Big Data Processing: homework 3

凌康伟 5140219295

May 10, 2017

Exercise 6.1.1

(a) In this setting, for items numbered beyond 20, there are less than 5 multiples for each number, therefore they are not frequent. The frequent items are: 1, 2, ..., 20;

Exercise 6.1.5

(a) The baskets that contains item 5 and 7 are 35, 70 (common multiples of 5 and 7), thus support($\{5,7\}$) = 2. Since 70 is also a multiple of 2, item 2 is in basket 70, support($\{2,5,7\}$) = 1. Therefore, conf($\{5,7\} \to 2$) = $1/2$.

(b) The baskets containing item 2, 3 and 4 are 12, 24, 36, 48, 60, 72, 84, 96, and among these basket, 60 contains 5. Therefore,

$$\begin{aligned}
\text{support}(\{2,3,4\}) &= 8 \\
\text{support}(\{2,3,4,5\}) &= 1 \\
\text{conf}(\{2,3,4\} \to 5) &= \frac{1}{8}
\end{aligned}$$

Exercise 11.1.3

Let A denote any symmetric 3x3 matrix:

$$\boldsymbol{A} = \begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix}$$

And for $\boldsymbol{A} - \lambda \boldsymbol{I}$:

$$\begin{aligned}
|\boldsymbol{A} - \lambda \boldsymbol{I}| = \begin{vmatrix} a-\lambda & b & c \\ b & d-\lambda & e \\ c & e & f-\lambda \end{vmatrix} &= (a-\lambda)[(d-\lambda)(f-\lambda) - e^2] - b[b(f-\lambda) - ce] + c[be - c(d-\lambda)] \\
&= -\lambda^3 + (a+d+f)\lambda^2 - (df + ad + af - e^2 - b^2 - c^2)\lambda \\
&\quad + adf + 2bce - ae^2 - fb^2 - dc^2
\end{aligned}$$

Therefore the equation can be expressed as:

$$\alpha_3\lambda^3 + \alpha_2\lambda^2 + \alpha_1\lambda + \alpha_0 = 0$$

where

$$\begin{aligned}
\alpha_3 &= -1 \\
\alpha_2 &= a + d + f \\
\alpha_1 &= -(df + ad + af - e^2 - b^2 - c^2) \\
\alpha_0 &= adf + 2bce - ae^2 - fb^2 - dc^2
\end{aligned}$$

Exercise 11.2.1

(a) $\boldsymbol{M}^T\boldsymbol{M} = \begin{bmatrix} 30 & 100 \\ 100 & 354 \end{bmatrix}$, $\boldsymbol{M}\boldsymbol{M}^T = \begin{bmatrix} 2 & 6 & 12 & 20 \\ 6 & 20 & 42 & 72 \\ 12 & 42 & 90 & 156 \\ 20 & 72 & 156 & 272 \end{bmatrix}$

(b) First sovle the equation $|\boldsymbol{M}^T\boldsymbol{M} - \lambda I| = 0$ for eigenvalues:

$$\begin{vmatrix} 30-\lambda & 100 \\ 100 & 354-\lambda \end{vmatrix} = (30-\lambda)(354-\lambda) - 10000 = \lambda^2 - 384\lambda + 620 = 0$$

$\lambda_1 \approx 1.6214, \lambda_2 \approx 382.3786$.

Second find the eigenvectors associated with each eigenvalue:

i. $\lambda_1 = 1.6214$. for the following equation:

$$\begin{bmatrix} 28.3786 & 100 \\ 100 & 352.3786 \end{bmatrix} \boldsymbol{x} = 0$$

a normalized solution: $\boldsymbol{x}_1 = \begin{bmatrix} 0.9620 & 0.2730 \end{bmatrix}^T$

ii. $\lambda_2 = 382.3786$, eigenvector $\boldsymbol{x}_2 = \begin{bmatrix} 0.2730 & 0.9620 \end{bmatrix}^T$

(c) The same eigenvalues as $\boldsymbol{M}^T\boldsymbol{M}$, $\lambda_1 \approx 1.6214, \lambda_2 \approx 382.3786$, and $\lambda_3 = \lambda_4 = 0$

