TURING 图灵计算机科学丛书

CAMBRIDGE

Data Mining and Analysis
Fundamental Concepts and Algorithms

数据挖掘与分析
概念与算法

[美] Mohammed J. Zaki [巴西] Wagner Meira Jr. 著
吴诚堃 译

专注于数据挖掘与分析的基本概念和算法，融合机器学习、统计学等相关学科知识，
涵盖频繁模式挖掘、聚类、分类等经典算法

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

“本书的两位作者都是数据挖掘领域享誉世界的专家，书中几乎涵盖了数据挖掘所有主题，从基本的统计学知识到高级的数据挖掘方法一应俱全，称得上是一本数据挖掘的百科全书。书中每个概念的介绍都完美地融合了直觉、算法演示和严格的数学分析。无论是作为教材还是作为参考书，本书都是明智之选。”

—— Christos Faloutsos

卡内基·梅隆大学计算机系教授，ACM SIGKDD年度最佳创新奖获得者

“本书是数据挖掘或数据分析课程的上佳教材。作者兼顾数据挖掘的基础知识和前沿话题，强调数学基础和算法的运用。每一章都配备了习题，相关的数据、课程幻灯片以及其他辅助材料都可以在本书网站上找到，十分方便。”

——Gregory Piatetsky-Shapiro

KDnuggets.com网站董事长、编辑，ACM SIGKDD共同创建者和组织者

- 数据挖掘与分析的**入门书**，针对初学者阐述所有关键概念，包括探索性数据分析、频繁模式挖掘、聚类和分类。
- 兼顾**前沿话题**，例如核方法、高维数据分析、复杂图和网络等。
- 提供算法对应的**开源实现方法**。
- 每章均有丰富**示例和练习**，帮助读者理解和巩固相关主题。
- 配备丰富**教辅资源**，包括课程幻灯片、教学视频、数据集等，可从以下网址获取：http://www.dataminingbook.info/pmwiki.php/Main/BookResources。

CAMBRIDGE
UNIVERSITY PRESS
www.cambridge.org

图灵社区：iTuring.cn
热线：(010) 51095186 转 600

分类建议 计算机 / 数据库

人民邮电出版社网址：www.ptpress.com.cn

ISBN 978-7-115-45842-1

ISBN 978-7-115-45842-1
定价：129.00元

TURING 图灵计算机科学丛书

Data Mining and Analysis
Fundamental Concepts and Algorithms

数据挖掘与分析
概念与算法

[美] Mohammed J. Zaki [巴西] Wagner Meira Jr. 著
吴诚堃 译

人民邮电出版社
北 京

图书在版编目（CIP）数据

数据挖掘与分析：概念与算法 /（美）穆罕默德·扎基（Mohammed J. Zaki），（巴西）小瓦格纳·梅拉（Wagner Meira Jr.）著；吴诚堃译. ——北京：人民邮电出版社，2017.9
（图灵计算机科学丛书）
ISBN 978-7-115-45842-1

Ⅰ. ①数… Ⅱ. ①穆… ②小… ③吴… Ⅲ. ①数据采集 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2017)第125598号

内 容 提 要

本书是专注于数据挖掘与分析的入门图书，内容分为数据分析基础、频繁模式挖掘、聚类和分类四个部分，每一部分的各个章节兼顾基础知识和前沿话题，例如核方法、高维数据分析、复杂图和网络等。每一章最后均附有参考书目和习题。

本书适合高等院校相关专业的学生和教师阅读，也适合从事数据挖掘与分析相关工作的人员学习参考。

◆ 著 [美] Mohammed J. Zaki
[巴西] Wagner Meira Jr.
译 吴诚堃
责任编辑 朱 巍
执行编辑 温 雪
责任印制 彭志环

◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路11号
邮编 100164 电子邮件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
北京鑫正大印刷有限公司印刷

◆ 开本：787×1092 1/16
印张：32.25
字数：765千字 2017年9月第1版
印数：1–3500册 2017年9月北京第1次印刷

著作权合同登记号 图字：01-2015-6180号

定价：129.00元

读者服务热线：(010)51095186转600 印装质量热线：(010) 81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字20170147号

