

一本使用Python进行机器学习的入门实战教程
助你成为一名机器学习的“老司机”和人工智能的开发者

Python

余本国 孙玉林 著

在机器学习中的应用

- **内容系统全面**：介绍并实现机器学习的经典和主流算法
- **原理浅显易懂**：使用循序渐进的方法阐述机器学习概念
- **配套视频教程**：提供视频教学讲解机器学习的理论与实践
- **应用实例代码**：使用 Python 3.6.X 代码来实现应用案例
- **结合丰富数据**：结合实际数据来帮助对算法的理解和应用



中国水利水电出版社

www.waterpub.com.cn

Python

在机器学习中的应用

余本国 孙玉林 著



中国水利水电出版社

www.waterpub.com.cn

· 北京 ·

内 容 提 要

随着大数据的兴起, Python 和机器学习迅速成为时代的宠儿。本书在内容编排上避免了枯燥的理论知识讲解, 依循“理论简述——实际数据集——Python 程序实现算法”分析数据的思路, 根据实际数据集的分析目的, 采用合适的主流机器学习算法来解决问题。全书共 12 章, 其中第 1 ~ 4 章介绍了机器学习的基础知识; 第 5 ~ 12 章讨论了在面对不同的数据时, 如何采用一些主流的算法来解决问题, 主要包括回归分析、关联规则、无监督学习、文本 LDA 模型、决策树和集成学习、朴素贝叶斯和 K 近邻分类、支持向量机和神经网络, 以及深度学习入门等内容。针对每个算法, 都给出 Python 代码实现算法建模的过程, 并结合可视化技术, 帮助读者更好地理解算法和分析结果。

《Python 在机器学习中的应用》是使用 Python 进行机器学习的入门实战教程, 可作为以 Python 为基础进行机器学习的本科生和研究生入门书籍, 也可供对 Python 机器学习感兴趣的研究人员参考阅读。

图书在版编目 (CIP) 数据

Python 在机器学习中的应用 / 余本国, 孙玉林著. —北京:
中国水利水电出版社, 2019.6
ISBN 978-7-5170-7483-0

I. ① P… II. ①余…②孙… III. ①软件工具—程序设计
IV. ① TP311.561

中国版本图书馆 CIP 数据核字 (2019) 第 031733 号

书 名	Python在机器学习中的应用 Python ZAI JIQI XUEXI ZHONG DE YINGYONG
作 者	余本国 孙玉林 著
出版发行	中国水利水电出版社 (北京市海淀区玉渊潭南路1号D座 100038) 网址: www.waterpub.com.cn E-mail: zhiboshangshu@163.com
经 售	电话: (010) 62572966-2205/2266/2201 (营销中心) 北京科水图书销售中心 (零售) 电话: (010) 88383994、63202643、68545874 全国各地新华书店和相关出版物销售网点
排 版	北京智博尚书文化传媒有限公司
印 刷	三河市龙大印装有限公司
规 格	170mm×230mm 16开本 21印张 422千字 2插页
版 次	2019年6月第1版 2019年6月第1次印刷
印 数	0001—5000册
定 价	79.80元

