MATHEMATICS FOR ML

SMAI (CSE471)

Content

Vectors

- By default we write **vectors as column vectors.**
- Adding two vectors a, b $\in R^n$ component-wise results in another vector: a + b = c $\in R^n$.
- Multiplying a $\in R^n$ by $\boldsymbol{\lambda} \in R$ results in a scaled vector $\boldsymbol{\lambda}$a $\in R^n$.

$$\boldsymbol{a} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \in \mathbb{R}^3$$

Linear Independence

Definition 2.12 (Linear (In)dependence). Let us consider a vector space V with $k \in \mathbb{N}$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in V$. If there is a non-trivial linear combination, such that $\mathbf{0} = \sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i$ with at least one $\lambda_i \neq 0$, the vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ are *linearly dependent*. If only the trivial solution exists, i.e., $\lambda_1 = \ldots = \lambda_k = 0$ the vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ are *linearly independent*.

Generating Set and Span

Definition 2.13 (Generating Set and Span). Consider a vector space $V = (\mathcal{V}, +, \cdot)$ and set of vectors $\mathcal{A} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\} \subseteq \mathcal{V}$. If every vector $\boldsymbol{v} \in \mathcal{V}$ can be expressed as a linear combination of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$, $\mathcal{A}$ is called a *generating set* of V. The set of all linear combinations of vectors in $\mathcal{A}$ is called the *span* of $\mathcal{A}$. If $\mathcal{A}$ spans the vector space V, we write $V = \mathrm{span}[\mathcal{A}]$ or $V = \mathrm{span}[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k]$.

Basis

Definition 2.14 (Basis). Consider a vector space $V = (\mathcal{V}, +, \cdot)$ and $\mathcal{A} \subseteq \mathcal{V}$. A generating set $\mathcal{A}$ of V is called *minimal* if there exists no smaller set $\tilde{\mathcal{A}} \subseteq \mathcal{A} \subseteq \mathcal{V}$ that spans V. Every linearly independent generating set of V is minimal and is called a *basis* of V.

Norm

Definition 3.1 (Norm). A *norm* on a vector space V is a function

$$\|\cdot\| : V \to \mathbb{R}\,, \tag{3.1}$$

$$\boldsymbol{x} \mapsto \|\boldsymbol{x}\|\,, \tag{3.2}$$

which assigns each vector $\boldsymbol{x}$ its *length* $\|\boldsymbol{x}\| \in \mathbb{R}$, such that for all $\lambda \in \mathbb{R}$ and $\boldsymbol{x}, \boldsymbol{y} \in V$ the following hold:

- *Absolutely homogeneous:* $\|\lambda\boldsymbol{x}\| = |\lambda|\|\boldsymbol{x}\|$
- *Triangle inequality:* $\|\boldsymbol{x} + \boldsymbol{y}\| \leqslant \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$
- *Positive definite:* $\|\boldsymbol{x}\| \geqslant 0$ and $\|\boldsymbol{x}\| = 0 \iff \boldsymbol{x} = \boldsymbol{0}$

Norm

L_1 Norm

The *Manhattan norm* on $\mathbb{R}^n$ is defined for $\boldsymbol{x} \in \mathbb{R}^n$ as

$$\|\boldsymbol{x}\|_1 := \sum_{i=1}^{n} |x_i| , \tag{3.3}$$

L_2 Norm

The *Euclidean norm* of $\boldsymbol{x} \in \mathbb{R}^n$ is defined as

$$\|\boldsymbol{x}\|_2 := \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}} \tag{3.4}$$

L_0 Norm

$$\|x\|_0 = \sqrt[0]{\sum_i x_i^0}$$

-> Gives the total number of non-zero elements

L_∞ Norm

$$\|x\|_\infty = \sqrt[\infty]{\sum_i x_i^\infty} = max|x_i|$$

L_p Norm

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

Matrix

A matrix is a rectangular array of numbers, symbols,or expressions, arranged in rows and columns

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad a_{ij} \in \mathbb{R}.$$

Rank of Matrix

The number of linearly independent columns of a matrix $A \in \mathbb{R}^{m \times n}$ equals the number of linearly independent rows and is called the *rank* of A and is denoted by $\mathrm{rk}(A)$.

