Math of Data Science: Lecture 3

Vlad Kobzar

APAM, Columbia

Sept. 13, 2022

Course progress

- Last time - diagonalization of square matrices, projections
- Today - singular value decomposition (SVD)

SVD - motivation

- Last time we studied diagonalization (eigendecomposition) of symmetric square matrices

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

- Non-symmetric square matrices
  - can be also diagonalized if they have n linearly independent eigenvectors,

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

  - but eigenvectors may not be orthogonal and the eigenvalues/eigenvectors may be complex-valued
  - To avoid these issues use SVD
- More generally use SVD for $A \in \mathbb{R}^{m\times n}$ of arbitrary dimension

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

where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal and $\Sigma \in \mathbb{R}^{m\times n}$ is "diagonal".

Background - LU factorization

- Previously encountered other factorizations of nonsquare matrices.
- For $A \in \mathbb{R}^{m \times n}$ with $m \leq n$, $Ax = b$ can be solved by LU factorization
  - *Elimination* leads to $Ux = L^{-1}x = c$ where $L \in \mathbb{R}^{m \times m}$ is the lower triangular matrix of multipliers of pivot rows,

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ l_{21} & 1 & 0 & 0 \\ l_{31} & l_{32} & 1 & 0 \\ l_{41} & l_{42} & l_{43} & 1 \end{bmatrix}$$

    and $l_{21} = a_{21}/a_{11}$, $l_{21} = a_{21}/a_{11}$, $l_{41} = a_{41}/a_{11}$, etc., and $U \in \mathbb{R}^{m \times n}$ is an upper triangular matrix of pivot rows.
  - *Backsubstitution* of $Ux = c$ leads to x
  - We have factored $A = LU$
- Not commonly used in practice when the system is underdetermined ($m < n$).
  - Instead use regularization (will study later) to fix a solution

Background - QR factorization

- A full rank $A = QR$ where $Q \in \mathbb{R}^{m \times m}$ is orthogonal and$R \in \mathbb{R}^{m \times m}$ is triangular
- Achieved by orthogonalizing $col(A)$ (*Gram-Schmidt*)
  - $q_1 = a_1$
  - $\hat{q}_i = a_i - \sum_{j=1}^{i-1} \langle q_j, a_i \rangle q_j$
  - $q_i = \hat{q}_i / \|\hat{q}_i\|$
- Therefore, each a_i is a linear combination of $q_1, \ldots, q_{i-1}$, i.e. R is upper triangular
- QR factorization can be generalized to nonsquare matrices
- Commonly used for least squares and related problems (if A is sparse, there are better algorithms) - will also study later

SVD -reduced form

- Another factorization $A = CR$ with rank r
  - The shape of CR is (m by n) = (m by r)(r by n)
  - C with r orthogonal columns, and
  - R with r orthogonal rows
- Normalization leads to the *reduced form* of the SVD

$$A = \begin{bmatrix} u_1 & u_2 & \cdots & u_r \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \cdots \\ v_r \end{bmatrix}$$
$$= U_r \Sigma_r V_r^T$$

- where $C = U_r \sqrt{\Sigma_r}$ and $R = \sqrt{\Sigma_r} V_r^T$, and $\sigma_i > 0$
- If you choose σ_i to be in descending order, then Σ_r is unique (but U and V are not necessarily unique)

Full SVD

- Add the $m-r$ orthogonal vectors that span $C(A)^{\perp}$ as columns to U_r
- Add the $n-r$ orthogonal vectors that span $N(A)$ as columns of V_r
- Add $\sigma_{r+1},...,\sigma_n = 0$ to Σ_r

$$A = \begin{bmatrix} u_1 & u_2 & \cdots & u_m \end{bmatrix} \left[\begin{array}{cccc|c} \sigma_1 & & & & \\ & \sigma_2 & & & 0 \\ & & \ddots & & \\ & & & \sigma_r & \\ \hline & & 0 & & 0 \end{array}\right] \begin{bmatrix} v_1 \\ v_2 \\ \cdots \\ v_n \end{bmatrix}$$

$$= U\Sigma V^T$$

where Σ is a $\mathbb{R}^{m\times n}$ rather than a square $\mathbb{R}^{r\times r}$ matrix.

