

Inclass 02: Linear Algebra and Probability

[SCS4049] Machine Learning and Data Science

Seongsik Park (s.park@dgu.ac.kr)

School of AI Convergence & Department of Artificial Intelligence, Dongguk University

Linear Algebra

- Notation for a column vector

$$\mathbf{x} = (x_1, x_2, \dots, x_n) = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} \in \mathcal{R}^n \quad (1)$$

- Transpose and a row vector

$$\mathbf{x}^T = (x_1, x_2, \dots, x_n)^T = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix} \quad (2)$$

- Magnitude of a vector (= 2-norm)

$$\|\mathbf{x}\| = \|\mathbf{x}\|_2 = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2} \quad (3)$$

$$\|\mathbf{x}\| = \|\mathbf{x}\|_2 = 0 \iff \mathbf{x} = \mathbf{0} \quad (4)$$

Vector operation

- Addition

$$\mathbf{x} + \mathbf{y} = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \dots \\ x_n + y_n \end{bmatrix} \quad (5)$$

where $\mathbf{x}, \mathbf{y} \in \mathcal{R}^n$

- Scalar product

$$\lambda \mathbf{x} = \lambda \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} = \begin{bmatrix} \lambda x_1 \\ \lambda x_2 \\ \dots \\ \lambda x_n \end{bmatrix} \quad (6)$$

Vector operation

- Inner product

$$\mathbf{x} = (x_1, x_2, \dots, x_n) \quad \mathbf{y} = (y_1, y_2, \dots, y_n) \quad (7)$$

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^T \mathbf{y} = \sum_{i=1}^n x_i y_i = \|\mathbf{x}\| \|\mathbf{y}\| \cos \theta \quad (8)$$

- Property

- Distributiveness: $(\mathbf{x} + \mathbf{y}) \cdot \mathbf{z} = \mathbf{x} \cdot \mathbf{z} + \mathbf{y} \cdot \mathbf{z}$ and $\mathbf{x} \cdot (\mathbf{y} + \mathbf{z}) = \mathbf{x} \cdot \mathbf{y} + \mathbf{x} \cdot \mathbf{z}$
- Linearity: $(\lambda \mathbf{x}) \cdot \mathbf{y} = \mathbf{x} \cdot (\lambda \mathbf{y}) = \lambda(\mathbf{x} \cdot \mathbf{y})$
- Symmetry: $\mathbf{x} \cdot \mathbf{y} = \mathbf{y} \cdot \mathbf{x}$
- Non-negativity: $\forall \mathbf{x} \neq \mathbf{0}, \mathbf{x} \cdot \mathbf{x} > 0$ and $\mathbf{x} = \mathbf{0} \iff \mathbf{x} \cdot \mathbf{x} = 0$

- A norm on a vector space Ω is a function $f: \Omega \rightarrow \mathcal{R}$ that satisfies following properties:
 - Positive scalability: $f(\lambda \mathbf{x}) = |\lambda|f(\mathbf{x})$
 - Triangle inequality: $f(\mathbf{x} + \mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$
 - If $f(\mathbf{x}) = 0$, then $\mathbf{x} = 0$
- Examples of norm
 - 1-norm: $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$
 - 2-norm: $\|\mathbf{x}\|_2 = (\sum_{i=1}^n |x_i|^2)^{\frac{1}{2}}$
 - p-norm: $\|\mathbf{x}\|_p = (\sum_{i=1}^n |x_i|^p)^{\frac{1}{p}}$
 - Infinity norm: $\|\mathbf{x}\|_\infty = \max_i |x_i|$

- Inner product revisited

$$\mathbf{x} = (x_1, x_2, \dots, x_n) \quad \mathbf{y} = (y_1, y_2, \dots, y_n) \quad (9)$$

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^T \mathbf{y} = \sum_{i=1}^n x_i y_i = \|\mathbf{x}\| \|\mathbf{y}\| \cos \theta \quad (10)$$

$$\|\mathbf{x}\|_2^2 = \sum_{i=1}^n |x_i|^2 = \mathbf{x} \cdot \mathbf{x} = \mathbf{x}^T \mathbf{x} \quad (11)$$