凡购买我社图书, 如有缺页、倒页、脱页的, 本社营销中心负责调换
版权所有·侵权必究

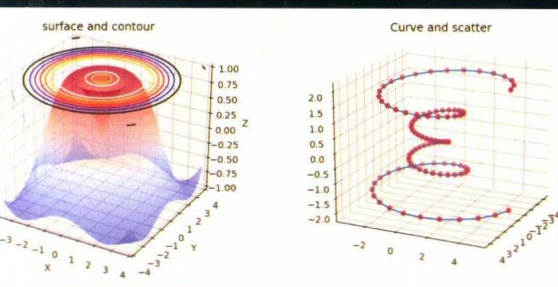


图2-15 三维数据可视化



◆ 图3-9 《红楼梦》词云

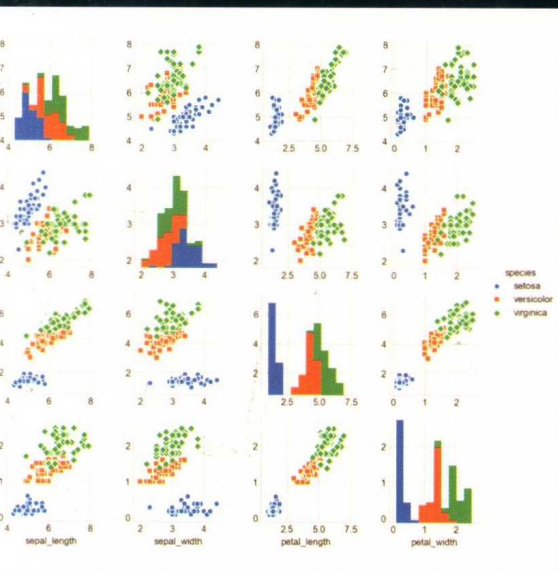
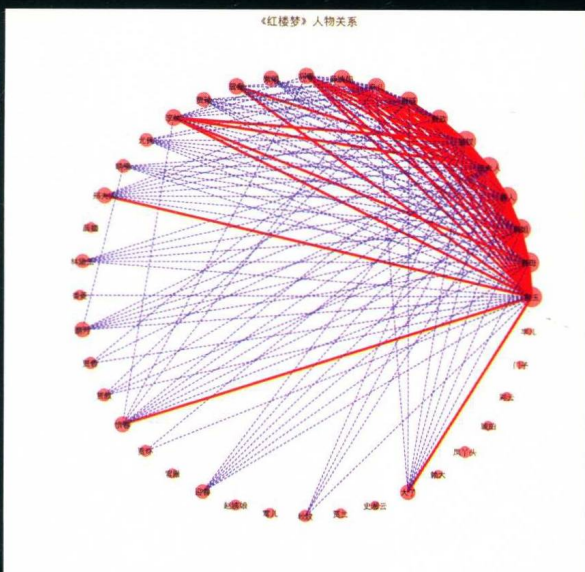


图3-6 鸢尾花数据的矩阵散点图



◆ 图3-10 《红楼梦》部分人物关系图

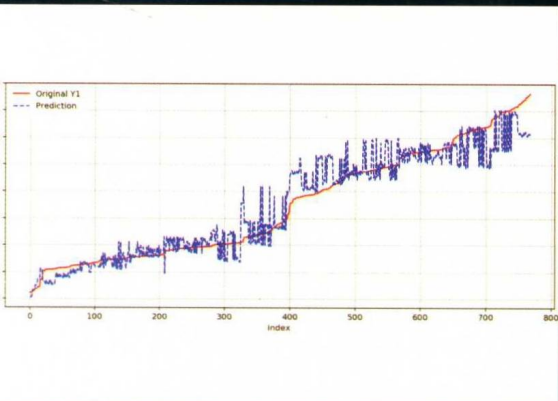
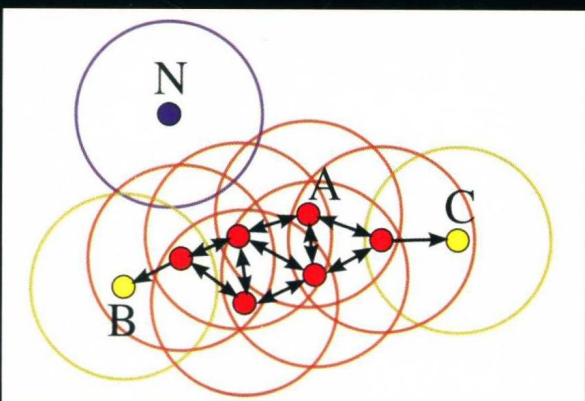


