

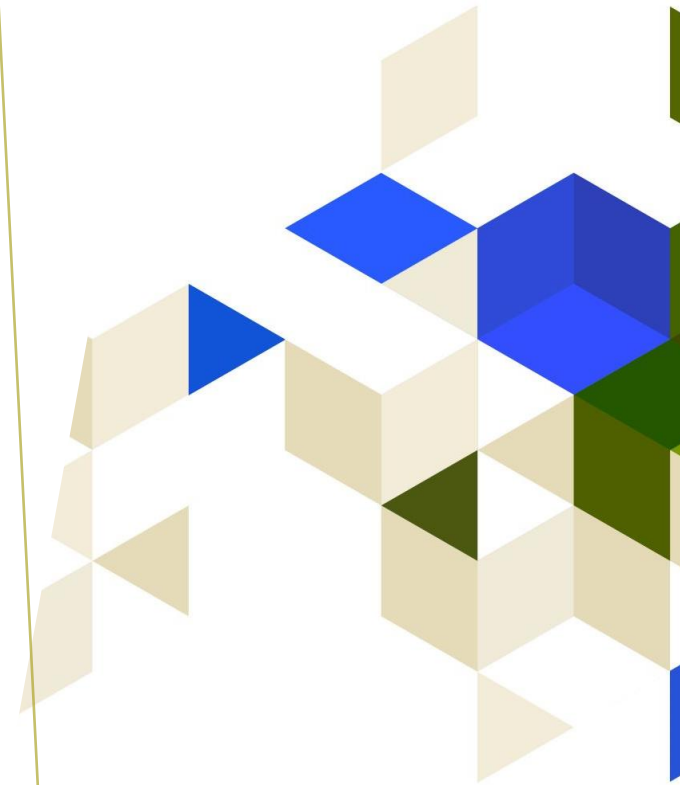
공지사항

- 교수학습지원센터 수업 컨설팅 9/20 (화)
 - 뒷모습 촬영
 - 간단한 설문조사

ARTIFICIAL INTELLIGENCE

WOOHWAN JUNG

Linear Algebra & Probability Theory



Today's Lecture

- Topics
 - Linear algebra
 - Probability theory
- Not a comprehensive study on linear algebra and probability theory
- Focused on the subset that is most relevant to deep learning

