日期: /

数据

一. 数据.

1. 数据集: 一组含有特征的数据对象的集合.
2. 样本: 由一组特征所描述的一个对象.
3. 特征: 描述样本或对象在某方面的表现或性质的事项 (对模型有用的部)

特征 (Features)

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

样本 (Samples)

例: 个人身高, 北京某时刻气温

属性同对象而异

或随时间变化

4. 数据的数学表示

$D=\{x_1, x_2, \cdots, x_m\}$ 表示 m 个样本的数据集

每个样本由 d 个属性描述; 样本 $x_i=(x_{i1}, x_{i2}, \cdots x_{in})$ 为 d 维样本空间 $\mathcal{X}$ 中的一个向量, $x_i \in \mathcal{X}$

x_{ij} 为 x_i 在第 j 个属性上的取值; (x_i, y_i) 表示第 i 个样例, y_i 是样本 x_i 的标记

日期: /

如下表，m=17，x_1 = (青绿；蜷缩；浊响；清晰；凹陷；硬滑)

x_{21} = “乌黑” 是 x_2 在第1个属性上的取值

描述西瓜的6个属性

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	否

标记（label）

二、特征类型

1. 特征取值

① 定义：给特征指派一个特征的数字或符号

② 相同的特征可以有不同取值
不同的特征可以有相同取值

- **样本-->姓名：张三、年龄：18、性别：男**
- **特征：**{姓名、年龄、性别}，**取值：**{张三、18、男}

2. 特征类型：定性，定量

① 定性属性：不具有数的大部分性质

② 定量属性：用数表示，具有数的大部分性质
可以是连续或离散

日期: /

属性类型		描述	例子	操作
分类的（定性的）	标称	标称属性的值只是不同的名字，即标称值只提供足够的信息以区分对象（=，≠）	邮政编码、雇员 ID 号、眼球颜色、性别	众数、熵、列联相关、χ^2检验
	序数	序数属性的值提供足够的信息确定对象的序（<，>）	矿石硬度{好，较好，最好}、成绩、街道号码	中值、百分位、秩相关、游程检验、符号检验
数值的（定量的）	区间	对于区间属性，值之间的差是有意义的，即存在测量单位（+，−）	日历日期、摄氏或华氏温度	均值、标准差、皮尔逊相关、t和F检验
	比率	对于比率变量，差和比率都是有意义的（*，/）	绝对温度、货币量、计数、年龄、质量、长度、电流	几何平均、调和平均、百分比变化

③

- **标称类型** Nominal
  - **只能区分样本之间的不同**，例如：学号、籍贯、邮政编码
- **序数类型** Ordinal
  - **能够对样本之间的顺序进行区分**，例如：排名、年级、衣服的号码{S, M, L, XL, XXL}
- **区间类型** Interval
  - **能够对样本在坐标系上的相对距离进行度量**，例如：日历上的日期、摄氏或华氏温度等
- **比率类型** Ratio
  - **能够对样本在坐标系上的绝对位置进行标定**，例如：开尔文温度、 长度、时间、质量、货币单位等

④ 本质区别：对应的操作不同

{ 相异性：=，≠　　加法：+，−
　 序：<，>　　乘法：*，/

标称：相异性（=，≠）

序数：相异性 序（=，≠，<，>）

日期: /

值之间差有意义.

区间: 相异性, 序, 加法 (=, ≠, <, >, +, -)

比率: 相异性, 序, 加法, 乘法

值之间差与比率均有意义.

- 摄氏温度?
  - 冰水混合物的温度定为0摄氏度，沸水的温度定为100摄氏度
- 华氏温度?
  - 氯化铵和水的混合物的冰点温度（即氨水结冰的温度）为温度计的零度，人体温度为温度计的100度。(华氏温度将水的冰点定为32度，沸点定为212度)
- 开尔文温度（绝对温度）？
  - 以绝对零度作为计算起点的温度，即将水三相点的温度准确定义为273.15K。
  - 20K是10K的2倍

摄氏温度或华氏温度的的零度是硬性规定的，其比率是无物理意义的

⑤ 区间类型和比率类型的特点.

