探索性数据分析

目　　录

第一章 探索性数据分析简介

本章结构

1.引言

2.探索性数据分析的定义

3.常用的数据变换

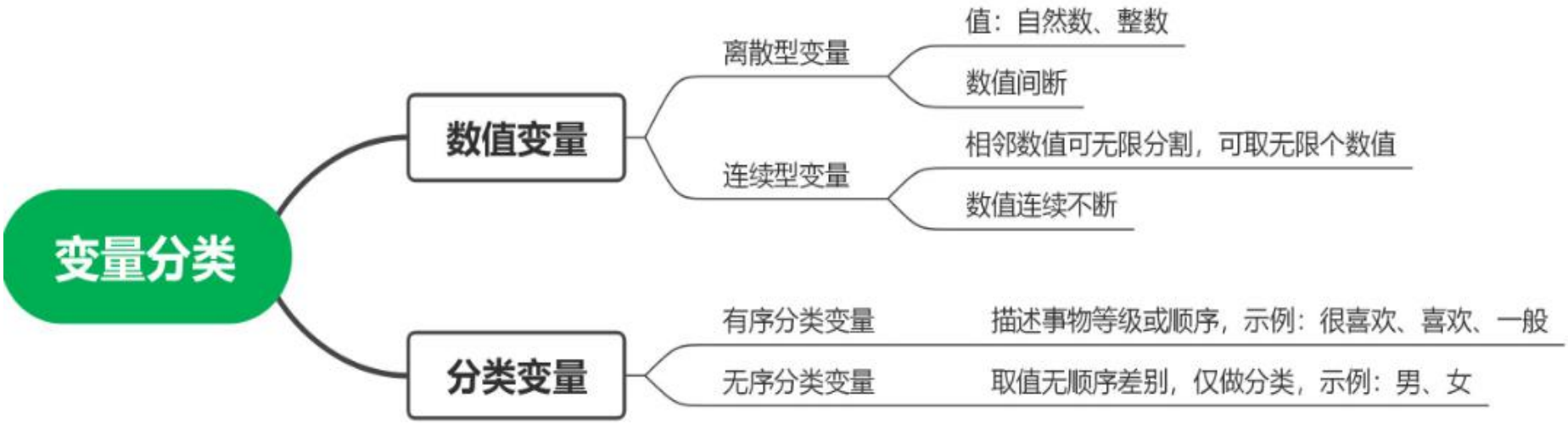
4.所用教材及软件

1. 引言

- 数据就是承载了信息的东西。
- 例如，数字、文本、图形、音频、视频、网页等。
- 对数据进行观察、研究，寻找其蕴含的规律，这就是数据分析。
- 数据分析的目的就是在本来彼此错综复杂的或者大量看似不相关的数据之间找到内在的、本质的、起作用的规律或特性。

1. 引言

- 数据
  - 数值型数据
  - 分类型数据
    - 顺序型数据
    - 名义型数据



1. 引言

- 数据分析更多注重数据指标的建立，数据的统计，数据之间的联系，数据的深度挖掘和机器学习。

- 探索性数据分析是对已有数据在尽量少的先验假设下通过作图、制表、方程拟合、计算特征量等手段探索数据的结构和规律的一种数据分析方法。

1. 引言

- 本学期，《探索性数据分析》课程32学时，2学分，学科基础课，必修课。
- 主要教学内容包括：
  - 第一章 探索性数据分析简介
  - 第二章 统计分析
  - 第三章 数据可视化
  - 第四章 方差分析
  - 第五章 典型相关分析
  - 第六章 判别分析
  - 第七章 聚类分析
  - 第八章 降维分析

1. 引言

- 通过该课程的教学，能够掌握探索性数据分析的基本概念、基本知识和基本方法，熟悉利用Matlab或者Python软件进行特征提取，培养运用探索性数据方法分析数据的能力。

2. 探索性数据分析的定义

- 探索性数据分析(Explorative Data Analysis, EDA)就是在较少预设或没有预设的前提下对数据进行分析。

- 探索性数据分析是对已有数据在尽量少的先验假设下通过作图、制表、方程拟合、计算特征量等手段探索数据的结构和规律的一种数据分析方法。

2. 探索性数据分析的定义

- 探索性数据分析(Explorative Data Analysis, EDA)属于统计学和数据分析，其思路是先探索数据，常采用描述性统计学、 科学可视化、 数据巡查、 降维等方法。这种探索没有任何预设观点或假设。

- 相反，这种方法使用探索的结果来引导和展开后续的假设检验和建模等。它与数据挖掘领域紧密关联，所讨论的很多EDA工具是知识发现和数据挖掘工具的一部分。

2. 探索性数据分析的定义

- 1977年，John W. Tukey是首个详细描述探索性数据分析(Explorative Data Analysis, EDA)的统计学家之一。
- 他定义EDA为“探查工作—数值探查工作—或计数探查工作或图形探查工作”。
- 这个数据分析体系的重点在于：研究人员在没有任何预先设想的情况下检视数据，以发现数据可以如何解释所研究的对象。

