

데이터 분석을 위한 기초 수학 및 통계

CONTENTS

1. 데이터 분석 개요 02

- 1) AI-ML-DL의 개념
- 2) ML 알고리즘과 학습 매커니즘
- 3) 데이터 분석/과학을 위한 수학 & 통계
- 4) 데이터 분석 vs. 데이터 과학

2. 데이터 분석을 위한 수학 리뷰 07

- 1) 선형대수학
- 2) 미분과 도함수
- 3) 유사도 측정법

3. 데이터 분석을 위한 통계학 리뷰 19

- 1) 기술 통계와 추론 통계
- 2) 척도
- 3) 모수와 통계량
- 4) 확률
- 5) 중심극한정리
- 6) 표집 분포
- 7) 상관 계수
- 8) 회귀 분석 & 로지스틱 회귀 분석
- 9) 가설 검정

4. Machine Learning 34

- 1) 머신 러닝의 작업 흐름
- 2) 전처리
- 3) 지도 학습과 비지도 학습
- 4) Regression 계열 알고리즘
- 5) Tree 계열 알고리즘
- 6) 모델 평가

5. Deep Learning 47

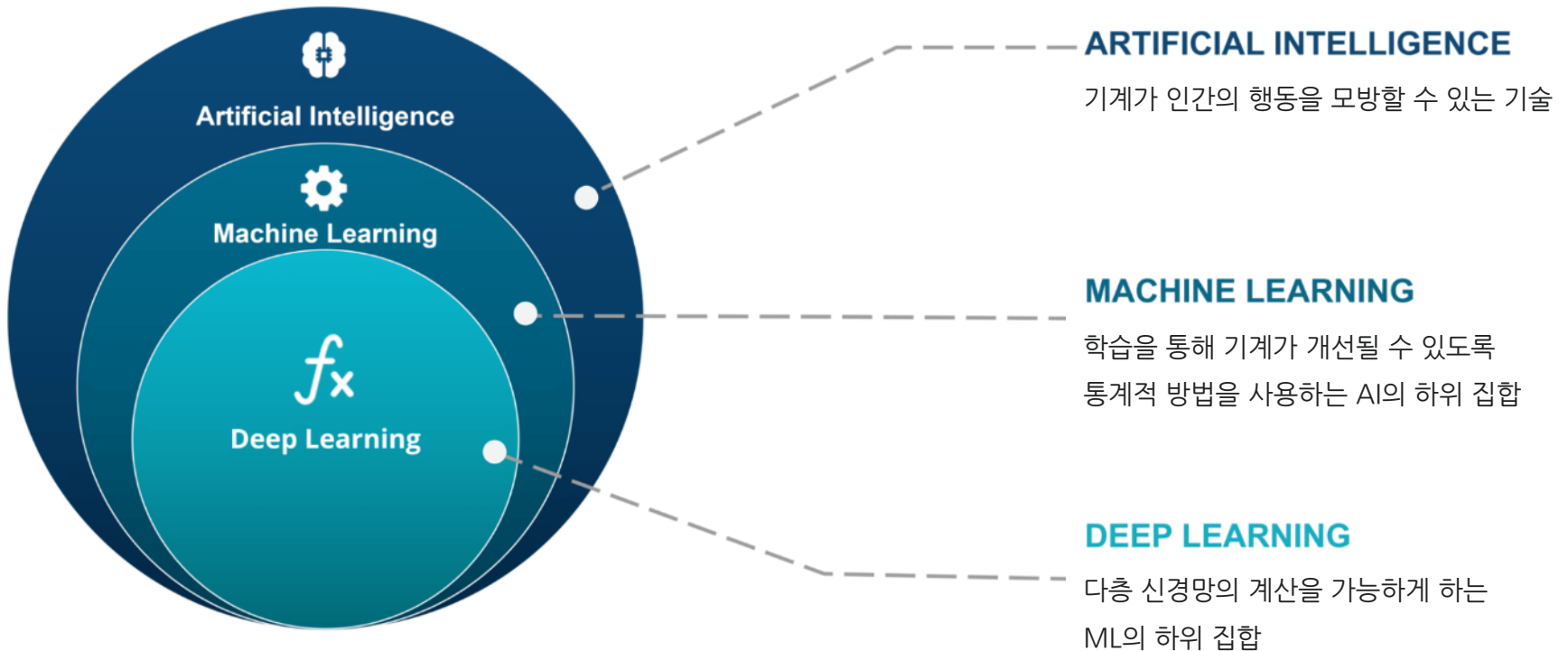
- 1) 인공 신경망
- 2) XOR 문제
- 3) 손실 함수
- 4) 활성화 함수
- 5) 소프트맥스 함수
- 6) 경사 하강법
- 7) 오차 역전파

1. 데이터 분석 개요

1. 데이터 분석 개요

1-1. AI-ML-DL의 개념

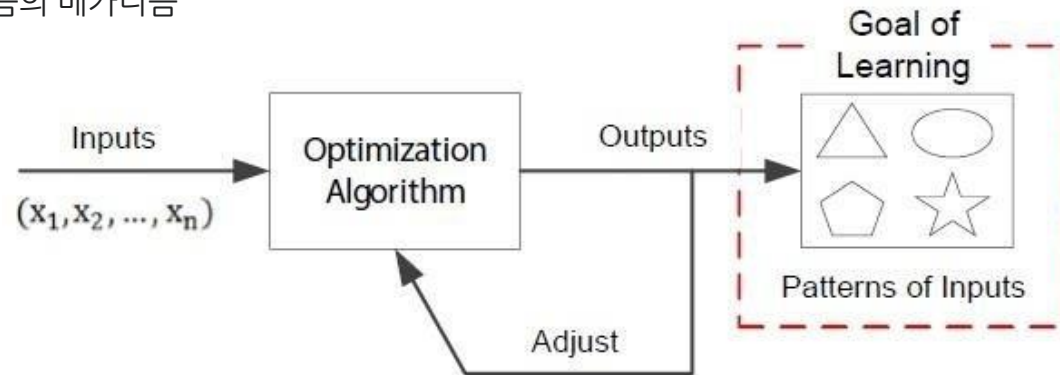
AI > Machine Learning > Deep Learning



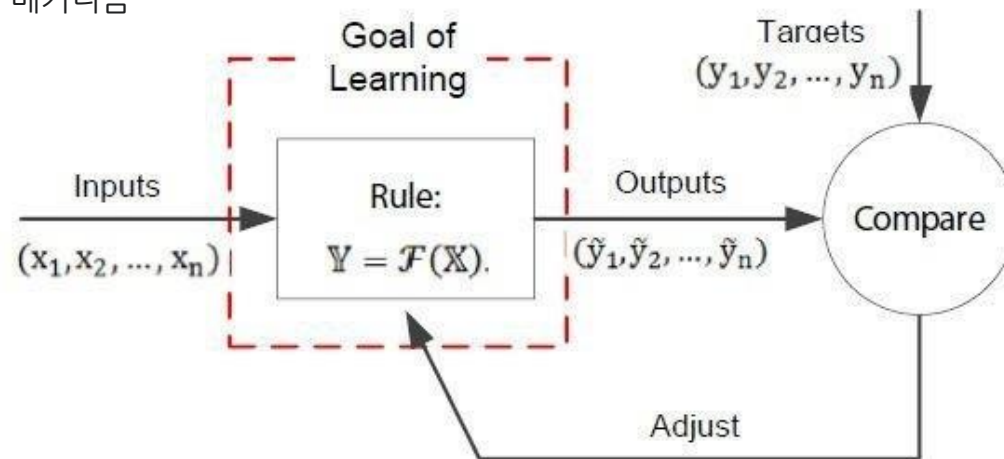
1. 데이터 분석 개요

1-2. ML 알고리즘과 학습 메카니즘

ML 지도학습 알고리즘의 메카니즘



ML 비지도학습 알고리즘의 메커니즘



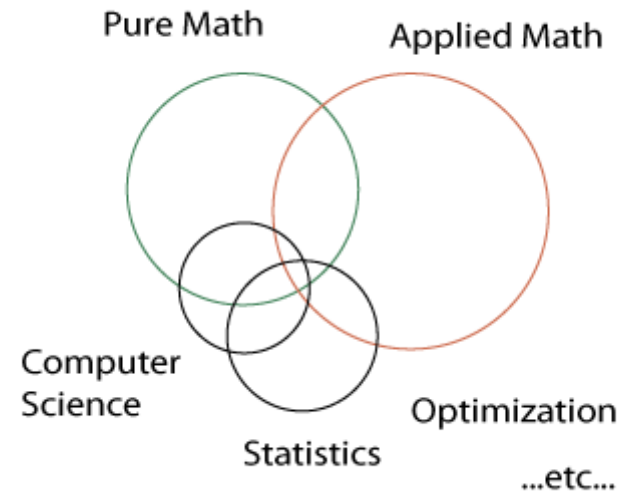
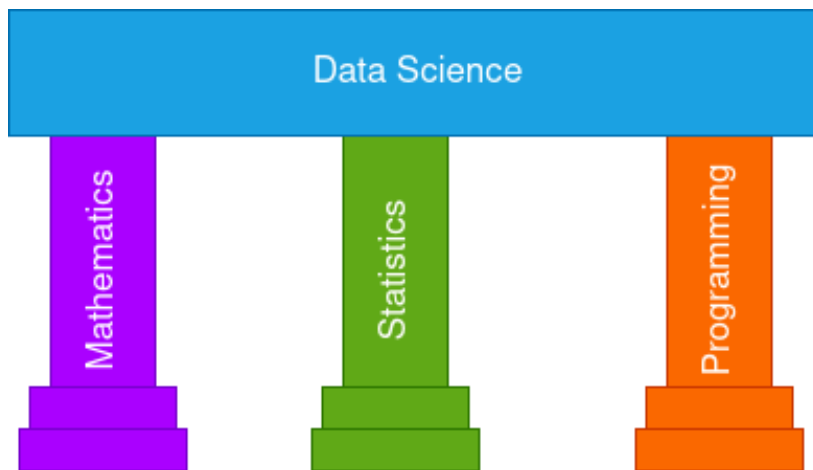
1. 데이터 분석 개요

1-3. 데이터 분석/과학을 위한 수학 & 통계

데이터 분석 및 과학은 수학과 통계, 프로그래밍 등 다양한 분야의 지식을 필요로 한다.