- 区间类型（interval-scaled）属性特点：
  - 例：温度属性，一般表示：10℃~15℃。
  - 1. 用相等的单位尺度度量，区间属性的值有序，可以为正、0、负。（值的秩评定）
  - 2. 允许比较与定量评估值之间的差。
  - 3. 区间标度属性是数值的，中心趋势度量中位数和众数，还可以计算均值。
- 比率类型（ratio-scaled）属性特点：
  - 1. 具有固有零点的数值属性。（也就是该种属性中会有固有的为 0 的值）
  - 2. 一个值是另一个的倍数（或比率）。
  - 3. 值是有序的。（可以计算差、均值、中位数、众数）
  - 例：度量重量、高度、速度和货币量（例如 100 元是 1 元的 100 倍）的属性。

17

⑥ 先标变换:

日期: /

属性类型		变　换	注　释
分类属性（定性属性）	标称	任何一对一变换，例如值的一个排列	邮政如果所有雇员的ID号都重新赋值，不会导致任何不同
	序数	值的保序变换，即 *new_value= f (old_value)* 其中*f*是单调函数	包括好，较好，最好的属性可以完全等价地用值{1, 2, 3}或{0.5, 1, 10}表示
数值属性（定量属性）	区间	*new_value=a*old_value+b* 其中*a*、*b*是常数	华氏和摄氏温度标度零度的位置和1度的大小（单位）不同
	比率	*new_value=a*old_value*	长度可以用米或英尺度量

3. 特征类型：离散与连续.（值的个数）.

① 离散 Discrete Feature.
- 具有有限或无限种可能值.（邮政编码…）
- 常常为整数变量.
- 可以为分类的，也可以是数值的.

② 连续 Continuous Feature.
- 取实数值的特征.（温度，高度…）
- 有限精度测量与表示.
- 浮点变量表示.

③ 二元特征 Binary Feature.
- 仅有2个值的特征.（0或1）（性别，对错）
- 是离散的特殊情形

④ 非对称特征. Asymmetric Binary Feature.

日期: /

{出现非0值才重要的特征，状态值为0不重要
非对称二元特征：只有非0值重要的二元特征
[体检：阳性(1) 阴性(0)

三、样本的相似性和相异性

1. 相似度和相异度基本概念

① 相似度：similarity

{两个样本相似特征的数值化度量
2个样本越相似，相似性越高
非负，取值通常在[0,1]

② 相异度：dissimilarity

{两个样本相异特征的数值化度量
2个样本越相似，相异性越低
非负，取值通常在[0,1]或[0,∞)

通常将距离作为相异度的同义词

2. 简单属性的相似度和相异度

属性类型	相异度	相似度
标称的	$d=\begin{cases}0 & \text{如果 } x=y\\1 & \text{如果 } x\neq y\end{cases}$	$s=\begin{cases}1 & \text{如果 } x=y\\0 & \text{如果 } x\neq y\end{cases}$
序数的	$d=\frac{\lvert x-y\rvert}{(n-1)}$ 值映射到整数0到$n-1$，其中n是值的个数	$s=1-d$
区间或比率的	$d=\lvert x-y\rvert$	$s=-d$，$s=\frac{1}{1+d}$，$s=e^{-d}$， $s=1-\frac{d-\min\_d}{\max\_d-\min\_d}$

日期: /

3. 距离计算

① 欧几里德距离

$$d(x,y)=\sqrt{\sum_{k=1}^{n}(x_k-y_k)^2}$$

n为对象数据维度(属性的个数), x_k和y_k分别是数据对象x和y的第k个属性.

② 闵可夫斯基距离

欧式距离是闵可夫斯基的特例

$$d(x,y)=\left(\sum_{k=1}^{n}|x_k-y_k|^r\right)^{\frac{1}{r}}$$

- $r=1$, 是哈达距离, L1范数 L1-norm
  $$d(x,y)=\sum_{k=1}^{n}|x_k-y_k|$$
- $r=2$, 欧式距离, L2范数, L2-norm
  $$d(x,y)=\sqrt{\sum_{k=1}^{n}|x_k-y_k|^2}$$
- $r=\infty$, 上确界距离, ∞范数 ∞-norm
  $$d(x,y)=\|x-y\|_\infty=\lim_{r\to\infty}\left(\sum_{k=1}^{n}|x_k-y_k|^r\right)^{\frac{1}{r}}$$

4. 距离特性:

① 非负性: 对于任意p和q, 存在$d(p,q)\geq 0$; 当且仅当$p=q$时$d(p,q)=0$

日期: /

② 对称性：对于任意 p 和 q，$d(p,q)=d(q,p)$

③ 三角不等式：对于任意 p、q 和 r，$d(p,r)\le d(p,q)+d(q,r)$

满足以上3条称为一种度量(metric)

