Machine Learning for Big Data: Singular Value Decomposition

Lionel Fillatre

fillatre@unice.fr

Topics

- Introduction
- Eigen-decomposition
- Singular Value Decomposition
- Some applications
- Conclusion

1 Introduction

What is feature reduction?

- A **feature** is an individual measurable property of a phenomenon being observed.
  - Examples: attribute, vector of reals, image patch, phonem, histogram, etc.

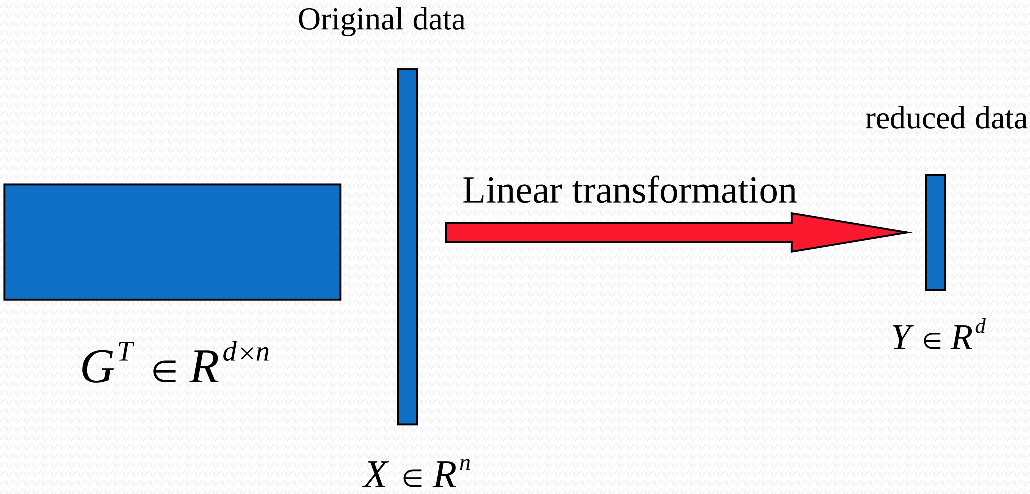
- Feature reduction refers to the mapping of the original high-dimensional data onto a lower-dimensional space.
  - Criterion for feature reduction can be different based on different problem settings.
    - Unsupervised setting: minimize the information loss
    - Supervised setting: maximize the class discrimination

- Given a set of p data points of n variables $\{x_1, x_1, \dots, x_p\}$

- Compute the linear transformation (projection)

$$G \in R^{n \times d} : x \in R^n \rightarrow y = G^T x \in R^n \ (d << n)$$

What is feature reduction?



$$G \in R^{n\times d} : X \to Y = G^T X \in R^n$$

Feature reduction versus feature selection

- Feature reduction
  - All original features are used
  - The transformed features are linear combinations of the original features.

- Feature selection
  - Only a subset of the original features are used.

2 Eigen-decomposition

Eigenvalues and Eigenvectors

- Eigenvectors (for a square $m \times m$ matrix S)

$$\boxed{\mathbf{S}\mathbf{v} = \lambda \mathbf{v}}$$

(right) eigenvector $\mathbf{v} \in \mathbb{R}^m \neq \mathbf{0}$

eigenvalue $\lambda \in \mathbb{R}$

Example

$$\begin{pmatrix} 6 & -2 \\ 4 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \end{pmatrix} = 2 \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

- How many eigenvalues are there at most?

$$\mathbf{S}\mathbf{v} = \lambda \mathbf{v} \iff (\mathbf{S} - \lambda \mathbf{I})\, \mathbf{v} = \mathbf{0}$$

only has a non-zero solution if $|\mathbf{S} - \lambda \mathbf{I}| = 0$

- This is a m-th order equation in λ which can have at most m distinct solutions (roots of the characteristic polynomial)
- Roots can be complex even though S is real.

Example

- Let $S = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$

Real, symmetric.

- Then $S - \lambda I = \begin{bmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{bmatrix} \Rightarrow (2-\lambda)^2 - 1 = 0.$
- The eigenvalues are 1 and 3 (nonnegative, real).
- The eigenvectors are orthogonal (and real):

$$\begin{pmatrix} 1 \\ -1 \end{pmatrix} \quad \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Plug in these values and solve for eigenvectors

Properties

- Eigenvalues and eigenvectors are only defined for square matrices
- Eigenvectors are not unique (e.g., if v is an eigenvector, so is $k\,v$)
- Suppose $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the eigenvalues of A, then:

(1) $\sum_i \lambda_i = tr(A)$

(2) $\prod_i \lambda_i = det(A)$

(3) if $\lambda = 0$ is an eigenvalue, then the matrix is not invertible

Eigenvalues & Eigenvectors

- For symmetric matrices, eigenvectors for distinct eigenvalues are orthogonal

$$Sv_{\{1,2\}} = \lambda_{\{1,2\}} v_{\{1,2\}}, \text{ and } \lambda_1 \neq \lambda_2 \Rightarrow v_1^T v_2 = 0$$

- All eigenvalues of a real symmetric matrix are real.