선형 대수학은 선형 방정식과 선형 방정식 그래프에 대한 연구로 통계 그래프를 이해하기 위한 기초이며

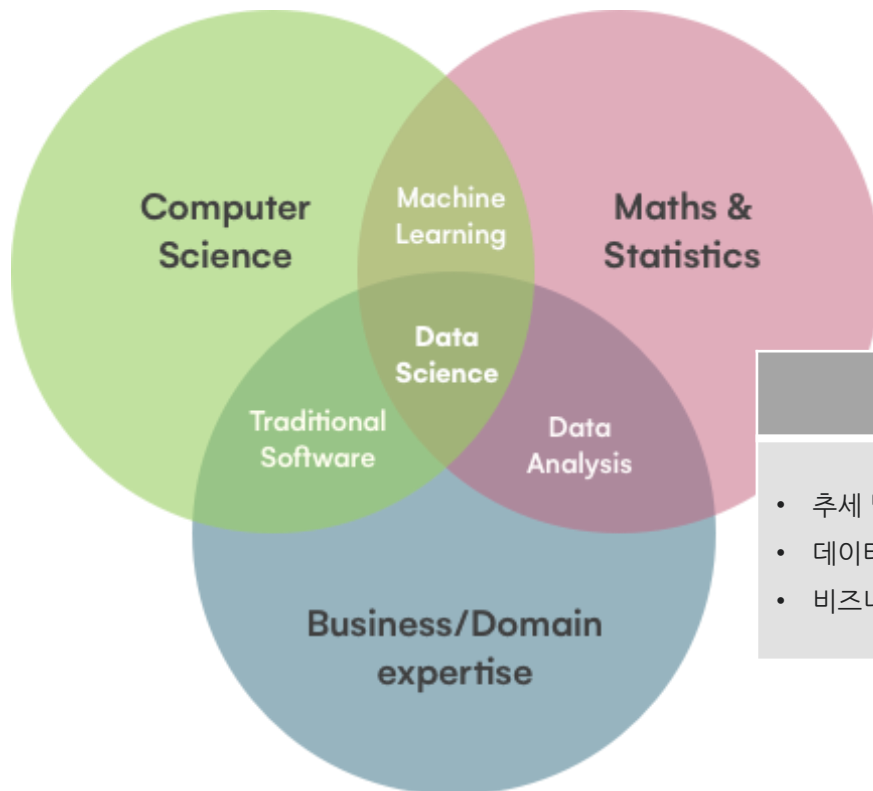
통계는 데이터를 이해하고 해석하고 제시하기 위한 기초 지식이다.



1. 데이터 분석 개요

1-4. 데이터 분석 vs. 데이터 과학

데이터 분석가는 데이터에서 가장 관련성이 높은 정보를 추출하는 반면,
데이터 과학자는 동일한 데이터에서 감지된 과거의 패턴을 기반으로 미래를 예측한다.



데이터 분석	데이터 과학
<ul style="list-style-type: none">추세 및 메트릭 탐색데이터 시각화비즈니스 지식 및 의사 결정 능력	<ul style="list-style-type: none">수학적 알고리즘 설계 및 구현기계 학습에 대한 지식비정형 데이터로 작업하는 경향

2. 데이터 분석을 위한 수학 리뷰

2. 데이터 분석을 위한 수학 리뷰

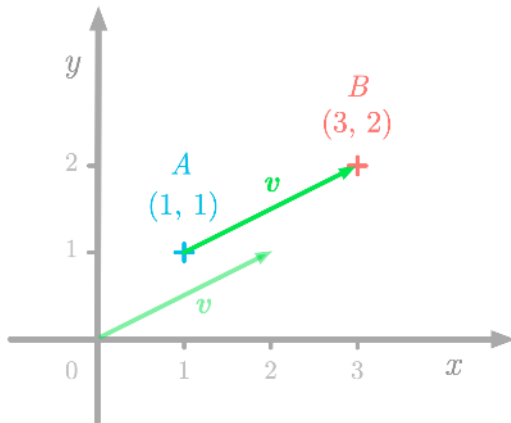
2-1. 선형대수학

벡터 (Vector)

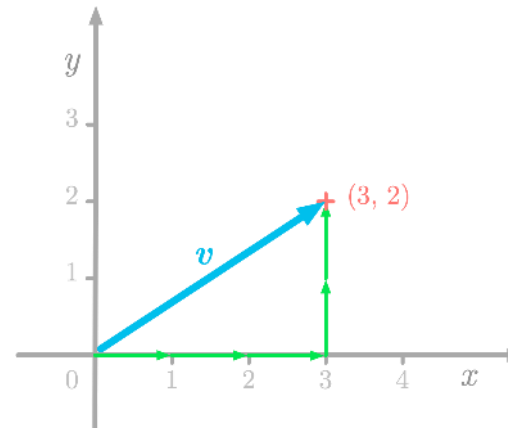
벡터는 분야에 따라 다양한 개념을 가지지만, 데이터 과학에서의 벡터는 데이터의 값을 저장하는 방법이다. 일반적으로 좌표계의 구성 요소로 설명된다.

- **기하 벡터** : 길이(크기)와 방향으로 정의되는 수학적 객체이다. 이러한 속성을 통해 좌표의 변위를 설명할 수 있다.
- **좌표 벡터** : 좌표에 해당하는 숫자의 배열이다. (좌표는 위치를 설명하는 값이다.)

기하 벡터
(Geometric Vectors)






좌표 벡터
(Coordinate Vectors)



2. 데이터 분석을 위한 수학 리뷰

2-1. 선형대수학

행렬 (Matrix)

Type	Scalar	Vector	Matrix	Tensor
Definition	a single number	an array of numbers	2-D array of numbers	k-D array of numbers
Notation	x	$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$	$\mathbf{X} = \begin{bmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,n} \\ X_{2,1} & X_{2,2} & \dots & X_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m,1} & X_{m,2} & \dots & X_{m,n} \end{bmatrix}$	\mathbf{X} $X_{i,j,k}$
Example	1.333	$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ \vdots \\ 9 \end{bmatrix}$	$\mathbf{X} = \begin{bmatrix} 1 & 2 & \dots & 4 \\ 5 & 6 & \dots & 8 \\ \vdots & \vdots & \ddots & \vdots \\ 13 & 14 & \dots & 16 \end{bmatrix}$	$\mathbf{x} = \begin{bmatrix} & & & [100 & 200 & 300] \\ & [10 & 20 & 30] & [00 & 600] \\ [1 & 2 & 3] & [50 & 60] & [00 & 900] \\ 4 & 5 & 6 & 80 & 90 \\ 7 & 8 & 9 & & \end{bmatrix}$
Python code example	<code>x = np.array(1.333)</code>	<code>x = np.array([1,2,3,4,5,6,7,8,9])</code>	<code>x = np.array([[1,2,3,4], [5,6,7,8], [9,10,11,12], [13,14,15,16]])</code>	<code>x = np.array([[[[1, 2, 3], [4, 5, 6], [7, 8, 9]], [[10, 20, 30], [40, 50, 60], [70, 80, 90]], [[100, 200, 300], [400, 500, 600], [700, 800, 900]]]])</code>
Visualization				 3-D Tensor

2. 데이터 분석을 위한 수학 리뷰

2-1. 선형대수학

내적 (Inner Dot)

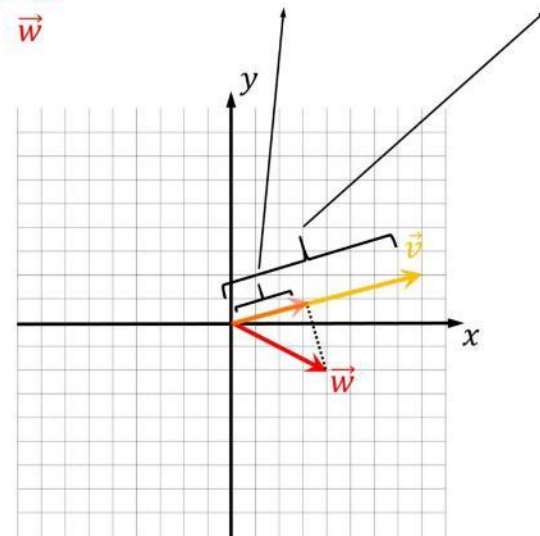
벡터의 내적은 \cdot (도트기호)로 표시되며 각 성분 쌍의 곱의 합으로 정의된다.

두 벡터 사이의 내적 결과는 스칼라이다.

$$\begin{bmatrix} 2 \\ 7 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 8 \\ 2 \\ 8 \end{bmatrix} = 2 \cdot 8 + 7 \cdot 2 + 1 \cdot 8 = 38$$

Dot product

$$\begin{bmatrix} 8 \\ 2 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ -2 \end{bmatrix} = (\text{Length of projected } \vec{w}) \cdot (\text{Length of } \vec{v})$$



2. 데이터 분석을 위한 수학 리뷰

2-1. 선형대수학

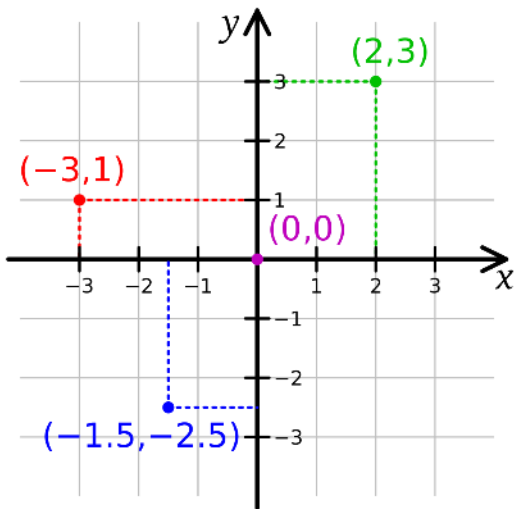
좌표계 (Coordinate System)

데이터를 표현하는 데 가장 널리 사용되는 좌표계는 데카르트(Cartesian) 좌표이다. 벡터의 공간에서 한 지점의 위치를 지정하려면 차원의 수 만큼 좌표가 필요하다.

