Mathematical Foundations: Linear Algebra

Ziping Zhao

School of Information Science and Technology
ShanghaiTech University, Shanghai, China

CS182: Introduction to Machine Learning (Fall 2022)
http://cs182.sist.shanghaitech.edu.cn

App. B of I2ML

Matrix

$$\mathbf{A} = [a_{ij}]_{m\times n} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = [\mathbf{a}_1\ \mathbf{a}_2\ \cdots\ \mathbf{a}_n]$$

- diagonal matrix:

$$\mathrm{diag}(a_{11}, a_{22}, \cdots, a_{nn}) = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix}$$

- identity matrix: $\mathbf{I} = \mathrm{diag}(1, 1, \cdots, 1)$
- trace: $\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$

Matrix Addition/Subtraction

If $\mathbf{C} = \mathbf{A} \pm \mathbf{B}$, then $[c_{ij}] = [a_{ij}] \pm [b_{ij}]$

- commutative: $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$
- associative: $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$

Multiply a Vector by a Matrix

$$\mathbf{A}\mathbf{x} = \mathbf{y}$$

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$y_i = \sum_{j=1}^{n} a_{ij} x_j$$

write $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n]$, then

$$\mathbf{y} = \sum_{j=1}^{n} x_j \mathbf{a}_j$$

- $\mathbf{y}$ can be written as a weighted sum of $\mathbf{A}$'s column vectors

Matrix Multiplication

If $\mathbf{C}_{m\times n} = \mathbf{A}_{m\times p}\mathbf{B}_{p\times n}$, then $[c_{ij}] = \sum_{k=1}^{p} a_{ik} b_{kj}$

- in general, non-commutative: $\mathbf{AB} \neq \mathbf{BA}$
- associative: $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$
- distributive: $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$

Transpose

- If $\mathbf{B} = \mathbf{A}^T$, then $b_{ij} = a_{ji}$
  - $\mathbf{A}^T$ is sometimes also denoted as $\mathbf{A}'$ or $\mathbf{A}^t$
- $(\mathbf{A}^T)^T = \mathbf{A}$, $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$, $(\mathbf{A}+\mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$
- symmetric matrix: $a_{ij} = a_{ji}$ or $\mathbf{A} = \mathbf{A}^T$
- Matrix $\mathbf{A}$ is orthogonal if $\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T = \mathbf{I}$ and $\mathbf{A}^T = \mathbf{A}^{-1}$

Determinant

- if $\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$, then $\det(\mathbf{A}) = |\mathbf{A}| = a_{11}a_{22} - a_{21}a_{12}$
- in general,
$$|\mathbf{A}| = \sum_{j=1}^{n} a_{ij}\mathrm{cof}(a_{ij}),$$
  - $\mathrm{cof}(a_{ij})$ is the cofactor of element a_{ij} and is defined as the product of $(-1)^{i+j}$ times the determinant of $\mathbf{A}$ after deleting its ith row and jth column

Properties:

- determinant is a scalar quantity
- if $|\mathbf{A}| = 0$ then $\mathbf{A}$ is singular, otherwise non-singular
- $|\mathbf{A}^T| = |\mathbf{A}|$
- $|\mathbf{AB}| = |\mathbf{BA}| = |\mathbf{A}||\mathbf{B}|$ (the last equality holds if $\mathbf{A}$ and $\mathbf{B}$ are symmetric)

Linear Dependence and Ranks

- A set of vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ is linearly dependent if there exist constants $c_1, c_2, \ldots, c_m$ (not all zero) such that

$$c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_m\mathbf{x}_m = \mathbf{0}$$

- For a $m \times n$ matrix $\mathbf{A}$, there are two sets of vectors: the columns and rows (of different sizes). Its column rank is the number of linearly independent columns, which is less than or equal to n; its row rank is similarly defined.
- The column rank is equal to the row rank. The rank of $\mathbf{A}$, denoted as $\text{rank}(\mathbf{A})$ is defined as either of them.
- If $\text{rank}(\mathbf{A}) = \min\{m, n\}$, it is full rank.

