第二章认识数据

Tuesday, September 28, 2021 6:43 PM

1. 数据对象与属性类型

- 文本、图像
- 网络
- 时序数据
  - 视频数据：图像序列
  - 时间数据：时间序列
  - 顺序数据：事务序列
  - DNA序列数据

a. 数据对象

- 数据集由数据对象组成
- 一个数据对象代表一个实体
- 数据对象用属性来描述
- 数据对象可以是一个记录、数据点、实体、样本、实例或者对象等。（如果数据对象存放在数据库中，它们是数据元组，即数据库中行对应数据对象，列对应于属性）

b. 属性类型

前言

- 一个属性是一个域，表示一个数据对象的一个特征。
- “属性”、“维度”、“特征”和“变量”这些词在语义上是可交换的。“维度”通常被用在数据仓库中，机器学习中倾向于使用“特征”；统计学倾向使用“变量”，数据挖掘和数据库经常使用“属性”。属性描述一个顾客对象，如：顾客ID，姓名，地址。
- 对给定属性的可观察值被称为观察。刻画一个给定对象的属性集合被称为属性向量（或特征向量）。

数据属性

- 标称属性（离散型）
  - 标称属性的值是事物的标号或者名称。
  - 每一个值表示类别、编码或者状态。
  - 值没有次序信息。
  - 在计算机领域，也可以称为枚举型。
- 二元属性（标称属性的特例）（离散型）
  - 只有两个类别或状态
  - 布尔属性
  - 对称的二元属性
  - 非对称的二元属性
- 序数属性（离散型）
  - 可能的值之间具有有意义的序或秩评定
  - 但是相继值之间的差是未知的

- 序数属性可以用来记录不能客观度量的主观质量评估
- 数值属性（连续型）
  - 定量的（整数或实数）
    - 区间标度属性
      - 用相等的单位尺度度量
      - 值有序，可以为正、负、0，允许定量评估值之间的误差
      - 没有真正的零值（如摄氏温度）
    - 比率标度属性
      - 具有固定的零点（如开式温度、字数、货币数量、工作年限）
      - 有序
      - 可以说一个值是另一个值的倍数

标称、二元和序数属性总结：

- 值之间具有有意义的序或秩评定
- 但是相继值之间的差是未知的（如small、medium、large）
- 序数属性可以用来记录不能客观度量的主观质量评估（如满意度等）

c. 对比：连续属性VS离散属性

- 离散属性：
  - 具有有限或者无限可数个值
  - 可以用或不用整数表示
- 连续属性
  - 连续值是实数
  - 实数值用有限位数字表示
  - 连续属性一般用浮点变量表示
  - 如果值不是离散的，则是连续的

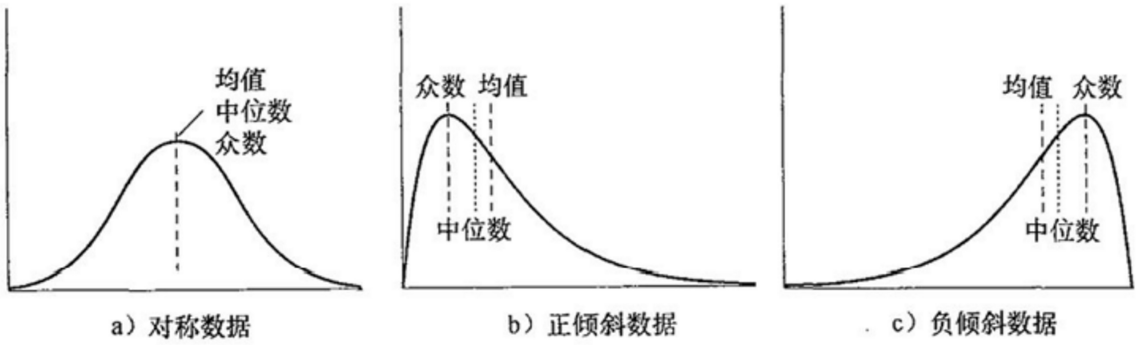
2. 数据的基本统计描述

a. 中心趋势度量

- 度量数据分布的中部或中心位置
- 均值、中位数、中列数、众数
  - 均值：加权均值、截尾均值（去掉高低极端值后的均值）
  - 中位数：数据量很大时可以用插值计算近似值
    - L1是中值区间的最低值，
    - N是数据值的个数，
    - (Σ $freq$) l是所有低于中值区间的所有区间的频率之和。
    - $freqmedian$是中值区间的频率（应该是频数，PPT可能写错了，上一句也是如此），
    - width是中值区间的宽度（极差）

$$median = L_1 + \left(\frac{\frac{N}{2} - (\sum freq)_l}{freq_{median}}\right) width$$

