Carnegie Mellon

15-826: Multimedia Databases and Data Mining

Lecture #17: SVD – part II – applications

C. Faloutsos

1

Carnegie Mellon

Problems

- Q1: How to find 'concepts' in a document collection?
- Q2: how to answer queries in English, when documents are in Spanish?
- Q3: how to compress a customer x day matrix
- Q4: how to interpret the rules/concepts
- Q5: KL transform?

2

Carnegie Mellon

Solutions

- Q1: How to find 'concepts' in a document collection?
- Q2: how to a[illegible] in English, when docume[illegible]
- Q3: how t[illegible] customer x day matrix
- Q4: how to interpret the rules/concepts
- Q5: KL transform?

SVD

3

Carnegie Mellon

Must-read Material

- MM Textbook Appendix D

4

Carnegie Mellon

Outline

Goal: 'Find similar / interesting things'

- Intro to DB
- Indexing - similarity search
- Data Mining

5

Carnegie Mellon

Indexing - Detailed outline

- primary key indexing
- secondary key / multi-key indexing
- spatial access methods
- fractals
- text
- Singular Value Decomposition (SVD)
- multimedia
- ...

6

Carnegie Mellon

SVD - Detailed outline

- Motivation
- Definition - properties
- Interpretation
- Complexity
- Case studies
- SVD properties
- Conclusions

7

Carnegie Mellon

SVD - Case studies

- multi-lingual IR; LSI queries
- compression
- PCA - 'ratio rules'
- Karhunen-Lowe transform
- query feedbacks
- google/Kleinberg algorithms

8

Carnegie Mellon

Case study - LSI

Q1: How to do queries with LSI?

Q2: multi-lingual IR (english query, on spanish text?)

9

Carnegie Mellon

Case study - LSI

Q1: How to do queries with LSI?

Problem: Eg., find documents with 'data'

$$
\begin{array}{c}
\\ \text{CS} \\ \\ \\ \\ \text{MD} \\ \\ \\
\end{array}
\begin{array}{c}
\begin{array}{ccccc} \text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} \end{array} \\
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
2 & 2 & 2 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 0 & 0 & 2 & 2 \\
0 & 0 & 0 & 3 & 3 \\
0 & 0 & 0 & 1 & 1
\end{bmatrix}
\end{array}
=
\begin{bmatrix}
0.18 & 0 \\
0.36 & 0 \\
0.18 & 0 \\
0.90 & 0 \\
0 & 0.53 \\
0 & 0.80 \\
0 & 0.27
\end{bmatrix}
\times
\begin{bmatrix}
9.64 & 0 \\
0 & 5.29
\end{bmatrix}
\times
\begin{bmatrix}
0.58 & 0.58 & 0.58 & 0 & 0 \\
0 & 0 & 0 & 0.71 & 0.71
\end{bmatrix}
$$

10

Carnegie Mellon

Case study - LSI

Q1: How to do queries with LSI?

A: map query vectors into 'concept space' – how?

$$
\begin{array}{c|c}
 & \begin{array}{ccccc} \text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} \end{array} \\
\begin{array}{c} \uparrow \\ \text{CS} \\ \downarrow \\ \uparrow \\ \text{MD} \\ \downarrow \end{array} &
\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}
\end{array}
= \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix}
\times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix}
\times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}
$$

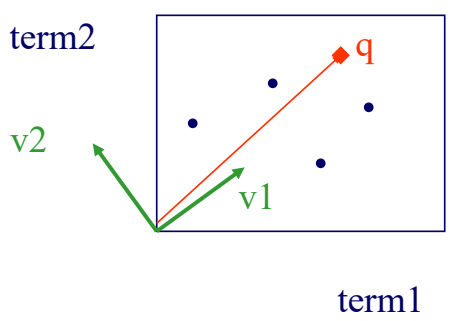
Carnegie Mellon

Case study - LSI

Q1: How to do queries with LSI?