- Notation for a 2D array of scalars

$$\mathbf{A} \in \mathcal{R}^{m \times n}, \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, (\mathbf{A})_{ij} = a_{ij} \quad (12)$$

- Identity matrix $\mathbf{I}_n$ is the $n \times n$ square matrix with ones on the main diagonal and zeros elsewhere

$$\mathbf{I}_m \mathbf{A} = \mathbf{A} \mathbf{I}_n = \mathbf{A} \quad (13)$$

Matrix operation

- Addition

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \cdots & a_{2n} + b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \cdots & a_{mn} + b_{mn} \end{bmatrix} \quad (14)$$

$$(\mathbf{A} + \mathbf{B})_{ij} = a_{ij} + b_{ij} \quad (15)$$

- Commutative: $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$
- Associative: $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$
- Transpose: $(\mathbf{A}^T)_{ij} = (\mathbf{A})_{ji}$
 - $(\mathbf{A}^T)^T = \mathbf{A}$
 - $(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T$

- Multiplication

$$(\mathbf{AB})_{ij} = \sum_k a_{ik} b_{kj} \quad (16)$$

where $\mathbf{A} \in \mathcal{R}^{m \times k}$, $\mathbf{B} \in \mathcal{R}^{k \times n}$, then $\mathbf{AB} \in \mathcal{R}^{m \times n}$

- Associative: $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$
- Distributive: $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$
- Not commutative: $\mathbf{AB} \neq \mathbf{BA}$

Matrix vector multiplication

For $\mathbf{x} \in \mathcal{R}^n, \mathbf{y} \in \mathcal{R}^m, \mathbf{A} \in \mathcal{R}^{m \times n}$,

$$\mathbf{y} = \mathbf{A}\mathbf{x} \quad (17)$$

$$= \begin{bmatrix} | & | & & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 + \cdots + \mathbf{a}_n x_n \quad (18)$$

$$= \begin{bmatrix} - & \tilde{\mathbf{a}}_1^T & - \\ - & \tilde{\mathbf{a}}_2^T & - \\ & \vdots & \\ - & \tilde{\mathbf{a}}_m^T & - \end{bmatrix} \mathbf{x} = \begin{bmatrix} \tilde{\mathbf{a}}_1^T \mathbf{x} \\ \tilde{\mathbf{a}}_2^T \mathbf{x} \\ \vdots \\ \tilde{\mathbf{a}}_m^T \mathbf{x} \end{bmatrix} \quad (19)$$

Probability

Random variables

- A **random variable** X takes on a defined set of values with different probabilities
 - For example, if you roll a die, the outcome is random (not fixed) and there are 6 possible outcomes, each of which occur with probability $1/6$
 - For example, if you poll people about their voting preferences, the percentage of the sample that responds “Yes on Proposition A” is a random variable (the percentage will be slightly different every time you poll)
- Roughly, **probability** is how frequently we expect different outcomes to occur if we repeat the experiment over and over

Discrete and continuous random variable

- **Discrete** random variables have a countable number of outcomes
 - Dead/live, dice, counts, etc.
 - Probability mass function (pmf, p.m.f.)
- **Continuous** random variables have an infinite continuum of possible values
 - Blood pressure, weight, real number, etc.
 - Probability density function (pdf, p.d.f.)
- **Probability distribution** $F_X(x) = \Pr(X \leq x)$
- **Probability mass function** $p_X(x) = \Pr(X = x)$
- **Probability density function** $f_X(x) = \frac{d}{dx}F_X(x)$