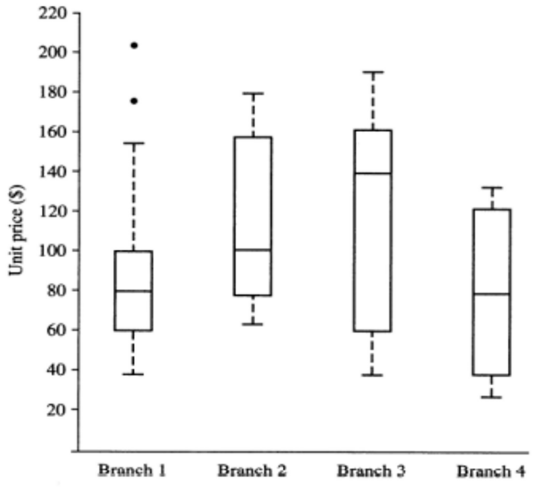
- 中列数
  - 数据集中最大值和最小值的平均值。
  - 可以用来评估数值型数据的中心性趋势
- 众数

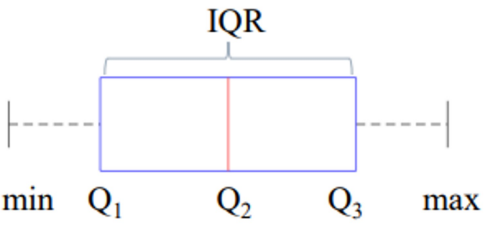


a）对称数据　b）正倾斜数据　c）负倾斜数据

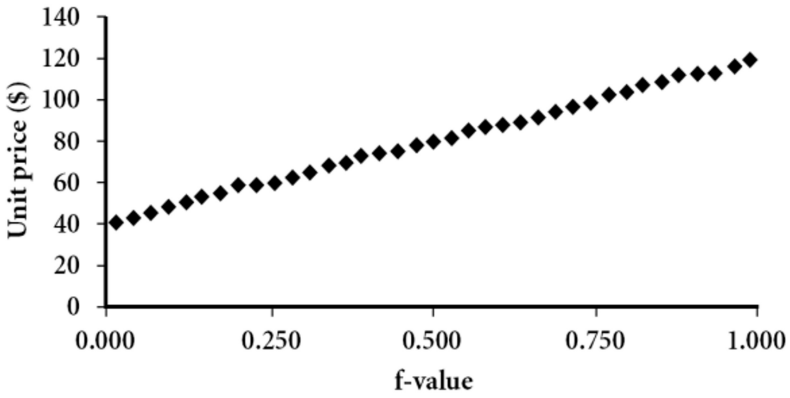
b. 度量数据分布

- 数据如何分散
- 1、极差、分位数、四分位数、四分位数极差
- 2、五数概括、盒图和离群点
- 3、方差和标准差
  - 分位数：每隔一定间隔上的点，把数据划分成基本上大小相等的连贯集合（计算方式看PPT42页）
    - 四分位数（Q1是25% percentile，Q3是75% percentile）
    - 四分位数极差：IQR=Q3-Q1
  - 五数概括：min、median、Q3、max
    - 鉴别可疑离群点的一个规则 是：挑选落在Q3以上或者Q1以下至少1.5*IQR以上的数据值
  - 盒图（min,Q1 ,Q2 ,Q3 ,max,须,离群点）
    - 盒的端点一般在四分位数上
    - 箱子的长度是四分位数极差IQR
    - 中位数是箱子中间的线
    - 箱子外面的两根须是观察的最小值和最大值
      - 规则：最大值和最小值不到1.5*IQR时扩展到他们。否则的话，须的末端是1.5*IQR处





- 分位数图：观察单变量数据分布的简单有效方法
  - 显示给定属性的所有数据
  - 汇出分位数信息
  - 对数据进行升序排列，每个观测值xi 与一个百分数fi配对 fi=(i-0.5)/N