Inverse

$$\mathbf{A}^{-1} = \frac{[\mathrm{cof}(\mathbf{A})]^T}{|\mathbf{A}|}$$

- $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
- $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
- $(\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T = \mathbf{A}^{-T}$

Vector Norm

Norm of a vector $\mathbf{x}$ is used to measure the length of $\mathbf{x}$
Examples of norm:

- 2-norm or Euclidean norm: $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$
- 1-norm or Taxicab norm or Manhattan norm: $\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$
- ∞-norm or maximum norm or sup-norm: $\|\mathbf{x}\|_\infty = \max_{i=1}^{n} |x_i|$
- p-norm ($p \geq 1$) or Hölder norm: $\|\mathbf{x}\|_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$

Inner Product, Outer Product

The inner product (dot product or scalar product) of two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y} = \mathbf{y}^T \mathbf{x} = \sum_{i=1}^{n} x_i y_i$$

- if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$, then $\mathbf{x}$ and $\mathbf{y}$ are orthogonal
- $\langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \cos\theta$, where θ is the angle between $\mathbf{x}$ and $\mathbf{y}$
- $\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^T \mathbf{x}}$

The outer product of two vectors $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$ is a matrix $\mathbf{A} = \mathbf{x}\mathbf{y}^T$, where

$$[a_{ij}] = [x_i y_j] = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_n \\ \vdots & \vdots & \vdots & \vdots \\ x_m y_1 & x_m y_2 & \cdots & x_m y_n \end{bmatrix}$$

Gradient Vector

Given: $f(\mathbf{x})$ is a real valued function

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \mathbf{g}(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \left[\frac{\partial f(\mathbf{x})}{\partial x_1}, \frac{\partial f(\mathbf{x})}{\partial x_2}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_n}\right]^T$$

- first order derivatives

Example

$\mathbf{x} = [x_1, x_2, x_3]^T$, $f(\mathbf{x}) = 2x_1^2 x_2 - x_1 x_3^3$

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial}{\partial x_1} f(\mathbf{x}) \\ \frac{\partial}{\partial x_2} f(\mathbf{x}) \\ \frac{\partial}{\partial x_3} f(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} 4x_1 x_2 - x_3^3 \\ 2x_1^2 \\ -3x_1 x_3^2 \end{bmatrix}$$

Gradient Vector: Properties

- $\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{y}) = \nabla_{\mathbf{x}}(\mathbf{y}^T\mathbf{x}) = \mathbf{y}$
- $\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{x}) = 2\mathbf{x}$
- $\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{y}) = \mathbf{A}\mathbf{y}$
- $\nabla_{\mathbf{x}}(\mathbf{y}^T\mathbf{A}\mathbf{x}) = \mathbf{A}^T\mathbf{y}$
- $\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{A}^T\mathbf{x}$ (if $\mathbf{A}$ is symmetric: $= 2\mathbf{A}\mathbf{x}$)

Hessian Matrix

Second order derivatives

$$\nabla^2_{\mathbf{x}} f(\mathbf{x}) = \mathbf{H}(\mathbf{x}) = \frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^T} = \left[\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \right]$$

$$= \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \vdots \\ \vdots & \vdots & & \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{bmatrix}$$

Obviously, the Hessian matrix is always symmetric

Positive Semidefinite Matrices - I

A symmetric matrix $\mathbf{A}$ is said to be

- positive semidefinite (PSD) if $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ for all $\mathbf{x}$
- positive definite (PD) if $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x}$ with $\mathbf{x} \neq 0$
- indefinite if both $\mathbf{A}$ and $-\mathbf{A}$ are not PSD

Notion:

- $\mathbf{A} \succeq \mathbf{0}$ means that $\mathbf{A}$ is PSD
- $\mathbf{A} \succ \mathbf{0}$ means that $\mathbf{A}$ is PD
- $\mathbf{A} \not\succeq \mathbf{0}$ means that $\mathbf{A}$ is indefinite

- if $\mathbf{A}$ is PD, then it is also PSD
- The concepts negative semidefinite and negative definite may be defined by reversing the inequalities or, equivalently, by saying $-\mathbf{A}$ is PSD or PD, respectively.

