

2022-2023秋季课程:数据科学与大数据导论

Introduction to Data Science and Big data


Chapter 3: Big Data Analytics Fundamentals

曹劲舟 助理教授

深圳技术大学 大数据与互联网学院

caojinzhou@sztu.edu.cn

2022年9月



Outline

□ Data Preprocessing 数据预处理 扩展学习

相似度函数：编辑距离

□编辑距离Edit Distance计算 – 动态规划算法

定义 $D(i,j)$ 为从字符串 $s1..si$ 到 $t1..tj$ 最少的编辑操作次数

$$= \min \begin{cases} D(i-1,j-1) + d(s_i,t_j) & //substitution/copy \quad \text{左上} \\ D(i-1,j)+1 & //insert \quad \text{上} \\ D(i,j-1)+1 & //delete \quad \text{左} \end{cases}$$

(其中 $d(c,d)=0$ 如果 $c=d$, 否则等于1)

另外初始化 $D(i,0)=i$ 以及 $D(0,j)=j$

相似度函数：编辑距离

□编辑距离Edit Distance计算 – 动态规划算法

■看一个实例

■假设有字符串s1为jary，和字符串s2为jerry，现在求s1和s2的编辑距离，也就是把s2转换为s1的最少编辑操作步数

■首先，我们建立如下的矩阵，并且初始化该矩阵

		j	a	r	y
	0	1	2	3	4
j	1				
e	2				
r	3				
r	4				
y	5				

■从源串的第一个字符(“j”)开始，从上至下与目标串进行对比

相似度函数：编辑距离

□编辑距离Edit Distance计算 - 动态规划算法

■Min (左上角+0或者1, 上+1, 左+1)

■比如, 第一次, 源串第一个字符“j”与目标串的“j”对比, 左+1、上+1、左上+0或者1三个值中取出最小的值0, 因为两字符相等, 所以填上0

■接着, 依次对比“j”→“e”、“j”→“r”、“j”→“r”、“j”→“y”等进行处理, 直到扫描完目标串, 得到的结果如下

		j	a	r	y
	0	1	2	3	4
j	1	0			
e	2	1			
r	3	2			
r	4	3			
y	5	4			

相似度函数：编辑距离

编辑距离Edit Distance计算 - 动态规划算法

按照上面的方法，遍历整个源串的各个字符，与目标串的各个字符对比，填写各个单元格，各个单元格的变化如下表所示

		j	a	r	y
	0	1	2	3	4
j	1	0	1		
e	2	1	1		
r	3	2	2		
r	4	3	3		
y	5	4	4		
		j	a	r	y
	0	1	2	3	4
j	1	0	1	2	
e	2	1	1	2	
r	3	2	2	1	
r	4	3	3	2	
y	5	4	4	3	
		j	a	r	y
	0	1	2	3	4
j	1	0	1	2	3
e	2	1	1	2	3
r	3	2	2	1	2
r	4	3	3	2	2
y	5	4	4	3	2

相似度函数：编辑距离

- 处理完最后一列，则最后一列的最后一个值，为最短编辑距离
 - 即jary和jerry的编辑距离为2
 - 也就是， jary插入r得到jarry，把a改成e得到jerry

		j	a	r	y
	0	1	2	3	4
j	1	0	1		
e	2	1	1		
r	3	2	2		
r	4	3	3		
y	5	4	4		
		j	a	r	y
	0	1	2	3	4
j	1	0	1	2	
e	2	1	1	2	
r	3	2	2	1	
r	4	3	3	2	
y	5	4	4	3	
		j	a	r	y
	0	1	2	3	4
j	1	0	1	2	3
e	2	1	1	2	3
r	3	2	2	1	2
r	4	3	3	2	2
y	5	4	4	3	2

