Foundations of Data Science

Lecture 5: Matrix Decomposition

MING GAO

DaSE@ECNU
(for course related communications)
mgao@sei.ecnu.edu.cn

Oct. 9, 2016

Outline

1. Singular value decomposition (SVD)
   - SVD
   - PCA
2. Matrix Factorization

Dimensionality reduction

Motivation

- High-dimension means many features.
  - Netflix: 480K users and 177K movies
  - Documents VS. words: thousands of words and billions of documents
  - Taobao: millions of users and millions of products
- Dimensionality reduction or compression
  - Discover hidden correlations, concepts or topics due to objects that occur commonly together.
  - Remove redundant and noisy features, not all features are useful
  - Easier storage and processing of the data

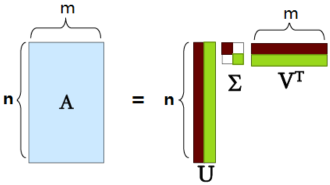
Outline

1. Singular value decomposition (SVD)
   - SVD
   - PCA
2. Matrix Factorization

SVD

Definition

Any real $m \times n$ matrix A can be decomposed uniquely as $A_{[n\times m]} \sim U_{[n\times r]} D_{[r\times r]} V^T_{r\times m}$ where U, V are orthogonal matrix, D is a diagonal matrix (non-negative real values called singular values).

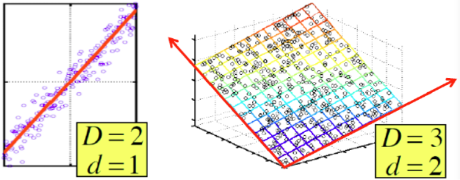


- A: an input $n \times m$ data matrix, e.g., n users and m items.
- U: left singular vectors in a $n \times r$ matrix, e.g., n users and r interests.
- D: a $r \times r$ diagonal matrix, e.g., strength of each interest.
- V: right singular vectors in a $r \times m$ matrix, e.g., r interests and m

SVD cont.

Properties

- It is always possible to decompose a real matrix A into $A = UDV^T$
  - U, D, V are unique, and D is a diagonal matrix.
  - U, V are column orthogonal (i.e., $U^T U = I$ and $V^T V = I$)
  - Entries of D are positive and sorted in decreasing order $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$.
- Axes of this subspace are effective representation of the data.



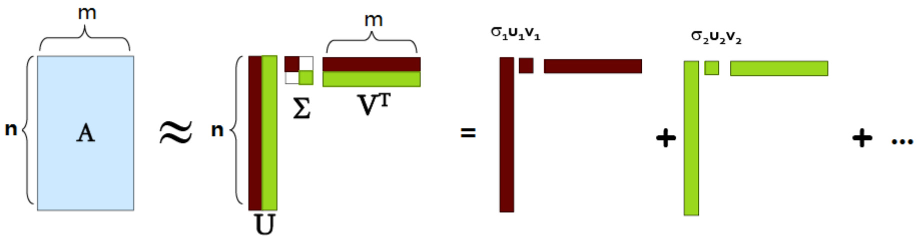
Assumption

- Data lies on or near a low $d-$dimensional subspace.
- Axes of this subspace are effective representation of the data.

Methodology of decomposition

Diagonalization

- $AA^T = UDV^TVDU^T = UD^2U^T$, where $D = diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2)$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$.
- $A^TA = VDU^TUDV^T = VD^2V^T$
- $A = UDV^T$. If $U = (\mathbf{u}_1\ \mathbf{u}_2\ \cdots \mathbf{u}_n)$ and $U = (\mathbf{v}_1\ \mathbf{v}_2\ \cdots \mathbf{v}_n)$, then $A = \sum_{i=1}^n \sigma_i \mathbf{u}_i \mathbf{v}_i^T$.



Methodology of dimensionality reduction

$$A = \left[\begin{array}{ccc|ccc} u_1 & \cdots & u_k & u_{k+1} & \cdots & u_m \end{array}\right] \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_k & \\ \hline & 0 & & 0 \end{array}\right] \left[\begin{array}{c} v_1^T \\ \vdots \\ v_k^T \\ \hline v_{k+1}^T \\ \vdots \\ v_n^T \end{array}\right]$$

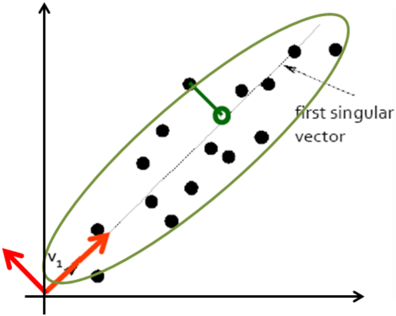
$$A = \left[\begin{array}{ccc} u_1 & \cdots & u_k \end{array}\right] \left[\begin{array}{ccc} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{array}\right] \left[\begin{array}{c} v_1^T \\ \vdots \\ v_k^T \end{array}\right]$$

Criteria

- $k = \arg\min_r \left\{ \frac{\sum_{i=1}^r \sigma_i}{\sum_{i=1}^n \sigma_i} > 90\% \right\}$, $A \approx \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^T$
- For example

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 3 & 3 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \approx \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

SVD interpretation



Explanation

- SVD gives best axis to project entities, where "best" means to minimize sum of squares of projection errors.
- SVD also gives the minimum reconstruction errors.