- All eigenvalues of a positive semidefinite matrix are non-negative

$$\forall w \in R^m, w^T S w \geq 0, \text{ then if } \mathrm{Sv} = \lambda \mathrm{v} \Rightarrow \lambda \geq 0$$

Eigen/diagonal Decomposition

- Let S be a square matrix with m linearly independent eigenvectors !

- Theorem: there exists an eigen-decomposition $S = U\Lambda U^{-1}$
  - Columns of U are eigenvectors of S
  - Diagonal elements of Λ are eigenvalues of S

$$\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_m), \qquad \lambda_i \geq \lambda_{i+1}$$

Sketch of proof

Let ***U*** have the eigenvectors as columns: $U = \begin{bmatrix} & & \\ v_1 & \cdots & v_n \\ & & \end{bmatrix}$

Then, ***SU*** can be written

$$SU = S\begin{bmatrix} & & \\ v_1 & \cdots & v_n \\ & & \end{bmatrix} = \begin{bmatrix} & & \\ \lambda_1 v_1 & \cdots & \lambda_n v_n \\ & & \end{bmatrix} = \begin{bmatrix} & & \\ v_1 & \cdots & v_n \\ & & \end{bmatrix}\begin{bmatrix} \lambda_1 & & \\ & \cdots & \\ & & \lambda_n \end{bmatrix}$$

Thus $\boldsymbol{SU=U\Lambda}$, or $\boldsymbol{U^{-1}SU=\Lambda}$

Diagonal decomposition - example

Recall $S = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}; \lambda_1 = 1, \lambda_2 = 3.$

The eigenvectors $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ form $U = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$

Inverting, we have $U^{-1} = \begin{bmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{bmatrix}$

Recall $UU^{-1} = 1$.

Then, $\boldsymbol{S=U\Lambda U^{-1}} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{bmatrix}$

Example continued

Let's divide ***U*** (and multiply ***U***$^{-1}$) by $\sqrt{2}$

Then, ***S***= $\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$

$\quad\quad\quad\quad\quad\quad\quad \boldsymbol{Q} \quad\quad\quad\quad\quad\quad \boldsymbol{\Lambda} \quad\quad\quad\quad\quad (\boldsymbol{Q}^{-1} = \boldsymbol{Q}^T)$

Symmetric Eigen-Decomposition

- Let S be a square symmetric matrix
- Theorem: there exists an unique eigen-decomposition

$$S = Q\Lambda Q^T$$

  - Q is an orthogonal matrix: columns of Q are normalized eigenvectors, columns are orthogonal.
  - Everything is real

Exercise

- Examine the symmetric eigen-decomposition, if any, for each of the following matrices:

$$\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \qquad \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

3 Singular Value Decomposition

Singular Value Decomposition

- For an $m \times n$ matrix A of rank r there exists a factorization (Singular Value Decomposition = SVD) as follows:

$$\underbrace{A = U}_{m\times m}\,\underbrace{\Sigma}_{m\times n}\,\underbrace{V^T}_{V \text{ is } n\times n}$$

- The columns of ***U*** are orthogonal eigenvectors of ***AA***^{T}.
- The columns of ***V*** are orthogonal eigenvectors of ***A***^{T}***A***.
- Eigenvalues $\lambda_1 \ldots \lambda_r$ of ***AA***^{T} are the eigenvalues of ***A***^{T}***A***.

