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

新媒体大数据分析

New Media Big Data Analysis

数据入门：数据清理

黄振亚，朱孟潇，张凯，唐秀秀

课程主页：
http://staff.ustc.edu.cn/~mxzhu/Course/NNM2026.html

助教：谭施越 朱嘉骏
bigdata_20262026@163.com

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

新媒体大数据分析

New Media Big Data Analysis

数据入门：数据集成

黄振亚，朱孟潇，张凯，唐秀秀

课程主页：
http://staff.ustc.edu.cn/~mxzhu/Course/NNM2026.html

助教：谭施越 朱嘉骏
bigdata_20262026@163.com



数据预处理

- 大数据环境下的数据特征
- 为什么需要进行预处理
- **预处理的基本方法**
  - 数据清理
  - **数据集成**
  - 数据变换
  - 数据规约



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：要求电商推荐
  - 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
  - 。 。 。

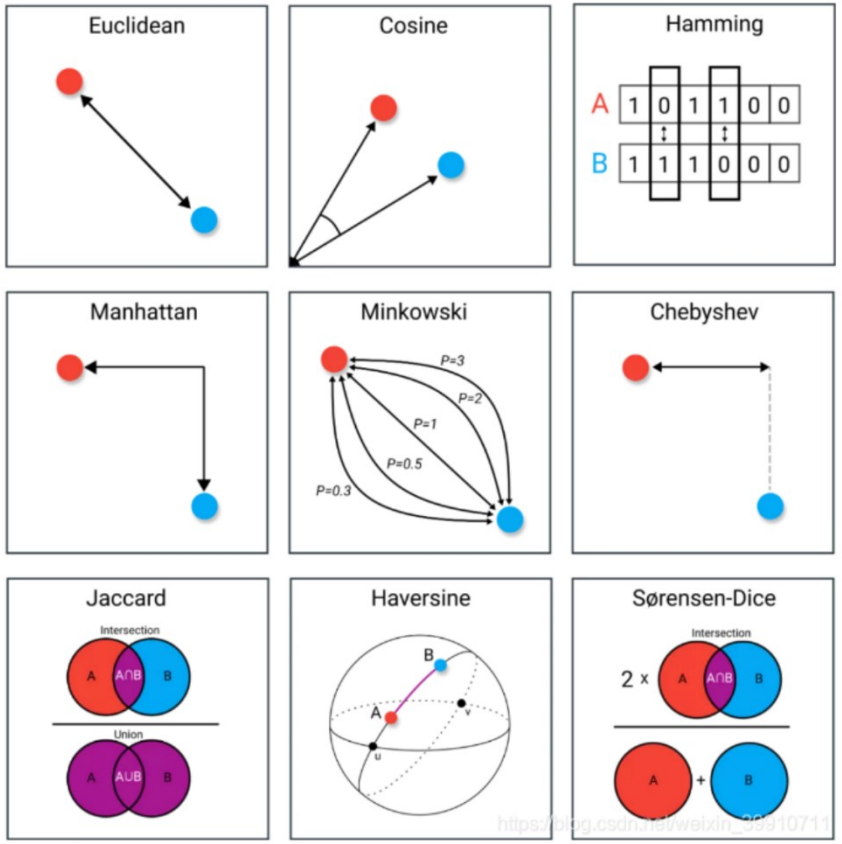


数据预处理：数据集成

- 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
    - 。。。

数据预处理：数据集成

- 检测冗余
  - 思想：数据样本（属性）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
    - 曼哈顿距离
    - 汉明距离
    - 明氏距离
    - ……
  - 方法：相似度计算
    - 余弦相似度
    - Jaccard相似度
    - ……



