本课件仅用于教学使用。未经许可，任何单位、组织和个人不得将课件用于该课程教学之外的用途(包括但不限于盈利等)，也不得上传至可公开访问的网络环境

数据科学导论

Introduction to Data Science

数据入门：数据相关性与集成

陈恩红，黄振亚
Email: cheneh@ustc.edu.cn, huangzhy@ustc.edu.cn

课程主页：
http://staff.ustc.edu.cn/~huangzhy/Course/DS2026.html

回顾：数据分析基础

- 数据采集　　Data Collection
- 数据预处理　　Data Preprocessing
- 特征工程　　Feature Engineering



数据预处理

- 大数据环境下的数据特点
- 为什么需要进行预处理
- 数据预处理
  - 数据清洗
  - 数据统计分析
  - **数据相关性与数据集成**
  - 数据变换

数据预处理：数据集成

- 数据集成
  - 将多个数据源的数据整合到一个一致的数据存储中
- 数据集成的目标
  - 获得更多的数据
  - 获得更完整的数据
  - 获得更全面的数据画像，如用户画像
- 例： 电商推荐-需求
  - **用户的购物记录：淘宝，美团，拼多多 等**
  - **用户的社交网络：微博，facebook等**
  - **用户的视频记录：爱奇艺，抖音等**
  - 。。。

数据预处理：数据集成

- 数据集成
  - 将多个数据源的数据整合到一个一致的数据存储中
  - 集成数据（库）时，经常出现冗余数据
    - 冗余数据带来的问题: 浪费存储、重复计算
    - 冗余的属性
    - 冗余的样本
  - 例如：
    - 用户的电商记录出现在很多app中
    - 用户的个人信息在多个app中
    - 。。。

数据相关性

- 数据相关性
  - 度量两个或多个属性（变量）之间的关联关系（正相关、负相关、不相关）
- 简单示例：某饮品店连续 5 天记录的气温与冰咖啡销量
  - 气温升高，销量明显升高。二者呈正相关（直观观察即可解决）
- 实际场景：生产流水线的参数，不同参数对产品质量的影响如何？
  - 参数数量成百上千，难以直接观察（**需要借助合适的算法分析数据相关性**）

