普通高等院校数据科学与大数据技术专业“十三五”规划教材

数据科学与数学建模

郝志峰 ■ 主编

SHUJU KEXUE YU SHUXUE JIANMO

華中科技大學出版社
http://www.hustp.com

普通高等院校数据科学与大数据技术专业“十三五”规划教材

数据科学与数学建模

主　编　郝志峰
副主编　李　杨　刘小兰　廖　芹

华中科技大学出版社
中国·武汉

内 容 简 介

本书内容分为8章，基本涵盖了目前较为常用的数据科学建模方法，包括现在热门的深度学习。书中不仅介绍了模型的理论基础，还将大量案例与现实数据结合，为读者展示数据分析中常见任务的处理流程，如分类、回归、聚类、推荐、图片识别等，帮助读者应用这些模型和方法解决实际问题。

第1章首先对数据科学的任务和重要性进行了概述，接着介绍数据科学的建模流程以及Python语言开发环境与常用库；第2章介绍了回归模型，包括线性回归和逻辑回归模型；第3章介绍了聚类模型，包括K-means算法、DBSCAN算法和DIANA算法；第4章介绍了关联规则分析，包括Apriori算法和FP-Growth算法；第5章介绍了决策树模型，包括ID3、C4.5和CART算法及树的剪枝方法；第6章介绍了支持向量机，包括线性和非线性支持向量机以及支持向量机的求解与多分类问题；第7章介绍了贝叶斯网络，包括朴素贝叶斯网络、TAN贝叶斯网络和无约束贝叶斯网络；第8章介绍了深度学习，包括卷积神经网络和循环神经网络。

图书在版编目(CIP)数据

数据科学与数学建模/郝志峰主编. —武汉：华中科技大学出版社，2019.1
ISBN 978-7-5680-4935-1

Ⅰ. ①数… Ⅱ. ①郝… Ⅲ. ①数据模型-高等学校-教材 Ⅳ. ①TP311.13

中国版本图书馆CIP数据核字(2019)第020601号

数据科学与数学建模 郝志峰 主编
Shuju Kexue yu Shuxue Jianmo

策划编辑：李 露 廖佳妮
责任编辑：李 露
封面设计：原色设计
责任校对：李 弋
责任监印：赵 月
出版发行：华中科技大学出版社(中国·武汉) 电话：(027)81321913
武汉市东湖新技术开发区华工科技园 邮编：430223
录 排：华中科技大学惠友文印中心
印 刷：武汉市洪林印务有限公司
开 本：787mm×1092mm 1/16
印 张：10.75
字 数：262千字
版 次：2019年1月第1版第1次印刷
定 价：26.80元

前言

PREFACE

数据科学与大数据技术专业作为一个热门专业，近年来引起了相关高校的关注，不少高校纷纷设立此专业。数据科学与大数据技术专业在2016年仅有3所高校（北京大学、对外经济贸易大学和中南大学）获批，2017年3月又有32所院校获批，包括编者所在的佛山科学技术学院，2018年又有248所院校获批。在2018年，教育部又设置了大数据管理与应用专业。可以预计，随着大数据与人工智能相关专业的兴起，数据科学课程的教学改革面临着诸多新的问题。而教育部2018年发布的《普通高等学校本科专业类教学质量国家标准》，对以数据科学与大数据技术专业为代表的专业人才培养方案提出了挑战。

编者郝志峰等曾出版《数据挖掘与数学建模》，该教材在华南理工大学应用数学专业（应用软件方向）、信息管理与信息系统专业的本科生教学中使用了近十年，也曾作为中国移动通信广东分公司的管理层培训材料，受到了广泛的欢迎。该教材结合具体的案例，从学习者的角度，渐进式地把大数据挖掘的技术和方法展示出来，使学习者有学习的热情。因此，大数据挖掘与数学建模的教学改革成了新的研究方向。不过，大数据挖掘所呈现出的不确定性，使得建模的价值，包括数学方法建模（简称数学建模）的价值，打了些折扣。本课程从大数据挖掘中，提炼出了科学的、可教学的、有模型的内容，将这些内容整合为一门数据科学与大数据技术专业的基础课呈现出来。这门课程的教材就是《数据科学与数学建模》。