数据预处理：数据集成

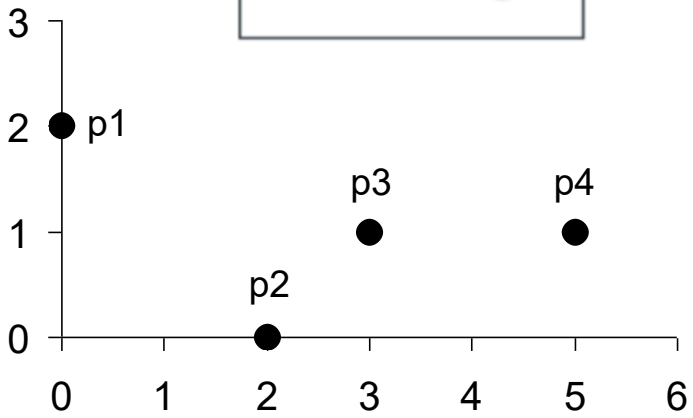


- 数据的距离度量
  - 欧几里得距离(Euclidean Distance)

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

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



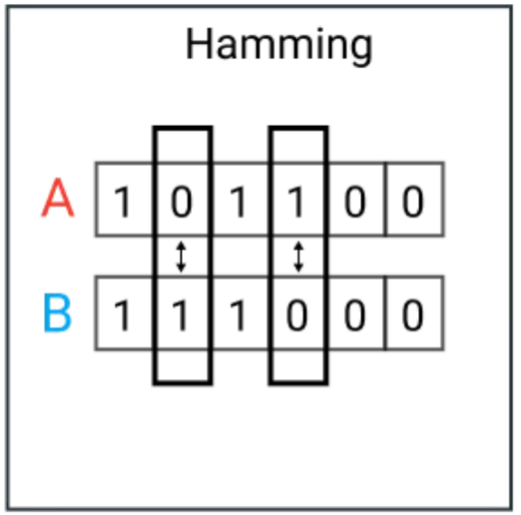
	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

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

数据预处理：数据集成

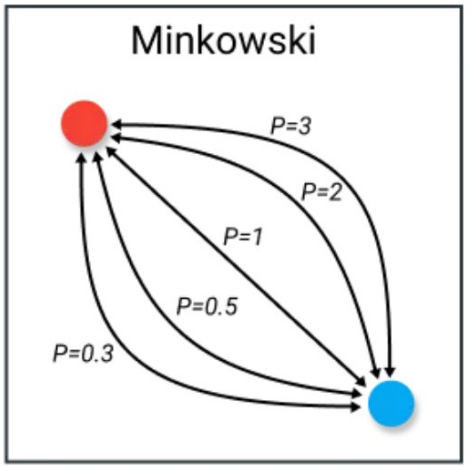
- 数据的距离度量
  - 汉明距离(Hamming Distance)
  - 两个向量之间不同值的个数。
    - 字符串比较：比较两个相同长度的二进制字符串
  - 要求：向量长度相同
  - 常用：HASH场景



