Matrix Factorization

常用在商品推薦系統

Otakus v.s. No. of Figures

A
B
C
D
E

There are some common *factors* behind otakus and characters.

http://www.quuxlabs.com/blog/2010/09/matrix-factorization-a-simple-tutorial-and-implementation-in-python/

肥宅

Otakus v.s. No. of Figures

The factors are latent.

link越粗代表買的數量越多

代表的vector

No one cares ……

代表的vector

代表的vector

match

A

B

C

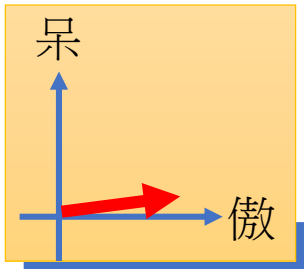
代表的vector

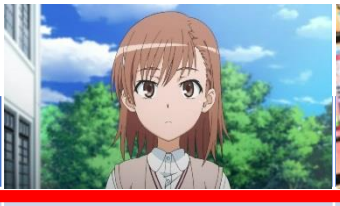
代表的vector

Not directly observable

代表的vector

代表的vector



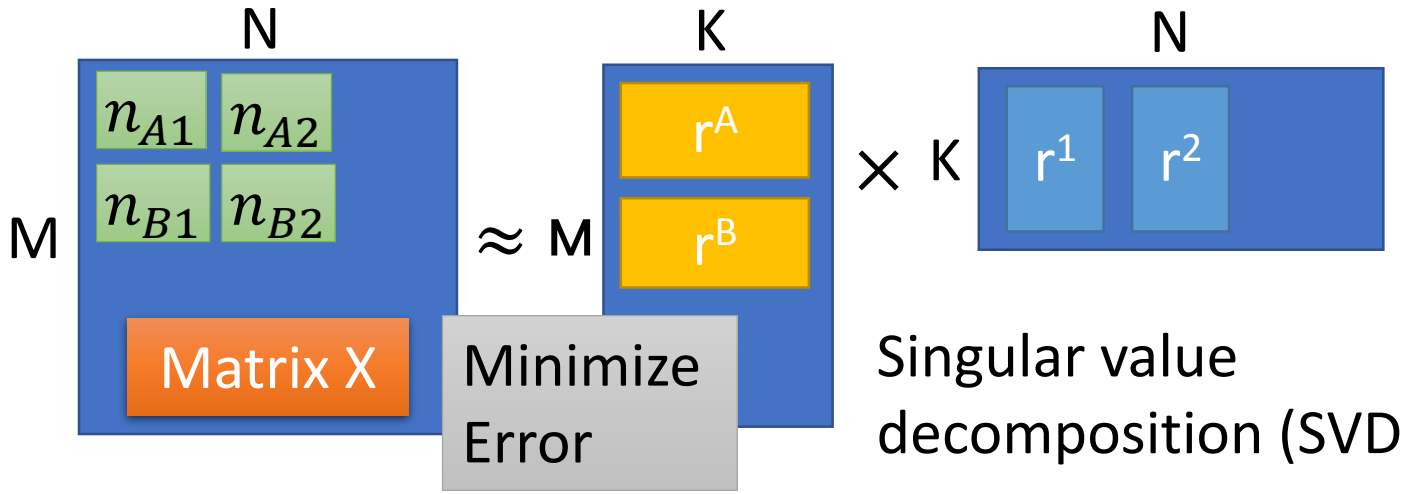
	r^1	r^2	r^3	r^4
r^A	A	5	3	0	1
r^B	B	4	3	0	1
r^C	C	1			5
r^D	D	1	0	4	4
r^E	E	0	1	5	4

Matrix X

No. of Otakus = M No. of characters = N No. of latent factor = K

代表的dimension數目

$$r^A \cdot r^1 \approx 5$$
$$r^B \cdot r^1 \approx 4$$
$$r^C \cdot r^1 \approx 1$$
$$\vdots$$



Singular value decomposition (SVD)

r^i \ r^j	r^1	r^2	r^3	r^4
				
r^A	A	5 n_{A1}	3	?	1
r^B	B	4	3	?	1
r^C	C	1	1	?	5
r^D	D	1	?	4	4
r^E	E	?	1	5	4

因為矩陣少了一些值，因此直接利用SVD是比較困難的

$r^A \cdot r^1 \approx 5$

$r^B \cdot r^1 \approx 4$

$r^C \cdot r^1 \approx 1$

⋮

Minimizing

Only considering the defined value

$$L = \sum_{(i,j)} \left(r^i \cdot r^j - n_{ij} \right)^2$$

Find r^i and r^j by gradient descent

		r^1	r^2	r^3	r^4
					
r^A	A	5	3	-0.4	1
r^B	B	4	3	-0.3	1
r^C	C	1	1	2.2	5
r^D	D	1	0.6	4	4
r^E	E	0.1	1	5	4

Assume the dimensions of r are all 2 (there are two factors)

假設二維

A	0.2	2.1
B	0.2	1.8
C	1.3	0.7
D	1.9	0.2
E	2.2	0.0

1 (春日)	0.0	2.2
2 (炮姐)	0.1	1.5
3 (姐寺)	1.9	-0.3
4 (小唯)	2.2	0.5

More about Matrix Factorization

- Considering the induvial characteristics

$$r^A \cdot r^1 \approx 5 \quad \Rightarrow \quad r^A \cdot r^1 + b_A + b_1 \approx 5$$

A本身的購買力

1公仔本身被買的能力

b_A: otakus A likes to buy figures

b_1: how popular character 1 is

Minimizing $L = \sum_{(i,j)} \left(r^i \cdot r^j + b_i + b_j - n_{ij}\right)^2$

Find r^i, r^j, b_i, b_j by gradient descent (can add regularization)

- Ref: Matrix Factorization Techniques For Recommender Systems

文件主題分析

Matrix Factorization for Topic analysis

character→document,
otakus→word

- Latent semantic analysis (LSA)

	Doc 1	Doc 2	Doc 3	Doc 4
投資	5	3	0	1
股票	4	0	0	1
總統	1	1	0	5
選舉	1	0	0	4
立委	0	1	5	4

Number in Table:
詞彙在文章中出現的次數有多少次
Term frequency
(weighted by inverse document frequency)
可以估測每個詞彙的重要性，以weight代表
Latent factors are topics
(財經、政治 ……)

- Probability latent semantic analysis (PLSA)
  - Thomas Hofmann, Probabilistic Latent Semantic Indexing, SIGIR, 1999
- latent Dirichlet allocation (LDA)
  - David M. Blei, Andrew Y. Ng, Michael I. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research, 2003