Lecture 02: Basic Mathematics

[SCS4049-02] Machine Learning and Data Science

Seongsik Park (s.park@dgu.edu)

AI Department, Dongguk University

Basics: linear algebra

Vector

- Notation for a column vector

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

- Transpose and a row vector

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

- Magnitude of a vector (= 2-norm)

$$\|\mathbf{x}\| = \|\mathbf{x}\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \tag{3}$$

$$\|\mathbf{x}\| = \|\mathbf{x}\|_2 = 0 \iff \mathbf{x} = 0 \tag{4}$$

Vector operation

- Addition

$$\mathbf{x} + \mathbf{y} = \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \cdots \\ x_n + y_n \end{bmatrix} \tag{5}$$

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

- Scalar product

$$\lambda\mathbf{x} = \lambda \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix} = \begin{bmatrix} \lambda x_1 \\ \lambda x_2 \\ \cdots \\ \lambda x_n \end{bmatrix} \tag{6}$$

Vector operation

- Inner product

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

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

- 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: $\forall \mathbf{x} \neq 0$, $\mathbf{x} \cdot \mathbf{x} > 0$ and $\mathbf{x} \iff \mathbf{x} \cdot \mathbf{x} = 0$

Vector Norm

- A norm on a vector space Ω is a function $f : \Omega \to \mathcal{R}$ 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: $\|\mathbf{x}\|_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{\frac{1}{2}}$
  - p-norm: $\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}$
  - Infinity norm: $\|\mathbf{x}\|_\infty = \max_i |x_i|$

Vector Norm

- Inner product revisited

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

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

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

Matrix

- Notation for a 2D array of scalars

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

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

$$\mathbf{I}_m \mathbf{A} = \mathbf{A} \mathbf{I}_n = \mathbf{A} \tag{13}$$

Matrix operation

- Addition

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

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

  - Communicative: $\mathbf{A}+\mathbf{B}=\mathbf{B}+\mathbf{A}$
  - Associative: $(\mathbf{A}+\mathbf{B})+\mathbf{C}=\mathbf{A}+(\mathbf{B}+\mathbf{C})$

Matrix operation

- Multiplication

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

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 communicative: $\mathbf{AB} \neq \mathbf{BA}$
- Transpose: $(\mathbf{A}^T)_{ij} = (\mathbf{A})_{ji}$
  - $(\mathbf{A}^T)^T = \mathbf{A}$
  - $(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T$

Matrix inversion

- Inverse matrix
  - Condition to have inverse matrix: square and non-singular
  - Inverse matrix of $\mathbf{A}$ is $\mathbf{A}^{-1}$

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n \text{ where non-singular } \mathbf{A} \in \mathcal{R}^{n \times n} \tag{17}$$

- Property
  - $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
  - $(\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T$
  - For an orthonormal matrix, $\mathbf{A}^{-1} = \mathbf{A}^T$
  - For a diagonal matrix, $\mathbf{D} = \text{diag}(d_1, d_2, ..., d_n)$

$$\mathbf{D}^{-1} = \text{diag}(d_1^{-1}, d_2^{-1}, ..., d_n^{-1}) \tag{18}$$

Basics: differentiation

Differentiation

- **Differentiation** is all about measuring change

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{(x + \Delta x) - (x)} \tag{19}$$

$$f(x) = x^n \qquad \to f'(x) = nx^{n-1} \tag{20}$$

$$f(x) = \lambda g(x) \qquad \to f'(x) = \lambda g'(x) \tag{21}$$

$$f(x) = g(x) + h(x) \qquad \to f'(x) = g'(x) + h'(x) \tag{22}$$

$$f(x) = g(x) \cdot h(x) \qquad \to f'(x) = g'(x) \cdot h(x) + g(x) \cdot h'(x) \tag{23}$$