日期	日最高温（℃）	冰咖啡销量（杯）
周一	20	45
周二	24	60
周三	28	95
周四	32	140
周五	36	190



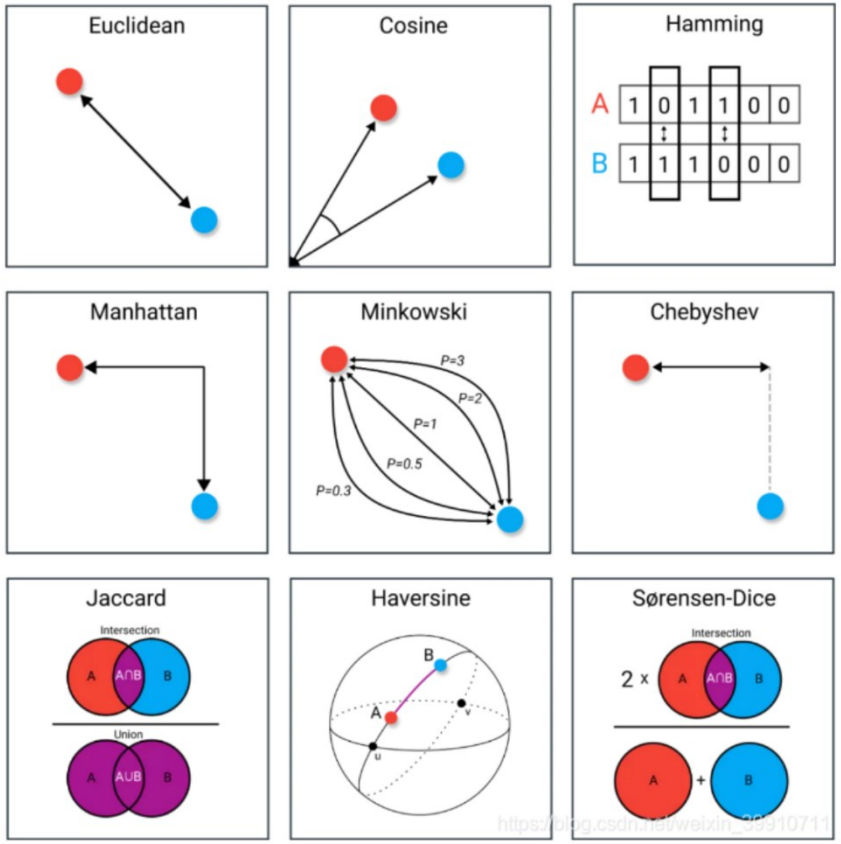
数据相关性

- 相关性分析方法：
  - 相似度、距离度量
  - 皮尔逊相关系数、斯皮尔曼相关系数
  - 卡方检验、互信息……
- 应用举例：
  - 数据集成：整合多源数据，识别冗余属性
  - 特征工程：构建算法模型，筛选关键特征
  - 异常检测：监控多维数据，发现违背关联的孤立点
  - 模式挖掘：遍历事务数据，提取频繁共现规律

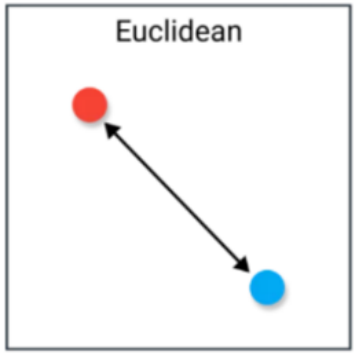
数据相关性分析是大数据预处理和特征挖掘中不可或缺的基础工具

数据相关性

- 检测冗余样本
  - 思想：数据样本之间的相关性，数据融合、去除冗余
  - **方法：距离度量**
    - 欧几里得距离
    - 汉明距离
    - 明氏距离
    - 马氏距离
    - ……
  - **方法：相似度计算**
    - 余弦相似度
    - Jaccard相似度
    - ……



数据相关性

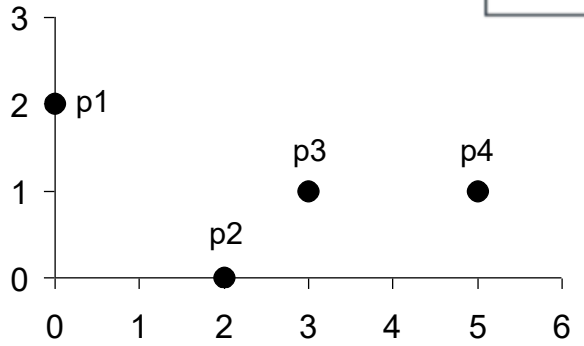


- 数据的距离度量
  - 欧几里得距离(Euclidean Distance)

$$dist = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$$

n 表示数据p和q维度数

p_k 和 q_k 表示数据p和q的第k个属性



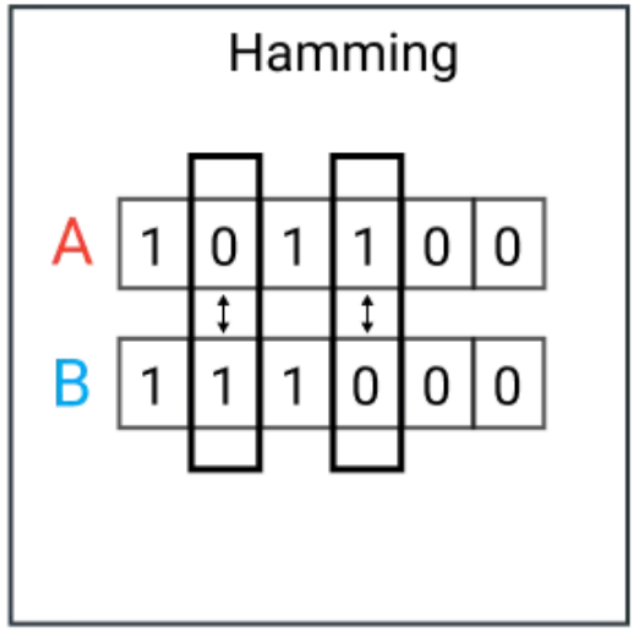
	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

