Lecture 05: Linear Algebra

Introduction to Machine Learning

Sajjad Amini

Department of Electrical Engineering
Sharif University of Technology

Contents

References

Except explicitly cited, the reference for the material in slides is:

- Murphy, K. P. (2022). *Probabilistic machine learning: an introduction.* MIT press.

Section 1

Basic Definitions

Vectors

In this course we assume column vectors represented by:

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = (x_1, x_2, \ldots, x_n)$$

Matrices

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix}$$

Basic Definitions

Matrix Rows

$$\boldsymbol{A} = \begin{bmatrix} -\boldsymbol{A}_{1,:}^T- \\ -\boldsymbol{A}_{2,:}^T- \\ \vdots \\ -\boldsymbol{A}_{m,:}^T- \end{bmatrix} = \left[\boldsymbol{A}_{1,:}^T \quad ; \quad \boldsymbol{A}_{2,:}^T \quad ; \quad \dots \quad ; \quad \boldsymbol{A}_{m,:}^T\right]$$

Matrix Columns

$$\boldsymbol{A} = \begin{bmatrix} | & | & & | \\ \boldsymbol{A}_{:,1} & \boldsymbol{A}_{:,2} & \dots & \boldsymbol{A}_{:,n} \\ | & | & & | \end{bmatrix} = \left[\boldsymbol{A}_{:,1} \quad , \quad \boldsymbol{A}_{:,2} \quad , \quad \dots \quad , \quad \boldsymbol{A}_{:,n}\right]$$

Vectorizing

Vectorizing Operator

$$vec(\boldsymbol{A}) = [\boldsymbol{A}_{:,1}; \ldots; \boldsymbol{A}_{:,n}] \in \mathbb{R}^{mn\times 1}$$

I-vectorizing Operator

$$\boldsymbol{A} = ivec(vec(\boldsymbol{A}), \mathcal{O})$$

Section 2

Vector Space

Vector Space

Vector Space

A vector space is a set of vectors $\boldsymbol{x} \in \mathbb{R}^n$, denoted $\mathcal{V}$, such that:

- It is closed under vector addition: $if\ \boldsymbol{x}, \boldsymbol{y} \in \mathcal{V} \Rightarrow \boldsymbol{x} + \boldsymbol{y} \in \mathcal{V}$
- It is closed under multiplication by a real scalar $c \in \mathbb{R}$: $if\ \boldsymbol{x} \in \mathcal{V} \Rightarrow c\boldsymbol{x} \in \mathbb{R}$

Linear Independence

A set of vectors $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ is said to be (linearly) dependent if:

$$\exists j :\ \boldsymbol{x}_j = \sum_{i, i \neq j} \boldsymbol{x}_i$$

Otherwise the set is said to be (linearly) independent.

Span

The span of a set of vectors $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ is defined as:

$$span(\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}) \triangleq \left\{ \boldsymbol{v} : \boldsymbol{v} = \sum_{i=1}^{n} \alpha_i \boldsymbol{x}_i, \alpha_i \in \mathbb{R} \right\}$$

Section 3

Products

Matrix-Vector Product

Matrix-Vector Product

Assume $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{x} \in \mathbb{R}^n$. Then the product vector $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \in \mathbb{R}^m$ can be viewed as follows:

View 1

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} = \begin{bmatrix} - & \widehat{\boldsymbol{a}}_1^T & - \\ - & \widehat{\boldsymbol{a}}_2^T & - \\ & \vdots & \\ - & \widehat{\boldsymbol{a}}_m^T & - \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} \widehat{\boldsymbol{a}}_1^T \boldsymbol{x} \\ \widehat{\boldsymbol{a}}_2^T \boldsymbol{x} \\ \vdots \\ \widehat{\boldsymbol{a}}_m^T \boldsymbol{x} \end{bmatrix}$$