A: map query vectors into 'concept space' – how?

$$
q = \begin{array}{c} \begin{array}{ccccc} \text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}
$$



Carnegie Mellon

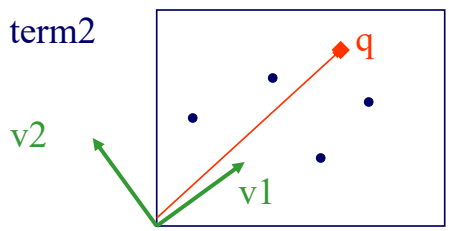
Case study - LSI

Q1: How to do queries with LSI?

A: map query vectors into 'concept space' – how?

data inf. retrieval brain lung

q= [1 0 0 0 0]



A: inner product (cosine similarity) with each 'concept' vector v_i

13

Carnegie Mellon

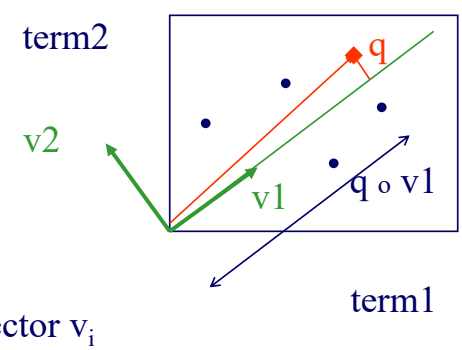
Case study - LSI

Q1: How to do queries with LSI?

A: map query vectors into 'concept space' – how?

data inf. retrieval brain lung

q= [1 0 0 0 0]



A: inner product (cosine similarity) with each 'concept' vector v_i

14

Carnegie Mellon

Case study - LSI

compactly, we have:

$$q\,\mathbf{V} = q_{concept}$$

Eg:

$$q = \begin{array}{c} \begin{array}{ccccc} \text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array} \begin{bmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \end{bmatrix} = \begin{array}{c} \text{CS-concept} \\ \downarrow \\ \begin{bmatrix} 0.58 & 0 \end{bmatrix} \end{array}$$

term-to-concept similarities

Carnegie Mellon

Case study - LSI

Drill: how would the document ('information' , 'retrieval') be handled by LSI?

Carnegie Mellon

Case study - LSI

Drill: how would the document ('information', 'retrieval') be handled by LSI? A: SAME:

$d_{concept} = d\,\mathbf{V}$

Eg:

$$d = \begin{array}{c} \begin{array}{ccccc} \text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \end{bmatrix} \end{array} \begin{bmatrix} 0.58 & 0 \\ 0.58 & 0 \\ 0.58 & 0 \\ 0 & 0.71 \\ 0 & 0.71 \end{bmatrix} = \begin{array}{c} \text{CS-concept} \\ \downarrow \\ \begin{bmatrix} 1.16 & 0 \end{bmatrix} \end{array}$$

term-to-concept similarities

17

Carnegie Mellon

Case study - LSI

Observation: document ('information', 'retrieval') will be retrieved by query ('data'), although it does not contain 'data' !!

$$\begin{array}{c} \begin{array}{ccccc} \text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} \end{array} \end{array}$$

$$d = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \end{bmatrix} \dashrightarrow \begin{array}{c} \text{CS-concept} \\ \downarrow \\ \begin{bmatrix} 1.16 & 0 \end{bmatrix} \end{array}$$

$$q = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix} \dashrightarrow \begin{bmatrix} 0.58 & 0 \end{bmatrix}$$

18

Carnegie Mellon

Solutions

- ✔Q1: How to find 'concepts' in a document collection?
- Q2: how to answer queries in English, when documents are in Spanish?
- Q3: how to compress a customer x day matrix
- Q4: how to interpret the rules/concepts
- Q5: KL transform?

Carnegie Mellon

Case study - LSI

Q1: How to do queries with LSI?

➡ Q2: multi-lingual IR (english query, on spanish text?)