目 录

CONTENTS

第1章 绪 论

1.1 数据科学概述

随着信息化建设的发展以及信息技术的应用，各领域都积累了海量的数据，大数据时代已经来临！而数据科学研究的核心内容就是数据。从结构化程度看，数据可分为结构化数据、半结构化数据和非结构化数据三种。结构化数据主要是指在传统关系型数据库中获取、存储、计算和管理的数据；半结构化数据介于结构化和非结构化数据之间，包括 HTML、XML 等数据；非结构化数据包括自然语言、电子邮件、音频、视频、图像等。仅仅是量大的数据并不能算是大数据，大数据的特征可以用 4 个 V 来描述：数据量大(Volume)、数据种类多(Variety)、价值密度低(Veracity)、速度快时效高(Velocity)。在大数据背景下，数据在人们的生活中必将发挥越来越大的作用。数据科学作为一门蓬勃发展的学科，它关注的是在大数据时代，如何运用与数据相关的技术和理论使数据发挥出更大的价值。

数据科学的研究范畴涉及数学、统计和计算机科学等学科，数学的基础知识包括线性代数、概率论、微积分和计算方法等，计算机科学的基础知识包括数据库、分布式系统、数据可视化技术等。此外，还包括统计学习、机器学习等学科，以及其他领域的知识。

数学建模是人们理解数据的重要途径之一，也是数据科学中的重要工具。数学模型是指在特定目标下，对特定对象的内在变化规律进行特征提取、假设表示、数学应用，从而得到的一个用数学符号表示各量关系的数学结构。统计分析方法是建立数学模型的常用方法。由于统计分析方法是从实验数据切入建模，所以当数据信息特征变化时，可以通过让学习形式变化，从而使模型参数也随之变化，即在一定条件下考虑了自适应数据特征变化以调整模型参数的特点，当采集样本扩大时，要通过对样本的不断学习，逐步调整模型参数，使模型能适应全域样本信息的安全评价。因此，统计分析方法在数学建模中具有更广泛的应用。

数据科学在不同领域得到了广泛的应用，并发挥了巨大的作用。

在商业领域，沃尔玛的“啤酒与尿布”是应用研究商品关联关系的“购物篮分析法”的一个经典案例。除此之外，在推荐场景中，阿里巴巴使用深度强化学习与自适应在线学习，通过持续机器学习和模型优化建立决策引擎，对海量用户行为以及百亿级商品特征进行实时分析，帮助每一个用户迅速发现宝贝，提高人和商品的配对效率，通过个性化推荐提高消费的概率。

在生物医学领域，“谷歌流感预测”是谷歌 2008 年推出的用于预警流感的即时网络服务。与美国疾病控制和预防中心通常需要花费数星期整理并发布流感疫情报告不同，谷歌的流感趋势报告每日更新。谷歌在美国的九个地区做了测试，并且发现它可比疾病控制和预防中心提前 7～14 天准确预测流感的爆发。谷歌的预测依据是汇总过的谷歌搜索数据，

搜索“流感”相关主题的人数与实际患有流感症状的人数之间存在着密切的关系。在医疗服务行业，随着医疗过程的电子化和数据化，医疗数据的增长也出现了井喷，这给“智慧医疗”的发展带来了契机，从临床业务到新药研发，从疾病预警到愈后监控，数据科学都得到了越来越广泛的应用，这些应用降低了医疗成本、改善了患者体验。

在智慧城市领域，瑞典首都斯德哥尔摩的“智能交通”项目，引入了 IBM 的“InfoSphere Streams”流计算平台，通过对装载 GPS 终端的出租车实时回传的位置数据进行实时分析，得出实时的道路拥堵状况，实现了为城区的通行车辆提供回避拥堵路线服务。应用此平台后，城市温室气体排放量减少了 10%，交通拥堵率减少了 20%，居民的外出开车时间也缩短了 50%。在中国，百度公司开发的“百度迁徙”，利用百度地图 LBS（基于位置服务）开放平台、百度天眼，对其拥有的 LBS 大数据进行计算分析，可以直观地展示节假日、小长假期间人口流动的趋势，是智慧城市中以“人群迁徙”为主题的大数据分析与可视化平台。