View 2

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} = \begin{bmatrix} | & | & & | \\ \boldsymbol{a}_1 & \boldsymbol{a}_2 & \dots & \boldsymbol{a}_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} | \\ \boldsymbol{a}_1 \\ | \end{bmatrix} x_1 + \begin{bmatrix} | \\ \boldsymbol{a}_2 \\ | \end{bmatrix} x_2 + \dots + \begin{bmatrix} | \\ \boldsymbol{a}_n \\ | \end{bmatrix} x_n$$

Matrix-Matrix Product

Matrix-Matrix Product

Assume $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{B} \in \mathbb{R}^{n \times p}$. Then the product vector $\boldsymbol{C} = \boldsymbol{AB} \in \mathbb{R}^{m \times p}$ can be viewed as follows:

View 1

$$\boldsymbol{C} = \boldsymbol{AB} = \begin{bmatrix} - & \widehat{\boldsymbol{a}}_1^T & - \\ - & \widehat{\boldsymbol{a}}_2^T & - \\ & \vdots & \\ - & \widehat{\boldsymbol{a}}_m^T & - \end{bmatrix} \begin{bmatrix} | & | & & | \\ \boldsymbol{b}_1 & \boldsymbol{b}_2 & \dots & \boldsymbol{b}_p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} \widehat{\boldsymbol{a}}_1^T \boldsymbol{b}_1 & \widehat{\boldsymbol{a}}_1^T \boldsymbol{b}_2 & \dots & \widehat{\boldsymbol{a}}_1^T \boldsymbol{b}_p \\ \widehat{\boldsymbol{a}}_2^T \boldsymbol{b}_1 & \widehat{\boldsymbol{a}}_2^T \boldsymbol{b}_2 & \dots & \widehat{\boldsymbol{a}}_2^T \boldsymbol{b}_p \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\boldsymbol{a}}_m^T \boldsymbol{b}_1 & \widehat{\boldsymbol{a}}_m^T \boldsymbol{b}_2 & \dots & \widehat{\boldsymbol{a}}_m^T \boldsymbol{b}_p \end{bmatrix}$$

View 2

$$\boldsymbol{C} = \boldsymbol{AB} = \begin{bmatrix} | & | & & | \\ \boldsymbol{a}_1 & \boldsymbol{a}_2 & \dots & \boldsymbol{a}_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} - & \widehat{\boldsymbol{b}}_1^T & - \\ - & \widehat{\boldsymbol{b}}_2^T & - \\ & \vdots & \\ - & \widehat{\boldsymbol{b}}_n^T & - \end{bmatrix} = \sum_{i=1}^{n} \boldsymbol{a}_i \widehat{\boldsymbol{b}}_i^T$$

Matrix-Matrix Product

Matrix-Matrix Product

Assume $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{B} \in \mathbb{R}^{n \times p}$. Then the product vector $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{B} \in \mathbb{R}^{m \times p}$ can be viewed as follows:

View 3

$$\boldsymbol{C} = \boldsymbol{A}\boldsymbol{B} = \boldsymbol{A} \begin{bmatrix} | & | & & | \\ \boldsymbol{b}_1 & \boldsymbol{b}_2 & \dots & \boldsymbol{b}_p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \boldsymbol{A}\boldsymbol{b}_1 & \boldsymbol{A}\boldsymbol{b}_2 & \dots & \boldsymbol{A}\boldsymbol{b}_p \\ | & | & & | \end{bmatrix}$$

View 4

$$\boldsymbol{C} = \boldsymbol{A}\boldsymbol{B} = \begin{bmatrix} - & \widehat{\boldsymbol{a}}_1^T & - \\ - & \widehat{\boldsymbol{a}}_2^T & - \\ & \vdots & \\ - & \widehat{\boldsymbol{a}}_m^T & - \end{bmatrix} \boldsymbol{B} = \begin{bmatrix} - & \widehat{\boldsymbol{a}}_1^T\boldsymbol{B} & - \\ - & \widehat{\boldsymbol{a}}_2^T\boldsymbol{B} & - \\ & \vdots & \\ - & \widehat{\boldsymbol{a}}_m^T\boldsymbol{B} & - \end{bmatrix}$$