Applications of SVD

Applications

- A square matrix A is nonsingular (i.e., $\sigma_i \neq 0$ for all i)
  - If A is a nonsingular matrix, then its inverse is given by $A^{-1} = V^T D^{-1} U$.
  - If A is singular or ill-conditioned, then we can use SVD to approximate its inverse by the following matrix: $A^{-1} = (UDV^T)^{-1} \approx VD_0^{-1}U^T$, where t is a small threshold and $D_0^{-1} = \begin{cases} \frac{1}{\sigma_i}, & \text{if } \sigma_i > t; \\ 0, & \text{otherwise.} \end{cases}$
- Consider linear system $Ax = b$, where $A \in \mathbb{R}^{n\times m}$. If $A^T A$ is ill-conditioned (small changes in b can lead to relatively large changes in the solution x) or singular, $x \approx VD_0^{-1}U^T b$.
- SVD can be helpful to similarity query or join.
- Data compression and anomaly detection.

Outline

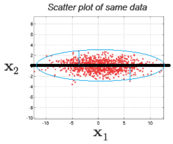
PCA: an important application of SVD

Motivation

PCA: Principle Component Analysis

- Problems arise when performing data mining or machine learning in a high-dimensional space (e.g., curse of dimensionality).
- Significant improvements can be achieved by first mapping the data into a lower-dimensionality space.
- Preserve as much information as possible

$$x = \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_N \end{bmatrix} -- > reduce\ dimensionality -- > y = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_K \end{bmatrix} \quad (K << N)$$



Goals

- Find a good representation for features (what?)
- Reduce redundancy in the data (how?)

PCA cont.

Criteria of good representation for features

- Minimize relation of the different dimensions.
- Keep the dimension as low as possible.

The best low-dimensional space

- It also gives best axis to project data, where "best" means to minimize sum of squares of projection errors.
- It also gives the minimum reconstruction errors.
- It can be determined by the "best" eigenvectors of the covariance matrix of x (i.e., the eigenvectors corresponding to the "largest" eigenvalues, also called "principal components").

Methodology

Methodology

Suppose $x_1, x_2, \cdots, x_n$ are $d \times 1$ vectors, then

1. $\overline{x} = \sum_{i=1}^{n} x_i$.
2. Subtract the mean: $y_i = x_i - \overline{x}$.
3. Form the matrix $A = [y_1 \ y_2 \ \cdots y_n]$ ($d \times n$ matrix), then compute

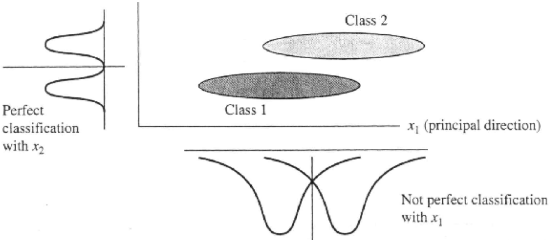
$$C = \frac{1}{n}\sum_{i=1}^{n} y_i {y_i}^T = AA^T (d \times d \text{matrix})$$

4. Compute the eigenvalues of $C : \lambda_1 > \lambda_2 > \cdots > \lambda_d$, and corresponding eigenvectors $u_1, u_2, \cdots, u_d$.
5. Keep only the terms corresponding to the k largest eigenvalues: $\widehat{x} - \overline{x} = \sum_{i=1}^{k} b_i u_i$, i.e., the representation of $\widehat{x} - \overline{x}$ into the basis $u_1, u_2, \cdots, u_k$.

Information loss

Analysis

- $\widehat{x} - \overline{x} = \sum_{i=1}^{k} b_i u_i$, i.e., $\widehat{x} = \sum_{i=1}^{k} b_i u_i + \overline{x}$
- It can be shown that the low-dimensional basis based on principal components minimizes the reconstruction error: $e = \|x - \widehat{x}\|$
- It can be shown that the error is equal to $e = \sum_{i=k+1}^{d} \lambda_i$.
- PCA is not always an optimal dimensionality-reduction procedure, e.g., classification problem.



SVD: Pros & Cons

Pros

- Optimal low-rank approximation in L_2 norm.
- There are many implementations, such as LINPACK, Matlab, SPlus, Mathematica...

Cons

- Conventional SVD is undefined for incomplete matrices.
- The complexity of computing SVD is $O(nm^2)$ or $O(n^2m)$. Less work if we want first k singular vecotrs or matrix is sparse.
- We need an approach that can simply ignore missing values and reduce the complexity.

Matrix factorization

Definition

Given a set of users U, and a set of items D, let $R \in \mathbb{R}^{|U| \times |D|}$ be the rating matrix. The matrix factorization is to find two matrices $P \in \mathbb{R}^{|U| \times K}$ and $Q \in \mathbb{R}^{|D| \times K}$ such that $R \approx PQ^T = \widehat{R}$,where K denotes the dimensionality of latent features.

- Each row of P would represent the strength of the associations between a user and the features.
- Each row of Q would represent the strength of the associations between an item and the features.
- Now, we have to find a way to obtain P and Q.

Problem formulation

Formal definition

- Each row of P would represent the strength of the associations between a user and the features.
- Each row of Q would represent the strength of the associations between an item and the features.
- Now, we have to find a way to obtain P and Q.

Take-home messages

- SVD
- Matrix factorization
  - Simple algorithm
  - PMF
  - NMF
- CUR
- PCA