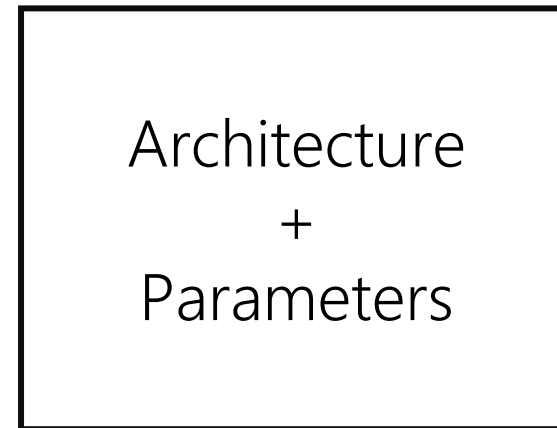
Learning Process

A lot of vector/matrix operations

Input



Model



Output

$$P(\text{Cat}) = 0.9$$

Label

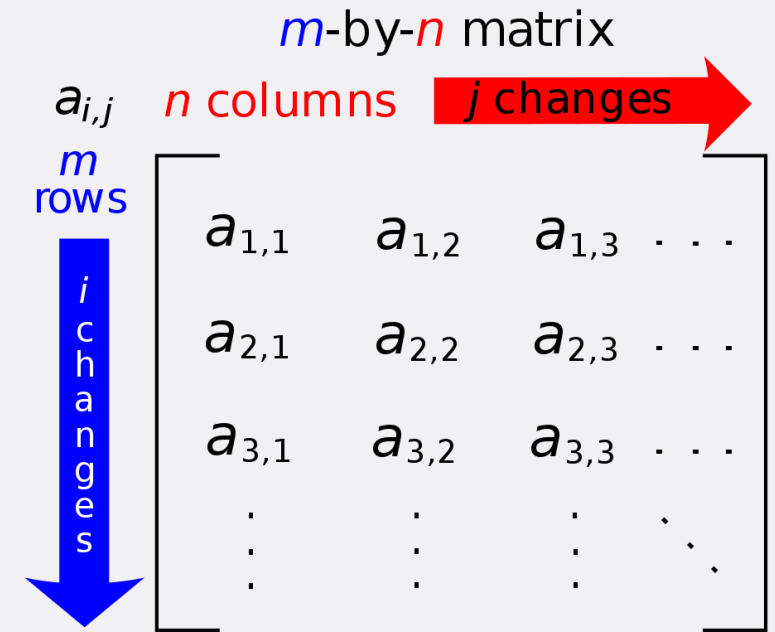
$$P(\text{Cat}) = 1$$

Update parameters
to minimize the loss

Compute loss



LINEAR ALGEBRA



Scalars and Vectors

- Scalars
 - A scalar is a single number
 - Usually denoted with italic font: a, n, x
 - Integers, real numbers, rational numbers, etc.
 - Example notation: $x \in \mathbb{R}, x \in \mathbb{Z}, x \in \mathbb{N}$
- Vectors
 - A vector is a 1-D array of numbers
 - Can be real, integer, etc.
 - Example notation for type and size:
 - $\mathbf{x} \in \mathbb{R}^n, \mathbf{x} \in \mathbb{Z}^n$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Matrices

- A matrix is a 2-D array of numbers:

$$\mathbf{A} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}$$

- Example notation for type and shape:

$$\mathbf{A} \in \mathbb{R}^{m \times n}$$

Tensors

- A tensor is an (multi-dimensional) array of numbers, that may have
 - Zero dimensions, and be a scalar
 - One dimension, and be a vector
 - Two dimensions, and be a matrix
 - And more dimensions

Matrix Transpose



$$\mathbf{A} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \\ A_{3,1} & A_{3,2} \end{bmatrix} \Rightarrow \mathbf{A}^T = \begin{bmatrix} A_{1,1} & A_{2,1} & A_{3,1} \\ A_{1,2} & A_{2,2} & A_{3,2} \end{bmatrix}$$

Figure 2.1: The transpose of the matrix can be thought of as a mirror image across the main diagonal.

Matrix Addition and Subtraction

- Adding or subtracting corresponding elements

- Let $\mathbf{A} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} B_{1,1} & B_{1,2} \\ B_{2,1} & B_{2,2} \end{bmatrix}$

- $\mathbf{A} + \mathbf{B} = \begin{bmatrix} A_{1,1} + B_{1,1} & A_{1,2} + B_{1,2} \\ A_{2,1} + B_{2,1} & A_{2,2} + B_{2,2} \end{bmatrix}$

- $\mathbf{A} - \mathbf{B} = \begin{bmatrix} A_{1,1} - B_{1,1} & A_{1,2} - B_{1,2} \\ A_{2,1} - B_{2,1} & A_{2,2} - B_{2,2} \end{bmatrix}$