Range and Null Spaces

Range of a Matrix

Assume $\boldsymbol{A} \in \mathbb{R}^{m \times n}$. The range or columns space of $\boldsymbol{A}$ is the span of the columns of $\boldsymbol{A}$ as:

$$\text{range}(\boldsymbol{A}) \triangleq \{\boldsymbol{v} \in \mathbb{R}^m : \boldsymbol{v} = \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} \in \mathbb{R}^n\}$$

Null Space of a Matrix

Assume $\boldsymbol{A} \in \mathbb{R}^{m \times n}$. The null space of $\boldsymbol{A}$ is the set of all vectors $\boldsymbol{x}$ that get mapped to the null vector when multiplied by $\boldsymbol{A}$ as:

$$\text{nullspace}(\boldsymbol{A}) \triangleq \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\}$$

Section 4

Norms

Vector Norms

Definition

Norm is any function $f : \mathbb{R}^n \to \mathbb{R}$ that satisfies the following properties:

1. $\forall \boldsymbol{x} \in \mathbb{R}^n \Rightarrow f(\boldsymbol{x}) \geq 0$ (non-negativity)
2. $f(\boldsymbol{x}) = 0$ iff $\boldsymbol{x} = 0$ (definiteness)
3. $\forall \boldsymbol{x} \in \mathbb{R}^n, \forall t \in \mathbb{R} \Rightarrow f(t\boldsymbol{x}) = |t| f(x)$ (absolute value homogeneity)
4. $\forall \boldsymbol{x} \in \mathbb{R}^n, \forall \boldsymbol{y} \in \mathbb{R}^n \Rightarrow f(\boldsymbol{x} + \boldsymbol{y}) \leq f(\boldsymbol{x}) + f(\boldsymbol{y})$ (triangle inequality)

Examples of Vector Norm

- p-norm (ℓ_p): $\|\boldsymbol{x}|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}, p \geq 1 \Rightarrow \begin{cases} \ell_1 : \|\boldsymbol{x}\|_1 = \sum_{i=1}^n |x_i| \\ \ell_2 : \|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2} \\ \ell_\infty : \|\boldsymbol{x}\|_\infty = \max_i |x_i| \end{cases}$
- 0-norm (ℓ_0): $|\boldsymbol{x}\|_0 = \sum_{i=1}^n \mathbb{I}(|x_i| > 0)$ (Pseudo norm due to inhomogeneity)

Matrix Norms

Examples of Matrix Norm

Assume matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, then:

- p-norm (ℓ_p): $\|\boldsymbol{A}\|_p = \max_{\boldsymbol{x} \neq 0} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_p}{\|\boldsymbol{x}\|_p} = \max_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}\boldsymbol{x}\|_p$
- Frobenius norm (ℓ_F): $\|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2} = \|\operatorname{vec}(\boldsymbol{A})\|_2$

Section 5

Matrix Operators

Trace of a Square Matrix

Definition

The trace of a square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, denoted $\mathrm{tr}(\boldsymbol{A})$, is the sum of diagonal elements in the matrix as:

$$\mathrm{tr}(\boldsymbol{A}) \triangleq \sum_{i=1}^{n} A_{ii}$$

Properties

Assume matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n \times n}$ and scalar $c \in \mathbb{R}$.

- $\mathrm{tr}(\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{A}^T)$
- $\mathrm{tr}(\boldsymbol{A} + \boldsymbol{B}) = \mathrm{tr}(\boldsymbol{A}) + \mathrm{tr}(\boldsymbol{B})$
- $\mathrm{tr}(c\boldsymbol{A}) = c\ \mathrm{tr}(\boldsymbol{A})$
- $\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{B}\boldsymbol{A})$

Trace of a Square Matrix

Cyclic Permutation Property

For real matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ where $\boldsymbol{ABC}$ is square, then we have:

$$\operatorname{tr}(\boldsymbol{ABC}) = \operatorname{tr}(\boldsymbol{CAB}) = \operatorname{tr}(\boldsymbol{BCA})$$

Determinant of a Square Matrix

Minor

Assume $\boldsymbol{A} \in \mathbb{R}^{n \times n}$. The (i, j) minor, denoted $\boldsymbol{A}_{ij}$ is the matrix obtained from $\boldsymbol{A}$ by deleting the i-th row and the j-th column.

