北京理工大学“双一流”建设精品出版工程

Data Mining Techniques And Applications

数据挖掘技术与应用

由育阳 ◎ 编著

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

北京理工大学“双一流”建设精品出版工程

数据挖掘技术与应用

由育阳　编著

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

图书在版编目（CIP）数据

数据挖掘技术与应用 / 由育阳编著. —北京：北京理工大学出版社，2021.6
ISBN 978-7-5682-9257-3

Ⅰ. ①数…　Ⅱ. ①由…　Ⅲ. ①数据采集　Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2020）第 226348 号

出版发行 / 北京理工大学出版社有限责任公司
社　　址 / 北京市海淀区中关村南大街 5 号
邮　　编 / 100081
电　　话 / (010)68914775(总编室)
　　(010)82562903(教材售后服务热线)
　　(010)68948351(其他图书服务热线)
网　　址 / http://www.bitpress.com.cn
经　　销 / 全国各地新华书店
印　　刷 / 保定市中画美凯印刷有限公司
开　　本 / 787 毫米×1092 毫米　1/16
印　　张 / 10.5
字　　数 / 247 千字
版　　次 / 2021 年 6 月第 1 版　2021 年 6 月第 1 次印刷
定　　价 / 68.00 元

责任编辑 / 孙　澍
文案编辑 / 孙　澍
责任校对 / 周瑞红
责任印制 / 李志强

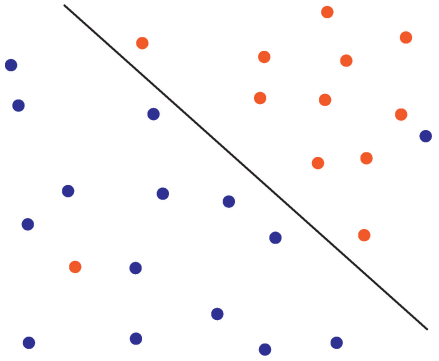

图 6 – 2　异常点影响超平面的选择

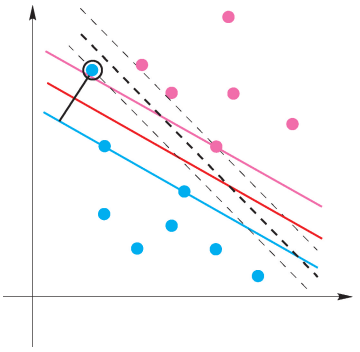

图 6 – 3　异常点影响模型预测效果

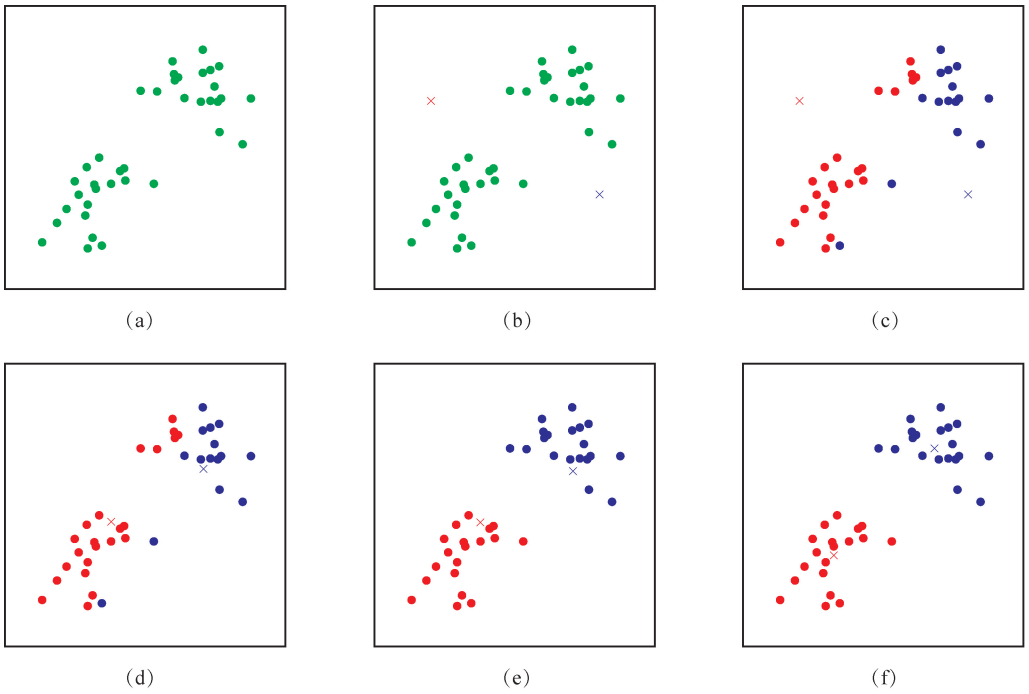

图 9 – 1　*K* – Means 的启发式方式

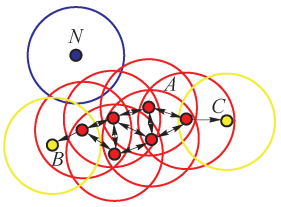

图 9 – 2　DBSCAN 基本概念

前 言

PREFACE