2. 探索性数据分析的定义

- Tukey对比了EDA和验证性数据分析(Confirmatory Data Analysis,CDA),CDA主要关注统计假设检验、 置信区间及估计等。

- 1979年，Hartwig和Dearing从社会学角度写了一本简短易读的关于EDA的书。他们认为：CDA模式会回答比如“数据能够验证假设XYZ吗？”这样的问题，而EDA往往会问“关于XYZ的关系，数据能够告诉我什么”。

2. 探索性数据分析的定义

- 1980年，Tukey扩展了他的将探索性和验证性数据分析结合在一起的思想，提出了一种典型的CDA线性方法学，步骤如下：
- 1.叙述所研究的问题；
- 2.设计实验解决问题；
- 3.根据设计的实验收集数据；
- 4.对数据进行统计分析；
- 5.得到答案。

2. 探索性数据分析的定义

- 1982年，Hoaglin在统计科学百科全书提供了EDA概述。他描述EDA技术为“灵活地寻找线索和证据”，称验证性数据分析为“评价了现有证据”。

- 1985年，Chatfield讨论了EDA的一些思想及其对统计教学的重要性。他将这个主题称为初始数据分析(Initial Data Analysis)或称IDA。Chatfield赞同EDA在数据分析中从无假设的方法开始，他还强调需要了解数据如何收集，分析的目的是什么，以及使用EDA/IDA作为整体方法进行统计推断。

3. 常用的数据变换

➢ 数据变换、预处理的目的

- 对于数据问题而言，在进行数据分析前，一般都需要对数据进行预处理，主要有 2 个目的：

（1）**无量纲化：**不同属性的数据往往具有不同的量纲，即使对同一属性，采用不同的计量单位，其数值也不同。因此，数据分析前，需对数据进行无量纲化

（2）**归一化：**不同属性的数据，其取值在数值大小有很大的差异，因此需要对数据进行预处理，将不同属性的数据变换到可比较的数值大小，可通过归一化处理，变换到[0, 1]

3. 常用的数据变换

- **线性变换**

对于原始数据矩阵为 $A=(a_{ij})_{m\times n}$，其每一列代表不同的数据属性，

$i=1,\cdots,m$，$j=1,\cdots,n$。设 $a_j^{\max}$ 是矩阵第 j 列中的最大值，则

$$b_{ij}=a_{ij}/a_j^{\max}$$

- 上述线性变换，可同时实现“无量纲”和“归一化”

3. 常用的数据变换

- **标准 0 - 1 变换**

与线性变换类似，也可以对于原始数据矩阵为 $A=(a_{ij})_{m\times n}$，进行标准 0 - 1 变换，$a_j^{\max}$和$a_j^{\min}$分别是矩阵第 j 列中的最大值和最小值，则

$$b_{ij}=\frac{a_{ij}-a_j^{\min}}{a_j^{\max}-a_j^{\min}},$$

- 同样，可同时实现“无量纲”和“归一化”

3. 常用的数据变换

- **规范化处理**

无论成本型属性还是效益型属性，向量规范化均用下式进行变换

$$b_{ij}=a_{ij}\Big/\sqrt{\sum_{i=1}^{m}a_{ij}^{2}}\ ,\ i=1,\cdots,m\ ,\ \ j=1,\cdots,n.$$

3. 常用的数据变换

- **标准化处理**

在实际问题中，每个变量都具有同等的表现力，可对数据进行标准化处理，即

$$b_{ij}=\frac{a_{ij}-\bar{a}_j}{s_j},\ i=1,2,\cdots,m,\ j=1,2,\cdots,n$$

其中$\bar{a}_j=\frac{1}{m}\sum_{i=1}^{m}a_{ij}$，$s_j=\sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(a_{ij}-\bar{x}_j)^2}$，$j=1,2,\cdots,n$。

4. 所用教材及软件

- 教材：
- 1.《Matlab数据探索性分析》，温迪等著，迟冬祥等翻译，北京：清华大学出版社，2018年9月。
- 2.《应用统计方法》，常兆光等编著，北京：石油工业出版社，2009年11月。
- 3.《数据分析方法》，梅长林等编，北京：高等教育出版社，2006年2月.
- 4.《Python程序设计—从基础开发到数据分析》，夏敏捷等，北京：清华大学出版社，2019年7月

4. 所用教材及软件

- **5.《Python数据分析与实践》，柳毅等，北京：清华大学出版社，2019年6月。**

- **应用软件：** Python, Matlab， R