Cofactor

Assume $\boldsymbol{A} \in \mathbb{R}^{n \times n}$. The (i, j) cofactor, denoted C_{ij} is: $C_{ij} = (-1)^{i+j} \det(\boldsymbol{A}_{ij})$, where $\det(\boldsymbol{A}_{ij})$ is the determinant of (i, j) minor.

Determinant

The determinant of a square matrix, denoted $\det(\boldsymbol{A})$ or $|\boldsymbol{A}|$, is a measure of how much it changes a unit volume when viewed as a linear transformation and is defined as:

$$\det(\boldsymbol{A}) = \sum_{i=1}^{n} a_{i1} C_{i1}$$

Condition Number of a Square Matrix

Condition Number

The condition number of a square matrix $\boldsymbol{A}$ is a measure for the stability of linear equation set $\boldsymbol{Ax} = \boldsymbol{b}$ and is defined as follows: $\kappa(\boldsymbol{A}) \triangleq \|\boldsymbol{A}\| \times \|\boldsymbol{A}^{-1}\|$. A suitable option for the matrix norm is ℓ_2 norm which result in $\kappa(\boldsymbol{A}) \geq 1$.

Matrix Conditioning

Assume square matrix $\boldsymbol{A}$. Based on the condition number, this matrix can be divided into two categories:

- $\boldsymbol{A}$ is ill-conditioned if $\kappa(\boldsymbol{A})$ is large.
- $\boldsymbol{A}$ is well-conditioned if $\kappa(\boldsymbol{A})$ is small (close to 1).

Condition Number of a Square Matrix

Frame Title

In a linear system of equations $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, assume we change $\boldsymbol{b}$ to $\boldsymbol{b}+\Delta\boldsymbol{b}$. Compute the change in $\boldsymbol{x}$ vector ($\Delta\boldsymbol{x}$) for the following two matrices:

- $\boldsymbol{A} = 0.1\boldsymbol{I}_{100\times 100}$ ($\kappa(\boldsymbol{A}) = 1, \det(\boldsymbol{A}) = 10^{-100}$):

$$\Delta\boldsymbol{x} = \boldsymbol{A}^{-1}\Delta\boldsymbol{b} = 10\boldsymbol{I}\Delta\boldsymbol{b} = 10\Delta\boldsymbol{b}$$

- $\boldsymbol{A} = 0.5\begin{bmatrix} 1 & 1 \\ 1+10^{-10} & 1-10^{-10} \end{bmatrix}$ ($\kappa(\boldsymbol{A}) = 2\times 10^{10}, \det(\boldsymbol{A}) = -2\times 10^{-10}$):

$$\Delta\boldsymbol{x} = \boldsymbol{A}^{-1}\Delta\boldsymbol{b} = 1\begin{bmatrix} \Delta b_1 - 10^{10}(\Delta b_1 - \Delta b_2) \\ \Delta b_2 + 10^{10}(\Delta b_1 - \Delta b_2) \end{bmatrix}$$

Section 6

Special Matrices

Diagonal Matrix

- Diagonal matrix:

$$\boldsymbol{D} = \begin{bmatrix} d_1 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & d_n \end{bmatrix} = \operatorname{diag}(d_1, d_2, \ldots, d_n)$$

- Block diagonal Matrix: A square matrix with square matrices in the main diagonal blocks and zero matrices in all off-diagonal blocks as:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 & & & \\ & \boldsymbol{A}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{A}_n \end{bmatrix} = \operatorname{diag}(\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_n)$$

Special Matrices

Band-diagonal Matrix

A band-diagonal matrix only has non-zero entries along the diagonal, and on k sides of the diagonal (k is known as bandwidth).

