A Brief Discussion on Singular Value Decomposition

Anqiao Ouyang

September 18, 2024

1 Supplementary Background Knowledge

To read this article, you should have a basic understanding of elementary algebra and linear algebra.

Due to time constraints, this is a brief introduction to fundamental definitions. Detailed applications will be covered later in the ML column when there is more time.

1.1 Unitary Matrices

We know that transposing a matrix is equivalent to flipping it along the diagonal, converting an $m \times n$ matrix M into an $n \times m$ matrix, denoted as $M^\top$. The conjugate transpose involves taking the transpose while also taking the complex conjugate of the elements in the matrix. For a matrix that takes the complex conjugate, we denote it as M^*, and the conjugate transpose of a matrix M is denoted as M^H (often written as $M^\dagger$ in quantum mechanics, but here in linear algebra).

A matrix M is a unitary matrix if and only if its conjugate transpose is equal to its inverse, i.e., $M^H = M^{-1}$. Its mathematical definition is $M^H M = MM^H = I_n$. In fact, unitary matrices can be understood as a generalization of orthogonal matrices to the complex domain.

Some obvious properties include, for unitary matrices M, N of order n: + The modulus of the determinant of M is 1 + All eigenvalues of a unitary matrix have a modulus of 1 + MN is also a unitary matrix + For a norm $\|\cdot\|$, for any matrix A, if $\|A\| = \|MAN\|$, then we call $\|\cdot\|$ a **unitary invariant norm**. Examples include the L2 norm and Frobenius norm.

For two complex square matrices P, Q, if there exists a unitary matrix M such that

$$P = MQM^H$$

then we say P and Q are **unitary equivalent**. For real square matrices P, Q, this is specifically referred to as **orthogonal equivalence**.

The set of unitary matrices forms a group over the complex domain, known as the **unitary group**.

1.1.1 Special Orthogonal Group

Consider a commutative ring R with $2^{-1} \in R$. Denote by $M_n(R)$ the set of all $n \times n$ matrices with elements in R. For any $n \times n$ matrix Q, the **orthogonal group** $O(n, R)$ is defined as

$$O(n, R) = \{W \in M_n(R) | Q^\top Q = I_n\}$$

equipped with matrix multiplication to form a group. The **special orthogonal group** is the normal subgroup of the orthogonal group consisting of elements with determinant 1, denoted as $SO(n, R)$. In the complex domain, this is referred to as the **special unitary group**.

The set of all orthogonal matrices over the real numbers $\mathbb{R}^{n\times n}$ forms a special orthogonal group $SO(n, \mathbb{R}^{n\times n})$; all unitary matrices over the complex numbers $\mathbb{C}^{n\times n}$ form a special unitary group $SU(n, \mathbb{C}^{n\times n})$. All unitary matrices over the complex numbers $\mathbb{C}^{n\times n}$ form a special orthogonal group $SO(n, \mathbb{C}^{n\times n})$.

1.2 Eigenvalues

For an $n \times n$ matrix M, if there exists a nonzero vector v and a scalar λ such that

$$Mv = \lambda v$$

then v is called an eigenvector, and λ is its eigenvalue. The set of all eigenvectors corresponding to the same eigenvalue, along with the zero vector, forms a linear space called the eigenspace.

For matrix M, we seek a set of eigenvectors that are orthogonal to each other (their inner product is 0), and these eigenvectors are unit vectors. When the matrix acts on this set of orthonormal vectors, the length of each eigenvector may change, but their orthogonality remains unchanged.

Diagonalization refers to the process of transforming a matrix into a diagonal matrix. For M, if there exists an invertible orthogonal matrix P such that

$$D = P^{-1}AP$$

then the matrix is diagonalizable. The prerequisite for diagonalization is that the original matrix must have n linearly independent eigenvectors, and the eigenvectors corresponding to each eigenvalue λ are the column vectors of matrix P.

1.3 Singular Values

Definition 1.1 (Singular Values)**.** Consider an $m\times n$ matrix M. We can construct two symmetric matrices:

$$A_1 = MM^\top \quad \text{and} \quad A_2 = M^\top M,$$

both of which have the same nonzero eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$, where $r = \text{rank}(M)$.

The *singular values* of M are defined as the square roots of these eigenvalues:

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

1.4 Spectral Theorem for Finite Dimensions

In general, the spectral theorem provides conditions under which an operator or matrix can be diagonalized. However, in linear algebra, we mainly study finite-dimensional linear spaces, while the spectral theorem is more broadly applicable to self-adjoint operators in infinite-dimensional spaces. Therefore, this article only discusses the finite-dimensional case; topics such as Hilbert spaces will be reserved for future functional analysis articles.

We call a matrix M **normal** if it satisfies $MM^H = M^H M$.

A **self-adjoint matrix**, or Hermitian matrix, is a conjugate symmetric matrix (an equivalent condition is $H^H = H$; for real matrices, this is equivalent to $H^\top = H$). It is easy to see that it is a special type of normal matrix. Every **normal matrix** H can be diagonalized by a unitary matrix, i.e., there exists a unitary matrix M and a diagonal matrix N such that

$$H = MNM^H$$

However, the eigenvalues of a normal matrix can be real or complex. Self-adjoint matrices H have real eigenvalues, meaning that the eigenvectors of a self-adjoint matrix can be chosen to be orthonormal. Thus, there exists a **standard orthonormal basis** V consisting of the eigenvectors of H.