对数据标准化非常重要

数据相关性

- 数据的距离度量
  - 汉明距离(Hamming Distance)
  - 定义：两个向量之间不同值的个数
    - 字符串比较：比较两个相同长度的二进制字符串
  - 要求：向量长度相同
  - 常用：HASH场景



Defu Lian, Haoyu Wang, **Enhong Chen,** Xing Xie. LightRec: a Memory and Search-Efficient Recommender System. **WWW 2020**.

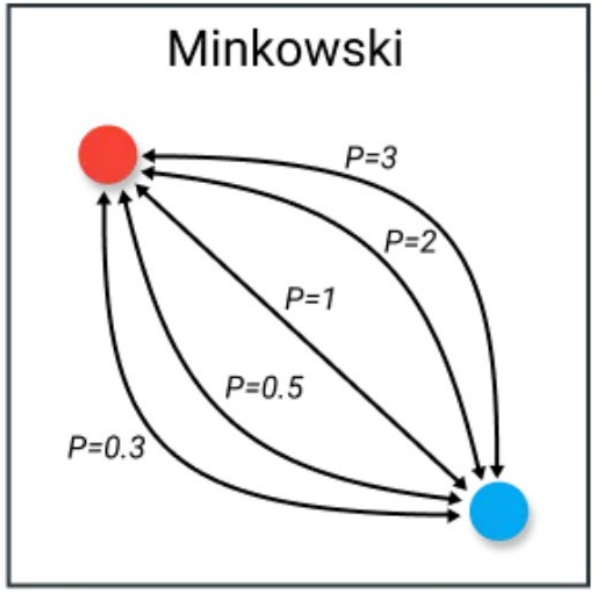
数据相关性

- 数据的距离度量
  - 明氏距离（Minkowski Distance）
    - 距离度量：通用表达形式

$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

r 是参数

n 表示数据p和q维度数，p_k 和 q_k 表示数据p和q的第k个属性

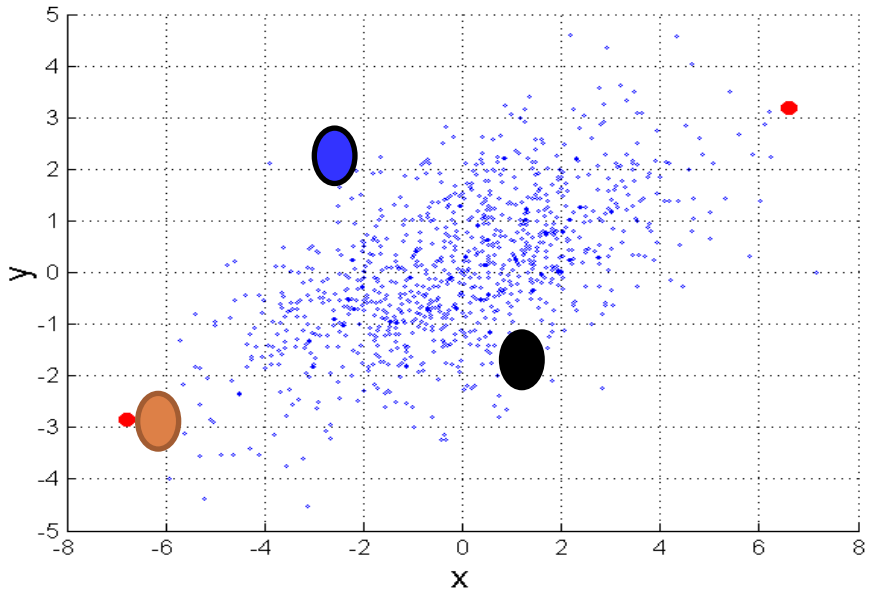


  - r=1：曼哈顿距离
  - r=2：欧氏距离
  - r=∞：切比雪夫距离

数据相关性

- 数据的距离度量
  - 马氏距离：数据的协方差距离
    - 欧氏距离的扩展，考虑到各种特性之间的联系（协方差）

$$s(p-q)=(p-q)\Sigma^{-1}(p-q)^T$$



Σ 是总体样本 ***X***的协方差矩阵

$$\Sigma_{j,k}=\frac{1}{n-1}\sum_{i=1}^{n}(X_{ij}-\overline{X}_j)(X_{ik}-\overline{X}_k)$$

- 确定未知样本集与已知样本集的相似度
- 它考虑了数据集的相关性，并且是比例不变的

红色的数据点, 欧氏距离为14.7, 马氏距离为6

数据相关性

- **马氏距离 vs 欧氏距离**
  - **假设：**以厘米为单位测量人的身高，以克（g）为单位测量人的体重。每个人被表示为一个两维向量。如：一个人身高173cm，体重50000g，表示为（173,50000），根据身高体重来判断人的体型的相似程度