- 分位数——分位数图（没理解）

3. 数据可视化

- 数据的基本统计描述的图形显示

  盒图、散点图等
- 层次可视化技术

  把所有维划分成子集
- 可视化复杂对象和关系

  标签云
- 可视化软件

  CiteSpace、Gephi
- t-SNE

- 深度模型画图工具

4. 度量数据的相似性和相异性

- 相似性和相异性都称邻近性
  - 相似性
    - 描述两个对象的相似程度
    - 通常在[0, 1]内
    - 值越高，越相似
  - 相异性
    - 差异程度
    - 通常是[0, 1]范围内
- 4.1 数据矩阵与相异性矩阵
  - 数据矩阵：行是一个对象，列是属性
  - 相异性矩阵：

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

- 4.2 标称属性的近邻性度量
  - 方法一：简单的匹配
    - 距离：d(i, j) = $\frac{p-m}{p}$
    - 相似度：sim(i, j) = 1 - d(i, j) = $\frac{m}{p}$
      - m: 匹配的数目（即i和j取值相同状态的属性数）
      - p: 刻画对象的属性总数
  - 方法二：one-hot编码
    - 距离都相同，一般均为√2
- 4.3 二元属性的邻近性度量
  - 对称二元属性：

    d(i, j) = $\frac{r+s}{q+r+s+t}$
  - 非对称二元属性：
    - 非对称的二元相异性：d(i, j) = $\frac{r+s}{q+r+s}$
    - 非对称的二元相似性：sim(i, j) = $\frac{q}{q+r+s}$
      - q是在对象i和j都取1的属性数
      - r是在对象i中取1，在j中取0的属性数
      - s是在对象i中取0，在j中取1的属性数
      - t是对象i和j都取0的属性数

表 2.3　二元属性的列联表

		对象 j		
		1	0	sum
对象 i	1	q	r	$q+r$
	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

- 4.4 序数属性的距离
  - 序数属性的值之间具有有意义的序或排位，相继值之间的量值未知。
  - 数值属性的值域可以映射到具有 M_f 个状态的序数属性 f.
    - 第i个对象的f 值为 x_{if} ，属性 f 有 M_f 个有序状态，用对应的排位r_{if} $\in\{1,\ldots,M_f\}$取代x_{if}
    - 将每个属性的值域映射到 [0, 1]上，以便每个属性都有相同的权重。$z_{if} = \frac{r_{if}-1}{M_f-1}$
    - 采用任意一种数值属性的距离度量计算。

point	attribute 1
x1	Excellent
x2	fair
x3	good
x4	Excellent

r_{if}

point	attribute 1	
x1	Excellent	3
x2	fair	1
x3	good	2
x4	Excellent	3

z_{if}

point	attribute 1	
x1	Excellent	1
x2	fair	0
x3	good	0.5
x4	Excellent	1

欧式距离

L2	x1	x2	x3	x4
x1	0			
x2	1	0		
x3	0.5	0.5	0	
x4	0	1	0.5	0

- 4.5 数值属性的距离
  - 欧几里得距离（Euclidean distance）
    - L_2范数
  - 曼哈顿距离（Manhattan distance）
    - L_1范数
  - 闵可夫斯基距离（Minkowski distance）
    - L_h范数或者L_p范数（L_0范数是非零元素的个数）

$$d(i,j)=\sqrt[h]{|x_{i1}-x_{j1}|^h+|x_{i2}-x_{j2}|^h+\cdots+|x_{ip}-x_{jp}|^h}$$

- 4.6 混合类型属性
  - 可以将所有属性类型一起处理

$$d(i,j)=\frac{\sum_{f=1}^{p}\delta_{ij}^{(f)}d_{ij}^{(f)}}{\sum_{f=1}^{p}\delta_{ij}^{(f)}}$$

- f 是标称或者二元的:
  $d_{ij}^{(f)}=0$ if $x_{if}=x_{jf}$, or $d_{ij}^{(f)}=1$ otherwise
- f 是数值的 $d_{ij}^{(f)}=\frac{|x_{if}-x_{jf}|}{max_h x_{hf}-min_h x_{hf}}$
- f 是序数的

$$z_{if}=\frac{r_{if}-1}{M_f-1}$$

- $\delta_{ij}^{(f)}=0$ if x_{if} 或者 x_{jf} 缺失，或者 $x_{if}=x_{jf}=0$；$\delta_{ij}^{(f)}=1$ otherwise

○ 4.7 余弦相似度

$$sim(v_1,v_2)=\frac{v_1\cdot v_2}{|v_1||v_2|}$$

▪ 向量v中第j个数值就是相应文档中第j个项的量度

v_1 = { 1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0 }
v_2 = { 0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,0,0,0,0,0,0,0 }
v_3 = { 1,1,1,0,1,0,1,1,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0 }

○ 4.8 度量学习

5. 小结