* 참고 : 주성분 분석(Principal Component Analysis; PCA)

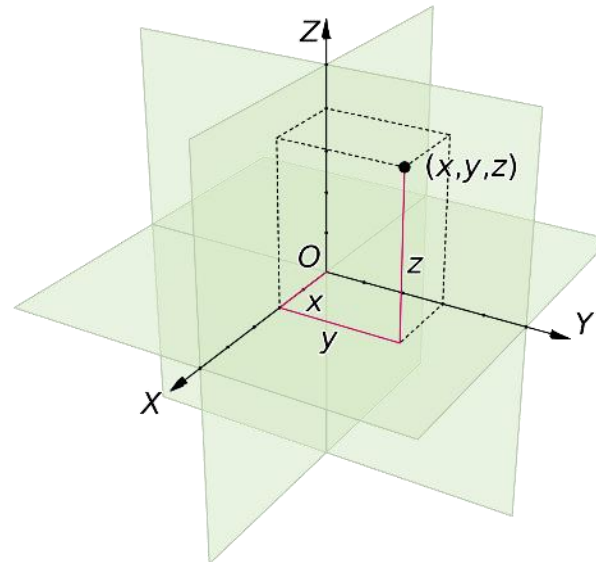
2차원 좌표계

(2-Dimensional Coordinate System)



3차원 좌표계

(3-Dimensional Coordinate System)



2. 데이터 분석을 위한 수학 리뷰

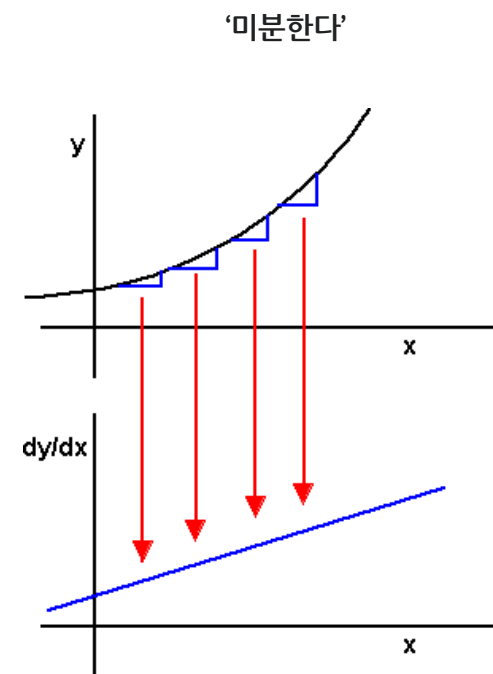
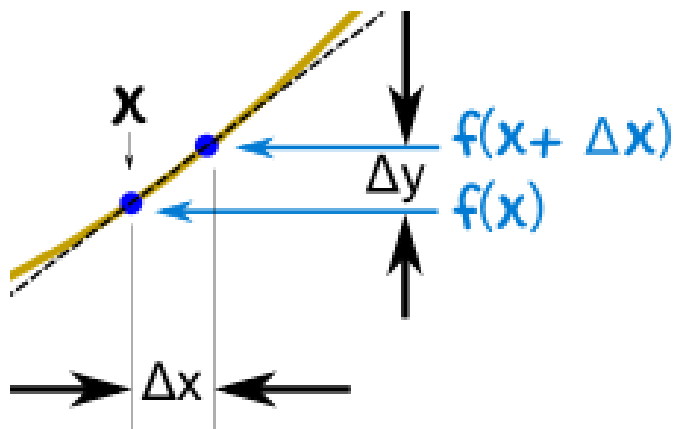
2-2. 미분과 도함수의 개념

미분 (Differentiation)과 도함수 (Derivatives)

도함수를 찾는 과정을 ‘미분 한다’고 한다.

도함수는 함수 곡선의 기울기 변화율을 말한다. 도함수는 아래와 같은 공식을 사용해 찾을 수 있으며 $f'(x)$ 로 표기된다.

$$\text{Slope} = \frac{\text{Change in } y}{\text{Change in } x} = \frac{\Delta y}{\Delta x} = \frac{dy}{dx} = \frac{f(x + \Delta x) - f(x)}{\Delta x}$$



2. 데이터 분석을 위한 수학 리뷰

2-2. 미분과 도함수의 개념

다변수 함수의 미분 (Multivariable Function Derivative)

$f(x, y, \dots)$ 와 같이 매개 변수가 여러 개인 경우 편미분(Partial Derivative)을 활용한다.

편미분은 특정 변수를 제외한 나머지 변수를 상수로 간주하여 미분하는 방법이다.

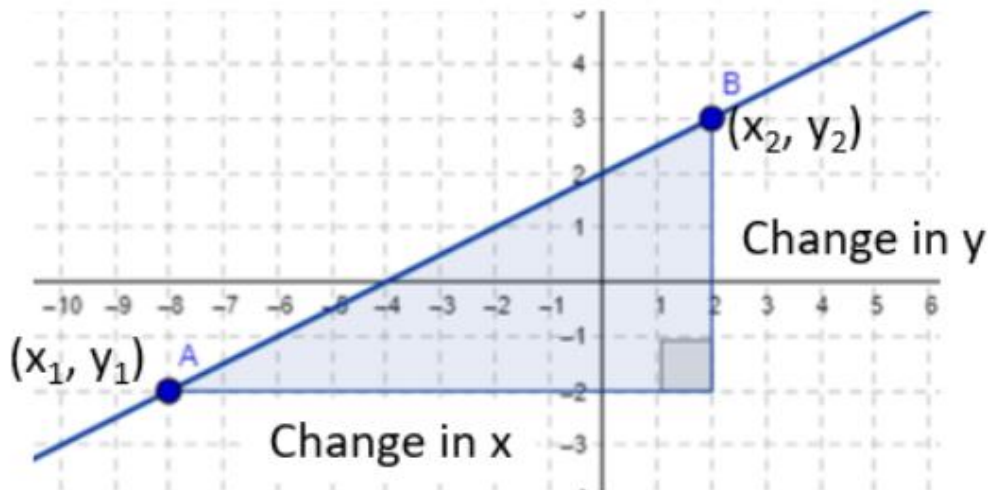
$$\begin{aligned} z = 2x^2y & \xrightarrow[\text{x를 상수 취급: } z = (2x^2)y]{\text{y에 대하여 편미분}} \frac{\partial z}{\partial y} = 2x^2 \times 1 \\ & = 2x^2 \\ & \xrightarrow[\text{y를 상수 취급: } z = (2y)x^2]{\text{x에 대하여 편미분}} \frac{\partial z}{\partial x} = 2y \times 2x \\ & = 4xy \end{aligned}$$

2. 데이터 분석을 위한 수학 리뷰

2-2. 미분과 도함수의 개념

그레디언트 (Gradient)

그레디언트 즉, 기울기 벡터는 x 방향으로의 편미분 값과 y 방향으로의 편미분 값을 원소로 하는 벡터이다. 벡터 공간에서 $f(x, y)$ 와 같은 스칼라 함수의 변화량을 알기 위해 사용된다.



$$\text{Gradient} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\text{Change in } y}{\text{Change in } x}$$

2. 데이터 분석을 위한 수학 리뷰

2-2. 미분과 도함수의 개념

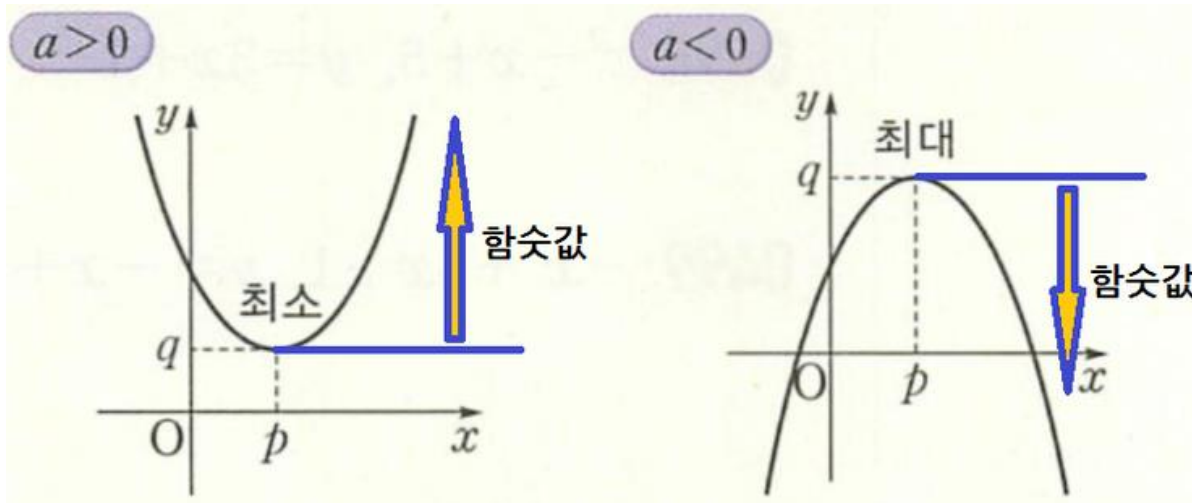
최소값 구하기

어떤 함수의 최소값 최대값은 y 의 범위를 구하면 알 수 있다. y 즉, 함수값이 가장 작은 값이 최소값, 가장 큰 값이 최대값이다.

x 의 범위가 없는 경우,

$a > 0$ 일 때, $x=p$ 에서 최소값 q 를 가지며 최대값은 구할 수 없다.

$a < 0$ 일 때, $x=p$ 에서 최대값 q 를 가지며 최소값은 구할 수 없다.



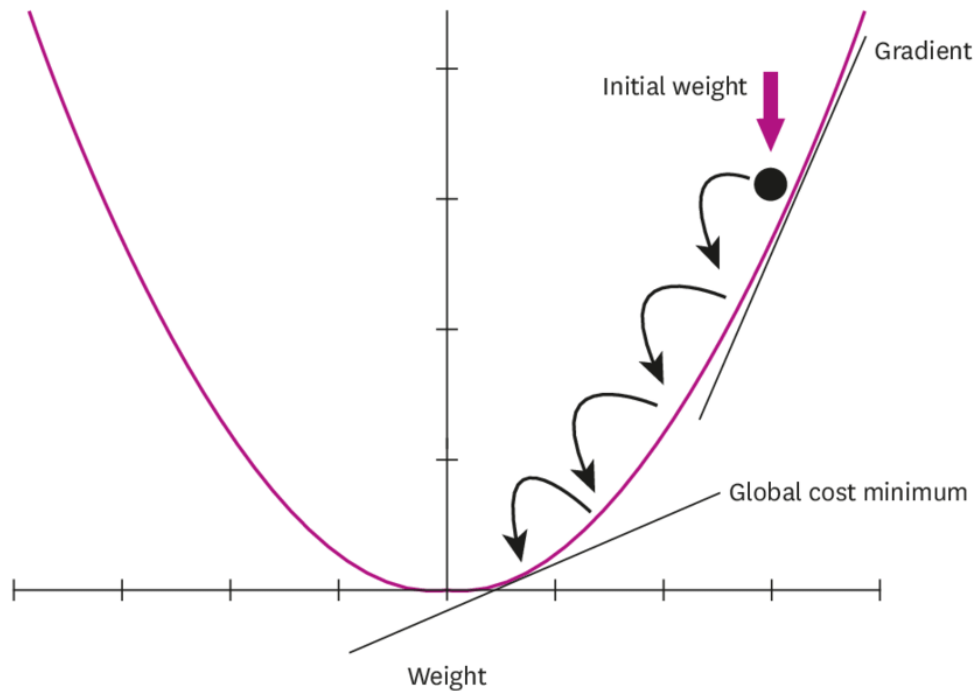
2. 데이터 분석을 위한 수학 리뷰

2-2. 미분과 도함수의 개념

경사 하강법 (Gradient Descent)

경사 하강법은 머신러닝 및 딥러닝에서 알고리즘을 학습시킬 때 사용되는 방법 중 하나이다.

함수의 기울기 즉, Gradient를 반복적으로 이동시키며, 미분 가능한 함수의 최소값을 찾는 방법이다.



