Chapter 1
SVD, PCA & Pre-processing

Part 1: Linear algebra and SVD
Contents

• Linear algebra crash course
• The singular value decomposition
• Normalization
• Selecting the rank
• The principal component analysis
Linear Algebra Crash Course
Matrices and vectors

• A vector is
  • a 1D array of numbers
  • a geometric entity with magnitude and direction
  • a matrix with exactly one row or column
**Norms and angles**

- The magnitude is measured by a (vector) **norm**
  - The **Euclidean** norm
    \[
    \|\mathbf{x}\| = \|\mathbf{x}\|_2 = \left(\sum_{i=1}^{n} x^2\right)^{1/2}
    \]
  - General \(L_p\) norm
    \[
    \|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x|^p\right)^{1/p}
    \]
- The direction is measured by the **angle**
  \[
  \angle 0.5880 \text{ rad (33.69°)}
  \]
Basic vector operations

• The **transpose** of $\mathbf{x}$, $\mathbf{x}^T$, transposes a row vector into a column vector and vice versa.

• A **dot product** of two vectors of the same dimension is $\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{n} x_i y_i$

• A.k.a. **scalar product** or **inner product**

• Same as $\langle \mathbf{x}, \mathbf{y} \rangle$, $\mathbf{a}^T \mathbf{b}$ (for column vectors), or $\mathbf{a} \mathbf{b}^T$ (for row vectors)
Orthogonality

• **Orthogonality** is a generalization of perpendicularity

• \( x \) and \( y \) are orthogonal if \( x \cdot y = 0 \)

• in Euclidean space: \( x \cdot y = ||x|| \ ||y|| \cos \theta \)
  
  • \( \theta \) is the angle between \( x \) and \( y \)
Matrix algebra

- Matrices in $\mathbb{R}^{n \times n}$ form a ring
  - Addition, subtraction, and multiplication
  - But usually no division
  - Multiplication is not commutative
    - $AB \neq BA$ in general
Matrix multiplication

• The product of two matrices, $A$ and $B$, is defined element-wise as

$$(AB)_{ij} = \sum_{\ell=1}^{k} a_{i\ell} b_{\ell j}$$

• The number of columns in $A$ and number of rows in $B$ must agree

• inner dimension
Intuition for Matrix Multiplication

- Element \((AB)_{ij}\) is the inner product of row \(i\) of \(A\) and column \(j\) of \(B\)

\[
C_{ij} = \sum_{l=1}^{k} a_{il} b_{lj}
\]
Intuition for Matrix Multiplication

• Column \( j \) of \( AB \) is the linear combination of columns of \( A \) with the coefficients coming from column \( j \) of \( B \)

\[
C = \begin{bmatrix}
\sum_{l=1}^{k} b_{l1} a_{l} \\
\sum_{l=1}^{k} b_{l2} a_{l} \\
\vdots \\
\sum_{l=1}^{k} b_{lm} a_{l}
\end{bmatrix}
\]
Intuition for Matrix Multiplication

- Matrix $\mathbf{A}\mathbf{B}$ is a sum of $k$ matrices $\mathbf{a}_l\mathbf{b}_l^T$ obtained by multiplying the $l$-th column of $\mathbf{A}$ with the $l$-th row of $\mathbf{B}$

$$\mathbf{C} = \sum_{l=1}^{k} \mathbf{a}_l \mathbf{b}_l^T$$
Matrix decompositions

• A decomposition of matrix $A$ expresses it as a product of two (or more) factor matrices
  
  • $A = BC$

• Every matrix has decomposition $A = AI$ (or $A = IA$ if $n < m$)

• The size of the decomposition is the inner dimension of the product
Matrices as linear maps

• Matrix $A \in \mathbb{R}^{n \times m}$ is a **linear mapping** from $\mathbb{R}^m$ to $\mathbb{R}^n$
  • $A(x) = Ax$

• If $A \in \mathbb{R}^{n \times k}$ and $B \in \mathbb{R}^{k \times m}$, then $AB$ is a mapping from $\mathbb{R}^m$ to $\mathbb{R}^n$

• The transpose $A^T$ is a mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$
  • $(A^T)_{ij} = A_{ji}$
  • $(AB)^T = B^T A^T$
Matrix inverse

