



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

1

数据科学导论

Introduction to Data Science

第二章 数据分析基础

黄振亚，陈恩红

Email: huangzhy@ustc.edu.cn, cheneh@ustc.edu.cn

课程主页:

<http://staff.ustc.edu.cn/~huangzhy/Course/DS2023.html>



数据预处理

28

- 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

29

□ 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

30

□ 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:

31

检测冗余属性

- 分析属性之间的相关性
- 相关性分析检测冗余

字段	说明	示例
ID_LAT_LON_YEAR_WEEK	地点、时间	ID_-0.510_29.290_2019_00
year	年份	2019
latitude	维度	-0.51
...	...	...

输入
78种字段

$$r_{A,B} = \frac{\sum (A - \bar{A})(B - \bar{B})}{(n-1)\sigma_A\sigma_B}$$

Pearson积矩相关系数，取值范围为 [-1; 1]

➤值大于 0，则属性 A 和 B 是正相关的，值越大相关性越强

因此，表明两个属性中有一个可以作为冗余删除

➤值为 0，则 A 和 B 是独立的，它们不存在相关性

➤值小于 0，则 A 和 B 是负相关的。

卡方检验：值越大，两个变量相关的可能性越大

$$\chi^2 = \sum_{i=1}^c \sum_{j=1}^r \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

卡方检验：*o_{ij}* 是联合事件 (*A_i*; *B_j*) 的观测频度（即实际计数），而 *e_{ij}* 是 (*A_i*; *B_j*) 的期望频度。卡方检验的原假设是 A 和 B 两个属性相互独立，如果可以拒绝该原假设，则我们说 A 和 B 是显著相关的。



数据预处理：数据集成

32

□ 检测冗余样本

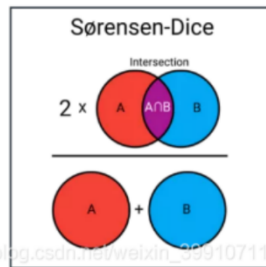
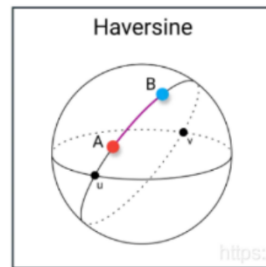
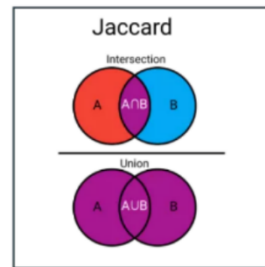
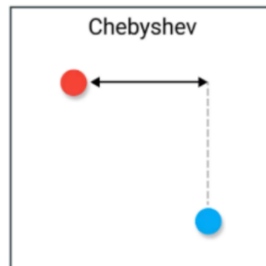
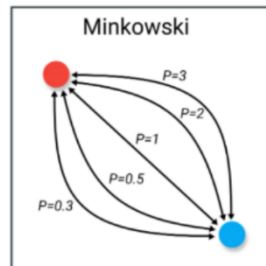
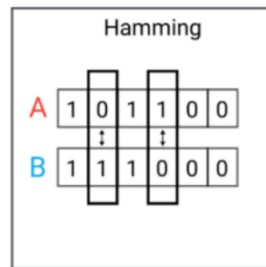
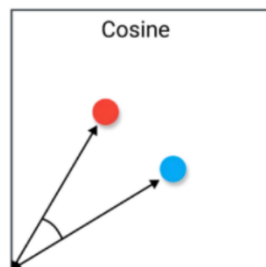
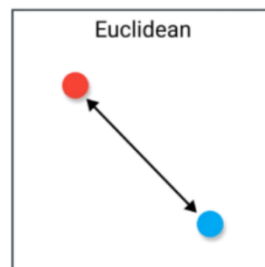
□ 思想：数据样本之间的相关性，数据融合、去除冗余