2. 데이터 분석을 위한 수학 리뷰

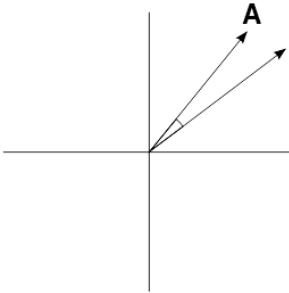
2-3. 유사도 측정법

코사인 유사도 (Cosine Similarity)

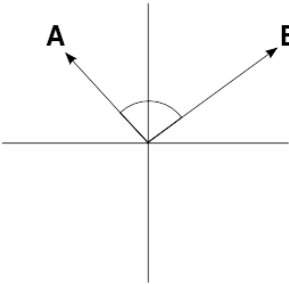
유사도 측정법은 클러스터링, 분류 등과 같은 머신 러닝의 비지도학습에서 많이 사용된다.
두 개체의 거리가 가까우면 유사도가 높고, 멀면 유사도가 낮은 것으로 본다.

코사인 유사도는 두 벡터 사이의 코사인 각도를 이용해 두 벡터의 유사도를 구하는 방식이다.

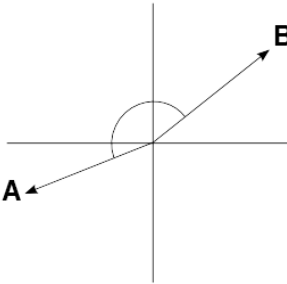
Similar



Unrelated



Opposite



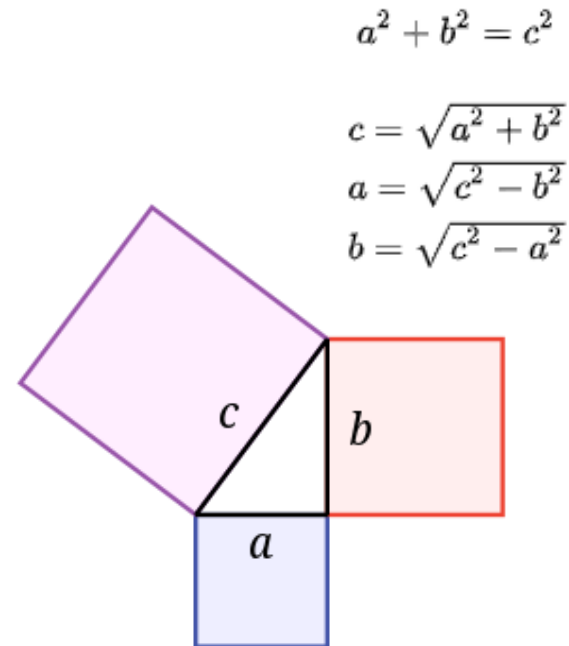
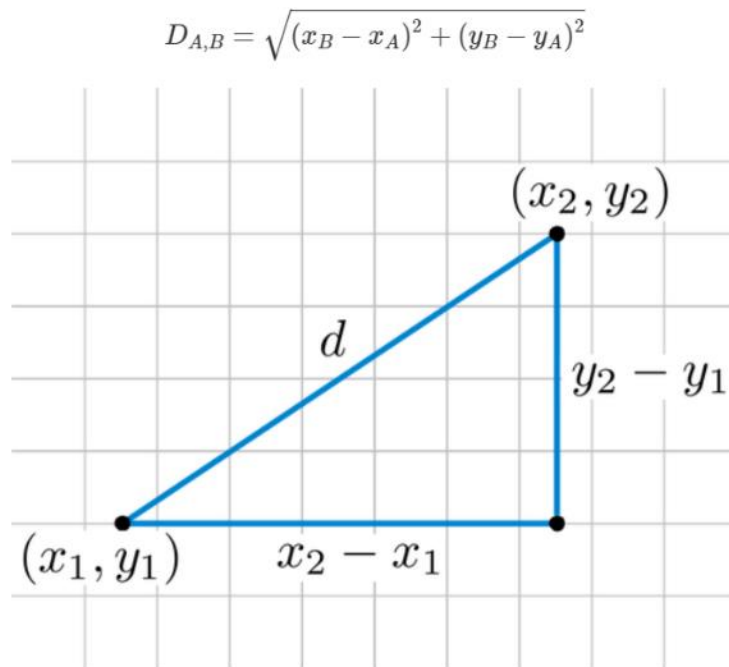
$$\text{similarity}(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n A_i^2} \times \sqrt{\sum_{i=1}^n B_i^2}}$$

2. 데이터 분석을 위한 수학 리뷰

2-3. 유사도 측정법

유클리디언 유사도 (Euclidean Similarity)

유클리디언 유사도는 유클리디언 거리를 이용해 두 점이 얼마나 멀리 떨어져 있는가에 대한 유사도를 측정하는 방식이다. 유클리디언 거리는 피타고라스의 정리에 기반한다.



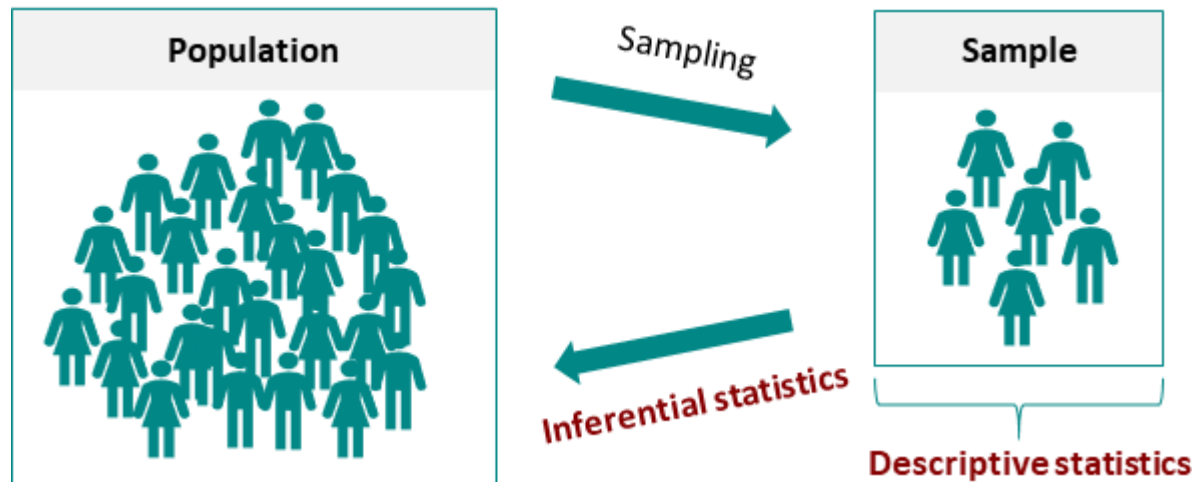
3. 데이터 분석을 위한 통계학 리뷰

3. 데이터 분석을 위한 통계학 리뷰

3-1. 기술 통계와 추론 통계 (Descriptive & Inferential Statistics)

통계 분석은 크게 기술 통계와 추론 통계로 나눌 수 있다.

- **기술 통계** : 수집한 데이터의 통계적 특성을 정리, 요약, 해석하는 통계 기법 (결론을 내거나 예측을 하지 않는다.)
- **추론 통계** : 수집한 데이터 표본(Sample)의 특성을 파악하여 모집단을 추론하는 통계 기법




3. 데이터 분석을 위한 통계학 리뷰

3-2. 척도 (Scales of Measurement)

- 명목 변수와 서열 변수는 범주형 이고 등간 변수와 비율 변수는 수치형이다.
- 범주형 데이터 보다 수치형 데이터에 대해 더 많은 통계 테스트를 해볼 수 있다.
- 등간 (예: 온도) 및 비율(예: 거리) 척도는 모두 동일하게 '간격'이라는 특성을 갖지만 비율 척도에만 절대 '0'이 있다.

	명목 척도	서열 척도	등간 척도	비율 척도
	Nominal	Ordinal	Interval	Ratio
Categories	●	●	●	●
Rank order		●	●	●
Equal spacing			●	●
True zero				●

 **The 4 levels of measurement**

3. 데이터 분석을 위한 통계학 리뷰

3-3. 모수와 통계량 (Parameter and Statistic)

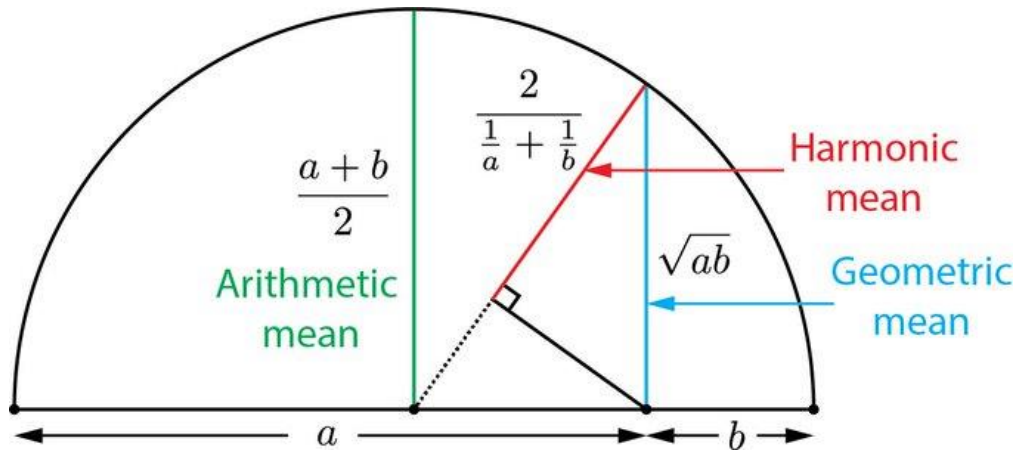
모집단의 특성 즉, 모수 파악하는 데는 시간과 비용의 제약이 따른다. 따라서, 표본의 통계량을 통해 모수를 추측하게 된다. 모집단의 모수를 알지 못하는 상태에서 통계량을 이용해 모수를 측정하는 과정을 통계적 추정(Statistical Estimation)이라고 한다.

- **모수** : 모집단 (Population)의 특성을 나타내는 척도
- **통계량** : 표본 (Sample)의 특성을 나타내는 척도

		모집단	표본
		Population	Sample
	# of subjects	N	n
평균	Mean	$\mu = \frac{\sum_{i=1}^N x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$
분산	Variance	$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$	$S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$
		Note: S^2 is the formula for unbiased sample variance, since we're dividing by $n - 1$.	
표준편차	Standard deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$	$S = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$
		Note: Finding S by taking $\sqrt{S^2}$ reintroduces bias.	

