38 条定理，1963 年证明了 52 条定理，并且纽厄尔（A.Newell）和西蒙因为这方面的工作获得了 1975 年的图灵奖。1958 年，华人数理逻辑学家王浩在 IBM-704 计算机上用 3～5 分钟证明了《数学原理》中有关命题演算的全部定理（220 条），并且还证明了谓词演算中 150 条定理的 85%。

（2）1956 年，IBM 公司的亚瑟·塞缪尔研制出了著名的西洋跳棋程序，该程序是使用判别函数法的典型例子。这个程序能从棋谱中学习，也能从下棋实践中提高棋艺。1959 年，它击败了塞缪尔本人。塞缪尔在 1956 年达特茅斯的人工智能研讨会上给出了一个新词——机器学习。

（3）这个时期，罗森布拉特（F.Rosenblatt）的感知机奠定了基于神经网络的“联结主义”的基础，基于神经网络的“联结主义”学习开始出现，基于逻辑表示的“符号主义”学习技术蓬勃发展，包括温斯顿（Winston）的“结构学习系统”“概念学习系统”和海斯·罗思（Hayes Roth）等的“归纳学习系统”等。

（4）20 世纪 70 年代末，中国科学院自动化研究所进行质谱分析和模式文法推断研究，

表明我国机器学习研究得以推进。1980 年，西蒙来华传播机器学习的火种后，我国机器学习研究出现了新局面。

第二阶段：20 世纪 80 年代初至 90 年代末。这个时期机器学习的发展与人工智能的发展都经历了起起落落，发生的主要事件如下。

(1) 1980 年，第一届机器学习国际研讨会在卡内基—梅隆大学(CMU)召开，标志着机器学习研究已在全世界兴起。1986 年，国际杂志《机器学习》创刊，迎来了机器学习蓬勃发展的新时期。

(2) 1996 年，IBM 公司的计算机深蓝与人类国际象棋世界冠军加里·卡斯帕罗夫对战，但是深蓝没有胜利。1997 年，IBM 公司升级了深蓝，运算速度达到 2^8 次/s，能够预测未来 8 步以上的棋局，战胜了加里·卡斯帕罗夫。事实上，深蓝并不具有学习能力，它是依靠计算速度和枚举在规则明确的游戏中取得了胜利。这次胜利成为人工智能和机器学习领域具有里程碑意义的对战。

(3) 这个时期，基于“联结主义”的研究成果丰富。神经网络的研究重新兴起，连接机制学习方法的研究方兴未艾，反向传播算法研究有了新进展；基于生物发育进化论的进化学习系统和遗传算法吸取了归纳学习和连接机制学习的优势而更加完善。还有，机器学习的一个重要研究课题“数据挖掘”的研究蓬勃发展，许多学习算法相继提出，例如决策树中的 ID3 算法、C4.5 算法，关联规则中的 Apriori 算法等。概念学习也从学习单个概念扩展到学习多个概念，示例归纳学习系统产生并快速发展，出现了第一个专家学习系统。再有，以支持向量机(support vector machine，SVM)为代表性成果的“统计学习”也成为机器学习的基本研究内容之一。

(4) 机器学习的应用广泛推广，其不仅应用在基于知识的各种应用系统中，也应用在模式识别、自然语言理解、机器视觉等许多领域。一个系统是否具有学习能力已经成为其是否具有“智能”的一个标志。

第三阶段：进入 21 世纪以来。机器学习蓬勃发展，一批重要学术成果相继出现，机器学习进入黄金发展期。各种机器学习方法如雨后春笋般蓬勃发展，并且获得许多成功的应用，包括遗传算法、支持向量机、协同过滤算法，等等，其中影响力最大的莫过于“深度学习”。

“深度学习”狂潮席卷机器学习和人工智能研究领域。2006 年，杰弗里·辛顿(Geoffrey Hinton)在《科学》杂志发表了论文 *Reducing the Dimensionality of Data with Neural Networks*，开启了深度学习浪潮。至今已经有多种深度学习框架，如深度神经网络、对抗生成网络、卷积神经网络等。这些深度学习框架在计算机视觉、自然语言处理、语音识别、自动驾驶等领域得到广泛应用，并取得了很好的效果。深度学习的内容将在第 7 章详细介绍。