大数据时代，数据挖掘技术被越来越广泛应用于解决工程应用和科学领域的复杂问题。近年来数据科学相关技术更新较快，但“数据挖掘”课程的教材出版较少。从探索培养应用型人才所需要的数据挖掘知识体系的角度出发，我们组织编写了这本质量过硬、新颖实用的教材。本教材适用于数据科学相关的本科、研究生相关专业，包括自动化、计算机、生物信息、机械类、材料类等。

本教材编写的指导思想是面向应用型数据挖掘人才的相关知识体系，充分考虑应用能力与理论知识相结合。书中大部分内容参考了英文资料，大部分专业术语给出了中文介绍，但仍有一些术语保留英文原文，以供读者参考。

北京理工大学动力学与先进控制实验室的硕士研究生刘国正、俞云开、单文婧、刘玉征同学参加了本书的部分图表制作和文字校对工作。

欢迎广大数据科学工作者和感兴趣的读者对本书提出宝贵意见，以促进本教材质量的提高。

编　者

2021 年 6 月

目　录

CONTENTS

第1章 绪 论

1.1 数据挖掘的定义

数据挖掘（Data Mining），通俗来讲就是在大量的数据中发现有用的信息。随着信息技术的发展，每天都会产生大量的数据，可以说我们正处于一个大数据的时代。面对如此海量的数据，传统的分析方法已不再适用，这就需要我们用新的技术工具从数据中找到隐藏的信息。数据挖掘这门新兴的学科涉及很多学科领域，它融合了统计学、人工智能、专家系统、智能数据库、知识获取、数据可视化及高性能计算等领域。

尽管目前对数据挖掘尚无明确的学科划定，从广义上来讲，数据挖掘先从巨大的数据体系或数据库里提炼出人们感兴趣的东西（可能在意料之中，也可能在意料之外）；或者说，从庞大的观察数据集中提炼并分析出不可轻易察觉或断言的关系，最后给出一个有用的、可以理解的结论。简单地说，数据挖掘就是在数据中发现数据之间的关系。数据挖掘也常被称为知识发现（Knowledge Discovery），因此许多知识发现中的算法——比如人工智能算法，常常被用于数据挖掘的过程中。尽管“数据挖掘”和“知识发现”的称谓在学术界并行，然而在产业界、媒体和数据库研究界，“数据挖掘”这一术语比“知识发现”更流行，因为前者更能够吸引投资者的视线，从而推动数据挖掘的使用和发展[1]。

数据挖掘有以下三个特点：

（1）数据量常常是巨大的。是否可以根据相关领域内的数据集找出数据关系即算法，使用全部数据还是随机数据或有目的地使用数据子集，能否高效地存取数据，这些问题都是数据挖掘工作者需要考虑的问题。

（2）数据挖掘面临的数据常常是为其他目的而收集的数据。这就为数据挖掘带来了一个问题，即收集数据时，可能没有收集一个或几个重要的变量，而这些变量在数据挖掘应用中被证明是有用的，甚至是至关重要的。