(d)

$$\begin{aligned}
\boldsymbol{x}_1 &= \begin{bmatrix} 0.5425 & 0.6527 & 0.3364 & -0.4079 \end{bmatrix}^T \\
\boldsymbol{x}_2 &= \begin{bmatrix} 0.0631 & 0.2247 & 0.4846 & 0.8429 \end{bmatrix}^T \\
\boldsymbol{x}_3 &= \begin{bmatrix} 0.6882 & -0.6882 & 0.2294 & 0 \end{bmatrix}^T \\
\boldsymbol{x}_4 &= \begin{bmatrix} 0.7960 & -0.5970 & 0 & 0.0995 \end{bmatrix}^T
\end{aligned}$$

Exercise 11.3.2

$$q = [0, 3, 0, 0, 4] \times \boldsymbol{V} = [1.74 \quad 2.84]$$

So Leslie is high in romantic interest, and a little bit interested in sci-fi. Use q to predict Leslie's rating for other movies.

$$r = q\boldsymbol{V}^T = \begin{bmatrix} 1.009 & 1.009 & 1.009 & 2.016 & 2.016 \end{bmatrix}$$

So it suggests that Leslie may like romantic movies(Titanic, Casablanca) very much, and also have some interests for other sci-fi movies.

Exercise 11.4.2

(a) The square of the Frobenius norm of $\boldsymbol{M}$ is $f_{\boldsymbol{M}} = 4(1 + 9 + 16 + 25) + 3(16 + 25 + 4) = 243$.

The columns for *The Matrix* and *Alien* each have a squared Frobenium norm of 51, therefore their probability is $51/243 = 0.210$. The scaler is $\sqrt{2P(j)} = 0.648$.

$$\boldsymbol{C} = \begin{bmatrix} \frac{\boldsymbol{c}_{matrix}}{\sqrt{2P(matrix)}} & \frac{\boldsymbol{c}_{alien}}{\sqrt{2P(alien)}} \end{bmatrix} = \begin{bmatrix} 1.54 & 4.63 & 6.17 & 7.72 & 0 & 0 & 0 \\ 1.54 & 4.63 & 6.17 & 7.72 & 0 & 0 & 0 \end{bmatrix}^T$$

The rows for Jim and John each have a squared Frobenium norm of 27, 48, respectively, therefore the probabilities are $27/243 = 0.111, 48/243 = 0.198$, respectively. The scalers are $\sqrt{2P(Jim)} = 0.471$, $\sqrt{2P(John)} = 0.629$.

$$\boldsymbol{R} = \begin{bmatrix} \frac{\boldsymbol{c}_{Jim}}{\sqrt{2P(Jim)}} \\ \frac{\boldsymbol{c}_{John}}{\sqrt{2P(John)}} \end{bmatrix} = \begin{bmatrix} 6.37 & 6.37 & 6.37 & 0 & 0 \\ 6.36 & 6.36 & 6.36 & 0 & 0 \end{bmatrix}$$

$\boldsymbol{W} = \begin{bmatrix} 3 & 3 \\ 4 & 4 \end{bmatrix}$. Use svd on $\boldsymbol{W}$:

$$\boldsymbol{W} = \boldsymbol{X}\boldsymbol{Z}\boldsymbol{Y}^T = \begin{bmatrix} 0.6 & -0.8 \\ 0.8 & 0.6 \end{bmatrix} \cdot \begin{bmatrix} \sqrt{50} & 0 \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0.707 & 0.707 \\ 0.707 & -0.707 \end{bmatrix}$$

$$\boldsymbol{Z}^+ = \begin{bmatrix} \frac{1}{\sqrt{50}} & 0 \\ 0 & 0 \end{bmatrix}$$

Then,

$$\boldsymbol{U} = \boldsymbol{Y}(\boldsymbol{Z}^+)^2\boldsymbol{X}^T = \begin{bmatrix} 0.0085 & 0.0113 \\ 0.0085 & 0.0113 \end{bmatrix}$$

Therefore, the CUR decomposition of $\boldsymbol{M}$ is:

$$\boldsymbol{M} \approx \boldsymbol{C} \cdot \boldsymbol{U} \cdot \boldsymbol{R} = \begin{bmatrix} 1.54 & 4.63 & 6.17 & 7.72 & 0 & 0 & 0 \\ 1.54 & 4.63 & 6.17 & 7.72 & 0 & 0 & 0 \end{bmatrix}^T \cdot \begin{bmatrix} 0.0085 & 0.0113 \\ 0.0085 & 0.0113 \end{bmatrix} \cdot \begin{bmatrix} 6.37 & 6.37 & 6.37 & 0 & 0 \\ 6.36 & 6.36 & 6.36 & 0 & 0 \end{bmatrix}$$