前　言

本书源自美国伦斯勒理工学院（RPI）和巴西米纳斯吉拉斯联邦大学（UFMG）数据挖掘课程讲义。自 1998 年起，RPI 每年秋季都会开设数据挖掘课程，UFMG 自 2002 年起也开设了这门课程。尽管有不少关于数据挖掘及相关话题的好书，但我们感觉大多数书的层次或难度太高。我们的目标是写一本专注于数据挖掘与分析的基本算法的入门书，通过解释所有初次碰到的关键概念，为学习数据挖掘的核心方法打下数学基础，并试图通过直观地阐述各种公式以辅助理解。

本书主要内容包括：探索性数据分析、频繁模式挖掘、聚类和分类。本书既能为以上任务打下良好的基础，又兼顾了前沿话题，例如核方法、高维数据分析、复杂图和网络等。本书融合了相关学科（如机器学习和统计学）中的相关概念，也非常适用于数据分析课程。绝大部分的必备知识都包含在本书之中，尤其是关于线性代数、概率和统计的知识。

本书使用了大量的例子来阐述主要的技术概念，同时每章末尾还附有习题（课上使用过的）。本书中涉及的所有算法作者都实现了一遍。建议读者使用自己喜欢的数据分析和挖掘软件来尝试书中给出的例子，并实现书中所描述的算法；我们推荐使用 R 或者 Python 的 NumPy 包。书中涉及的所有数据集及其他参考材料，如课程项目构思以及课堂讲义等，都可以在以下网址找到：

http://dataminingbook.info/pmwiki.php

理解了数据挖掘和数据分析的基本原理和算法之后，读者将完全有能力开发自己的方法或者使用更高级的技术。

建议阅读路线

本书各章之间的依赖关系如图 0-1 所示。下面给出阅读本书或在课程中使用本书的几种典型路线图。对于本科生课程，建议讲授第 1~3 章、第 8 章、第 10 章、第 12~15 章、第 17~19 章，以及第 21~22 章。对于不讲探索性数据分析的本科生课程，建议讲授第 1 章、第 8~15 章、第 17~19 章及第 21~22 章。对于研究生课程，可以快速把第一部分过一遍，或将其当作背景知识阅读，然后直接讲授第 9~22 章；本书的其他部分，即频繁模式挖掘（第二部分）、聚类（第三部分）和分类（第四部分），可以按任意顺序讲授。对于讲数据分析的课程，必须讲授第 1~7 章、第 13~14 章、第 15 章的第 2 节，以及第 20 章。最后，对于强调图和核的课程，建议讲授第 4~5 章、第 7 章（第 1~3 节）、第 11~12 章、第 13 章（第 1~2 节）、第 16~17 章和第 20~22 章。