（3）数据挖掘工作者常常不愿把先验知识预先嵌入算法内，因为这样就等于做“假设检验”。数据挖掘常常要求算法可以主动地揭示一些数据的内在关系，结论的新颖性是衡量数据挖掘算法好坏的一个重要标准。当然，这些新颖性的结论必须是可以被人理解的，绝对不应该是漫无边际的奇怪结论。

1.2 为什么进行数据挖掘

1.2.1 数据挖掘的背景

在海量数据中发掘隐藏的信息这一需求是数据挖掘产生的前提。从古至今，人类就刻意地在生活的方方面面中搜索有用的数据信息。然而，随着信息量的快速增长，需要越来越多的更加自动、有效的数据挖掘方法。早期的方法如18世纪的贝叶斯定理和19世纪的回归分析是最早用于数据挖掘的技术。20世纪，随着计算机的普及与计算机技术的不断发展，数据采集变得越来越容易，数据存储空间也显著扩大。随着数据规模的扩大和复杂性的增加，数据分析的难度也越来越大，相应地产生了自动数据处理的方法，这些方法给数据分析提供了诸多帮助。数据处理方法包括20世纪50年代的神经网络、聚类、遗传算法，20世纪60年代的决策树，20世纪80年代的支持向量机等[2]。

数据挖掘技术已经被企业、科研机构和政府使用很多年。它被用于筛选大量数据，如航空公司乘客旅行信息、人口数据和营销数据以生成市场研究报告，尽管这些报告有时不被认为是数据挖掘。

数据挖掘通常涉及四类任务：①分类，将数据划分成预定义的组；②聚类，类似于分类，但数据的分组不是预先定义好的，算法会尝试将相似的数据分组在一起；③回归，试图找到一个对数据建模误差最小的函数，并求解这个函数；④关联规则学习，寻找变量之间的关系[3]。

数据挖掘的功能包括数据特征描述、数据识别、关联分析、分类、聚类、离群值分析和数据演化分析等。数据特征描述是对目标数据类型特征的总结；数据识别是将目标类对象的一般特征与一个或一组对比类对象的一般特征进行比较；关联分析是发现数据关联规则的过程，关联规则显示属性值条件，属性值条件在给定的一组数据中经常同时出现；分类是寻找一组描述和区分数据类或概念的模型的过程，目的是使用模型预测类标签未知的对象的类；聚类分析数据对象时不参考已知的类模型；离群值和数据演化分析描述并模型化行为随时间变化的对象的规律或趋势[4]。

1.2.2 数据挖掘的意义

数据挖掘涉及有效的数据收集存储和计算处理。数据挖掘表现为使用数据处理算法分割数据和估计未来事件发生的概率。

数据挖掘是当前计算机行业热门的研究领域之一，涉及多种学科技术，如数据库技术、统计学、机器学习、高性能计算、模式识别、神经网络、数据可视化、信息检索、图像和信号处理以及空间数据分析等。随着计算机技术的发展，数据挖掘技术将会更加广泛和深入地应用到各个学科领域。

数据挖掘是从大量的、复杂的、不规则的、随机的、模糊的数据中获取隐含的、有潜在价值的知识和信息的过程。若将此项技术科学、合理地应用于商业领域之中，那么业界人士将能够在大量的数据信息之中获取对自己或者对企业有利用价值的信息数据，以此为标准制定商业决策，可以优化调整生产经营活动，创造较高的经济效益。无论是从科技的角度来讲

还是从商业的角度来讲，数据挖掘技术的研发与应用都是非常有意义的。随着大数据的数据管理和检索技术研究的进步，数据挖掘技术将迎来巨大的发展机遇，其应用也将更加广泛，数据挖掘的工具也将更加强大。深入研究数据挖掘技术在各个领域之中如何有效地应用，发现数据挖掘技术的不足之处，能够为今后更加深入地研究和创新该项技术创造条件。