机器学习已经发展成为一个学科领域，它是一个多学科交叉的研究领域，包括统计学、计算机科学、生物学、神经学等学科。计算机科学的分支学科领域中都有机器学习的身影，例如图形学、软件工程、多媒体等。同时，它还为许多交叉学科提供了重要的技术支撑，机器学习已成为最重要的技术进步源泉之一。

5.2 机器学习的分类

事物一般都有许多种分类方式，从不同的角度有不同的分类方法。例如，对人的分类，可以按照性别分类，可以按照所在地区分类，也可以按照职业分类，等等。机器学习方法的

分类亦是如此，可以按照学习目标进行分类，分为概念学习、规则学习、函数学习、类别学习等；可以按照数据形式进行分类，分为结构化学习、非结构化学习和半结构化学习等；也可以按照训练方法进行分类，分为监督学习、无监督学习和强化学习。以下主要介绍根据训练方法的分类，即监督学习、无监督学习和强化学习。

1. 监督学习

从一般意义上理解，监督学习就是有老师在一旁监督的学习，无监督学习就是没有老师监督的学习。在机器学习中，早期监督学习一般指有人工干预的学习，但现在多指学习对象带有指定标签的学习。例如，让机器来学习水果的识别方法时，用来作为学习对象的水果都带有水果的名称（即带着苹果、香蕉、西瓜、橘子等名称标识），这些名称就称为标签，分类对象中既有特征表示，又有类别标识，机器利用这些特征和标签建立一个模型，这个模型具有分类的作用，也称为分类器。这个分类器就是机器学习的目标，这个机器再见到一个新的水果时，就会根据这个分类器对水果进行分类，来识别这个新水果究竟是什么水果。监督学习是指学习时使用带有标签的对象或数据，生成的是能够执行相应任务的函数、规则或模型。机器学习中的绝大多数学习算法都是监督学习算法。例如我们常用的分类方法都是监督学习方法。

监督学习过程就如同我们教小孩学习认识水果的过程，我们拿出一个苹果给小孩看（相当于输入），然后告诉他这是一个苹果（相当于输出），完成一次训练；再拿出一个香蕉给他看（相当于输入），然后告诉他这是一个香蕉（相当于输出），又完成一次训练……如此继续下去，各种水果反复出现进行训练。小孩就会对水果有相应的认识，进而形成分类方法。这样的学习过程就是监督学习。

监督学习就是在已知输入和输出的情况下训练出一个模型，将输入映射到输出。监督学习的学习过程就是建立模型的过程，根据给出的具有输入输出的数据训练出模型。

2. 无监督学习

无监督学习与监督学习相对应，无监督学习就是不受监督的学习。无监督学习是指学习时使用不带有标签的对象或数据的学习。也就是说，无监督学习不需要人类进行数据标注，而是通过模型不断地自我认知、自我巩固，最后进行自我归纳来实现其学习过程。

无监督学习过程就如同有一堆包含不同品种的水果，水果上面没有标签，我们不知道是什么水果，但是可以根据水果的特点和性质，将具有相同或相近特点的水果聚在一起，进而认识这些水果。这就是无监督学习。

当前，无监督学习的应用远远没有监督学习应用广泛，但无监督学习具有很多明显的优势。首先，无监督学习面对的是学习对象本身，没有人为的主观因素，学习的结果更加客观。其次，无监督学习不需要标注数据，也就省去了数据标注的工作。要知道，监督学习用到的标注数据是大量的，甚至是巨大量的，甚至达到百万、亿级别的数据，标注这些数据需要大量的人力、物力和财力，也消耗大量的时间成本。

聚类方法就是一种无监督学习方法，5.4.2 节中将详细介绍。

3. 强化学习

强化学习的核心思想是模仿有机生命体对环境进行探索和与之交互时，做出正确行为时会得到奖励，做出错误行为时会得到处罚，进而不断强化对正确行为的选择，在执行任务时制定出最优的决策。强化学习是人工智能研究中行为主义流派的典型学习方法。