参考如下箭头

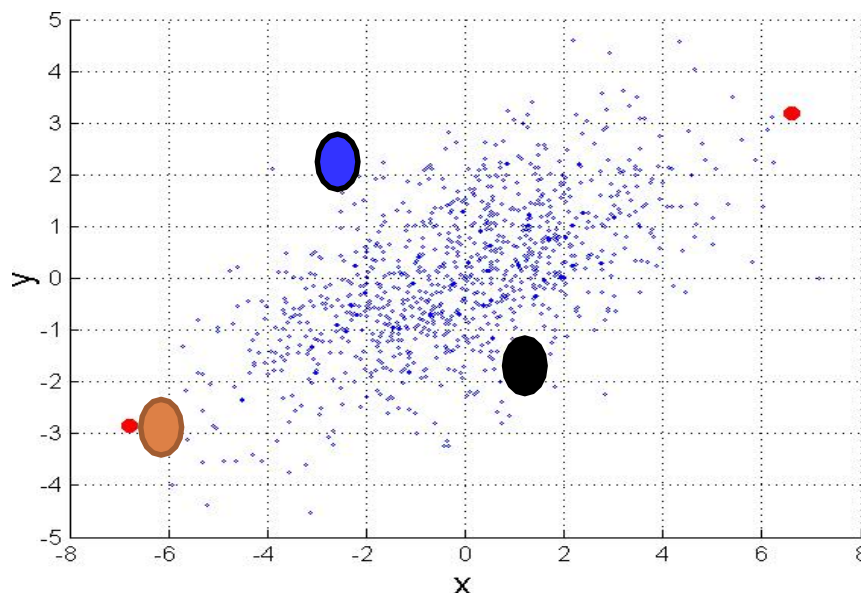
		j	a	r	y
	0	1	2	3	4
j	1	0	1	2	3
e	2	1	1	2	3
r	3	2	2	1	2
r	4	3	3	2	2
y	5	4	4	3	2

马氏距离

■ 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）

$$\text{mahalanobis}(p, q) = (p - q)^T \Sigma^{-1} (p - q)$$



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

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。**

Jaccard相关系数

■ 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

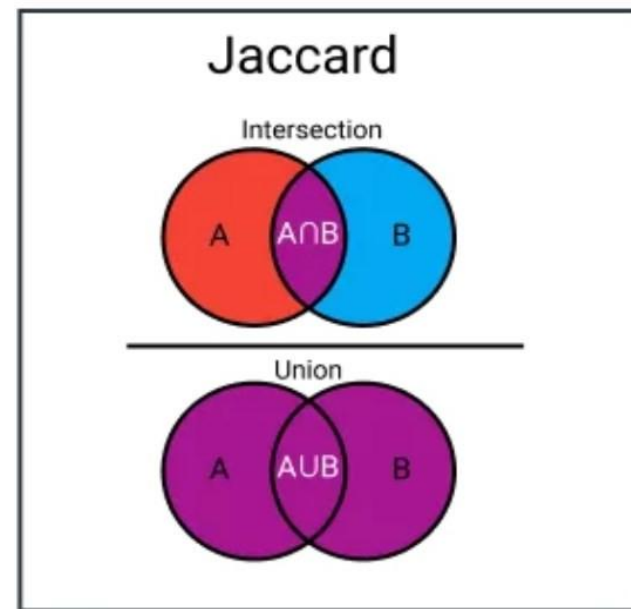
- 例：数据p和q，定义如下4个变量
- F01: p为0, q为1的属性数量
- F10: p为1, q为0的属性数量
- F00: p为0, q为0的属性数量
- F11: p为1, q为1的属性数量

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

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

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

$$\begin{aligned} \text{Jaccard} &= \text{number of 11 matches} / \text{number of non-zero attributes} \\ &= (F11) / (F01 + F10 + F11) \end{aligned}$$



Jaccard相关系数

■简单匹配Simple Matching VS Jaccard相关系数

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

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

p和q是否相关?

$$F_{01} = 2 \quad (\text{p为0, q为1的属性数量})$$

$$F_{10} = 1 \quad (\text{p为1, q为0的属性数量})$$

$$F_{00} = 7 \quad (\text{p为0, q为0的属性数量})$$

$$F_{11} = 0 \quad (\text{p为1, q为1的属性数量})$$

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

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

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

数据预处理：数据集成

- 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、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

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

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

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

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

有序数据相关性分析

□ 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

有序数据相关性分析

□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)

有序数据相关性分析

□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:

- $IDCG = 7+4.42+3.5+1.29+1.16+0.36 = 17.73$

□计算NDCG@6:

- $NDCG@6 = 13.85/17.73 = 0.78$

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

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

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

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

□ 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)。

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

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

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

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

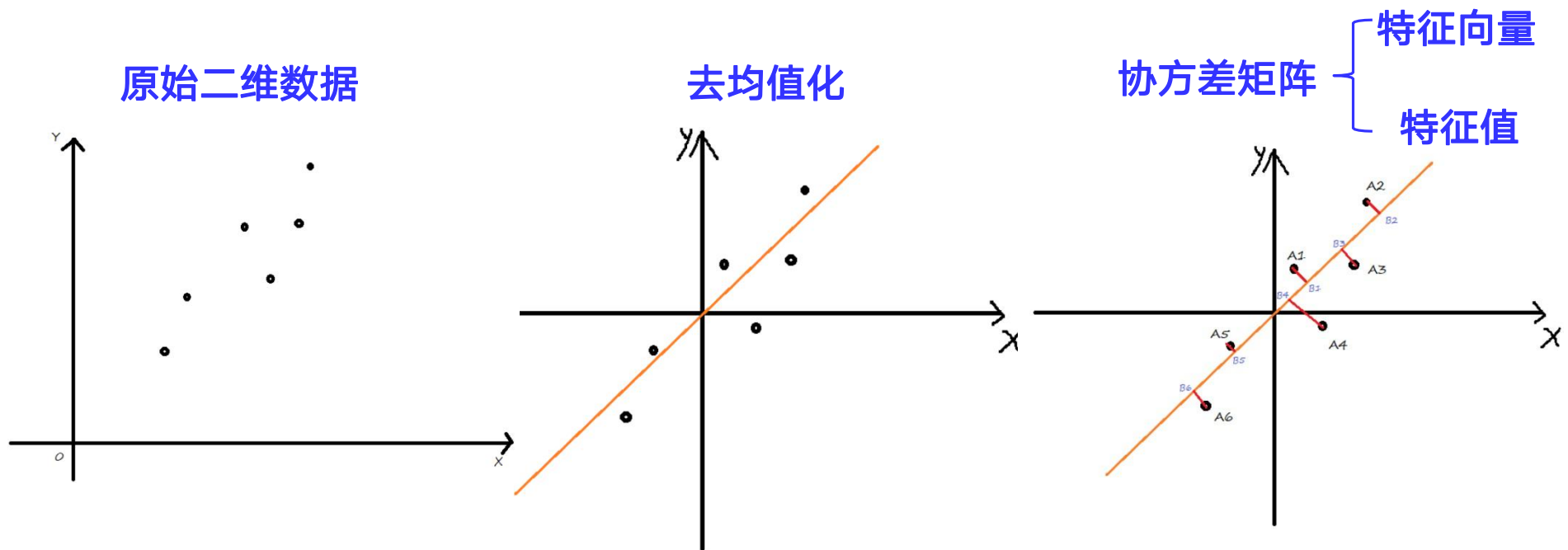
➤ 0.9746

➤ 0.9889

主成分分析PCA

■例：如何把二维数据降维至一维

- 思想：方差越大，数据间的差异越大。让新数据的方差尽可能地大，使得新数据尽可能地不丢失原有数据的信息



课后思考：一元线性回归与PCA的关系？

主成分分析PCA

□主成分分析(principal component analysis, PCA)

■原数据为X，降维后新数据为Y

- 不同维度相互独立：要求数据Y的协方差矩阵为对角阵 $B = \frac{1}{m}Y^T Y \in \mathbb{R}^{k \times k}$

$$B = \frac{1}{m}Y^T Y = \frac{1}{m}(XP)^T (XP) = \frac{1}{m}P^T X^T X P = P^T C P \quad C = \frac{1}{m}X^T X \in \mathbb{R}^{n \times n}$$

- 最大化每个维度内的样本方差：Y的协方差矩阵中的对角元素越大越好
 - 对C 进行特征值分解，将求解得到的特征值从大到小依次作为协方差矩阵的对角线元素，特征向量组成变换矩阵 (P)
- 保留最大的k 个特征值：第K个主成分就是第K大的特征值对应的特征向量

■存储空间：

- 对于原始的N*M（N维M个样本）的数据，原始存储空间是N*M
- PCA以后为：K*M（M个K维样本）+N*K(K个特征向量)

主成分分析PCA

□主成分分析(principal component analysis, PCA)

Algorithm 5 PCA 算法

Input: 原始的样本矩阵 $X \in \mathbb{R}^{m \times n}$

Output: 压缩后的样本矩阵 $Y \in \mathbb{R}^{m \times k}$

- 1: 对样本矩阵进行去均值化 $\mathbf{x}_i \leftarrow \mathbf{x}_i - \frac{1}{m} \sum_{i=1}^m \mathbf{x}_i, \forall i \in \{1, 2, \dots, m\}$
 - 2: 计算协方差矩阵 $C = \frac{1}{m} X^\top X$
 - 3: 通过特征值分解求解 C 的特征值和特征向量
 - 4: 将特征值从大到小排序, 取最大的 k 个特征值对应的特征向量作为列向量构成变换矩阵 $P \in \mathbb{R}^{n \times k}$
 - 5: 将原始数据转换到新的空间中 $Y = XP$
-