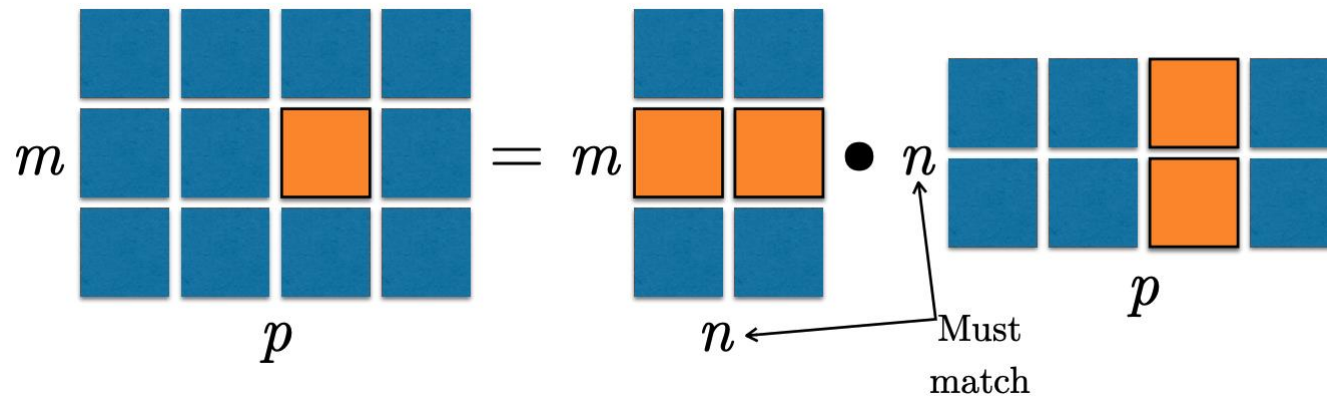
Matrix Multiplication

- For $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$

$$C = AB.$$

$$C_{i,j} = \sum_k A_{i,k} B_{k,j}.$$

$$C \in \mathbb{R}^{m \times p}$$



Shape of the Result of Vector/Matrix Multiplication

- For $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$, and $C = AB$
 - $C \in \mathbb{R}^{m \times p}$
- For $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$, and $y = Ax$
 - $y \in \mathbb{R}^m$

Norms

- Functions that measure how “Large” a vector is
 - Lp norm

$$\|\mathbf{x}\|_p = \left(\sum_i |x_i|^p \right)^{\frac{1}{p}}$$

- Most popular norm: L2 norm ($p=2$)
 - L1 norm ($p=1$): $\|\mathbf{x}\|_1 = \sum_i |x_i|$
 - Max norm ($p = \infty$): $\|\mathbf{x}\|_\infty = \max_i |x_i|$
- Frobenius norm of a matrix
 - $\|\mathbf{A}\|_F = \sqrt{\sum_i \sum_j A_{ij}^2}$

Distance Between a Pair of Vectors

- Norm of $\mathbf{x} - \mathbf{y}$
- Lp distance

$$\|\mathbf{x} - \mathbf{y}\|_p = \left(\sum_i |x_i - y_i|^p \right)^{\frac{1}{p}}$$

PROBABILITY THEORY

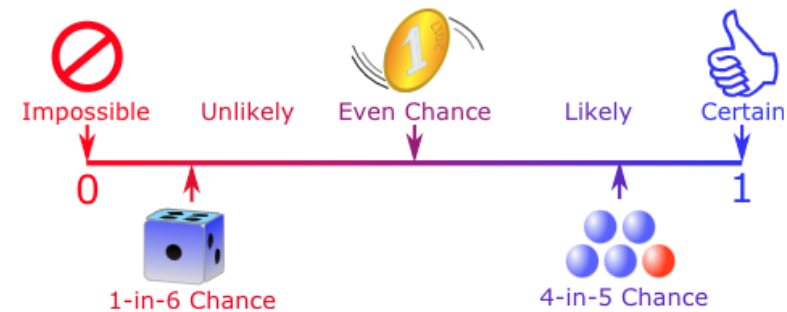


Random Variable

- A **variable** whose values depend on outcomes of a **random** phenomenon
- Discrete random variables
 - Bernoulli r.v.
 - Categorical r.v.
- Continuous random variables
 - Gaussian (Normal) r.v.
 - Laplace r.v.