强化学习既不同于监督学习，又不同于无监督学习，它不需要数据的标签，而是需要相应的奖惩策略，并用它来决定学习的方向。强化学习就是在学习过程中不断地尝试，错了就扣分，对了就奖励，进而得到在环境中的最好决策。强化学习的目标就是研究在与环境交互的过程中，如何学习一种行为策略，以得到最大化的积累奖赏。

强化学习过程类似人类训练狗时狗的学习过程。狗没有先验经验，它只是在与人交互的奖惩中积累经验，进而做出相应的行为决策。人给狗一个指令，狗做对了，就给予狗一个奖赏（例如一块肉、一个鸡腿等），做错了就不给吃的或者揍它。在长期交互中，狗就会在面对各种环境时形成一个行为策略，以获得最大的奖赏。

下面对监督学习、无监督学习和强化学习作一个简单总结。监督学习主要针对有标签的数据，一般完成回归或分类等任务。无监督学习主要针对无标签的数据，一般完成聚类和降维等任务。与监督学习和无监督学习相比，强化学习最大的区别就是不要求预先给定任何数据，而是通过接收环境对动作的奖惩获得信息，并更新相关的行为策略。下面给出监督学习、无监督学习和强化学习在学习依据、数据来源和学习目标三方面的区别，如表 5-1 所示。

表 5-1　监督学习、无监督学习和强化学习的区别

类　别	学习依据	数据来源	学习目标
监督学习	基于监督（标签）信息	给定的带标签的数据	输入到输出的映射
无监督学习	基于对数据结构的假设	给定的数据	数据的分布模式
强化学习	基于奖励评估	在交互中产生的数据	选择获取最大收益状态到行为的映射

5.3　机器学习系统的基本结构

讨论机器学习系统的基本结构时，我们必须坚持“问题导向”和“系统观念”。那么怎么思考这个问题？机器学习是要研究使机器具有学习能力，那么人类是如何学习的？

首先，“孟母三迁”是为给孩子一个好的学习环境，所以环境是学习的一个重要因素。学习环境有老师教授的环境（例如学校等），也有没有老师的环境（例如野外探险等），但是有一点是明确的：环境给人提供了信息，人是在这些信息的基础上开展学习的。其次，学习什么？有老师教时，学习知识或模型形成认知或判断的能力；没有老师教时，学习就是应用从环境中得到的信息，采用总结、归纳、推理等方法获得规律性的知识、模型或行为方式策略。这个过程依据学习方法的不同，可能需要多次反复的学习或训练。学习之后怎么用？有时学过的知识、模型或行为策略一直在用，不作任何修改；有时学过的知识、模型或行为策略使用之后，有偏差，根据情况调整或修改所用的知识、模型或行为策略；也有边学边用边修改的时候。这里有几个关键因素：环境、知识或模型的形成方法和表示方式、知识的使用和完善方式。

我们从人类社会的视角去理解学习过程，有老师教的叫作“学习”，老师会告诉你对错；没有老师教的叫作“研究”，没有老师告诉你对错，但有相关的性能指标，你可以根据这些性能指标自己去判断是否合适；边做边学的叫作“技术”，在实践中积累经验。有老师教的前提是已经有了知识，例如教孩子认识苹果前老师已经有了苹果的知识（标签数据），教孩子的是

已有的知识（苹果的分类），就是我们常说的“学习”，在机器学习中是监督学习；没有老师教时，面对的是陌生的对象（不带标签的数据），通过相关的方法去认识这些陌生的对象，例如聚类方法，我们不知道一些对象是怎样分类的，但我们可以通过它们的性质、特点将它们聚类，之后将相同或相似的对象聚在一起，形成一个新的类，给这个新类起个名字，则可以粗浅地认为发现了一个类，定义了一个新的名称，即研究了这些陌生的对象，在机器学习中是无监督学习。

