《数据科学算法》教学大纲

课程编号：071613B

课程类型：□通识教育必修课 □通识教育选修课

□学科基础课 ☑专业核心课

□专业提升课 □专业拓展课

总学时：48 讲课学时：32 实验（上机）学时：16

学　分：3

考试类型： □考试 ☑考查

适用对象：数据科学与大数据

□是 ☑否 适合作为其他专业学生的个性化选修课

先修课程：程序设计基础、数据结构等

一、教学目标

目标1：立德树人，通过课程思政的方式，培养社会主义事业的建设者。

目标2：通过本课程的学习，要求学生掌握数据科学算法的基本概念和基本原理，熟练掌握一些常用的数据科学算法实用开发方法和技术。

目标3：了解数据科学算法各领域的发展方向，掌握如何用数据科学算法开发软件项目。

目标4：能够利用所学知识去解决一些实际的应用问题（或模拟案例）。

二、教学内容及其与毕业要求的对应关系

**1.对应毕设要求的教学内容：**

（1）对应毕业要求1即为社会主义现代化建设培养人才的要求，本课程的教学内容1为：针对政府机关企事业单位的大数据分析业务需求，培养大数据分析人才，使其具备大数据分析相关的知识与技能。

（2）对应毕业要求2即掌握专业知识，本课程的教学内容2为使学生掌握大数据分析算法的原理与实现，包括大数据的处理、回归分析、分类分析、聚类分析、降维与特征提取、推荐与协同过滤、预测与时间序列预测、神经网路与深度学习、智能系统集成。

（3）对应毕业要求3即掌握专业技能，本课程的教学内容3为使学生能够运用上述专业知识，独立完成大数据分析相关的毕业设计。

（4）对应毕业要求4即具备实践能力，本课程的教学内容4为使学生能够运用上述专业知识，独立完成大数据分析相关的企业真实项目。

（5）对应毕业要求5即具备良好的沟通协作能力，本课程的教学内容5为使学生能够以团队协作的方式完成上述知识、技能、项目实践的过程。

**2.教学方法和手段**

（1）传统课堂+课程思政模式：以真实项目为背景，对理论知识体系进行设计，引发学生兴趣，使学生清晰了解整个知识体系架构，清楚理解原理。（2）教改创新实时互动实践教学：实现与学生高频率PDCA互动教学，提升学生课堂学习效果。（3）线上线下混合教学：在中国大学慕课面向全国发布网课，使学生课前和课后都具有充足的预习和复习资料，与全国同学们互动过程中，提升自己对知识的理解和对技能的掌握，培养学生的乐于助人、为社会做贡献的沟通能力、团队协作能力和管理能力。

**3.学习要求**

1．为有效学习本课程，要求学生首先具备程序设计基础、数据结构等方面的基本知识，较熟练掌握系统开发设计等方面的完整知识。自学时应该熟读大纲，提纲挈领地掌握数据科学算法的内容，随后，按照大纲熟读教材，通过课后思考和练习题进行多角度和层次的反复学习。

2．上机实践，本门课程有很多上机内容，每章的内容都需要借助于算法设计与实现进行学习，建议自学时安装这些软件进行自学，完成书上案例的内容。通过上机学习本课程的理论，掌握数据科学算法。

三、各教学环节学时分配

以表格方式表现各章节的学时分配，表格如下：

**教学课时分配**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 章节内容 | 讲课 | 实验 | 其他 | 合计 |
| 1 | 数据科学算法概述 | 2 | 1 |  | 3 |
| 2 | 时间亚线性算法 | 2 | 1 |  | 3 |
| 3 | 空间亚线性算法 | 2 | 1 |  | 3 |
| 4 | 外存算法概述 | 2 | 1 |  | 3 |
| 5 | MapReduce算法 | 2 | 1 |  | 3 |
| 6 | 众包算法 | 2 | 1 |  | 3 |
| 7 | 大数据配套技术架构 | 2 | 1 |  | 3 |
| 8 | 回归、分类、聚类 | 2 | 1 |  | 3 |
| 9 | 降维分析 | 2 | 1 |  | 3 |
| 10 | 推荐与协同过滤 | 2 | 1 |  | 3 |
| 11 | 预测与时间序列预测 | 2 | 1 |  | 3 |
| 12 | 神经网络与深度学习 | 2 | 1 |  | 3 |
| 13 | 大数据分析结果评价 | 2 | 1 |  | 3 |
| 14 | 大数据分析可视化 | 2 | 1 |  | 3 |
| 15 | 大数据分析综合案例1 | 2 | 1 |  | 3 |
| 16 | 大数据分析综合案例2 | 2 | 1 |  | 3 |
| **合计** |  | **32** | **16** |  | **48** |

1. 教学内容

第1章 大数据分析算法基础

## 大数据分析算法概述

### 教学任务

第一节 大数据概述

理解大数据的概念与特点；

了解大数据在各个行业领域的应用。

第二节 大数据算法

了解大数据上求解问题的过程；

掌握大数据算法的定义与特点；

第三节 大数据算法设计与分析

了解大数据算法设计技术；

了解大数据算法分析技术。

### 教学重点、难点：

重点理解大数据上求解问题的过程，难点在于大数据算法与传统算法的区别和特点。

### 课程的考核要求：

掌握大数据的概念和特点、具有采用大数据算法和技术解决实际问题的能力。

### 复习思考题：

谈谈对“大数据”的理解，以及对业界竞争关系的分析和未来发展方向的判断。

请针对你所了解的推荐系统，讨论推荐系统中需要哪些大数据算法和技术。

## 时间亚线性算法

### 教学任务

第一节 时间亚线性算法概述

掌握平面图直径问题和排序链表搜索的亚线性算法；

掌握两个多边形交集问题的多项式时间算法；

第二节 最小生成树代价估计

了解连通分量个数估计算法；

能运用最小生成树代价估计算法解决实际问题；

第三节 时间亚线性判定算法概述

了解时间亚线性判定算法的基本原理和思想；

理解数组有序的判定算法和串相等判定算法的基本原理，掌握用该算法解决问题的思路和方法。

### 教学重点、难点：

重点掌握时间亚线性算法原理和思路，难点在于应用该算法解决适用问题的能力。

### 课程的考核要求：

能用编程语言实现最小生成树代价估计算法和时间亚线性判定算法的基本原理，难点在于能给用编程语言给出两种算法的描述。

### 复习思考题：

在输入的正文串（长度为n）中查找某一字符是否出现，若出现，输出1，否则输出0。设计时间复杂度o(n)的算法求解这个问题。

设计亚线性算法检查两个字符串是否相同。

设计时间复杂度为o(n)的算法，判定一个长度为n的0，1数组是否为0，1交替的循环数组。

## 空间亚线性算法

### 教学任务

第一节 空间亚线性算法概述

了解空间亚线性算法基本思想和原理；

理解水库抽样的基本思想和原理；

第二节 寻找频繁元素的非随机算法

了解频繁元素的精确解概念；

能运用频繁元素的Misra-Gries算法解决相关问题；

第三节 寻找频繁元素的随机算法

了解估算不同元素数量的基本算法和改建算法；

掌握略图法和计数-最小略图方法；

理解频率矩的AMS估计算法，并能运用基于拔河略图的频率矩估计；

### 教学重点、难点：

重点掌握空间亚线性算法原理和思路，难点在于应用该算法解决适用问题的能力。

### 课程的考核要求：

能用编程语言实现寻找频繁元素的非随机算法和非随机算法的基本过程，难点在于能给运用编程语言给出两种算法的描述。

### 复习思考题：

对于一个二进制流，设计空间复杂度为O(n)的算法计算长度为N的滑动窗口中1的个数。

解释如何将估计频率矩的AMS算法推广到现金注册模式，证明算法的正确性。

我们学习了两种估计数据流中的不同元素数量的方法，但是它们都没有产生一个我们定义的“略图”。如何解决这个问题？

## 外存算法与外存查找结构

### 教学任务

第一节 外存存储结构与外存算法概述

第二节 外存算法示例：外存排序算法

了解外存归并排序算法基本原理和思想；

掌握外存多路快速排序算法的思想并能运用其解决实际问题，并给出算法示例；

第三节 外存数据结构示例：外存搜索树

理解外存数据结构；

能运用外存数据结构解决外存搜索树算法相关问题；

### 教学重点、难点：

重点掌握外存排序算法的原理和思路，难点在于应用外存数据结构的外存搜索树解决适用问题的能力。

### 课程的考核要求：

能用编程语言实现外存排序算法的基本过程，难点在于能给运用编程语言给出外存排序算法和外存搜索树的描述。

### 复习思考题：

设计外存有效的算法，求N个元素中是否存在两个数的和等于K，假定N很大，N个元素无法全部放入内存。

给定a、b两个文件，各存放50亿个url，每个url各占64字节，内存限制是4GB，找出a、b文件共同的url。

怎么在海量数据中找出重复次数最多的一个？

## 外存图数据算法

### 教学任务

第一节 线性表排名及其应用

理解线性表排名问题、欧拉回路的主要思想；

掌握父子关系判定的一般方法；

了解前序技术和计算子数大小算法的主要思路和描述；

第二节 时间前向处理方法

理解DAG形式逻辑表达式计算问题；

掌握最大独立集合算法主要思想并能给出相应算法描述；

第三节 缩图法

理解基于缩图法的图连通分量计算半外存和全外存的算法思想；

掌握基于缩图法的图连通分量计算全外存算法；

掌握最小生成树算法，能用编程语言实现其描述，能用该算法解决相关实际问题；

### 第四节 广度优先搜索与深度优先搜索

理解有向图的BFS和DFS算法思想；

掌握无向图的BFS算法，并能给出算法描述；

理解无向图更高效的BFS算法主要思想，能运用其解决实际应用问题。

### 教学重点、难点：

重点掌握有向图和无向图算法的原理和思路，难点在于使学生能应用无向图更高效算法BFS的思想解决适用问题。

### 课程的考核要求：

能用编程语言实现有向图和无向图算法的基本过程，给出其算法描述；难点在于能运用编程语言给出无向图及更高效的BFS算法解决实际问题。

### 复习思考题：

设计外存有效算法，分别求图中极大匹配和最大匹配。

设计外存有效算法，计算图中三角形个数，并求图中最大团。

## MapReduce算法及例析

### 教学任务

第一节 MapReduce算法基础

了解MapReduce的基本框架；

掌握mapper和reducer的特点和组成结构；

第二节 MapReduce算法设计方法

了解局部聚合的概念；

掌握两种重要的算法设计模式—词对法和条块法；

掌握MapReduce算法设计与算法实现技巧；

### 第三节 图算法

了解连接算法的基本思想；

了解图算法的基本思路，能给出基于广度优先搜索的MapReduce图处理算法；

掌握PageRank的MapReduce算法主要思想，并能运用其解决实际问题；

第四节 基于图处理平台的并行算法

了解并行节点计算的概念及计算平台；

掌握基于并行结点计算的单源最短路径算法的设计与实现方法；

### 教学重点、难点：

重点掌握MapReduce的基本模型及算法设计模式，难点在于使学生能应用基于广度优先搜索的图处理算法的思想解决实际问题。

课程的考核要求：能用编程语言设计实现MapReduce算法，并掌握其实现技巧；能运用PageRank的MapReduce算法解决在互联网排序过程中遇到的实际问题。