3. 데이터 분석을 위한 통계학 리뷰

3-3. 모수와 통계량 (Parameter and Statistic)



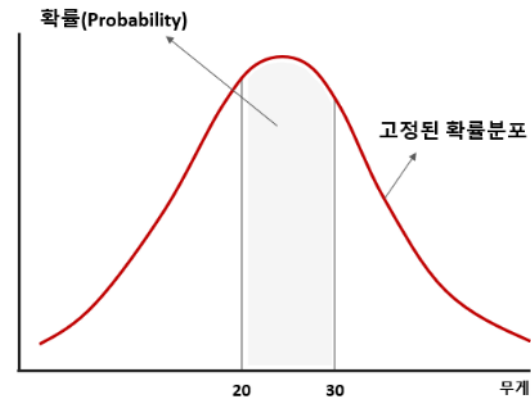
- **산술평균** (Arithmetic Mean)
관측된 값을 자료의 수 n 으로 나눈 값
- **기하평균** (Geometric Mean)
 n 개의 관측 값에 대한 곱의 n 제곱근
- **조화평균** (Harmonic Mean)
 n 개의 관측 값에 대해서 관측 값의 역수들을 산술 평균한 것에 다시 역수를 취한 대푯값

3. 데이터 분석을 위한 통계학 리뷰

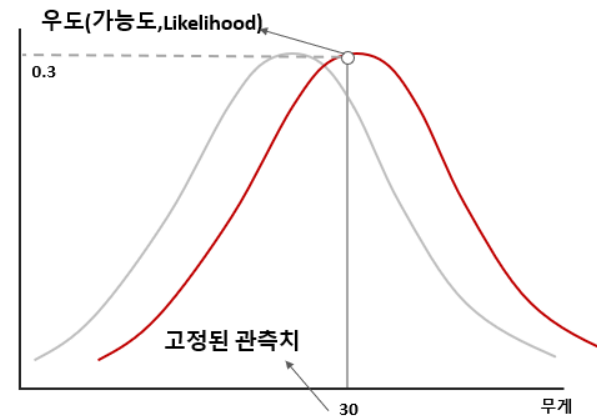
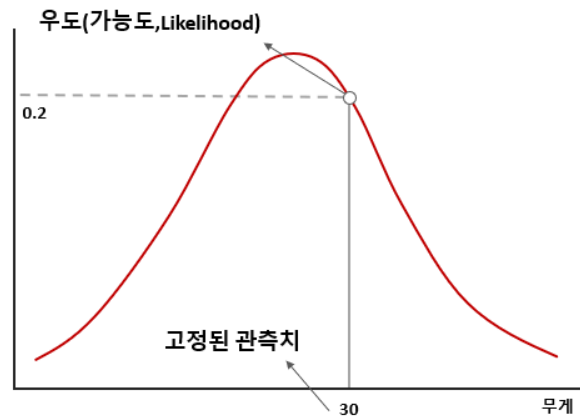
3-4. 확률

확률과 우도 (Probability & Likelihood)

- **확률** : 고정된 확률분포에서 어떠한 관측 값이 나타나는지에 대한 확률



- **우도** (또는 가능도) : 고정된 관측 값이 어떠한 확률분포에서 어느 정도의 확률로 나타나는지에 대한 확률



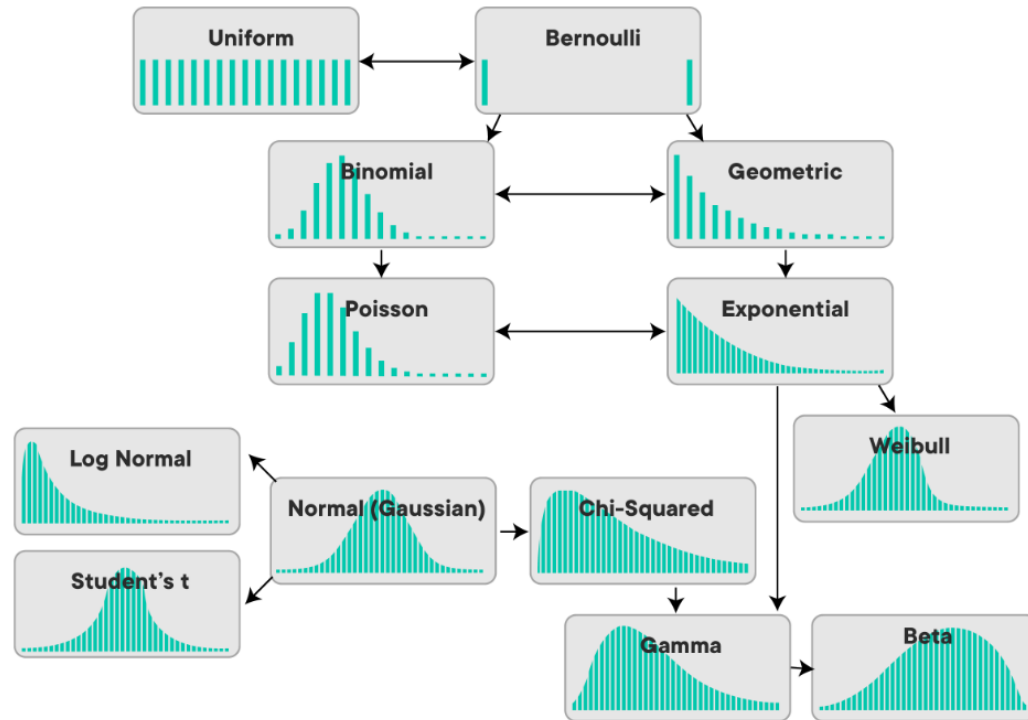
3. 데이터 분석을 위한 통계학 리뷰

3-4. 확률

확률 분포 (Probability Distribution)

자연에는 다양한 확률 분포가 존재한다. 확률 분포란 이벤트가 발생할 확률을 나타내는 함수이다.

결과 값을 셀 수 있는지에 따라 크게 **이산 분포** (Discrete Distribution)(예: 주사위 던지기), **연속 분포**(Continuous Distribution)(예: 날씨)로 나누어 볼 수 있다.



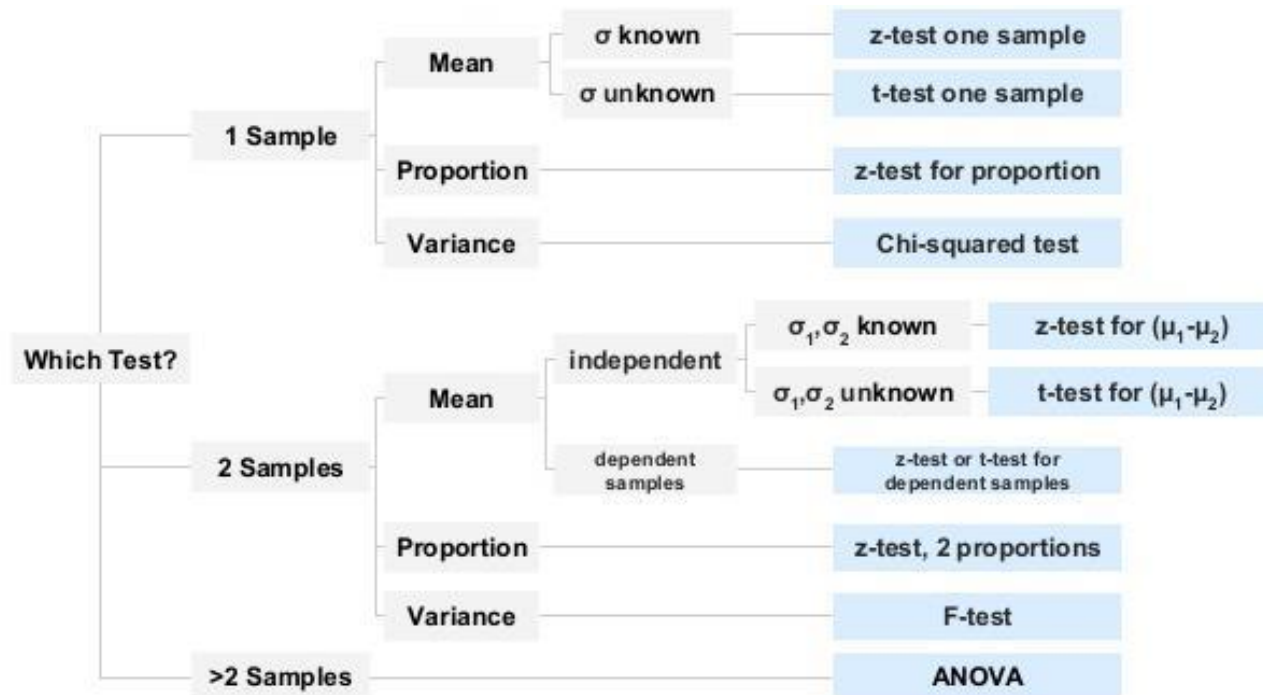
3. 데이터 분석을 위한 통계학 리뷰

3-4. 확률

확률 분포와 검정 통계량 (Probability Distribution & Test Statistic)

검정 통계량은 가설검정을 위한 통계량이다.

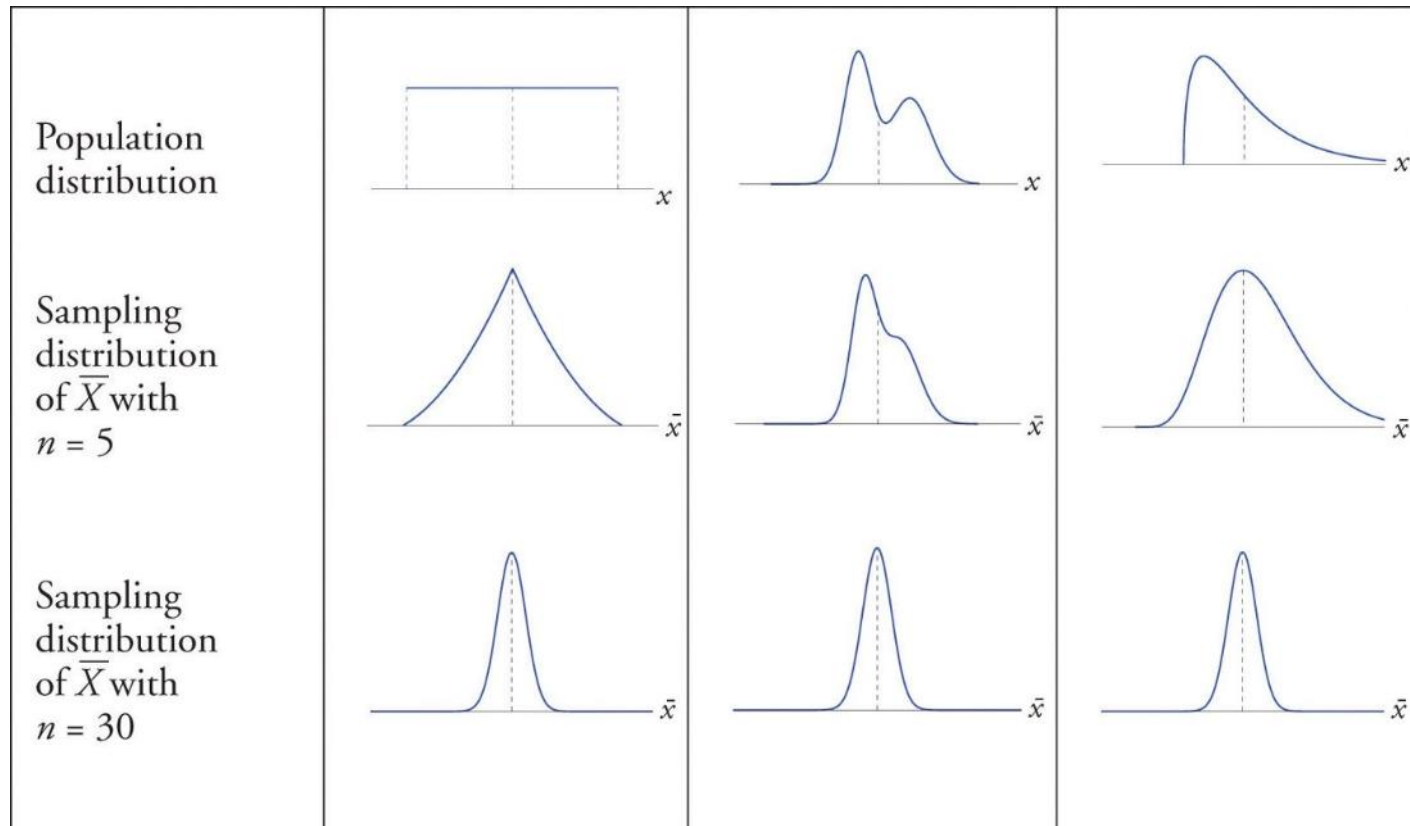
통계적 유의성을 설정하고 귀무가설 (Null Hypothesis)을 기각할지 여부를 결정하는데 사용된다. 가설 검정을 위해 확률 분포를 활용하는데, 데이터의 형태에 따라 알맞은 검정 통계량을 선택해야 한다.