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

数据预处理：数据集成

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



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

r 是参数

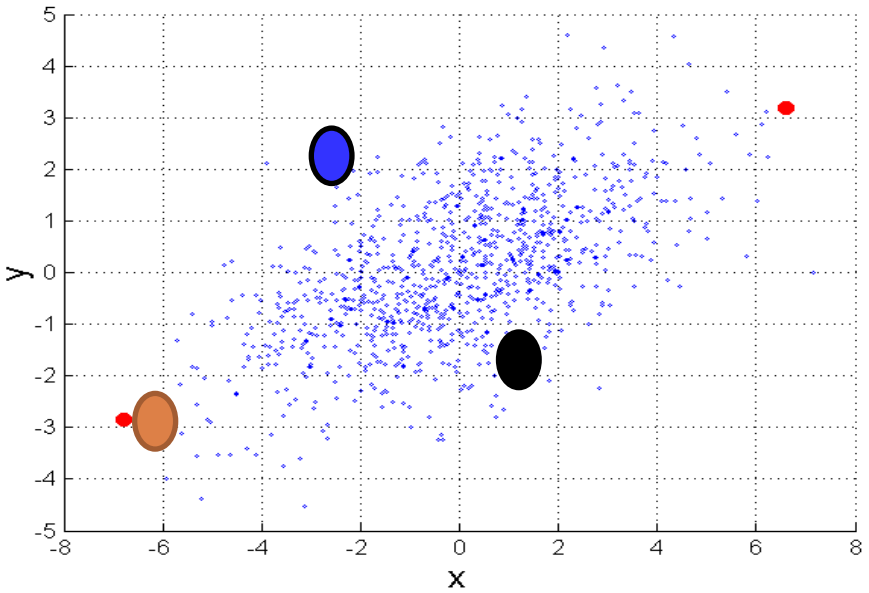
n 表示数据p和q维度数，p_k 和 q_k 表示数据p和q的第k个属性

- r=1：曼哈顿距离
- r=2：欧氏距离
- r=∞：切比雪夫距离

数据预处理：数据集成

- 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

数据预处理：数据集成

- 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

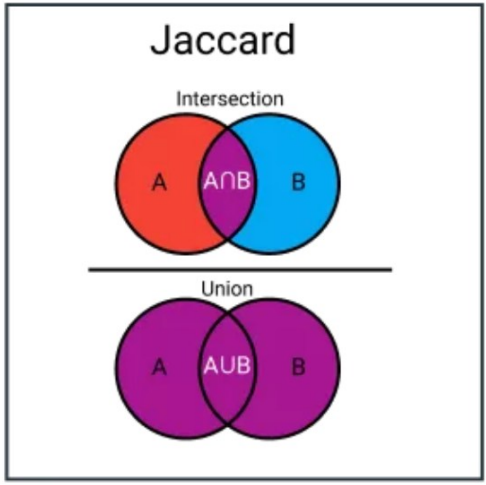
- 以克（g）为单位测量人的体重，数据分布比较分散，即方差大，
- 以厘米为单位来测量人的身高，数据分布就相对集中，方差小

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

数据预处理：数据集成

- 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 = **F11属性数量**/ **非0属性数量**
= (F11) / (F01 + F10 + F11)

数据预处理：数据集成

- 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)

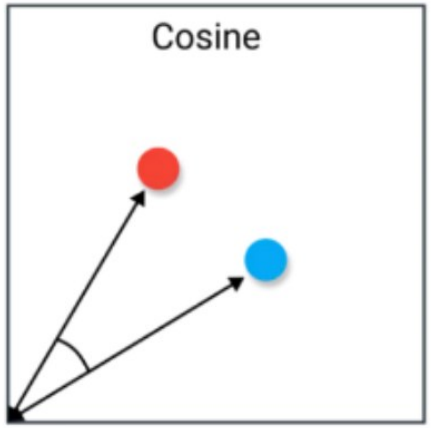
= (0+7) / (2+1+0+7) = 0.7

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

数据预处理：数据集成

- 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

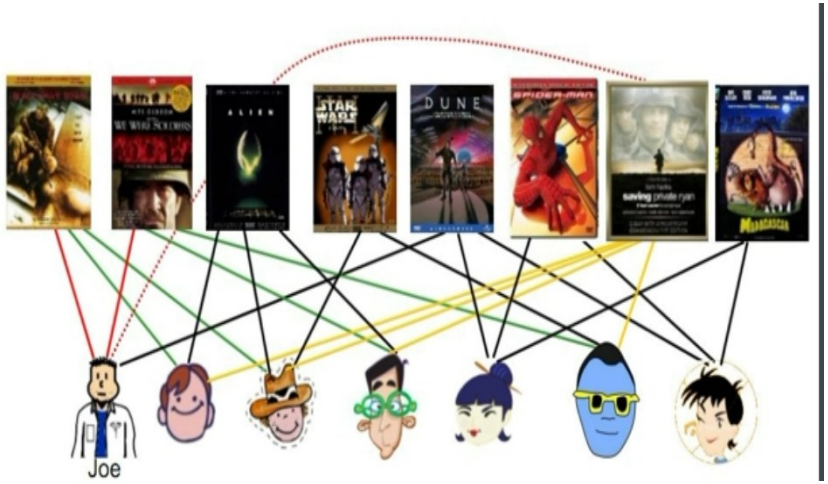
$||B||$ = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2) $^{0.5}$ = (6) $^{0.5}$ = 2.245