Tridiagonal Matrix

Tridiagonal matrix is a band-diagonal matrix with $k = 1$. A sample 6×6 tridiagonal matrix is:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & 0 & 0 & 0 & 0 \\ a_{21} & a_{22} & a_{23} & 0 & 0 & 0 \\ 0 & a_{32} & a_{33} & a_{34} & 0 & 0 \\ 0 & 0 & a_{43} & a_{44} & a_{45} & 0 \\ 0 & 0 & 0 & a_{54} & a_{55} & a_{56} \\ 0 & 0 & 0 & 0 & a_{65} & a_{66} \end{bmatrix}$$

Triangular Matrix

Lower Triangular Matrix

$$
\boldsymbol{L} = \begin{bmatrix}
l_{11} & & & & \\
l_{21} & l_{22} & & & \\
l_{31} & l_{32} & l_{33} & & \\
\vdots & \vdots & \ddots & \ddots & \\
l_{n1} & l_{n2} & \dots & l_{n(n-1)} & l_{nn}
\end{bmatrix}
$$

Upper Triangular Matrix

$$
\boldsymbol{U} = \begin{bmatrix}
u_{11} & u_{12} & u_{13} & \dots & u_{1n} \\
& u_{22} & u_{23} & \dots & u_{2n} \\
& & \ddots & \ddots & \vdots \\
& & & \ddots & u_{(n-1)n} \\
& & & & u_{nn}
\end{bmatrix}
$$

Definite and Indefinite Matrices

Symmetric Matrix

Matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is symmetric iff $\boldsymbol{A} = \boldsymbol{A}^T$ (We usually show this by $\boldsymbol{A} \in \mathbb{S}^n$)

Definite and Indefinite Matrices

Suppose $\boldsymbol{A} \in \mathbb{S}^n$ and arbitrary nonzero vector $\boldsymbol{v} \in \mathbb{R}^n \setminus \{0\}$ then:

- $\boldsymbol{A}$ is positive definite (PD), denoted $\boldsymbol{A} \succ 0$, iff: $\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v} > 0$
- $\boldsymbol{A}$ is positive semidefinite (PSD), denoted $\boldsymbol{A} \succeq 0$, iff: $\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v} \geq 0$
- $\boldsymbol{A}$ is negative definite (ND), denoted $\boldsymbol{A} \prec 0$, iff: $\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v} < 0$
- $\boldsymbol{A}$ is negative semidefinite (NSD), denoted $\boldsymbol{A} \preceq 0$, iff: $\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v} \leq 0$
- $\boldsymbol{A}$ is indefinite iff it is none of the above.

Orthogonal Square Matrices

$\boldsymbol{A} = \begin{bmatrix} | & | & & | \\ \boldsymbol{a}_1 & \boldsymbol{a}_2 & \dots & \boldsymbol{a}_n \\ | & | & & | \end{bmatrix}$ is orthogonal iff:

$$\boldsymbol{a}_i^T \boldsymbol{a}_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

Section 7

Inverse Matrix

Inverse Matrix

Inverse Matrix

The inverse of a square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, denoted $\boldsymbol{A}^{-1}$, is the unique matrix such that:

$$\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}$$

Singular Matrix

$\boldsymbol{A}^{-1}$ exists iff $\det(\boldsymbol{A}) \neq 0$. If $\det(\boldsymbol{A}) = 0$, $\boldsymbol{A}$ is called a singular matrix.

Section 8

Eigenvalue Decomposition

Eigenvalue and Eigenvector

Eigenvalue and Eigenvector

Assume a square matrix $\boldsymbol{A} \in \mathbb{R}^{2\times 2}$, we say that $\lambda \in \mathbb{R}$ is an eigenvalue of $\boldsymbol{A}$ and $\boldsymbol{u} \in \mathbb{R}^n$ is the corresponding eigenvector if:

$$\boldsymbol{A}\boldsymbol{u} = \lambda \boldsymbol{u}, \ \boldsymbol{u} \neq \boldsymbol{0}$$

"The" Eigenvector

For any eigenvector $\boldsymbol{u} \in \mathbb{R}^n \backslash \{\boldsymbol{0}\}$ and scalar $c \in \mathbb{R}\backslash\{0\}$, $c\boldsymbol{u}$ is also an eigenvector. "The" eigenvector is normalized to have unit length.