□ 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
-





数据预处理：数据集成

33

数据的距离度量

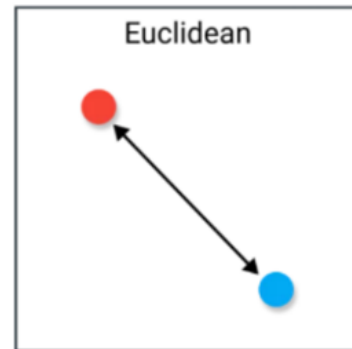
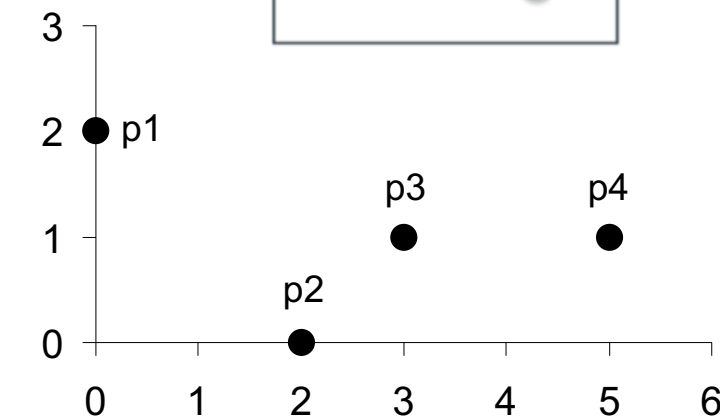
欧几里得距离(Euclidean Distance)

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

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

34

□ 数据的距离度量

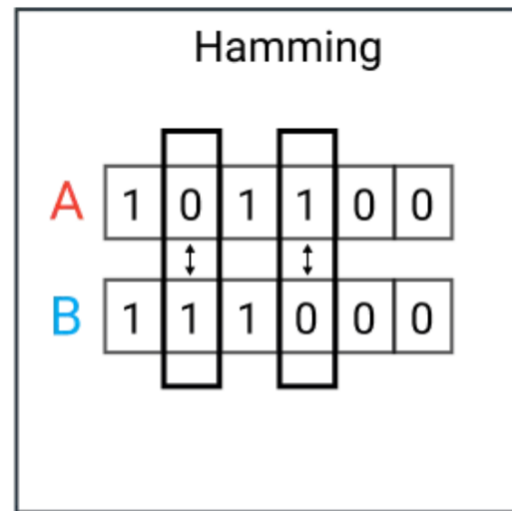
- 汉明距离(Hamming Distance)

- 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**.



