



任务1.1 认识数据分析

目录

1	任务描述
2	知识准备

1.1 认识数据分析

【任务描述】

数据分析是大数据技术的重要组成部分。近年来，随着大数据技术的发展，数据分析技能也是被认为是数据科学领域中数据从业人员需要的技能之一。同时数据分析师也成为当下最热门的职业之一。掌握数据分析技能是一个循序渐进的过程，明确数据分析概念、流程等相关知识是掌握数据分析的第一步。

【知识准备】

- **1.掌握数据分析的概念**

- 数据分析是指用适当的分析方法对收集来的大量数据进行分析，提取有用信息和形成结论，对数据加以详细研究和概括总结的过程。随着计算机技术的全面发展，企业生产、收集、存储和处理数据的能力大大提高，数据量与日俱增。而在现实生活中，需要将这些繁多、复杂的数据通过统计分析进行提炼，以此研究出数据的发展规律，进而帮助企业管理层做出决策。

- 数据分析是指根据分析目的，采用对比分析、分组分析、交叉分析和回归分析等分析方法，对收集的数据进行处理与分析，提取有价值的信息，发挥数据的作用，得到一个特征统计量结果的过程。

【知识准备】

- **2.掌握数据分析的流程**

- 数据分析已经逐渐演化成为一种解决问题的过程，甚至是一种方法论。虽然每个公司都会根据自身需求和目标创建最适合的数据分析流程，但是数据分析的核心步骤是一致的。

【知识准备】

- 1. 数据获取

- 数据获取是数据分析工作的基础，是指根据需求分析的结果提取、收集数据。数据获取主要有两种方式：网络数据与本地数据。网络数据是指存储在互联网中的各类视频、图片、语音和文字等信息。本地数据则是指存储在本地数据库中的生产、营销和财务等系统的数据。本地数据按照数据时间又可以划分为两部分，分别是历史数据与实时数据。历史数据是指系统在运行过程中遗存下来的数据，其数据量随系统运行时间的增加而增长；实时数据是指最近一个单位时间周期（月、周、日、小时等）内产生的数据。在数据分析过程中，具体使用哪种数据获取方式，需要依据需求分析的结果而定。

【知识准备】

• 2. 数据预处理

- 数据预处理是指对数据进行数据合并、数据清洗、数据标准化和数据变换，并直接用于分析建模的这一过程的总称。其中，数据合并可以将多张互相关联的表格合并为一张；数据清洗可以去除重复、缺失、异常、不一致的数据；
- 数据标准化可以去除特征间的量纲差异；数据变换则可以通过离散化、哑变量处理等技术满足后期分析与建模的数据要求。在数据分析的过程中，数据预处理的各个过程互相交叉，并没有明确的先后顺序。

【知识准备】

- 3. 分析与建模
 - 分析与建模是指通过对比分析、分组分析、交叉分析、回归分析等分析方法，以及聚类模型、分类模型、关联规则、智能推荐等模型与算法，发现数据中有价值的信息，并得出结论的过程。
 - 分析与建模的方法按照目标不同可以分为几大类。如果分析目标是描述客户行为模式的，那么可以采用描述型数据分析方法，同时还可以考虑关联规则、序列规则和聚类模型等。如果分析目标是量化未来一段时间内某个事件发生概率的，那么可以使用两大预测分析模型，即分类预测模型和回归预测模型。在常见的分类预测模型中，目标特征通常为二元数据，如欺诈与否、流失与否、信用好坏等。在回归预测模型中，目标特征通常为连续型数据，常见的有股票价格预测等。

【知识准备】

- 4. 模型评价与优化

- 模型评价是指对于已经建立的一个或多个模型，根据其模型的类别，使用不同的指标评价模型性能优劣的过程。常用的聚类模型评价指标有ARI评价法（兰德系数）、AMI评价法（互信息）、V-measure评分、FMI评价法和轮廓系数等。常用的分类模型评价指标有准确率（Accuracy）、精确率（Precision）、召回率（Recall）、F1值（F1 Value）、ROC和AUC等。常用的回归模型评价指标有平均绝对误差、均方误差、中值绝对误差和可解释方差值等。

- 模型优化则是指模型性能在经过模型评价后已经达到了要求，但在实际生产环境应用过程中，发现模型的性能并不理想，继而对模型进行重构与优化的过程。多数情况下，模型优化和分析与建模的过程基本一致。

【知识准备】

• 3.熟悉python数据分析工具

• 1. 了解数据分析常用工具

- 目前主流的数据分析语言主要有Python、R、MATLAB这3种。其中，Python具有丰富和强大的库，同时Python常被称为胶水语言，能够将使用其他语言制作的各种模块（尤其是C/C++）轻松地连接在一起，是一门更易学、更严谨的程序设计语言。R语言通常用于统计分析、绘图。R是属于GNU系统的一个自由、免费、源代码开放软件。MATLAB的作用是进行矩阵运算、绘制函数与数据、实现算法、创建用户界面和连接其他编程语言的程序等，其主要应用于工程计算、控制设计、信号处理与通信、图像处理、信号检测、金融建模设计与分析等领域。
- Python数据分析的主要包括以下5个方面的优势。