2 Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD) is an important matrix decomposition method in linear algebra. It is commonly used in PCA for feature extraction, dimensionality reduction, and improving model generalization.

2.1 General Description

We decompose an $m \times n$ matrix M with elements in $\mathbb{K}$ (i.e., real or complex numbers) into three matrices:

$$M = U\Sigma V^H$$

where:

- U: An $m \times m$ orthogonal matrix (i.e., unitary matrix in the case of complex numbers),
- Σ: An $m \times n$ diagonal matrix with non-negative real numbers, where each element on the main diagonal is called a singular value, equal to the square root of the positive eigenvalues of MM^H or $M^H M$ (note that these are real numbers, so they are non-negative square roots),
- V^H: The conjugate transpose of an $n \times n$ unitary matrix V.

This decomposition is called **Singular Value Decomposition**. We typically refer to the columns of U as left singular vectors, the columns of V as right singular vectors, and the diagonal elements of Σ, σ_i, as the singular values of the matrix. Since V is orthogonal, we can also write the decomposition in another form:

$$MV = U\Sigma$$

Additionally, it is important to note that the singular value decomposition of a matrix is not unique, but it is often convenient to order the singular values from largest to smallest.

When a square matrix can be diagonalized, it means there exists a basis where its action can be represented as the action of a diagonal matrix. Geometrically, we can understand Singular Value Decomposition as transforming vectors from the original space to a new basis (where this basis consists of the eigenvectors of $M^H M$); scaling the vectors in the new basis, with each direction's scaling factor determined by the singular values; and transforming the scaled vectors back to the target space using a new basis (where this basis consists of the eigenvectors of MM^H). Therefore, Singular Value Decomposition can be seen as a generalization of diagonalization.

2.2 Orthogonality and Extension

The previous definition might seem a bit abrupt, so let us elaborate. Continuing from the previous definition, let u_i and v_i denote the columns of U and V, respectively, and let $r = \text{rank}(M)$. Singular Value Decomposition (SVD) provides orthogonal bases for four subspaces of the matrix M:

- $u_1, u_2, \ldots, u_r$ form an orthogonal basis for the column space of M,
- $v_1, v_2, \ldots, v_r$ form an orthogonal basis for the row space of M,
- $u_{r+1}, \ldots, u_m$ form an orthogonal basis for the left null space of M,
- $v_{r+1}, \ldots, v_n$ form an orthogonal basis for the null space of M.

Note that for $u_{m-r}, \ldots, u_m$ and $v_{n-r}, \ldots, v_n$, since the null space and left null space are orthogonal to the row and column spaces, these vectors are orthogonal to the first r basis vectors. In constructing the orthogonal matrices U and V, these vectors are included to ensure they are square matrices, making the equation $MV = U\Sigma$ valid.

2.3 Optimal Low-Rank Approximation

Optimal Low-Rank Approximation refers to finding a rank-k matrix that minimizes the error with respect to a given matrix and a specified rank k under some norm.

2.3.1 Eckart–Young–Mirsky Theorem

Low-rank approximation of a matrix is achieved by retaining the largest k singular values and their corresponding singular vectors. For a given $k < r$, the best rank-k approximation matrix M_k of matrix M can be expressed as

$$M_k = \sum_{i=1}^{k} \sigma_i u_i v_i^*$$

This approximation is optimal in minimizing the Frobenius norm error, i.e.,

$$\|M - M_k\|_F = \min_{\text{rank}(N) \leq k} \|M - N\|_F$$

The Eckart–Young–Mirsky theorem states that for any positive integer k (where $k \leq \min(m, n)$), there exists a rank-k matrix M_k such that the Frobenius norm $\|M - M_k\|_F$ is the minimum among all rank-k matrices N with respect to the Frobenius norm to M.

2.3.2 A Greedy Algorithm

Here's a brief overview of a greedy algorithm approach.

For an $m \times n$ matrix M, where each row vector m_i can be viewed as a point in an n-dimensional space, we choose a unit vector v to represent the direction of a line through the origin. The projection length of each m_i in the direction of v is $|m_i \cdot v|$, and we seek the line that maximizes the sum of squared projection lengths $|Mv|^2$, i.e., the line that minimizes the squared distance of each corresponding point.

Thus, the first singular vector is defined as

$$v_1 = \arg \max_{\|v\|=1} |Mv|$$

The corresponding first singular value is

$$\sigma_1 = \max_{\|v\|=1} |Mv_1|$$

Since we are using the sum of squared projection lengths, any 2-dimensional subspace containing v_1 will have a projection length squared equal to the sum of the squared projection in the v_1 direction plus the squared length in the direction orthogonal to v_1. Similarly,

$$v_2 = \arg \max_{\|v\|=1, v \perp v_1} |Mv|$$

Similarly, finding the k-th singular vector v_k involves maximizing $|Mv|^2$ among all unit vectors orthogonal to previously found singular vectors

$$v_k = \arg \max_{\|v\|=1, v \perp v_1, v_2, \ldots, v_{k-1}} |Mv|$$

The corresponding singular value is

$$\sigma_k = \max_{\|v\|=1, v \perp v_1, v_2, \ldots, v_{k-1}} |Mv_k|$$