数据预处理：数据集成

35

□ 数据的距离度量

□ 明氏距离（Minkowski Distance）

■ 距离度量：通用表达形式

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

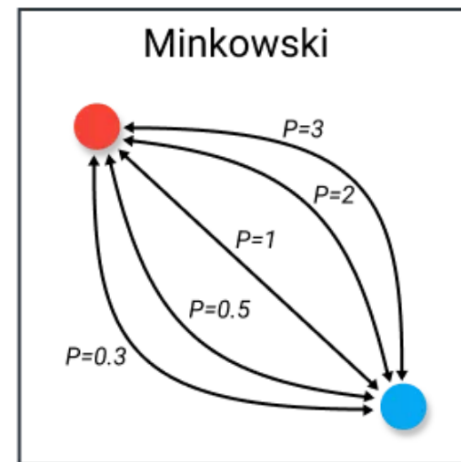
r 是参数

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

□ $r=\infty$ ：切比雪夫距离





数据预处理：数据集成

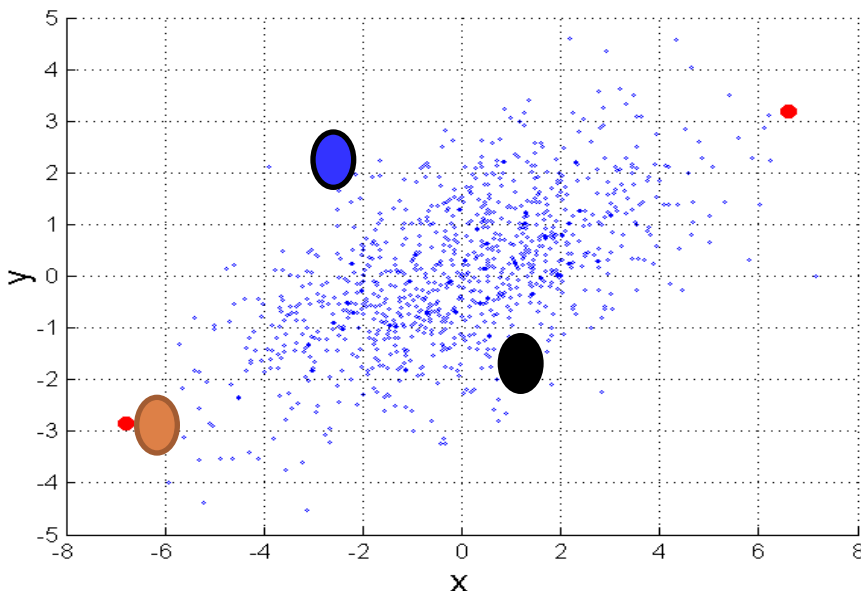
36

□ 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

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

- 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

37

□ 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二维向量。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

38

□ 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

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

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

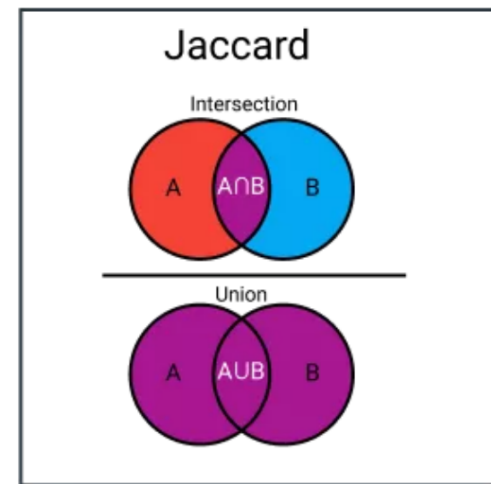
■ F01: p为0、q为1的属性数量

■ 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

$$\begin{aligned} \text{SMC} &= \text{number of matches} / \text{number of attributes} \\ &= (F11 + F00) / (F01 + F10 + F11 + F00) \end{aligned}$$



$$\begin{aligned} \text{Jaccard} &= \text{number of F11 matches} / \text{number of non-zero attributes} \\ &= (F11) / (F01 + F10 + F11) \end{aligned}$$



数据预处理：数据集成

39

□ 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?

$F_{01} = 2$ (p为0, q为1的属性数量)

$F_{10} = 1$ (p为1, q为0的属性数量)

$F_{00} = 7$ (p为0, q为0的属性数量)

$F_{11} = 0$ (p为1, q为1的属性数量)

