EE445 Mod3-Lec2: SVD & Low Rank Approximation

References:

- [CE-OptMod]: Chapter: 5

Outline

1. **M3-L1**: Review Eigenvalues & Eigenvectors
2. **M3-L1**: Symmetric Matrices
3. **M3-L2** (this lecture): Singular value decomposition `SVD` & low rank approximation

Overview

- We just talked about special classes of matrices that have a nice decomposition in terms of their eigenvalues—namely, symmetric PSD matrices.
- Now, we will talk about a matrix decomposition that every matrix has—i.e., `SVD`
- And, it is fundamentally related to a key ML analysis tool: `PCA`

Matrix Decomposition

- Matrix decomposition, also known as matrix factorization, involves describing a given matrix using its constituent elements.
- Recall that you saw QR decomposition in **Module 1** and then its use in **Module 2** (e.g., solving least squares, in particular sparse problems)
- Perhaps the most known and widely used matrix decomposition method is the **Singular-Value Decomposition**, or `SVD`.
- All matrices have an `SVD`, which makes it more stable than other methods, such as the eigen-decomposition.
- We will see the `SVD` is useful for computing the pseudoinverse efficiently and for dimensionality reduction

Singular Value Decomposition

What is SVD?

- One can generalize eigenvalues/vectors to non-square matrices, in which case they are called singular vectors and singular values.
- The `SVD` is a unique matrix decomposition that exists for every matrix $A \in \mathbb{R}^{m\times n}$:

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

  where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are *unitary* matrices, and $\Sigma \in \mathbb{R}^{m\times n}$ is a matrix with non-negative entries on the diagonal and zeros on the off diagonal.
- Unitary: $UU^\top = I$ and $VV^\top = I$

SVD as a Dyadic Exanpsion

An equivalent way to express the SVD $A = U\Sigma V^\top$ is as a dydic expansion:

$$A = \sum_{i=1}^{\min\{m,n\}} \sigma_i \cdot u_i v_i^\top$$

(i.e., weighted sum of dyads)

- That is, the SVD expresses A as a nonnegative linear combination of $\min\{m, n\}$ rank-1 matrices
- the singular values provide the multipliers
- the outer products of the left and right singular vectors provide the rank-1 matrices.

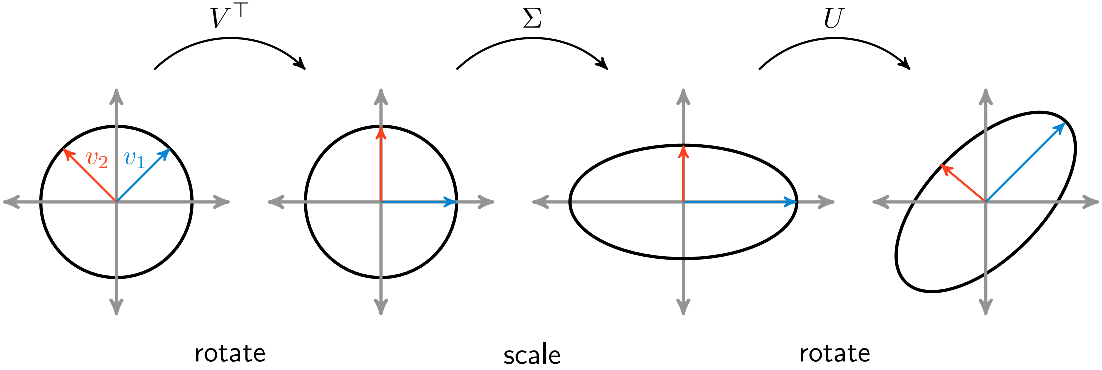
SVD

- The diagonal entries of Σ are called the singular values of A
- The column vectors of V are called the right singular vectors of A
- The column vectors of U are called the left singular vectors of A.
- The number of nonzero singular values is equal to the rank of the matrix A.

$$\underbrace{A}_{m \times n} = \underbrace{U}_{m \times m} \underbrace{\Sigma}_{m \times n} \underbrace{V^\top}_{n \times n}$$

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_n \\ 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 \end{bmatrix}$$

Geometric View of SVD



Unpacking the SVD

- Let $A \in \mathbb{R}^{m\times n}$
- **Fact 1**. Both $A^\top A \in \mathbb{R}^{n\times n}$ and $AA^\top \in \mathbb{R}^{m\times m}$ are symmetric square matrices:

$$(A^\top A)^\top = A^\top (A^\top)^\top = A^\top A \quad \text{and} \quad (AA^\top)^\top = (A^\top)^\top A^\top = AA^\top$$