- For symmetric PSD matrices $U = V$ by the eigendecomposition, so it's a special case of the SVD
- For other symmetric matrices, the SVD generalizes eigendecomposition modulo the sign(s) of σ_i, v_i, u_i.

SVD

- The proof of the SVD existence is constructive and based on the eigendecomposition of symmetric matrices
- $A^T A$ and AA^T which are positive semidefinite and have the same nonzero eigenvalues

$$A^T A = V \Lambda V^T = (V \Sigma U^T)(U \Sigma V^T)$$
$$AA^T = U \Lambda U^T = (U \Sigma V^T)(V \Sigma^T U^T)$$

where $\sigma_k = \sqrt{\lambda_k}$ for $\lambda_k > 0$ and the remaining entries of Σ are zero.

SVD

- By the previous page

$$A^T A = V \Lambda V^T [= (V \Sigma U^T)(U \Sigma V^T)]$$

  where $\sigma_k = \sqrt{\lambda_k}$ for $\lambda_k \neq 0$ and the remaining entries of Σ are zero.
- To determine u_k we require $Av_k = \sigma_k u_k$
- This would imply $AV = \Sigma U$, and therefore the existence of SVD

$$Av_k = \sigma_k u_k \Rightarrow u_k = \frac{Av_k}{\sigma_k}$$

- Add the $m - r$ orthogonal vectors $u_{r+1}, \dots, u_m$ that span $C(A)^\perp$ as columns to U
- And add the $n - r$ orthogonal vectors $v_{r+1}, \dots, v_n$ that span $N(A)$ as columns of V to get the full SVD

SVD

- To confirm that u_k are eigenvectors of AA^T, i.e.,

$$AA^T = U\Lambda U^T = (U\Sigma V^T)(V\Sigma^T U^T)$$

we take

$$AA^T u_k = AA^T \frac{Av_k}{\sigma_k} = A\frac{A^T Av_k}{\sigma_k} = A\frac{\sigma_k^2 v_k}{\sigma_k} = \sigma_k^2 u_k$$

- To confirm that u_k are orthonormal:

$$u_j^T u_k = \left(\frac{Av_j}{\sigma_j}\right)^T \frac{Av_k}{\sigma_k} = \frac{v_j^T (A^T A v_k)}{\sigma_j \sigma_k} = \frac{\sigma_k}{\sigma_j} v_j^T v_k = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{if } j \neq k \end{cases}$$

Geometric interpretation of SVD

- SVD can be represented as rotation x stretching x rotation

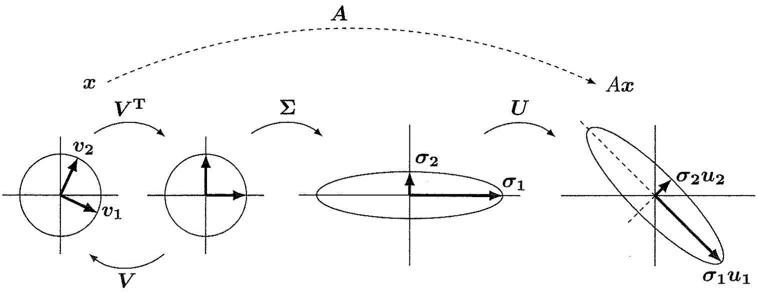

Figure: Fig I.10 from [1]

- V or U can also entail reflections along an $n-1$ dimensional hyperplane (if $\det A < 0$)

