

# 第 2 章：数据

- 《机器学习》(第1-2章), 周志华, 清华大学出版社, 2016
- 《数据挖掘导论》(第2版), 陈封能等著, 段磊等译, 机械工业出版社, 2019
- 《机器学习精讲》(看第1章), 安德烈-布可夫著, 人民邮电出版社, 2020

# 什么是数据?

- **数据集**: 一组含有**特征(属性)**的数据对象(样本)的集合
- **样本**: 由一组**特征(属性)**所描述的一个对象
  - 样本也成为**示例**、记录、点、向量、模式、事件、案例、实例、观测、实体
- **特征(属性)**: 描述**样本或对象**在某方面的表现或性质的事项
  - 例如: 某个人的身高、北京某一时刻的气温等等
  - **特征**也叫做**属性**、变量、特性、字段、维
  - 属性因对象而已, 或随时间而变化

**特征 (Features)**

<i>Tid</i>	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)**

# 数据的数学表示

- 一般地，令  $D = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$  表示  $m$  个样本的数据集，每个样本由  $d$  个属性描述，则每个样本  $\mathbf{x}_i = (x_{i1}; x_{i2}; \dots; x_{id})$  是  $d$  维样本空间  $\mathcal{X}$  中的一个向量， $\mathbf{x}_i \in \mathcal{X}$ ，其中  $x_{ij}$  是  $\mathbf{x}_i$  在第  $j$  个属性上的取值， $d$  称为样本  $\mathbf{x}_i$  的维数。 $(\mathbf{x}_i, y_i)$  表示第  $i$  个样例， $y_i$  是样本  $\mathbf{x}_i$  的标记(有时也称标签或标注)。
- 如下表，样本数  $m=17$ ， $\mathbf{x}_1 = (\text{青绿}; \text{蜷缩}; \text{浊响}; \text{清晰}; \text{凹陷}; \text{硬滑})$
- $x_{21} = \text{“乌黑”}$  是  $\mathbf{x}_2$  在第1个属性上的取值

描述西瓜的6个特征/属性

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜	标记 (label)
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	

# 特征(属性)的类型

# 特征的取值

- 特征的取值是指派给一个特征的数字或符号
  - 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}
- 相同的特征可以有不同的取值
  - **特征**：{姓名、身高}，**取值**：{张三、175cm}; {李四、180cm}
- 不同的特征可以有相同的取值
  - **特征**：{年龄、排名}，**取值**：{18、18}

# 特征(属性)的类型 (测量标度的类型)

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、 $\chi^2$ 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

- 定性属性不具有数的大部分性质
- 定量属性用数表示，并且具有数的大部分性质，可以是连续值或整数值

# 特征的类型（测量标度的类型）

- **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

# 特征的类型本质区别

- 特征类型的本质区别是其所对应的**操作**不同

- 相异性:**  $=$   $\neq$

- 序:**  $<$   $>$

- 加法:**  $+$   $-$

- 乘法:**  $*$   $/$


- 不同的特征类型适用不同的操作

- 标称类型:** 相异性

- 序数类型:** 相异性、序

- 区间类型:** 相异性、序、加法

- 比率类型:** 相异性、序、加法、乘法



下方属性类型拥有  
上方属性类型上的  
所有性质与操作

(属性类型的定义是累积的)



# 特征的类型

- **标称类型 Nominal**

适用符号：= ≠

- 例如：学号、籍贯、邮政编码

- **序数类型 Ordinal**

适用符号：= ≠ < >

- 例如：排名、年级、衣服的号码{S, M, L, XL, XXL}

- **区间类型 Interval**

- 有顺序，可以比大小，数据的差值有意义，但比例没有意义，可以加减，不能乘除（但可以算平均值）
- 例如：日历上的日期、摄氏或华氏温度等

- **比率类型 Ratio**

- 有顺序，可以比大小，数据差值和比率都有意义，可以四则运算。例如：开尔文温度、长度、时间、年龄、质量等

# 区间类型和比率类型的区别

• 举个栗子：“**10°的水的温度是5°水温度的两倍**” 这句话是否成立？

## • 摄氏温度？

- 冰水混合物的温度定为0摄氏度（0不代表无，通常是一个分界值），沸水的温度定为100摄氏度。有顺序，可以比大小，数据的差值有意义，可以加减，不能乘除。可以说10°C比5°C高，且高5°C，但是不能说是两倍

## • 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（0不代表无），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（0代表无），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零度是硬性规定的，其比率是无物理意义的