### 复习思考题：

输入社交网络G和定点x，G上任意两点u、v的密切程度为定义的wu、v，设计MapReduce算法求解顶点x和它的哪些邻居关系最密切。

Hadoop的社区包实现了重分区连接，但其思想是将待连接的所有输出值都读取到内存中，然后进行多路连接，考虑优化该连接的方案。

PageRank迭代终止的可选条件有哪些？

## 众包算法

### 教学任务

第一节 众包的定义

了解众包的定义及实例；

第二节 众包的要素及关键技术

了解众包的基本流程；

掌握众包的报酬机制及其中的关键技术；

能运用众包中的关键算法和技术解决实际问题，并给出算法例析。

### 教学重点、难点：

重点掌握众包的流程和报酬机制，难点在于使学生能应用众包算法的思想解决实际问题。

### 课程的考核要求：

能用编程语言设计实现众包算法；难点在于能运用众包的思想设计某应用领域的实际问题。

### 复习思考题：

请给出众包工作的一般流程。

请给出设计众包算法时要考虑的因素，并分析为什么要考虑这些因素。

请选择一个众包平台，熟悉使用方法。注意体会如何设计HIT，如何发布任务、收集答案，如何选择完成众包的工人等。

# 2

第2章 数据结构、操作与可视化

## 本章工作任务

掌握Python语言的分支、循环语句和函数的使用。应用分支、循环语句和函数实现简单程序的编写，例如简单的四则运算和阶乘。

## 本章技能目标

掌握基本的分支语句

掌握基本的循环语句

使用函数实现基本的运算功能

使用分支语句、循环语句实现简单程序的编写

## 本章简介

**分支语句：**满足特定条件时，执行不同语句的一种基础语句。**循环语句：**一组被重复执行的语句，由循环体及循环的终止条件两部分组成。**函数：**函数是可重复使用的、用来实现特定功能的代码片段。函数提高了应用的模块性和代码的重复利用率。函数分为内建函数和自定义函数两类。其中，内建函数是指Python语言预定义好的函数，例如print()和input()等；自定义函数是指由用户自己定义的函数。

**分支语句可以解决的实际应用问题是：**根据实际问题可能出现的几种可能性（选择乘法或除法），设置判断条件（输入1时进行乘法运算，否则进行除法运算）。满足条件时，执行条件对应的语句。**循环语句可以解决的实际应用问题是：**已知某个语句需要重复执行多次（如阶乘），并存在终止判断条件时（如某一正整数的阶乘从自身开始一直乘到1结束），运用循环语句可以实现重复执行多次的操作。**函数可以解决的实际应用问题是：**已知某种方法可以解决一类问题时（加法运算），将该方法编写成函数，可以在后续再次使用该方法时，通过调用该函数解决此类问题，减少相同代码的重复编写。

**本章的重点是：**分支、循环语句和函数的理解和使用。

## 本章总结

本章实现的工作是：首先采用Python语言的分支语句实现了基本的乘除运算。然后采用Python语言的循环语句实现了阶乘运算。最后运用函数实现了基础的加法运算。

本章掌握的技能是：（1）使用分支语句，在不同判断条件下执行不同的语句；（2）使用循环语句，执行重复语句时明确起始和终止条件；（3）定义和调用函数，解决可以应用相同方法的问题。

## 本章作业

实现本章的案例，即运用分支语句实现乘除运算；运用循环语句实现阶乘运算；运用函数实现加法运算。

设计一个简单的计算器，运用分支、循环语句和函数，实现加减乘除、阶乘和指数等基本运算。

# 【绿色部分不要修改，以后会隐藏，用于实现章节号的自动编号】

第3章 数据结构、操作与可视化

## 本章工作任务

熟练使用Python中常见的八类数据结构：（1）元组tuple、（2）列表list、（3）数组ndarray、（4）矩阵matrix、（5）集合set、（6）字典dict、（7）序列Series和（8）数据框DataFrame，实现各类数据结构的创建、元素的增删改查、不同数据结构之间的相互转换以及各种数据结构的特色操作，实现数据可视化。

## 本章技能目标

掌握Python原生、NumPy库和pandas库中的八种数据结构

掌握Python中不同数据结构的创建

掌握Python中不同数据结构的增删改查操作

掌握Python中不同数据结构的特色操作

掌握Python中不同数据结构的相互转换方法

掌握Python中matplotlib中基本的数据可视化方法

## 本章简介

**数据结构是指：** 计算机存储和组织数据的方式。常用的数据结构有：元组(tuple)、列表(list)、数组（ndarray）、矩阵（matrix）、序列（Series）和数据框（DataFrame）这八种重要的数据结构。其中，**元组是指：**一个元素不可修改的线性数据结构，数据项放在“( )”内。**列表是指：**一个有序可变长度的线性数据结构，列表项放在“[ ]”中。**数组是指：**储存多维相同类型数据的数据结构，与其他编程语言中的数组类似，通过NumPy包创建，NumPy包提供了很多功能方法。**矩阵是指：**一个二维的数据结构，具有线性代数中的计算功能，例如求行列式、特征矩阵、逆矩阵等。**集合是指：**一个元素无序且元素不可重复的数据结构，集合中的元素放在“{ }”。**字典是指：**一组具有多个键-值对组成的数据结构，形式为“{键1:值1，键2:值2…}”。**序列是指：**由一列索引和一列数据组成的数据结构，通过pandas包创建，pandas包提供了很多功能方法。**数据框是指：**一种具有索引列的表格型数据结构，它含有一个索引列和多个数据列，通过pandas包创建，pandas包提供了很多功能方法。**数据操作是指：**创建数据结构存放数据、数据结构中增删改查元素、各数据结构的特色操作（例如矩阵乘法、求逆）和不同数据结构之间的转换。**数据可视化是指：**采用图形化的方式呈现数据，通常用于描述性统计分析，例如：折线图、条形图、饼状图等。

**掌握数据结构的特点（可以解决的问题）：**tuple保证了元素数据的完整性，也是函数之间用于传递参数的数据结构。list由众多列表项构成，常用于存储结构相同的数据，例如网页上抓取的大量相同结构的内容，常用于遍历列表项的数据。ndarray可以实现多维数据，NumPy为其提供了很多功能。matrix便于进行矩阵数据的特征值及特征向量计算，用于解决线性代数问题。set中的元素具有无序性和不重复性。字典类型dict中的键值对具有映射关系，便于根据键查找值，实现类似数据库中字段名与字段内容之间的关系。Series中的每个值都具有索引，可以实现根据索引产生的各种操作。例如，若一个Series类型对象的索引是时间，则该对象可以用于处理一维时间序列。DataFrame不仅具有表格功能，还具有索引，是升级版本的Series，可以方便地与外界的表格文件进行数据交换。

**本章的重点是：**数据结构的理解和操作。

## 本章总结

本章实现的工作是：采用Python语言创建tuple、list、set、dict、ndarray、matrix、Series和DataFrame八种数据结构的对象，并进行增、删、改和查等操作以及数据结构之间的转换，最终绘制八种统计图形，为后续的算法学习打下基础。

本章掌握的技能是：（1）熟悉Python原生、NumPy库和pandas库中的八种数据结构；（2）使用Python读取和存储数据，（4）数据的切片与合并；（5）使用matplotlib库实现数据的可视化，掌握绘制八种统计图形的方法。

## 本章作业

实现本章的全部案例。

使用seanborn包中的乘客数据集，将其转换为各种数据结构进行存储，进行数据可视化，分别绘制直方图、箱线图以及散点图等，并给图示增加标题。

# 【绿色部分不要修改，以后会隐藏，用于实现章节号的自动编号】

第4章 缺失值填充

## 本章工作任务

采用均值填充、前采样后采样填充和机器学习（MLP神经网络算法，即多层感知器算法）填充三种方法，对数据中的缺失值进行预测填充。（1）算法的输入是：9名学生的语文、英语、数学成绩数据（含有缺失值）；（2）算法模型需要求解的是：均值填充方法与前后采样填充方法无需求解模型，机器学习（MLP神经网络算法）填充方法求解的是最佳的权值与偏置项，使得预测结果误差最小；（3）算法的结果是：缺失值对应的预测结果。

## 本章技能目标

掌握均值填充及前后采样填充的含义与使用

掌握机器学习（MLP神经网络算法）填充的原理

掌握使用MLP算法填充缺失值的方法

使用Python对预测结果进行填充并展示填充后的结果

## 本章简介

**均值填充是指：**把缺失值所在列用不含缺失值部分的数据的平均值或指定列的数据的平均值来填充。**前采样后采样填充是指：**通过数据内部采集，把缺失值处的值填充为缺失值所在列或行的前一个值或后一个值。**机器学习（MLP神经网络算法）填充是指：**通过建立多层感知器（MLP，Multi-Layer Perceptron，主要由输入层、隐藏层、输出层组成）来填充缺失值的办法，该方法通过训练模型得到隐藏层所需的各项参数之后，将输入层的数据代入隐藏层参数组成的公式中进行计算，即可得到对应的输出层数据（缺失值的预测结果）。

**缺失值填充可以解决如下科学问题：**已知属性A、B下分别记录了个数据，，其中B属性下的个数据中包含个缺失值，结合属性A的数据，运用三种缺失值填充方式对属性B的数据的处缺失值进行预测填充，以保证数据的完整性与有效性。

**缺失值填充可以解决的实际应用问题是：**在对多位学生的各科成绩进行分析时，记录和整理过程中产生的遗漏数据会影响分析结果的准确性。利用三种缺失值填充方法可以补充遗漏的数据，保证数据的完整性和有效性，从而提高分析的准确性。

**本章的重点是：**三种缺失值填充方法的理解与使用。

## 本章总结

本章实现的工作是：首先使用Python语言导入学生语文、英语、数学成绩的样本数据。然后分别使用均值填充、前后采样填充和机器学习（MLP神经网络算法）填充这三种方法对输入数据进行处理，进而得到填充缺失值后的样本数据。

本章掌握的技能是：（1）使用Python导入学生成绩数据CSV文件；（2）使用均值填充、前后采样填充的方法实现缺失值的填充；（3）使用MLP算法中的MLC多层感知器分类器实现缺失值的填充；（4）使用iloc[]方法实现所需数据的提取；（5）使用Python实现填充后样本数据的展示。

## 本章作业

实现本章的案例，即运用均值填充、前后采样填充和机器学习（MLP神经网络算法）填充这三种缺失值填充方法对学生成绩数据中的缺失值进行填充完善，并将填充结果展示。

设计一个带有缺失值的数据，包含20名学生对三类课程的评分。分别运用均值填充、前后采样填充和机器学习填充三种方法，实现对缺失值的填充。

# 【绿色部分不要修改，以后会隐藏，用于实现章节号的自动编号】

第5章 线性回归

## 本章工作任务

采用线性回归算法编写程序，对股票价格进行预测。（1）算法的输入是：“时间-股价”数据，（2）算法模型需要求解的是：线性回归系数，（3）算法的结果是：未来、过去、期间更多时间点对应的股价。