在影视娱乐领域，美国影视租赁公司 Netflix 在《纸牌屋》这部剧播放之前，就通过对海量的用户数据的分析成功预测了这部剧的走红，提前购买了版权。《纸牌屋》的数据库包含了 3000 万个用户的收视选择、400 万条评论、300 万次主题搜索。最终，拍什么、谁来拍、谁来演、怎么播，都由数千万观众的喜好统计决定。可以说，《纸牌屋》的成功得益于 Netflix 对海量用户数据的积累和分析。

就连围棋这一人类传统强势项目的世界冠军，也被人工智能围棋机器人 AlphaGo 打败。2016 年 3 月，AlphaGo 与围棋世界冠军、职业九段棋手李世石进行围棋人机大战，并以 4∶1 的总比分获胜。AlphaGo 的主要工作原理是深度学习，并结合了监督学习和强化学习的优势，通过训练形成一个策略网络（policy network），将棋盘上的局势作为输入信息，并对所有可行的落子位置生成一个概率分布。然后，训练出一个价值网络（value network）对自我对弈进行预测，以 -1（对手的绝对胜利）到 1（AlphaGo 的绝对胜利）的标准，预测所有可行落子位置的结果。这两个网络自身都十分强大，而 AlphaGo 将这两个网络整合进基于概率的蒙特卡罗树搜索（MCTS）中，实现了它真正的优势。

1.2 数据科学的建模流程

获取数据→数据分析和可视化→数据准备→建立模型→结果评估是进行数据科学建模的基本流程，在应用过程中可根据实际情况增加或减少步骤，有时也会重复进行某几个步骤，需要数据科学工作者依据自身经验和知识具体实施，本书在案例实现中基本遵循这一流程。

1. 获取数据

获取数据是进行数据科学建模的第一步。从很多网站可以获取免费或收费的数据集。例如，UCI 数据集的网站上就有很多面向不同分析任务的公开数据集。网址为 http://archive.ics.uci.edu/ml/datasets.html。

其他数据来源还包括一些数据挖掘竞赛官网、学术论文以及技术博客等，如阿里天池大数据竞赛官网（网址为 https://tianchi.aliyun.com/），在每个赛题中都可以下载相对应的数

据集。

以获取 UCI 数据集为例，在 UCI 网站上，每个数据集所对应的网页会包含以下信息。

- 数据集特征(data set characteristics)：多变量(multivariate)、文本(text)、时间序列(time-series)、空间数据集(spatial)等。
- 属性特征(attribute characteristics)：分类属性(categorical)、实数(real)、整数(integer)等。
- 相关任务(associated tasks)：分类、聚类、回归分析等。
- 实例数目(number of instances)。
- 属性数目(number of attributes)。
- 是否有缺失值(missing values?)。
- 领域(area)。
- 数据捐赠日期(date donated)。
- 网络点击次数(number of web hits)。
- 数据集下载链接。
- 相关的论文成果。

2. 数据分析和可视化

获取数据后，需要对数据进行分析和可视化，以便对数据的分布等特性有一个大致的了解。可以对数据进行具有统计学意义的计算，来获得一些统计学系数。

例如，对鸢尾花数据集的统计结果如表 1.1 所示。

表 1.1　鸢尾花数据集统计结果

	最小值	最大值	均值	标准差	类别相关系数
花萼长度	4.3	7.9	5.84	0.83	0.7826
花萼宽度	2.0	4.4	3.05	0.43	−0.4194
花瓣长度	1.0	6.9	3.76	1.76	0.9490
花瓣宽度	0.1	2.5	1.20	0.76	0.9565