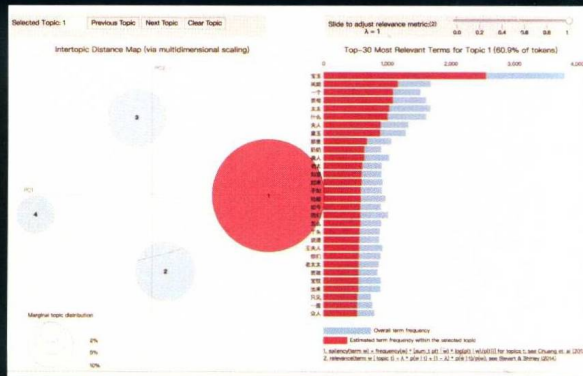
图5-3 多元回归预测结果图



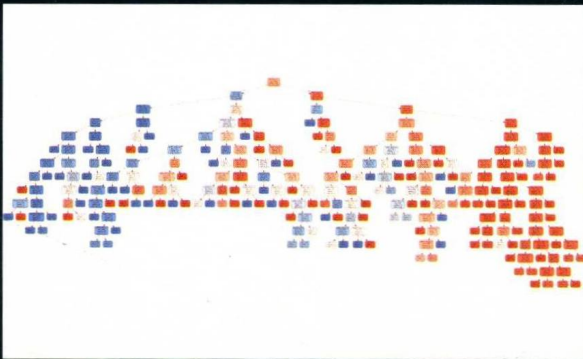
◆ 图7-1 核心点示意图



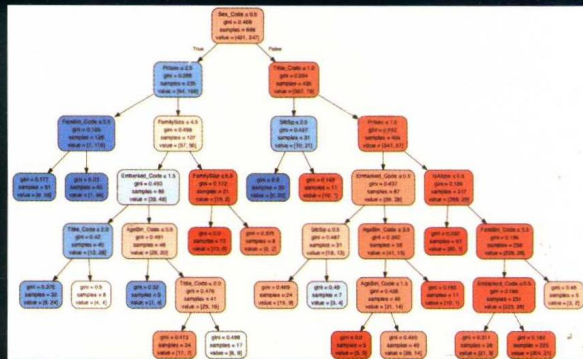
◆ 图7-7 密度聚类结果



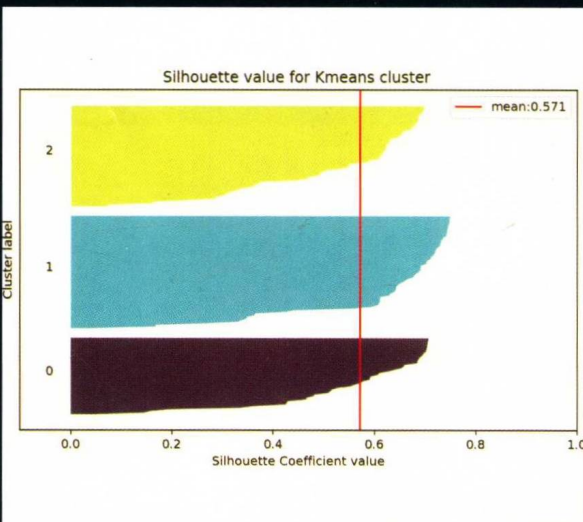
◆ 图8-1 《红楼梦》4个主题



◆ 图9-4 决策树图



◆ 图9-5 剪枝后决策树模型



◆ 图7-6 K-均值聚类轮廓图



◆ 图12-19 人脸特征点检测



前言

Preface

自从 2016 年 AlphaGo 战胜人类围棋顶尖高手后，机器学习、深度学习“忽如一夜春风来”，在互联网上迅速走红，成为民众茶余饭后讨论最多的话题，当然我也不例外。不过很多人可能苦于不知如何下手，或者考虑到算法中的数学知识，而产生了放弃的念头。对此在《基于 Python 的大数据分析基础及实战》一书中做过解释，“很多人对数据分析有畏难心理，主要是因为怕用到很多的数学知识，所以一提到数学估计很多人连勇气都没有了，直接放弃逃跑。”有鉴于此，我在完成《基于 Python 的大数据分析基础及实战》之后，与我的学生孙玉林共同完成了这本书的初稿，剔除了枯燥乏味的数学原理以及推导过程，用浅显易懂的 Python 代码去实现这些经典和主流算法。虽然市面上已经有许多机器学习的书籍，但它们大多要么过于偏重理论，要么过于偏重应用，要么过于“厚重”。《Python 在机器学习中的应用》一书致力于将理论与实践相结合，在讲述理论的同时，避免复杂的数学公式推导，而利用 Python 这一简明有力的编程语言，进行一系列的实践与应用。

本书共分 12 章，前 4 章主要介绍基于 Python 进行机器学习的预备知识，后面 8 章分模块地介绍了机器学习的主流算法和经典应用。本书内容全面、系统，讲解



循序渐进，案例经典实用，代码清晰易懂，只需具备一些基本的 Python 程序设计语言基础，跟随本书学习并多加实践，相信你也能成为一名机器学习的“老司机”和人工智能的开发者。

