UPPSALA
UNIVERSITET

Information Retrieval (5LN712)

Matrix decomposition and latent semantic indexing

2020-05-25

Ali Basirat
Department of Linguistics and Philology

UPPSALA UNIVERSITET

Today

- Linear algebra
- Latent semantic indexing

UPPSALA UNIVERSITET

Linear algebra

- The rank of a matrix is the number of linearly independent rows or columns in it
- The vectors $\vec{v}_1, \ldots, \vec{v}_n$ are linearly independent if $a_1\vec{v}_1 + \cdots + a_n\vec{v}_n = 0$ has no non-zero solution
- None of the vectors can be formed by linear combination of other vectors
- If $a_1\vec{v}_1 + \cdots + a_n\vec{v}_n = 0$ has a non-zero solution, then $\vec{v}_1, \ldots, \vec{v}_n$ are linearly dependent
- If C is an M×N matrix, then

$$Rank(C) \leq \min\{M, N\}$$

UPPSALA UNIVERSITET

Linear algebra

- A square matrix is diagonal if all its off-diagonal elements are zero

$$\begin{matrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{matrix}$$

- The rank of an $r \times r$ diagonal matrix is r
- If the diagonal elements are 1, then the matrix is called the identity matrix, I_r

UPPSALA UNIVERSITET

Linear algebra

- A matrix is a linear operator that maps a vector space onto another space

$$A\vec{x} = \vec{y}$$

- The direction of some vectors is not changed under the matrix operation (given that A is a square matrix)

$$A\vec{x} = \lambda\vec{x}$$

- $\vec{x}$ is an eigenvector of A and λ is the corresponding eigenvalue of A
- The number of nonzero eigenvalues of A is at most rank(A)

UPPSALA UNIVERSITET

Linear algebra

- Example: the matrix

$$\begin{matrix} 30 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 1 \end{matrix}$$

has three eigenvalues and eigenvectors

$$\lambda_1 = 30, \lambda_1 = 20, \lambda_1 = 1$$

$$x_1 = \begin{matrix} 1 \\ 0 \\ 0 \end{matrix}, \quad x_2 = \begin{matrix} 0 \\ 1 \\ 0 \end{matrix}, \quad x_3 = \begin{matrix} 0 \\ 0 \\ 1 \end{matrix}$$

UPPSALA UNIVERSITET

Linear algebra

- Any vector can be expressed as a linear combination of eigenvectors

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = a\times\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + b\times\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + c\times\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

UPPSALA UNIVERSITET

Linear algebra

- If $\vec{x}_1, \ldots, \vec{x}_n$ are the eigenvectors corresponding to the real nonzero eigenvalues $\lambda_1, \ldots, \lambda_n$ of the matrix A of rank n and $\vec{v}$ is an n-dimensional vector, then

$$A\vec{v} = \sum_{i=1}^{n} v_i \lambda_i \vec{x}_i$$

- Example: if $A = \begin{bmatrix} 30 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ and $\vec{v} = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}$, then

$$A\vec{v} = 2{\times}30{\times}\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + 4{\times}20{\times}\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + 6{\times}1{\times}\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 60 \\ 80 \\ 6 \end{bmatrix}$$

- The effect of small eigenvalues on a matrix-vector product is small

UPPSALA UNIVERSITET

Linear algebra

- Matrix decomposition: a matrix is factored into the product of matrices
- Eigen decomposition: every square real-valued matrix $A_{M\times M}$ of rank M can be decomposed to

$$A = U\Lambda U^{-1}$$

where $U = [\vec{x}_1 \quad \dots \quad \vec{x}_m]$ is a matrix of eigenvectors and

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_m \end{bmatrix}$$

is a diagonal vector of eigenvalues

UPPSALA UNIVERSITET

Linear algebra

- Symmetric diagonal decomposition: every square, symmetric, real-valued matrix $A_{M\times M}$ of rank M can be decomposed to

$$A = Q\Lambda Q^T$$

where columns of Q are orthonormal eigenvectors of A and Λ is a diagonal matrix of eigenvalues

UPPSALA UNIVERSITET

Linear algebra

- Singular value decomposition: every $M \times N$ matrix A of rank r can be decomposed to

$$A = U\Sigma V^T$$

where

- $U_{\mathrm{M}\times r}$ is the matrix of eigenvalues of AA^T
- $\Sigma_{\mathrm{r}\times r}$ is a diagonal matrix with $\sigma_{i,i} = \sqrt{\lambda_i}$
- $V_{\mathrm{N}\times r}$ is the matrix of eigenvalues of $A^T A$

UPPSALA UNIVERSITET

Linear algebra

- Truncated SVD: the small singular values and their correpsonding righ and left singular vectors are discarded

$$\tilde{A} = U_t \Sigma_t V_t^T$$

where t is smaller than the rank of A

- $\tilde{A}$ is a rank t approximation of A
- $\tilde{A}$ is a rank t matrix with minimum discrepancy with A

$$\tilde{A} = \arg\min_{A_t} ||A - A_t||$$

UPPSALA UNIVERSITET

Latent semantic indexing

- A term-document matrix is not necessarily a square and/or symmetric matrix
- We can estimate its low rank approximation using truncated SVD
- The rank of the matrix can be very high
- LSI uses truncated SVD to construct a low-rank approximation of the matrix
- The size of the matrix does not change

UPPSALA UNIVERSITET

Latent semantic indexing

- Documents are represented based on the singular values and vectors

$$C_k = U_k \Sigma_k V_k^T$$

  - Terms are represented by left singular vectors
  - Documents are represented by right singular vectors
  - Σ_k show the variance along the bases of document and term spaces
  - C_k has the same size as C

UPPSALA UNIVERSITET

Latent semantic indexing

- A query vector $\vec{q}$ is projected to the LSI space

$$\vec{q}_k = \Sigma_k^{-1} U_k^T \vec{q}$$

- $\vec{q}_k$ is a dense vector and it needs significant computations
- Similarly, we can generate dense document vectors from word counts

UPPSALA UNIVERSITET

Latent semantic indexing

- The SVD of term-documents matrices is computationally expensive
- LSI works well on the domains that there is little overlap between queries and documents
- LSI provides a solution to synonymy and polysemy
- LSI can be considered as soft document clustering

UPPSALA UNIVERSITET

Summary

- Eigenvectors are the basis of a vector space
- Different matrix decompositions
- Singular value decomposition for term-document matrices
- Low-rank matrix factorization
- Latent semantic indexing
- How to project query/document vectors to a low-rank vector space