1.3　数据挖掘的应用

实际上数据挖掘技术从一开始就是面向应用的。目前，在很多重要的领域，数据挖掘技术都发挥着积极的作用。尤其是在银行、电信、保险、交通、零售（如超级市场）等商业应用领域，数据挖掘技术取得了显著的成就。数据挖掘能够帮助解决许多典型的商业问题，其中包括：数据库营销、客户群体划分、背景分析、交叉销售等市场分析行为，以及客户流失性分析、客户信用评分、欺诈发现等。

数据挖掘技术在企业市场营销中得到了比较广泛的应用，它以市场营销学的市场细分原理为基础，其基本假设是“消费者过去的行为是其今后消费倾向的最好说明”。通过收集、加工和处理涉及消费者消费行为的大量信息，确定特定消费群体或个体的兴趣、消费习惯、消费倾向和消费需求，进而推断出相应消费群体或个体下一步的消费行为。然后，以此为基础，对所识别出来的消费群体进行特定内容的定向营销。这与传统的不区分消费者对象特征的大规模营销手段相比，大大节省了营销成本，提高了营销效果，从而为企业带来更多的利润。

消费者的信息来自市场中的各种渠道。例如，每当使用信用卡消费时，企业就可以在信用卡结算过程中收集消费者的信息，记录下我们消费的时间、地点、感兴趣的商品或服务、愿意接受的价格水平和支付能力等数据。当我们在申办信用卡、办理汽车驾驶执照、填写商品保修单和其他需要填写个人信息的时候，我们的个人信息就存入了相应的业务数据库，企业除了自行收集相关业务信息之外，还可以从其他公司或机构购买此类信息为自己所用。

组合来自各种渠道的数据后，人们使用超级计算机，利用并行处理、神经元网络、模型化算法和其他信息处理技术处理组合后的数据，从中得到商家用于向特定消费群体或个体进行定向营销的决策信息。这些决策信息是如何应用的呢？例如，当银行通过对业务数据进行挖掘后，发现一个银行账户持有者突然要求申请双人联合账户，并且确认该消费者是第一次申请联合账户时，银行会推断该用户可能要结婚了，它就会向该用户定向推销用于购买房屋、支付子女学费等长期投资业务，银行甚至可能将该信息卖给专营婚庆商品和服务的公司。

在市场经济比较发达的国家和地区，许多公司都开始在原有信息系统的基础上通过数据挖掘对业务信息进行深加工，以构筑自己的竞争优势，增加自己的营业额。美国运通公司有一个用于记录信用卡业务的数据库，数据量达到上亿字符，并仍在随着业务增长不断更新。运通公司通过对这些数据进行挖掘，制定了“关联结算优惠”的促销策略，即如果一个顾客在一个商店用运通卡购买一套时装，在同一个商店再买一双鞋，就可以得到比较大的折扣，这样既可以增加商店的销售量又可以增加运通卡在该商店的使用率。例如，居住在英国伦敦的持卡消费者如果最近刚刚乘英国航空公司的航班去过巴黎，那么他可能会得到一张周末前往美国纽约的机票打折优惠卡。

商家通过数据挖掘技术制定营销策略，向消费者发出与其以前消费行为相关的推销材料。卡夫食品公司建立了一个拥有 3 000 万客户资料的数据库，数据库是通过收集对公司发

出的优惠券等促销手段做出积极反应的客户的销售记录建立起来的，卡夫公司通过数据挖掘了解特定客户的兴趣和口味，并以此为基础向他们发送特定产品的优惠券，还为他们推荐符合客户口味和健康状况的卡夫产品食谱。美国的《读者文摘》出版公司运行着一个积累了多年的业务数据库，业务数据库中包含了全球一亿多位客户的消费信息，数据库每天 24 h 连续运行，保证数据不断得到实时更新。正是基于对客户资料数据库进行数据挖掘的优势，《读者文摘》出版公司才能够从通俗杂志扩展到专业杂志、书刊和音像制品等的出版业务。

数据挖掘还有其他的一些应用。

（1）在对客户进行分析方面：银行信用卡和保险行业，利用数据挖掘将市场分成有意义的群组和部门，从而协助市场经理和业务执行人员更好地集中于对效益有促进作用的活动并开拓新的市场。