3. 데이터 분석을 위한 통계학 리뷰

3-5. 중심극한정리 (Central Limit Theorem, CLT)

모집단 분포의 모양에 관계 없이 표본의 크기가 증가함에 따라 평균의 표본 분포가 정규 분포에 접근 한다.

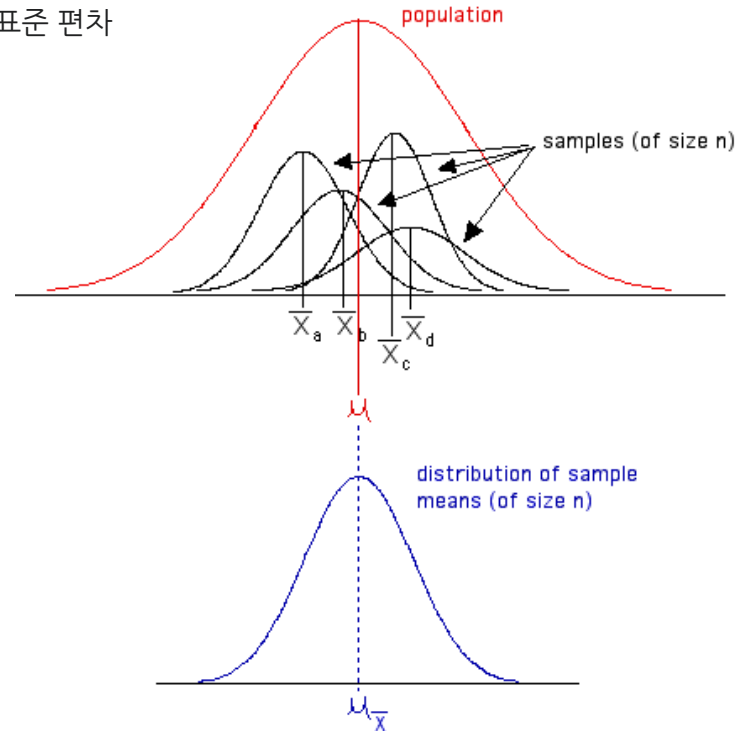


3. 데이터 분석을 위한 통계학 리뷰

3-6. 표집 분포 (Sampling Distribution)

표집 분포는 모집단에서 얻을 수 있는 특정 크기(n)의 가능한 모든 무작위 표본에 대한 평균 분포를 말한다.

- 표집 분포의 모양은 정규 분포인 경향이 있다. ($n > 30$ 일 경우, 정규 분포를 따른다.)
- **기대값** (Expected Value) : 모든 표본 평균의 평균
- **표준 오차** (Standard Error) : 표집 분포의 표준 편차



3. 데이터 분석을 위한 통계학 리뷰

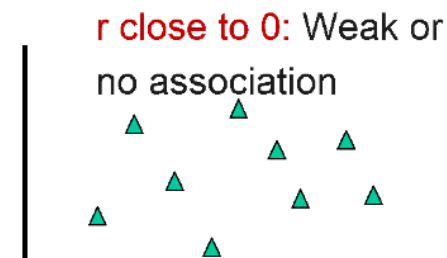
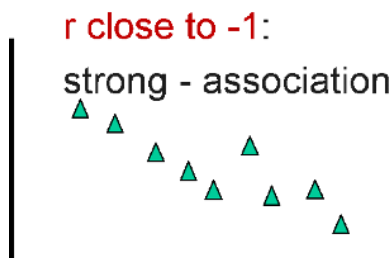
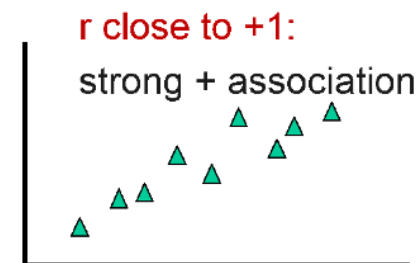
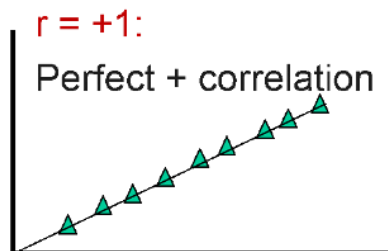
3-7. 상관 계수 (Correlation Coefficient)

상관계수는 두 변수 사이의 통계적 관계를 표현하기 위해 특정한 상관 관계의 정도를 -1에서 +1 사이 값으로 나타낸 계수이다. (Wikipedia)

* 참고: 거리와 상관계수

$$r = \frac{\text{Cov}(X, Y)}{\sqrt{s_x^2 s_y^2}}$$

$$\text{Cov}(X, Y) = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n-1}$$

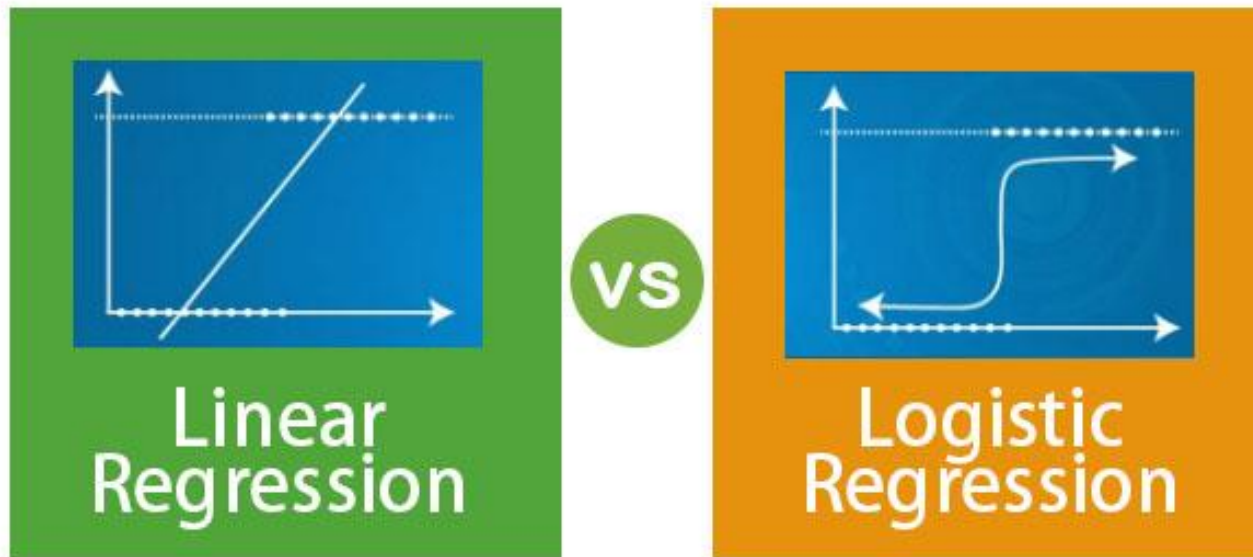


3. 데이터 분석을 위한 통계학 리뷰

3-8. 회귀 분석 & 로지스틱 회귀 분석

선형 회귀 (Linear Regression)는 직선을 사용하여 두 변수 간의 관계를 설정하는 회귀 분석으로 독립 변수의 변경에 따른 연속 종속 변수(예: 나이, 키, 가격 등)를 예측하는데 사용된다.

로지스틱 회귀 (Logistic Regression)는 범주형 결과(예: Yes / No)를 예측 하는데 사용되며 시그모이드 활성화 함수(Sigmoid Active Function)를 통해 얻은 S자 비선형 곡선이 활용된다.



3. 데이터 분석을 위한 통계학 리뷰

3-9. 가설 검정 (Hypothesis Test)

T-검정 (T-Test)

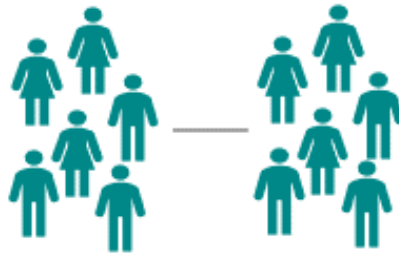
T-검정은 모집단과 그룹의 평균 또는, 두 그룹의 평균 차이가 통계적으로 유의미한 지 확인할 때 사용된다.
T-검정을 사용하기 위해서는 표본의 정규성, 등분산성, 독립성이 만족되어야 한다.