$$SMC = (F_{11} + F_{00}) / (F_{01} + F_{10} + F_{11} + F_{00})$$

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

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



数据预处理：数据集成

40

□ 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

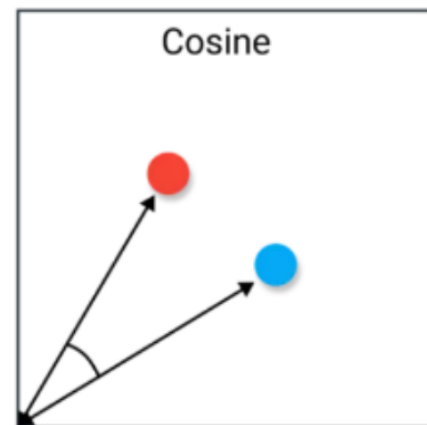
B = 1 0 0 0 0 0 0 1 0 2

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

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

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

$$\cos(A, B) = 0.3150$$



思考：余弦相似度是不是一种距离？



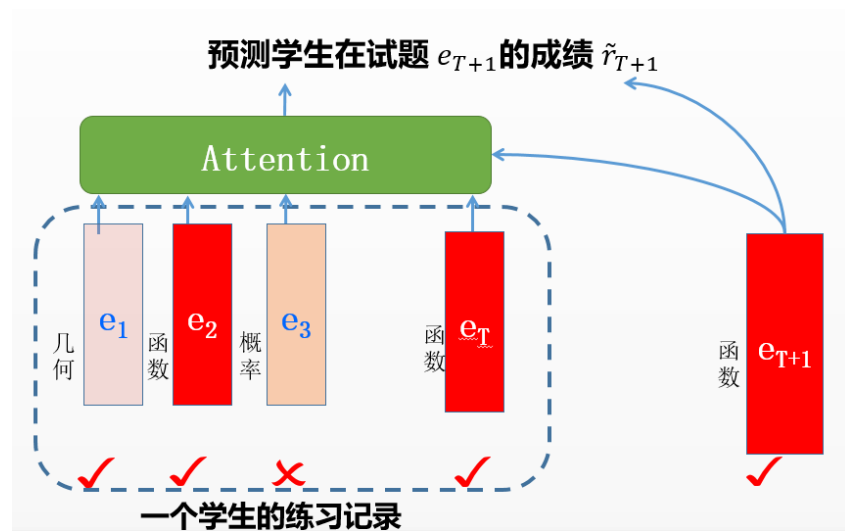
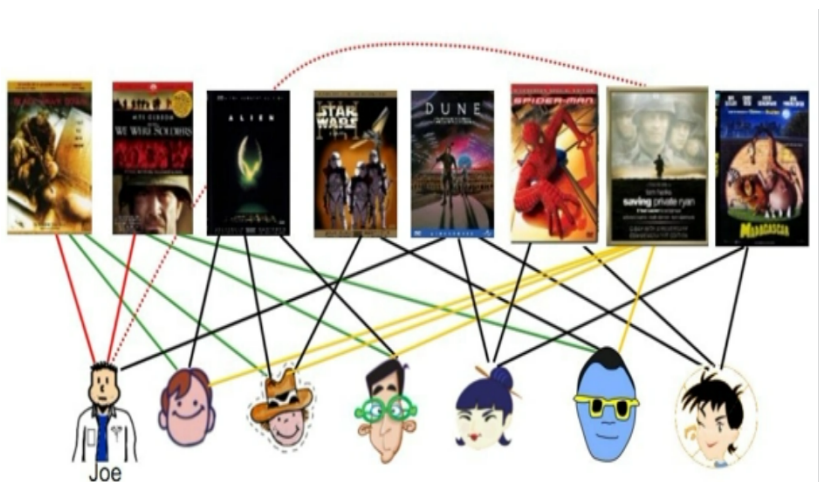
数据预处理：数据集成

41

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





数据预处理：数据集成

42

□ 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)

