大数据基础与应用
BIG DATA FOUNDATION AND APPLICATION

数据分析算法

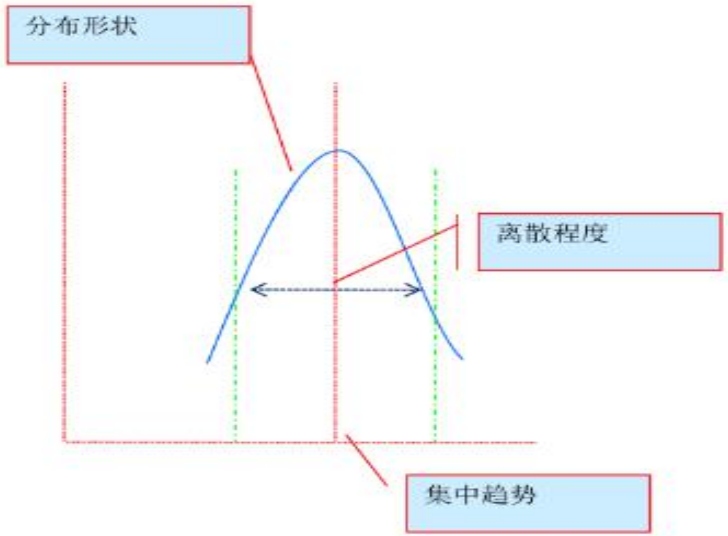
北京理工大学计算机学院　孙新

2019年1月

目　录

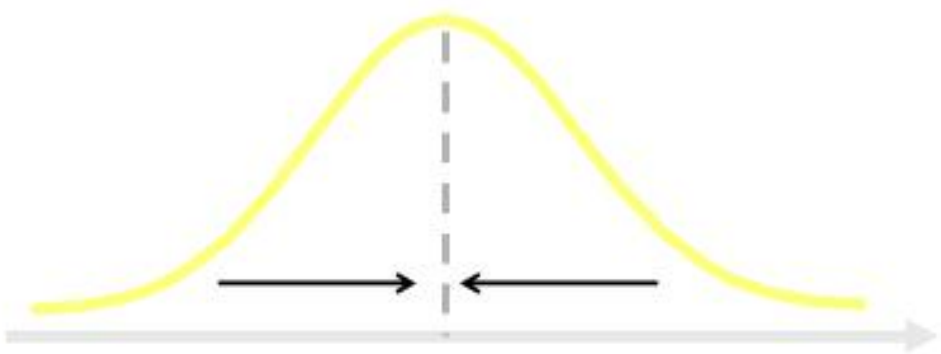
数据描述行分析

- 在大数据分析中，获取到数据后，第一时间往往是需要从宏观角度来观察数据，也就是分析数据的特征。
- 这些能够概括数据位置特性、分散性、关联性等数字特征，以及能够反映数据整体分布特征的分析方法，称为数据描述性分析
- 数据分布的特征有三种：



- 2.1数据的中心趋势度量
  - 均值（mean）、加权算数均值、中位数（median）、截断均值、众数（mode）、中列数（midrange）
- 2.2数据的离散程度度量（离中趋势）
  - 四分位数（quartiles）、四分位极差（InterQuartiles Range，IQR）、方差（variance）和标准差等
- 2.3数据的分布度量
  - 偏态(skewness)及其测度
  - 峰态(kurtosis)及其测度
- 2.4图形化分析方法

- 中心趋势度量(也叫做: 集中趋势测度)
  - 一组数据向其中心值靠拢的倾向和程度
  - 中心趋势度量就是寻找数据水平的代表值或中心值
  - 对于不同类型的数据, 我们可以采用不同的集中趋势测度值
    - (1) 均值
    - (2) 加权算数均值
    - (3) 中位数
    - (4) 截断均值
    - (5) 众数
    - (6) 中列数



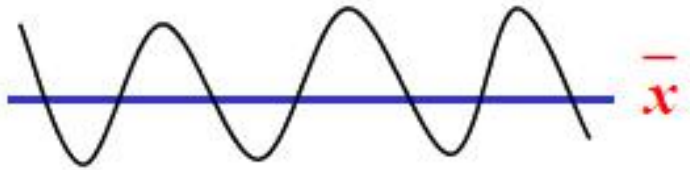
(1)均值（也叫算术均值）：假设一组数据共有n个一维数据，分别是$x_1, x_2,\ldots,x_n$，则均值可以表示为