SVD and spectral norms

- For any matrix $A \in \mathbb{R}^{m \times n}$ with left singular vectors $u_1, u_2, \ldots, u_r$ corresponding to the nonzero singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$,

$$
\begin{aligned}
\sigma_1 &= \max_{||u||_2 = 1} ||A^T u||_2, \\
u_1 &= \arg \max_{||u||_2 = 1} ||A^T u||_2, \\
\sigma_k &= \max_{\substack{||u||_2 = 1 \\ u \perp u_1, \ldots, u_{k-1}}} ||A^T u||_2, \quad 2 \leq k \leq r, \\
u_k &= \arg \max_{\substack{||u||_2 = 1 \\ u \perp u_1, \ldots, u_{k-1}}} ||A^T u||_2, \quad 2 \leq k \leq r.
\end{aligned}
$$

SVD and spectral norms

- *Soln:* If $A = U\Sigma V^T$ is a reduced form SVD then

$$AA^T = U\Sigma V^T V\Sigma U^T = U\Sigma^2 U^T,$$

where Σ^2 is a diagonal $\mathbb{R}^{r\times r}$ matrix containing $\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq \sigma_r^2$ in its diagonal.
- The result now follows from applying the optimization-based formulation of eigendecomposition we discussed in Lecture 2 to the quadratic form

$$uAA^T u = ||A^T u||_2^2.$$

SVD is the best k-rank approximation

- Unlike other matrix factorizations, SVD has a property that is often exploited in data science applications
- Let $A_k = \sigma_1 u_1 v_1^T + ... + \sigma_k u_k v_k^T$.
- It is the best k-rank approximation of A, i.e.,

$$\|A - A_k\| \leq \|A - B\|$$

for all B with rank k.

SVD is the best k-rank approximation in the spectral norm

- Let's prove this for the spectral, or l^2, norm:

$$\|A\|_2 = \max_{\|x\|=1} \|Ax\| = \sigma_1$$

- Note that $A - A_k = \sigma_{k+1} u_{k+1} v_{k+1}^T + ... + \sigma_r u_r v_r^T$.
- Therefore taking $x = v_{k+1}$, we have $\|A - A_k\| = \sigma_{k+1}$.
- Now we just need to show that

$$\|A - B\| \leq \sigma_{k+1}$$

  for all B with rank k.

SVD is the best k-rank approximation in the spectral norm

- The nullspace of B has dim $\geq n - k$ since B has rank $\leq k$.
- Also $v_1, \ldots, v_{k+1}$ span a $k + 1$ dimensional subspace.
- We have $\geq n - k$ and $k + 1$ dimensional subspaces in an n dimensional space.
- Then by standard linear algebra, $\text{span}(v_1, \ldots, v_{k+1})$ and $N(B)$ must intersect.

SVD is the best k-rank approximation

- Choose nonzero unit norm vector in this intersection

$$x = \sum_{i=1}^{k+1} c_i v_i \in N(B) \cap \text{span}(v_1, \ldots, v_{k+1})$$

- Then since $x \in N(B)$ and $\|x\|^2 = \sum_{i=1}^{k+1} c_i^2 = 1$, we have

$$\|(A - B)x\|^2 = \|Ax\| = \| \sum_{i=1}^{k+1} c_i \sigma_i {v_i}^T \|^2 = \sum_{i=1}^{k+1} c_i^2 \sigma_i^2 \geq \sigma_{k+1}^2$$

for all B with rank k.

Next steps

- Review of probability and optimization
- PCA

References I

[1] Strang, Linear Algebra and Learning from Data, Wellesley Cambridge Press, 2019 2012

[2] Carlos Fernandez-Granda, *DS-GA 1013 / MATH-GA 2821 Mathematical Tools for Data Science, Lecture Notes*, 2020

[3] Carlos Fernandez-Granda, *Probability and Statistics for Data Science, Lecture Notes*, 2017 `https://cims.nyu.edu/~cfgranda/pages/stuff/probability_stats_for_DS.pdf`