Eigenvalue and Eigenvector

Characteristic Equation

$(\lambda, \boldsymbol{u})$ is (eigenvalue,eigenvector) pair if:

$$(\lambda \boldsymbol{I} - \boldsymbol{A})\boldsymbol{u} = \boldsymbol{0}, \ \boldsymbol{u} \neq \boldsymbol{0}$$

Thus:

- $\boldsymbol{u}$ is in the nullspace of $\lambda \boldsymbol{I} - \boldsymbol{A}$.
- $\det(\boldsymbol{A}) = 0$

Equation $\det(\boldsymbol{A}) = 0$ is called characteristic equation.

Characteristic Equation

- The order of characteristic equation is n.
- Characteristic equation has n roots, denoted $\lambda_1, \ldots, \lambda_n$, possibly complex.
- $\boldsymbol{u}_i$ corresponding to λ_i can be easily found by finding the nullspace of $\lambda_i \boldsymbol{I} - \boldsymbol{A}$ matrix.

Eigenvalue and Eigenvector

Eigenvalue and Eigenvector

Find the eigenvalues and eigenvectors of $\boldsymbol{A} = \begin{bmatrix} 0.8 & 0.3 \\ 0.2 & 0.7 \end{bmatrix}$.

Solution:

$$\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = \det\left(\begin{bmatrix} 0.8-\lambda & 0.3 \\ 0.2 & 0.7-\lambda \end{bmatrix}\right) = (\lambda - 1)(\lambda - 0.5) = 0$$

$$\Rightarrow \begin{cases} \lambda_1 = 1 \\ \lambda_2 = 0.5 \end{cases}$$

$$(\boldsymbol{A} - \lambda_1 \boldsymbol{I})\boldsymbol{u}_1 = \mathbf{0} \Rightarrow \begin{bmatrix} -0.2 & 0.3 \\ 0.2 & -0.3 \end{bmatrix} \boldsymbol{u}_1 = \mathbf{0} \Rightarrow \boldsymbol{u}_1 = \begin{bmatrix} 1.5 \\ 1 \end{bmatrix}$$

$$(\boldsymbol{A} - \lambda_2 \boldsymbol{I})\boldsymbol{u}_2 = \mathbf{0} \Rightarrow \begin{bmatrix} 0.3 & 0.3 \\ 0.2 & 0.2 \end{bmatrix} \boldsymbol{u}_2 = \mathbf{0} \Rightarrow \boldsymbol{u}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Eigenvalue and Eigenvector

Rank

The rank of matrix $\boldsymbol{A}$ is equal to the number of non-zero eigenvalues of $\boldsymbol{A}$.

Connection to Trace and Determinant

Assume $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$. Then:

- The rank of $\boldsymbol{A}$ equals to the number of non-zero eigenvalues of $\boldsymbol{A}$.
- $\boldsymbol{A}^{-1}$ shares the eigenvector with $\boldsymbol{A}$ while its eigenvalues are $\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \ldots, \frac{1}{\lambda_n}$.
- Symmetric matrix $\boldsymbol{A}$ is PD iff $\lambda_i > 0, i = 1, \ldots, n$.
- Symmetric matrix $\boldsymbol{A}$ is PSD iff $\lambda_i \geq 0, i = 1, \ldots, n$.
- $\mathrm{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} \lambda_i$
- $\det(\boldsymbol{A}) = \prod_{i=1}^{n} \lambda_i$

Diagonalizable

Diagonalizable

As we see: $\boldsymbol{A}\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i, i = 1, \ldots, n$
We can write the above equalities as:

$$\boldsymbol{A}\boldsymbol{U} = \boldsymbol{U}\boldsymbol{\Lambda}$$

where:

- $\boldsymbol{U} \in \mathbb{R}^{n \times n} = \begin{bmatrix} | & | & & | \\ \boldsymbol{u}_1 & \boldsymbol{u}_2 & & \boldsymbol{u}_n \\ | & | & & | \end{bmatrix}$
- $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$

Now assume that matrix $\boldsymbol{U}$ is invertible. Then:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{-1}$$

A matrix that can be written in this form is called diagonalizable.