- Square matrix $A$ is **invertible** if there is a matrix $B$ s.t. $AB = BA = I$
  - $B$ is the inverse of $A$, denoted $A^{-1}$
  - Usually the transpose is **not** the inverse
- Non-square matrices don’t have general inverses
  - Can have left or right inverse:
    - $AR = I$ or $LA = I$
Linear independency

• Vector $u$ is **linearly dependent** on a set of vectors $V = \{v_i\}$ if $u$ is a linear combination of $v_i$
  
  • $u = \sum_i a_i v_i$ for some $a_i$

• If $u$ is not linearly dependent, it is **linearly independent**

• Set $V$ of vectors is **linearly independent** if all $v_i$ are linearly independent of $V \setminus \{v_i\}$
Matrix ranks

• The **column rank** of a matrix $A$ is the number of linearly independent columns of $A$

• The **row rank** of $A$ is the number of linearly independent rows of $A$

• The **Schein rank** of $A$ is the least integer $k$ such that $A$ can be expressed as a sum of $k$ rank-1 matrices
  
  • Rank-1 matrix is an outer product of two vectors
Orthogonal matrices

• Set of vectors \( \{ \mathbf{v}_i \} \) is **orthogonal** if all \( \mathbf{v}_i \) are mutually orthogonal, i.e. \( \langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0 \) for all \( i \neq j \)

  • If \( ||\mathbf{v}_i||_2 = 1 \) for all \( \mathbf{v}_i \), the set is **orthonormal**

• Square matrix \( \mathbf{A} \) is orthogonal if its columns form a set of orthonormal vectors

  • Non-square matrices can be row- or column-orthogonal

• If \( \mathbf{A} \) is orthogonal, then \( \mathbf{A}^{-1} = \mathbf{A}^T \)
Properties of orthogonal matrices

• The inverse of orthogonal matrices is easy to compute

• Orthogonal matrices perform a rotation
  • Only the angle of the vector is changed, the length stays the same
Matrix norms

- **Matrix norms** measure the magnitude of the matrix
- the magnitude of the values or the image

- **Operator norms:**
  \[ \| A \|_p = \max\{ \| Mx \|_p : \| x \|_p = 1 \} \text{ for } p \geq 1 \]

- **Frobenius norm:**
  \[ \| A \|_F = \left( \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij}^2 \right)^{1/2} \]
Singular Value Decomposition
“The SVD is the Swiss Army knife of matrix decompositions”

– Diane O’Leary, 2006
The definition

• **Theorem.** For every $A \in \mathbb{R}^{n\times m}$ there exists an $n$-by-$n$ orthogonal matrix $U$ and an $m$-by-$m$ orthogonal matrix $V$ such that $U^TAV$ is an $n$-by-$m$ diagonal matrix $\Sigma$ that has values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min\{n,m\}} \geq 0$ in its diagonal

• i.e. every $A$ has decomposition $A = U\Sigma V^T$

• The **singular value decomposition** of $A$
$\mathbf{A} = \mathbf{U} \Sigma \mathbf{V}^T$

- $\mathbf{v}_i$ are the right singular vectors
- $\mathbf{u}_i$ are the left singular vectors
- $\sigma_i$ are the singular values
Some useful equations

• $A = UΣV^T = \sum_i \sigma_i u_i v_i^T$
  • Expresses $A$ as a sum of rank-1 matrices

• $A^{-1} = (UΣV^T)^{-1} = VΣ^{-1}U^T$ (if $A$ is invertible)

• $A^TAv_i = \sigma_i^2 v_i$ (for any $A$)

• $AA^Tu_i = \sigma_i^2 u_i$ (for any $A$)
Truncated SVD

- The rank of the matrix is the number of its non-zero singular values (write $A = \sum_i \sigma_i u_i v_i^T$)
- The truncated SVD takes the first $k$ columns of $U$ and $V$ and the main $k$-by-$k$ submatrix of $\Sigma$
  - $A_k = U_k \Sigma_k V_k^T$
  - $U_k$ and $V_k$ are column-orthogonal
Truncated SVD

\[
\begin{align*}
\text{Full} & : \quad A = U \Sigma V^T \\
\text{Truncated} & : \quad A \approx U \Sigma V^T
\end{align*}
\]
Why is SVD important?

