CSC 4020 Fundamental of Machine Learning: Introduction to Linear Algebra

Baoyuan Wu
School of Data Science, CUHK-SZ

January 18/20, 2021

Outline

Scalar, vector, matrix

- Scalar: we denote a scalar as $x \in \mathbb{R}$ (plot below)

Scalar, vector, matrix

- Scalar: we denote a scalar as $x \in \mathbb{R}$ (plot below)

- Vector: we denote a vector as $\boldsymbol{x} = [x_1; x_2; \ldots; x_d] \in \mathbb{R}^d$. Note that when we say a vector, if no specification, it always means a column vector. (plot below)

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

Scalar, vector, matrix

- Scalar: we denote a scalar as $x \in \mathbb{R}$ (plot below)

- Vector: we denote a vector as $\boldsymbol{x} = [x_1; x_2; \ldots; x_d] \in \mathbb{R}^d$. Note that when we say a vector, if no specification, it always means a column vector. (plot below)

- Matrix: we denote a matrix as $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{m \times n}$. (plot below)

$m \left[x_1, x_2 \cdots x_n \right]$

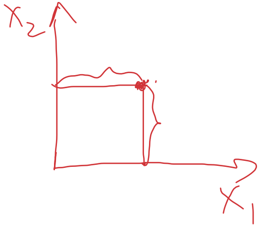
Norms of scalar, vector and matrix

- Absolute-value norm: $||x|| = |x|$
- ℓ_0 norm: $\|\boldsymbol{x}\|_0 = \sum_i^d \mathbb{I}(x_i \neq 0)$, the number of non-zero entries in a vector

$$\begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_d \end{bmatrix} \quad \begin{array}{l} x_1 \rightarrow 0 \\ \\ x_i \rightarrow \neq 0 \\ \\ \\ \end{array}$$

Norms of scalar, vector and matrix

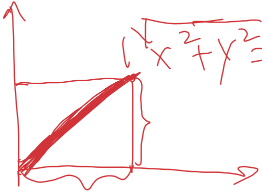
- Absolute-value norm: $||x|| = |x|$
- ℓ_0 norm: $\|\boldsymbol{x}\|_0 = \sum_i^d \mathbb{I}(x_i \neq 0)$, the number of non-zero entries in a vector
- ℓ_1 norm: $\|\boldsymbol{x}\|_1 = \sum_i^d |x_i|$, the summation of absolute values of all entries



Norms of scalar, vector and matrix

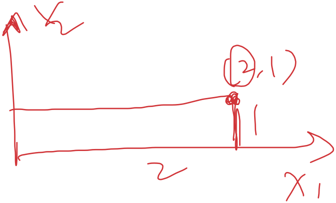
- Absolute-value norm: $||x|| = |x|$
- ℓ_0 norm: $\|\boldsymbol{x}\|_0 = \sum_i^d \mathbb{I}(x_i \neq 0)$, the number of non-zero entries in a vector
- ℓ_1 norm: $\|\boldsymbol{x}\|_1 = \sum_i^d |x_i|$, the summation of absolute values of all entries
- ℓ_2 norm: $\|\boldsymbol{x}\|_2 = (\sum_i^d x_i^2)^{\frac{1}{2}}$, which is also called **Euclidean norm**

$\sqrt{x^2+y^2} = \|z\|$



Norms of scalar, vector and matrix

- Absolute-value norm: $||x|| = |x|$
- ℓ_0 norm: $\|\boldsymbol{x}\|_0 = \sum_i^d \mathbb{I}(x_i \neq 0)$, the number of non-zero entries in a vector
- ℓ_1 norm: $\|\boldsymbol{x}\|_1 = \sum_i^d |x_i|$, the summation of absolute values of all entries
- ℓ_2 norm: $\|\boldsymbol{x}\|_2 = (\sum_i^d x_i^2)^{\frac{1}{2}}$, which is also called **Euclidean norm**
- ℓ_∞ norm: $\|\boldsymbol{x}\|_\infty = \max_i |x_i|$



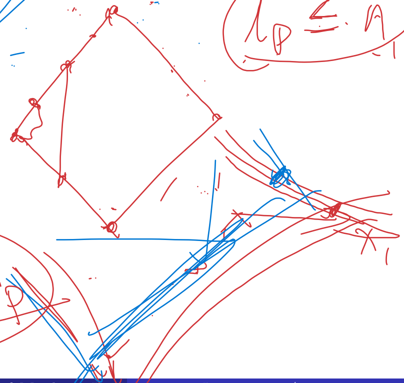
- Absolute-value norm: $||x|| = |x|$
- ℓ_0 norm: $\|\boldsymbol{x}\|_0 = \sum_i^d \mathbb{I}(x_i \neq 0)$, the number of non-zero entries in a vector
- ℓ_1 norm: $\|\boldsymbol{x}\|_1 = \sum_i^d |x_i|$, the summation of absolute values of all entries
- ℓ_2 norm: $\|\boldsymbol{x}\|_2 = (\sum_i^d x_i^2)^{\frac{1}{2}}$, which is also called **Euclidean norm**
- ℓ_∞ norm: $\|\boldsymbol{x}\|_\infty = \max_i |x_i|$
- ℓ_p norm: $\|\boldsymbol{x}\|_p = (\sum_i^d |x_i|^p)^{\frac{1}{p}}$