数据可视化也是数据分析中必不可少的工具，可以通过绘制盒须图来可视化一组数据的统计信息，体现最大值、最小值、中位数以及四分位数等。

图 1.1 中标示了盒须图中每条线的含义。

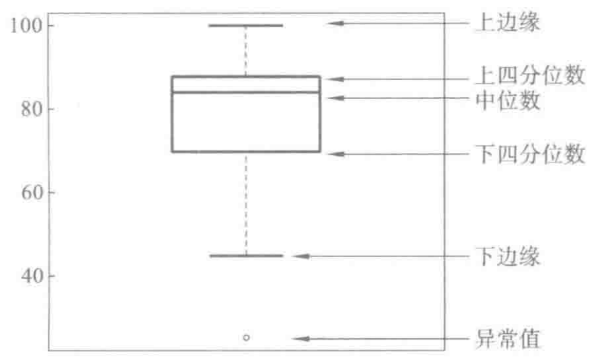

图 1.1　盒须图

长方形盒子的顶部和底部分别是上、下四分位数，上、下四分位数是指数据按值排序后处于25%和75%位置上的数据值。上边缘和下边缘的两条横线分别是最大值和最小值，超出上、下边缘的部分被认为是异常值，在盒子内的数据被认为是大多数数据所在的地方（占50%以上），再结合中位数，可以直观地看出数据的分布情况。

有些情况下，统计信息并不能完整地揭示数据的所有特性。例如，统计学家F. J. Anscombe构造了四组在统计特性上极为相似的数据集(X,Y)，X的均值都为9.0，Y的均值都为7.5；X的方差都为10.0，Y的方差都为3.75；X，Y的相关度都为0.816，线性拟合结果都为$Y=3+0.5X$，仅以统计学指标难以区分四组数据之间的差异。

图1.2利用散点图可视化这四个数据集，在每个散点图中，横坐标表示X，纵坐标表示Y，每一个点代表一个数据，把所有数据点绘制在二维平面上，可以用来观察数据的分布。可以从散点图上明显地看出四组数据在分布上的不同。

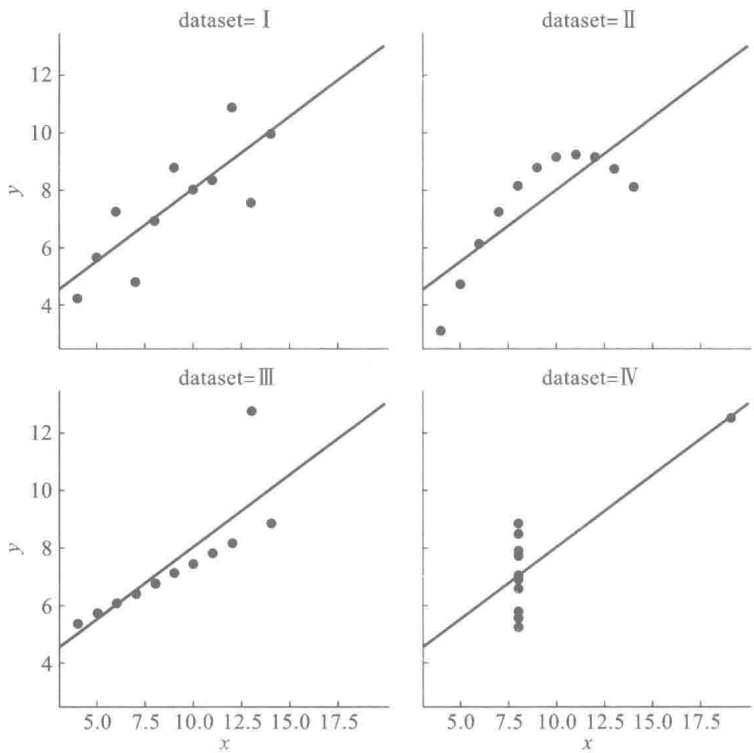

图1.2 四组数据的散点图

散点图矩阵是散点图的扩展，可以用于多维变量的数据可视化，对于n维的数据，采用n^2个散点图逐一表示n个属性之间的两两关系，可揭示数据特定属性上的分布特点。图1.3展示了一个四维数据（100多种花的四个属性）的散点图矩阵实例。