Properties of Determinant

- Swapping two rows/columns changes the sign of det(A).
- $\det(\boldsymbol{A}) = \det(\boldsymbol{A}^\top)$.
- $\det(\boldsymbol{AB}) = \det(\boldsymbol{A})\det(\boldsymbol{B})$.
- $\det(\lambda \boldsymbol{A}) = \lambda^n \det(\boldsymbol{A})$.
- *A square matrix* $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ *has* $\det(\boldsymbol{A}) \neq 0$ *if and only if* $\mathrm{rk}(\boldsymbol{A}) = n$. *In other words,* $\boldsymbol{A}$ *is invertible if and only if it is full rank.*

Trace

Trace is defined as:

$$\mathrm{tr}(\boldsymbol{A}) := \sum_{i=1}^{n} a_{ii}\,,$$

i.e. , the trace is the sum of the diagonal elements of A.

Properties of Trace:

$$\mathrm{tr}(\boldsymbol{A}+\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{A}) + \mathrm{tr}(\boldsymbol{B}) \text{ for } \boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$$
$$\mathrm{tr}(\alpha\boldsymbol{A}) = \alpha\mathrm{tr}(\boldsymbol{A})\,, \alpha \in \mathbb{R} \text{ for } \boldsymbol{A} \in \mathbb{R}^{n\times n}$$
$$\mathrm{tr}(\boldsymbol{I}_n) = n$$
$$\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{B}\boldsymbol{A}) \text{ for } \boldsymbol{A} \in \mathbb{R}^{n\times k}, \boldsymbol{B} \in \mathbb{R}^{k\times n}$$

System Of Equations: Matrix Formulation

If we consider the system of linear equations

$$
\begin{aligned}
2x_1 + 3x_2 + 5x_3 &= 1 \\
4x_1 - 2x_2 - 7x_3 &= 8 \\
9x_1 + 5x_2 - 3x_3 &= 2
\end{aligned}
\tag{2.35}
$$

and use the rules for matrix multiplication, we can write this equation system in a more compact form as

$$
\begin{bmatrix} 2 & 3 & 5 \\ 4 & -2 & -7 \\ 9 & 5 & -3 \end{bmatrix}
\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}
=
\begin{bmatrix} 1 \\ 8 \\ 2 \end{bmatrix} .
\tag{2.36}
$$

Triangular Matrix

$$A = \begin{bmatrix} a_{00} & 0 & 0 & 0 & 0 \\ a_{10} & a_{11} & 0 & 0 & 0 \\ a_{20} & a_{21} & a_{22} & 0 & 0 \\ a_{30} & a_{31} & a_{32} & a_{33} & 0 \\ a_{40} & a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

Lower Triangular Matrix

$$A = \begin{bmatrix} a_{00} & a_{01} & a_{02} & a_{03} & a_{04} \\ 0 & a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & 0 & a_{22} & a_{23} & a_{24} \\ 0 & 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & 0 & 0 & a_{44} \end{bmatrix}$$

Upper Triangular Matrix

Solve **Lx = b**

$$\begin{bmatrix} a_{11} & 0 & \dots & 0 \\ a_{21} & a_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

We solve for x as:

$$\begin{aligned} x_1 &= \frac{b_1}{a_{11}} \\ x_2 &= \frac{b_2 - a_{21}x_1}{a_{22}} \\ x_3 &= \frac{b_3 - a_{31}x_1 - a_{32}x_2}{a_{33}} \\ &\vdots \\ x_n &= \frac{b_n - a_{n1}x_1 - a_{n2}x_2 - \dots - a_{n,n1}x_{n1}}{a_{nn}}. \end{aligned}$$

Each addition, multiplication, division is considered 1 flop.