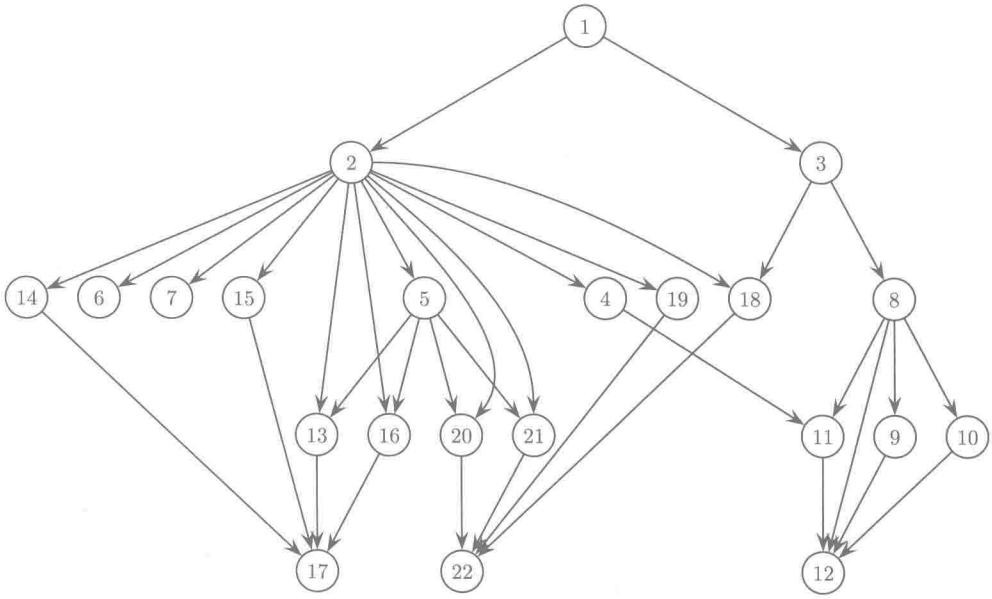

图 0-1 各章依赖关系

致谢

本书的初稿已在若干数据挖掘课程中使用过。参与试用的教师和学生提供了很多宝贵的意见和建议，特此致谢：

- Muhammad Abulaish，印度国立伊斯兰大学
- Mohammad Al Hasan，印第安纳大学与普渡大学印第安纳波里斯联合分校
- Marcio Luiz Bunte de Carvalho，巴西米纳斯吉拉斯联邦大学
- Loïc Cerf，巴西米纳斯吉拉斯联邦大学
- Ayhan Demiriz，土耳其萨卡里亚大学
- Murat Dundar，印第安纳大学与普渡大学印第安纳波里斯联合分校
- Jun Luke Huan，堪萨斯大学
- Ruoming Jin，肯特州立大学
- Latifur Khan，得克萨斯州大学达拉斯分校
- Pauli Miettinen，德国马克斯·普朗克计算机科学研究所
- Suat Ozdemir，土耳其加齐大学
- Naren Ramakrishnan，弗吉尼亚理工学院暨州立大学
- Leonardo Chaves Dutra da Rocha，巴西圣若昂–德尔雷伊联邦大学
- Saeed Salem，北达科塔州立大学
- Ankur Teredesai，华盛顿大学塔科马分校
- Hannu Toivonen，芬兰赫尔辛基大学
- Adriano Alonso Veloso，巴西米纳斯吉拉斯联邦大学
- Jason T.L. Wang，新泽西理工学院

• Jianyong Wang，清华大学
• Jiong Yang，凯斯西储大学
• Jieping Ye，亚利桑那州立大学

我们还要感谢参加了 RPI 和 UFMG 的数据挖掘课程的学生，以及为各章提供了技术性建议的匿名审稿人。感谢 RPI 和 UFMG 的计算机科学系以及卡塔尔计算研究所的合作与支持性氛围。此外，还要感谢美国国家科学基金会、巴西国家科学技术发展委员会、巴西高等教育人员促进会、巴西米纳斯吉拉斯州研究支持基金会、巴西国家网络科技研究所，以及巴西科学无国界计划的支持。特别感谢本书编辑、剑桥大学出版社的 Lauren Cowles 为本书的出版给予的指导和耐心的帮助。

最后，从个人角度而言，Mohammed J. Zaki 将此书献给他的妻子 Amina，以感谢她的爱、耐心与多年来的支持；也献给他的孩子 Abrar 和 Afsah，以及他的父母。Wagner Meira Jr. 将此书献给他的妻子 Patricia、孩子 Gabriel 和 Marina，以及父母 Wagner 和 Marlene，感谢他们的爱、鼓励和启发。

目　录

第三部分 聚类

第四部分 分类

第 1 章　数据挖掘与分析

数据挖掘是从大规模数据中发现新颖、深刻、有趣的模式和具有描述性、可理解、能预测的模型的过程。本章首先讨论数据矩阵的基本性质。我们会强调数据的几何和代数描述以及概率解释。接下来讨论数据挖掘的主要任务，包括探索性数据分析、频繁模式挖掘、聚类和分类，从而为本书设定基本的脉络。

1.1　数据矩阵

数据经常可以表示或者抽象为一个 $n \times d$ 的*数据矩阵*，包含 n 行 d 列，其中各行代表数据集中的实体，而各列代表了实体中值得关注的特征或者属性。数据矩阵中的每一行记录了一个给定实体的属性观察值。$n \times d$ 的矩阵如下所示：

$$\boldsymbol{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \boldsymbol{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \boldsymbol{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}\right)$$

其中 $\boldsymbol{x}_i$ 表示第 i 行的一个如下 d 元组：

$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{id})$$

而 X_j 表示第 j 列的一个如下 n 元组：

$$X_j = (x_{1j}, x_{2j}, \cdots, x_{nj})$$

根据应用领域的不同，数据矩阵的行还可以被称作*实体*、*实例*、*样本*、*记录*、*事务*、*对象*、*数据点*、*特征向量*、*元组*，等等。同样，列可以被称作*属性*、*性质*、*特征*、*维度*、*变量*、*域*，等等。实例的数目 n 被称作数据的*大小*，属性的数目 d 被称作数据的*维度*。针对单个属性进行的分析，称作*一元分析*；针对两个属性进行的分析，称作*二元分析*；针对两个以上的属性进行的分析，称作*多元分析*。