  - **已知：**小明(160,60000)；小王(160,59000)；小李(170，60000)。小明与谁的体型更相似？

分析：根据常识可以知道小明和小王体型相似。但是如果根据**欧氏距离**来判断，小明和小王的距离要远大于小明和小李之间的距离，即小明和小李体型相似

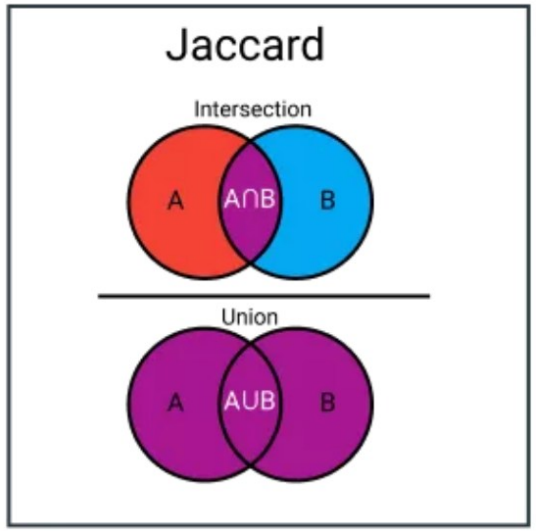
原因：不同特征的**度量标准之间存在差异**而导致判断出错
以克（g）为单位测量人的体重，数据分布比较分散，即方差大，以厘米为单位来测量人的身高，数据分布就相对集中，方差小

马氏距离把方差归一化，使得特征之间的关系更加符合实际情况

数据相关性

- 数据的相似度计算
  - **简单匹配 Simple Matching VS Jaccard相关系数**
  - 离散数据，属性的取值表示为0或1
  - 例：数据p和q，定义如下4个变量
    - F01：p为0、q为1的属性数量
    - F10：p为1、q为0的属性数量
    - F00：p为0、q为0的属性数量
    - F11：p为1、q为1的属性数量

$$p = (1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0)$$
$$q = (0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1)$$



SMC = number of matches / number of attributes
= (F11 + F00) / (F01 + F10 + F11 + F00)

Jaccard = **number of F11 matches** / **number of non-zero attributes**
= (F11) / (F01 + F10 + F11)

数据相关性

- 数据的相似度计算
  - **简单匹配 Simple Matching VS Jaccard相关系数**

    假设：存在该属性为1，不存在该属性为0

    p = (1 0 0 0 0 0 0 0 0 0)

    q = (0 0 0 0 0 0 1 0 0 1)

    p和q是否相关?