平行坐标（parallel coordinates）采用相互平行的坐标轴，每个坐标轴代表数据的一个属性。对于高维数据，可以利用平行坐标可视化方法，通过变换纵轴的排列顺序，观察数据特征之间的关系，如图1.4（图片来源：http://visual.ly/nutrient-content-parallel-coordinates）所示。

3. 数据准备

对数据的特性进行分析之后，需要对数据进行预处理，例如，利用插值等方法处理数据

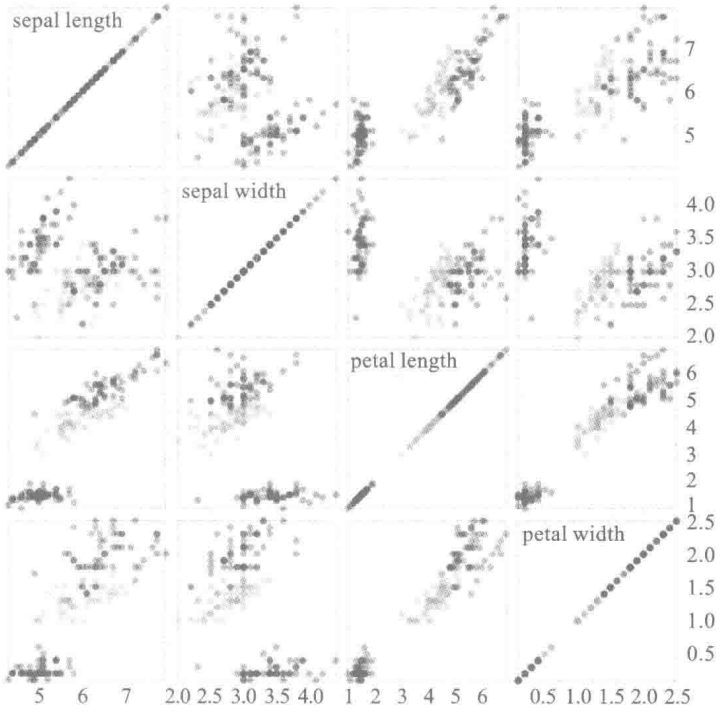

图 1.3　多维数据的散点图

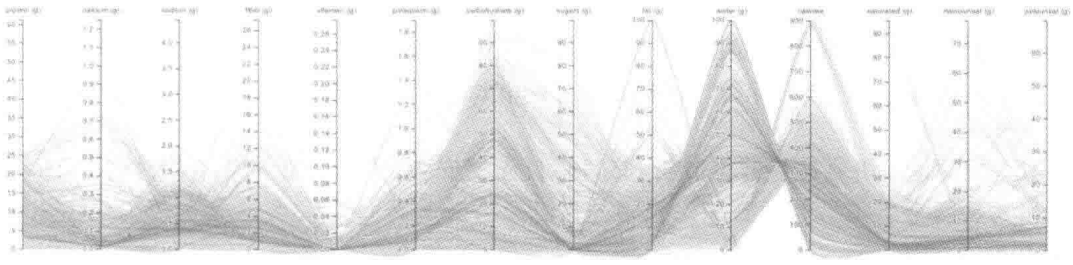

图 1.4　高维数据的平行坐标图

中的缺失特征，利用抽样减少数据集的大小，删除一些异常数据或无用特征，对某些特征进行联合、拼凑、清理等处理。

4. 建立模型

根据数据分析的具体任务、数据的特性(噪声大小、值的分布状况等)，设计合适的模型对数据集建模，并进行参数调优等工作。

5. 结果评估

最后需要对建立的模型进行评估。例如，在分类问题中，模型可以用于预测一条数据最有可能属于哪一类，将预测结果和实际类别进行对比，看看模型预测的分类是否准确，可以得到分类准确率。在回归问题中，对于一条数据，也可对模型预测的值和这条数据的实际值进行对比。为了进行更有说服力的对比，训练数据集和测试数据集最好没有太多的交集。交叉验证法、留出法等是常用的划分训练集、测试集的方法。