例 1.1　表 1-1 列举了鸢尾花（iris）数据集的一部分数据。完整的数据集是一个 150×5 的矩阵。每一行代表一株鸢尾花，包含的属性有：萼片长度、萼片宽度、花瓣长度、花瓣宽度（以厘米计），以及该鸢尾花的类型。第一行是一个如下的五元组：

$$\boldsymbol{x}_1 = (5.9, 3.0, 4.2, 1.5, \texttt{iris-versicolor})$$

并非所有的数据都是以矩阵的形式出现的。复杂一些的数据还可以以序列（例如 DNA 和蛋白质序列）、文本、时间序列、图像、音频、视频等形式出现，这些数据的分析需要专门

表 1-1 鸢尾花数据集的一部分数据

	萼片长度	萼片宽度	花瓣长度	花瓣宽度	类型
	X_1	X_2	X_3	X_4	X_5
$\boldsymbol{x}_1$	5.9	3.0	4.2	1.5	iris-versicolor
$\boldsymbol{x}_2$	6.9	3.1	4.9	1.5	iris-versicolor
$\boldsymbol{x}_3$	6.6	2.9	4.6	1.3	iris-versicolor
$\boldsymbol{x}_4$	4.6	3.2	1.4	0.2	iris-setosa
$\boldsymbol{x}_5$	6.0	2.2	4.0	1.0	iris-versicolor
$\boldsymbol{x}_6$	4.7	3.2	1.3	0.2	iris-setosa
$\boldsymbol{x}_7$	6.5	3.0	5.8	2.2	iris-versicolor
$\boldsymbol{x}_8$	5.8	2.7	5.1	1.9	iris-versicolor
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{x}_{149}$	7.7	3.8	6.7	2.2	iris-virginica
$\boldsymbol{x}_{150}$	5.1	3.4	1.5	0.2	iris-setosa

的技术。然而，在大多数情况下，即使原始数据不是一个数据矩阵，我们还是可以通过特征提取将原始数据转换为一个数据矩阵。例如，给定一个图像数据库，我们可以创建这样一个数据矩阵：每一行代表一幅图像，各列对应图像的特征，如颜色、纹理等。有些时候，某些特定的特征可能蕴含了特殊的语义，处理的时候需要特别对待。比如，时序和空间特征通常都要区别对待。值得指出的是，传统的数据分析假设各个实体或实例之间是相互独立的。但由于我们生活在一个互联的世界里面，这一假设并不总是成立。一个实例可能通过各种各样的关系与其他实例相关联，从而形成一个*数据图*：图的节点代表实例，图的边代表实例间的关联关系。

1.2 属性

属性根据它们所取的值主要可以分为两类。

1. *数值型属性*

*数值型属性*是在实数或者整数域内取值。例如，取值域为 $\mathbb{N}$ 的属性Age（年龄），即是一个数值型属性，其中 $\mathbb{N}$ 代表全体的自然数（即所有的非负整数）；表 1-1 中的花瓣长度同样也是一个数值型属性，其取值域为 $\mathbb{R}^+$（代表全体正实数）。取值范围为有限或无限可数集合的数值属性又被称作*离散型的*；取值范围为任意实数的数值属性又被称作*连续型的*。作为离散型的特例，取值限于集合 $\{0,1\}$ 的属性又被称为*二元属性*。数值属性又可以进一步分成如下两类。

- *区间标度类*：对于这一类属性，只有差值（加或减）有明确的意义。例如，属性温度（无论是摄氏温度还是华氏温度）就是属于区间标度型的。假设某一天是 20°C，接下来一天是 10°C，那么谈论温度降了 10°C 是有意义的，却不能说第二天比第一天冷两倍。
- *比例标度类*：对于这一类属性，不同值之间的差值和比例都是有意义的。例如，对

于属性年龄，我们可以说一个 20 岁的人的年龄是另一个 10 岁的人的年龄的两倍。

2. 类别型属性

类别型属性的定义域是由一个定值符号集合定义的。例如，Sex（性别）和Education（教育水平）都是类别型属性，它们的定义域如下所示：

$$\text{domain}(\texttt{Sex}) = \{\texttt{M}, \texttt{F}\}$$

$$\text{domain}(\texttt{Education}) = \{\texttt{HighSchool}, \texttt{BS}, \texttt{MS}, \texttt{PhD}\}$$