    F01 = 2 (p为0，q为1的属性数量)

    F10 = 1 (p为1，q为0的属性数量)

    F00 = 7 (p为0，q为0的属性数量)

    F11 = 0 (p为1，q为1的属性数量)

    SMC = (F11 + F00) / (F01 + F10 + F11 + F00)

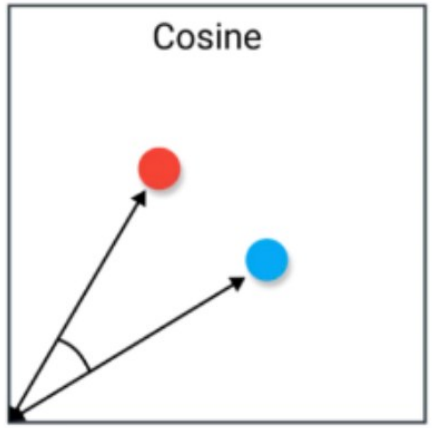
    = (0+7) / (2+1+0+7) = 0.7

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

数据相关性

- 数据的相似度计算
  - **余弦相似性 (Cosine Similarity)**

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}.$$



  - **例：**

A = 3 2 0 5 0 0 0 2 0 0
B = 1 0 0 0 0 0 0 1 0 2

$A \bullet B = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

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

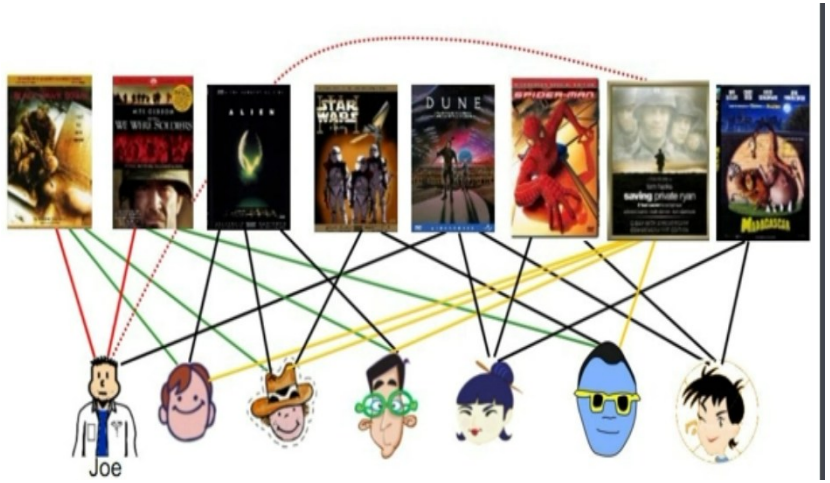
$||B|| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.245$