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

在表示数据的位置特征方面，均值是最常用的指标之一

均值是度量数据集中心的最常用、最有效的数值度量方法

- 一组数据的均衡点所在
- 易受极端值的影响



- 均值可以用来反映数据的平均水平，如下表中给出的两个名人的微博数据中

两个比较受欢迎的微博名人在 2018 年 3～5 月间的一部分微博数据

微博名人 W			微博名人 Z		
发布时间	点赞数量	转发数量	发布时间	点赞数量	转发数量
18. 3. 18	3056	187	18. 5. 8	3398	1175
18. 4. 29	1169	511	18. 3. 6	4849	253
18. 3. 3	2743	177	18. 3. 18	4246	211
18. 4. 29	1616	215	18. 4. 28	4342	113
18. 2. 22	2391	92	18. 3. 14	3464	206
18. 3. 19	930	119	18. 5. 2	1819	1067
18. 4. 8	968	331	18. 5. 1	2300	1056
18. 5. 2	1011	51	18. 4. 8	2955	120
18. 5. 10	1386	36	18. 4. 17	3023	104
18. 4. 18	936	38	18. 3. 14	2560	229

均值1620.6　　均值3295.6

(2) 加权算数均值：又称加权算术平均

集合中每个值与一个权值相关联，权值反映对应值的显著性、重要性或出现频率。

在这种情况下，使用加权算数均值（weighted arithmetic mean）

W表示权值

$$\overline{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \ldots\ldots + w_N x_N}{w_1 + w_2 + \ldots\ldots + w_N}$$

加权平均数例题分析

$$\bar{x} = \frac{\sum_{i=1}^{k} W_i X_i}{n} = \frac{22200}{120} = 185$$

某电脑公司销售量数据分组表

按销售量分组	组中值(x_i)	频数(w_i)	$x_i w_i$
140~150	145	4	580
150~160	155	9	1395
160~170	165	16	2640
170~180	175	27	4725
180~190	185	20	3700
190~200	195	17	3315
200~210	205	10	2050
210~220	215	8	1720
220~230	225	4	900
230~240	235	5	1175
合计	—	120	22200

(3) 中位数：

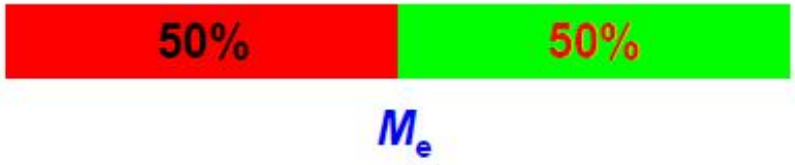
一组数据按从小到大(或从大到小)的顺序依次排列,处在中间位置的一个数(或最中间两个数据的平均数)

中位数M可以表示为

$$M=\begin{cases} x_{\frac{n+1}{2}}, & n\text{为奇数} \\ \frac{1}{2}\left(x_{\frac{n}{2}}+x_{\frac{n}{2}+1}\right), & n\text{为偶数} \end{cases}$$

备注:奇数个数的话取中间的数字，偶数个数的话取中间两个数的平均数

- (3) 中位数(median)--M_e
  - 排序后，处于中间位置上的值



  - 不受极端值的影响
  - 主要用于顺序数据，也可用数值型数据，但不能用于分类数据
- 相比均值，中位数有着更好的抗干扰性

99个年收入10万的人群=>均值(平均年收入)=10万
1个年收入1000万的人

均值（平均年收入）=19.9万

例题：顺序数据的中位数

甲城市家庭对住房状况评价的频数分布		
回答类别	甲城市	
	户数（户）	累计频数
非常不满意	24	24
不满意	108	132
一般	93	225
满意	45	270
非常满意	30	300
合计	300	—

解： 中位数的位置为

(300+1)/2＝150.5

从累计频数看，中位数在“一般”这一组别中

中位数为

M_e=一般

例题：数值型数据的中位数

9个家庭的人均月收入数据

原始数据:	1500	750	780	1080	850	960	2000	1250	1630
排　序:	750	780	850	960	**1080**	1250	1500	1630	2000
位　置:	1	2	3	4	5	6	7	8	9

$$\text{位置} = \frac{n+1}{2} = \frac{9+1}{2} = 5$$

中位数 = 1080

例题：数值型数据的中位数

10个家庭的人均月收入数据

排序:	660	750	780	850	960	1080	1250	1500	1630	2000
位置:	1	2	3	4	5	6	7	8	9	10

$$位置 = \frac{n+1}{2} = \frac{10+1}{2} = 5.5$$

$$中位数 = \frac{960+1080}{2} = 1020$$