One sample t-test



Is there a **difference** between a **group** and the **population**

Independent samples t-test



Is there a **difference** between **two groups**

Paired samples t-test



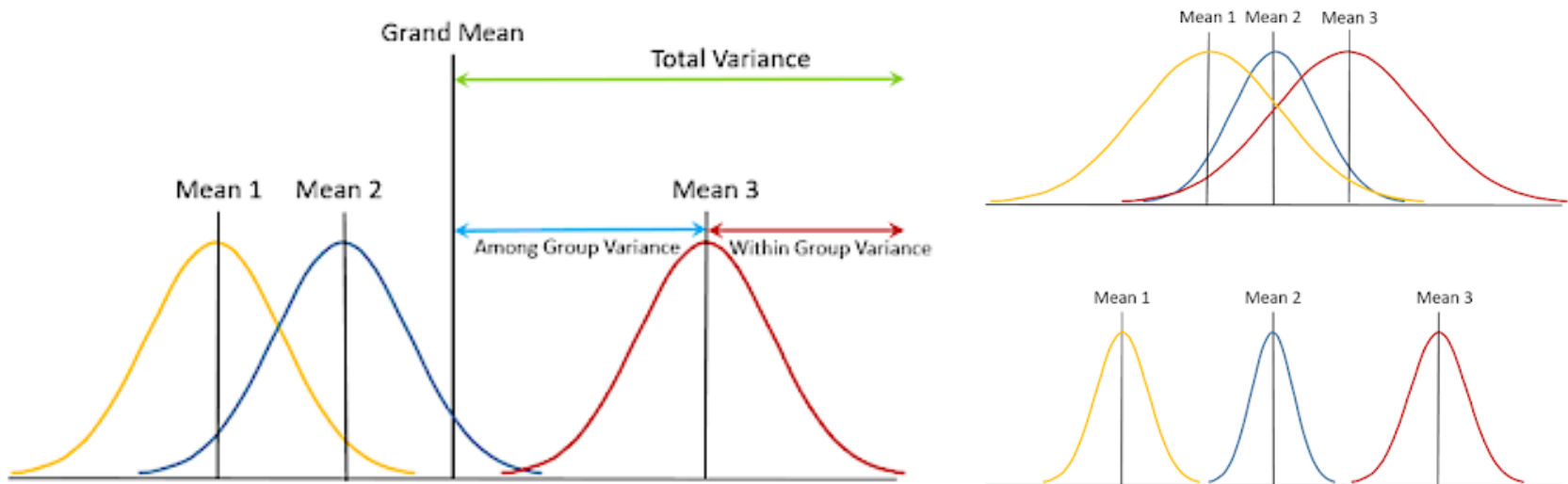
Is there a **difference** in a **group** between **two points in time**

3. 데이터 분석을 위한 통계학 리뷰

3-9. 가설 검정 (Hypothesis Test)

ANOVA 분석 (Analysis of Variance, ANOVA)

그룹이 3개 이상일 경우에는 그룹간 차이를 비교하기 위해 T-검정(평균 분석) 대신 ANOVA(분산 분석)를 사용한다. F-value가 클수록 두 그룹의 분산에 차이가 있음을 의미한다.



3. 데이터 분석을 위한 통계학 리뷰

3-9. 가설 검정 (Hypothesis Test)

ANOVA 분석 (Analysis of Variance, ANOVA)

1. F-value 계산법은 다음과 같다.

$$F = \frac{MSR}{MSE} = \frac{RSS/k}{SSE/(n-k-1)}$$

where:

RSS = regression sum of squares

SSE = sum of squared errors

MSR = mean regression sum of squares

MSE = mean squared error

2. 분자와 분모의 자유도는 다음과 같다.

$$df_{\text{numerator}} = k$$

$$df_{\text{denominator}} = n - k - 1$$

where:

n = number of observations

k = number of independent variables

3. F-검정의 결정 규칙은 다음과 같다.

Decision rule: reject H_0 if F (test-statistic) $> F_c$ (critical value)

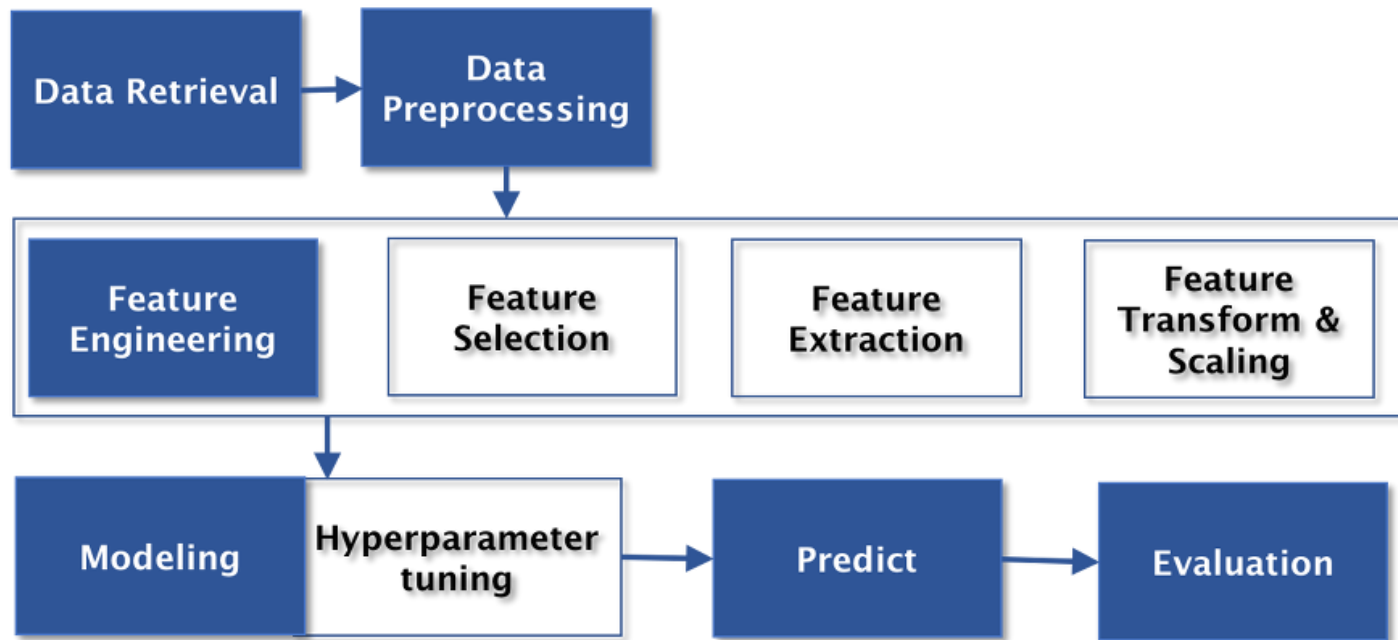
4. Machine Learning

4. Machine Learning

4-1. 머신 러닝의 작업 흐름

머신 러닝의 전체 흐름에서 가장 중요한 단계는 Data Preprocessing과 Feature Engineering이다.

비정형 데이터를 정형 데이터로 변환, 결측치/이상치 처리, 중복된 데이터를 제거, 데이터 차원의 축소 등의 작업을 통해 모델링 단계에서의 계산 시간과 비용을 절약하고 모델의 성능을 향상 시킬 수 있다.



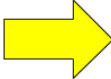
4. Machine Learning

4-2. 전처리 (Preprocessing)

데이터 분석의 안정적인 결과와 성능 향상을 위해서 주어진 데이터를 분석에 적합하게 가공하는 작업이다.
대표적인 작업으로는 필터링, 클리닝, 결측치 처리, 이상치 처리, 데이터 형태 변경 등이 있다.

- Filtering / Cleaning / Missing Value / Outlier
- 범주형 데이터 인코딩 : 레이블 인코딩 (Label Encoding) & 원핫 인코딩 (One-hot Encoding)
- Feature 스케일링 : 표준화 (Normalization) & 정규화 (Standardization)
- Data Shape : Long Data, Wide Data

One-hot Encoding

Color		Red	Yellow	Green
Red				
Red		1	0	0
Yellow		1	0	0
Green		0	1	0
Yellow		0	0	1

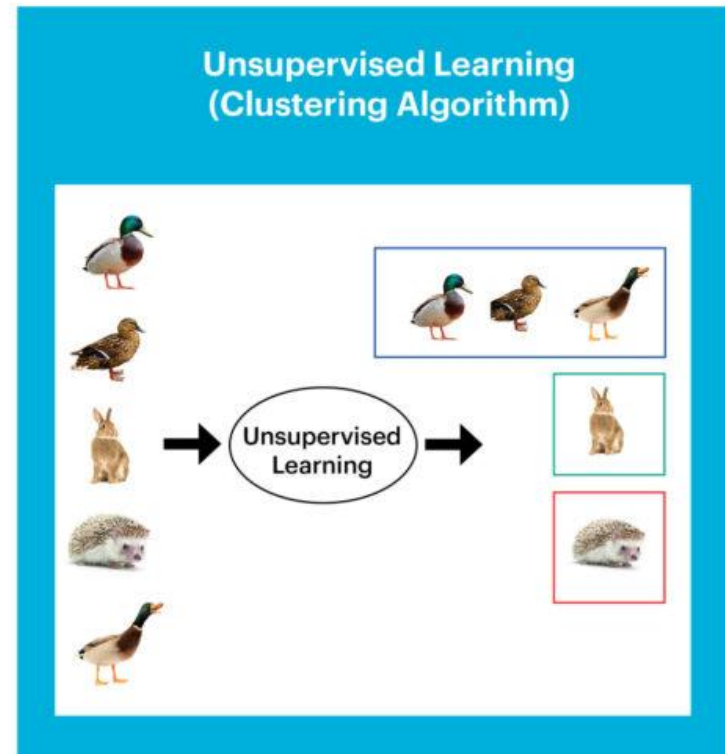
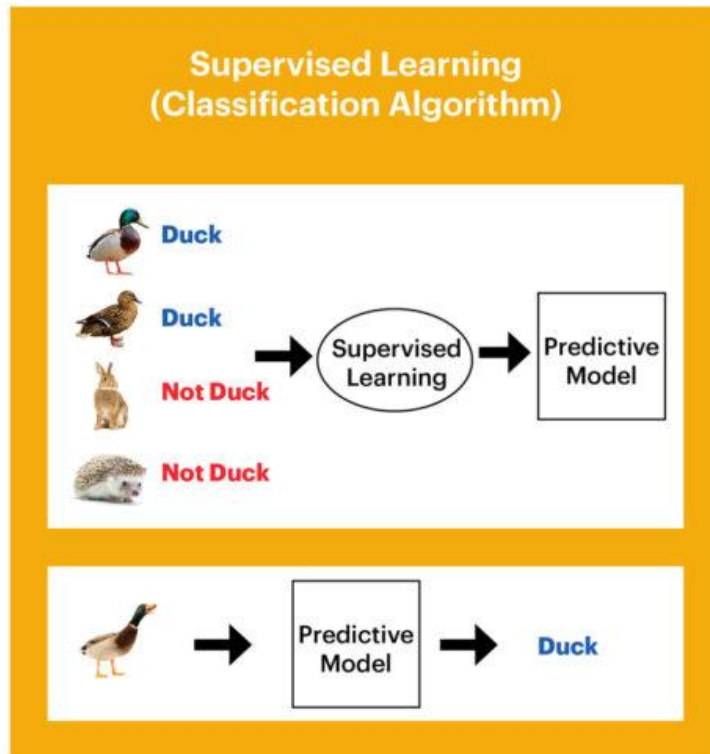
Feature Scaling

Standardisation	Normalisation
$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation}(x)}$	$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$

4. Machine Learning

4-3. 지도 학습과 비지도 학습 (Supervised Learning & Unsupervised Learning)