Flops Count :

$1 + 3 + 5 + \dots + (2n-1) = n^2$

Positive Definite (PD) Matrix

A matrix is said to be positive definite if it satisfies

1. The matrix is Symmetric
    - A symmetric matrix is a square matrix that is equal to its transpose
2. $x^T A x > 0 \; \forall x$

$A^T A$ is Positive Definite

$A^T A$ is Symmetric & Square

Quadratic form

$$x^T A^T A x = (Ax)^T (Ax) = \| Ax \|^2 > 0$$

Example

$$\mathbb{A} = \begin{bmatrix} 7 & 2 \\ 2 & 1 \end{bmatrix}$$

$$\mathbf{x}^T \mathbb{A} \mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 7 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 7x_1 + 2x_2 \\ 2x_1 + x_2 \end{bmatrix}$$

$$= 7x_1^2 + 2x_1x_2 + 2x_1x_2 + x_2^2$$

$$= 7x_1^2 + 4x_1x_2 + x_2^2 = f(x_1, x_2)$$

Cholesky Decomposition

Theorem 4.18 (Cholesky Decomposition). *A symmetric, positive definite matrix* $\boldsymbol{A}$ *can be factorized into a product* $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\top$, *where* $\boldsymbol{L}$ *is a lower-triangular matrix with positive diagonal elements:*

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} = \begin{bmatrix} l_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ l_{n1} & \cdots & l_{nn} \end{bmatrix} \begin{bmatrix} l_{11} & \cdots & l_{n1} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & l_{nn} \end{bmatrix} . \tag{4.44}$$

$\boldsymbol{L}$ *is called the Cholesky factor of* $\boldsymbol{A}$, *and* $\boldsymbol{L}$ *is unique.*

$$\begin{bmatrix} a_{11} & A_{21}^T \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} l_{11} & L_{21}^T \\ 0^T & L_{22} \end{bmatrix}$$

A matrix is postive semi definite (PSD) if the second constraint is relaxed as $x^T A x \geq 0 \; \forall x$

$$\begin{bmatrix} a_{11} & A_{21}^T \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} l_{11} & L_{21}^T \\ 0^T & L_{22}^T \end{bmatrix}$$

Now, we can form the following set of equations from above

- $a_{11} = l_{11} \cdot l_{11} \implies l_{11} = \sqrt{a_{11}}$
- $A_{21} = L_{21} \cdot l_{11} \implies L_{21} = \frac{A_{21}}{l_{11}}$
- $A_{22} = L_{21}L_{21}^T + L_{22}L_{22}^T \implies (A_{22} - L_{21}L_{21}^T) = L_{22}L_{22}^T$

Algorithm: Cholesky

1. Calculate $l_{11} = \sqrt{a_{11}}$
2. Calculate $L_{21} = \frac{A_{21}}{l_{11}}$
3. Use **Cholesky** to compute L_{22} as $(A_{22} - L_{21}L_{21}^T) = L_{22}L_{22}^T$

Question

Compute Cholesky Factorization of : $\begin{bmatrix} 1 & 2 & 3 \\ 2 & 20 & 26 \\ 3 & 26 & 70 \end{bmatrix}$

Solution

- $l_{11} = \sqrt{a_{11}} = 1$
- $L_{21} = \frac{A_{21}}{l_{11}} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$. Now, L $= \begin{bmatrix} 1 & 0 & 0 \\ 2 & l_{22} & 0 \\ 3 & l_{32} & l_{33}. \end{bmatrix}$
- We have to do Cholesky factorization of A_{22} - $\mathrm{L}_{21}\mathrm{L}_{21}{}^{\mathrm{T}} = \begin{bmatrix} 16 & 20 \\ 20 & 61 \end{bmatrix} = \begin{bmatrix} l_{22} & 0 \\ l_{32} & l_{33}. \end{bmatrix} \cdot \begin{bmatrix} l_{22} & l_{32} \\ 0 & l_{23} \end{bmatrix}$
- $l_{22} = \sqrt{a_{22}} = 4$
- $l_{23} = \frac{20}{4} = 5$
- The matrix is now $\begin{bmatrix} 1 & 0 & 0 \\ 2 & 4 & 0 \\ 3 & 5 & l_{33} \end{bmatrix}$
- We have to factorize 61 - $5 \cdot 5 = 36$. This gives l_{33} as 6
- The final answer is
$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 20 & 26 \\ 3 & 26 & 70 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 4 & 0 \\ 3 & 5 & 6 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{bmatrix}$$