Probability Mass Function (PMF): $P(x)$

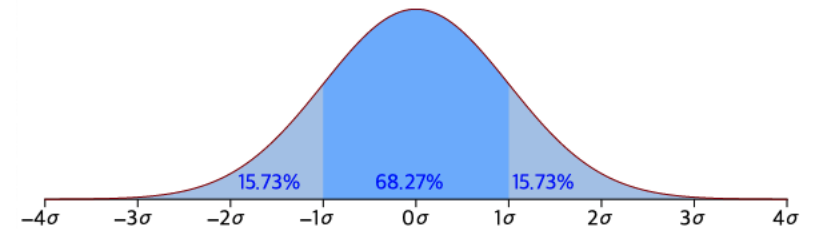
- A function that gives the probability that a **discrete random variable** is equal to some value
- For all x , $0 \leq P(X = x) \leq 1$
- $\sum P(X = x) = 1$
- Example: discrete uniform distribution
 - $P(X = x) = \frac{1}{k}$ where k is the number of possible values



Probability is always between 0 and 1

Probability Density Function (PDF): $p(x)$

- The PDF of a **continuous random variable** gives the *relative* likelihood of any outcome x
- Properties
 - $p(x) \geq 0$, (Note: we do not require $p(x) \leq 1$)
 - $\int p(x)dx = 1$
- Example: continuous uniform distribution $u(a,b)$



$$p(x) = \frac{1}{b - a}$$

Marginal Probability Distribution

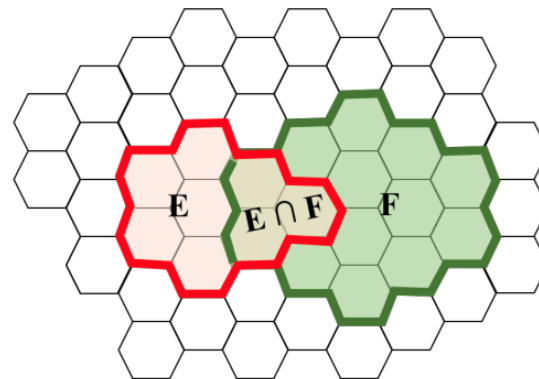
- A probability distribution over the subset of variables
- Sum rule

$$\forall x \in \mathbf{x}, P(\mathbf{x} = x) = \sum_y P(\mathbf{x} = x, y = y).$$

$$p(x) = \int p(x, y) dy.$$

Conditional Probability

$$P(y = y \mid x = x) = \frac{P(y = y, x = x)}{P(x = x)}.$$



$$P(E) = \frac{8}{50} \approx 0.16$$

$$P(E|F) = \frac{3}{14} \approx 0.21$$

Chain Rule



$$P(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}) = P(\mathbf{x}^{(1)}) \prod_{i=2}^n P(\mathbf{x}^{(i)} \mid \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(i-1)}).$$

- Example (n=4)

$$\begin{aligned} & P(X^{(1)}, X^{(2)}, X^{(3)}, X^{(4)}) \\ &= P(X^{(1)})P(X^{(2)} \mid X^{(1)})P(X^{(3)} \mid X^{(1)}X^{(2)}) \end{aligned}$$

Independence

- Two random variables x and y are **independent** if and only if

$$\forall x \in \mathbf{x}, y \in \mathbf{y}, p(\mathbf{x} = x, \mathbf{y} = y) = p(\mathbf{x} = x)p(\mathbf{y} = y).$$

- If x and y are **conditionally independent**

$$\forall x \in \mathbf{x}, y \in \mathbf{y}, z \in \mathbf{z}, p(\mathbf{x} = x, \mathbf{y} = y \mid \mathbf{z} = z) = p(\mathbf{x} = x \mid \mathbf{z} = z)p(\mathbf{y} = y \mid \mathbf{z} = z).$$

Expectation

$$\mathbb{E}_{\mathbf{x} \sim P}[f(x)] = \sum_x P(x) f(x),$$

$$\mathbb{E}_{\mathbf{x} \sim p}[f(x)] = \int p(x) f(x) dx.$$

$$\mathbb{E}_{\mathbf{x}}[\alpha f(x) + \beta g(x)] = \alpha \mathbb{E}_{\mathbf{x}}[f(x)] + \beta \mathbb{E}_{\mathbf{x}}[g(x)],$$