• It gives us the **dimensions of the fundamental subspaces**
• It lets us compute various norms
• It tells about **sensitivity of linear systems**
• It gives us optimal solutions to **least-squares linear systems**
• It gives us the **least-error rank-$$k$$ decomposition**
• Every matrix has one
Fundamental theorem of linear algebra

• **Theorem.** Every $n$-by-$m$ matrix $A$ induces four fundamental subspaces
  
  • The **range** of dimension $\text{rank}(A) = r$
    
    • The set of all linear combinations of columns of $A$
  
  • The **kernel** of dimension $m - r$
    
    • The set of all vectors $x$ for which $Ax = 0$
  
  • The **coimage** of dimension $r$
    
    • The **cokernel** of dimension $n - r$
Fundamental subspaces

• The bases for the fundamental subspaces are:
  • Range: the first $r$ columns of $U$
  • Kernel: the last $(m - r)$ columns of $V$
  • Coimage: the first $r$ columns of $V$
  • Cokernel: the last $(n - r)$ columns of $U$
SVD and norms

• Let $A = U\Sigma V^T$ be the SVD of $A$.

• $\|A\|_F^2 = \sum_{i=1}^{\min\{n,m\}} \sigma_i^2$

• $\|A\|_2 = \sigma_1$

• Therefore $\|A\|_2 \leq \|A\|_F \leq \sqrt{\min\{n,m\}} \|A\|_2$

• For truncated SVD, $\|A_k\|_F^2 = \sum_{i=1}^{k} \sigma_i^2$
Sensitivity of linear systems

- The solution for system $Ax = b$ is $x = A^{-1}b$
- Requires that $A$ is invertible
- Hence $x = (U\Sigma V^T)^{-1}b = \sum_{i=1}^{n} \frac{u_i^T b}{\sigma_i} v_i$
- Small changes in $A$ or $b$ yield large changes in $x$ if $\sigma_n$ is small
- Can we characterize this sensitivity?
Condition number

- The condition number $\kappa_p(A)$ of a square matrix $A$ is $||A||_p ||A^{-1}||_p$
  - Particularly $\kappa_2(A) = \sigma_1(A)/\sigma_n(A)$
  - $\kappa_2(A) = \infty$ for singular $A$
- If $\kappa$ is large, the matrix is **ill-conditioned**
  - The solution is sensible for small perturbations
Least-squares linear systems

- **Problem.** Given \( A \in \mathbb{R}^{n \times m} \) and \( b \in \mathbb{R}^n \), find \( x \in \mathbb{R}^m \) minimizing \( \|Ax - b\|_2 \).

- If \( A \) is invertible, \( x = A^{-1}b \) is an exact solution.

- For non-invertible \( A \) we have to find other solution.
The Moore–Penrose pseudo-inverse

- $n$-by-$m$ matrix $B$ is the Moore–Penrose pseudo-inverse of $n$-by-$m$ matrix $A$ if
  - $ABA = A$ (but possibly $AB \neq I$)
  - $BAB = B$
  - $(AB)^T = AB$ ($AB$ is symmetric)
  - $(BA)^T = BA$
  - Pseudo-inverse of $A$ is denoted by $A^+$
Pseudo-inverse and SVD

- If $A = U\Sigma V^T$ is the SVD of $A$, then
  $$A^+ = V\Sigma^{-1}U^T$$

- $\Sigma^{-1}$ replaces non-zero $\sigma_i$’s with $1/\sigma_i$ and transposes the result

- N.B. not a real inverse

- **Theorem.** Setting $x = A^+y$ gives the optimal solution to $||Ax - y||$
The Eckart–Young theorem

- **Theorem.** Let $A_k = U_k \Sigma_k V_k^T$ be the rank-$k$ truncated SVD of $A$. Then $A_k$ is the closest rank-$k$ matrix of $A$ in the Frobenius sense, that is,

\[ ||A - A_k||_F \leq ||A - B||_F \]

for all rank-$k$ matrices $B$

- Holds for any unitarily invariant norm
That’s all for today

• Next week: normalization and selecting the rank
  • Lecture starts at 12:00 sharp
  • Will end earlier as well
• But SVD will return...