目录

2 数据理解

由欧盟机构的跨行业数据挖掘标准流程可以看出, 数据理解是整个数据分析流程中非常重要的一环, 就像医生诊断时需要详细地了解病人的病情, 或者厨师做料理前应仔细考察手中的原材料. 这里所说的数据理解, 就是要理解数据所拥有的特点. 统计学最大的一个特点是“分而治之”, 即针对特定的分析目标和数据特点, 发展和应用特定的统计学方法. 只有把握数据特点, 才能选择恰当的数据预处理及后续的数据分析方法.

2.1 数据特点

可以从多个角度分析数据特点, 包括但不限于以下几个重要的角度:

(1) 来源. 数据是个人或企业收集的一手资料, 还是别人收集好甚至经过加工整理后的二手资料? 是调查数据、实验数据还是观察数据? 在哪些地点, 什么时间段, 由谁, 用什么设备收集的? 若我们试着还原数据收集现场和整个收集过程, 可能会对理解数据很有帮助.

(2) 范围. 数据涵盖的群体范围是什么, 是总体数据, 还是样本数据? 若是样本数据, 其对总体的代表性如何? 若存在样本结构与总体结构不相符的问题, 是否需要在分析中加权调整? 这些问题都需要在预处理和分析中予以考虑.

(3) 结构性. 结构化数据具有明确的测量维度, 例如消费者刷一次银行卡, 可以记录刷卡时间、地点、商户名、金额等数据. 对于结构化的数据, 我们需要考察变量是否准确、完整地反映了分析目标. 非结构化数据通常不具有明确的维度, 例如图片、语音和视频等, 需要借助于一定的

转换才能成为计算机可处理的结构化数据. 此时, 我们需要考量转换方法和转换结果的合理性. 当然, 结构化数据也并非都可直接应用于建模, 可能需要先作特征衍生等预处理.

(4) 时序关系. 按照时序关系, 可以将数据分为横截面数据、时间序列数据和面板数据. 横截面数据是指在一个时间点截取的多个对象的数据, 它们仅反映一个时间点的情况. 时间序列数据是指获取了某一对象多个时间点的数据, 这些数据一般按时间顺序排列, 形成一个序列. 面板数据则综合了横截面数据和时间序列数据. 时序关系是设计数据分析方案最重要的线索之一, 针对每一类时序关系都有专门的统计学方法.

(5) 粒度. 主要可从时间和空间两个角度来看. 从时间粒度看, 数据可以是高粒度、高频的,例如以毫秒或秒为时间间隔单位采集数据, 也可以是低粒度、低频的, 例如以月、季度和年为时间间隔单位. 从空间粒度看, 可以按国家、省、市、县、镇、小区、家庭等不同的粒度采集数据. 按照粒度进行数据上卷或下钻, 可以压缩或扩张数据, 这也是数据预处理的一部分.

(6) 同质性. 数据之间是同质还是异质, 这可能极大地影响了分析效果. 例如, 缺失的数据与已经观测到的数据是否同质, 决定了缺失值的处理方式. 此外, 当我们对不同数据源的数据进行合并时, 也需要考虑它们是否同质, 是否可以直接合并.

(7) 规模. 一般从两个角度了解数据规模. 第一个角度是样本量规模, 它影响了预处理和分析方法的选择. 例如, 对于样本量特别大的数据, 需要考虑是否通过采样等预处理进行数据压缩. 而对于样本量特别小的数据, 例如少数类样本, 可以考虑在预处理阶段衍生数据. 第二个角度是变量维数规模, 它对数据预处理的影响尤其大. 对于高维或超高维数据, 一般需要先作特征选择等预处理.

(8) 计量尺度. 统计学“分而治之”最重要的一个划分标准就是变量的类型, 即变量通过何种计量尺度获得. 将在2.2节具体阐述.

(9) 质量. 数据的质量直接影响了数据预处理工作, 数据预处理最重要的目的就是改进数据质量. 将在2.3节具体阐述.

2.2 计量尺度

不同类型的变量来源于不同的计量尺度. 变量类型直接决定了数据预处理与后续分析方法的选择. 数据的计量尺度包含两大类, 定性尺度和定量尺

度, 其中定性尺度可分为定类尺度和定序尺度, 定量尺度可分为定距尺度和定比尺度.