当然，对于机器来说，不管是有老师教的、没有老师教的、边做边学的都叫作机器学习。

根据以上讨论，机器学习系统的基本结构应该主要包括环境和机器（有的教材也称为智能体），环境给机器提供信息，机器从信息中获取知识、模型或行为策略，知识、模型或行为策略形成机器的指令及行为，这些指令或行为反作用于环境，被机器作用的环境产生新的信息，再反馈给机器，用于训练或改善知识、模型或行为策略，循环往复形成了机器学习系统的过程，如图 5-1 所示。

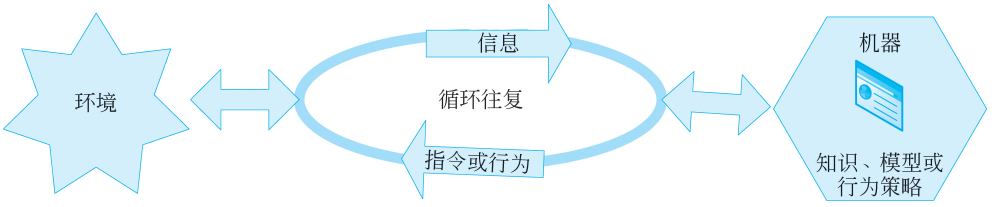

图 5-1　机器学习系统的基本结构示意图

理想的具有学习能力的机器应该是：无论什么样的环境，机器都能根据所处的环境选择相应的学习方法，学会相关的知识、模型或行为策略，进而成为机器对环境的认知、判断和行为能力。当机器面对不同的环境时，会自主选用不同的学习方法去学习和适应环境。

当前，机器学习系统一般只是针对某个问题或某个领域设计的，系统中的学习方法一般只涉及少数几种方法，目标只针对解决的问题。此时，机器学习系统的结构一般是由环境、学习方法、知识或模型构成，其中环境提供相关的信息，学习方法是从信息中获得“知识、模型或行为策略”，同时学习方法还应具有修改和完善“知识、模型或行为策略”的能力。

5.4　机器学习方法应用举例

当前的机器学习一般是用指定的方法或模型去完成学习任务，例如线性回归学习方法是指定了所使用的方法是线性方法，也就是，假定了研究对象具有某种线性关系，学习只是通过相关的数据确定线性模型（方程），进而使用这个方程去预测未知事物。再如，为了解决某个复杂问题要应用深度学习方法，学习之前就已经确定了要使用这个方法，学习的目标只是将模型确定下来，并应用于实际。具体的过程可以描述为：首先要搭建一个深度学习框架，使用已有的训练数据去训练这个框架，应用相关的性能指标确定其中的各种模型参数，进而确定模型，完成学习任务（详细内容见第 7、8 章）。本节主要介绍两种机器学习的方法，一种是监督学习之线性模型方法，另一种是无监督学习之聚类方法。

在监督学习中，学习任务主要是根据环境或实际问题选择相应的学习方法，建立任务模型。在学习过程中，经常会把数据集拆分为训练集和测试集。训练集用来训练模型，调整模

型参数；测试集用来验证模型的性能，具体的性能指标是由相关的问题决定的，例如分类问题的性能指标是准确率、错误率等，线性回归的性能指标是残差平方和、R^2 等。

5.4.1 线性模型方法

在现实生活中，往往需要分析若干变量之间的关系，例如碳排放量与气候变暖之间的关系、蛋糕大小与蛋糕价格的关系等，这种分析不同变量之间存在关系的研究叫作回归分析，刻画不同变量之间关系的模型被称为回归模型。如果这个模型是线性的，则称为线性回归模型，也简称为线性回归(linear regression，LR)。在机器学习算法中，线性回归模型简单，是最基础的机器学习模型。我们首先从最简单的一元线性回归开始，再介绍多元线性回归。线性回归是一种通过拟合自变量与因变量之间最佳线性关系来预测目标变量的方法。回归过程也称为用自变量来解释因变量的变化。