【知识准备】

- **（1）语法简单精练。**对于初学者来说，比起其他编程语言，Python更容易上手。
- **（2）含有大量功能强大的库。**结合在编程方面的强大实力，可以只使用Python这一种语言去构建以数据为中心的应用程序。
- **（3）功能强大。**从特性观点来看，Python是一个混合体。丰富的工具集使Python介于传统的脚本语言和系统语言之间。Python不仅具备所有脚本语言简单和易用的特点，而且提供了编译语言所具有的高级软件工程工具。
- **（4）Python是一门胶水语言。**Python程序能够以多种方式轻易地与其他语言的组件“粘接”在一起。例如，Python的C语言 API可以帮助Python程序灵活地调用C程序，这意味着用户可以根据需要给Python程序添加功能或在其他环境系统中使用Python。

【知识准备】

- 2. 了解python数据分析常用库
- 使用Python进行数据分析时所用到的类库主要有Numpy、pandas、Matplotlib、seaborn、pyecharts、scikit-learn 等。

【知识准备】

- 1) Numpy
 - Numpy是Numerical Python的简称，是一个Python科学计算的基础包。Numpy主要提供了以下内容。
 - (1) 快速高效的多维数组对象ndarray。
 - (2) 对数组执行元素级计算和直接对数组执行数学运算的函数。
 - (3) 读/写硬盘上基于数组的数据集的工具。
 - (4) 线性代数运算、傅里叶变换和随机数生成的功能。
 - (5) 将C、C++、Fortran代码集成到Python的工具。
 - 除了为Python提供快速的数组处理能力外，Numpy在数据分析方面还有另外一个主要作用，即作为算法之间传递数据的容器。
 - 对于数值型数据，使用Numpy数组存储和处理数据要比使用内置的Python数据结构高效得多。
 - 此外，由低级语言（如C和Fortran）编写的库可以直接操作Numpy数组中数据，无须进行任何数据复制工作。

【知识准备】

- 2) pandas

- pandas是Python的数据分析核心库，最初被作为金融数据分析工具而开发出来。pandas为时间序列分析提供了很好的支持，它提供了一系列能够快速、便捷地处理结构化数据的数据结构和函数。Python之所以成为强大而高效的数据分析环境与它息息相关。
- pandas兼具Numpy高性能的数组计算功能以及电子表格和关系型数据库（如SQL）灵活的数据处理功能。它提供了复杂精细的索引功能，以便完成重塑、切片与切块、聚合和选取数据子集等操作。pandas将是本书中使用的主要工具。

【知识准备】

• 3) Matplotlib

- Matplotlib是较为流行的用于绘制数据图表的Python库，是Python的2D绘图库。Matplotlib的操作比较容易，用户只需用几行代码即可生成直方图、功率谱图、条形图、错误图和散点图等图形。Matplotlib提供了pylab的模块，其中包括了Numpy和pyplot中许多常用的函数，方便用户快速进行计算和绘图。Matplotlib与IPython的结合，提供了一种非常好用的交互式数据绘图环境。Matplotlib绘制的图表也是交互式的，读者可以利用绘图窗口中工具栏中的相应工具放大图表中的某个区域，或对整个图表进行平移浏览。

【知识准备】

- 4) **seaborn**

- seaborn是基于Matplotlib的图形可视化Python库，它提供了一种高度交互式界面，便于用户能够做出各种有吸引力的统计图表。seaborn是在Matplotlib的基础上进行了更高级的API封装，使得作图更加容易。seaborn不需要了解大量的底层代码，即可使图形变得精致。在大多数情况下，使用seaborn能做出很具有吸引力的图，而使用Matplotlib能制作具有更多特色的图。因此，可将seaborn视为Matplotlib的补充，而不是替代物。同时，seaborn能高度兼容Numpy与pandas数据结构以及scipy与statsmodels等统计模式，可以在很大程度上帮助用户实现数据可视化。

【知识准备】

• 5) pyecharts

- Echarts是一个由百度 开源的数据可视化工具，凭借着良好的交互性，精巧的图表设计，得到了众多开发者的认可。而Python是一门富有表达力的语言，很适合用于数据处理。pyecharts是Python与Echarts的结合。pyecharts可以展示动态交互图，对于展示数据更方便，当鼠标悬停在图上时，即可显示数值、标签等。pyecharts支持主流Notebook环境，如Jupyter Notebook、JupyterLab等；可轻松集成至Flask、Django等主流Web框架；高度灵活的配置项，可轻松搭配出精美的图表。

- pyecharts囊括了30多种常见图表，如Bar（柱形图/条形图）、Boxplot（箱形图）、Funnel（漏斗图）、Gauge（仪表盘）、Graph（关系图）、HeatMap（热力图）、Radar（雷达图）、Sankey（桑基图）、Scatter（散点图）、WordCloud（词云图）等。

【知识准备】

- 6) **scikit-learn**

- `scikit-learn`是一个简单有效的数据挖掘和数据分析工具，可供用户在各种环境下重复使用。`scikit-learn`建立在Numpy、SciPy和Matplotlib基础之上，对一些常用的算法方法进行了封装。目前，`scikit-learn`的基本模块主要有数据预处理、模型选择、分类、聚类、数据降维和回归6个。在数据量不大的情况下，`scikit-learn`可以解决大部分问题。对算法不精通的用户在执行建模任务时，并不需要自行编写所有的算法，只需要简单地调用`scikit-learn`库里的模块即可。

任务1.2 Anaconda环境搭建

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/338034004103006073>