- **Fact 2**. Both $A^\top A$ and $AA^\top$ share the same non-zero eigenvalues: let (λ, v) be an eigenvalue-eigenvector pair for $A^\top A$ so that

$$A^\top A v = \lambda v \implies AA^\top \underbrace{Av}_{u} = \lambda A v \implies (\lambda, u) \text{ is eigenpair of } AA^\top$$

Unpacking the SVD

- According to the othogonally diagonalizable property of symmetric matrices, the matrices $A^\top A$ and $AA^\top$ can be decomposed as following:

$$A^\top A = V\Lambda V^\top \quad \text{and} \quad AA^\top = U\Lambda U^\top$$

- Indeed, if $A = U\Sigma V^\top$ then

$$A^\top A = (U\Sigma V^\top)^\top (U\Sigma V) = V\Sigma^\top U^\top U\Sigma V^\top = V\Sigma^\top \Sigma V^\top$$

- Can compute by diagonalizing the PSD symmetric matrices $A^\top A$ and $AA^\top$

Using the SVD to Compute Pseudo Inverses

- It turns out that using the SVD we have a very easy way to compute the pseudo-inverse of A—i.e., $A^\dagger = (A^\top A)^{-1} A^\top$ which we saw in **Mod1 & Mod2**
- Indeed

$$A^\dagger = (V\Sigma^\top \Sigma V^\top)^{-1} V\Sigma U^\top = V(\Sigma^\top\Sigma)^{-1} V^\top V \Sigma U^\top = V(\Sigma^\top \Sigma)^{-1}\Sigma U^\top$$

since Σ is a diagonal matrix, its pseudo-inverse is just a diagonal matrix with the reciprocals of the nonzero elements on the diagonal.

Matrix Norms and Connections to Singular values

- Matrix norms and singular values have special relationships.
- **Forbenius Norm**:

$$\|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{1/2} = (\mathsf{Tr}(A^\top A))^{1/2}$$

- **Matrix p-norm**: matrix p-Norm is defined as the largest scalar that you can get for a unit vector

$$\|A\|_p = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} = \max_{\|x\|_p = 1} \|Ax\|_p$$

- Aside: supremum $\sup(\cdot)$ is the “ least upper bound" of its argument

Spectral Radius

- **Definition (Spectral Radius)**: The spectral radius $\rho(A)$ is the maximum modulus of the eigenvalues of A—i.e., $\rho(A) = \max_{i=1,\ldots,n} |\lambda_i(A)|$.
- It is not an induced norm (since $\rho(A) = 0$ does not imply $A = 0$), however we do have the property that $\rho(A) \leq \|A\|_p$ for any p.
- Indeed, letting (λ_i, v_i) where $v_i \neq 0$ be an eigenpair of A, we have

$$\|A\|_p \|v_i\|_p \geq \|Av_i\|_p = \|\lambda_i v_i\|_p = |\lambda_i| \|v_i\|_p \implies |\lambda_i| \leq \|A\|_p \ \forall i$$

Common Matrix Norms

Other norms of interest include the 1-norm and ∞-norms.

- 1-**norm** (ℓ_1): consider $x \in \mathbb{R}^n$. The ℓ_1-norm is given by $\|x\|_1 = \sum_{i=1}^n |x_i|$
- ∞-**norm** (ℓ_∞): consider $x \in \mathbb{R}^n$. The ℓ_∞-norm is given by $\|x\|_\infty = \max_{i=1,\ldots,n} |x_i|$

We can define induced norms from these ℓ_p norms:

$$\|A\|_1 = \max_{\|x\|_1=1} \|Ax\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^m |a_{ij}| \quad \text{i.e., the max column sum}$$

$$\|A\|_\infty = \max_{\|x\|_\infty=1} \|Ax\|_\infty = \max_{i=1,\ldots,m} \sum_{j=1}^n |a_{ij}| \quad \text{i.e., the max row sum,}$$

Matrix Norms: Spectral Norm

- **Spectral Norm (Matrix 2-norm)**: Largest singular value of the matrix $\sigma_1(A)$

$$\begin{aligned}\max_{\|x\|_2=1}\|Ax\|_2 = \max_{\|x\|_2=1}(x^\top A^\top Ax)^{1/2} &= \max_{\|x\|_2=1}(x^\top V\Sigma^2 \underbrace{V^\top x}_{=:y})^{1/2}\\ &= \max_{\|y\|_2=1}(y^\top \Sigma^2 y)^{1/2} = \sigma_1(A)\end{aligned}$$

  where in the last equality we choose x to be the eigenvector of $A^\top A$ corresponding to the largest eigenvalue.