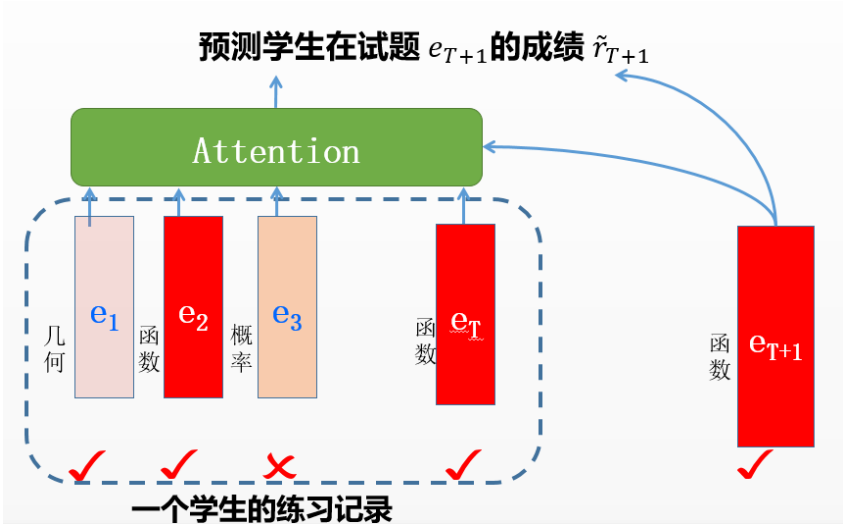
cos(A, B) = 0 .3150

思考：余弦相似度是不是一种距离？

数据预处理：数据集成

- 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

数据预处理：数据集成

- 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

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

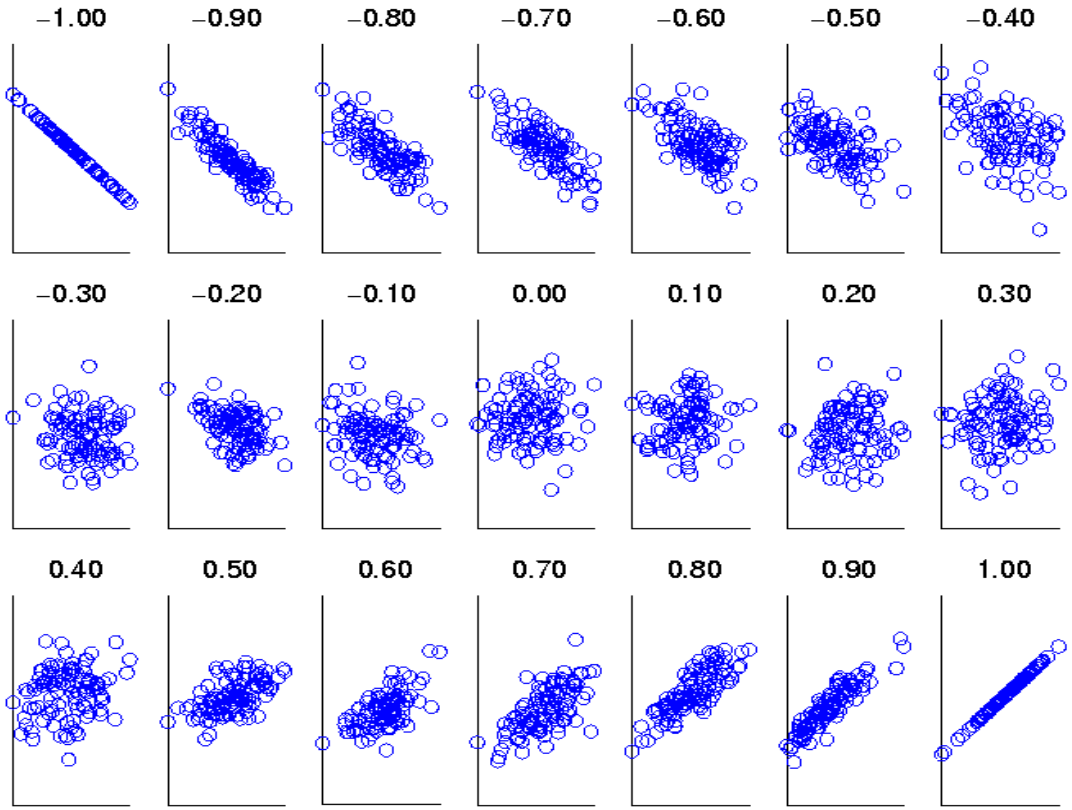
$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