## 本章技能目标

掌握线性回归原理

使用Python构造模拟时间序列股票数据

使用Python实现线性回归模型建模与求解

使用Python实现线性回归模型计算与预测

使用Python对线性回归结果进行可视化展示

## 本章简介

**回归分析是指：**一种确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。**一元线性回归是指：**如果回归分析中只包括一个自变量和一个因变量，且二者的关系可用一条直线近似表示的情况。**多元线性回归是指：**如果回归分析中包括两个或两个以上的自变量，且因变量和自变量之间是线性关系的情况。

**一元线性回归算法可以解决如下科学问题：**已知个自变量，与之对应个因变量，希望找到自变量和因变量之间的关系：。实现根据样本数据和，找到自变量和因变量之间的映射关系（规律）。例如：。

**回归算法可以解决的实际应用问题是：**根据已知的自变量（时间）和因变量（股票价格）样本数据，找到因变量和自变量之间的映射关系（从样本数据回归出规律模型和模型的参数），然后根据模型计算更多自变量（更多时刻、过去时刻、未来时刻）对应的因变量（预测的股票价格）。

**本章的重点是：**一元线性回归方法的理解和使用。

## 本章总结

本章实现的工作是：首先采用Python语言生成了时间-股票价格的样本数据。然后采用一元线性回归算法，对样本数据进行回归计算，得到回归系数。进而生成更多时刻点，并采用一元线性回归模型进行预测。最后将预测结果进行显示。

本章掌握的技能是：（1）使用NumPy库生成连续的时刻数值和正弦数值数组；（2）使用NumPy库线性回归函数进行回归计算和预测计算；（3）使用Matplotlib库实现数据的可视化，绘制散点图、折线图。

## 本章作业

实现本章的案例，即生成样本数据，实现一元线性回归模型的建模、预测和数据可视化。

设计一个GDP、进出口额、学习成绩等时间序列数据，运用一元线性回归方法，实现数值的预测过去、未来和期间值。

# 【绿色部分不要修改，以后会隐藏，用于实现章节号的自动编号】

第6章 多元线性回归

## 本章工作任务

采用多元线性回归算法编写程序，根据2支股票的历史价格预测第3支股票的价格。（1）算法的输入是：2支股票价格，（2）算法模型需要求解的是：多元线性回归模型的回归系数，（3）算法的结果是：第3支股票的价格。

## 本章技能目标

掌握多元线性回归原理

使用Python导入样本数据并划分为训练集和测试集

使用Python实现多元线性回归模型的建模与求解

使用Python实现多元线性回归模型的预测

使用Python对多元线性回归结果进行可视化展示

## 本章简介

**回归分析是指：**一种确定因变量与自变量间线性关系的统计分析方法。**一元线性回归是指：**如果回归分析中只包括一个自变量和一个因变量，且二者的关系可用一条直线近似表示的情况。**多元线性回归是指：**如果回归分析中包括两个或两个以上的自变量，且因变量和自变量之间是线性关系的情况。

**多元线性回归算法可以解决如下科学问题：**已知个样本数据，每个样本数据由个输入值（特征），其中，对应的输出值（标签）为，其中。希望找到样本数据和因变量之间的关系，即。

**回归算法可以解决的实际应用问题是：**根据已知的多个自变量（2支股票的价格）和因变量（第3支股票的价格）构成的样本数据，找到因变量和自变量之间的映射关系（从样本数据回归出规律模型和模型的参数），然后根据模型计算更多自变量（2支股票更多的价格）对应的因变量（第3支股票的价格）。

**本章的重点是**：多元线性回归方法的理解和使用。

## 本章总结

本章实现的工作是：首先采用Python语言导入3支股票的价格数据，并将数据划分为训练集和测试集。然后采用二元线性回归算法，对样本训练集进行回归计算，得到回归系数。进而采用二元线性回归模型对测试样本进行预测。最后对预测结果进行可视化展示。

本章掌握的技能是：（1）使用Python导入样本数据，使用loc、iloc方法将样本数据划分为训练数据和测试数据；（2）使用NumPy库中的线性回归函数进行回归计算和预测计算；（3）使用Matplotlib库实现回归结果的可视化，绘制散点图、折线图、3D图。

## 本章作业

实现本章的案例，即导入样本数据，划分样本训练集和测试集，实现二元线性回归模型的建模、预测并将数据可视化。

从sklearn库中加载波士顿房价数据集（boston）。以城镇犯罪率、一氧化氮浓度、住宅平均房间数、到中心区域的加权距离等为自变量，以波士顿平均房价为因变量的多元线性回归，运用多元线性回归方法，实现对房价的预测并将数据可视化。

# 【绿色部分不要修改，以后会隐藏，用于实现章节号的自动编号】

第7章 K近邻算法

## 本章工作任务

采用K近邻(K-Nearest Neighbors, KNN)算法编写程序，根据每一位同学的数学和英语成绩将学生划分为不同类别（理科生，综合生，文科生）。（1）算法的输入是：600位同学每位同学的英语和数学成绩及分类信息；（2）算法模型需要配置的参数是，决定分类结果的近邻数量；（3）算法的结果是：对学生的预测分类结果及分类结果的准确率。

## 本章技能目标

掌握K近邻分类算法原理

使用Python导入学生成绩数据

使用Python实现K近邻模型建模、参数配置与求解

使用Python实现KNN算法对样本数据集的分类

使用Python对算法分类结果进行可视化展示

## 本章简介

**KNN分类算法：**是一种相对简单的分类方法，如果一个样本在特征空间中的个最相邻的样本中的大多数样本都属于某一类别，则该样本也属于类别。

**KNN分类算法可以解决如下科学问题：**已知包含个正确分类样本的样本数据集，找到离待分类样本点距离（以空间中的欧式距离）最近的个样本，统计个样本中出现频率最高的标签值，即为待分类样本的预测分类结果。

**KNN分类算法可以解决的实际应用问题是：**如果根据某个学生的成绩对该学生进行分类，可根据已知的学生成绩属性值（语文成绩及数学成绩），找到待分类学生成绩欧氏距离最近的的个学生，统计这个学生中出现频率最高的分类标签，即为待分类学生的分类结果。

**本章的重点是：**KNN分类算法的理解和使用。

## 本章总结

本章实现的工作是：首先用Python导入包含学生成绩和学生类别的样本数据。然后采用K近邻分类算法，配置算法模型中的值，以维空间的欧式距离为度量标准，求解待分类学生样本的预测标签；将预测标签与真实标签进行对比得出分类结果准确率。最后将预测结果可视化。

本章掌握的技能是：（1）通过编写KNN算法的底层代码,实现KNN算法模型的构建及参数配置，对样本数据进行预测分类并计算分类结果准确率；（2）使用sklearn模块中的neighbors最近邻模块，导入其中的KNeighborsClassifier模型实现预测分类和模型打分；（3）使用Matplotlib库实现数据的可视化，绘制散点图、折线图。

## 本章作业

实现本章的案例，导入样本数据，实现K近邻分类算法建模、预测、打分和数据可视化显示。

引用sklearn库中的鸢尾花案例，使用KNN算法对其进行分类。

第6章 逻辑回归

## 本章工作任务

采用逻辑回归算法编写程序，根据学生的语文、数学、英语成绩对学生所属类别进行预测。（1）算法的输入是：600位学生的语文、数学、英语成绩和学生类别标签值（1表示文科生、2表示理科生、3表示综合生），（2）算法模型需要求解的是：逻辑回归模型的参数，（3）算法的结果是：根据逻辑回归模型预测的学生类别标签值（1表示文科生、2表示理科生、3表示综合生）。

## 本章技能目标

掌握逻辑回归原理

使用Python对数据划分训练集和测试集

使用Python实现逻辑回归模型建模与求解

使用Python通过逻辑回归模型实现对不同学生的分类

使用Python对逻辑回归结果进行可视化展示

## 本章简介

**逻辑回归分析是指：**一种分类方法，首先根据样本训练集计算出logistic函数的系数，然后将测试集带入logistic函数得到输出结果，再将该输出结果映射为0或1，最终实现分类。**二分类逻辑回归是指：**分类结果只有两种类别的逻辑回归。**多分类逻辑回归是指：**分类结果有多种类别的逻辑回归，是逻辑回归的推广。

**逻辑回归算法可以解决如下科学问题：**已知个样本数据,每个样本数据有个输入属性，即，这个样本数据对应的类别标签值分别为。希望找到自变量和因变量之间的关系：。例如：样本自变量有3个属性时，模型为，其中从而根据上述模型和测试样本的输入数据，预测测试样本的类别标签值。

**逻辑回归算法可以解决的实际应用问题是：**根据600名学生的语文、数学、英语成绩以及学生的类别标签值，建立逻辑回归模型，根据上述模型和测试集中学生的各科成绩，可以预测出对应学生的类别标签值。

**本章的重点是：**逻辑回归算法的理解和使用。

## 本章总结

本章实现的工作是：首先导入样本数据，包括学生的语文、数学、英语成绩及学生所属类别（文科生、理科生、综合生），并划分训练集与测试集。然后采用逻辑回归算法，用样本数据构建回归模型，进而利用模型和更多学生的成绩进行分类预测。最后将预测与真实类别标签值进行对比显示，检验模型预测正确率。

本章掌握的技能是：（1）读取样本数据并划分样本为训练集与测试集；（2）使用LogisticRegressionCV模块构建逻辑回归模型；（3）使用Matplotlib库绘制散点图，实现数据的可视化。

## 本章作业

实现本章的案例，即导入数据集并划分样本数据，实现逻辑回归模型的建模、预测和数据可视化。

登录Kaggle官网，针对名为Titanic：Machine Learning From Disaster的经典逻辑回归问题，运用本章所学逻辑回归方法，利用训练集数据训练逻辑回归模型，预测测试集中样本的分类结果，并计算预测准确率。访问链接www.kaggle.com，注册账号并登陆到Kaggle官方网站中，访问数据集的链接<https://www.kaggle.com/c/titanic>找到数据集。

# 【绿色部分不要修改，以后会隐藏，用于实现章节号的自动编号】

第8章 贝叶斯算法

## 本章工作任务

采用贝叶斯算法编写程序，对含有侮辱性词语的邮件进行过滤。（1）算法的输入是：40个随机词语，（2）算法模型需要求解的是：某词语出现在含有侮辱性词语邮件的概率，（3）算法的结果是：词语的类型（属于侮辱性词语还是非侮辱性词语）。

## 本章技能目标

掌握贝叶斯算法原理

掌握概率分布原理

使用Python实现文本分类

使用Python实现朴素贝叶斯分类器训练

## 本章简介

**贝叶斯算法是指：**一种根据新样本的已有特征在数据集中的**条件概率**来判断新样本所属类别的算法。**条件概率是指：**已知事件B已经发生的条件下，事件A发生的概率。

**贝叶斯算法可以解决如下科学问题：**已知A个结果和B个导致该结果的样本，得到在具有某特征的样本发生的条件下某结果发生的概率。根据，得到在某结果出现时某个样本可能发生的概率。

**贝叶斯算法可以解决的实际应用问题是：**根据已知标签的（被过滤邮件和非被过滤邮件）样本数据，输入待检验的词汇，通过贝叶斯算法，判断其是否为侮辱类词语（根据样本数据中词语出现在被过滤邮件和非被过滤邮件的概率）。

**本章的重点是：**贝叶斯算法的理解和使用。

## 本章总结

本章实现的工作是：首先采用Python语言生成了单词词组。然后采用贝叶斯算法，对样本数据进行计算得到每个单词出现的概率。进而采用贝叶斯模型对待分类的邮件进行判断。最后将判断结果进行展示。