$$f(x) = g(h(x)) \qquad \to f'(x) = g'(h(x)) \cdot h'(x) \tag{24}$$

Differentiation: example

$$f(x) = x^3 \qquad \rightarrow f'(x) = 3x^2 \tag{25}$$

$$f(x) = 2x^3 \qquad \rightarrow f'(x) = 2 \cdot 3x^2 = 6x^2 \tag{26}$$

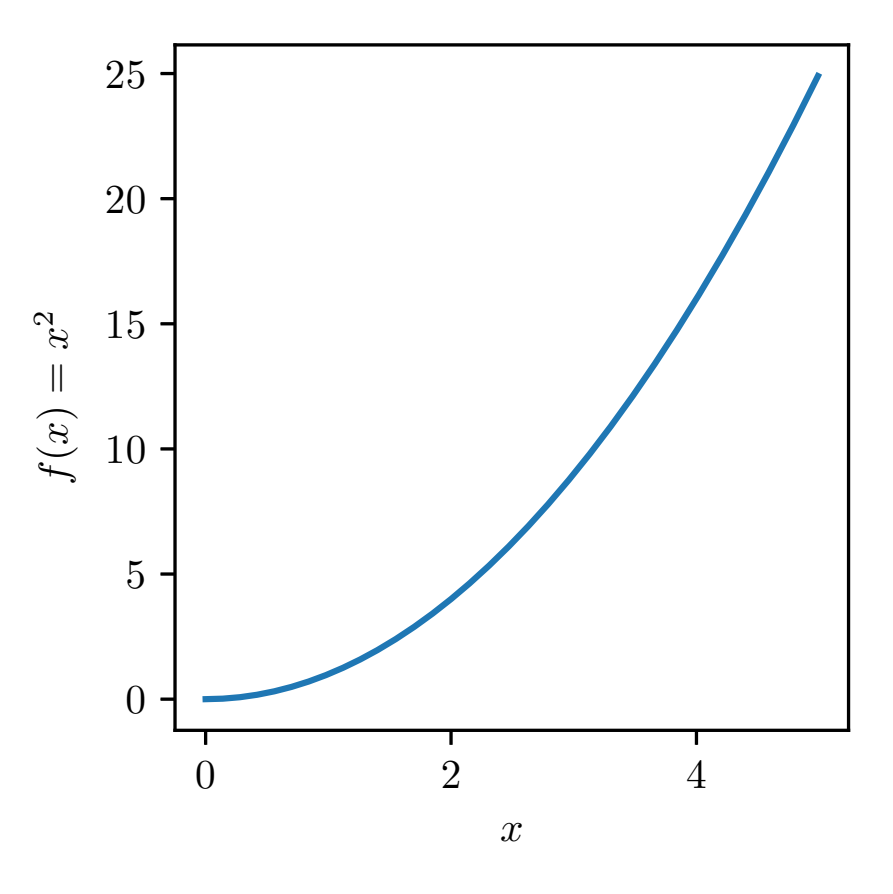
$$f(x) = 2x^3 + 5x + 10 \qquad \rightarrow f'(x) = 6x^2 + 5 \tag{27}$$

$$f(x) = (2x^2 + 3x) \cdot (5x + 10) \qquad \rightarrow f'(x) = (4x + 3) \cdot (5x + 10) \tag{28}$$

$$+ (2x^2 + 3x) \cdot 5 \tag{29}$$

$$f(x) = (3x^2 + 2x)^3 \qquad \rightarrow f'(x) = 3 \cdot (3x^2 + 2x)^2 \cdot (6x + 2) \tag{30}$$