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

数据预处理：数据集成

- 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]



数据预处理：数据集成

- 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)

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

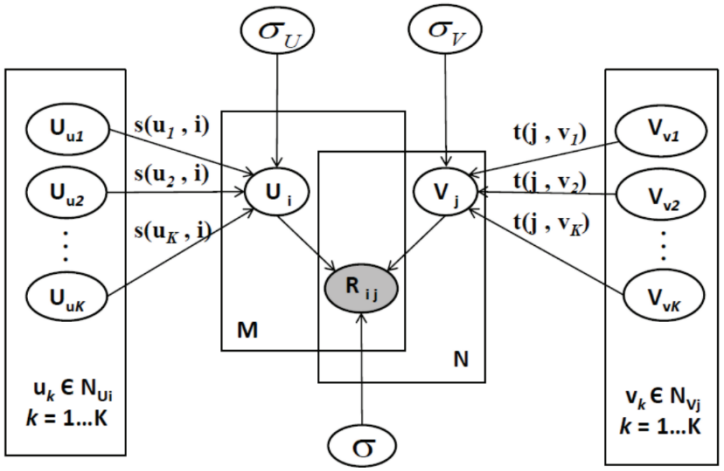
- Mean(X) = 0, Mean(Y) = 4
- Correlation= ？
  - = (-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+3(5)= 0

数据预处理：数据集成

- 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

数据预处理：数据集成

- 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)

数据预处理：数据集成

数据的相关性分析

- **无序数据：每个数据样本的不同维度是没有顺序关系的**
  - 余弦相似度、相关度、欧几里得距离、Jaccard
- **有序数据：对应的不同维度(如特征)是有顺序(rank)要求的**
  - 在信息检索中，如何判断不同检索方法返回的页面序列的优劣
  - 在推荐系统中，如何判断不同推荐序列的好坏
    - **Spearman Rank(斯皮尔曼等级)相关系数**
    - **归一化的折损累计增益(NDCG)**
    - 肯德尔相关性系数
      - kendall correlation coefficient
- 课外阅读：PageRank算法

i	相关度
1	3
2	3
3	2
4	0
5	1
6	2

方法返回结果

i	相关度
1	3
2	3
3	2
4	2
5	1
6	0

真实结果

数据预处理：数据集成

数据的相关性分析—举例

- 已知：6个网页的相关度是3，2，3，0，1，2，所以在信息检索中，最好的返回结果应当如(a)所示。
- 如果我们设计了两个检索算法，返回结果分别是(b)和(c)，请问哪个方法的结果与真实结果更相似？

i	相关度
1	3
2	3
3	2
4	2
5	1
6	0

(a)真实结果

i	相关度
1	3
2	3
3	0
4	2
5	2
6	1

(b)方法1返回结果

i	相关度
1	3
2	3
3	2
4	0
5	2
6	1

(c)方法2返回结果

数据预处理：数据集成

- 有序数据的距离度量(信息检索、推荐系统等)
  - Spearman Rank(斯皮尔曼等级)相关系数
    - 比较两组变量的相关程度
    - 当关系是非线性时，它是两个变量之间关系评价的更好指标

$$\rho_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

    - ρ_s：表示斯皮尔曼相关系数
    - d_i^2：表示每一对样本之间等级的差
    - n：表示样本容量
  - ρ_s的范围：-1 to 1 (正相关:$\rho_s > 0$，负相关:$\rho_s < 0$，不相关:$\rho_s = 0$)