Carnegie Mellon

Case study - LSI

- Problem:
  - given many documents, translated to both languages (eg., English and Spanish)
  - answer queries across languages

21

Carnegie Mellon

Case study - LSI

- Solution: ~ LSI

$$
\begin{array}{c|cccccc}
 & \text{data} & \text{inf.} & \text{retrieval} & \text{brain} & \text{lung} & \text{datos} & \text{informacion} \\
\hline
\text{CS} & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\
 & 2 & 2 & 2 & 0 & 0 & 1 & 2 \\
 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\
 & 5 & 5 & 5 & 0 & 0 & 5 & 5 \\
\text{MD} & 0 & 0 & 0 & 2 & 2 & 0 & 0 \\
 & 0 & 0 & 0 & 3 & 3 & 0 & 0 \\
 & 0 & 0 & 0 & 1 & 1 & 0 & 0
\end{array}
\qquad
\begin{array}{cc}
1 & 1 \\
2 & 2 \\
1 & 1 \\
4 & 0 \\
2 & 2 \\
2 & 3 \\
1 & 1
\end{array}
$$

22

Carnegie Mellon

Solutions

- ✔Q1: How to find 'concepts' in a document collection?
- ✔Q2: how to answer queries in English, when documents are in Spanish?
- Q3: how to compress a customer x day matrix
- Q4: how to interpret the rules/concepts
- Q5: KL transform?

Carnegie Mellon

Case study: compression

[Korn+97]

Problem:

- given a matrix
- compress it, but maintain 'random access'

(surprisingly, its solution leads to data mining and visualization...)

Flip Korn, H. V. Jagadish, and Christos Faloutsos. *Efficiently supporting ad hoc queries in large datasets of time sequences.* SIGMOD '97, 289-300.

Carnegie Mellon

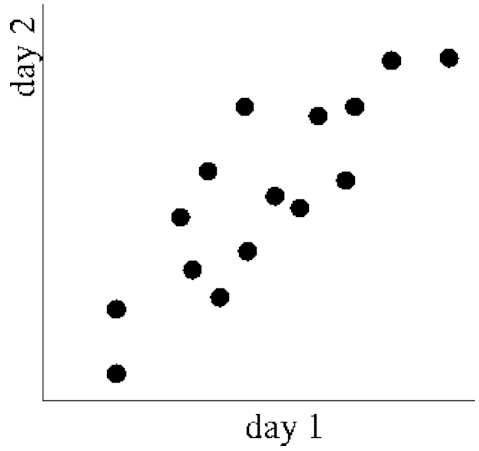
Problem - specs

- ~10**6 rows; ~10**3 columns; no updates;
- random access to any cell(s) ; small error: OK

day / customer	We 7/10/96	Th 7/11/96	Fr 7/12/96	Sa 7/13/96	Su 7/14/96
ABC Inc.	1	1	1	0	0
DEF Ltd.	2	2	2	0	0
GHI Inc.	1	1	1	0	0
KLM Co.	5	5	5	0	0
Smith	0	0	0	2	2
Johnson	0	0	0	3	3
Thompson	0	0	0	1	1

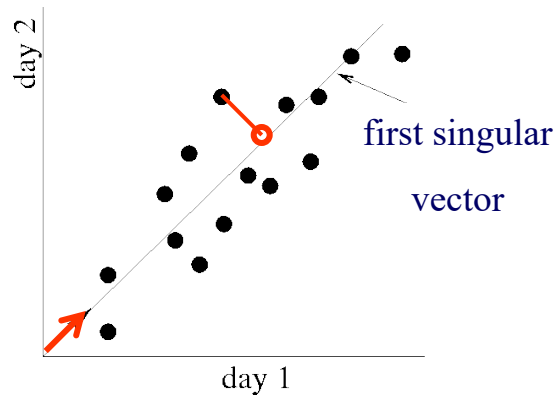
Carnegie Mellon

Idea



Carnegie Mellon

SVD - reminder



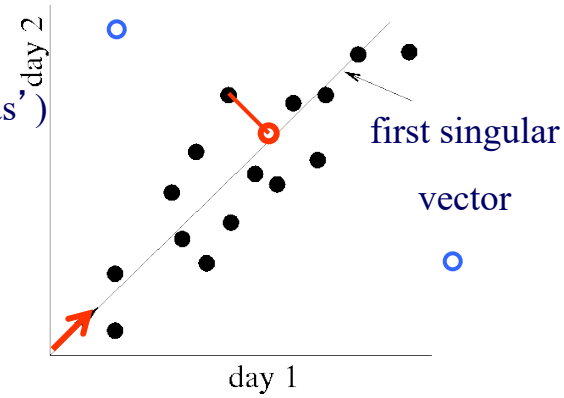
- space savings: 2:1
- minimum RMS error