本章掌握的技能是：（1）构造并使用朴素贝叶斯分类器；（2）使用sklearn模块中的naive\_bayes库进行文本分类。

## 本章作业

实现本章的案例，即生成样本数据，实现贝叶斯模型的建模、预测。

引用sklearn库中的鸢尾花案例，运用贝叶斯方法对其进行分类。

# 【绿色部分不要修改，以后会隐藏，用于实现章节号的自动编号】

第9章 决策树

## 本章工作任务

采用ID3和CART决策树算法编写程序，依据学生的数学、英语成绩，对学生进行分类（类别包括：文科生、理科生和综合生）。（1）算法的输入是：600名学生的成绩与所属类别，（2）算法模型需要求解的是：决策树（各节点代表的属性与分支条件），（3）算法的结果是：根据待测样本成绩预测的分类结果。

## 本章技能目标

掌握决策树原理

使用Python配置决策树模型的参数

使用Python实现决策树模型建模

使用Python实现决策树模型学习与预测

使用Python对决策树分类结果进行可视化展示

## 本章简介

**决策树是指：**一种树形结构的分类算法，该算法依据各节点属性与分支条件对数据进行分类。（例如：根据学生各科成绩对学生进行分类时，先按照数学成绩分为两类，数学成绩即为该节点分支依据的属性，小于80分的学生分到左支，小于80分即为分支条件）。**ID3算法是指：** 一种以信息增益作为指标，判断各节点放什么属性的决策树算法。**CART算法是指：**一种以基尼系数作为指标，判断各节点放什么属性的决策树算法。

**决策树算法可以解决如下科学问题：**已知个样本，对于第个样本具有个属性和1个标签值，通过找到属性值与类别之间的关系，实现根据第个测试样本得到分类结果。

**决策树算法可以解决的实际应用问题是：**实现根据学生的各科成绩和学生所属类别，找到各科成绩与学生类别的关系（建立模型、求解模型），然后根据更多的学生成绩使用模型对其所属类别进行预测。

**本章的重点是：**决策树算法的理解和使用。

## 本章总结

本章实现的工作是：首先使用Python语言导入了学生英语、数学成绩及学生所属类别（文科生、理科生、综合生）的样本数据。然后分别采用ID3和CART决策树算法，对样本数据进行模型拟合，求出各属性值的信息增益或基尼系数，进而求解决策树模型，并采用求解后的模型进行预测。最后将训练样本与测试样本进行展示。

本章掌握的技能是：（1）使用Python导入学生成绩的数据；（2）使用pydotplus包配置决策树绘制方法的参数；（3）使用sklearn中决策树分类工具实现ID3和CART决策树模型建立、求解与预测；（4）使用matplotlib库对ID3和CART决策树分类结果进行可视化展示。

## 本章作业

实现本章的案例，即实现ID3和CART决策树模型的建模、预测和数据可视化。

引用sklearn库中的鸢尾花数据集。采用sklearn中的决策树算法工具对其进行分类。

# 【绿色部分不要修改，以后会隐藏，用于实现章节号的自动编号】

第10章 支持向量机

## 本章工作任务

采用支持向量机算法编写程序，根据400名学生的语文成绩和数学成绩对这400名学生进行分类，将其划分为文科生和理科生。（1）算法的输入是：“400名学生的语文成绩和数学成绩以及正确的分类信息（文科生、理科生）”，（2）算法模型需要求解的是：线性核函数下支持向量机的超平面方程的系数和，（3）算法的结果是：待测样本中学生的分类结果。

## 本章技能目标

掌握支持向量机分类原理。

使用Python读取样本数据并划分为训练集和测试集。

使用Python实现支持向量机分类模型的建模与求解。

使用Python实现使用支持向量机分类模型对样本数据进行分类。

使用Python将支持向量机的分类结果进行可视化展示。

## 本章简介

**支持向量机（SVM, Support Vector Machine）是指：**是一种线性二分类器，它的工作原理是在一组仅含两个类别的（n个类别）样本数据中找出一个（n个）超平面作为决策边界，使得分类误差尽可能的小。我们用上述表格罗列出支持向量机模型解决分类问题时所用的原理，使大家能更直观的理解支持向量机分类器的本质——线性二分类器。

**支持向量机算法可以解决如下科学问题：**已知个样本数据，每个数据具有M个输入属性，即,个样本的分类标签值分别是。根据样本数据建立支持向量机模型，可以找到样本属性和样本分类标签之间的关系，之后对于新的任意个样本数据，都可以推算出该样本所属的类别（分类标签值）。

**支持向量机算法可以解决的实际应用问题是：** 已知个样本（学生），每个样本具有个输入属性（数学成绩和语文成绩）,每个样本对应的分类标签值分别是文科生、理科生和综合生。通过建立支持向量机模型，可以找到样本属性（成绩）与样本分类标签（文科生、理科生）之间的关系，从而对于任意个新的样本（学生），都可以利用该样本所具有的属性推算出该样本所属的分类标签。

**本章的重点是：**支持向量机分类算法的理解和使用。

## 本章总结

本章实现的工作是：首先导入含有学生数学成绩、语文成绩和学生所属类别（文科生、理科生和综合生）的样本数据，利用Python语言中的iloc方法将样本划分为训练集和测试集。然后调用SVM方法集中的SVC算法创建支持向量机分类器（SVC）模型，接着输入训练数据，求解支持向量机分类器模型，进而在输入更多学生的数学成绩和语文成绩时，能利用已求解的支持向量机分类器模型去预测其分类结果。最后将预测结果进行可视化展示。

本章掌握的技能是：（1）使用Python方法读取csv文件中的数据；（2）使用iloc方法划分训练集和测试集；（3）使用sklearn库中的支持向量机模块解决非线性多分类问题；（4）使用Matplotlib库实现数据的可视化，绘制散点图。

## 本章作业

实现本章的案例，即生成样本数据，实现非线性支持向量机回归模型的建模、预测和数据可视化。

从UCI数据库中下载Iris（鸢尾花）原始数据集，运用支持向量机分类算法，实现根据鸢尾花的任意两个特征对其进行分类。

# 【绿色部分不要修改，以后会隐藏，用于实现章节号的自动编号】

第11章 随机森林

## 本章工作任务

采用随机森林分类算法编写程序，根据600名学生的英语、数学成绩，对这600名学生进行分类，将其划分为文科生、理科生或综合生。（1）算法的输入是：600个学生的英语、数学成绩以及学生的所属类别（文科生、理科生、综合生），（2）算法模型需要求解的是：6棵决策树（包括各节点属性与分支条件），（3）算法的结果是：待测样本中学生的分类结果。

## 本章技能目标

掌握随机森林分类原理

使用Python实现随机森林分类模型的建模

使用Python优化随机森林分类模型的参数

使用Python实现随机森林分类模型的求解与预测

使用Python对随机森林分类结果进行可视化展示

## 本章简介

**集成学习是指：**一种分类方法，其核心思想是将若干个弱分类器（分类器的分类准确率在60%-80%之间）按照一定的结合策略组合起来，得到一个分类性能显著优越的强分类器（分类器的分类准确率在90%以上）。集成学习算法主要分为三类。

**随机森林分类算法是指：**一种Bagging集成学习算法，当一棵决策树无法很好地对待测样本进行分类时，采用多棵决策树对待测样本进行分类。该分类算法从给定训练集中有放回地均匀抽取多组样本，每组样建立一棵决策树，进而形成一个森林（由多棵不同的决策树构成），森林中的每棵决策树都会生成一个分类结果，接着将多个分类结果进行投票统计，得票最多的类别即为最终的分类结果。

**随机森林分类算法可以用于解决如下科学问题：**已知有个样本数据，每个样本数据均具有个输入属性，即,个样本的分类标签值分别是。根据样本数据建立随机森林分类模型，可以找到样本属性和样本分类标签值之间的关系，之后对于新的任意第个样本数据，都可以推算出该样本所属的类别（分类标签值）。

**随机森林分类算法可以解决的实际应用问题是：**已知个样本（学生），每个样本具有个输入属性（英语成绩和数学成绩）,每个样本（学生）对应的分类标签值分别是1-文科生、2-理科生和3-综合生。通过建立随机森林分类模型，可以找到样本属性（英语成绩和数学成绩）与样本分类标签值（1-文科生、2-理科生和3-综合生）之间的关系，从而对于任意第个新的样本（学生），都可以根据样本属性（英语成绩和数学）推算出样本的分类标签值（1-文科生、2-理科生和3-综合生）。

**本章的重点是：**随机森林分类算法的理解和使用。

## 本章总结

本章实现的工作是：首先采用Python导入含有600个学生的英语成绩、数学成绩以及学生所属类型（文科生、理科生、综合生）的样本数据。然后建立随机森林分类模型，利用训练样本训练该模型，得到6棵决策树，进而采用改进后的随机森林分类模型对待测样本数据进行预测。最后将预测结果进行显示。

本章掌握的技能是：（1）使用NumPy库实现对样本数据的读取；（2）使用sklearn库实现训练集和测试集的划分；（3）使用sklearn库中的随机森林分类算法解决分类问题；（5）使用Pydotplus库实现数据的可视化，绘制树形图；（6）使用Matplotlib库实现数据的可视化，绘制分类图。

## 本章作业

实现本章的案例，即导入样本数据，实现随机森林分类模型的建模、预测和数据可视化。

引用sklearn库中的鸢尾花案例，运用随机森林分类算法对其分类。

# 【绿色部分不要修改，以后会隐藏，用于实现章节号的自动编号】

第12章 AdaBoost算法

## 本章工作任务

采用AdaBoost分类算法编写程序，根据600名学生的语文成绩和数学成绩对这600名学生进行分类，将其划分为文科生、理科生或综合生。（1）算法的输入是：600名学生的“英语成绩和数学成绩以及正确的学生分类信息（文科生、理科生和综合生）”数据。（2）算法模型需要求解的是：每个弱分类器和分类器的权值。（3）算法的结果是：根据学生的成绩预测出的分类结果。

## 本章技能目标

掌握AdaBoost算法原理

使用Python读取数据并划分为训练集和测试集

使用Python实现AdaBoost算法模型的建模与优化

使用Python实现AdaBoost算法模型的计算与预测

使用Python对AdaBoost算法模型结果进行可视化展示

## 本章简介

**AdaBoost算法是：**一种集成算法，其核心思想是针对同一个训练集反复学习得到一系列弱分类器（子分类器），每个弱分类器都具有分类能力，但是对最终的分类结果的影响能力不同（权值不同），每个弱分类器通过加权后得到强分类器（总分类器）。它的特点是：在每一次迭代时，被弱分类器错误分类的样本的权值会增大，正确分类的样本的权值会减小，权值变更后的样本用来训练下一个弱分类器。**弱分类器（子分类器）是指：**每次训练迭代过程中所产生的分类器。**强分类器（总分类器）是指：**由多个弱分类器的分类结果进行加权计算，得到的最终分类器（输出最终分类结果）。最终分类结果（总分类器输出），其中表示第个分类器的分类结果，表示第个分类器的权值。