交叉验证的基本思想是先将原始数据分组，一部分作为训练集，另一部分作为测试集，

首先用训练集对模型进行训练，再利用测试集对模型的预测结果进行评测，以此来作为评价模型准确率的指标。例如，10 折交叉验证将初始样本等分为 10 个子样本，每一个子样本作为一次测试集，其他 9 个样本作为训练集，重复 10 次，得到 10 组评测指标，将最终的平均值或者加权值作为评测结果。10 折交叉验证的数据划分如图 1.5 所示。

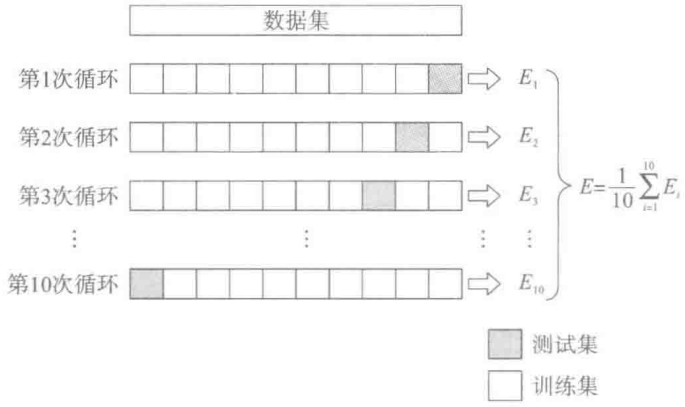

图 1.5 10 折交叉验证的数据划分

1.3 Python 语言开发环境与库入门

1.3.1 开发环境

使用 Python 语言进行数据分析和建模，首先要安装 Python 开发环境。目前 Python 有 2.x 和 3.x 两个版本。考虑到越来越多的人已经由 2.x 转向使用 3.x，本书代码均在 3.x 版本下开发。

以 Windows 操作系统为例，打开 Python 官网 https://www.python.org/downloads/windows/，根据自己的操作系统版本选择可执行安装程序或下载压缩包，下载完成后按提示安装到本机指定路径下，并把该路径加入到 Windows 系统的环境变量 Path 中。

安装 Python 以后，打开 Python 命令行交互环境，进行 Python 代码的编写。打开系统的命令行(cmd.exe)，如果已经配置好系统的环境变量，输入 Python 即可进入与 Python 交互的命令行。如果出现找不到该命令的错误，请先检查自己的环境变量是否配置正确。输入 exit()退出 Python 交互环境，回到系统命令行。

使用 Python 命令行交互环境的优点是可以即时得到运行结果，缺点在于无法进行保存操作。如果需要一段重复运行的代码，命令行交互就显得无能为力了。除了命令行，还可以使用文本编辑器来编写代码，可以把代码写在一个文本文件中，并将其保存为扩展名是.py 的文件。使用 Python 加文件路径的命令即可运行这些代码。

除了命令行和文本编辑器，还可以使用 PyCharm 等集成开发环境(IDE)进行 Python 代码的编写和调试。PyCharm 可以从官网 https://www.jetbrains.com/pycharm/下载，社区版(community)是免费的并且可以满足基本的开发需求。

下载安装完成后打开 PyCharm，PyCharm 以项目为单位进行管理，创建一个 Python 项目，单击 File→New Project，选择 Pure Python，填写项目存放的路径和名字，PyCharm 会扫描系统存在的 Python 解释器，选择安装好的 Python，点击 Create 进行创建，如图 1.6 所示。