28

Carnegie Mellon

Case study: compression

outliers?
A: treat separately
(SVD with 'Deltas')



29

Carnegie Mellon

Compression - Performance

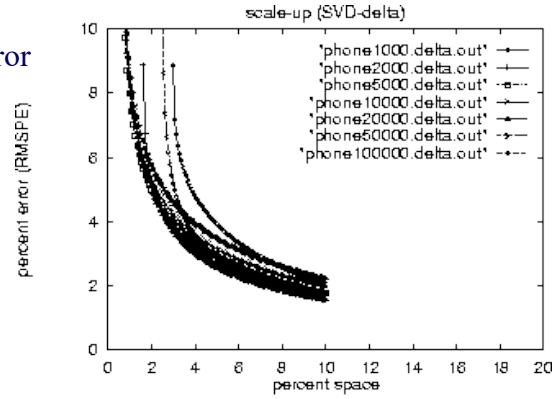
- 3 pass algo (-> scalability) (HOW?)
- random cell(s) reconstruction
- 10:1 compression with < 2% error

30

Carnegie Mellon

Performance - scaleup

error



space

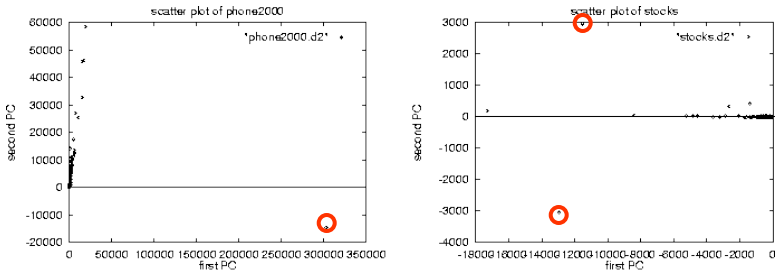
SVDD.

31

Carnegie Mellon

Compression - Visualization

- no Gaussian clusters; Zipf-like distribution



(a) 'phone2000' (b) 'stocks'

Carnegie Mellon

Solutions

- ✔ Q1: How to find 'concepts' in a document collection?
- ✔ Q2: how to answer queries in English, when documents are in Spanish?
- ✔ Q3: how to compress a customer x day matrix
- Q4: how to interpret the rules/concepts
- Q5: KL transform?

Carnegie Mellon

PCA - 'Ratio Rules'

[Korn+98]

Typically: '*Association Rules*' (eg.,

{bread, milk} -> {butter}

But, can we discover more details? like:

$-bread : $-milk : $-butter ~ $2 : $4 : $3

Flip Korn, Alexandros Labrinidis, Yannis Kotidis, and Christos Faloutsos. *Ratio Rules: A New Paradigm for Fast, Quantifiable Data Mining*. (VLDB '98), 582-593.

35

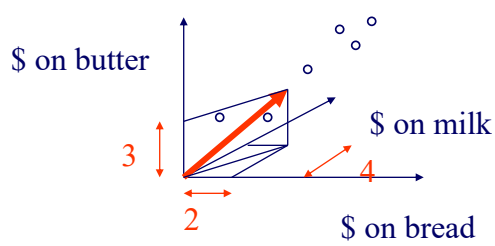
Carnegie Mellon

PCA - 'Ratio Rules'

Idea: try to find 'concepts':

- singular vectors dictate rules about ratios:

bread:milk:butter = 2:4:3



36

Carnegie Mellon

PCA - 'Ratio Rules'

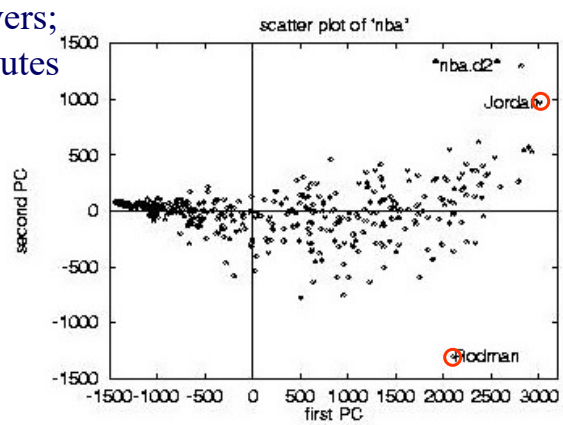
Identical to PCA = Principal Components Analysis

- √ – Q1: which set of rules is 'better'?
- √ – Q2: how to reconstruct missing/corrupted values?
- √ – Q3: is there need for binary/bucketized values? NO
- ➡ – Q4: how to interpret the rules (= 'principal components')?