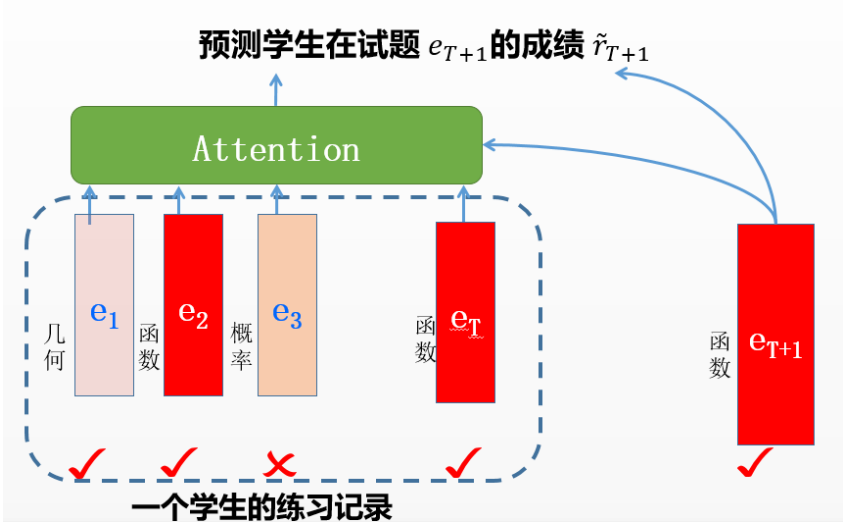
$\cos(A, B) = 0.3150$

思考：余弦相似度是不是一种距离？

数据相关性

- 数据的相似度计算
  - **余弦相似度 (Cosine Similarity)**
    - 推荐系统中，协同过滤算法(UCF, ICF)—经典算法
      - 用户(向量)的相似度度量，产品(向量)的相似度度量
    - 深度学习中，训练Attention（注意力机制）的权重
      - 基于注意力机制的学生成绩预测模型





Qi Liu, Zhenya Huang, Enhong Chen,, EKT: Exercise-aware Knowledge Tracing for Student Performance Prediction, TKDE

数据相关性

- 数据的相关性分析
  - **Pearson相关系数**
    - 衡量两个数据对象之间的线性关系
    - 数据标准化
  - 可以简单理解为：p和q的协方差/(p的标准差*q的标准差)

$$p_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \tilde{X})(Y_i - \tilde{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \tilde{X})^2 \sum_{i=1}^{n}(Y_i - \tilde{Y})^2}}$$

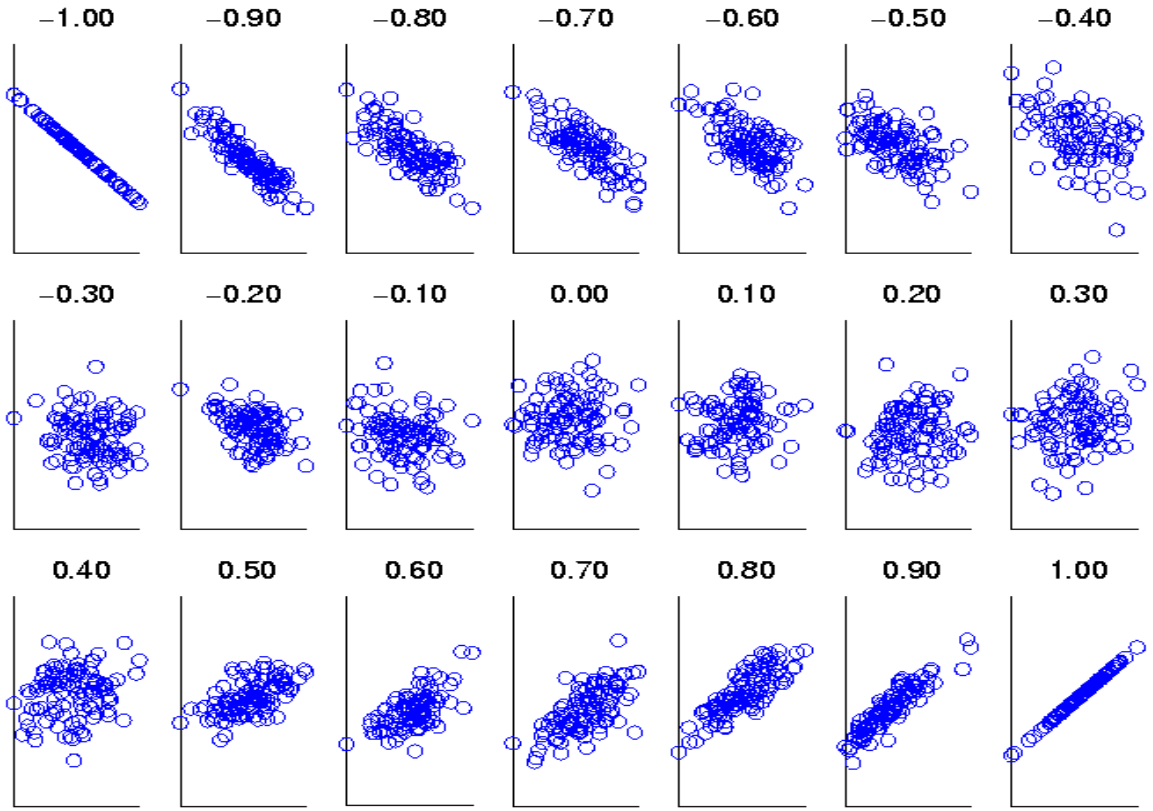
$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y},$$