**AdaBoost算法可以解决如下科学问题：**已知个样本数据，每个样本数据具有个输入属性，即，，…，，个样本数据的标签值分别是 。根据上述样本数据建立AdaBoost分类算法模型，可以找到属性值和标签值之间的关系，之后对于新的任意第个测试样本，可以推算出该样本的所属类别（标签值）。

**AdaBoost分类算法可以解决的实际应用问题是：**根据已知的样本特征值（学生成绩）和其对应的标签值（学生类别）找到样本属性值和标签值的关系，从而对于任意个新的样本（学生），都可以利用该样本所具有的属性值推算出该样本所属的类别。

**本章的重点是：** AdaBoost算法的理解和使用。

## 本章总结

本章实现的工作是：首先采用Python语言读取数据并构造训练集和测试集。然后建立Adaboost模型，利用训练集训练该模型，接着优化最大迭代次数，得到Adaboost优化模型，求解得到每个弱分类器和它的权重，进而得到最终的分类器，实现计算模型准确率。最后将预测的分类结果可视化。

本章掌握的技能是：（1）使用os库改变工作目录；（2）使用pandas库读取数据；（3）使用train\_test\_split函数构造训练集和测试集；（4）使用Matplotlib库实现数据的可视化，绘制分类图；（5）使用sklearn.ensemble库中的AdaBoostClassifier函数建立Adaboost模型，进行模型拟合，计算模型准确率；（7）使用sklearn.model\_selection中的GridSearchCV函数对Adaboost参数中的最大迭代次数进行自动调参。

## 本章作业

实现本章的案例，即构造训练集和测试集，实现AdaBoost模型的建模、测试和分类结果可视化。

设计一个通过鸢尾花的花萼长度，花萼宽度，花瓣长度，花瓣宽度4个属性预测鸢尾花卉属于（Setosa，Versicolour，Virginica）三个种类中的哪一类的案例

# 【绿色部分不要修改，以后会隐藏，用于实现章节号的自动编号】

第13章 GBDT

## 本章工作任务

采用梯度提升决策树（Gradient Boosting Decision Tree）算法编写程序，根据700名学生的“数学”、“英语”成绩对其进行分类，将其划分为“文科生”、“理科生”和“综合生”。（1）算法的输入是：700名学生的“数学”和“英语”成绩以及相应的学生类型，（2）算法模型需要求解的是：颗残差树（每颗残差树需要求解所有分支，每个分支结点需要求解该分支的属性及分支的阈值），（3）算法的结果是：待测样本中学生的分类。

## 本章技能目标

掌握GBDT原理

使用Python读取样本数据，并划分训练集和测试集

使用Python实现GBDT的建模与求解

掌握GBDT模型的参数调整方法

使用GBDT模型实现预测

使用Python实现对GBDT分类结果的可视化展示

## 本章简介

**梯度提升决策树(GBDT)是指：** 一种集成算法，由多个子分类器（弱分类器）的分类结果进行累加，从而得到总分类器（强分类器）的分类结果。GBDT子分类器的特点是后一个子分类器是对前一个子分类器得到的分类结果与目标值之间的差值进行的拟合，即后一个子分类器是对前一个子分类器得到的残差值的矫正。

**梯度提升决策树(GBDT)算法可以解决如下科学问题：**已知个样本数据，每个样本数据具有个输入属性(样本特征)，即,个样本数据的标签值分别是。根据上述样本数据建立梯度提升决策树算法模型，然后根据这个模型，对于新的任意个测试样本，可以推算出该样本的所属类别。

**梯度提升决策树（GBDT）算法可以解决的实际应用问题是：**已知个样本数据，样本特征是学生的数学成绩和英语成绩，样本标签是学生的类型。通过建立GBDT模型对样本数据进行训练，找到样本特征和样本标签之间的关系，从而预测出个新的样本（学生）的类型。

**本章的重点是**：梯度提升决策树方法的理解和使用。

## 本章总结

本章实现的工作是：首先采用Python语言读取含有英语成绩、数学成绩以及学生所属类型的样本数据。然后将样本数据划分为训练集和测试集，接着采用GBDT算法，对训练集数据进行拟合，最后在输入更多学生的数学成绩和英语成绩后，使用已求解的最优模型去预测其分类结果。

本章掌握的技能是：（1）使用NumPy包读取连续的样本数据；（2）使用sklearn库model\_selection模块中的model\_selection函数实现训练集和测试集的划分；（3）使用Matplotlib库实现数据的可视化，绘制树状图。

## 本章作业

实现本章的案例，即生成样本数据，实现梯度提升决策树模型的建模、参数调整、预测和数据可视化。

利用Iris（鸢尾花）原始数据集，运用GBDT算法，实现根据鸢尾花的任意两个特征对其进行分类。

# [绿色部分不要修改，以后会隐藏，用于实现章节号的自动编号]

第14章 XGBoost

## 本章工作任务

采用XGBoost（eXtreme Gradient Boosting）算法编写程序，根据700名学生的“数学”、“英语”成绩对其进行分类，将其划分为“文科生”、“理科生”和“综合生”。（1）算法的输入是：600名学生的“数学”和“英语”成绩以及相应的学生类型，（2）算法模型需要求解的是：颗残差树（每颗残差树需要求解所有分支，每个分支结点需要求解该分支的属性及分支的阈值），（3）算法的结果是：待测样本中学生的分类。

## 本章技能目标

掌握XGBoost原理

使用Python读取学生成绩数据

使用Python实现XGBoost模型建模与求解

使用Python实现XGBoost模型计算与预测

使用Python对XGBoost结果进行可视化展示

## 本章简介

**XGBoost是指：**一种经过改进的梯度提升决策树（GBDT）算法，XGBoost算法也是构造棵分类树（子分类器），后一个子分类器对前一个子分类器的残差进行修正，各子分类器结果求和构成总分类器。**XGBoost与GBDT算法的最显著区别是：XGBoost的目标函数由损失函数的二阶泰勒展开项与复杂度惩罚项构成。**XGBoost被普遍认为是最经典的分类算法，应用于大量数据分析和各类学术竞赛。XGBoost支持多种线性分类器，收敛速度快，参数学习空间大，可预防过拟合，可处理缺失值，支持并行计算，近似计算。

**XGBoost算法可以解决如下科学问题：**已知个样本数据，每个样本数据具有个特征值，即，和个样本数据的标签值分别是。根据上述样本数据建立梯度提升决策树模型，然后根据这个模型，对于新的任意个测试样本，可以推算出该样本的所属类别。

**XGBoost算法可以解决的实际应用问题是：**已知一部分学生（样本数据）的各科成绩（特征值）和学生类别（标签），通过XGBoost算法找到学生各科成绩（样本数据特征值）与学生的分类值（样本数据的标签值）之间的对应（分类）规律（模型），从而可以根据另一批学生的成绩（测试样本的特征值）测去预测出这些学生的分类（测试样本的标签）。

**本章的重点是：**XGBoost方法的理解和使用。

## 本章总结

本章实现的工作是：首先pandas库读取含有英语成绩、数学成绩以及学生所属类型的样本数据。然后将样本数据划分为训练集和测试集，接着采用XGBoost算法，对训练集数据进行拟合，最后使用Matplotlib库实现数据的可视化，绘制散点图。

本章掌握的技能是：（1）使用NumPy包读取连续的样本数据；（2）使用XGBoost库的XGBoost函数进行回归计算和预测计算；（3）使用Matplotlib库实现数据的可视化，绘制树状图。

## 本章作业

实现本章的案例，即生成样本数据，实现XGBoost模型的建模、参数调整、预测和数据可视化。

利用Iris（鸢尾花）原始数据集，分别运用AdaBoost、GBDT、XGBoost方法对该数据集进行分类，并比较结果的准确率和所用时间。

# 【绿色部分不要修改，以后会隐藏，用于实现章节号的自动编号】

第15章 K-means聚类算法

## 本章工作任务

采用K-means聚类算法编写程序，对100名学生的数学、语文成绩进行聚类。（1）算法的输入是：100名学生的数学、语文成绩数据；（2）算法模型需要求解的是：各个聚类质心（即聚类的中心）；（3）算法的结果是：各个数据（每个学生）所属簇（一类数据对象的集合）。

## 本章技能目标

掌握K-means聚类算法原理

使用Python实现对K-means聚类算法建模与求解

依据K-means聚类算法模型计算样本数据所属簇

使用Python对聚类结果进行可视化展示

## 本章简介

**聚类是指：**将相似的对象归到同一簇中的过程。K-means聚类是指：一种迭代求解的聚类分析算法，算法的核心内容是，找到个聚类中心，使得各样本点距离其聚类中心距离最近。

**K-means算法可以解决如下科学问题：**拟将个样本点聚为类，希望得到个聚类中心，对于任意一个样本点，如果与的距离最近，则的类别即为的聚类结果。

**K-means算法可以解决的实际应用问题是：**已知100同学的语文和数学成绩，准备将这些同学聚为3类，聚类结束后，进一步根据第个类别中学生的成绩分析出第个类别的特征（例如，如果第个类别的学生，数学成绩均大于90分，语文成绩均小于60分，则可以将第个类别解释为理科生）。

**本章的重点是：**K-means聚类过程的理解和使用。

## 本章总结

本章实现的工作是：首先采用Python语言读取样本数据（学生的语文、数学成绩）。然后建立K-means聚类算法模型并求解模型（得到了质心坐标），同时得到每个样本点所属的簇。在二维平面中绘制样本点，样本点的形状由上述聚类的结果决定，最后画出质心。

本章掌握的技能是：（1）掌握Python求解欧式距离的方法；（2）掌握求解K-means模型的方法；（3）使用Matplotlib库实现聚类结果的可视化。

## 本章作业

实现本章的案例，即读取样本数据，实现K-means聚类算法的建模、求解和将数据可视化。

运用K-means聚类算法，利用与地区经济发展相关的两个数据（如地区生产总值、货物净出口额），将不同的地区聚为类（很发达、较为发达、落后等）。

# 【绿色部分不要修改，以后会隐藏，用于实现章节号的自动编号】

第16章 DBSCAN聚类算法

## 本章工作任务

采用DBSCAN聚类算法编写程序，依据100位学生的数学成绩、英语成绩对学生进行聚类。（1）算法的输入是：100位学生的数学成绩、英语成绩，（2）DBSCAN求解前需要配置的模型参数是：半径和最小样本数（3）算法的结果是：学生聚类后的标签值、不同标签值的个数。

## 本章技能目标

掌握DBSCAN 聚类算法原理

掌握DBSCAN算法启动前的配置方法

使用DBSCAN算法实现聚类。

掌握DBSCAN聚类结果可视化的方法

## 本章简介

**DBSCAN算法是指：**一种基于数据空间密度的聚类算法，该算法将样本空间内具有足够密度的样本点聚成一类，聚类前不需要指定聚类数量，算法会基于数据密度分布特点推断出聚类数量。例如：如果样本点属于类别，则如果在样本点周围半径为的空间中，其他样本点的数量大于算法运行前配置的最小样本点数量，则认为上述个样本点与样本点都属于类别。