Carnegie Mellon

PCA - Ratio Rules

NBA dataset
~500 players;
~30 attributes



Carnegie Mellon

PCA - Ratio Rules

- PCA: get singular vectors v1, v2, ...
- ignore entries with small abs. value
- try to interpret the rest

39

Carnegie Mellon

PCA - Ratio Rules

NBA dataset - **V** matrix (term to 'concept' similarities)

field	RR_1	RR_2	RR_3
minutes played	.808	−.4	
field goals			
goal attempts			
points	.406	.199	
total rebounds		−.489	.602
assists			−.486
steals			−.07

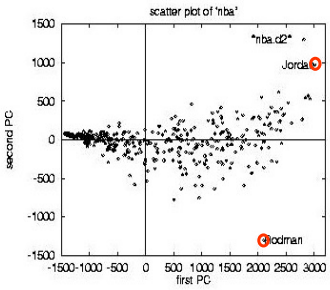
v1

40

Carnegie Mellon

Ratio Rules - example

- RR1: minutes:points = 2:1
- corresponding concept?



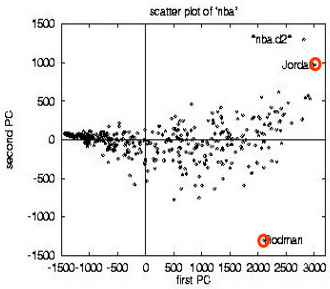
→ v1

41

Carnegie Mellon

Ratio Rules - example

- RR1: minutes:points = 2:1
- corresponding concept?



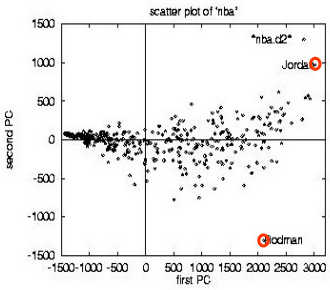
→ v1

42

Carnegie Mellon

Ratio Rules - example

- RR1: minutes:points = 2:1
- co
  co



v1

Carnegie Mellon

Ratio Rules - example

- RR1: minutes:points = 2:1
- corresponding concept?
- A: 'goodness' of player

Carnegie Mellon

Ratio Rules - example

- RR2: points: rebounds negatively correlated(!)

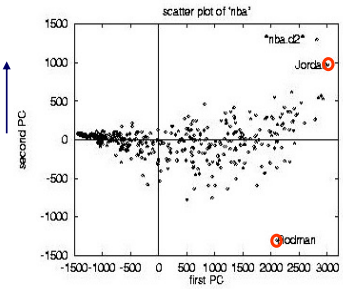
field	RR_1	RR_2	RR_3
minutes played	.808	−.4	
field goals			
goal attempts			
points	.406	.199	
total rebounds		−.489	.602
assists			−.486
steals			−.07

Carnegie Mellon

Ratio Rules - example

- RR2: points: rebounds negatively correlated(!) - concept?

v2



Carnegie Mellon

Ratio Rules - example

- RR2: points: rebounds negatively correlated(!) - concept?
- A: position: offensive/defensive

47

Carnegie Mellon

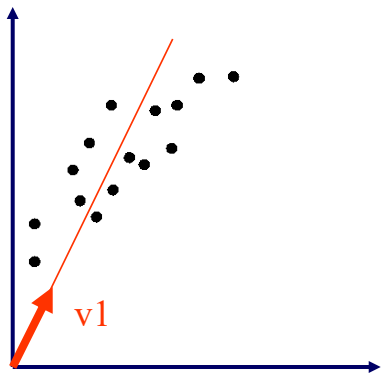
Solutions

- ✔Q1: How to find 'concepts' in a document collection?
- ✔Q2: how to answer queries in English, when documents are in Spanish?
- ✔Q3: how to compress a customer x day matrix
- ✔Q4: how to interpret the rules/concepts
- Q5: KL transform?