1. 一元线性回归

(1) 问题的提出。

假如家人的生日快到了，我们想买个生日蛋糕庆贺，买多大的，什么价位的？我们只知道一些尺寸对应的蛋糕价格，如表 5-2 所示，不知道所有尺寸的蛋糕价格。根据具体情况，我们要购买 14 英寸蛋糕，但不知道 14 英寸蛋糕的价格。怎么办呢？我们只好用已经了解的情况，预测一下我们需要的蛋糕价格，好做相应的安排。

表 5-2 蛋糕店的“生日蛋糕”价目表

尺寸/英寸	6	8	10	12	16	18
价格/元	38	48	68	98	168	189

用二维坐标图表示蛋糕尺寸与价格的对应关系，如图 5-2 所示。

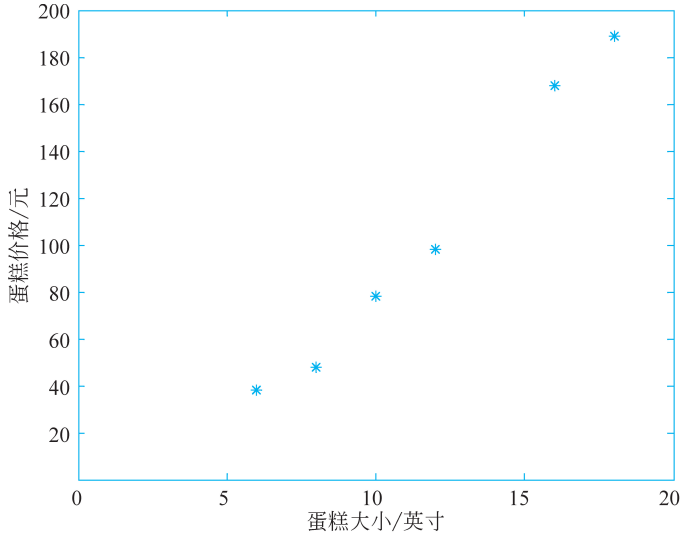

图 5-2 蛋糕尺寸/价格对应图

那么，怎样来预测蛋糕的尺寸与价格的关系？

设已有的数据集

$$D=\{(x_1,y_1)(x_2,y_2),\cdots,(x_6,y_6)\}$$
$$=\{(6,38),(8,48),(10,68),(12,98),(16,168),(18,198)\}$$

显然,我们可以猜想(或称估计)蛋糕的尺寸与价格之间具有线性关系。

(2) 怎样用一元线性回归方法解决问题?

我们可以根据这些数据建立一个蛋糕尺寸与蛋糕价格的一元线性模型,进行蛋糕价格的预测。即 $f(x)=w_1x+w_0$,其中 x 为自变量(表示蛋糕尺寸),$f(x)$为因变量(表示蛋糕价格),w_i 是系数($i=0,1,\cdots,n$),用此来预测各种尺寸蛋糕的价格。现在只需要确定系数 w_1、w_0,使用前面两个点(6,38)、(8,48),就可以确定一条直线,进而预测 14 英寸蛋糕的价格。即解下面的方程组:

$$\begin{cases}38=6w_1+w_0\\48=8w_1+w_0\end{cases}$$

解得 $w_0=8$、$w_1=5$,进而得到一元线性方程为

$$f(x)=5x+8$$

这个一元线性方程与实际价格对比情况如图 5-3 所示,其中符号"*"表示原始数据。可以看出,只用两点给出的一元线性方程来拟合蛋糕尺寸与价格的模型效果很差,蛋糕尺寸在 10 英寸以上时,预测价格与实际价格相差很大。预测出 14 英寸蛋糕价格为 78 元。显然是不合适的,因为这个价格比 12 英寸蛋糕的价格 98 元还低。