- Aside: singular values are the square roots of the eigenvalues of $A^\top A$
- One can also show that $\|A\|_F = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2(A)}$ using the fact that $\|A\|_F = \sqrt{\mathsf{Tr}(A^\top A)}$

Reduced SVD & Low Rank Approximation

- Rank of Λ is $r \implies$ there are r non-zero eigenvalues of the matrices $A^\top A$ and $AA^\top$
- Reduced SVD:

$$\underbrace{A}_{m\times n} = \underbrace{U_r}_{m\times r}\underbrace{\Sigma_r}_{r\times r}\underbrace{V_r^\top}_{r\times n}$$

Low Rank Structure

$$\underbrace{A}_{m \times n} = \underbrace{Y}_{m \times r} \underbrace{Z^\top}_{r \times n}$$

Low Rank Structure

$$A = uv^\top = \begin{bmatrix} — & u_1 v^\top & — \\ — & u_2 v^\top & — \\ & \vdots & \\ — & u_m v^\top & — \end{bmatrix} = \begin{bmatrix} | & & | \\ v_1 u & \cdots & v_n u \\ | & & | \end{bmatrix}$$

$$A = uv^\top + wz^\top = \begin{bmatrix} — & u_1 v^\top + w_1 z^\top & — \\ — & u_2 v^\top + w_2 z^\top & — \\ & \vdots & \\ — & u_m v^\top + w_m z^\top & — \end{bmatrix} = \begin{bmatrix} | & | \\ u & w \\ | & | \end{bmatrix} \begin{bmatrix} — & v^\top & — \\ — & z^\top & — \end{bmatrix}$$

Low Rank Approximation

- Low Rank Approximation: take only top k-singular values and corresponding dyads in the dyadic expansion

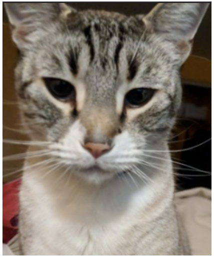
$$A \approx A_k = \sum_{i=1}^{k} \sigma_i \cdot u_i v_i^\top \quad \text{equivalently} \quad A_k = U_k \Sigma_k V_k^\top$$

- Low Rank Approximation is an important tool for many applications including
  - Linear system identification: approximating matrix is Hankel structured. (You saw this in `M2-N2.ipynb`)
  - ML: feature space dimensionality reduction
  - Recommender systems: matrix completion
  - Distance matrix completion where there is a positive definiteness constraint.
  - Natural language processing where the approximation is non-negative.
  - Image or video compression

Example: Compression

- **Compression.** A low-rank approximation provides a (lossy) compressed version of the data matrix.
  - The original matrix A is described by mn numbers, while describing Y and $Z^\top$ requires only $k(m+n)$ numbers.
  - When k is small relative to m and n, replacing the product of m and n by their sum is a big win.
  - With images, a modest value of k (say 100 or 150) is usually enough to achieve approximations that look a lot like the original image.

bender in color



gray scale

rank 2 bender

rank 20 bender

rank 100 bender

Optimality of Low Rank Approximation

- The low rank approximation obtained via SVD is optimal in the following sense.
- Recall the Forbenius norm:

$$\|A\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2\right)^{1/2} = (\mathrm{Tr}(A^\top A))^{1/2}$$

- This is just the ℓ_2-norm (i.e., usual Euclidean norm) applied to the matrix as if it were a vector
- **Theorem [Eckat-Young-Mirsky].** $A_k = \sum_{i=1}^{k} \sigma_i \cdot u_i v_i^\top$ is the closest matrix of rank k to the matrix A: i.e,

$$\|A - A_k\|_F \le \|A - B\|_F \quad \forall \text{ rank-}k \text{ matrices } B \in \mathbb{R}^{m\times n}$$

How to choose k?

- When producing a low-rank matrix approximation, we have been taking as a parameter the target rank k.
- **Ideal Setting**: the singular values of A give strong guidance
  - if the top few singular values are big and the rest are small, then the obvious solution is to take k equal to the number of "big values".
- **Less Ideal Setting**: take k as small as possible subject to obtaining a useful approximation, where what "useful" means depends on the application.
  - e.g., a common rule of thumb is to choose k such that the sum of the top k singular values is at least c times as big as the sum of the other singular values, where c is a domain-dependent constant (like 10, say).

Next Up

- Next lecture we will talk about PCA, and show that PCA reduces to SVD and is fundamentally connected to low rank approximations.