由于计算机技术的迅猛发展，书中的疏漏及不足之处在所难免，敬请广大读者批评指正、不吝赐教。也欢迎加入 QQ 群一起交流，QQ 群号：25844276。

Contents





第 1 章 机器学习简介	1
1.1 机器学习的任务	2
1.2 机器学习的三种方式	3
1.3 机器学习系统的建立	8
1.4 机器学习实例	9
第 2 章 Python 常用库介绍	18
2.1 Python 的安装 (Anaconda)	19
2.1.1 Spyder	22
2.1.2 Jupyter Notebook.....	23
2.2 Python 常用库.....	26
2.2.1 Numpy 库.....	27
2.2.2 Pandas 库.....	32
2.2.3 Matplotlib 库.....	37
2.2.4 Statsmodels 库	45
2.2.5 Scikit-learn 库.....	47
2.3 其他 Python 常用的数据库	48
2.4 Python 各种库在机器学习中的应用	49
第 3 章 数据的准备和探索	52
3.1 数据预处理.....	53
3.2 数据假设检验	59
3.3 数据间的关系	65
3.4 数据可视化.....	69
3.5 特征提取和降维	79
第 4 章 模型训练和评估	90
4.1 模型训练技巧	91
4.2 分类效果的评价	98
4.3 回归模型评价	102
4.4 聚类分析评估	104

第 5 章 回归分析	108
5.1 回归分析简介	109
5.2 多元线性回归分析	111
5.2.1 多元线性回归	111
5.2.2 逐步回归	114
5.3 Lasso 回归分析	118
5.4 Logistic 回归分析	122
5.5 时间序列预测	125
第 6 章 关联规则	134
6.1 关联规则简介	135
6.2 使用关联规则找到问卷的规则	136
6.3 关联规则可视化	142
第 7 章 无监督学习	147
7.1 无监督学习介绍	148
7.2 系统聚类	152
7.3 K-均值聚类	155
7.4 密度聚类	160
7.5 Mean Shift 聚类	163
7.6 字典学习图像去噪	165
第 8 章 文本 LDA 模型	175
8.1 文本分析简介	176
8.2 中文分词	177
8.3 LDA 主题模型分析《红楼梦》.....	179
8.4 红楼梦人物关系	185
第 9 章 决策树和集成学习	194
9.1 模型简介	195
9.2 泰坦尼克号数据预处理	198
9.3 决策树模型	204



9.4 决策树剪枝.....	207
9.5 随机森林模型	210
9.6 AdaBoost 模型	215
第 10 章 朴素贝叶斯和 K 近邻分类	221
10.1 模型简介.....	222
10.2 垃圾邮件数据预处理	224
10.3 贝叶斯模型识别垃圾邮件	227
10.4 基于异常值检测的垃圾邮件查找	233
10.4.1 PCA 异常值检测	234
10.4.2 Isolation Forest 异常值检测	236
10.5 数据不平衡问题的处理	238
10.6 K 近邻分类.....	239
第 11 章 支持向量机和神经网络.....	252
11.1 模型简介.....	253
11.2 肺癌数据可视化	256
11.3 支持向量机模型	259
11.4 全连接神经网络	264
第 12 章 深度学习入门.....	278
12.1 深度学习介绍	279
12.2 卷积和池化.....	281
12.3 CNN 人脸识别.....	290
12.4 CNN 人脸检测.....	303
12.5 深度卷积图像去噪	309
12.5.1 空洞卷积.....	309
12.5.2 图像与图像块的相互转换.....	310
12.5.3 一种深度学习去噪方法.....	312

Chapter

01

第1章

机器学习简介

进入 21 世纪之后，计算机技术的发展给人们的生产和生活带来了巨大的变化。计算机的迅速普及也使机器学习得到了快速的发展，计算能力强大的计算机和大量的数据集在多种算法的应用下，对各行各业的发展产生了巨大的影响。如今已进入大数据时代，而掌握机器学习技术的人，则是大数据时代的“弄潮儿”，不断带给人们一个又一个惊喜。同时这也是我们进入机器学习的最佳时机，利用机器学习算法结合各种开源库，你也能成为时代的宠儿，每个人都可以利用机器学习算法书写属于自己的篇章。

本章主要介绍机器学习的概念、算法的类型，以及如何建立机器学习系统等方面的内容。



1.1 机器学习的任务

面对不同的问题，可以有多种不同的解决方法，如何使用合适的机器学习算法去完美地解决问题，是一个需要经验与技术的过程，而这些都是建立在对机器学习的各种方法有了充分了解的基础之上。