(1) 定类尺度(nominal scale)

定类尺度又称为类别尺度, 是最粗略、计量层次最低的一种尺度. 在定类尺度下, 按照事物的某种属性对其进行平行的分类或分组, 注意类别应穷尽和互斥. 各类别可以指定数字代码表示, 但这里的数字表示类别的编号, 不具有量化的涵义. 由定类尺度所刻画的变量即为定类变量, 其值仅具有=或≠的数学特性, 不能比较大小, 也不能进行加减乘除运算. 如对于“性别”这一变量, 以”1”代表男, ”2”代表女. 我们可以说1不等于2, 即二者性别不同, 但不能比较1和2的大小, 也不能进行加减乘除运算.

(2) 定序尺度(ordinal scale)

定序尺度又称为顺序尺度, 是对事物之间等级或顺序差别的测度. 在定序尺度下, 人们在对事物分类的同时, 给出各类别的顺序. 因而, 定序尺度比定类尺度更精确. 在定序尺度下, 类别之间虽然有序, 但它们之间的差异无法准确度量, 相邻类别的差异度也未必相同. 由定序尺度所刻画的变量即为定序变量, 其值具有=或≠, >或<的数学特性. 请注意, 定序变量的取值不能进行加减乘除运算. 例如, 在使用3级量表的满意度调查中, 以1、2、3分别表示“不满意”、“一般”、“满意”, 可以比较两位被调查者的满意度等级是否相同, 也可以比较谁的等级更高, 但不能通过减法运算获得等级之间的差异, 也不能由加法运算获得平均满意度, 等等. 但是, 由于定序变量的分析方法较少, 人们常将定序变量转换为定量变量使用. 这被称为等级变换(ranking transformation), 是数据变换的一种. 实际上, 作等级变换是基于两个相邻等级之间差距为1的假定. 我们应该意识到这种潜在的假定及其对数据分析的影响.

(3) 定距尺度(interval scale)

定距尺度又称为间隔尺度, 是对事物间间隔的准确测度, 比定序尺度精确. 在了解定距尺度的特点之前, 有必要先理解绝对零点.

定义2.1. 绝对零点是一个零值, 它代表所测量的物体的特征不能出现或不能观测到的一点, 即零值表示“没有”或“不存在” (nil or nothing).

由定距尺度所刻画的变量即为定距变量, 它没有绝对零点, 也就是说, 零值并不代表特定的“没有”或“不存在”. 这样的零值往往是不固定的, 可以设置和改变. 例如, 华氏温度以盐水的冰点为0 (华氏度) , 摄氏温度则以水的冰点为0 (摄氏度) , 这两种温度的零值都是设置的, 不代表“没有”或“不存在”.

表 2.1: 计量尺度

级别	类型	是否有序	有无绝对零点	适用的数学运算
低级	定类尺度	无序	无	=, ≠
↓	定序尺度	有序	无	=, ≠, >, <
	定距尺度	有序	无	=, ≠, >, <, +, −
高级	定比尺度	有序	有	=, ≠, >, <, +, −, ×, ÷

此外, 由定序变量通过等级变换而得到的变量也属于定距变量. 例如, 在3级量表的满意度调查中, 只有1、2、3等三个数值, 并没有零值, 因而没有绝对零点. 定距变量的取值表现为数值, 具有=或≠, >或<, +或–的数学特性, 注意, 定距变量不具有乘除的运算特性(固定了零值的乘除法才有意义). 例如, 我们可以说今天的温度(摄氏度或华氏度) 比昨天高2度, 但不能说今天的温度增长了10%.

(4) 定比尺度(ratio scale)

定比尺度又称为比率尺度, 是对事物的准确测度. 相比于定距尺度, 定比尺度具有绝对零点, 即零值表示“没有”或“不存在”, 如: 收入、产量、人数等. 值得一提的是, Kelvin温度是定比尺度的产物, 这是因为Kelvin零度为理想气体压强为零、气体动能为零时的温度(对应于-273.15摄氏度) , 从而零值代表没有动能, 为绝对零度. 由定比尺度所刻画的变量即为定比变量, 其取值也表现为数值, 具有=或≠, >或<, +或–, ×或÷的数学特性. 例如, 可以说今年的产量比去年多200千克, 也可以说今年的产量增长了10%.

上述四个计量尺度具有等级顺序, 从低级到高级依次为: 定类尺度、定序尺度、定距尺度、定比尺度. 表2.1总结了它们的特点.