Zhenya Huang, Qi Liu, Enhong Chen, et al, Question Difficulty Prediction for READING Problems in Standard Tests, AAAI'2017

数据相关性

- 数据的相关性
  - Pearson相关系数：衡量数据对象之间的线性关系

散点图显示相似度[-1, 1]



数据相关性

- 数据的相关性分析
  - Pearson**相关系数**：衡量数据对象之间的线性关系

- 例： **问：X与Y有没有关系？**
  - X = (-3, -2, -1, 0, 1, 2, 3)
  - Y = (9, 4, 1, 0, 1, 4, 9)

$$p_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \tilde{X})(Y_i - \tilde{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \tilde{X})^2 \sum_{i=1}^{n}(Y_i - \tilde{Y})^2}}$$

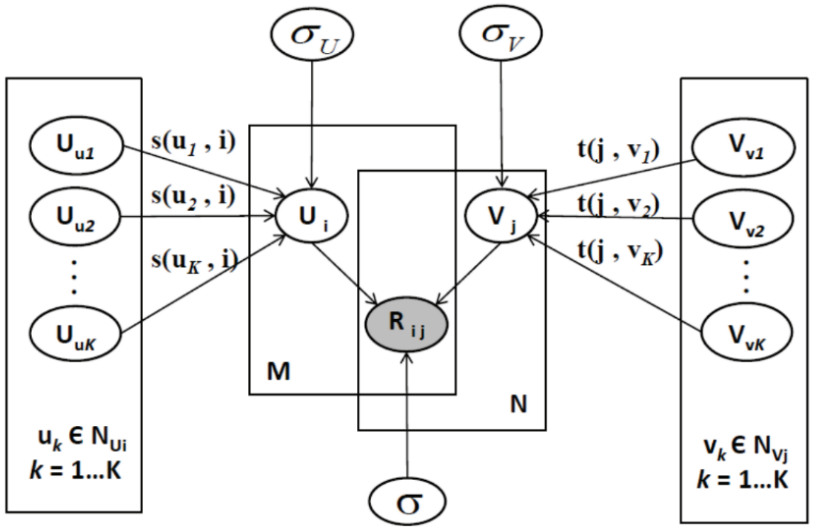
- Mean(X) = 0, Mean(Y) = 4
- Correlation= ？
  - = (-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+3(5)= 0

数据相关性

- 数据的相关性分析
  - 有时，不同的属性产生的影响不同
  - 在计算距离，相似度时，可以赋予数据属性的权重不同 (w_k)

$$\text{similarity}(\mathbf{x},\mathbf{y}) = \frac{\sum_{k=1}^{n} w_k \delta_k s_k(\mathbf{x},\mathbf{y})}{\sum_{k=1}^{n} \delta_k}$$

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



Le Wu, Enhong Chen, Qi Liu, Leveraging Tagging for Neighborhood-aware Probabilistic Matrix Factorization. CIKM2012

数据相关性

- 数据的相关性分析
- 练习题1：给定数据x和y，计算指定的相似性或距离
  - 余弦相似度、pearson相关度、欧几里得距离、Jaccard
    - 已知：**X = (0, 1, 0, 1)，　Y = (1, 0, 1, 0)**
    - 问：**分析X和Y的相关性**

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}.$$

$$p_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \tilde{X})(Y_i - \tilde{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \tilde{X})^2 \sum_{i=1}^{n}(Y_i - \tilde{Y})^2}}$$

$$dist = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$$

J = (F11) / (F01 + F10 + F11)