# 区间类型和比率类型的区别

- 类似的例子

- 小明的身高比全班的平均身高高3公分，小磊的身高比平均身高高6公分
- 是否能说：“小磊的身高是小明身高的2倍”？

- |                |                     |
|----------------|---------------------|
| • <b>区间类型：</b> | 向异性、序、加法            |
| • <b>比率类型：</b> | 向异性、序、加法、 <b>乘法</b> |

# 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°C~15°C。
  - 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。

# 特征的类型

- 不同类型的特征对应不同的坐标变换方式

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

# 特征类型的其它分类方式：离散与连续

- 用值的个数描述属性
- **离散特征 (Discrete Feature)**
  - 具有有限或无限可数个值， 例： 邮政编码, 计数, 所采集文档的单词集...
  - 常常使用整数变量表示
  - 可以是分类的（定性的）也可以是数值的（定量的）
- **连续特征 (Continuous Feature)**
  - 取实数值的特征， 例： 温度、高度、重量...
  - 在实际使用中， 实数只能用有限的精度测量与表示
  - 连续特征常常用浮点变量表示

# 特征类型的其它分类方式：离散与连续

## • 二元特征 (Binary Feature)

- 仅仅具有两个值的特征，常常使用0或1表示
- 例：性别、对错、选课与否...
- 二元属性是离散特征的一种特殊情形

## • 非对称特征 (Asymmetric Binary Feature)

- 出现非0值才重要的特征，其状态的结果不是同等重要
- 只有非0值才重要的二元特征称为非对称二元特征
- 例：体检结果阳性(1)与阴性(0)，大部分情况下该属性都为0，因此我们一般只关注属性为1的情况，所以这个就是非对称的二元特征(属性)。

# 数据质量的重要性

- 在样本小于10000的情况下，保证数据标注的一致性，可以提升效果。达到相同的效果，需要的数据量也会减少。
- 无法获得更多数据的时候，提高数据质量至关重要。
- 清洗脏数据与扩大一倍数据集带来的提升效果相当。
- 大部分的情况，训练数据都在10000条以下（小数据）。提升数据质量带来的效果提升比大数据集更明显
- 即使像网络搜索、自动驾驶和推荐系统这些大数据任务，其中有很多的长尾事件其实也是小数据问题。

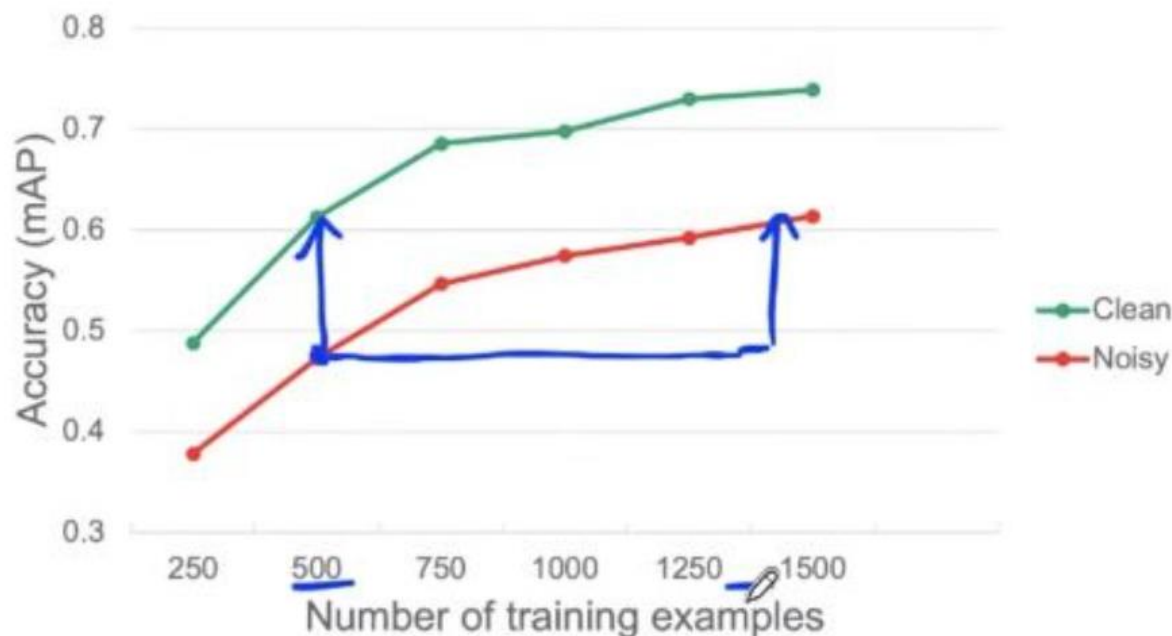


# 数据的质量的重要性

- 调优模型的效果提升 < 调优数据质量的效果提升
- 调优数据比调优模型更有效



## Example: Clean vs. noisy data



Note: Big data problems where there's a long tail of rare events in the input (web search, self-driving cars, recommender systems) are also small data problems.