1. 机器学习的定义

什么是机器学习？简单地说，机器学习是指计算机程序随着经验的积累而自动提高性能，使系统自我完善的过程。换句话说，可以认为机器学习是一个从大量的已知数据中，学习如何对未知的新数据进行预测，并且通过对学习内容的增加（如已知训练数据的增加），提高对未来数据预测的准确性的过程。

可以发现，数据是决定机器学习的一个因素，数据量爆炸式增长是机器学习快速发展的原因之一。数据有多种形式，如传统的数据库、数据表格等结构化数据，以及图片、视频、音频、文本等非结构化数据，都是可以学习的对象。机器的另一重要因素就是算法，正因为研究者们提出了各种各样的算法，才得以对各种形式的的数据以不同的目标进行挖掘。例如，基于贝叶斯模型的垃圾邮件分类、基于关联规则的商品推荐、基于深度学习的图像识别等。也可以说，针对某些问题可能无法找到最好的算法，但总可以找到合适的算法。

2. 学习问题的描述

对于学习问题可以简单地描述为：针对要解决的任务 T 和它所对应的性能度量 P ，如果计算机程序能够因为数据量 D 的增加（即经验 E 的积累）而不断地自我提高性能，那么可以认为该程序在进行机器学习。需要注意的是，针对学习问题应用的算法应该在可接受的时间内，从数据中学习到的有效的结果，这样的学习过程才是有意义的。通常，为了较好地定义一个学习问题，我们需要明确任务的种类、衡量任务学习能力提高的标准、经验的来源 3 个特征。

例如，针对手写数字 0 ~ 9 的识别学习问题。

- 任务 T ：识别图像中的手写数字 0 ~ 9。

- 性能度量标准P：识别的正确率，判断正确的百分比。
- 训练经验E：已经带有标签的手写数字图像数据集D。

这里对学习的定义很宽泛，基本包括了在机器学习领域会遇到的各种问题，如自动驾驶、购物篮分析、欺诈检测、语音识别等。

1.2 机器学习的三种方式

机器学习方法中，根据它们学习的方式不同，可以简单地归为3类：无监督学习 (Unsupervised Learning)、有监督学习 (Supervised Learning) 和半监督学习 (Semi-supervised Learning)，如图 1-1 所示。接下来将详细地介绍这3种方法之间的区别和相关应用。



图 1-1 3 种机器学习类型

1. 无监督学习

无监督学习和其他两种学习方式的主要区别是：无监督学习不需要提前知道数据集的类别标签。常用的无监督学习算法有各种聚类算法（如 K-均值聚类、系统聚类等）、数据降维（如主成分分析）等。

（1）数据聚类发现数据的类别。

聚类是一个把数据对象集划分为多个组或簇的过程。簇内对象不仅具有很高的相似性，还要和其他簇的对象有明显区别。即使在相同的数据集上，使用不同的聚类算法，也可能产生不同的聚类结果。因为聚类分析在分为不同的簇时，不需要提前知道每个数据的类别标签，所以整个聚类过程是无监督的。

聚类分析已经在许多领域得到了广泛的应用，包括商务智能、图像模式识别、Web 搜索等。尤其是在商务领域中，聚类可以把大量的客户划分为不同的组，各



组内的客户具有相似的特性，这对商务策略的调整、产品的推荐、广告投放等是有利的。

现有的聚类算法有很多种，如基于划分方法的K-均值聚类、K中值聚类；基于层次方法的层次聚类、概率层次聚类；基于密度划分方法的高密连通区域算法(DBSCAN算法)、基于密度分布的聚类等。