5. 相似性

① 若 $s(x,y)$ 为数据点 x 和 y 之间的相似度

{ 仅当 $x=y$ 时，$s(x,y)=1$ $(0\le s\le 1)$ [非负]
对于所有 x 和 y，$s(x,y)=s(y,x)$ [对称性]

② 二元数据的相似度度量

对2个二元向量 x 和 y 上的 n，值取0或1

f_{ij} 表示 $x=i$，$y=j$ 时的属性数目 $(i,j=0,1)$

ⅰ, SMC：简单匹配系数

$$SMC=\frac{\text{值匹配的属性个数}}{\text{属性个数}}$$

$$=\frac{f_{11}+f_{00}}{f_{01}+f_{00}+f_{10}+f_{11}}$$

{ 通常 $SMC\in[0,1]$，0表示不相似，1表示相似
适用：对称二值离散型属性
0和1同等重要(例如性别)

日期: /

② Jaccard 系数.

$$J = \frac{\text{匹配个数}}{\text{不匹配中不涉及的属性个数}}$$

- 如果每个非对称的二元属性对应于商店的一种商品，则1表示该商品被购买，0表示该商品未被购买。由于未被顾客购买的商品数远大于被购买的商品数，因此，简单匹配系数(SMC)会判定所有的事务都是类似的。Jaccard值越大说明相似度越高。
- Jaccard系数来处理仅包含非对称的二值离散型属性。不对称的二值离散型属性是指属性取值为1或者0不是同等重要，例如：是否是癌症的结果，因此通常用1来表示阳性结果，而用0来表示阴性结果。

32

$x = (1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0)$
$y = (0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1)$

$SMC = (F11 + F00) / (F01 + F10 + F11 + F00)$
$= (0+7) / (2+1+0+7) = 0.7$

$J = (F_{11}) / (F_{01} + F_{10} + F_{11}) = 0 / (2 + 1 + 0) = 0$

③ 余弦相似度 Cosine Similarity

x,y为文档向量

$$\cos(x,y) = \frac{\langle x,y\rangle}{\|x\|\,\|y\|} = \frac{\langle x',y'\rangle}{\|x\|\,\|y\|},$$

'表示转置，$\langle x,y\rangle$ 表示向量内积.

日期: /

$\langle x, y \rangle = \sum_{k=1}^{n} x_k y_k = x^T y$

且 $\|x\|$ 是向量 x 的长度，$\|x\| = \sqrt{\sum_{k=1}^{n} x_k^2} = \sqrt{\langle x, x \rangle}$

应用：非对称属性，可以处理非二元向量

欧式距离衡量空间点的直线距离，余弦距离衡量点在空间的方向差异

•Example:

$d_1 = 3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0$
$d_2 = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2$

$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$
$\|d_1\| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$
$\|d_2\| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.245$

$\cos(d_1, d_2) = .3150$

6. 相关性

① 应用：用于测量两个变量之间或两个对象之间的关系。

② 若2个数据对象中的值来自不同属性，使用相关性来度量属性之间相似度

日期: /

- 皮尔森相关系数 Pearson's Correlation
  - 度量两个变量之间的线性相关性

$$\mathrm{corr}(\boldsymbol{x},\boldsymbol{y}) = \frac{\mathrm{covariance}(\boldsymbol{x},\boldsymbol{y})}{\mathrm{standard\_deviation}(\boldsymbol{x}) \times \mathrm{standard\_deviation}(\boldsymbol{y})} = \frac{s_{xy}}{s_x s_y}$$

协方差与标准差之比

$$\mathrm{covariance}(\boldsymbol{x},\boldsymbol{y}) = s_{xy} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y})$$

$$\mathrm{standard\_deviation}(\boldsymbol{x}) = s_x = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})^2}$$

$$\mathrm{standard\_deviation}(\boldsymbol{y}) = s_y = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k-\bar{y})^2}$$

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n}x_k \text{ 是 } \boldsymbol{x} \text{ 的均值} \qquad \bar{y} = \frac{1}{n}\sum_{k=1}^{n}y_k \text{ 是 } \boldsymbol{y} \text{ 的均值}$$

- 对于**非线性的相关性**难以建模
- X = (-3, -2, -1, 0, 1, 2, 3)
- Y = (9, 4, 1, 0, 1, 4, 9)

Y = X^2

负相关　　正相关

- Mean(X) = 0, Mean(Y) = 4

- 皮尔森相关系数 Correlation
  = (-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+3(5)
  = **-15 + 0 + 3 + 0 - 3 + 0 + 15**
  = 0 (即相关度为0)