数据科学基础

Foundations of Data Science

1.2 数据类型

陈振宇
南京大学智能软件工程实验室
www.iselab.cn

数据类型

数据类型通常可以分为四类：

- 定类数据（Nominal Data）
- 定序数据（Ordinal Data）
- 定距数据（Interval Data）
- 定比数据（Ratio Data）

定类数据

- 通常只用来代表不同的分类。数据相应的数据没有数量的含义, 只用来识别种类。
- 它是没有顺序大小之分的比较低级的一类数据。
- 比如我们使用1代表男性、0代表女性, 但不代表男性比女性更好。
- 我们比较不同程序语言,分别将C、C++、Java定义为1, 2, 3; 也可定义为0, 1, 2。定类数据之间的数学关系就是等于(=)或者不等于(≠)。我们可以说Java≠C, 但不能说Java>C。

定序数据

- 定序数据是量化尺度的最基本形式, 通常采用数字表示顺序。
- 定序数据的每个分类不但有差别, 而且有等级之分。例如, 更大、更快、更高、更强等等。
- 定序数据之间的数学关系除了=和≠, 还有>, <等。
- 比如产品等级分为优等品, 合格品和不合格品, 分别记为2, 1, 0; 也可以将优等品, 合格品和不合格品分别记为1, 2, 3。

定距数据

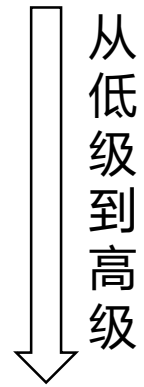
- 定距变量,描述事物类别或次序之间的间距。
- 定距变量不仅能将事物区分为不同类型并进行排序, 而且可以准确地指出类别之间的差距是多少。
- 定距变量的数据是一种真正数量化的数值, 即可以对这些数据进行加、减、乘、除等运算。
- 在定距变量中, 0是强行规定的, 它不代表完全没有的意思。

定比数据

- 定比变量是在定距变量的基础上, 扩展可作为比率的基数而成。
- 定比变量一般需要统一的单位, 如米、厘米、公斤、秒等等。
- 我们前面讲的身高、体重等都是一种定比变量, 它们对应于长度和重量度量。
- 和定距变量的一个根本区别是定比变量的零点代表了完全没有的含义。比如0米代表没有长度, 0公斤代表没有重量。此时, 我们可以说体重为80公斤的人是40公斤的人的两倍重。

小结

数据类型	基本特征	关系和运算	举例
定类数据	无次序分类	=, ≠	性别，政党等
定序数据	有次序分类	=, ≠, >, <	Bug严重级别，年级等
定距数据	有距离度量没有绝对零点	+, -, ×, ÷（数值除法）	温度，成绩等
定比数据	具有绝对零点	可以是比例除法运算	长度、重量、年龄等



这四种类型是从低到高的递进关系，高级的类型可以用低级类型的数据分析方法来分析，而反过来却不行。

课程讨论：数据类型

- 结构化数据

  例如：关系型数据库的数据

- 半结构化数据

  例如：XML，JSON

- 非结构化数据

  例如：长文本，图片，视频，音频等

更多详情，请咨询课程助教。