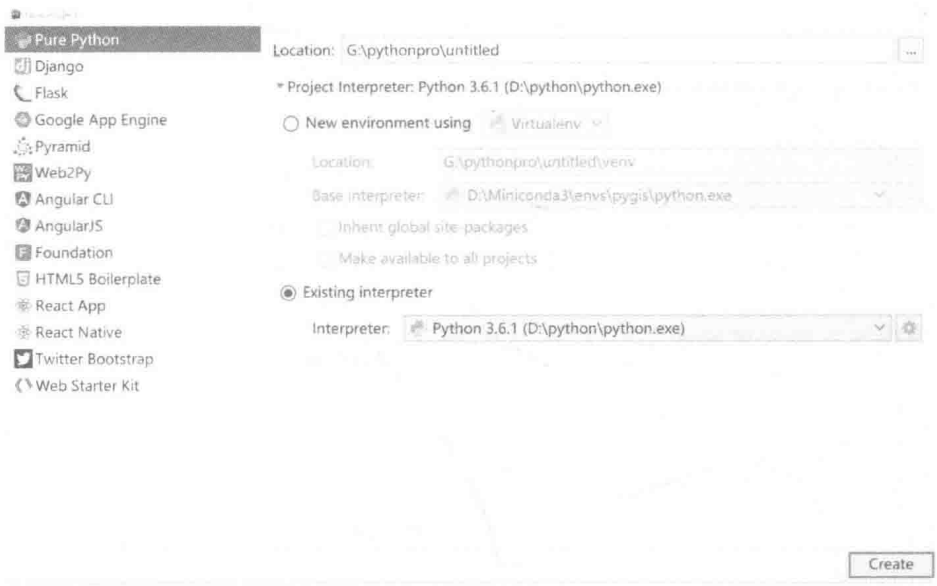

图 1.6　PyCharm 创建项目界面

创建项目后，进入主界面，在左侧的项目导航栏中选择刚刚创建的项目，右击，选择 New→Python File 创建一个文件，并命名为 example.py，如图 1.7 所示。

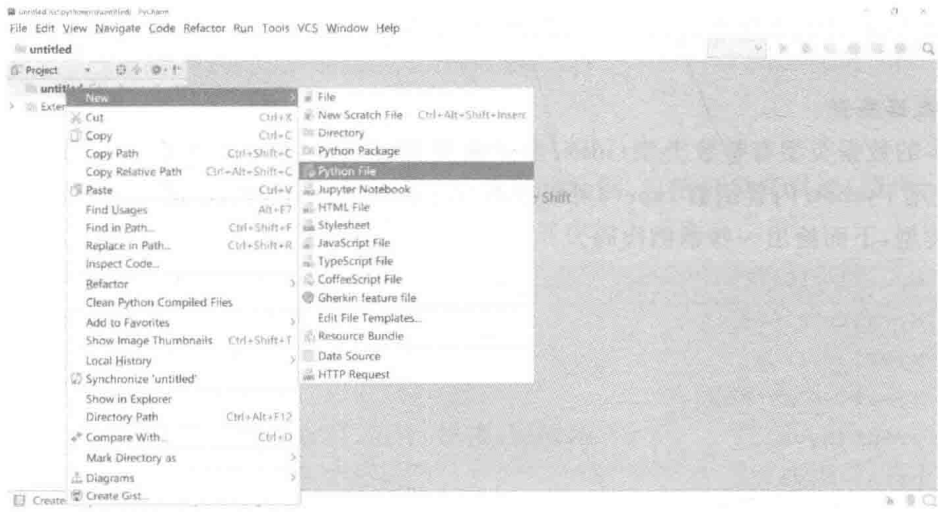

图 1.7　PyCharm 创建文件界面

开始进行代码的编写，如图 1.8 所示。

完成编写后右击编写界面，点击 Run 运行程序，界面下方会输出程序的运行结果，如图 1.9 所示。

图 1.8 PyCharm 编写代码界面

图 1.9 PyCharm 运行结果界面

1.3.2 基本语法

本节介绍 Python 的基本语法。Python 的语法相对简单，变量不需要定义就可以直接使用。

1. 数据类型

基本的数据类型有整数类型(int)、浮点数类型(float)、布尔类型(bool)和字符串类型(str)，使用 Python 内置函数 type()可以查看变量数据类型，使用 print()函数可以打印数据的值或类型，下面给出一些示例代码。

```
Myval1= 10+ 2
Myval2= 2.66* 4
Myval3= True
Myval4= "pythonStr"
print(Myval1)
print(Myval2)
print(Myval3)
print(Myval4)
print(type(Myval4))
print("hello world")
```