**DBSCAN算法可以解决如下科学问题：**已知样本空间中的个样本点，希望将上述样本点中高密度聚集的那些样本点归为一类，此时采用DBSCAN聚类算法。

**DBSCAN算法可以解决的实际应用问题是：**已知体育课上自由活动的学生坐标，用DBSCAN算法对这些学生进行聚类，聚类后的同一类学生呈现出共同特征。已知一个班上所有学生的数学成绩、英语成绩，用DBSCAN算法对学生成绩聚类，得到理科生、文科生、综合生3种学生类别。

**本章的重点是：**DBSCAN算法的理解和使用。

## 本章总结

本章实现的工作是：首先导入100位学生的数学成绩、英语成绩，然后建立DBSCAN模型，配置模型参数，对样本数据进行聚类，得到学生聚类后的类别数和每个学生的标签值，最后将聚类结果可视化。

本章掌握的技能是：（1）使用DBSCAN函数实现数据聚类；（2）使用Matplotlib库实现数据的可视化，绘制图形。

## 本章作业

实现本章的案例，即导入样本数据，实现DBSCAN模型的建模和数据可视化。

导入sklearn包中的鸢尾花数据集，运用DBSCAN聚类算法进行聚类。

# 【绿色部分不要修改，以后会隐藏，用于实现章节号的自动编号】

第17章 凝聚类层次聚类

## 本章工作任务：

采用凝聚类层次聚类编写程序，根据语文、数学、英语成绩对不同学生进行分层。（1）算法的输入是：20名学生的语文，数学和英语成绩；（2）算法模型需要求解的是：任意两个数据集之间的距离；（3）算法的结果是：20个数据组被先后聚合，最终逐层聚成一类。

## 本章技能目标

掌握凝聚类层次聚类的原理

使用Python计算出欧式距离

使用Python画出聚类树图

使用Python画出热度图

## 本章简介

**凝聚类层次聚类分析指：**一种分层聚类的方法，对目标数据集使用某种标准，先后进行聚合，直至聚合到最后一层，得到一个由所有数据集逐层聚合成的集合。

**凝聚类层次聚类分析算法可以解决如下科学问题：**已知个数据组，先计算出任意两组原始数据组之间的距离后，根据任意两个数据组之间距离最近（组内元素之间距离最近）的标准进行聚合，得到新的数据组 ；再计算新数据组任意两个数据组之间的距离，将距离最近的先聚，再次得到一个新的数据组，以此类推，最终聚成一类。

**凝聚类层次聚类分析算法可以解决的实际应用问题是：**已知二十位学生的语文、数学和英语三科成绩（每一个学生的三科成绩相当于一个原始数据组），计算出任意两个数据组之间的距离后，将距离最近的先聚合，得到新的数据组，再计算任意两个数据组的距离，将距离最近的先聚合，以此类推，聚成一类。

**本章的重点是：**凝聚类层次聚类分析的理解和使用。

## 本章总结

本章实现的工作是：首先导入20位学生的三科成绩，然后根据优先聚合距离最近的两个数据组的聚类标准，得到由20位学生划分成的不同簇类。

本章掌握的技能是：（1）使用SciPy包计算欧式距离；（2）使用dendrogram实现画出聚类树图；（3）使用Matplotlib包实现数据可视化，绘制热度图。

## 本章作业

实现本章的案例，即求出任意两个数组之间的距离，画出聚类树图和热度图。

运用凝聚类层次聚类分析，实现通过消费者工资和花销，对消费者分层。数据集链接如下（作业只需要使用其中2列：annual income和spending score）http://www.kankanyun.com/data/Mall\_Customers.csv。

# 【绿色部分不要修改，以后会隐藏，用于实现章节号的自动编号】

第18章 主成分分析与因子分析

## 本章工作任务

采用主成分分析（PrincipalComponentAnalysis，PCA）和因子分析（FactorAnalysis，FA）算法编写程序，对学生的成绩属性（语文、历史、政治、数学、物理、化学共6门课）进行降维处理。（1）算法的输入是：100名学生的6门课成绩。（2）算法模型需要求解的是：主成分分析的降维矩阵与方差最大化的因子载荷矩阵。（3）算法的结果是：每位同学降维后的2维指标值（对应着每位同学降维前的六科成绩）。

## 本章技能目标

掌握主成分分析和因子分析的原理

掌握主成分分析和因子分析的详细求解过程（主成分分析需要求解相关矩阵、特征值、特征向量，根据累计贡献度确定最佳降维维度；因子分析需要求解方差最大化时的因子载荷矩阵、因子得分）

使用Python掌握将高维数据转为低维数据的方法（计算将六维数据转换为二维矩阵的转换矩阵）

根据方差最大化的因子载荷矩阵得出各因子和各维度之间的关系，从而得到各因子的含义。

使用Python对降维结果进行可视化展示

## 本章简介

**主成分分析是指：**一种可以将原始数据的多维属性转换为少量低维指标（成分）的降维方法，降维后的数据尽可能多的保留了原始多维数据的特征信息，从而使低维指标的数据可以代替多维属性的数据。（例如：将100名学生的六门课成绩用两个指标来表示，即将原始6维变量降为2维来描述，以便于数据的分析）

**因子分析是指：**一种可以将原始数据的多维属性转换为少量低维指标（因子）的降维方法，降维过程中尽可能保留原数据的特征信息，降维结果中可找到低维指标（因子）与原始数据的高维属性之间的对应关系，且使得低维指标的每个因子可以被解释。（例如：把100名学生的六门课成绩属性降维至两个指标，这两个指标分别被解释为文科和理科）

**主成分分析和因子分析算法可以解决如下科学问题：**已知个样本数据，每个样本数据具有个输入属性，，根据上述样本数据建立主成分分析和因子分析模型，根据模型将*M*维属性表示的样本数据降低至可由*P*个特征表示的样本数据，其中，且可以很大程度上保留原始数据的信息。

**主成分分析和因子分析算法可以解决的实际应用问题是：**根据已知的学生的各科成绩（语文、历史、政治、数学、物理、化学）和学生分类结果的样本数据求解特征值和特征向量，根据累计贡献度找到最佳降维维数（主成分、因子个数），降维的结果可以用于分类、聚类、可视化。进一步进行因子分析，计算因子载荷矩阵，对因子载荷矩阵进行方差最大化变换，进而实现降维，且得到的每个因子都可以被原始数据很好的解释。

**本章的重点是：**主成分分析、因子分析方法的理解和使用。

## 本章总结

本章实现的工作是：首先导入含有100名学生的6门课成绩与学生分类信息的样本数据。然后采用主成分分析，对样本数据进行降维（本例中为6维降至2维），确定最佳降维维数，进而对数据进行因子分析，求解方差最大化的因子载荷矩阵，实现降维操作，其中提取出的每个因子都可以被原数据很好的解释，最后将降维结果可视化。

本章掌握的技能是：（1）使用NumPy库、pandas库求解数据相关矩阵、特征值及特征向量。（2）通过求解特征值得到累积贡献率，确定最佳降维维度。（3）使用PCA对数据进行降维并进行可视化处理。（4）求解方差最大化的因子载荷矩阵并对数据进行降维，且提取出的因子可以找到和原数据属性之间的对应关系。（5）使用matplotlib库实现降维结果的可视化。

## 本章作业

实现本章的案例，即对100名学生六门课成绩的样本数据进行降维，实现主成分分析和因子分析算法。

查找出一个地区的GDP、教育水平、消费指数、环境污染指数等相关变量的数据信息，运用主成分分析和因子分析方法，实现对数据的降维操作。

# 【绿色部分不要修改，以后会隐藏，用于实现章节号的自动编号】

第19章 奇异值分解

## 本章工作任务

采用奇异值分解（Singular Value Decomposition, SVD）算法编写程序，对100名学生的6科成绩进行降维。（1）算法的输入是：100名学生的6门成绩数据（语文、历史、政治、数学、物理、化学）；（2）算法模型需要求解的是：奇异值分解模型的降维矩阵；（3）算法的结果是：每位同学的六科成绩对应的降维后的二维指标值。

## 本章技能目标

掌握奇异值分解原理

使用Python实现奇异值分解的详细过程

使用Python实现将高维数据降维到低维指标数据

使用Python对降维结果进行可视化展示

## 本章简介

**奇异值分解降维算法是指：**一种通过对样本数据进行奇异值分解，在尽可能保留样本数据特征信息的基础上，将高维数据降低为低维数据的方法。

**奇异值分解降维算法可以解决如下科学问题：**已知个样本数据，每个样本数据具有个输入属性，即

。根据上述样本数据建立奇异值分解模型，求出降维矩阵，根据降维矩阵将样本数据从原始的维降低至维

，其中。降维后的第个样本保留了降维前第个样本中的绝大部分特征信息。

**奇异值分解算法可以解决的实际应用问题是：**根据已知的100名同学的六科成绩对学生进行分类比较困难，如果可以用少数维度指标代替学生的六科成绩（特征信息相近），例如仅根据两个指标对学生进行分类，则上述分类的工作将被简化。

**本章的重点是：**奇异值分解降维方法的理解和使用。

## 本章总结

本章实现的工作是：首先导入学生成绩及分类数据作为样本数据。然后采用奇异值分解算法，对样本数据进行降维,进而将六维数据降到二维指标。最后，依照分类结果将低维指标可视化。

本章掌握的技能是：（1）求解输入数据的协方差矩阵及其特征值和特征向量；（2）通过奇异值分解得到奇异值矩阵；（3）通过计算累计贡献度，确定最佳的降维数目；（4）使用Matplotlib库实现数据的可视化，绘制散点图。

## 本章作业

实现本章的案例，即实现学生成绩样本矩阵的奇异值降维和数据可视化。

读取鸢尾花特征（花萼长度、花萼宽度、花瓣长度、花瓣宽度）和种类（Setosa，Versicolour，Virginica）的样本数据，运用奇异值分解方法，实现数据的降维和可视化。

# 

第19章 线性判别分析

## 本章工作任务

采用线性判别分析（Linear Discriminant Analysis，LDA）的算法编写程序，对100名同学的六科成绩进行降维。（1）算法的输入是：100个学生的语文成绩、历史成绩、政治成绩、数学成绩、物理成绩、化学成绩；（2）算法模型需要求解的是：线性判别分析模型的降维矩阵；（3）算法的结果是：每位同学的六科成绩对应的降维结果（二维指标或一维指标）。

## 本章技能目标

掌握LDA算法的降维原理。

使用Python导入样本数据。

使用Python建立线性判别分析的模型

使用Python中线性判别分析算法进行降维

使用Python对降维后的分类结果进行可视化展示

## 本章简介

**降维算法是指：**一种高维数据向低维数据映射的方法，即将原高维空间中的数据点映射到低维度空间中的数据点。降维后的数据保存了高维数据的绝大部分特征信息，从而使用低维度的数据代替高维度的数据。**LDA线性判别分析是指：**将高维样本映射到低维坐标轴上，降维后的效果为：同一类别数据尽量靠近，不同类别数据尽量分开。