Eigenvalues and Eigenvectors of Symmetric Matrices

Based on *Spectral Theorem*, for symmetric matrices we have:

- All eigenvalues are real
- Eigenvectors are orthonormal ($\boldsymbol{U}$ is orthogonal thus $\boldsymbol{U}^{-1} = \boldsymbol{U}^T$)

Then we have:

$$
\begin{aligned}
\boldsymbol{A} =& \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T = \begin{bmatrix} | & | & & | \\ \boldsymbol{u}_1 & \boldsymbol{u}_2 & & \boldsymbol{u}_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \begin{bmatrix} - & \boldsymbol{u}_1^T & - \\ - & \boldsymbol{u}_2^T & - \\ & \vdots & \\ - & \boldsymbol{u}_m^T & - \end{bmatrix} \\
=& \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T
\end{aligned}
$$

Data Whitening Using Eigenvectors

Data Whitening Using Eigenvectors

Suppose we have a dataset $\boldsymbol{X} \in \mathbb{R}^{N\times D}$ where the empirical mean vector is zero and empirical covariance matrix is $\boldsymbol{\Sigma} = \frac{1}{N}\boldsymbol{X}^T\boldsymbol{X}$. Find matrix $\boldsymbol{W} \in \boldsymbol{R}^{D\times D}$ such that empirical covariance matrix for transformed vector $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ is $\boldsymbol{I}$.

Solution: Matrix $\boldsymbol{\Sigma}$ is symmetric, thus $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$. Assume $\boldsymbol{W} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{U}^T$, then the covariance matrix for $\boldsymbol{y}$ is:

$$
\begin{aligned}
\mathrm{Cov}[\boldsymbol{y}] =& \frac{1}{N}\boldsymbol{Y}^T\boldsymbol{Y} = \frac{1}{N}(\boldsymbol{X}\boldsymbol{W}^T)^T(\boldsymbol{X}\boldsymbol{W}^T) = \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T \\
=& \boldsymbol{D}^{-\frac{1}{2}}\underbrace{\boldsymbol{U}^T\boldsymbol{U}}_{\boldsymbol{I}}\boldsymbol{D}\underbrace{\boldsymbol{U}^T\boldsymbol{U}}_{\boldsymbol{I}}\boldsymbol{D}^{-\frac{1}{2}} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{D}\boldsymbol{D}^{-\frac{1}{2}} = \boldsymbol{I}
\end{aligned}
$$

Section 9

Matrix Calculus

Gradient

Gradient

Assume function $f : \mathbb{R}^n \to \mathbb{R}$. The gradient vector of this function at a point $\boldsymbol{x}$ is the vector of partial derivatives as:

$$\boldsymbol{g} = \frac{\partial f}{\partial \boldsymbol{x}} = \nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

To emphasize the gradient evaluation point we write:

$$\boldsymbol{g}(\boldsymbol{x}^\star) \triangleq \left.\frac{\partial f}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}=\boldsymbol{x}^\star}$$

Hessian

Hessian

Assume function $f : \mathbb{R}^n \to \mathbb{R}$. The Hessian matrix of this function is the matrix of second partial derivatives as:

$$\boldsymbol{H}_f = \frac{\partial^2 f}{\partial \boldsymbol{x}^2} = \nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \cdots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

Jacobian

Jacobian

Assume function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$. The Jacobian matrix of this function is an $m \times n$ matrix of partial derivatives as:

$$\boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{x}) = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}^T} \triangleq \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla f_1(\boldsymbol{x})^T \\ \vdots \\ \nabla f_m(\boldsymbol{x})^T \end{bmatrix}$$