# 数据的质量的重要性

- 好的数据是指：
  - 标签定义的一致性：（定义标签 $y$ 是不清晰的）
  - 覆盖所有的代表性案例：（对输入 $x$ 的覆盖面）
  - 生产环境数据分布变化的及时反馈（分布覆盖数据漂移和概念漂移）
  - 合适的数据集大小

# 样本的相似性和相异性

参考：《数据挖掘导论》(第2版)，陈封能等著，段磊等译，机械工业出版社，2019，第2章

# 样本相似性和相异性

---

- 相似度和相异度的基本概念
- 相异度 (距离(Distances))
- 相似度 (Similarities)
- 相关性 (Correlation)

# 相似度和相异度的度量

- 相似度 (similarity)

- 两个样本相似程度的**数值化**度量
- 两个样本越相似，他们之间的相似性就越高
- 相似度是非负的，**通常**取值范围在 $[0, 1]$

- 相异度 (dissimilarity)

- 两个样本之间差异程度的**数值化**度量
- 两个样本越相似，他们之间的相异性越低
- 相异度是非负的，取值在 $[0, 1]$ 和 $[0, \infty)$ 均有

- 通常术语**距离 (distance)** 用作**相异度**的同义词

# 数据相似性和相异性

---

- 相似度和相异度的基本概念
- **距 离 (Distances)**
- 相似性 (Similarities)
- 相关性 (Correlation)

# 简单属性(单一特征)的相似度和相异度

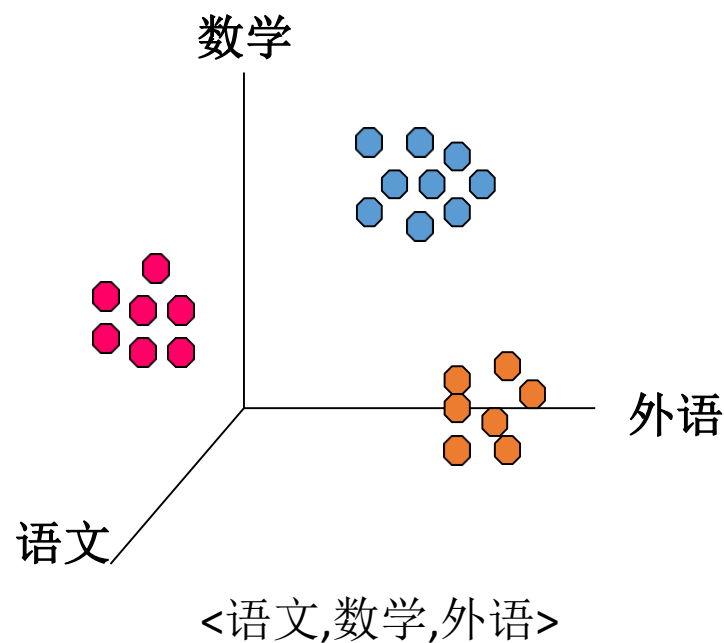
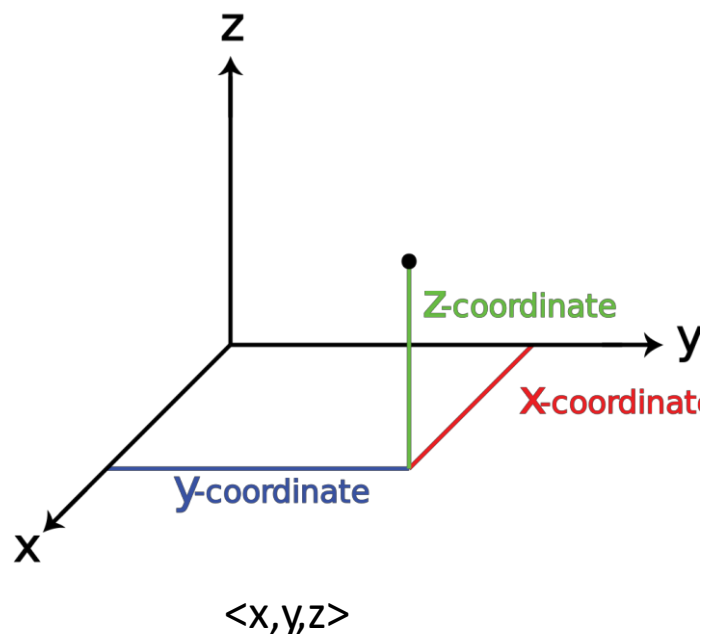
- $x$  和  $y$  分别对应两个样本的某一个属性值