$0 < p < 1$ $\quad p > 1 \ldots$ $\quad \infty$

$\ell_p \leq 1$

$\|\hat{y} - y\|_p^2$

$\|\theta\|_p$ $\quad \|\theta\|_p$



Norms of scalar, vector and matrix

- Absolute-value norm: $||x|| = |x|$
- ℓ_0 norm: $\|\boldsymbol{x}\|_0 = \sum_i^d \mathbb{I}(x_i \neq 0)$, the number of non-zero entries in a vector
- ℓ_1 norm: $\|\boldsymbol{x}\|_1 = \sum_i^d |x_i|$, the summation of absolute values of all entries
- ℓ_2 norm: $\|\boldsymbol{x}\|_2 = (\sum_i^d x_i^2)^{\frac{1}{2}}$, which is also called **Euclidean norm**

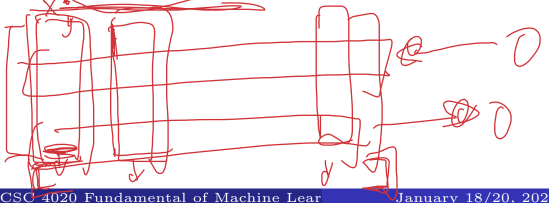
$L_1 \Rightarrow$ Sparsity

- ℓ_∞ norm: $\|\boldsymbol{x}\|_\infty = \max_i |x_i|$
- ℓ_p norm: $\|\boldsymbol{x}\|_p = (\sum_i^d |x_i|^p)^{\frac{1}{p}}$

$\boxed{L_{2,1}} \Rightarrow$

- $\ell_{p,q}$ norm: $\|\boldsymbol{X}\|_{p,q} = \left(\sum_j^n \left(\sum_i^m |x_{ij}|^p\right)^{\frac{q}{p}}\right)^{\frac{1}{q}}$

$p = 2$, $q = 1$



Norms of scalar, vector and matrix

- Absolute-value norm: $||x|| = |x|$

- ℓ_0 norm: $\|\boldsymbol{x}\|_0 = \sum_i^d \mathbb{I}(x_i \neq 0)$, the number of non-zero entries in a vector
- ℓ_1 norm: $\|\boldsymbol{x}\|_1 = \sum_i^d |x_i|$, the summation of absolute values of all entries
- ℓ_2 norm: $\|\boldsymbol{x}\|_2 = (\sum_i^d x_i^2)^{\frac{1}{2}}$, which is also called **Euclidean norm**
- ℓ_∞ norm: $\|\boldsymbol{x}\|_\infty = \max_i |x_i|$
- ℓ_p norm: $\|\boldsymbol{x}\|_p = (\sum_i^d |x_i|^p)^{\frac{1}{p}}$

- $\ell_{p,q}$ norm: $\|\boldsymbol{X}\|_{p,q} = \left(\sum_j^n \left(\sum_i^m |x_{ij}|^p\right)^{\frac{q}{p}}\right)^{\frac{1}{q}}$
- $\ell_{Frobenius}$ norm: $\|\boldsymbol{X}\|_F = \sqrt{\sum_i^m \sum_j^n |x_{ij}|^2} = \sqrt{\text{trace}(A^\top A)}$, where $\text{trace}(\boldsymbol{X})$ calculates the sum of diagonal entries of a square matrix.

- Product of two vectors (plot):

$$\boldsymbol{x}_1^\top \boldsymbol{x}_2 = \sum_i^d x_1(i) \cdot x_2(i) = \|\boldsymbol{x}_1\| \cdot \|\boldsymbol{x}_2\| \cdot \cos(\theta) \tag{1}$$

Product

- Product of two vectors (plot):

$$\boldsymbol{x}_1^\top \boldsymbol{x}_2 = \sum_i^d x_1(i) \cdot x_2(i) = \|\boldsymbol{x}_1\| \cdot \|\boldsymbol{x}_2\| \cdot \cos(\theta) \tag{1}$$

$p \times q$

- Product of matrix and vector (plot): $\Rightarrow m \times q$

$m \times n$ $A \cdot B \cdot C \cdot D \cdot E \cdot F$

n

$$\boldsymbol{A}\boldsymbol{x} = [\sum_j^n a_{1j} x_j; \ldots; \sum_m^n a_{ij} x_j] \tag{2}$$

$n \times 1$

$a_1 \quad a_2 \quad --- \quad a_n$

m

x

Product

- Product of two vectors (plot):

$$\boldsymbol{x}_1^\top \boldsymbol{x}_2 = \sum_i^d x_1(i) \cdot x_2(i) = \|\boldsymbol{x}_1\| \cdot \|\boldsymbol{x}_2\| \cdot \cos(\theta) \tag{1}$$