由定性尺度所刻画的变量为定性变量, 包含定类变量和定序变量; 由定量尺度所刻画的变量为定量变量, 包含定距变量和定比变量. 四种变量也具有等级顺序, 从低级到高级依次为: 定类变量、定序变量、定距变量、定比变量. 一般, 高级变量可转换为低级变量, 例如, 可通过离散化(分箱) 将定量变量转换为定序变量. 因此, 适用于低级变量的统计学方法也可以适用于高级变量(先将高级变量转换为低级变量, 然后应用该方法). 需要注意, 将高级变量转换为低级变量必然会损失信息. 另一方面, 由于信息不足, 难以将低级变量转换为高级变量, 除非加入一些假定(例如, 等级变换中的假定). 通常情况下可不区分定距变量和定比变量, 但应注意一些特殊场合(例如基于乘除法衍生特征), 日常生活中也应避免一些不恰当的表述.

理解数据的关键一步是介绍每一个变量的信息, 包括: 变量名称、变量含

表 2.2: golf数据集

序号	Temperature	Humidity	Windy	Outlook	Play
1	85	85	False	Sunny	No
2	80	90	True	Sunny	No
3	83	86	False	Overcast	Yes
4	70	96	False	Rainy	Yes
5	68	80	False	Rainy	Yes
6	65	70	True	Rainy	No
7	64	65	True	Overcast	Yes
8	72	95	False	Sunny	No
9	69	70	False	Sunny	Yes
10	75	80	False	Rainy	Yes
11	75	70	True	Sunny	Yes
12	72	90	True	Overcast	Yes
13	81	75	False	Overcast	Yes
14	71	91	True	Rainy	No

义、变量类型、取值范围与取值含义(尤其是定性变量的取值) 等. 表**??**是一个例子. 下面, 给出另一个例子.

例2.1. 采用表2.2中的golf数据集, 描述变量信息, 展示在表2.3中.

2.3 数据质量

数据的质量直接决定了数据分析的结果, 也是数据预处理阶段的重要关注点. 我们都希望获得高质量的数据, 但现实中的数据往往存在各种质量问题. 研究显示, 不良的数据使企业额外花费15%-25%的成本[1], 企业管理者认为不良数据造成的损失约为平均每年1500万美元[2].

衡量数据质量的指标有很多. Wang & Strong (1996) 通过调查, 归纳出四类对数据用户较为重要的质量指标, 包括:

(1) 内在数据质量(intrinsic data quality). 这些指标用于衡量数据本身的

[1]中国信息通信研究院云计算与大数据研究所, CCSA TC601 大数据技术标准推进委员会(2019) 数据资产管理实践白皮书(4.0版).

[2]Gartner (2018) How to Create a Business Case for Data Quality Improvement, https://www.gartner.com/smarterwithgartner/how-to-create-a-business-case-for-data-quality-improvement.

表 2.3: golf数据集中的变量信息

序号	变量名称	变量含义	变量类型	取值(范围) 及其含义
1	Temperature	当天气温	定量	64-85 (华氏度)
2	Humidity	当天湿度	定量	65-96 (%)
3	Windy	是否有风	定类	False:无风; True:有风
4	Outlook	天气情况	定类	Overcast:多云；Rainy:有雨; Sunny:晴朗
5	Play	打高尔夫情况	定类	No:不打高尔夫; Yes:打高尔夫

质量, 包括: 准确性(accuracy)、客观性(objectivity)、可信性(believability)、声誉(reputation) 等.

(2) 外联数据质量(contextual data quality). 这些指标用于衡量数据适用于分析任务的情况, 包括: 完整性(completeness)、及时性(timeliness)、关联性(relevancy)、增值(value-added) 和适量(appropriate amount of data).

(3) 表示数据质量(representational data quality). 这些指标主要衡量数据系统的质量, 包含可解释性(interpretability) 和易理解性(ease of understanding) 等数据含义方面的指标, 以及一致性(representational consistency) 和简洁性(concise representation) 等数据格式方面的指标.

(4) 可达数据质量(accessibility data quality). 与表示数据质量指标类似, 可达数据质量指标也主要衡量数据系统的质量, 侧重于评估数据是否可获得且不受损害, 主要包含数据的可达性(accessibility) 和安全性(access security).

总体而言, 准确性、完整性、一致性和及时性等是较为典型的数据质量指标(Ballou & Tayi,1999). 我们认为, 在数据分析实践中, 尤其需要关注的数据质量指标有:

(1) 代表性: 样本是否能代表总体.

(2) 准确性: 选取的变量能否准确反映问题, 变量取值的合理性等.

(3) 完整性: 选取的变量是否全面地反映问题, 数据缺失的情况和缺失的原因等.