## 常用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	$d = \frac{ x-y }{(n-1)}$ <p>值映射到整数 0 到 <math>n-1</math>，其中 <math>n</math> 是值的个数</p>	$s = 1 - d$
区间或比率的	$d =  x - y $	$s = -d, \quad s = \frac{1}{1+d}, \quad s = e^{-d},$ $s = 1 - \frac{d - \min\_d}{\max\_d - \min\_d}$

# 坐标系与维度

- 在真实空间的坐标系下，空间中每一个点都可以表达为  $(x, y, z)$  的三维向量。
- 类似地，对于数据当中的每一个样本，都可以看作以特征为坐标系的高维空间中的一个点。





- 欧几里德距离

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

其中  $n$  是对象的数据维度(属性的个数),  $x_k$  和  $y_k$  分别是数据对象  $\mathbf{x}$  和  $\mathbf{y}$  的第  $k$  个属性。

四个点的  $x$  和  $y$  坐标

点	$x$ 的坐标	$y$ 的坐标
p1	0	2
p2	2	0
p3	3	1
p4	5	1

# 闵氏距离 Minkowski Distance

- 闵可夫斯基距离

- 闵氏距离的欧式距离的一种**泛化**，欧式距离是闵氏距离的一种特例

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

其中  $n$  是对象的数据维度(属性的个数)， $x_k$  和  $y_k$  分别是数据对象  $\mathbf{x}$  和  $\mathbf{y}$  的第  $k$  个属性。



闵可夫斯基  
1864—1909

# 闵氏距离 Minkowski Distance

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

- $r=1$ , 曼哈顿距离, **L1范数, L1-norm**

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_1 = \sum_{k=1}^n |x_k - y_k|$$

- $r=2$ , 欧氏距离, **L2范数, L2-norm**

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{\sum_{k=1}^n |x_k - y_k|^2}$$

- $r=\infty$ , 上确界距离,  **$\infty$ 范数, L-norm**

- 对象各个属性之间的最大距离, 即上确界。

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

# 距离的典型特性

- 距离（如欧几里得距离）满足以下三个特性：
  - 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  $d(p, q) = 0$ .
  - 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  $r$ ,  $d(p, r) \leq d(p, q) + d(q, r)$ .
- 其中  $d(p, q)$  是  $p$  和  $q$  之间的距离。
- 满足上述三个特性的距离也成为一种度量 (**metric**)

# 数据相似性

---

- 相似度和相异度的基本概念
- 距 离 (Distances)
- **相似性 (Similarities)**
- 相关性 (Correlation)

# 数据对象之间的相似度性质

如果  $s(\mathbf{x}, \mathbf{y})$  是数据点  $\mathbf{x}$  和  $\mathbf{y}$  之间的相似度

- 1) 仅当  $\mathbf{x} = \mathbf{y}$  时  $s(\mathbf{x}, \mathbf{y}) = 1$ 。 $(0 \leq s \leq 1)$  (非负性)
- 2) 对于所有  $\mathbf{x}$  和  $\mathbf{y}$ ,  $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$ 。(对称性)

# 二元数据的相似度量

- 对两个二元向量  $\mathbf{x}$  和  $\mathbf{y} \in \mathbb{R}^n$ ，值取0或1，两个对象的比较可以生成如下的四个量：

$f_{01}$  =  $\mathbf{x}$  取 0 且  $\mathbf{y}$  取 1 的属性数,  $f_{10}$  =  $\mathbf{x}$  取 1 且  $\mathbf{y}$  取 0 的属性数,  $f_{00}$  =  $\mathbf{x}$  取 0 且  $\mathbf{y}$  取 0 的属性数,  $f_{11}$  =  $\mathbf{x}$  取 1 且  $\mathbf{y}$  取 1 的属性数