- Product of matrix and vector (plot):

$$\boldsymbol{A}\boldsymbol{x} = [\sum_j^n a_{1j} x_j; \ldots; \sum_m^n \underline{a_{ij} x_j}] \tag{2}$$

- Product of two matrices:

$$\underline{\boldsymbol{AB}} =? \tag{3}$$

$[Ab_1, Ab_2 - - - Ab_q]$

Derivative

$y = ax + b$

$\frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix}$

- Vector by scalar: $\frac{\partial \boldsymbol{y}}{\partial x} = [\frac{\partial y_1}{\partial x}; \ldots; \frac{\partial y_m}{\partial x}] \in \mathbb{R}^m$
- Scalar by vector: $\frac{\partial y}{\partial \boldsymbol{x}} = [\frac{\partial y}{\partial x_1}; \ldots; \frac{\partial y}{\partial x_n}] \in \mathbb{R}^n$
- Vector by vector: $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = [\frac{\partial \boldsymbol{y}}{\partial x_1}; \ldots; \frac{\partial \boldsymbol{y}}{\partial x_n}] \in \mathbb{R}^{m \times n}$, with $\boldsymbol{y} \in \mathbb{R}^m, \boldsymbol{x} \in \mathbb{R}^n$
- Scalar by matrix: $\frac{\partial x}{\partial \boldsymbol{A}} = [\frac{\partial x}{\partial \boldsymbol{a}_1}; \ldots; \frac{\partial x}{\partial \boldsymbol{a}_n}] \in \mathbb{R}^{m \times n}$, with $\boldsymbol{A} = [\boldsymbol{a}_1; \ldots; \boldsymbol{a}_n]$
- Vector by matrix: $\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{A}} = \frac{\partial x_1}{\partial \boldsymbol{A}}; \ldots; \frac{\partial x_d}{\partial \boldsymbol{A}} \in \mathbb{R}^{d \times m \times n}$, with $\boldsymbol{x} \in \mathbb{R}^d, \boldsymbol{A} \in \mathbb{R}^{m \times n}$.

$\frac{\partial x}{\partial A} = \begin{bmatrix} \frac{\partial x}{\partial a_{11}} & \cdots & \\ & \ddots & \\ & & \frac{\partial x}{\partial a_{ij}} \end{bmatrix}$

Eigenvalue,eigenvector

- For a square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, we have

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}, \tag{4}$$

where λ is one eigenvalue of $\boldsymbol{A}$, while $\boldsymbol{x}$ is the corresponding eigenvector.

λ_1

$\vdots$

λ_n

$[x_1 \; - \; - \; \cdots \; x_n]$

$\forall i \; \lambda_i \geq 0$

$x^T A x \geq 0$

Eigenvalue,eigenvector

- For a square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, we have

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}, \tag{4}$$

  where λ is one eigenvalue of $\boldsymbol{A}$, while $\boldsymbol{x}$ is the corresponding eigenvector.
- Besides, $\boldsymbol{A}$ can be decomposed as follows:

$$\boldsymbol{A} = \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^{-1}, \tag{5}$$

  where the columns of $\boldsymbol{W} \in \mathbb{R}^{n \times n}$ are n eigenvectors, while the diagonal values of $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ are n corresponding eigenvalues.

Eigenvalue,eigenvector

- For a square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, we have

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}, \tag{4}$$

  where λ is one eigenvalue of $\boldsymbol{A}$, while $\boldsymbol{x}$ is the corresponding eigenvector.
- Besides, $\boldsymbol{A}$ can be decomposed as follows:

$$\boldsymbol{A} = \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^{-1}, \tag{5}$$

  where the columns of $\boldsymbol{W} \in \mathbb{R}^{n \times n}$ are n eigenvectors, while the diagonal values of $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ are n corresponding eigenvalues.
- If $\boldsymbol{A} = \boldsymbol{A}^\top$, then we have

$$\boldsymbol{A} = \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^\top. \tag{6}$$

SVD decomposition

- For a rectangular matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$,

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \tag{7}$$

SVD decomposition

- For a rectangular matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$,

$$\boldsymbol{A} = \boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{V}^\top \tag{7}$$

- $\boldsymbol{U} \in \mathbb{R}^{m \times m}, \boldsymbol{V} \in \mathbb{R}^{n \times n}$ are orthogonal.

SVD decomposition

$x_i^\top x_j = 0$

x_i x_j

- For a rectangular matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$,

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \tag{7}$$

- $\boldsymbol{U} \in \mathbb{R}^{m \times m}, \boldsymbol{V} \in \mathbb{R}^{n \times n}$ are orthogonal.
- $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{m \times n}$, and $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$ indicate singular values of $\boldsymbol{A}$, and $p = \min(m, n)$.

p

p