（2）在客户关系管理方面：数据挖掘可以帮助商家找出产品的使用模式和了解客户行为，从而改进通道管理（如银行分支和 ATM 机等）。例如，正确时间销售就是基于顾客生活周期模型来实施的。

（3）在零售业方面：数据挖掘用于顾客购货篮的分析可以协助商家布置货架、安排促销活动时间、组合促销商品以及了解滞销和畅销商品状况等商业活动。通过对一种商品在各连锁店的市场共享、客户统计以及历史状况的分析，可以确保销售和广告业务的有效性。

（4）在产品质量保证方面：数据挖掘协助管理大数据变量之间的相互作用，并能自动发现某些不正常的数据分布，揭示制造和装配操作过程中的变化情况和各种因素，从而协助质量工程师及时地注意到问题发生的范围并采取改正措施。

（5）在网络容量利用方面：数据挖掘可以让企业了解客户使用聚集服务的结构和模式，从而指导企业人员对网络设施做出最佳的投资决策。

在各个企事业部门的业务中，数据挖掘在假伪检测、险灾评估、失误回避、资源分配、市场销售预测和广告投资等很多方面起着很重要的作用。例如在化学及制药行业，将数据挖掘用于大量化学信息可以发现新的有用的化学成分；在遥感领域，利用每天从卫星上及其他方面来的海量数据，数据挖掘能对气象预报、臭氧层监测等起很大的作用。自 20 世纪 90 年代开始出现数据挖掘商用软件以来，据不完全统计，1998 年年底 1999 年年初，已有 50 多个厂商从事数据挖掘系统的软件开发工作，美国数据挖掘产品市场在 1994 年达到 5 000 万美元，1997 达到 3 亿美元。从产品的类型来看，通常有以下五类产品。

（1）能够提供广泛的数据挖掘能力的产品，典型的有：IBM 公司的 Intelligent Miner、SAS 公司的 Enterprise Miner。

（2）旨在为某个部门求解问题的产品，典型的有：Unica 公司的 Response Modeler Segmentor、IBM 公司的 Business Application 等。

（3）与提供服务联系在一起的产品，典型的有：NeoVista、Hyperparallel、HNC Marksman。

（4）黑匣工具，典型的有：GroupModel、ModelMax、Predict。

（5）解决客户问题的产品，典型的有：Marketier Paregram、Exchange Application。

数据挖掘（知识发现）的目的是为企业决策提供正确的依据，从分析数据、发现问题到做出决策、采取行动这一系列操作是一个单位的动作行为，利用计算机及信息技术完成整

体行动，是发挥机构活力和赢得竞争优势的唯一手段。人们将这种机构的手段称为“商业智能”（Business Intelligent，BI），BI 系统能极大地提高决策的质量和及时性，从而提高机构的生产率以发挥竞争优势。近年来，一些大公司将数据分析和数据挖掘工具及其有关技术组合起来，形成所谓的商业智能软件 BIS。其中 SAS 公司的 Enterprise Miner 就是将数据源、数据预处理、数据存储、数据分析与发掘、信息表示与应用等技术结合形成一个复杂的数据挖掘系统。

IBM 公司更全面地考虑了 BI 系统的结构和功能，与其他公司共同合作开发了各类 BI 软件和工具。开发 BI 软件需要从多方面加以考虑。首先必须有一个良好的数据库，为了能使企业管理与决策机制覆盖管理与决策的全阶段，IBM 提出了一个统一的数据库系统——DB2 和一个可视化数据仓库（Visual Data Warehouse，VDW）。它可以将各种应用和各部门的信息融为一体，利用可视化仓库联机分析处理（Online Analytical Processing，OLAP）工具可以生成实时报告。在信息发现和数据发掘工具方面，提出能对结构型和非结构型数据进行挖掘的一整套智能矿工家族。由于 BI 手段只有在好的数据基础上才能见效，因此 IBM 公司提出数据重组工具。由于向用户提供言之有据的信息是做出正确决策的前提，因此 IBM 公司又提出能支持异形数据库的 DataJointer（数据接合）。BI 系统是从数据到知识再到决策的进程中更深入的一步，展示了真正实用的智能信息系统的雏形。