49

Carnegie Mellon

K-L transform



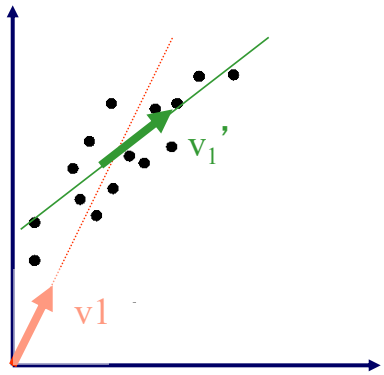
[Duda & Hart]; [Fukunaga]

A subtle point:
SVD will give vectors that go through the origin

50

Carnegie Mellon

K-L transform

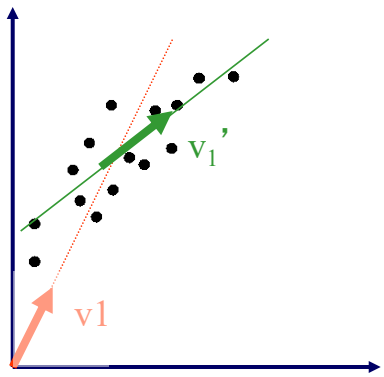


A subtle point:
SVD will give vectors that go through the origin
Q: how to find v_1' ?

51

Carnegie Mellon

K-L transform



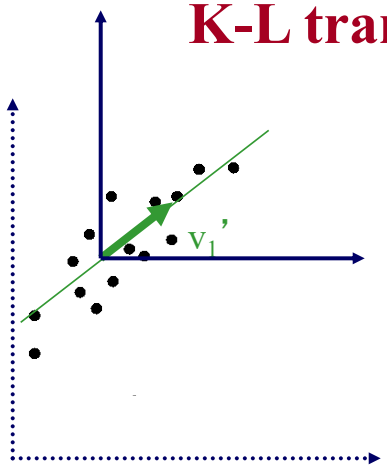
A subtle point:
SVD will give vectors that go through the origin
Q: how to find v_1' ?

A: 'centered' PCA, ie., move the origin to center of gravity

52

Carnegie Mellon

K-L transform



A subtle point:
SVD will give vectors that go through the origin
Q: how to find v_1' ?

A: 'centered' PCA, ie., move the origin to center of gravity
and THEN do SVD

53

Carnegie Mellon

K-L transform

- How to 'center' a set of vectors (= data matrix)?
- What is the covariance matrix?
- A: see textbook
- ('whitening transformation')

Carnegie Mellon

Conclusions

- SVD: popular for dimensionality reduction / compression
- SVD is the 'engine under the hood' for PCA (principal component analysis)
- ... as well as the Karhunen-Lowe transform
- (and there is more to come ...)

Carnegie Mellon

Solutions

- ✔Q1: How to find 'concepts' in a document collection?
- ✔Q2: how to a[illegible] n English, when docume[illegible]
- ✔Q3: how t[illegible] customer x day matrix
- ✔Q4: how to interpret the rules/concepts
- ✔Q5: KL transform?

SVD

56

Carnegie Mellon

References

- Duda, R. O. and P. E. Hart (1973). Pattern Classification and Scene Analysis. New York, Wiley.
- Fukunaga, K. (1990). Introduction to Statistical Pattern Recognition, Academic Press.
- Jolliffe, I. T. (1986). Principal Component Analysis, Springer Verlag.

57

Carnegie Mellon

References

- Korn, F., H. V. Jagadish, et al. (May 13-15, 1997). Efficiently Supporting Ad Hoc Queries in Large Datasets of Time Sequences. ACM SIGMOD, Tucson, AZ.
- Korn, F., A. Labrinidis, et al. (1998). Ratio Rules: A New Paradigm for Fast, Quantifiable Data Mining. VLDB, New York, NY.

Carnegie Mellon

References

- [Korn+, ’00] Korn, F., A. Labrinidis, et al. (2000). "Quantifiable Data Mining Using Ratio Rules." VLDB Journal 8(3-4): 254-266.
- Press, W. H., S. A. Teukolsky, et al. (1992). Numerical Recipes in C, Cambridge University Press.