数据预处理：数据集成

- 有序数据的距离度量(信息检索、推荐系统等)
  - **Spearman Rank(斯皮尔曼等级)相关系数**
  - X = (a, b, c, d, e, f)
  - Y = (c, a, e, d, f, b)

$$d_i = Y_i - X_i$$

  - $d_i^2 = (4, 1, 4, 0, 1, 16)$

$$\rho_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

  - $\rho = 1 - \frac{6(26)}{6(36-1)} \approx 1 - 0.743 = 0.257$

数据预处理：数据集成

数据的相关性分析—**课后思考**

- Spearman Rank相关度与Pearson相关度之间的联系与区别？

$$\rho_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

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

数据预处理：数据集成

数据的相关性分析—课后思考

- Spearman Rank相关度与Pearson相关度之间的联系与区别？

$$\rho_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \qquad \rho_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

$$\rho_s = \frac{\sum_{i=1}^{N}(R_i-\bar{R})(S_i-\bar{S})}{[\sum_{i=1}^{N}(R_i-\bar{R})^2 \sum_{i=1}^{N}(S_i-\bar{S})^2]^{\frac{1}{2}}} \qquad p_{X,Y} = \frac{\sum_{i=1}^{n}(X_i-\tilde{X})(Y_i-\tilde{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\tilde{X})^2 \sum_{i=1}^{n}(Y_i-\tilde{Y})^2}}$$

斯皮尔曼相关系数被定义成等级数据变量(rank/order variables)之间的皮尔逊相关系数

数据预处理：数据集成

$$\rho_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

数据的相关性分析——练习题2 （计算Spearman）

- 已知6个网页的相关度是3，2，3，0，1，2，所以在信息检索中，最好的返回结果应当如(a)所示。如果我们设计了两个检索算法，返回结果分别是(b)和(c)，
- 请问：哪个方法的结果与真实结果更相似？

i	相关度
1	3
2	3
3	2
4	2
5	1
6	0

(a)真实结果

i	相关度
1	3
2	3
3	0
4	2
5	2
6	1

(b)方法1返回结果

i	相关度
1	3
2	3
3	2
4	0
5	2
6	1

(c)方法2返回结果

只考虑了每个位置的数据与真实数据的顺序差异，但是没有考虑到不同位置(position)的重要性差异

数据预处理：数据集成

- **NDCG：有序数据的度量**（信息检索、推荐系统等）
  Normalized Discounted cumulative gain
  - G(增益)：每个结果的相关度
  - CG(累计增益)：所有（k）结果的累计增益 $$CG@K = \sum_{i=1}^{K} rel_i$$
  - DCG（折损累计增益）：引入折损因子，对每一个结果相关性进行位置折损后累计。排名靠前的结果更重要！

$$DCG@K = \sum_{i=1}^{K} \frac{rel_i}{\log_2(i+1)} \qquad DCG@K = \sum_{i=1}^{K} \frac{2^{rel-i}-1}{\log_2(i+1)}$$

  - i是结果的位置，i越大，表示该结果的结果列表排名越靠后，结果列表越差，DCG越小

Qi Liu, Yong Ge, Enhong Chen, and Hui Xiong. Personalized Travel Package Recommendation. ICDM'2011, **(Best Research Paper Award)**