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

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



数据预处理：数据集成

43

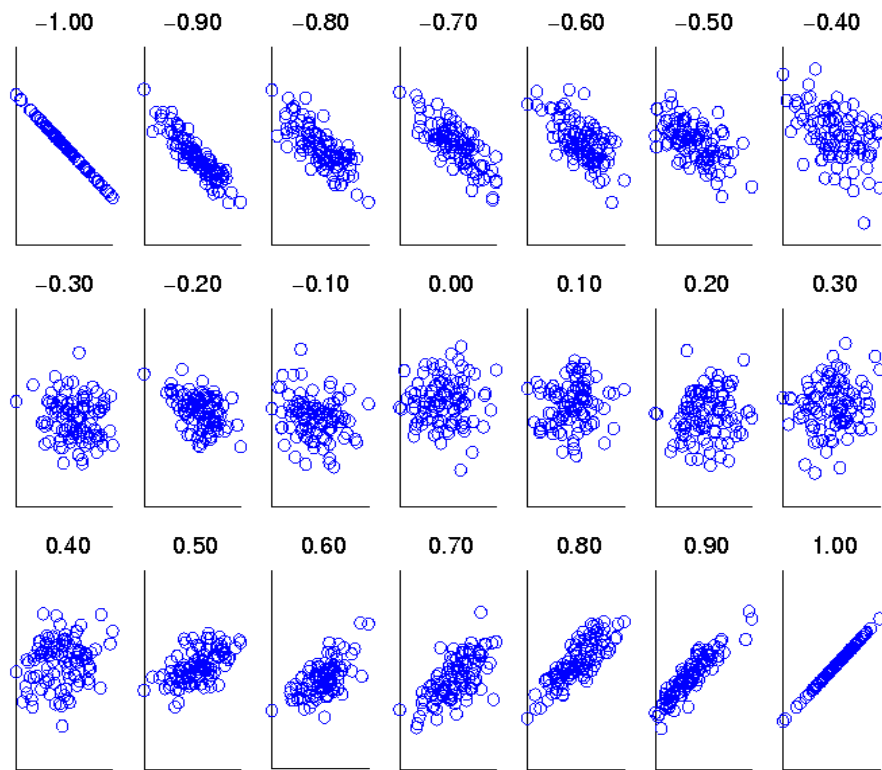
□ 数据的相关性

□ 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 $[-1, 1]$

两个数据样本 x, y 各有30个属性，这些属性值随机产生，使得 x 和 y 的相关度从-1到1。

图中每个小圆圈代表30个属性中的一个，其 x 坐标是 x 的一个属性的值， y 坐标是 y 的相同属性的值





数据预处理：数据集成

44

□ 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？

□ $X = (-3, -2, -1, 0, 1, 2, 3)$

□ $Y = (9, 4, 1, 0, 1, 4, 9)$

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

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

□ 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

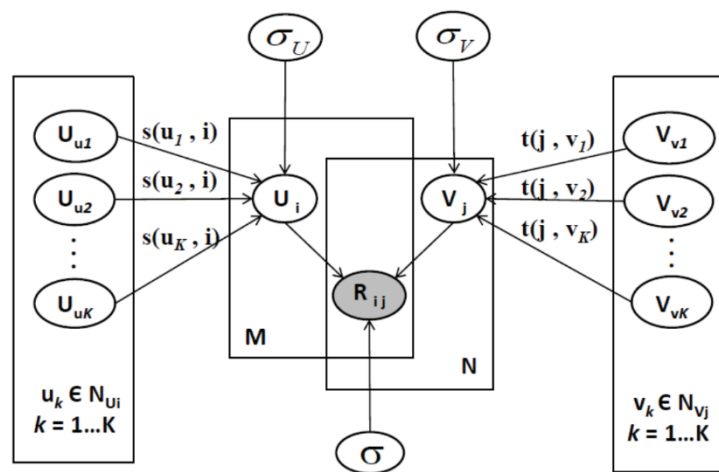
45

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)

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

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





数据预处理：数据集成

47

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

49

□ 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

50

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

52

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？

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



$$\rho_s = \frac{\sum_{i=1}^N (R_i - \bar{R})(S_i - \bar{S})}{[\sum_{i=1}^N (R_i - \bar{R})^2 \sum_{i=1}^N (S_i - \bar{S})^2]^{\frac{1}{2}}}$$



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

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

54

□ 有序数据的距离度量(信息检索、推荐系统等)

□ NDCG(Normalized Discounted cumulative gain)

- **CG(累计增益)**: 只考虑到了相关性的关联程度, 没有考虑每个推荐结果处于**不同位置**对整个推荐效果的影响

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

rel_i 表示处于位置 i 的推荐结果的相关性

- **DCG(折损累计增益)**: 就是在每一个CG的结果上处以一个折损值, 目的就是为了让排名越靠前的结果越能影响最后的结果

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

- i 表示推荐结果的位置, i 越大, 则推荐结果在推荐列表中排名越靠后 推荐效果越差, DCG越小



数据预处理：数据集成

55

□ 有序数据的距离度量(信息检索、推荐系统等)

□ NDCG(Normalized Discounted cumulative gain)

- **NDCG**: 由于搜索结果随着检索词的不同, 返回的数量不一致, 而 DCG 是一个累加的值, 没法针对两个不同的搜索结果进行比较, 因此需要**标准化**处理, 这里是除以IDCG:

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

IDCG为理想 (ideal) 情况下最大的DCG值, 指推荐系统为某一用户返回的最好推荐结果列表(或者, 真实的数据序列)



数据预处理：数据集成

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

56

- 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
 - $CG@6 = 3+2+3+0+1+2$
 - $DCG@6 = 7+1.89+3.5+0+0.39+1.07 = 13.85$
- 假如我们实际召回了8个物品，除了上面的6个，还有两个物品，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:
 - $NDCG@6 = 13.85/17.73 = 0.78$

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

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**



数据预处理：数据集成

58

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)
- 信息检索经典研究：PageRank算法