Variance & Covariance

- Variance

$$\text{Var}(f(x)) = \mathbb{E} \left[(f(x) - \mathbb{E}[f(x)])^2 \right] .$$

$$= E[f(x)^2] - E[f(x)]^2$$

- Standard deviation: square root of the variance
- Covariance

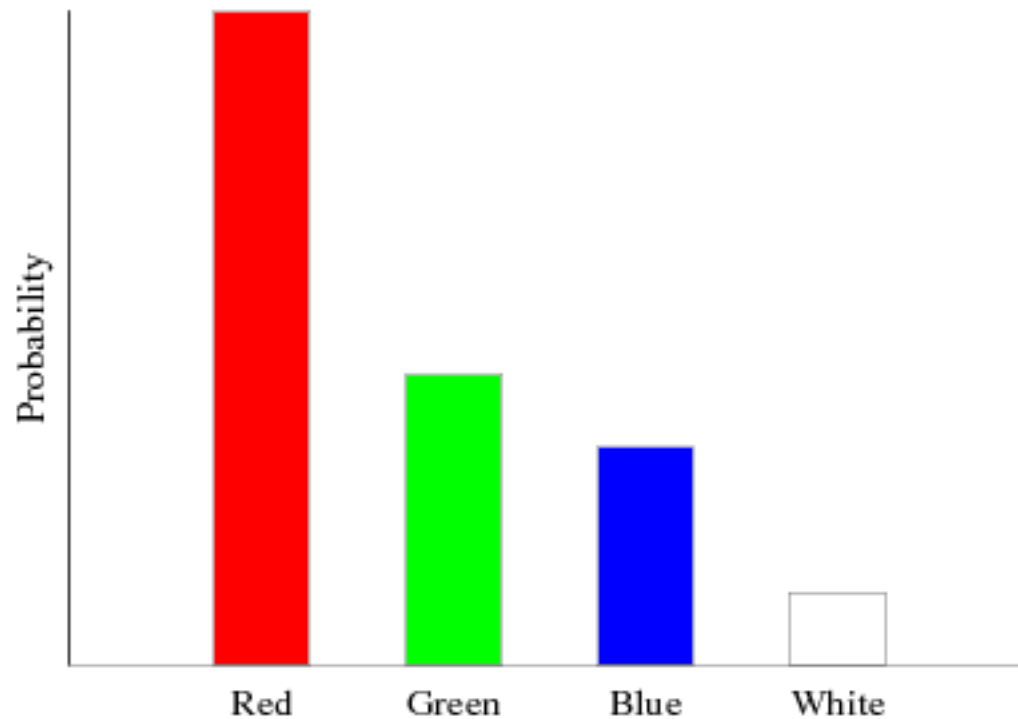
$$\text{Cov}(f(x), g(y)) = \mathbb{E} [(f(x) - \mathbb{E}[f(x)]) (g(y) - \mathbb{E}[g(y)])] .$$

Bernoulli Distribution

- PMF: $P(X = x) = \begin{cases} \phi & \text{if } x = 1 \\ 1 - \phi & \text{if } x = 0 \end{cases}$
- Expectation: $E[X] = \phi$
- Variance: $Var[X] = \phi(1 - \phi)$

Categorical Distribution

- A.k.a. multinoulli distribution



$$P(X = k) = p_k$$

Gaussian Distribution

- A.k.a Normal distribution
- Parameterized by variance σ^2 :