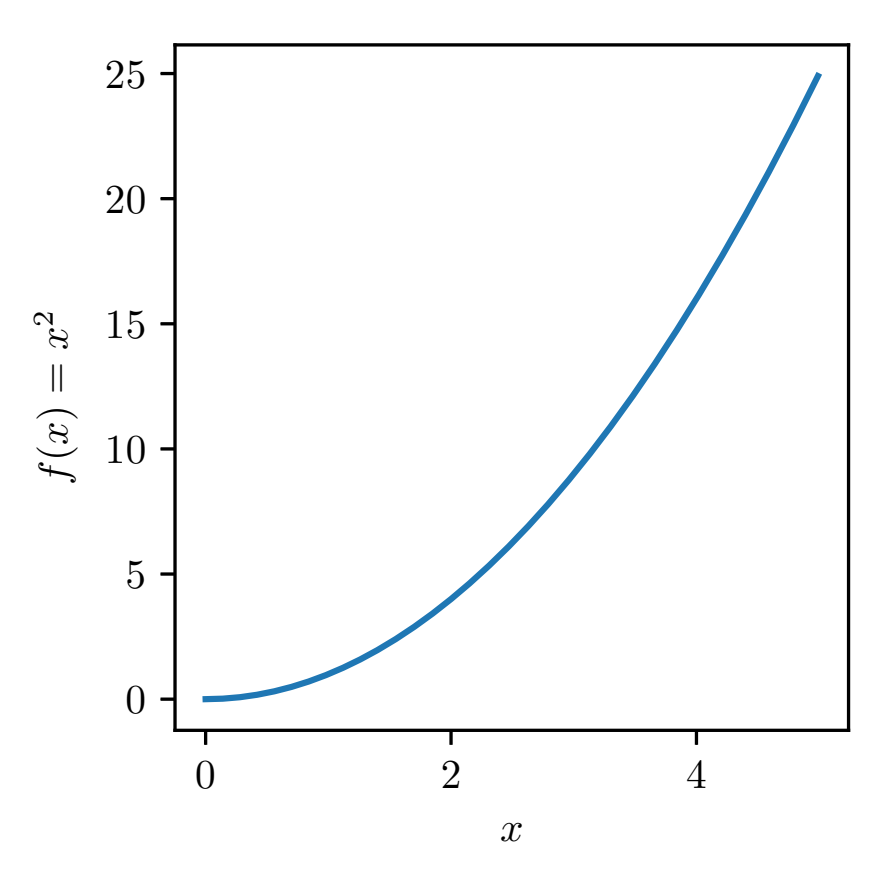
Differentiation for $f(x) = x^2$



Slope between $x_1 = 2$ and $x_2 = 4$, or $x + \Delta x = 4$ and $x = 2$

$$\text{slope} = \frac{f(x_2) - f(x_1)}{x_2 - x_1} = \frac{f(x + \Delta x) - f(x)}{(x + \Delta x) - (x)} = \frac{4^2 - 2^2}{4 - 2} = 6 \qquad (31)$$

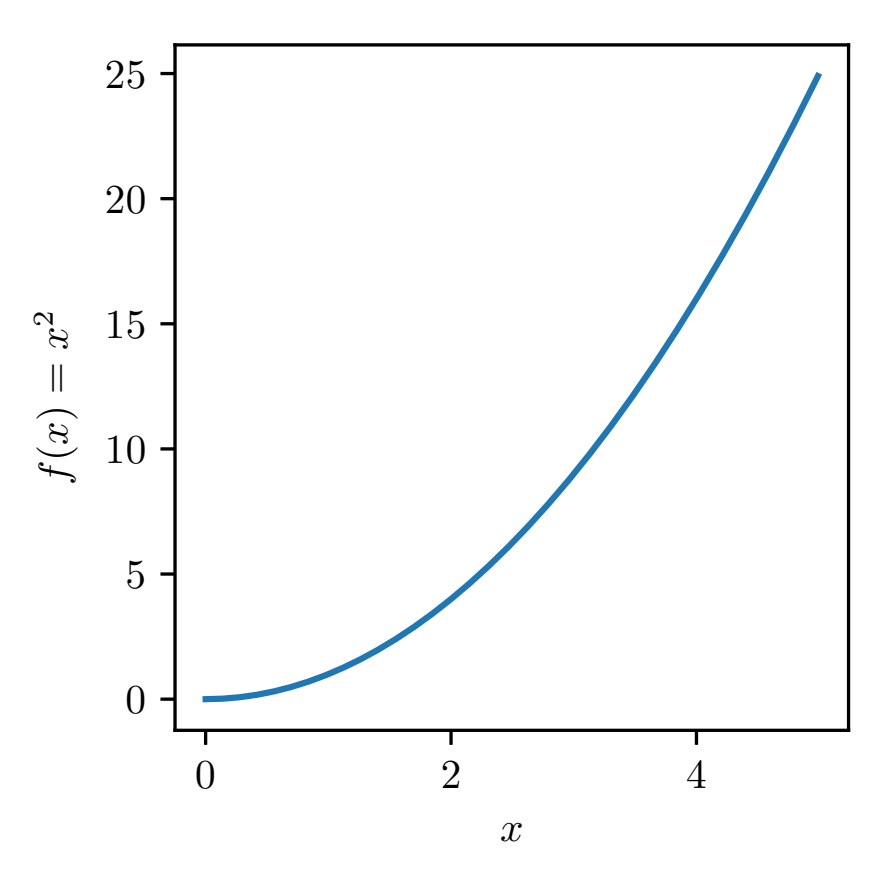
Differentiation for $f(x) = x^2$



Slope between $x + \Delta x = 2.1$ and $x = 2$

$$\text{slope} = \frac{f(x + \Delta x) - f(x)}{(x + \Delta x) - (x)} = \frac{2.1^2 - 2^2}{2.1 - 2} = 4.1 \tag{32}$$

Differentiation for $f(x) = x^2$



Slope at $x = 2$

$$\text{slope} = \lim_{x \to \Delta x} \frac{f(2 + \Delta x) - f(2)}{(2 + \Delta x) - (2)} = f'(2) = 4 \tag{33}$$

Basics: probability

Axioms of probability

Axioms for events

1. Ω is an event.
2. For every sequence of events $A_1, A_2, \ldots$, the union $\bigcup_{n=1}^{\infty} A_n$ is an event.
3. For every event A, the complement A^c is an event.