**LDA算法可以解决如下的科学问题**：已知原始数据矩阵，由于维度过高，不便于特征的提取，所以采用降维以方便特征的提取，降维的方法是：找到一个降维数据矩阵，通过矩阵运算，得到低维数据矩阵。不同于PAC的是，LDA进行降维时样本数据的类别信息参与计算，降维后低维空间中数据点的特征是：同一类别数据尽量靠近，不同类别数据尽量分开。

**LDA算法可以解决的实际应用问题是：**根据已知的100个学生的语文成绩、历史成绩、政治成绩、数学成绩、物理成绩、化学成绩和学生类别（文科生，理科生和综合生）样本数据，通过线性判别分析（LDA）进行降维，用低维指标反映这100名学生的高维成绩信息。

**本章的重点是：**LDA算法的理解与使用。

## 本章总结

本章实现的工作是：首先导入100名同学的6科成绩，然后使用LDA降维算法对样本数据进行处理，得到降维后的100名同学的二维与一维指标，最后将二维与一维结果可视化。

本章掌握的技能是：（1）使用pandas库导入csv文件。（2）使用LDA方法将数据降维。（3）使用Matplotlib库将降维后的数据可视化，绘制散点图。

## 本章作业

实现本章的案例，即导入样本数据，建立LDA模型并降维，实现数据的可视化。

导入白葡萄酒数据集，建立LDA模型并降维。

第20章 基于项目的协同过滤

## 本章工作任务

采用基于项目的协同过滤算法编写程序，根据多个用户的电影喜好评分，为用户A推荐电影（将与用户A喜欢的电影相似的电影推荐给用户A）。（1）算法的输入是：“不同用户对多部电影的评分”数据；（2）算法模型需要求解的是：电影相似度；（3）算法的结果是：为用户推荐的电影。

## 本章技能目标

掌握基于项目的协同过滤原理

使用Python实现基于项目的协同过滤

## 本章简介

**协同过滤是指：**根据用户A与其他用户共同喜好的内容为用户A进行推荐内容的算法。**基于项目的协同过滤是指：**找到与用户A喜欢的项目相似的项目，推荐给用户A。

**基于项目的协同过滤算法可以解决的实际应用问题是：**在电子商城中，协同过滤推荐算法被广泛应用，可根据当前用户A对项目（历史购买商品）的评分和不同项目（其他商品）之间的相似度为用户A推荐商品。

**本章的重点是：**基于项目的协同过滤算法的理解和使用。

## 本章总结

本章实现的工作是：首先采用Python语言生成了“不同用户对多个项目的评分”的样本数据。然后对样本数据进行相似度计算，得到不同项目之间的相似度矩阵。进而得出某一用户对未打分项目的预测兴趣度，最终得到推荐的项目。

本章掌握的技能是：（1）在Python中获取“不同用户对多个项目的评分”数据；（2）使用NumPy库进行矩阵运算，计算不同项目之间相似度；（3）使用Python 计算用户对未接触项目的预测兴趣度。

## 本章作业

实现本章的案例，即生成“不同用户对多部电影的评分”样本数据，实现相似度计算、预测兴趣度计算和兴趣度排序。

按照不同用户对多部图书的评分数据，运用基于项目的协同过滤算法，为不同用户推荐图书。【绿色部分不要修改，以后会隐藏，用于实现章节号的自动编号】

第21章 基于用户的协同过滤

## 本章工作任务

采用基于用户的协同过滤算法编写程序，根据多个用户的电影喜好评分，为用户A推荐电影（将与A喜欢的电影类型相似的用户推荐给A）。（1）算法的输入是：“不同用户对多部电影的评分”数据（2）算法模型需要求解的是：用户的相似度（3）算法的结果是推荐给用户的电影。

## 本章技能目标

掌握基于用户的协同过滤原理

使用Python实现基于用户的协同过滤

## 本章简介

**协同过滤是指：**根据用户A与其他用户的共同喜好，为用户A推荐其可能喜好的算法。**基于用户的协同过滤是指：**找到与用户A喜欢的项目相似的用户，将其喜好的项目推荐给用户A。

**基于用户的协同过滤算法可以解决的实际应用问题是：**在电子商城中，协同过滤推荐算法被广泛应用，可根据当前用户A对项目（历史购买商品）的评分和用户间的相似度为用户A推荐商品。

**本章的重点是：**基于用户的协同过滤算法的理解和使用。

## 本章总结

本章实现的工作是：首先采用Python语言生成了“不同用户对多个项目评分”的样本数据。然后对样本数据进行相似度计算，得到不同用户之间的相似度。进而得出对某一用户未打分项目的推荐序列表。

本章掌握的技能是：（1）使用Python构造不同用户对多部电影的评分数据；（2）使用皮尔逊相关系数计算用户之间的相似度；（3）使用nearstUser函数，计算与当前用户最相似的n个用户；（4）使用recommand函数，将所有当前用户没有看过的电影，按照评分从高到低推荐给用户。

## 本章作业

实现本章的案例，通过不同用户对多部电影的评分数据，实现用户之间的相似度计算、获得最相似的n个用户并生成电影推荐列表。

按照下列表格中不同用户对多部图书的评分数据，运用基于用户的协同过滤算法，为不同用户推荐图书。我看了下要求是要粘贴一个连接过来，能保证读者可以从该链接直接下载数据资源。

# 【绿色部分不要修改，以后会隐藏，用于实现章节号的自动编号】

第23章 ARIMA

## 本章工作任务

采用ARIMA模型编写程序，对航空乘客数量进行预测。（1）算法的输入是：不同时刻对应的乘客数量；（2）算法模型需要求解的是：ARIMA模型的回归系数序列和，（3）算法的结果是：过去、未来更多时间点对应的乘客数量。

## 本章技能目标

掌握ARIMA原理

使用Python构造含有时间序列与对应的乘客数量

使用Python检验时间序列的稳定性

使用Python去除趋势和季节性因素的干扰

使用ARIMA模型对平稳的时间序列数据建立预测模型

使用Python求解预测模型

使用预测模型预测时间序列的未来值

使用Python将预测结果可视化

## 本章简介

**ARIMA是指：**利用时间序列的多个历史时刻对应的值预测未来时刻对应值的一种方法。AR为自回归算法，其特点是未来值与历史值之间存在线性组合关系。MA为移动平均算法，其特点是未来值与历史随机事件值（白噪声序列）之间存在线性组合关系。ARIMA中的I表示差分阶数，即当原数据非平稳时，采用1阶或多阶差分解决非平稳问题。

**ARIMA模型可以解决如下科学问题：ARIMA模型为：**

其中表示时刻的数据，即预测值；表示时刻的历史数据；表示常数项；表示时刻随机波动数据；表示时刻的数据对预测值的影响权重；表示时刻随机波动数据对预测值的影响权重；表示AR模型的阶数，表示MA模型的阶数。模型利用预测值的历史时间数据和与历史值无关的随机波动数据对时刻进行预测。

**ARIMA可以解决的实际应用问题是：**根据一个月以来的入院人数和出院人数预测未来时间的出院人数。例如：已知过去连续100天每天住院人数和出院人数，建立时间序列模型，预测未来10天出院的人数。

**本章的重点是：**ARIMA模型的理解和使用。

## 本章总结

本章实现的工作是：首先采用Python语言导入了不同时间点所对应的乘客量的时间序列数据。然后对数据进行差分和平稳化处理。将平稳的数据定阶后构建ARIMA模型并进行模型训练和预测，最后将预测值进行反n阶差分运算得到最终预测数据。

本章掌握的技能是：（1）使用ADF（单位根检验法）检验时序数据的平稳性（2）使用差分法对非平稳数据进行平稳化处理；（3）使用AIC/BIC对ARIMA模型进行定阶（AIC和BIC定阶：用于描述模型优化的程度，AIC/BIC的数值越小，阶数定的越好。）（4）使用反差分对数值进行计算（将差分的结果进行逆运算，将平稳序列还原为带趋势的序列，和原始数据保持一致）

## 本章作业

实现本章的案例，即对数据进行平稳化检验和处理，实现ARIMA的定阶、建模、预测和数据可视化。

设计一个时间、全国社会零售品总额的时间序列数据，运用ARIMA模型，实现半年后全国社会零售品总额的预测。

# 【绿色部分不要修改，以后会隐藏，用于实现章节号的自动编号】

第24章 神经网络（多层感知机MLP）

## 本章工作任务

采用多层感知机（Multilayer perceptron，MLP，一种神经网络）编写程序，解决分类和拟合两个问题，分类解决数字图像识别问题，拟合解决股票价格预测问题。（1）分类算法输入是：图像（个点，每个点用1或0表示黑或白）和标签值；拟合算法的输入是：由时刻和每个时刻对应的股票价格构成的时间序列。（2）算法模型需要求解的是：神经网络的权值。（3）分类算法的结果是：对新的图像分类；拟合算法的结果是：过去、期间、未来更多时刻对应的股票价格。

## 本章技能目标

掌握多层感知机（MLP）的基本原理（以BP神经网络为例）

使用Python构造图像（0、1矩阵）数据和股票价格的时间序列数据

使用Python多层感知机对数字图像分类

使用Python多层感知机预测更多时刻的股票价格

使用Python对神经网络分类和预测的结果进行可视化展示

## 本章简介

**人工神经网络（Neural Network）是指：**模仿人脑生物结构、人脑学习过程、人脑分类与预测过程的数学模型。神经网络的结构特点是：由多层神经元构成，各层神经元之间相互连接，相互连接的两个神经元之间的影响力用连接权值表示。神经网络的工作原理特点是：神经网络通过误差反向传播、梯度下降等方法进行学习（训练），通过前向计算（从输入层到输出层）等方法计算出预测值（分类和拟合的结果）。经典的神经网络包括：多层感知机、BP神经网络、RBF神经网络、hopfield神经网络、卷积神经网络、循环神经网络等。

**多层感知机（MLP）是指：**一种经典的神经网络（结构简单），其特点是：网络从前往后，每一层的输入都是前一层的加权，每一层都存在激活函数。

**BP神经网络是指：**用了“BP算法”进行训练的“多层感知机模型”。在本章中以BP神经网络为例讲解多层感知机。BP神经网络通过信息前向传播，误差反向传播训练神经网络的训练，使用梯度下降的方法调整参数，使得预测误差逐渐减小，达到预测效果最佳的目的。本章中使用sigmoid函数作为每层神经元之间的激活函数，具体解释参见后文算法详解部分。

**多层感知机可以解决如下科学问题：**

变量定义：已知一共有个训练样本。每个样本（例如第个训练样本）有个特征值，记做。每个样本（例如第个训练样本）对应一个标签值（在分类问题中，特征值是样本的各属性值，标签值是样本的类别信息；在拟合问题中，特征值是样本的自变量，标签值是样本的因变量），记做。

算法功能：希望神经网络对训练样本的特征值和标签值进行学习（训练），从而使得神经网络能得到一组权值，使得对训练样本的预测值接近训练样本的标签值（获得最小的损失函数），即对样本数据特征值与标签值之间的映射规律进行提取。进而根据上述学习结果（得到的权值），通过对测试样本的特征值进行前向计算，得到该测试样本的预测值。