2. 运算符

Python 中的运算符与大多数编程语言的运算符类似，包括算术运算符、关系运算符、赋

值运算符、逻辑运算符。下面给出一些示例代码。

```
Myval1= 2+ 2                    # 加法运算
print(Myval1* * 2)              # 指数幂运算
Myval2= 3                       # 赋值运算
print(Myval1/Myval2)            # 除法运算
print(Myval1//Myval2)           # 地板除,删除小数点后的商
Myval3= 2> 1                    # 关系运算
Myval4= 2= = 3                  # 关系运算
Myval5= Myval3|Myval4           # 逻辑运算
print(Myval3)
print(Myval4)
print(Myval5)
```

3. 控制语句和代码块

与C语言中使用{}限定一个代码块不同,Python的语法中利用缩进完成同样的功能,下面给出Python选择和循环语句的例子。

```
a= True
if a:
  print("true value")
else:
  print("false value")

for i in range(5):
  if(i= = 3):
      break                     # 当i等于3,退出循环
  else:
      print(i)
```

4. 容器类型

Python中有两种比较重要的容器类型:List和Dict。

List是一个有序列表,可以用于存放任意类型的对象,包括自定义对象,使用append()函数将对象加到列表尾。下面创建一个List,并向其中加入元素。

```
a= [1,5,3]
a.append(7)
a.append("iu")
print(a)
```

也可以简单地对列表进行索引、切片、拼接,示例如下。

```
a= [1,2,3,4,5]
b= [6,7,8]
print(a[2])
print(a[1:4])
c= a+ b
print(c)
```

可以使用内置函数len()返回容器类型的元素个数,使用for循环可以方便遍历列表

List，示例如下。

```
print(len(c))
for i in c:
  print(i)
```

Dict 则是 Dictionary（字典）的缩写，里面存放的数据是一个个无序的键值对（key/value），下面创建一个 Dict，并加入一些元素。

```
a= {1:"value1",2:"value2"}
a[3]= "value3"
a["test"]= 23
print(a)
```

通过字典名字和 key 可以获得对应的值：

```
print(a["test"])
```

1.3.3 常用库和功能

本节介绍利用 Python 进行数据分析、科学计算时常用的库及其功能。使用 pip 软件的“pip install＋库名”可以进行 Python 第三方库的安装。

1. Numpy

Numpy 常用于进行与线性代数相关的运算，支持高维数组、矩阵等数据类型，并且提供了大量相关的函数。Numpy 内部运算由 C 语言实现，因此运算速度十分快。要想使用它，首先要导入 Numpy 库，一般约定导入时的别名为 np：

```
import numpy as np
```

Numpy 中有两种重要的数据类型，分别是数组和矩阵。首先创建一个二维数组（numpy.ndarray），打印该数组的元素类型、形状、大小：

```
array= np.arange(6).reshape(2,3)
print(array)
print(array.dtype)
print(array.shape)
print(array.size)
```

通过 Numpy 中的数组可以批量进行数据运算，下面对数组与标量进行数学运算，这些运算相当于对数组内的每一个元素都进行运算：

```
array1= array+ 2
array2= array* * 2
print(array1)
print(array2)
```

Numpy 中的数组也可以进行索引和切片。对一维数组而言，索引和切片方式与 Python 中的 List 相同。对于高维数组，可以使用逗号隔开不同维度的索引来访问某个元素：

```
print(array[1,2])
```

如果忽略掉后面的索引，会返回比高维数组维度低一点的数组：

```
print(array[1])
```

当数组为二维数组时，可以把它看作一个矩阵，进行线性代数相关运算，先创建一个二维数组：