Axioms for probability

1. $\Pr\{\Omega\} = 1$.
2. For every event A, $\Pr\{A\} \geq 0$.
3. The probability of the union of any sequence $A_1, A_2, \ldots$ of disjoint events is given by

$$\Pr\left\{\bigcup_{n=1}^{\infty} A_n\right\} = \sum_{n=1}^{\infty} \Pr\{A_n\}$$

Axioms of probability theory (cont'd)

Corollaries

$$\Pr\{\emptyset\} = 0$$

$$\Pr\left\{\bigcup_{n=1}^{m} A_n\right\} = \sum_{n=1}^{m} \Pr\{A_n\} \qquad \text{for } A_1, ..., A_m \text{ disjoint}$$

$$\Pr\{A^c\} = 1 - \Pr\{A\} \qquad \text{for all } A$$

$$\Pr\{A\} \leq \Pr\{B\} \qquad \text{for all } A \subseteq B$$

$$\sum_{n} \Pr\{A_n\} \leq 1 \qquad \text{for all } A_1, ... \text{ disjoint}$$

Conditional probability

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

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

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 distribution function** $f_X(x) = \frac{d}{dx} F_X(x)$

Conditional probability

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

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

Discrete rv's

$$p_{XY}(x_i, y_j) = p_X(x_i)p_Y(y_j)$$
$$p_{X|Y}(x_i \mid y_j) = p_X(x_i)$$

Continuous rv's

$$f_{XY}(x, y) = f_X(x)f_Y(y)$$
$$f_{X|Y}(x \mid y) = f_X(x)$$

Expectation

The expected value $\mathrm{E}[X]$ 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

$$\mathrm{E}[X] = \sum_x x p_X(x).$$

- Discrete case: $\mathrm{E}(X) = \sum_x x p_X(x)$
- Continuous case: $\mathrm{E}(X) = \int_x x f_X(x) dx$

Variance and standard deviation

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

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

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

$$\mathrm{E}(X) = \sum_x x p_X(x) \tag{34}$$

$$= 10 \cdot 0.4 + 11 \cdot 0.2 + 12 \cdot 0.2 + 13 \cdot 0.1 + 14 \cdot 0.1 \tag{35}$$

$$= 11.3 \tag{36}$$

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

$$\begin{aligned}
\mathrm{E}\left[(X-\mathrm{E}(X))^2\right] &= \sum_x (x-11.3)^2 p_X(x) && (37)\\
&= (10-11.3)^2 \cdot 0.4 + (11-11.3)^2 \cdot 0.2 && (38)\\
&\quad + (12-11.3)^2 \cdot 0.2 + (13-11.3)^2 \cdot 0.1 && (39)\\
&\quad + (14-11.3)^2 \cdot 0.1 && (40)\\
&= 1.81 && (41)
\end{aligned}$$

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

$$\begin{aligned}
\mathrm{E}\left[X^2\right] - \{\mathrm{E}(X)\}^2 &= \sum_x x^2 p_X(x) - 11.3^2 && (42)\\
&= 10^2 \cdot 0.4 + 11^2 \cdot 0.2 + 12^2 \cdot 0.2 + 13^2 \cdot 0.1 && (43)\\
&\quad + 14^2 \cdot 0.1 - 11.3^2 && (44)\\
&= 40 + 24.2 + 28.8 + 16.9 + 19.6 - 127.69 && (45)\\
&= 1.81 && (46)
\end{aligned}$$

Appendix

HW #1

Due: 9월 13일 일요일, 23시 59분까지

- pdf로 업로드하세요.
- 손으로 작성한 파일을 스캔앱(Adobe scan, Office lens 등)을 써서 pdf로 저장해주세요.
- 컴퓨터로 작성(latex, word, ppt, 한글 등)한 파일도 가능합니다. pdf로 저장해주세요.
- 가독성이 떨어지는 파일도 불량처리 합니다.

1. 확률변수 X를 주사위 두개를 던져 나온 두 수의 합이라고 할 때 아래의 문제에 대해 답하세요. (25점)

   1.1 아래의 표를 완성하세요. (5점)

X	2	3	4	5	6	7	8	9	10	11	12
$p_X(x)$											

   1.2 기대값 $\mathrm{E}[X]$를 구하세요. (5점)

   1.3 분산 σ_X^2를 구하세요. (5점)

   1.4 확률분포 함수 $F_X(x)$를 그리세요. (5점)

   1.5 확률 질량 함수 (probability mass function) $p_X(x)$를 그리세요. (5점)