CSE 847: Machine Learning— Singular Value Decomposition
Instructor: Jiayu Zhou

1 Singular Value Decomposition

- The singular value decomposition (SVD) is a powerful technique in many applications. Using the SVD of a matrix rather than the original matrix has the advantage of being more robust to numerical errors. Many practical applications exploit key properties of the SVD, i.e., its relation to the rank of a matrix and its optimal low-rank approximation of a given matrix.
  - Deerwester, S., Dumais, S., Landauer, T., Furnas, G. and Harshman, R. (1990) Indexing by Latent Semantic Analysis. Journal of the American Society of Information Science 41(6):391-407.
  - Alter O, Brown PO, Botstein D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci U S A, 97, 10101-6.
  - Liu L., Hawkins D.M., Ghosh S., and Young S.S. (2003) Robust singular value decomposition analysis of microarray data. Proc Natl Acad Sci U S A, 100, 13167-13172.
  - Singular Value Decomposition – A Primer by Sonia Leach
- The SVD decomposes the matrix A by writing it as a product of three matrices.
- Let A be an $m \times n$ matrix, with $m \geq n$. It can be factorized as

$$A = U \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} V^T,$$

  where $U \in \mathrm{I\!R}^{m \times m}$ and $V \in \mathrm{I\!R}^{n \times n}$ are orthogonal, and $\Sigma \in \mathrm{I\!R}^{m \times n}$ is diagonal

$$\Sigma = \mathrm{diag}\left(\sigma_1, \sigma_2, \cdots, \sigma_n\right), \quad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0.$$

  - The quantities σ_i's are called the **singular values** of A and the columns of U and V are called the **left and right singular vectors** of A respectively.
  - The following can be derived from SVD:

$$\begin{aligned} Av_j &= \sigma_j u_j, \\ A^T u_j &= \sigma_j v_j. \end{aligned}$$

  - The columns of U are the eigenvectors of AA^T. The columns of V are the eigenvectors of $A^T A$.
  - If A is symmetric and positive semi-definite, then $A = U \Sigma U^T$, i.e., A has the same left and right singular vectors. Show that if A is symmetric and positive semi-definite, then we can express A as $A = BB^T$ for some matrix B.
  - Assume that the rank of A is r. That is $\sigma_{r+1} = \cdots = \sigma_n = 0$. Partition U and V as follows: $U = [U_1 \; U_2]$ and $V = [V_1 \; V_2]$, where $U_1 \in \mathrm{I\!R}^{m \times r}$, $U_2 \in \mathrm{I\!R}^{m \times (m-r)}$, $V_1 \in \mathrm{I\!R}^{n \times r}$, $V_2 \in \mathrm{I\!R}^{n \times (n-r)}$. What is the null space of A? What is the orthogonal complement of the null space of A? What is the range space of A? What is the orthogonal complement of the range space of A?

- Compute the norm of the matrix A:

$$||A||_2 = \sigma_1, \quad ||A||_F = \sqrt{\sum_{i=1}^{n} \sigma_i^2}.$$

- The trace norm (or nuclear norm) of the matrix A is defined as:

$$||A||_* = \sum_{i=1}^{n} \sigma_i.$$

- The trace norm has become very popular in recent years for matrix completion.
  - ∗ E. J. Candés and T. Tao. The power of convex relaxation: Near-optimal matrix completion. IEEE Trans. Inform. Theory, 56(5), 2053-2080.
  - ∗ E. J. Candés and B. Recht. Exact matrix completion via convex optimization. Found. of Comput. Math., 9 717-772.
- Compact SVD: Only the r column vectors of U and r row vectors of V^T corresponding to the non-zero singular values are calculated.

$$U = (U_1 \ U_2)\,\Sigma\,(V_1 \ V_2)^T = U_1 \Sigma_1 V_1^T$$

  where Σ_1 includes all nonzero singular values.
- An example:

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

- Truncated SVD: Only the t column vectors of U and t row vectors of V^T corresponding to the non-zero singular values are calculated. The rest of the matrix is discarded. This can be much quicker and more economical than the compact SVD if $t < r$.

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

  where U_t and V_t consist of the first t columns of U and V, respectively.
- The truncated SVD is no longer an exact decomposition of the original matrix A. We will show that the truncated SVD produces the closest approximation to A that can be achieved by a matrix of rank t.
- Outer product form:

$$\begin{aligned} A &= U_1 \Sigma_1 V_1^T = (u_1, \cdots, u_r) \begin{pmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \cdots & \\ & & & \sigma_r \end{pmatrix} \begin{pmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_r^T \end{pmatrix} \\ &= (u_1, \cdots, u_r) \begin{pmatrix} \sigma_1 v_1^T \\ \sigma_2 v_2^T \\ \vdots \\ \sigma_r v_r^T \end{pmatrix} = \sum_{i=1}^{r} \sigma_i u_i v_i^T. \end{aligned}$$

Each term $\sigma_i u_i v_i^T$ in the sum is a rank one matrix. This is thus a sum of rank one matrices.