Definition For any two events A and B (with $\Pr\{B\} > 0$), the conditional probability of A , conditional on B , is defined by

$$\Pr\{A \mid B\} = \Pr\{A \cap B\} / \Pr\{B\}$$

Definition Two events, A and B , are statistically independent if

$$\Pr\{A \cap B\} = \Pr\{A\}\Pr\{B\}$$

Two rv's, say X and Y , are *statistically independent* if

$$F_{XY}(x, y) = F_X(x)F_Y(y) \text{ for each value } x_i \text{ of } X \text{ and } y_j \text{ of } Y$$

Discrete rv's

$$p_{XY}(x_i, y_j) = p_X(x_i)p_Y(y_j)$$

$$p_{X|Y}(x_i | y_j) = p_X(x_i)$$

Continuous rv's

$$f_{XY}(x, y) = f_X(x)f_Y(y)$$

$$f_{X|Y}(x | y) = f_X(x)$$

Expectation

The expected value $E[X]$ of a random variable X is also called the expectation or the mean and is frequently denoted as $\bar{X}$. Considering nonnegative discrete rv's, the expected value is then given by

$$E[X] = \sum_x xp_X(x).$$

- Discrete case: $E(X) = \sum_x xp_X(x)$
- Continuous case: $E(X) = \int_x xf_X(x)dx$

Variance and standard deviation

The *variance* is denoted by σ_X^2 or $\text{Var}[X]$. It is given by

$$\sigma_X^2 = \text{E}[(X - \bar{X})^2] = \text{E}[X^2] - \bar{X}^2$$

The *standard deviation* σ_X of X is the square root of the variance and provides a measure of dispersion of the rv around the mean. Thus the mean is a rough measure of typical values for the outcome of the rv, and σ_X is a measure of the typical difference between X and $\bar{X}$.

Example: expectation and variance

Table 1: Example: discrete random variable

X	10	11	12	13	14
$p_X(x)$	0.4	0.2	0.2	0.1	0.1

- Expectation

$$E(X) = \sum_x x p_X(x) \quad (20)$$

$$= 10 \cdot 0.4 + 11 \cdot 0.2 + 12 \cdot 0.2 + 13 \cdot 0.1 + 14 \cdot 0.1 \quad (21)$$

$$= 11.3 \quad (22)$$

Example: expectation and variance

Table 2: Example: discrete random variable

X	10	11	12	13	14
$p_X(x)$	0.4	0.2	0.2	0.1	0.1

- Variance

$$E[(X - E(X))^2] = \sum_x (x - 11.3)^2 p_X(x) \quad (23)$$

$$= (10 - 11.3)^2 \cdot 0.4 + (11 - 11.3)^2 \cdot 0.2 \quad (24)$$

$$+ (12 - 11.3)^2 \cdot 0.2 + (13 - 11.3)^2 \cdot 0.1 \quad (25)$$

$$+ (14 - 11.3)^2 \cdot 0.1 \quad (26)$$

$$= 1.81 \quad (27)$$

Example: expectation and variance

Table 3: Example: discrete random variable

X	10	11	12	13	14
$p_X(x)$	0.4	0.2	0.2	0.1	0.1

- Variance

$$E[X^2] - \{E(X)\}^2 = \sum_x x^2 p_X(x) - 11.3^2 \quad (28)$$

$$= 10^2 \cdot 0.4 + 11^2 \cdot 0.2 + 12^2 \cdot 0.2 + 13^2 \cdot 0.1 \quad (29)$$

$$+ 14^2 \cdot 0.1 - 11.3^2 \quad (30)$$

$$= 40 + 24.2 + 28.8 + 16.9 + 19.6 - 127.69 \quad (31)$$

$$= 1.81 \quad (32)$$

Uniform distribution (discrete)

$$\mathcal{U}(x) = \frac{1}{n} \quad (33)$$

Gaussian distribution (continuous)

$$p(x) = p(x; \mu, \sigma^2) = \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\} \quad (34)$$

Bayes' theorem

Bayes' theorem

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)} \quad (35)$$

where

- $\Pr(A)$: prior probability of A
- $\Pr(A \mid B)$: posterior probability of A given B
- $\Pr(B \mid A)$: likelihood of A given B