1.4 数据挖掘的对象和常用方法

1.4.1 数据挖掘的对象

数据挖掘与传统的数据分析（如查询、报表、联机应用分析）的本质区别是数据挖掘是在没有明确假设的前提下去挖掘信息、发现知识的，因此数据挖掘所得到的信息应具备先前未知性、有效性和实用性三个特征。

先前未知的信息是指该信息是预先未曾预料到的，即数据挖掘要发现那些不能靠直觉发现的信息或知识，甚至是违背直觉的信息或知识，挖掘出的信息越是出乎意料，就可能越有价值。

数据挖掘可以针对任何类型的数据库进行，既包括传统的关系数据库，也包括非数据库组织的文本数据库、Web 数据库以及复杂的多媒体数据库等[5]。

1. 关系数据库

关系数据库具有坚实的数据基础、统一的组织结构、完整的规范化理论和一体化的查询语言等优点，是当前数据挖掘最重要、最流行、信息最丰富的数据源，是人们进行数据挖掘研究的主要形式之一。

2. 数据仓库

数据仓库是数据库技术发展的高级阶段，它是面向主题的、集成的、内容相对稳定的、随时间变化的数据集合，可以用来支持管理决策的制定。数据仓库允许将各种应用系统、多个数据库集成在一起，为统一的历史数据分析提供坚实的平台。

数据挖掘需要有良好的数据组织和“纯净”的数据，数据的质量直接影响到数据挖掘的效果，而数据仓库的特点恰恰最符合数据挖掘的要求。它从各类数据源中抓取数据，经过

清洗、集成、选择、转换等处理，为数据挖掘所需要的高质量数据提供了保证。可以说，数据仓库为数据挖掘准备了良好的数据源，数据挖掘为数据仓库提供了有效的分析处理手段。因此，随着数据仓库与数据挖掘的协调发展，数据仓库必然成为数据挖掘的最佳环境。

3. 文本数据库

文本数据库所记载的内容均为文字，这些文字并不是简单的关键词，而是长句子、段落甚至全文。文本数据库多数为非结构化的，也有些是半结构化的，如 HTML、E－mail等。Web 网页也是文本信息，由众多的 Web 网页组成的数据库就是最大的文本数据库。当然，如果文本数据具有良好的结构，也可以使用关系数据库来实现。

4. 复杂类型的数据库

以复杂类型的数据库是指非单纯文本的数据库或能够表示动态序列数据的数据库，主要有以下几类[6]。

（1）空间数据库：主要指存储空间信息的数据库，其中数据可能以光栅格式提供，也可能用矢量图形数据表示。例如，地理信息数据库、卫星图像数据库、城市地下管道、下水道以及各类地下建筑分布数据库等。对空间数据库的挖掘可以为城市规划、生态规划、道路修建提供决策支持。

（2）时序数据库：主要用于存放与时间相关的数据，它可以用来反映随时间变化的即时数据或不同时间发生的不同事件，例如连续存放即时的股票交易信息、卫星轨道信息等。对时序数据的挖掘可以发现数据随时间的发展趋势、事物的演变过程和隐藏属性，这些信息对事件的计划、决策和预警是非常有用的。

（3）多媒体数据库：主要指用于存放图像、声音和视频信息的数据库。随着多媒体技术的发展以及相关研究（如可视化信息检索、虚拟现实技术）的进步，多媒体数据库也逐渐普及并应用于许多重要研究领域。目前，多媒体数据的挖掘主要集中在对图像数据的检索和匹配上，随着研究的深入将会拓展到对声音、视频信息的挖掘。

1.4.2 数据挖掘的常用方法

常用的数据挖掘方法有四大类，分别对应四个问题，这四个问题是数据挖掘的基础，分别是聚类挖掘、分类挖掘、关联模式挖掘和异常值检测。这四个问题很重要，因为它们涵盖了表示数据矩阵条目之间不同种类的正面、负面、监督或无监督关系的详尽情景。这些问题也以各种方式相互关联。