– Solve least squares problem using SVD: $\min_{x \in {\rm I\!R}^n} ||Ax - b||_2^2$.

  $*$ Assume the SVD of A is given by

$$A = (U_1 \;\; U_2) \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} V^T,$$

  where $\Sigma \in {\rm I\!R}^{n\times n}$ is diagonal, $U_1 \in {\rm I\!R}^{m\times n}$ and $U_2 \in {\rm I\!R}^{m\times(m-n)}$ have orthonormal columns and $V \in {\rm I\!R}^{n\times n}$ is orthogonal.

$$\begin{aligned} ||Ax - b||_2^2 &= \left\| (U_1 \;\; U_2) \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} V^T x - b \right\|_2^2 = \left\| \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} V^T x - (U_1 \;\; U_2)^T b \right\|_2^2 \\ &= \left\| \begin{pmatrix} \Sigma y \\ 0 \end{pmatrix} - \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \right\|_2^2 = \|\Sigma y - b_1\|_2^2 + \|b_2\|_2^2, \end{aligned}$$

  where $y = V^T x$, $b_1 = U_1^T b$, and $b_2 = U_2^T b$. The optimal y is given by $y = \Sigma^{-1} b_1$. Thus, the least squares solution is given by

$$x = Vy = V\Sigma^{-1} b_1 = V\Sigma^{-1} U_1^T b = \sum_{i=1}^{n} \frac{u_i^T b}{\sigma_i} v_i.$$

2 Existence of SVD

Denote $\sigma_1 = \|A\|_2$. We know the largest eigenvalue of $A^T A$ is σ_1^2. Let x be the corresponding (normalized) eigenvector, i.e., $\|x\|_2 = 1$. Define a vector y as follows:

$$y = \frac{1}{\sigma_1} Ax.$$

It is easy to verify that $\|y\|_2 = \sqrt{\frac{1}{\sigma_1^2} x^T A^T A x} = 1$. Thus, y is also normalized. From x and y, we can construct two orthogonal matrices as follows:

$$Z_1 = (y, Z_2) \in {\rm I\!R}^{m\times m}, \quad W_1 = (x, W_2) \in {\rm I\!R}^{n\times n},$$

where $Z_2 \in {\rm I\!R}^{(m-1)\times m}$ with all its columns orthogonal to y, and $W_2 \in {\rm I\!R}^{(n-1)\times n}$ with all its columns orthogonal to x. Then, we have

$$Z_1^T A W_1 = \begin{pmatrix} y^T \\ Z_2^T \end{pmatrix} A\,(x, W_2) = \begin{pmatrix} y^T A x & y^T A W_2 \\ Z_2^T A x & Z_2^T A W_2 \end{pmatrix}.$$

In addition, we have

$$\begin{aligned} y^T A x &= \frac{1}{\sigma_1} x^T A^T A x = \sigma_1, \\ Z_2^T A x &= Z_2^T \sigma_1 y = \sigma_1 Z_2^T y = 0, \\ y^T A W_2 &= \frac{1}{\sigma_1} x^T A^T A W_2 = \frac{1}{\sigma_1} \sigma_1^2 x^T W_2 = \sigma_1 x^T W_2 = 0. \end{aligned}$$

Combing all of these, we have

$$Z_1^T A W_1 = \left(\begin{array}{c} y^T \\ Z_2^T \end{array} \right) A\,(x, W_2) = \left(\begin{array}{cc} \sigma_1 & 0 \\ 0 & Z_2^T A W_2 \end{array} \right).$$

Thus, after one operation (a left multiplication of an orthogonal matrix Z_1^T and a right multiplication of an orthogonal matrix W_1), we get a block diagonal form (the first block is of size 1). We can repeat the same procedure multiple times to obtain a diagonal form. That is, we can obtain two orthogonal matrices U and V such that $U^T AV$ is diagonal with nonnegative diagonal entries. Thus, we obtain an SVD for A.

3 Matrix Approximation

Theorem 3.1 *Let $U_k = (u_1, \cdots, u_k)$, $V_k = (v_1, \cdots, v_k)$, and $\Sigma_k = diag(\sigma_1, \sigma_2, \cdots, \sigma_k)$. Define*

$$A_k = U_k \Sigma_k V_k^T.$$

Then

$$\min_{B: rank(B) \le k} ||A - B||_F = ||A - A_k||_F = \sqrt{\sum_{i=k+1}^{n} \sigma_i^2}.$$

- A_k is the best approximation of rank k for the matrix A.
- This low rank approximation is useful for
  - Compression
  - Noise reduction
  - finding “concepts” or “topics” (text mining/LSI)
  - data exploration and visualizing data
  - classification (e.g. handwritten digits)
- SVD appears under different names:
  - Principal Component Analysis (PCA)
  - Latent Semantic Indexing (LSI)/Latent Semantic Analysis (LSA)
  - Karhunen-Loeve expansion/Hotelling transform (in image processing)

3.1 Outline of the Proof

- We can first simplify the problem by transforming A into a diagonal matrix. Why does the following hold?

  $$||A - B||_F^2 = ||U \Sigma V^T - B||_F^2 = ||\Sigma - U^T BV||_F^2$$

  Denote $\tilde{B} = U^T BV$. The only constraint on $\tilde{B}$ is that rank$(\tilde{B}) = k$ (since U and V are orthogonal matrices, they are nonsingular).