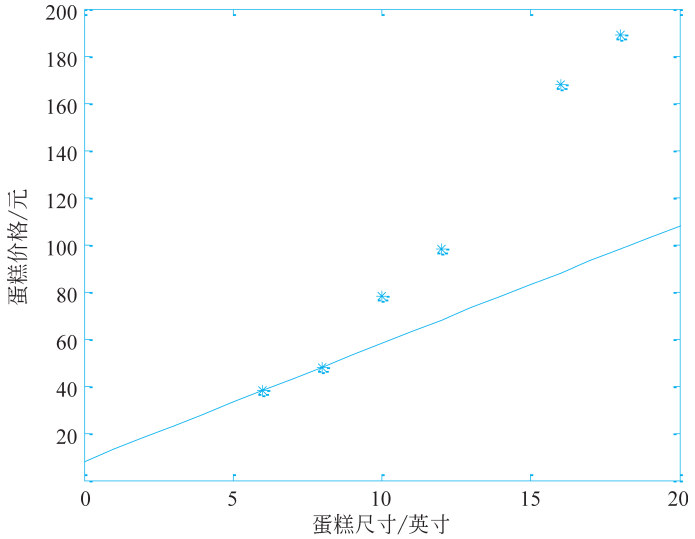

图 5-3　应用两点给出的一元线性方程与实际价格对比图

如何解决这个问题?

假设用一元线性模型 $f(x)=w_1x+w_0$ 来预测各个尺寸蛋糕的价格,我们希望模型的预测值 $f(x_i)$与实际值 y_i 最接近,也就是它们的均方误差最小。即使得 $\sum_{i=1}^{N}(f(x_i)-y_i)^2$ 最小的 w_1、w_0,我们记作 w_1^*、w_0^*,即

$$w_1^*,w_0^*=\underset{w_1,w_0}{\arg\min}\sum_{i=1}^{N}(f(x_i)-y_i)^2=\underset{w_1,w_0}{\arg\min}\sum_{i=1}^{N}(w_1x_i+w_0-y_i)^2 \qquad (5\text{-}1)$$

其中，arg min 表示求最小值点，也就是使 $\sum_{i=1}^{N}(f(x_i)-y_i)^2$ 最小的 w_1、w_0。

这个求取 w_1^*、w_0^* 的过程称为参数估计，用均方误差进行估计的方法称为“最小二乘法”。

如何求得 w_1^*、w_0^* 的值？

设函数

$$g(w_1,w_0)=\sum_{i=1}^{N}(w_1x_i+w_0-y_i)^2 \tag{5-2}$$

这是关于 w_1、w_0 的二次函数，其最小值是在对 w_1、w_0 的偏导数为 0 时取得的。因此，对 w_1、w_0 求偏导数，使其等于 0。得到下面两个方程：

$$\frac{\partial(g(w_1,w_0))}{\partial w_1}=\frac{\partial\left(\sum_{i=1}^{N}(w_1x_i+w_0-y_i)^2\right)}{\partial w_1}=2\sum_{i=1}^{N}(w_1x_i+w_0-y_i)x_i=0 \tag{5-3}$$

$$\frac{\partial(g(w_1,w_0))}{\partial w_0}=\frac{\partial\left(\sum_{i=1}^{N}(w_1x_i+w_0-y_i)^2\right)}{\partial w_0}=2\sum_{i=1}^{N}(w_1x_i+w_0-y_i)=0 \tag{5-4}$$

可以解得

$$w_0=\frac{1}{N}\sum_{i=1}^{N}(y_i-w_1x_i)=\frac{1}{N}\sum_{i=1}^{N}y_i-w_1\frac{1}{N}\sum_{i=1}^{N}x_i \tag{5-5}$$

其中，$\frac{1}{N}\sum_{i=1}^{N}x_i$ 与 $\frac{1}{N}\sum_{i=1}^{N}y_i$ 分别是数据集中 x_i 和 y_i 的均值，令 $\bar{y}=\frac{1}{N}\sum_{i=1}^{N}y_i$，$\bar{x}=\frac{1}{N}\sum_{i=1}^{N}x_i$，将其代入上式，可得

$$w_0=\bar{y}-w_1\bar{x} \tag{5-6}$$

进而可得

$$w_1=\frac{\sum_{i=1}^{N}x_iy_i-N\bar{x}\bar{y}}{\sum_{i=1}^{N}x_i^2-N\bar{x}^2}$$

$$w_0=\bar{y}-\bar{x}\left(\frac{\sum_{i=1}^{N}x_iy_i-N\bar{x}\bar{y}}{\sum_{i=1}^{N}x_i^2-N\bar{x}^2}\right) \tag{5-7}$$

