一、相似度:

- 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]

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

1）仅当 $\boldsymbol{x}=\boldsymbol{y}$ 时 $s(\boldsymbol{x}, \boldsymbol{y})=1$。$(0\leqslant s\leqslant 1)$（非负性）

2）对于所有 $\boldsymbol{x}$ 和 $\boldsymbol{y}$，$s(\boldsymbol{x}, \boldsymbol{y})=s(\boldsymbol{y}, \boldsymbol{x})$。（对称性）

简单匹配系数 SMC：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

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

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

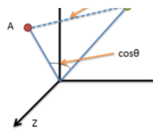
Jaccard 系数：应用于非对称二元属性；不对称的二值离散型属性是指属性取值为 1 或者 0 不是同等重要

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

余弦相似度：cos<x,y>：衡量点在空间的方向差异，适应于非对称属性，还可以处理非二元向量

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

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



其中′表示向量或者矩阵的转置，$<\boldsymbol{x}, \boldsymbol{y}>$表示两个向量的内积：

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

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

二、相异度:（距离）

- 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, ∞)均有

• 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。

属性类型	相异度	相似度
标称的	$d=\begin{cases}0 & 如果\ x=y \\ 1 & 如果\ x\neq y\end{cases}$	$s=\begin{cases}1 & 如果\ x=y \\ 0 & 如果\ x\neq y\end{cases}$
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	$d=\lvert x-y\rvert$	$s=-d$，$s=\frac{1}{1+d}$，$s=e^{-d}$， $s=1-\frac{d-\text{min\_d}}{\text{max\_d}-\text{min\_d}}$

• 欧几里德距离

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

• 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

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

• r=1, 曼哈顿距离，**L1范数，L1-norm**

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

• **r=2, 欧氏距离，L2范数，L2-norm**

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

• r=∞， 上确界距离，**∞范数，L-norm**

• 对象各个属性之间的最大距离，即上确界。

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

三、相关性：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，使用相关性来度量属性之间的相似度

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

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

协方差与标准差之比

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

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

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

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

- 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)