数据预处理：数据集成

- **NDCG：有序数据的度量**（信息检索、推荐系统等）
  Normalized Discounted cumulative gain
  - 不同结果列表的数量不一致（如，由于搜索结果随着检索词的不同，返回的数量不一致），需要标准化(Normalized)处理:

$$NDCG@K = \frac{\mathrm{DCG@K}}{\mathrm{IDCG@K}}$$

  - IDCG为理想（ideal）情况下最大的DCG值，即最好结果列表的DCG分数
    - 某一用户返回的最好推荐结果列表
    - 真实的数据序列

数据预处理：数据集成

$$NDCG@K = \frac{DCG@K}{IDCG@K}$$

- 例，假设一个推荐系统为用户推荐了3部电影，顺序为A，B，C，用户实际对这三部电影的偏好为B > A > C, 假定A，B，C三部电影的相关性分数分别为2，3，1，那么对于系统返回的结果有：
  - CG@3 = 2 + 3 + 1= 6
  - DCG@3 = 3 + 4.42 + 0.5 = 7.92

$$CG_k = \sum_{i=1}^{k} rel_i$$

- 理想情况下，系统给出的电影排序应该为B，A，C
  - IDCG@3 = 7 + 1.89 + 0.5 = 9.39
- 可以计算NDCG@3
  - NDCG@3 = 7.92 / 9.39 = 0.84

$$DCG@K = \sum_{i=1}^{K} \frac{2^{rel-i} - 1}{\log_2(i+1)}$$

i	movie	rel	$\frac{2^{rel_i} - 1}{log_2(i+1)}$
1	A	2	3
2	B	3	4.42
3	C	1	0.5

方法返回结果

i	movie	rel	$\frac{2^{rel_i} - 1}{log_2(i+1)}$
1	B	3	7
2	A	2	1.89
3	C	1	0.5

真实结果

数据预处理：数据集成

$$NDCG_k = \frac{DCG_k}{IDCG_k}$$

- 例，假设搜索返回的6个物品，其相关性分别是 3、2、3、0、1、2
  - CG@6 = 3+2+3+0+1+2
  - DCG@6 = 7+1.89+3.5+0+0.39+1.07 = 13.85
- 假如用户真实选择了8个物品，除了上面的6个，还有2个物品，第7个相关性为3，第8个相关性为0。那么在理想情况下的相关性分数排序应该是
  - 3、3、3、2、2、1、0、0。
- 计算IDCG@6：
  - IDCG = 7+4.42+3.5+1.29+1.16+0.36 = 17.73
- 可以计算NDCG@6：
  - NDCG@6 = 13.85/17.73 = 0.78

$$CG_k = \sum_{i=1}^{k} rel_i \qquad DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i}-1}{log_2(i+1)}$$

i	rel
1	3
2	2
3	3
4	0
5	1
6	2

方法返回结果

i	rel
1	3
2	3
3	3
4	2
5	2
6	1

真实结果

课堂练习：数据集成

$$CG_k = \sum_{i=1}^{k} rel_i$$

$$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i}-1}{log_2(i+1)}$$

$$NDCG_k = \frac{DCG_k}{IDCG_k}$$

数据相关性分析——练习题3

- 已知6个网页的相关度是3，2，3，0，1，2，所以在信息检索中，最好的返回结果应当如(a)所示。如果我们设计了两个检索算法，它们的返回结果分别是(b)和(c)，请问哪个方法的结果与真实结果更相似（根据**NDCG的计算结果**）。