Positive Semidefinite Matrices - II

- If $\mathbf{A} = \mathbf{X}^T\mathbf{X}$ and $\mathbf{X}$ is $m \times n$, with $\text{rank}(\mathbf{A}) = n < m$, then $\mathbf{A}$ is positive definite. If $\text{rank}(\mathbf{A}) < \min\{m, n\}$, then $\mathbf{A}$ is positive semidefinite.
- A positive definite matrix can be "factored" as $\mathbf{A} = \mathbf{T}^T\mathbf{T}$, where $\mathbf{T}$ is a nonsingular upper triangular matrix. One way to obtain $\mathbf{T}$ is by Cholesky decomposition.

Eigenvalue λ ; Eigenvector $\mathbf{v}$

$$\begin{aligned}
\mathbf{A}\mathbf{v} &= \lambda\mathbf{v} \\
(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} &= \mathbf{0} \\
|\mathbf{A} - \lambda\mathbf{I}| &= 0 \quad \text{(characteristic equation)}
\end{aligned}$$

Solutions (λ) to the characteristic equation are called eigenvalues and their corresponding $\mathbf{v}$ eigenvectors

- The eigenvalues of a positive definite matrix are all positive.
- The eigenvalues of a positive semidefinite matrix are all positive or zero, with the number of nonzero eigenvalues equal to the rank of the matrix.
- If $\mathbf{A}$ and $\mathbf{B}$ are both square and of the same size, the eigenvalues of $\mathbf{AB}$ and $\mathbf{BA}$ are the same, though the eigenvectors may be different.
  - This result holds even if $\mathbf{AB}$ and $\mathbf{BA}$ are both square but of different sizes.

Eigendecomposition (Spectral Decomposition)

- If a square $n \times n$ matrix A has an eigendecomposition (spectral decomposition), it can be written as

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$$

  where $\mathbf{Q}$ is a square $n \times n$ matrix whose columns are the eigenvectors of $\mathbf{A}$ ordered in terms of decreasing eigenvalues. $\mathbf{\Lambda}$ is a diagonal $n \times n$ matrix whose diagonal elements are the corresponding eigenvalues. The eigenvectors are generally normalized so that $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$.
- If $\mathbf{A}$ is symmetric, its eigenvectors are mutually orthogonal and the eigenvalues are real.

Singular Value Decomposition

- Singular value decomposition is a factorization method that can be viewed as a generalization of the spectral decomposition to rectangular matrices.
- If $\mathbf{A}$ is $m \times n$, it can be written as

$$\mathbf{A} = \mathbf{U\Sigma V}^T$$

where $\mathbf{U}$ is $m \times m$ and contains the orthonormal eigenvectors of $\mathbf{AA}^T$ in its columns, $\mathbf{V}$ is $n \times n$ and contains the orthonormal eigenvectors of $\mathbf{A}^T\mathbf{A}$ in its columns, and the $m \times n$ matrix $\mathbf{\Sigma}$ contains the $k = \min(m, n)$ singular values, σ_i, $i = 1, \ldots, k$, on its diagonals that are the square roots of the nonzero eigenvalues of both $\mathbf{AA}^T$ and $\mathbf{A}^T\mathbf{A}$; the rest of $\mathbf{\Sigma}$ is zero.
- We have

$$\mathbf{AA}^T = (\mathbf{U\Sigma V}^T)(\mathbf{U\Sigma V}^T)^T = \mathbf{U\Sigma V}^T\mathbf{V\Sigma}^T\mathbf{U}^T = \mathbf{U}(\mathbf{\Sigma\Sigma}^T)\mathbf{U}$$
$$\mathbf{A}^T\mathbf{A} = (\mathbf{U\Sigma V}^T)^T(\mathbf{U\Sigma V}^T) = \mathbf{V}(\mathbf{\Sigma}^T\mathbf{\Sigma})\mathbf{V}$$

where $\mathbf{\Sigma\Sigma}^T$ and $\mathbf{\Sigma}^T\mathbf{\Sigma}$ are of different sizes but are both square and contain on their diagonal and 0 elsewhere.