类别型属性也可分为两种类型。

- 名义类：这类属性的定义域是无序的，只有相等性比较才有意义。也就是说，我们只能判断两个不同实例的属性值是否相等。例如性别就是一个名义类的属性。表 1-1 中的类别属性也是名义类的，其定义域 domain（class）={iris-setosa，iris-versicolor，iris-virginica}。
- 次序类：这类属性的定义域是有序的，因此相等性比较（两个值是否相等）与不等性比较（一个值比另一个值大还是小）都是有意义的，尽管有的时候不能够量化不同值之间的差。例如，教育水平是一个次序类属性，因为它的定义域是按受教育水平递增排序的。

1.3 数据的几何和代数描述

若数据矩阵 $\boldsymbol{D}$ 的 d 个属性或者维度都是数值型的，则每一行都可以看作一个 d 维空间的点：

$$\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{id}) \in \mathbb{R}^d$$

或者，每一行可以等价地看作一个 d 维的列向量（所有向量都默认为列向量）：

$$\boldsymbol{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{pmatrix} = (x_{i1} \quad x_{i2} \quad \cdots \quad x_{id})^{\mathrm{T}} \in \mathbb{R}^d$$

其中 T 是矩阵转置算子。

d 维笛卡儿坐标空间是由 d 个单位向量定义的，又被称作标准基，每个轴方向上一个。第 j 个标准基向量 $\boldsymbol{e}_j$ 是一个 d 维的单位向量，该向量的第 j 个分量是 1，其他分量是 0。

$$\boldsymbol{e}_j = (0, \cdots, 1_j, \cdots, 0)^{\mathrm{T}}$$

$\mathbb{R}^d$ 中的任何向量都可以由标准基向量的线性组合来表示。例如，每一个点 $\boldsymbol{x}_i$ 都可以用如下的线性组合来表示：

$$\boldsymbol{x}_i = x_{i1}\boldsymbol{e}_1 + x_{i2}\boldsymbol{e}_2 + \cdots + x_{id}\boldsymbol{e}_d = \sum_{j=1}^{d} x_{ij}\boldsymbol{e}_j$$

其中标量 x_{ij} 是沿着第 j 个轴的坐标值或者第 j 个属性。

例 1.2　考虑表 1-1 中的鸢尾花数据。如果我们将所有数据映射到前两个属性，那么每一行都可以看作二维空间中的一个点或是向量。例如，五元组 $\boldsymbol{x}_1$=（5.9, 3.0, 4.2, 1.5, `iris-versicolor`）在前两个属性上的投影如图 1-1a 所示。图 1-2 给出了所有 150 个数据点在由前两个属性构成的二维空间中的散点图。图 1-1b 将 $\boldsymbol{x}_1$ 显示为三维空间中的一个点和向量，该三维空间将数据映射到前三个属性。(5.9, 3.0, 4.2) 可以看作 $\mathbb{R}^3$ 中标准基线性组合的系数：

$$\boldsymbol{x}_1 = 5.9\boldsymbol{e}_1 + 3.0\boldsymbol{e}_2 + 4.2\boldsymbol{e}_3 = 5.9\begin{pmatrix}1\\0\\0\end{pmatrix} + 3.0\begin{pmatrix}0\\1\\0\end{pmatrix} + 4.2\begin{pmatrix}0\\0\\1\end{pmatrix} = \begin{pmatrix}5.9\\3.0\\4.2\end{pmatrix}$$

