Latent Semantic Indexing and Dimension Reduction

Yan Liu

Outline

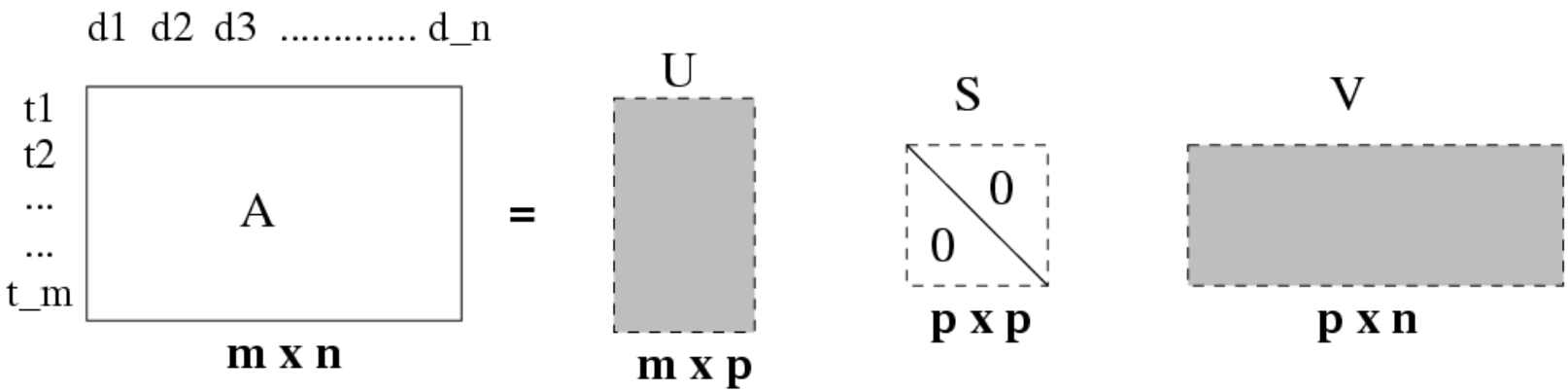
- Mathematical derivation
  - Singular Value Decomposition (SVD)
- Application
  - Latent Semantic Indexing
    - Language Modeling
  - Principle component analysis
    - Clustering

Matrix Decomposition

- Eigen Decomposition
  - Non-singular square matrix W
- Singular Value Decomposition
  - An m * n matrix W
- Generalized Singular Value Decomposition
  - An m * n matrix A and p * n matrix B

Latent Semantic Indexing

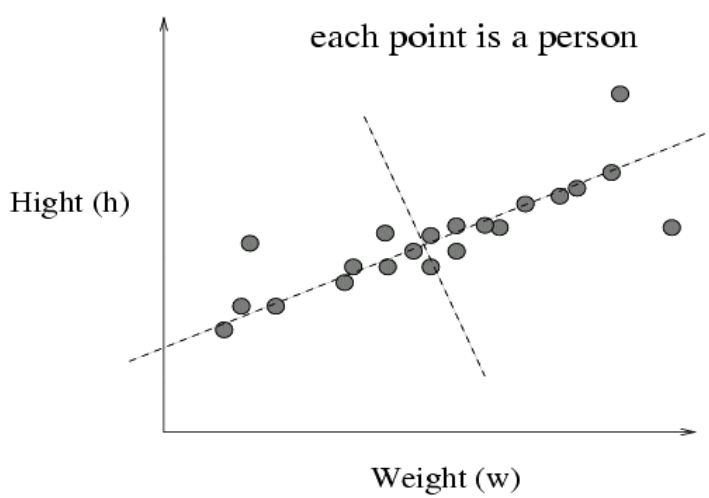
- Word-document matrix representation W
- Singular value decomposition
  - $W = USV^T$
- Document representation
  - $d_i' = d_i US^{-1}$
- Word-to-word correlation, document-to-document correlation
  - $WW^T = US^2U^T$, $W^TW = VS^2V^T$


d1 d2 d3 ..... d_n
t1
t2
...
...
t_m
A
m x n
=
U
m x p
S
0
0
p x p
V
p x n

Principle Component Analysis

- to exploit major patterns in word usage and to ignore miner differences
- to discover semantic structures and better understand the data (by the user)
- to have more transparent mathematical properties (singular vectors "reserving" main variances in data)

A toy example for Singular Value Documposition



What is the Principal Singular Vector?