这里的 w_1、w_0 就是我们要求的 w_1^*、w_0^*，进而得到一元线性模型为

$$f(x)=w_1x+w_0=\left(\frac{\sum_{i=1}^{N}x_iy_i-N\bar{x}\bar{y}}{\sum_{i=1}^{N}x_i^2-N\bar{x}^2}\right)x+\left(\bar{y}-\bar{x}\left(\frac{\sum_{i=1}^{N}x_iy_i-N\bar{x}\bar{y}}{\sum_{i=1}^{N}x_i^2-N\bar{x}^2}\right)\right) \tag{5-8}$$

应用这个模型，我们针对生日蛋糕数据集求出一元线性回归模型，数据集如下

$$\begin{aligned}D&=\{(x_1,y_1)(x_2,y_2),\cdots,(x_6,y_6)\}\\&=\{(6,38),(8,48),(10,68),(12,98),(16,168),(18,198)\}\end{aligned}$$

有

$$w_1=\frac{\sum_{i=1}^{6}x_iy_i-6\bar{x}\bar{y}}{\sum_{i=1}^{6}x_i^2-6\bar{x}^2}=13.3820 \tag{5-9}$$

$$w_0=\bar{y}-\bar{x}\left(\frac{\sum_{i=1}^{6}x_iy_i-6\bar{x}\bar{y}}{\sum_{i=1}^{6}x_i^2-6\bar{x}^2}\right)=-52.9565 \tag{5-10}$$

可得关于“生日蛋糕”尺寸与价格的预测模型为

$$f(x)=w_1x+w_0=13.3820x-52.9565 \tag{5-11}$$

数据集中原始数据、用数据集中前两点求得的直线、应用“最小二乘法”求得的一元线性回归模型的对比情况如图 5-4 所示。从图中可以看出,应用“最小二乘法”求得的一元线性回归模型很好,14 英寸蛋糕的价格按照这个模型预测为 135 元左右,较为合理。

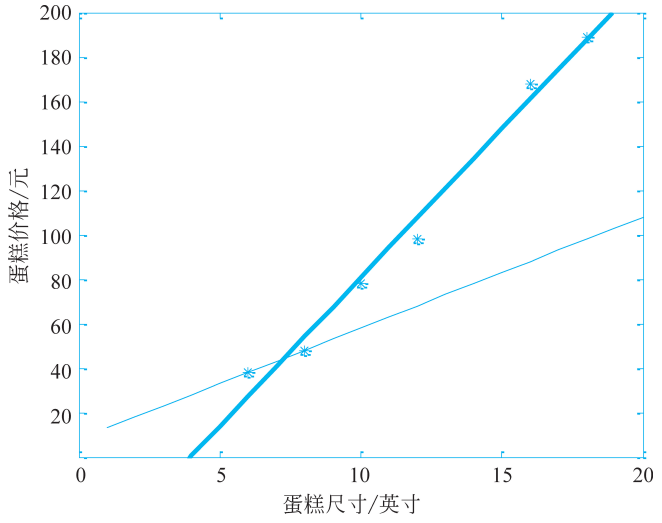

图 5-4　拟合情况图

拟合情况的好坏只是我们观察后得到的结论,具有主观性。那么,衡量线性回归模型性能好坏的性能指标是什么?

(3) 如何衡量给出的线性回归模型的性能?

各个蛋糕价格有一定的随机性,我们的目标要给出一个最佳的一元线性模型,能够反映蛋糕尺寸与价格的关系。“怎样衡量最佳模型”就成了关键因素。这里引入统计学的几个概念,给出衡量线性回归模型的性能指标。

总偏差平方和(sum of squares for total,SST)是每个因变量的实际值与其平均值的差的平方,即

$$\mathrm{SST}=\sum_{i=1}^{N}(y_i-\bar{y})^2 \tag{5-12}$$

总偏差平方和的值反映了因变量取值的总体波动情况,其值越大,说明原始数据本身具有越