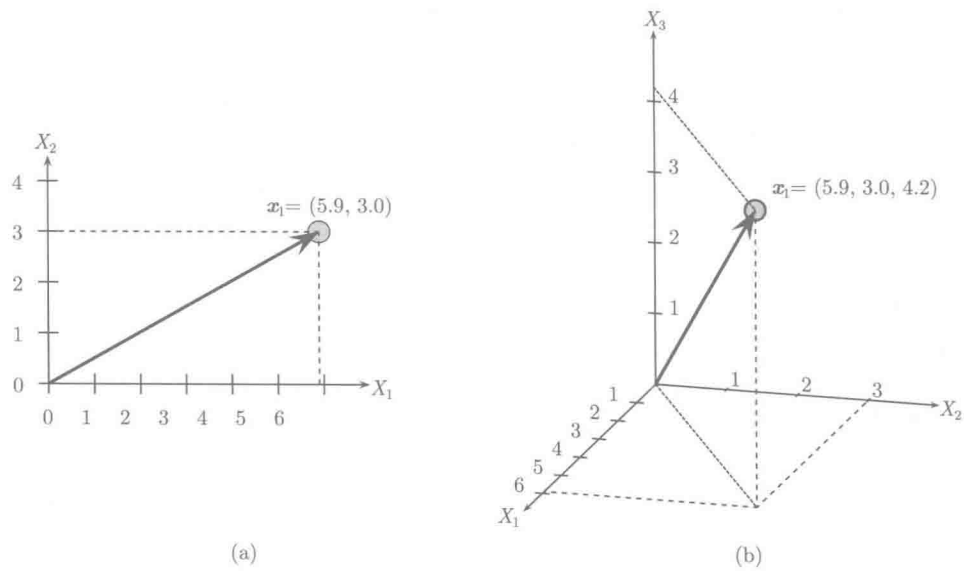

图 1-1　行 $\boldsymbol{x}_1$ 在不同空间中分别作为一个点和一个向量：(a)$\mathbb{R}^2$；(b)$\mathbb{R}^3$

每一个数值型的列或属性还可以看成 n 维空间 $\mathbb{R}^n$ 中的一个向量：

$$X_j = \begin{pmatrix}x_{1j}\\x_{2j}\\\vdots\\x_{nj}\end{pmatrix}$$

如果所有的属性都是数值型的，那么数据矩阵 $\boldsymbol{D}$ 事实上是一个 $n \times d$ 的矩阵，可记作 $\boldsymbol{D} \in \mathbb{R}^{n\times d}$，如以下公式所示：

$$\boldsymbol{D} = \begin{pmatrix}x_{11} & x_{12} & \cdots & x_{1d}\\x_{21} & x_{22} & \cdots & x_{2d}\\\vdots & \vdots & \ddots & \vdots\\x_{n1} & x_{n2} & \cdots & x_{nd}\end{pmatrix} = \begin{pmatrix}-\boldsymbol{x}_1^{\mathrm{T}}-\\-\boldsymbol{x}_2^{\mathrm{T}}-\\\vdots\\-\boldsymbol{x}_n^{\mathrm{T}}-\end{pmatrix} = \begin{pmatrix}| & | & & |\\X_1 & X_2 & \cdots & X_d\\| & | & & |\end{pmatrix}$$

我们可以将整个数据集看成一个 $n \times d$ 的矩阵，或是一组行向量 $\boldsymbol{x}_i^{\mathrm{T}} \in \mathbb{R}^d$，或是一组列向量 $X_j \in \mathbb{R}^n$。

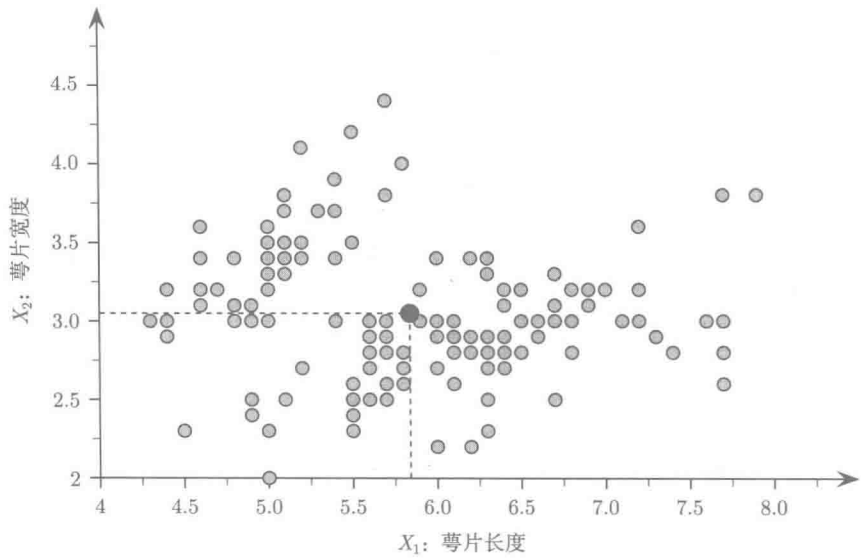

图 1-2 萼片长度与萼片宽度的散点图，实心圈代表平均点

1.3.1 距离和角度

将数据实例和属性用向量来描述或者将整个数据集描述为一个矩阵，可以应用几何与代数的方法来辅助数据挖掘与分析任务。

假设 $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^m$ 是如下的两个 m 维向量：

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} \quad \boldsymbol{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