1. 分类技术

从分类问题的提出至今，已经衍生出很多具体的分类技术，下面介绍数据挖掘中四种最常用的分类技术。本章节尽量用简单易理解的语言来表述这些技术，之后的章节中我们会再次给读者详细讲解各种算法和相关原理。

在学习这些算法之前必须清楚一点，分类算法不会百分百准确。每个算法在测试集上的运行都会有一个准确率的指标。使用不同的算法建模的分类器（Classifier），在不同的数据集上也会有不同的表现[7]。

1）*K* 最近邻分类算法

K 最近邻（K－Nearest Neighbor，KNN）分类算法可以说是整个数据挖掘分类技术中最简单的方法。所谓 *K* 最近邻，就是 *K* 个最近的邻居，说的是每个样本都可以用它最接近的 *K*

个邻居来代表。

我们用一个简单的例子来说明 KNN 算法的概念。如果您住在一个市中心的住宅内，周围若干个小区的同类大小房子售价都在 280 万 ~ 300 万元，那么我们可以把您的房子和它的近邻们归类到一起，估计售价在 280 万 ~ 300 万元。同样，您的朋友住在郊区，他周围的同类房子售价都在 110 万 ~ 120 万元，那么他的房子和近邻的同类房子归类之后，售价也在 110 万 ~ 120 万元。

KNN 算法的核心思想是如果一个样本在特征空间中 K 个最近邻的样本中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别的特性。该方法在确定分类决策上只依据最邻近的一个或者几个样本的类别来决定待分类样本所属的类别。KNN 算法在类别决策时，只与极少量的相邻样本有关。由于 KNN 算法主要靠周围有限的邻近样本，而不是靠判别类域的方法来确定所属类别，因此对于类域的交叉或重叠较多的待分样本集来说，KNN 算法比其他方法更为适合。

2）决策树

如果说 KNN 是最简单的算法，那决策树应该是最直观最容易理解的分类算法。最简单的决策树的形式是 if – then（如果 – 就）决策方式的树形分叉。

决策树上的每个结点要么是一个新的决策结点，要么是一个代表分类的叶子，而每一个分支则代表一个测试的输出。决策结点上做的是对属性的判断，而所有的叶子结点属于一类。决策树要解决的问题就是用哪些属性充当这棵树的各个结点，其中最关键的是根结点，在它的上面没有其他结点，其他所有的结点都是它的后继结点。

大多数分类算法（如下面介绍的神经网络、支持向量机（SVM）等）都是一种类似于黑盒子式的输出结果，你无法搞清楚具体的分类方式，而决策树让人一目了然，十分方便。决策树按分裂准则的不同可分为基于信息论的方法和最小 GINI 指数方法。

3）神经网络

在 KNN 算法和决策树算法之后，下面介绍神经网络[8]。神经网络就像一个爱学习的孩子，你教他的知识他不会忘记，而且会学以致用。我们把学习集中的每个样本输入到神经网络中，并告诉神经网络输出应该是什么分类。在全部学习集都运行完成之后，神经网络就根据学习集构建好了神经网络模型，构建模型的过程可以看作一个黑盒。然后，我们就可以把测试集中的测试例子用神经网络来分别做测试，如果测试通过（如 80% 或 90% 的正确率），那么神经网络就构建成功了，就可以用这个神经网络来判断事物的类别。

神经网络是通过对人脑的基本单元——神经元的建模和连接，探索模拟人脑神经系统功能的模型，这种模型是一种具有学习、联想、记忆和模式识别等智能信息处理功能的人工系统。神经网络的一个重要特性是它能够从环境中学习，把学习的结果分别存储于网络的突触连接中。神经网络的学习是一个过程，在其所处环境的激励下，相继给网络输入一些样本模式，并按照一定的规则（学习算法）调整网络各层的权值矩阵，网络各层权值都收敛到一定值时学习过程结束。

4）支持向量机

与上面的三种算法相比，支持向量机算法可能会有一些抽象。因此，可以这样理解，尽量把样本中从更高的维度看起来在一起的样本分为一类。例如，在一维（直线）空间里的样本从二维平面上可以把它们分成不同类别，而在二维平面上分散的样本如果我们从三维