Solving linear equations by Cholesky factorization

Solve $AX = b$ where A is PD matrix of order n. $LL^T x = b$

1. *Cholesky factorization:* factor A as LL^T ($(1/3)n^3$ flops).
2. *Forward substitution:* Solve $Lw = b$ (n^2 flops).
3. *Back substitution:* Solve $L^T x = w$ (n^2 flops).

Total cost is $(1/3)n^3 + 2n^2$ or roughly $(1/3)n^3$.

Eigenvectors and Eigenvalues

- An eigenvector is a vector whose direction remains unchanged when a linear transformation is applied to it.
- If the vector space V is finite-dimensional, then the linear transformation T can be represented as a square matrix A (n x n), and the vector x by a non-zero column vector (n x 1), then the eigenvalue equation or eigenequation is given by:

$$Ax = \lambda x \tag{1}$$

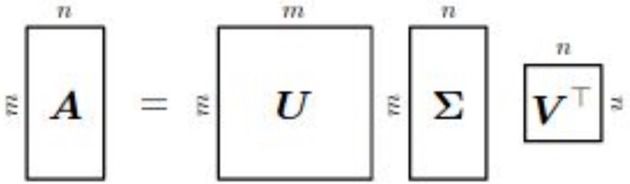
$$Ax - \lambda x = 0 \tag{2}$$

$$(A - \lambda I)x = 0 \tag{3}$$

- If x is non-zero, then the above eqn. will have a solution if $|A - \lambda I| = 0$. This equation is called the characteristic equation of A, and is an n^{th} order polynomial in λ with n roots.

SVD decomposition

Theorem 4.22 (SVD Theorem). *Let $A^{m\times n}$ be a rectangular matrix of rank $r \in [0, \min(m,n)]$. The SVD of A is a decomposition of the form*



(4.64)

with an orthogonal matrix $U \in \mathbb{R}^{m\times m}$ with column vectors u_i, $i = 1, \ldots, m$, and an orthogonal matrix $V \in \mathbb{R}^{n\times n}$ with column vectors v_j, $j = 1, \ldots, n$. Moreover, Σ is an $m \times n$ matrix with $\Sigma_{ii} = \sigma_i \geqslant 0$ and $\Sigma_{ij} = 0,\ i \neq j$.

$U^TU = UU^T = I_{m\times m}$, $V^TV = VV^T = I_{n\times n}$

SVD Decomposition

- The left-singular vectors of $\boldsymbol{A}$ are eigenvectors of $\boldsymbol{A}\boldsymbol{A}^\top$
- The right-singular vectors of $\boldsymbol{A}$ are eigenvectors of $\boldsymbol{A}^\top\boldsymbol{A}$.
- The non-zero singular values of $\boldsymbol{A}$ are the square roots of the non-zero eigenvalues of $\boldsymbol{A}\boldsymbol{A}^\top$ and are equal to the non-zero eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$.

Let λ be an eigenvalue of $A^T A$, i.e.

$$A^T A x = \lambda x$$

for some $x \neq 0$. We can multiply A from the left and get

$$A A^T (A x) = \lambda (A x).$$

In the SVD, the entries in the diagonal matrix Σ are all real and non-negative.
The diagonal entries σ_i , i = 1,..., r, of Σ are called the singular values.
By convention, the singular values are ordered, i.e., $\sigma_1 > \sigma_2 > \sigma_r > 0$.

Computing Inverse using SVD

Let us assume $A = UDV^T$

$$A^{-1} = VD^{-1}U^T$$

SVD Example:

http://www.d.umn.edu/~mhampton/m4326svd_example.pdf