Matrix Decompositions

April 20, 2023

1 Matrix Approximation

Partial SVD can also help us to represent $\boldsymbol{A}$ as a sum of simpler (low-rank) matrices $\boldsymbol{A}_i$, which lends itself to a matrix approximation scheme that is cheaper to compute than the full SVD.

We construct a rank-1 matrix $\boldsymbol{A}_i \in \mathbb{R}^{m \times n}$ as

$$\boldsymbol{A}_i := \boldsymbol{u}_i \boldsymbol{v}_i^T \tag{1}$$

which is formed by the outer product of the i-th orthogonal column vector of $\boldsymbol{U}$ and $\boldsymbol{V}$.

A matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ of rank r can be written as a sum of rank-1 matrices $\boldsymbol{A}_i$ so that

$$\boldsymbol{A} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T = \sum_{i=1}^{r} \sigma_i \boldsymbol{A}_i \tag{2}$$

where the outer-product matrices $\boldsymbol{A}_i$ are weighted by the i-th singular value σ_i. The diagonal structure of the singular value matrix $\boldsymbol{\Sigma}$ multiplies only matching left- and right-singular vectors $\boldsymbol{u}_i \boldsymbol{v}_i^T$ and scales them by the corresponding singular value σ_i. All terms $\sum_{ij} \boldsymbol{u}_i \boldsymbol{v}_i$ vanish for $i \neq j$ because $\boldsymbol{\Sigma}$ is a diagonal matrix. Any terms $i > r$ vanish because the corresponding singular values are 0.

We can sum up r different rank-1 matrices to obtain a rank-r matrix $\boldsymbol{A}$. If the sum does not run over all matrices $\boldsymbol{A}_i, i = 1, \ldots, r$, but only up to an intermediate $k < r$, we obtain a rank-k approximation

$$\hat{\boldsymbol{A}} := \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T = \sum_{i=1}^{k} \sigma_i \boldsymbol{A}_i \tag{3}$$

of $\boldsymbol{A}$ with $rk(\hat{\boldsymbol{A}}(k)) = k$.
To measure the difference (error) between $\boldsymbol{A}$ and its rank-k approximation $\hat{\boldsymbol{A}}(k)$, we need the notion of a norm.

Definition - Spectral Norm of a Matrix
For $\boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$, the *spectral norm* of a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is defined as

$$\|\boldsymbol{A}\|_2 := \max_x \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \tag{4}$$

The spectral norm determines how long any vector $\boldsymbol{x}$ can at most become when multiplied by $\boldsymbol{A}$.

Theorem
The spectral norm of $\boldsymbol{A}$ is its largest singular value σ_1.

Theorem - Eckhart-Young Theorem
Consider a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ of rank r and let $\boldsymbol{B} \in \mathbb{R}^{m \times n}$ be a matrix of rank k. For any $k \leqslant r$ with $\hat{\boldsymbol{A}}(k) = \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$ it holds that

$$\begin{aligned} \hat{\boldsymbol{A}}(k) &= argmin_{rk(\boldsymbol{B})=k} \|\boldsymbol{A} - \boldsymbol{B}\|_2 \\ \|\boldsymbol{A} - \hat{\boldsymbol{A}}(k)\|_2 &= \sigma_{k+1} \end{aligned} \tag{5}$$

The Eckhart-Young theorem states explicitly how much error we introduce by approximating $\boldsymbol{A}$ using rank-k approximation. We can interpret the rank-k approximation obtained with the SVD as a projection of the full-rank matrix $\boldsymbol{A}$ onto a lower-dimensional space of rank-at-most-k matrices. Of all possible projections, the SVD minimizes the error (with respect to the spectral norm) between $\boldsymbol{A}$ and any rank-k approximation.
We observer that the difference between $\boldsymbol{A} - \hat{\boldsymbol{A}}(k)$ is a matrix containing the sum of the remaining rank-1 matrices

$$\boldsymbol{A} - \hat{\boldsymbol{A}}(k) = \sum_{i=k+1}^{T} \sigma_i \boldsymbol{u}_i \boldsymbol{v}^T \tag{6}$$

We immediately obatin σ_{k+1} as the spectral norm of the difference matrix. If we assume that there is another matrix $\boldsymbol{B}$ with $rk(\boldsymbol{B}) \leqslant k$, such that

$$\|\boldsymbol{A} - \boldsymbol{B}\|_2 < \|\boldsymbol{A} - \hat{\boldsymbol{A}}(k)\|_2 \tag{7}$$

then there exists an at least $(n-k)$-dimensional null space $Z \subseteq \mathbb{R}^n$, such that $\boldsymbol{x} \in Z$ implies that $\boldsymbol{B}\boldsymbol{x} = \boldsymbol{0}$. Then it follows that

$$\|\boldsymbol{A}\boldsymbol{x}\|_2 = \|(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{x}\|_2 \tag{8}$$

and by using a version of the Cauchy-Schwartz inequality that encompasses norms of matrices, we obtain

$$\|\boldsymbol{A}\boldsymbol{x}\|_2 \leqslant \|\boldsymbol{A} - \boldsymbol{B}\|_2 \|\boldsymbol{x}\|_2 < \sigma_{k+1} \|\boldsymbol{x}\|_2 \tag{9}$$

However, there exists a $(k+1)$-dimensional subspace where $\|\boldsymbol{A}\boldsymbol{x}\|_2 \geqslant \sigma_{k+1} \|\boldsymbol{x}\|_2$, which is spanned by the right-singular vectors $\boldsymbol{v}_j, j \leqslant k+1$ of $\boldsymbol{A}$. Adding up dimensions of these two spaces yields a number greater than n, as there must be a nonzero vector in both spaces. This is a contradiction of the rank-nullity theorem.
The Eckart-Young theorem implies that we can use SVD to reduce a rank-r matrix $\boldsymbol{A}$ to a rank-k matrix $\hat{\boldsymbol{A}}$ in a principled, optimal (in the spectral norm sense) manner. We can interpret the approximation of $\boldsymbol{A}$ by a rank-k matrix as a form of lossy compression. Therefore, the low-rank approximation of a matrix appears in many machine learning applications, e.g., image processing, noise filtering, and regularization of ill-posed problems. Furthermore, it plays a key role in dimensionality reduction and principal component analysis.