□ 不足之处

- 当原始数据的维度 n 特别大的时候, 计算协方差时的 $X^\top X$ 已经具有相当大的计算量
- 针对协方差矩阵 C 的特征值求解过程计算效率不高

主成分分析PCA

样本矩阵(10个城市样本, 8个属性)的转置 X^T

省份	GDP X_1	居民消 费水平 X_2	固定资 产投资 X_3	职工平 均工资 X_4	货物周 转 量 X_5	居民消费 价格指数 X_6	商品零售 价格指数 X_7	工业总 产 值 X_8
北京	1394.89	2505	519.01	8144	373.9	117.3	112.6	843.43
天津	920.11	2720	345.46	6501	342.8	115.2	110.6	582.51
河北	2849.52	1258	704.87	4839	2033.3	115.2	115.8	1234.85
山西	1092.48	1250	290.9	4721	717.3	116.9	115.6	697.25
内蒙	832.88	1387	250.23	4134	781.7	117.5	116.8	419.39
辽宁	2793.37	2397	387.99	4911	1371.1	116.1	114	1840.55
吉林	1129.2	1872	320.45	4430	497.4	115.2	114.2	762.47
黑龙江	2014.53	2334	435.73	4145	824.8	116.1	114.3	1240.37
上海	2462.57	5343	996.48	9279	207.4	118.7	113	1642.95
江苏	5155.25	1926	1434.95	5943	1025.5	115.8	114.3	2026.64

数据归一化并得
到协方差矩阵

三个主成分 (8*3)

第一特征向量 a_1	第二特征向量 a_2	第三特征向量 a_3
0.470641	0.107995	0.19241
0.456708	0.258512	0.109819
0.424712	0.287536	0.19241
-0.31944	0.400931	0.397525
0.312729	-0.40431	0.24505
0.250802	0.498801	-0.24777
0.240481	-0.48868	0.332179
-0.26267	0.167392	0.723351

特征值
分解

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	1.000	.267	.951	.191	.617	-.274	-.264	.874
X_2	.267	1.000	.426	.718	-.151	-.234	-.593	.363
X_3	.951	.426	1.000	.400	.431	-.282	-.359	.792
X_4	.191	.718	.400	1.000	-.356	-.134	-.539	.104
X_5	.617	-.151	.431	-.356	1.000	-.255	.022	.659
X_6	-.274	-.234	-.282	-.134	-.255	1.000	.760	-.126
X_7	-.264	-.593	-.359	-.539	.022	.760	1.000	-.192
X_8	.874	.363	.792	-.104	.659	-.126	-.192	1.000

特征子集选择

□ 维度规约

■ 属性子集选择（特征子集）

- 做法：删除不相关或冗余的属性来减少维度与数据量
- 目标：找到最小属性集，使得数据的概率分布尽可能接近使用所有属性得到的原分布
- 理解：从全部属性中选取一个特征属性子集，使构造出来的模型更好

