

Big Data Processing

Homework 3

作业

- 完成指定的题目
- 编写报告
- **单人不组队** (本次作业都是书后题目, 不涉及到代码的编写以及程序的部署, 所以不组队)
- 本次作业为 <http://www.mmds.org/> 书中
- 6.1.1(a) & 6.1.5
- 11.1.3
- 11.2.1
- 11.3.2
- 11.4.2(a)

Mmnds 6.1.1(a) & 6.1.5

Exercise 6.1.1: Suppose there are 100 items, numbered 1 to 100, and also 100 baskets, also numbered 1 to 100. Item i is in basket b if and only if i divides b with no remainder. Thus, item 1 is in all the baskets, item 2 is in all fifty of the even-numbered baskets, and so on. Basket 12 consists of items $\{1, 2, 3, 4, 6, 12\}$, since these are all the integers that divide 12. Answer the following questions:

- (a) If the support threshold is 5, which items are frequent?

Exercise 6.1.5: For the data of Exercise 6.1.1, what is the confidence of the following association rules?

- (a) $\{5, 7\} \rightarrow 2$.
 (b) $\{2, 3, 4\} \rightarrow 5$.

11.1.3

Exercise 11.1.3: For any symmetric 3×3 matrix

$$\begin{bmatrix} a - \lambda & b & c \\ b & d - \lambda & e \\ c & e & f - \lambda \end{bmatrix}$$

there is a cubic equation in λ that says the determinant of this matrix is 0. In terms of a through f , find this equation.

11.2.1

Exercise 11.2.1: Let M be the matrix of data points

$$\begin{bmatrix} 1 & 1 \\ 2 & 4 \\ 3 & 9 \\ 4 & 16 \end{bmatrix}$$

- (a) What are $M^T M$ and MM^T ?
- (b) Compute the eigenpairs for $M^T M$.
- ! (c) What do you expect to be the eigenvalues of MM^T ?
- ! (d) Find the eigenvectors of MM^T , using your eigenvalues from part (c).

11.3.2

	Titanic				
	Casablanca				
	Star Wars				
	Alien				
	Matrix				
Joe	1	1	0	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

Figure 11.6: Ratings of movies by users

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 0 & 0 & 2 & 2 \end{bmatrix} = \begin{bmatrix} .14 & 0 \\ .42 & 0 \\ .56 & 0 \\ .70 & 0 \\ 0 & .60 \\ 0 & .75 \\ 0 & .30 \end{bmatrix} \begin{bmatrix} 12.4 & 0 \\ 0 & 9.5 \end{bmatrix} \begin{bmatrix} .58 & .58 & .58 & 0 & 0 \\ 0 & 0 & 0 & .71 & .71 \end{bmatrix}$$

$M \qquad U \qquad \Sigma \qquad V^T$

Figure 11.7: SVD for the matrix M of Fig. 11.6

Exercise 11.3.2: Use the SVD from Fig. 11.7. Suppose Leslie assigns rating 3 to Alien and rating 4 to Titanic, giving us a representation of Leslie in “movie space” of $[0, 3, 0, 0, 4]$. Find the representation of Leslie in concept space. What does that representation predict about how well Leslie would like the other movies appearing in our example data?

11.4.2(a)

	<i>The Matrix</i>	<i>Alien</i>	<i>Star Wars</i>	<i>Casablanca</i>	<i>Titanic</i>
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

Figure 11.12: Matrix M , repeated from Fig. 11.6

! Exercise 11.4.2: Find the CUR-decomposition of the matrix of Fig. 11.12 when we pick two “random” rows and columns as follows:

- (a) The columns for *The Matrix* and *Alien* and the rows for Jim and John.

报告要求

- 使用Word, Pages, LaTeX或者markdown等编写都可以, 但最后提交时转成PDF文件格式。
- (本次作业涉及到数学公式的排版, 建议采用LaTeX编写、配合markdown使用mathjax、使用word自带的公式编辑或mathtype)

提交

- 作业提交位置
 - <ftp://public.sjtu.edu.cn> username: shen_yao password: public
 - 提交到ftp中/upload/CS426/hw3/ 目录下
- 作业提交时间
 - ddl: 5月11号23:59:59
 - 晚交惩罚：每超时24小时，该次作业总分扣除20%成绩，不满24小时按照24小时计算，5月14日23:59:59之后提交的作业一概不接收。
 - 时间根据ftp服务器接收到文件的时间为准。
- 作业命名规则
 - 学号_姓名_hw3.pdf

评分标准（满分10分）

- 6.1.1(a)和6.1.5一共2分
- 剩下每题2分

- (分数的档次只会有0, 0.5, 1.0, 1.5, 2.0以此递推)

完成作业过程遇到任何问题, 请发邮件到gdshen@sjtu.edu.cn