(4) 有效性: 所拥有的数据是否适用研究目标的时间、空间和人群等.

(5) 充分性: 对于研究目标和分析方法而言, 样本观测和变量是否充足. 例如, 普通最小二乘回归中, 要求样本量大于变量维数, 若不满足, 则表明样本观测可能不够充分.

某些质量问题可以在数据预处理阶段, 通过探索性数据分析方法予以识别, 或按照一定的规则通过计算机程序排查. 例如, 若发现数据存在不一致(inconsistency)问题, 即两个样本观测的自变量取值完全相同但因变量取值不同, 则表明可能存在数据不准确或变量不充足(现有的自变量未能体现样本观测的差异性) 的问题. 但某些质量问题可能直到建模完成以后可能才

会发现. 例如, 当发现模型的效果特别好时, 应该反思建模中是否使用了“过于充足”的信息, 比如, 实施预测时无法获得的一些自变量信息(包括因变量的衍生信息等).

注记2.1. 建议按照以下流程理解数据:

(1) 回顾数据的收集思路与过程;

(2) 梳理变量间的层次结构和关联;

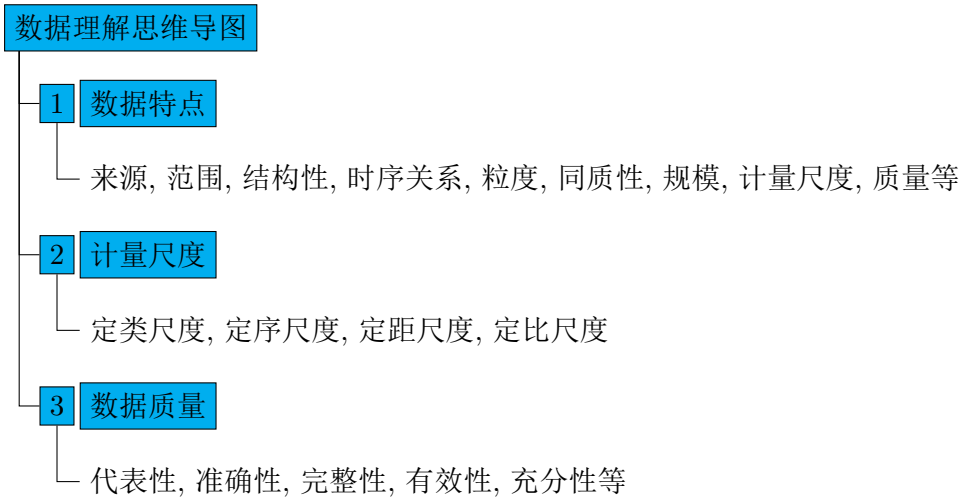
(3) 介绍每一个变量的信息, 包括: 变量名称、变量含义、变量类型、取值范围与取值含义(尤其是定性变量的取值) 等;

(4) 采用探索性分析等方法全面地了解数据质量;

(5) 分析数据的其他特点.

本章小结

思维导图



延伸阅读

2015年, 美国科学院院士、美国艺术与科学院院士、加州大学伯克利分校统计系郁彬教授应邀发表了《Data Wisdom for Data Science》[3], 论述了

[3]英文发表于《Big Data, New Data Management Technologies and Data Science》, http://www.odbms.org/2015/04/data-wisdom-for-data-science/: 《中国计算机学会通讯》2016年第1期刊登了中译版《数据科学中的数据智慧》

大数据时代下统计数据分析工作者的使命, 以及如何正确应用统计方法解决实际问题, 从十个方面讨论如何形成和培养"数据智慧". 文中, 郁彬教授强调对业务理解和对数据理解的重要性, 并指出良好的人际交流技巧有助于更好地理解业务和数据. 感兴趣的同学可以搜索阅读.

习题

1. 请描述数据的不一致问题. 如何诊断数据的不一致问题?
2. 针对天猫用户重复购买预测案例, 按照2.3节中的流程, 分析数据特点.

提示: 该案例来源于阿里云天池的学习赛, 具体介绍可参看https://tianchi.aliyun.com/competition/en

References

[1] 吴翌琳, 房祥忠(2016) 大数据探索性分析. 北京:中国人民大学出版社.

[2] Ballou D.P., Tayi G.K. (1999) Enhancing data quality in data warehouse environments. Communications of the ACM, 42(1):73-78.

[3] Wang R.Y., Strong, D.M. (1996) Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems 12(4): 5 – 34.