图1-2展示了基于两个特征PC1和PC2在聚类算法下的类别归属情况。图中的数据点被分成了3类，分别为红色、绿色、蓝色。

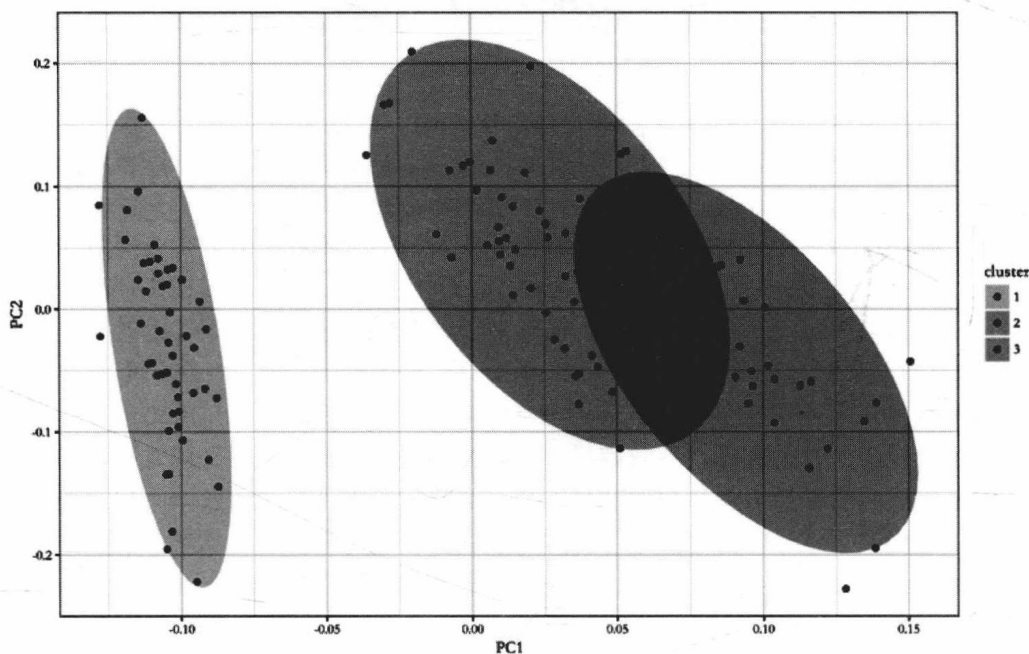


图1-2 数据聚类

(2) 数据降维减少数据的维度。

在机器学习中，数据降维是无监督学习中的另一个领域。数据降维是指在某些限定条件下，降低数据集特征的个数，得到一组新的特征的过程。在大数据时代，通常数据都是高维的(每一个样例都会有很多特征)，高维数据往往会带有冗余信息，而数据降维的一个重要作用就是去除冗余信息，保留必要的信息。如果数据维度过高，会大大拖慢算法的运行速度，此时就体现出了数据降维的重要性。数据降维的算法有很多，如主成分分析(PCA)是通过正交变换，将原来的特征进行线性组合生成新的特征，并且只保留前几个主要特征的方法；核主成分分析(KPCA)则是

基于核技巧的非线性降维的方法；而流形学习则是借鉴拓扑结构的数据降维方法。

数据降维对数据的可视化有很大的帮助。通常我们很难看到高维数据之间的依赖和变化关系，通过数据降维可以将数据投影到二维或三维空间，能够更加方便地观察数据之间的变化趋势。如图 1-3 所示，人脸图像经过 PCA 降维到二维空间后，通过各个人脸在空间中的位置分布，可以发现不同类型的图片在空间中的分布是有规律的。正是通过降维可视化，我们发现了这种规律，以方便后续的学习与研究。