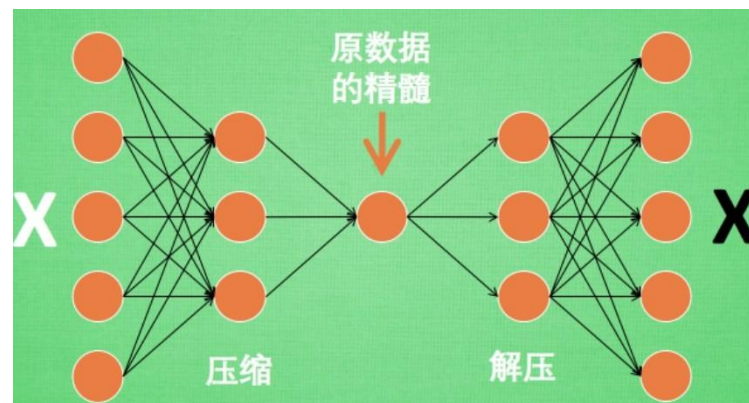
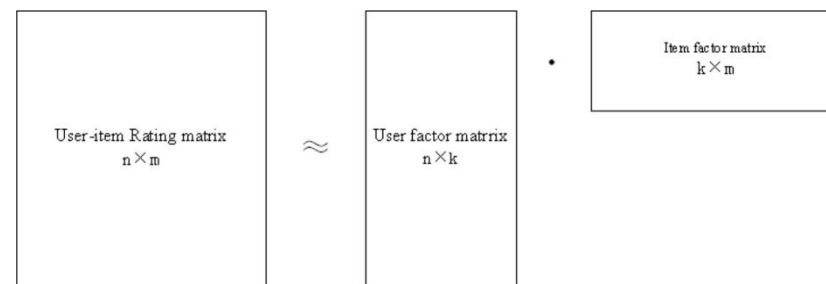
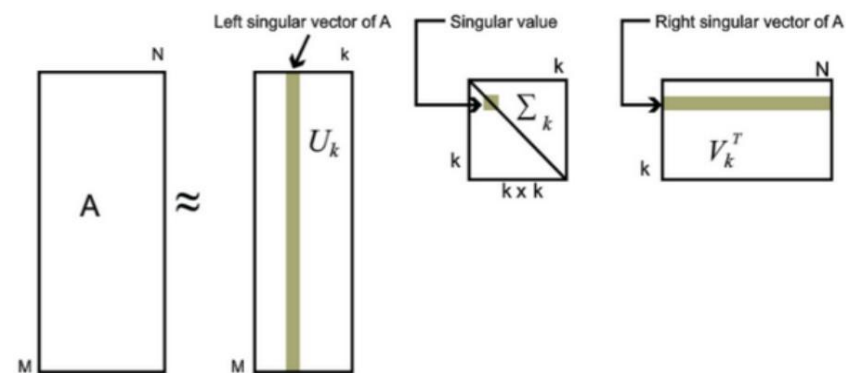
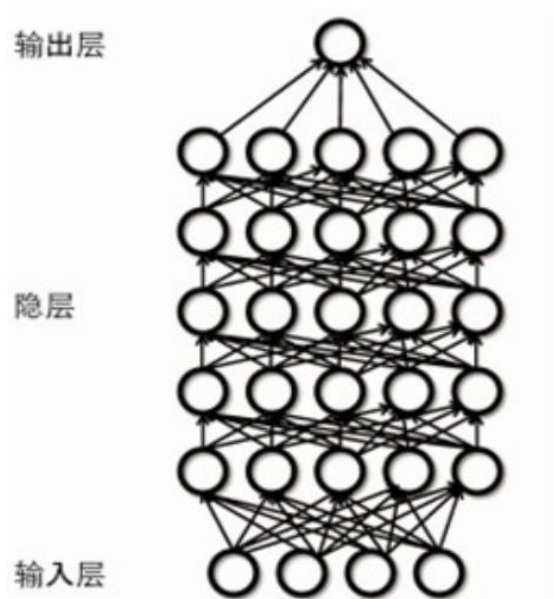
■ 启发式步骤

- 建立子集集合
- 构造评价函数
- 构建停止准则
- 验证有效性
- 例如：决策树

特征子集选择

□ 常用方法

- 奇异值分解(SVD)
- 矩阵分解(PMF)
- 深度学习(Deep Learning)
 - DNN, AutoEncoder, etc



数据规约：数值规约

□数值规约

■通过选择替代的、较小的数据表示形式来减少数据量

- 参数化方法
 - 使用一个参数模型估计数据，最后只要存储参数即可，不用存储数据（除了可能的离群点）
 - 常用方法
 - 线性回归方法；多元回归；对数线性模型；
- 非参数化方法
 - 不使用模型的方法存储数据
 - 常用方法：直方图，聚类，抽样

资料推荐

- ❑ 数据挖掘导论第2章：数据，人民邮电出版社
- ❑ 数据挖掘原理与算法第2章，清华大学出版社
- ❑ T.C. Redman Data Quality: The Field Guide. January 2001
- ❑ I.T.Jolliffe. Principal Component Analysis. Springer Verlag, 2nd edition,
❑ October 2002.
- ❑ Feature selection algorithms: A survey and experimental evaluation, ICDM 2003
- ❑ Zhenya Huang, Qi Liu, Enhong Chen, Learning or Forgetting? A Dynamic Approach for
- ❑ Tracking the Knowledge Proficiency of Students, ACM TOIS
- ❑ Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang,, Neural Cognitive Diagnosis for
Intelligent Education Systems, AAI'2020