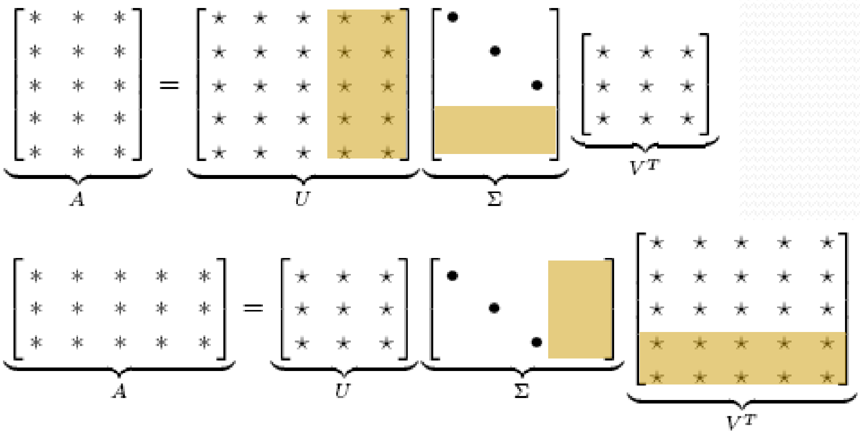
$$\sigma_i = \sqrt{\lambda_i}$$

$$\Sigma = \mathrm{diag}(\sigma_1 \ldots \sigma_r)$$

Singular values $\geq 0.$

Singular Value Decomposition

- Illustration of SVD dimensions and sparseness



SVD example

Let $A = \begin{bmatrix} 1 & -1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$

Thus m=3, n=2. Its SVD is

$$\begin{bmatrix} 0 & 2/\sqrt{6} & 1/\sqrt{3} \\ 1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3} \\ 1/\sqrt{2} & 1/\sqrt{6} & -1/\sqrt{3} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \sqrt{3} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$$

Typically, the singular values arranged in decreasing order.

4 Some applications

SVD and Inverses

- Why is SVD so useful? To invert a matrix!
- Assume A is an invertible matrix
- Compute the SVD: $A = U\Sigma V^T$
- $A^{-1} = (V^T)^{-1}\Sigma^{-1}U^{-1} = V\Sigma^{-1}U^T$
  - Using fact that inverse = transpose for orthogonal matrices
  - Since Σ is diagonal, Σ^{-1} is also diagonal with reciprocals of entries of Σ

SVD and Pseudo-inverses

- Assume that A is a singular square matrix nxn (rank(A)=r<n)
- The inversion fails because some σ_i are zero for $i > r$
- Pseudoinverse: if $\sigma_i = 0$, set $\frac{1}{\sigma_i} = 0$, hence

$$\Sigma^- = \text{diag}\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_r}, 0, \ldots, 0\right)$$

- $A^+ = V\Sigma^- U^T$ is called the pseudo-inverse of A (“closest” matrix to inverse)
  - It is equal to $(A^TA)^{-1}A^T$ if A^TA invertible
  - It satisfies $AA^+A = A$, $A^+AA^+ = A^+$, $(AA^+)^T = AA^+$ and $(A^+A)^T = A^+A$

- Defined for all (even non-square, singular, etc.) matrices

SVD and linear systems

- Assume A is a matrix of size nxm

- Solving $Ax = b$ by least squares: $\hat{x} = \underset{x}{\operatorname{argmax}} \|Ax - b\|_2^2$

  Then $\hat{x} = A^+ b$

- In fact, all the solutions of $Ax = b$ are given by

$$\hat{x} = A^+ b + (I_m - A^+ A) w$$

where w is any vector in R^m

  - Solutions exist if and only if $AA^+ b = b$

Low-rank Approximation

- SVD can be used to compute optimal low-rank approximations.
- Approximation problem: find A_k of rank k such that

$$A_k = \underset{X:rank(X)=k}{\arg\min} \|A - X\|_F$$

Frobenius norm

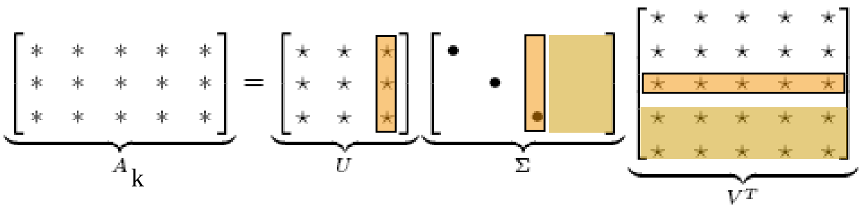
$$\|A\|_F \equiv \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2}.$$

  - A_k and X are both $m \times n$ matrices.
  - Typically, want $k << r$.

Low-rank Approximation

- Solution via SVD

$$A_k = U \operatorname{diag}(\sigma_1, \ldots, \sigma_k, \underbrace{0, \ldots, 0}_{\text{set smallest (r-k) singular values to zero}}) V^T$$



$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

column notation: sum of rank 1 matrices

Approximation error

- How good (bad) is this approximation?
- It's the best possible, measured by the Frobenius norm of the error:

$$\min_{X:rank(X)=k} \left\| A - X \right\|_F = \left\| A - A_k \right\|_F = \sum_{i \geq k+1} \sigma_i$$

where the σ_i are ordered such that $\sigma_i \geq \sigma_{i+1}$.

4 Conclusion

Conclusion

- SVD is a very useful tool
- Very efficient algorithms to compute SVD
- One of the most famous data analysis method
- A huge number of applications (PCA for example)