1. 点乘

$\boldsymbol{a}$ 和 $\boldsymbol{b}$ 的点乘定义为如下的标量值：

$$\begin{aligned} \boldsymbol{a}^{\mathrm{T}}\boldsymbol{b} &= \begin{pmatrix} a_1 & a_2 & \cdots & a_m \end{pmatrix} \times \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} \\ &= a_1 b_1 + a_2 b_2 + \cdots + a_m b_m \\ &= \sum_{i=1}^{m} a_i b_i \end{aligned}$$

2. 长度

向量 $\boldsymbol{a} \in \mathbb{R}^m$ 的欧几里得范数或长度定义为：

$$\|\boldsymbol{a}\| = \sqrt{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}} = \sqrt{a_1^2 + a_2^2 + \cdots + a_m^2} = \sqrt{\sum_{i=1}^{m} a_i^2}$$

$\boldsymbol{a}$ 方向上的单位向量定义为：

$$\boldsymbol{u} = \frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} = \left(\frac{1}{\|\boldsymbol{a}\|}\right)\boldsymbol{a}$$

根据定义，单位向量的长度为 $\|\boldsymbol{u}\| = 1$，它又可被称为正则化向量，在某些分析中可以代替向量 $\boldsymbol{a}$。

欧几里得范数 L_p 是范数的特例，定义为：

$$\|\boldsymbol{a}\|_p = (|a_1|^p + |a_2|^p + \cdots + |a_m|^p)^{\frac{1}{p}} = \left(\sum_{i=1}^{m} |a_i|^p\right)^{\frac{1}{p}}$$

其中 $p \neq 0$。因此，欧几里得范数是 $p = 2$ 的 L_p 范数。

3. 距离

根据欧几里得范数，我们可以定义两个向量 $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 之间的欧氏距离如下：

$$\delta(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{a} - \boldsymbol{b}\| = \sqrt{(\boldsymbol{a} - \boldsymbol{b})^{\mathrm{T}}(\boldsymbol{a} - \boldsymbol{b})} = \sqrt{\sum_{i=1}^{m} (a_i - b_i)^2} \tag{1.1}$$

因此，向量 $\boldsymbol{a}$ 的长度即是它到零向量 $\mathbf{0}$ 的距离（零向量的所有元素都为 0），亦即 $\|\boldsymbol{a}\| = \|\boldsymbol{a} - \mathbf{0}\| = \delta(\boldsymbol{a}, \mathbf{0})$。

根据 L_p 范数的定义，我们可以定义对应的 L_p 距离函数如下：

$$\delta_p(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{a} - \boldsymbol{b}\|_p \tag{1.2}$$

若没有指明 p 的具体值，如公式 (1.1)，则默认 $p = 2$。

4. 角度

两个向量 $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 之间的最小角的余弦值，被称作余弦相似度，由如下公式定义：

$$\cos\theta = \frac{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b}}{\|\boldsymbol{a}\|\|\boldsymbol{b}\|} = \left(\frac{\boldsymbol{a}}{\|\boldsymbol{a}\|}\right)^{\mathrm{T}}\left(\frac{\boldsymbol{b}}{\|\boldsymbol{b}\|}\right) \tag{1.3}$$

因此，向量 $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 间角度的余弦可以通过 $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 的单位向量 $\frac{\boldsymbol{a}}{\|\boldsymbol{a}\|}$ 和 $\frac{\boldsymbol{b}}{\|\boldsymbol{b}\|}$ 的点乘来计算。

柯西–施瓦茨（Cauchy-Schwartz）不等式描述了对于 $\mathbb{R}^m$ 中的任意向量 $\boldsymbol{a}$ 和 $\boldsymbol{b}$，若满足如下关系：

$$|\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b}| \leqslant \|\boldsymbol{a}\| \cdot \|\boldsymbol{b}\|$$

则根据柯西–施瓦茨不等式马上可以得到：

$$-1 \leqslant \cos\theta \leqslant 1$$

由于两个向量之间的最小角 $\theta \in [0°, 180°]$ 且 $\cos\theta \in [-1, 1]$，余弦相似度取值范围为 $+1$（对应 0° 角）到 -1（对应 180° 角，或是 π 弧度）。