**多层感知机可以解决的实际应用问题是：**（1）分类问题：通过对10张写有0-9不同数值的图像（每个数值对应一个类别）进行训练，训练结束后，算法可以实现对一张新的图像进行分类。（2）拟合问题：通过对股票的时刻-股价数据进行学习，根据学习的结果，推算出任意时刻对应的股价。

**本章的重点是：**多层感知机（以BP神经网络为例）的理解和应用。

## 本章总结

本章实现的工作是：首先采用Python语言生成了用于数字图像检测的训练样本输入值矩阵、输出值和用于股票价格预测的时间、对应股价数据。然后使用训练样本，对神经网络进行训练，得到神经网络模型。进而输入更多的测试样本，采用训练后的神经网络进行预测。最后将预测结果进行可视化。

本章掌握的技能是：（1）使用NumPy库生成连续的时刻数值和正弦数值数组；（2）使用sklearn.neural\_network库中多层感知机MLPClassifier和MLPRegressor进行分类计算和拟合计算；（3）使用Matplotlib库实现数据的可视化，绘制散点图、折线图。

## 本章作业

实现本章的案例，即生成样本数据，实现BP神经网络模型的建模、预测和数据可视化。

设计一个模拟空间内有人物、物体和没有人物或物体的矩阵数据（同数字图像模拟方法），运用BP神经网络算法，实现图像的识别，达到输入数列，能够输出图像类别的目的。

# 【绿色部分不要修改，以后会隐藏，用于实现章节号的自动编号】

第25章 XPath

## 本章工作任务

采用XPath语言编写程序，对人民网新闻详情页的信息（新闻的标题、时间、来源和内容）进行抓取。（1）算法的输入是：需要抓取的新闻页面的URL地址;（2）算法模型需要求解的是：目标信息所在网页文件中的特征标签和属性名称;（3）算法的结果是：URL网页中所有的新闻标题、时间、来源和内容，以列表的形式呈现。

## 本章技能目标

掌握爬虫原理

掌握XPath （XML路径语言）方法原理

使用requests包实现对网页HTML代码的获取

使用XPath的方法实现对HTML中内容的抓取

使用pandas库实现对抓取结果的文件存储

## 本章简介

**数据抓取是指：**一种将非结构化的数据（网站、微博等网页中的网络文字、图片等），按照一定规则进行获取，并将获取的数据以结构化数据（表格等）的形式进行保存和应用的技术。**XPath是指：**一种用来确定XML文档（网页文档）中某个标签位置（例如标题在网页文档中所属标签的位置）的查询语言。例如：若网页的源代码为<标签开始符号 属性名=”属性值”>标签内容<标签结束符号>格式，可通过编写代码“对象.方法('标签开始符号', 属性名='属性值')”定位内容标签位置，通过后续方法，获取标签内容。

**XPath可以解决如下科学问题：**对于网页内容，如果内容与标签存在如下关系：<标签开始符号 属性名=”属性值”

>标签内容<标签结束符号>，则根据标签符号，将“标签内容”进行提取。

**XPath可以解决的实际应用问题：**例如我们想获得人民日报网站上的各新闻信息（包括新闻标题、时间、来源和内容信息），则可以使用XPath语言进行逐一定位并获取，并将结果放入列表（List）中，并保存为文件。

**本章的重点是：**XPath中各方法的理解和使用

## 本章总结

本章实现的工作是：本章实现的工作是：首先对人民网新闻的标题、时间、来源和内容部分的路径进行定位，确定待抓取内容所在页面的URL地址然后分析待抓取内容的特征，根据特征编写XPath抓取程序对新闻的标题、时间、来源和内容进行抓取，最后将抓取结果存储为.xls和.csv格式文档。

本章掌握的技能是：（1）使用requests包获取全部网页内容；（2）通过模拟浏览器访问网页的方法绕过反爬虫机制；（3）使用Lxml解析，将获取到的网页的内容解析成html格式：（4）使用XPath定位信息所在路径并抓取，将抓取结果通过pandas库存入csv与xls。

## 本章作业

实现本章的案例，即根据给定的网址，用XPath语言抓取人民网新闻的标题、时间、来源和内容。

抓取新浪新闻详情页最新的一条新闻内容，并将结果导入.csv和.xls文件中。

# 【绿色部分不要修改，以后会隐藏，用于实现章节号的自动编号】

第26章 Beautiful Soup

## 本章工作任务

本章运用Beautiful Soup对象提供的方法编写程序，对人民日报网详情页中的新闻相关信息进行爬取。（1）算法的输入是：人民日报网新闻详情页网址；（2）网络爬虫需要配置的参数是：新闻详情页各部分信息的特征标签及其属性值；（3）算法的结果是：该新闻详情页中的所有新闻发布时间、标题与内容，以列表的形式呈现。

## 本章技能目标

掌握爬虫原理

掌握Beautiful Soup对象的使用方法

使用Requests对象实现网页内容的获取

使用Beautiful Soup对象中的方法实现对网页内容的提取

使用pandas包实现对抓取数据的文件存储

将抓取结果进行可视化展示

## 本章简介

**数据抓取是指：**一种将非结构化的数据（网站、微博等网页中的网络文字、图片等），按照一定规则进行获取，并将获取的数据以结构化数据（表格等）的形式进行保存和应用的技术。**Beautiful Soup是指：**一种含有爬取数据方法的对象，使用者可通过Beautiful Soup对象中各类方法，对获取到的HTML和XML的树状结构进行解析，并且以每一个节点为对象，对其中内容进行定向查找与获取。例如：若网页的源代码为<标签开始符号 属性名=”属性值”>标签内容<标签结束符号>格式，可通过编写代码“Beautiful Soup对象.匹配方法('标签名称', 属性名='属性值')”定位到具有指定属性名称的标签，获取该标签中的内容。

**Beautiful Soup可以解决如下科学问题：**对于网页内容，如果内容与标签存在如下关系：<标签开始符号 属性名=”属性值”>标签内容<标签结束符号>，则根据标签符号及属性值，将“标签内容”进行提取。

**Beautiful Soup算法可以解决的实际应用问题是：**例如我们想获得人民日报网站上的新闻信息，可以使用Beautiful Soup对象中的方法进行逐一抓取，将结果放入列表（List）中，并保存为文件。

**本章的重点是：**Beautiful Soup对象中各方法的理解和使用。

## 本章总结

本章实现的工作是：首先采用Requests对象获取人民日报网的单条新闻详情页，将网页HTML解析为树状结构文本。然后使用Beautiful Soup对象中的一系列方法，对所需内容进行抓取，每一条新闻以字典形式存储，不同新闻构成了字典数组。采用pandas包，将抓取到的内容存入后缀名为csv与后缀名为xls的文件。

本章掌握的技能是：（1）使用Requests包获取网页HTML代码（树状结构）；（2）使用Beautiful Soup对象中的方法对HTML代码中的关键信息进行提取；（3）使用pandas包实现将抓取的数据存入后缀名为csv或后缀名为xls的文件。

## 本章作业

实现本章的案例，即生成样本数据，实现一元线性回归模型的建模、预测和数据可视化。

设计抓取程序，采用Beautiful Soup对象，抓取首都经济贸易大学教务处的通知信息，并且将上述通知信息（标题，时间，内容）可视化，并存入Excel文档中。

# 【绿色部分不要修改，以后会隐藏，用于实现章节号的自动编号】

第27章 Tkinter

## 本章工作任务

采用Tkinter图形用户界面模块（GUI），实现案例1加法计算器的操作和结果显示、案例2“时间-股票价格”预测工具的配置和结果显示。（1）算法的输入是：案例1加法器的输入是通过文本框输入的两个加数，案例2股票预测功能的输入是预测算法的阶数（通过文本框输入）和“时间-股票价格”样本数据。（2）Tkinter模型是：按钮的事件响应机制。（3）算法的结果是：案例1加法器输出的是两个加数的和，并显示在一个文本框中，案例2预测功能输出的是根据输入的阶数计算出的预测的股票价格，最后将时间-股票预测价格可视化。

## 本章技能目标

掌握Tkinter模块图形界面开发的基本原理

掌握输入框、按钮等Tkinter控件对象布局，实现数据的界面输入

掌握按钮的单击事件的实现机制

掌握通过文本框、图像框实现计算结果的可视化

## 本章简介

**Tkinter模块是指：**一种Python语言中最常用的图形开发界面库，用户可以使用Tkinter模块快速创建图形用户界面（GUI）。

**Tkinter可以解决如下科学问题：**为用户提供图形用户界面（GUI）中的按钮，画布，框架，标签，菜单，滚动条，输入等基础控件。

**Tkinter可以解决的实际应用问题是：**案例1用户可以通过图形界面输入两个加数，通过点击按钮调用两个数相加求和并显示结果的方法，将结果显示在文本框中，实现加法器；在案例2中，用户可以通过图形界面输入一元线性回归模型的阶数，来预测出更多（未来等）时刻的股票价格，并将预测结果以散点图、折线图的形式显示在用户界面的图像窗口中，实现股票价格的预测与显示。

**本章的重点是：**Tkinter模块的理解和使用。

## 本章总结

本章实现的工作是：首先采用Tkinter模块实现各种控件的布局，根据用户的输入，在图形用户界面中完成加法计算。然后根据用户输入的阶数进行“时间-股票价格”模型的创建并求解，并对测试集数据进行预测，最终将预测结果绘制在图形用户界面（GUI）中。

本章掌握的技能是：（1）使用Tkinter模块实现控件布局；（2）使用Tkinter模块实现事件响应（加法器事件和股票价格预测事件）；（3）使用Tkinter模块实现数据可视化。

## 本章作业

实现本章的案例，即完成加法器和股票价格预测器。

设计一个科学计算器，用户输入相关数值后，可以实现四则运算、阶乘和正余弦等计算。

设计一个房价预测器，应用NumPy包中的arange方法随机生成一组“时间-房价”训练集数据，用户输入回归阶数，完成模型创建及拟合后，随机生成更多的时间数据，对房价进行预测，最后实现预测结果的可视化。

五、考核方式、成绩评定

1. 课程考核应掌握的知识点和技能

需要掌握的知识点：数据科学算法的原理与实现，包括大数据的处理、回归分析、分类分析、聚类分析、降维与特征提取、推荐与协同过滤、预测与时间序列预测、神经网路与深度学习、智能系统集成。

需要掌握的技能：理解并实现上述知识点，独立完成数据科学算法相关毕业设计和企业真实项目。

2．考试形式

（1）论文审核：模拟毕业设计论文

（2）答辩考核：模拟毕业设计答辩、模拟应聘面试

3．考试时间

（1）第15教学周提交论文、模拟毕业设计答辩（20%不通过，需二次答辩）

（2）第16教学周模拟应聘面试、二次答辩

六、主要参考书及其他内容

[1] Harrington,Peter.《Machine Learning in Action》[M]. Manning Publications, 2015.

[2] 刘经纬.《Python人工智能与大数据》[M].机械工业出版社(北京). 首都经济贸易大学出版社，2020.

执笔人：刘经纬 教研室主任：范烺 系教学主任审核签名：范烺