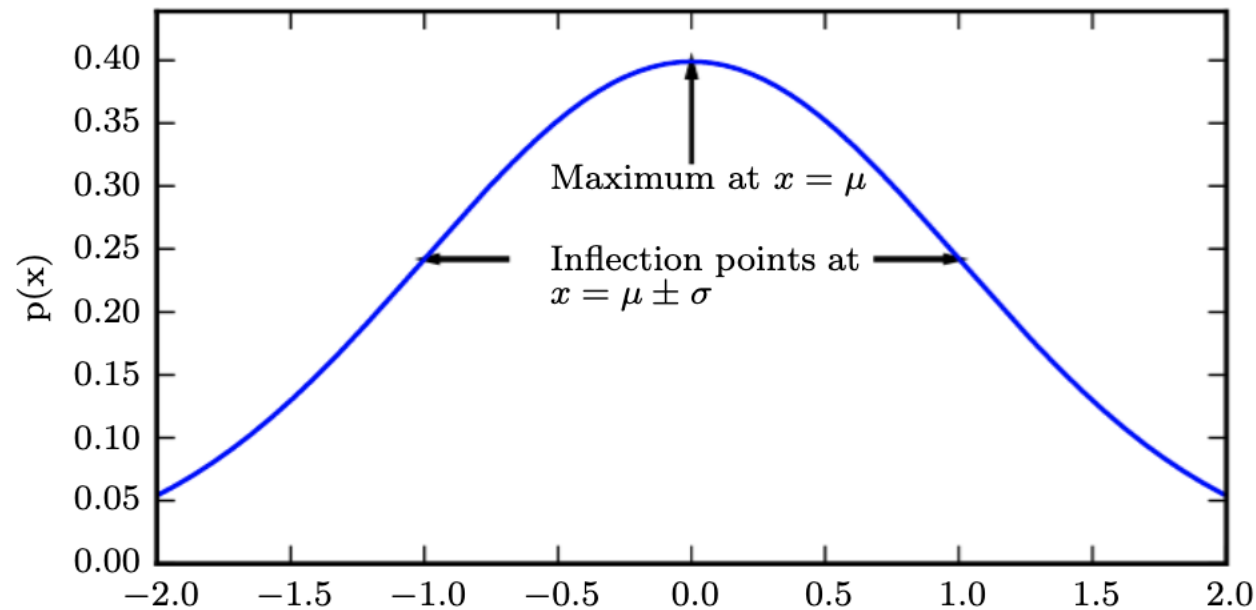
$$\mathcal{N}(x; \mu, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right).$$

- Parameterized by precision $\beta = \frac{1}{\sigma^2}$

$$\mathcal{N}(x; \mu, \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{1}{2}\beta(x - \mu)^2\right).$$

Standard Normal Distribution

- Normal (gaussian) distribution $\mathcal{N}(x; \mu, \sigma^2)$ with $\mu = 0$ and $\sigma = 1$

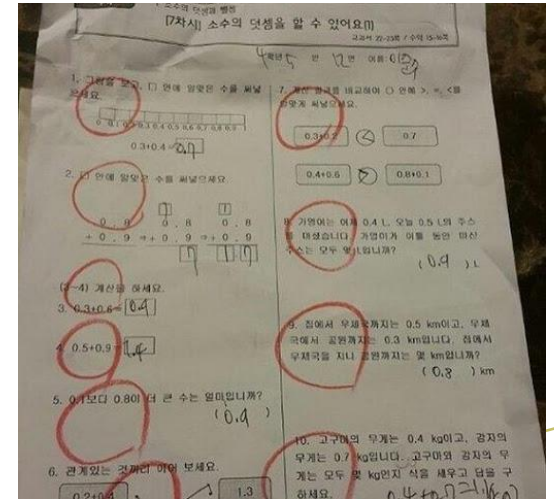


Gaussian is Indeed Normal!

- Gaussian distributions are sensible choice for many application
 - Especially in the absence of prior knowledge

Central limit theorem

“The sum of many independent random variables is approximately **normally** distributed”



Bayes' Rule

$$\text{Posterior } P(x | y) = \frac{\text{Prior } P(x) \text{ Likelihood } P(y | x)}{P(y)}.$$

Q: 국가 A는 선진국일까? 아닐까?

국가 A

기대수명 (남)	기대수명 (여)
80.5	86.5

OECD

기대수명 (남)	기대수명 (여)
77.9	83.3

전세계

기대수명 (남)	기대수명 (여)
74.8	79.4