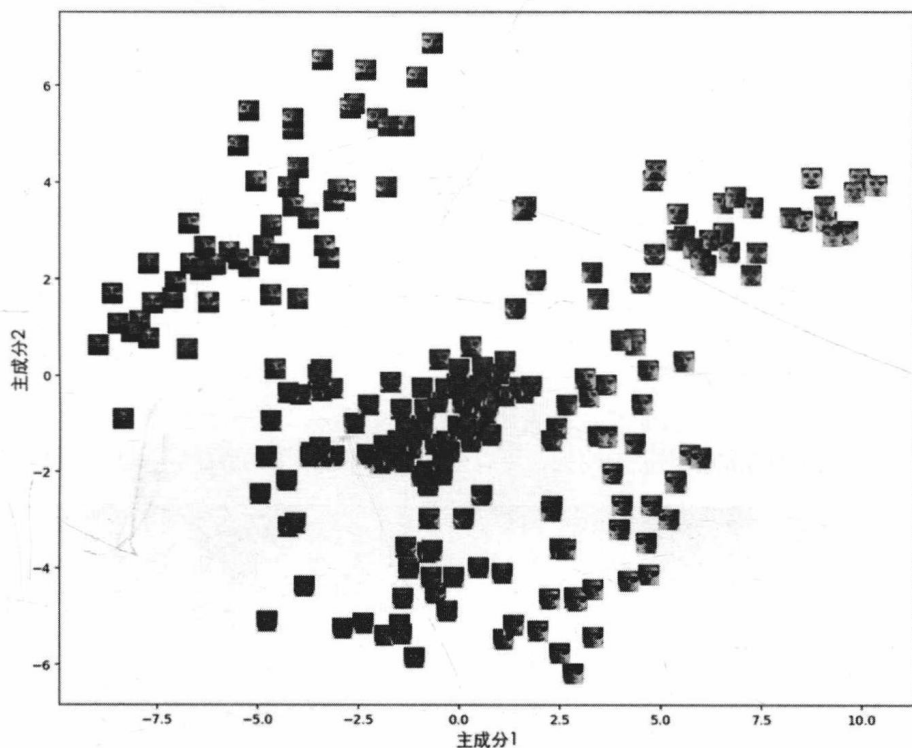


图 1-3 图像降维可视化

2. 有监督学习

有监督学习的主要特性是：使用有标签的训练数据来建立模型，用来预测新的未知标签的数据。用来指导模型建立的标签可以是类别数据、连续数据等。相应地，如果标签是可以分类的，如 0 ~ 9 手写数字的识别、判断是否为垃圾邮件等，这样的有监督学习称为分类；如果标签是连续的数据，如身高、年龄、商品的价格等，这样的有监督学习称为回归。图 1-4 展示了有监督学习的一般过程。

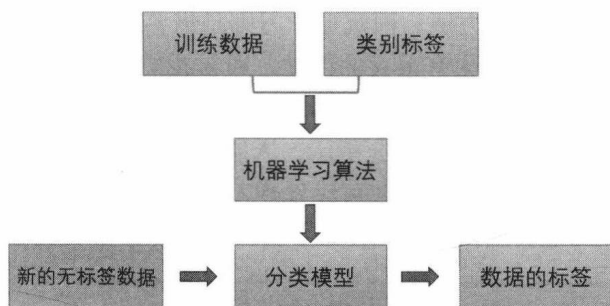


图 1-4 有监督学习的一般过程

(1) 分类。

分类是最常见的有监督学习方式之一。如果数据的类别只有两类——是或否 (1 或 0)，则这类问题称为二分类问题。常见的有是否存在欺诈、是否为垃圾邮件、是否患病等问题。二分类常用的算法有朴素贝叶斯算法 (常用于识别是否为垃圾邮件)、逻辑斯蒂回归算法等。如果数据的标签多于两类，这类情况常常称为多分类问题，如人脸识别、手写字体识别等。在多分类中常用的方法有神经网络、K 近邻、随机森林、深度学习等算法。

图 1-5 展示的是二维空间中 (主成分 1 和主成分 2) 两类数据被一条空间曲线分为两类的示例。如果有新的数据被观测到，可以根据它在平面中的位置确定其应属的类别。

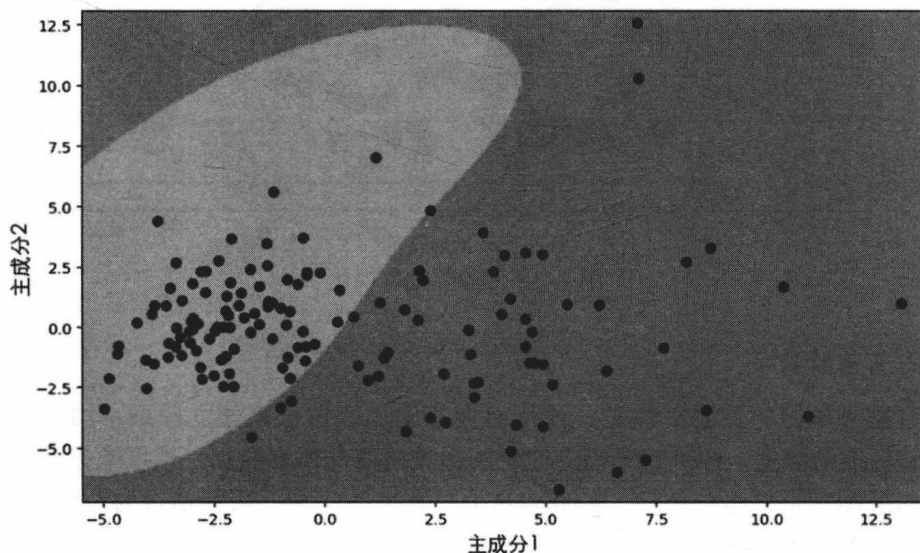


图 1-5 二分类问题示意图