- Since Σ is diagonal, rank$(\tilde{B})$ should be diagonal in order to minimize the Frobenius error. Denote rank$(\tilde{B}) = S$, for some diagonal matrix. Thus, $B = USV^T$. Note that the entries in S may not be ordered as in SVD.

- Denote $S = \text{diag}(s_1, s_2, \cdots, s_n)$. Then we have:

$$||A - B||_F^2 = \sum_{i=1}^{n} (\sigma_i - s_i)^2.$$

  Since only k diagonal entries of S are nonzero, the optimal S minimizing the Frobenius error is given by $s_i = \sigma_i$, for $i = 1, \cdots, k$, and $s_i = 0$ for $i > k$.

3.2 A Rigorous Proof

Let $B \in \mathbb{R}^{m \times n}$ be of rank k. Thus, the dimension of the null space $\text{Null}(B)$ is $n - k$. Consider the subspace V_k spanned by $\{v_1, v_2, \cdots, v_{k+1}\} \in \mathbb{R}^n$. Since $(n - k) + (k + 1) > n$, $\text{Null}(B)$ and V_k must intersect. Let z be a unit vector in the intersection. Thus, $Bz = 0$. In addition, assume $z = [v_1, v_2, \cdots, v_{k+1}]\alpha$ for some vector $\alpha \in \mathbb{R}^{k+1}$. We can show that α is of unit length:

$$1 = \|z\|_2^2 = \alpha^T [v_1, v_2, \cdots, v_{k+1}]^T [v_1, v_2, \cdots, v_{k+1}]\alpha = \alpha^T \alpha = \|\alpha\|_2^2.$$

Thus,

$$\|A - B\|_2^2 \geq \|(A - B)z\|_2^2 = \|Az\|_2^2 = \|U\Sigma V^T z\|_2^2 = \|\Sigma_{k+1}\alpha\|_2^2 \geq \sigma_{k+1}^2,$$

where Σ_{k+1} is the $(k + 1)$-th principal sub-matrix of Σ.

It is clear that the lower bound can be achieved by choosing $B = A_k$. This completes the proof.

- Let $B \in \mathbb{R}^{m \times n}$ be of rank k. Thus, the dimension of the null space $\text{Null}(B)$ is $n - k$.
- Consider the subspace V_k spanned by $\{v_1, v_2, \cdots, v_{k+1}\} \in \mathbb{R}^n$.
- Since $(n - k) + (k + 1) > n$, $\text{Null}(B)$ and V_k must intersect.
- Let z be a unit vector in the intersection. Thus, $Bz = 0$.
- In addition, assume $z = [v_1, v_2, \cdots, v_{k+1}]\alpha$ for some vector $\alpha \in \mathbb{R}^{k+1}$.
- We can show that α is of unit length:

$$1 = \|z\|_2^2 = \alpha^T [v_1, v_2, \cdots, v_{k+1}]^T [v_1, v_2, \cdots, v_{k+1}]\alpha = \alpha^T \alpha = \|\alpha\|_2^2.$$

  Thus,

$$\|A - B\|_2^2 \geq \|(A - B)z\|_2^2 = \|Az\|_2^2 = \|U\Sigma V^T z\|_2^2 = \|\Sigma_{k+1}\alpha\|_2^2 \geq \sigma_{k+1}^2,$$

  where Σ_{k+1} is the $(k + 1)$-th principal sub-matrix of Σ.

- It is clear that the lower bound can be achieved by choosing $B = A_k$.

4 How to Compute SVD?

- The most commonly used SVD algorithm is found in Matlab and in the LAPACK linear algebra library (http://www.netlib.org/lapack/). It is a revised version of one that appeared in Golub and Van Loan.

  – J. Demmel, W. Kahan, "Accurate Singular Values of Bidiagonal Matrices," S IAM J. Sci. Stat. Comput., 11(1990) pp. 873-912.

  - Available at http://www.netlib.org/lapack/lawnspdf/lawn03.pdf

- The SVD of a matrix A is typically computed by a two-step procedure.
  - In the first step, the matrix is reduced to a bidiagonal matrix. This takes $O(mn^2)$ floating-point operations, if we assume that $m \geq n$.
  - In the second step, we compute the SVD of the bidiagonal matrix. This step can only be done with an iterative method (as with eigenvalue algorithms). The second step takes $O(n)$ iterations, each costing $O(n)$ flops. Thus, the first step dominates the computation, and the overall cost is $O(mn^2)$ flops (Trefethen & Bau III 1997, Lecture 31).
- The first step can be done using Householder transformations. How?
- The second step can be done by a variant of the QR algorithm (Golub & Kahan 1965). The LAPACK subroutine DBDSQR implements this iterative method, with some modifications to cover the tricky case where the singular values are very small (Demmel & Kahan 1990).