- 简单匹配系数 (Simple Matching Coefficient, **SMC**) 定义如下：

$$\begin{aligned}\text{SMC} &= \text{值匹配的属性个数} / \text{属性个数} \\ &= (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})\end{aligned}$$

- SMC通常在0和1之间取值，0代表对象一点也不相似，1代表对象完全相同。SMC对出现和不出现的都计数，因此，SMC对仅包含是非题的测验中发现问题回答相似的学生。
- 如果样本的属性都是**对称的二值离散型属性**，则样本间的距离可用简单匹配系数计算。
- 对称的二值离散型属性**是指属性取值为1或者0同等重要，例如：性别就是一个对称的二值离散型属性，即：用1表示男性，用0表示女性；或者用0表示男性，用1表示女性是等价的，属性的两个取值没有主次之分。

- Jaccard 系数

$$J = \frac{\text{匹配个数}}{\text{00匹配中不涉及的属性个数}} \\ = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

- 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。



# 二元数据的相似度度量

- SMC和Jaccard系数

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

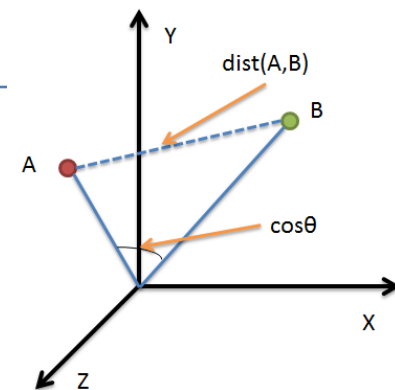
$$y = (0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1)$$

$$\begin{aligned} \text{SMC} &= (F_{11} + F_{00}) / (F_{01} + F_{10} + F_{11} + F_{00}) \\ &= (0+7) / (2+1+0+7) = 0.7 \end{aligned}$$

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

- 当你和你的朋友在商场相遇时，你会不会说：
  - “咱俩真是有猿粪！商场里面的上万种商品，咱们基本上都没有买。”

# 余弦相似度 Cosine Similarity



如果  $\mathbf{x}$  和  $\mathbf{y}$  是两个文档向量，则

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

其中'表示向量或者矩阵的转置， $\langle \mathbf{x}, \mathbf{y} \rangle$ 表示两个向量的内积：

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

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

- 在向量的点积当中，0-0匹配是没有贡献的，因此，**余弦相似度和Jaccard系数一样，适用于非对称属性**。同时，余弦相似度**还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。

# 余弦相似度 Cosine Similarity

•Example:

$$d_1 = \mathbf{3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0}$$

$$d_2 = \mathbf{1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2}$$

$$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

$$\|d_1\| = (3*3 + 2*2 + 0*0 + 5*5 + 0*0 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0)^{0.5} = (42)^{0.5} = 6.481$$

$$\|d_2\| = (1*1 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 1*1 + 0*0 + 2*2)^{0.5} = (6)^{0.5} = 2.245$$

$$\cos(d_1, d_2) = 0.3150$$

# 数据相似性

---

- 相似度和相异度的基本概念
- 距 离 (Distances)
- 相似性 (Similarities)
- **相关性 (Correlation)**

# 相关性 Correlation

- 相关性被用于测量两个变量(高度和重量)之间或两个对象之间的关系
- 若两个数据对象中的值来自不同的属性, 可使用相关性来度量属性之间的相似度

相似/相异

Features 相关

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

抽烟和肺病的相似性

抽烟和肺病的相关性

# 相关性 Correlation

- 皮尔森相关系数 Pearson's Correlation

- 度量两个变量之间的线性相关性

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

协方差与标准差之比

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

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

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

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

# 皮尔森系数的局限性

- 对于**非线性的相关性**难以建模

- $X = (-3, -2, -1, 0, 1, 2, 3)$
- $Y = (9, 4, 1, 0, 1, 4, 9)$

$Y = X^2$

<div style="border: 2px solid red; padding: 5px; display: inline-block;"><math>(-3, -2, -1, 0)</math></div> 负相关	<div style="border: 2px solid red; padding: 5px; display: inline-block;"><math>(1, 2, 3)</math></div> 正相关
--	--

- $\text{Mean}(X) = 0, \text{Mean}(Y) = 4$

- 皮尔森相关系数 Correlation

$$= (-3)(5) + (-2)(0) + (-1)(-3) + (0)(-4) + (1)(-3) + (2)(0) + 3(5)$$

$$= -15 + 0 + 3 + 0 - 3 + 0 + 15$$

$$= 0 \text{ (即相关度为0)}$$