i	相关度
1	3
2	3
3	2
4	2
5	1
6	0

(a)真实结果

i	相关度
1	3
2	3
3	0
4	2
5	2
6	1

(b)方法1返回结果

i	相关度
1	3
2	3
3	2
4	0
5	2
6	1

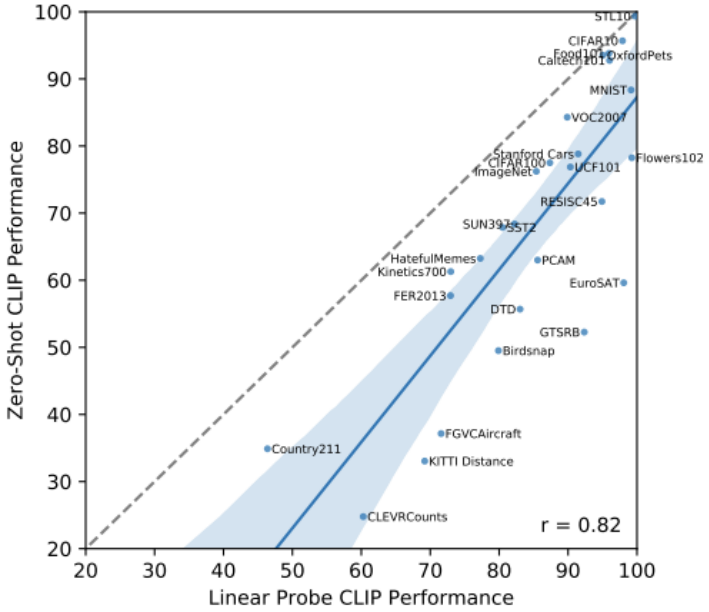
(c)方法2返回结果

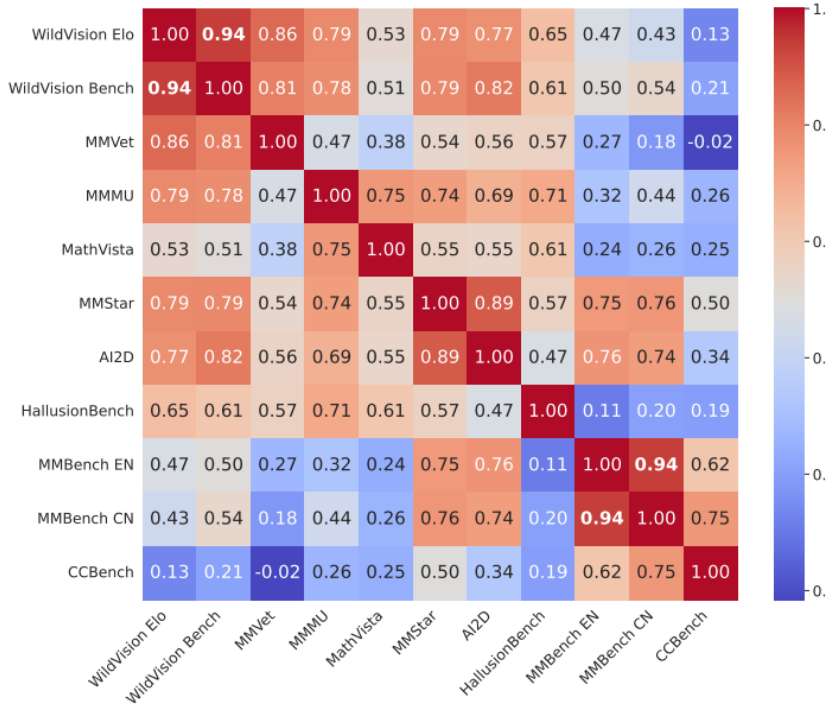
可以只列出计算公式，不用给出计算结果

- **0.9746**
- **0.9889**

有序数据相关性

- 相关性分析研究案例：
  - 相关系数：量化变量间相关性的方向与强度
  - 散点图：直观呈现变量间的关系、趋势及异常值
  - 热力图：展示多变量间的相关结构与强弱差异





课堂练习：数据集成

- NDCG的用途相当广泛
  - 两个列表的相关性
  - 冗余数据的相关性
  - 搜索引擎：评价搜索结果的优劣
  - 推荐系统：评价推荐结果的好坏
  - 大模型Agent：工具调用的好坏
  - 。 。 。



数据预处理：数据集成

课后阅读

- Defu Lian, Haoyu Wang, Enhong Chen, Xing Xie. LightRec: a Memory and Search-Efficient Recommender System. WWW 2020.
- Qi Liu, Zhenya Huang, Enhong Chen,, EKT: Exercise-aware Knowledge Tracing for Student Performance Prediction, TKDE
- Zhenya Huang, Qi Liu, Enhong Chen, et al, Question Difficulty Prediction for READING Problems in Standard Tests, AAAI'2017
- Qi Liu, Yong Ge, Enhong Chen, and Hui Xiong. Personalized Travel Package Recommendation. ICDM'2011, (Best Research Paper Award)
- PageRank算法