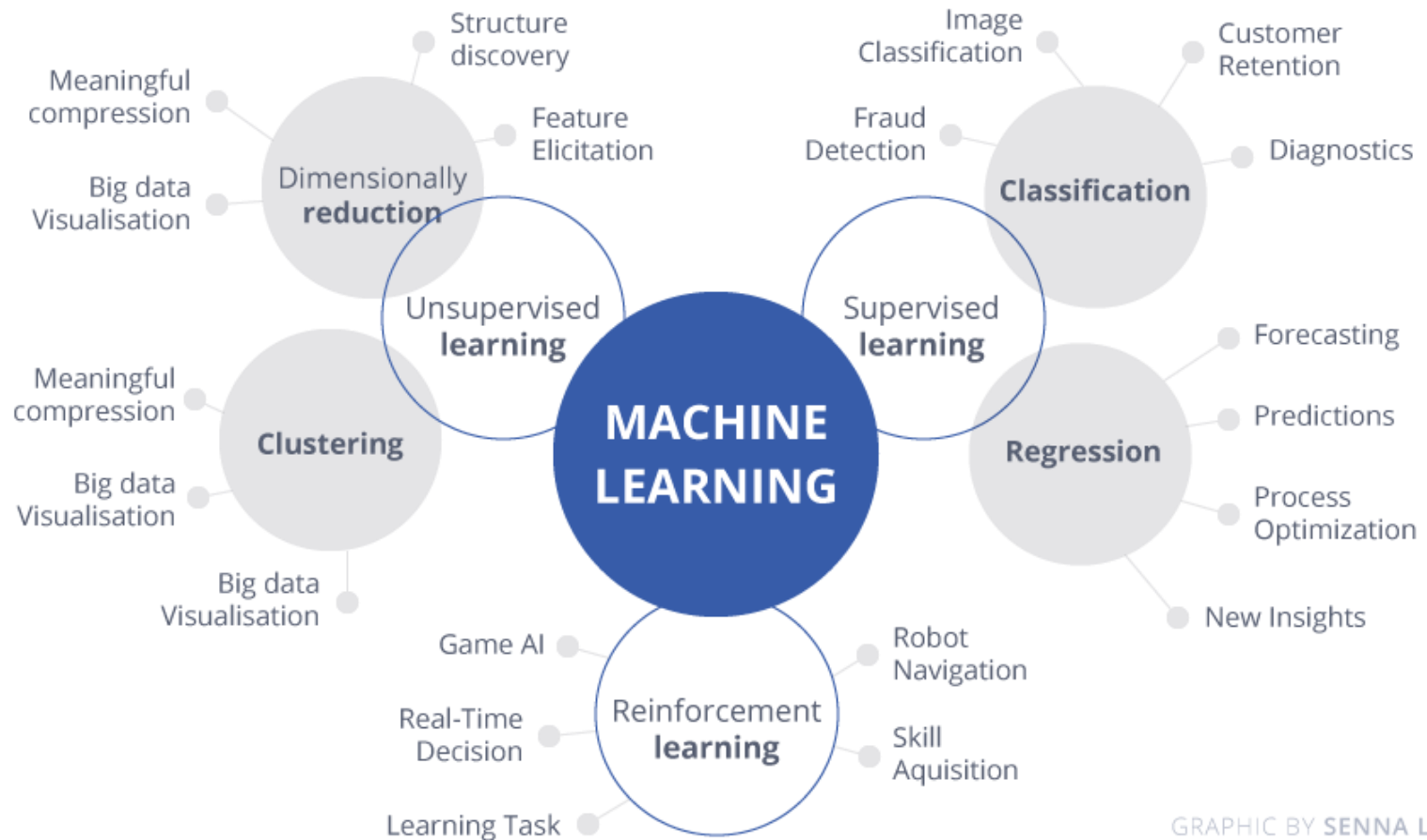
지도 학습과 비지도 학습의 가장 큰 차이점은 학습 데이터에서의 레이블(Label) 유무이다.



Western Digital.

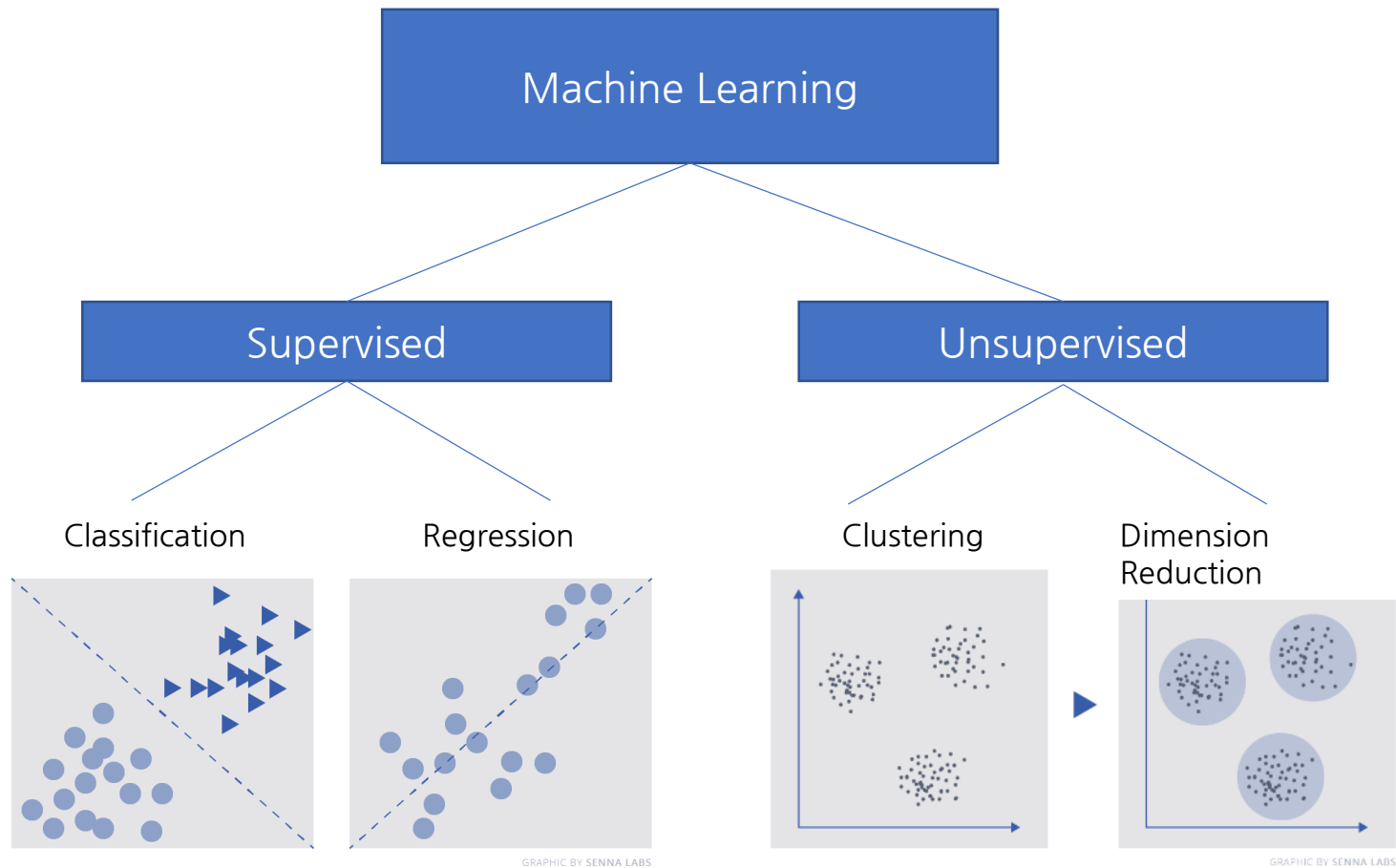
4. Machine Learning

4-3. 지도 학습과 비지도 학습 (Supervised Learning & Unsupervised Learning)



4. Machine Learning

4-3. 지도 학습과 비지도 학습 (Supervised Learning & Unsupervised Learning)



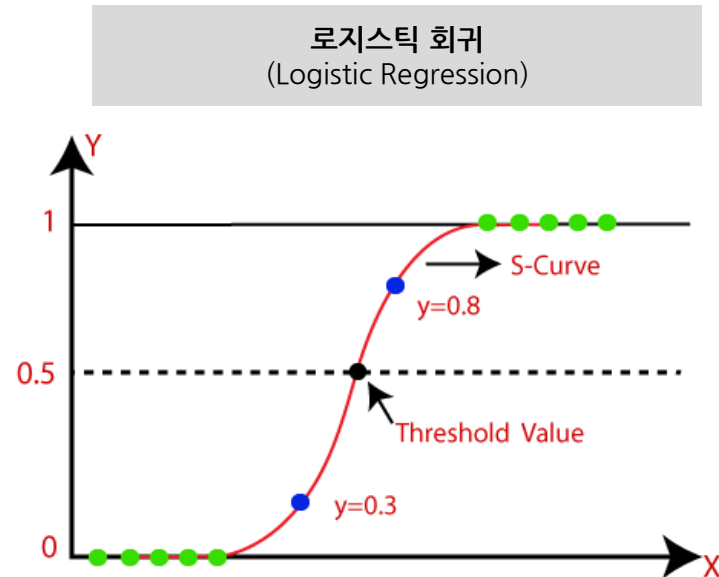
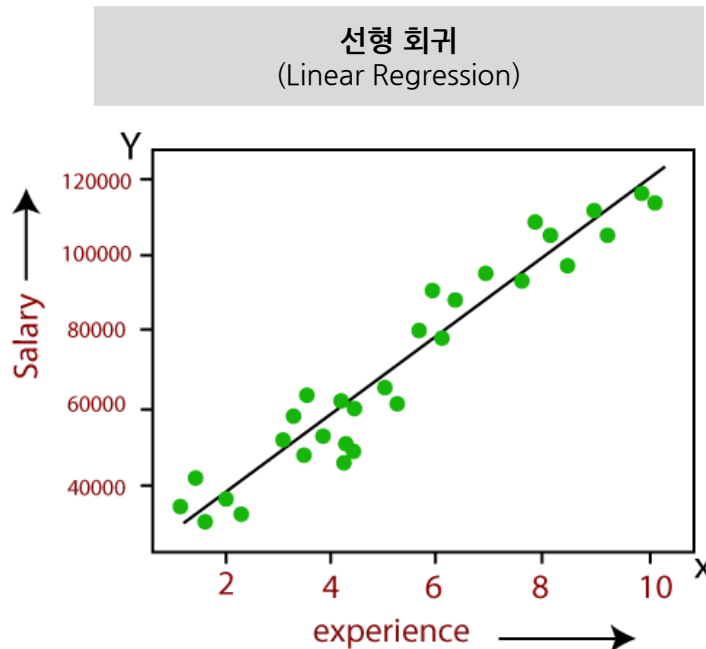
4. Machine Learning

4-4. Regression 계열 알고리즘

회귀 (Regression) 알고리즘은 레이블링 된 데이터를 기반으로 결과 값을 예측하는데 사용되며 지도학습에 해당된다. 회귀 모델은 'Yes / No'등 범주형 값을 예측하는 **로지스틱 회귀** (Logistic Regression)과 온도, 주가 등 연속되는 값을 예측하는 **선형 회귀** (Linear Regression)으로 구분된다.

* 참고 : Ridge / Lasso / Elastic Net

SVM (Support Vector Machine)



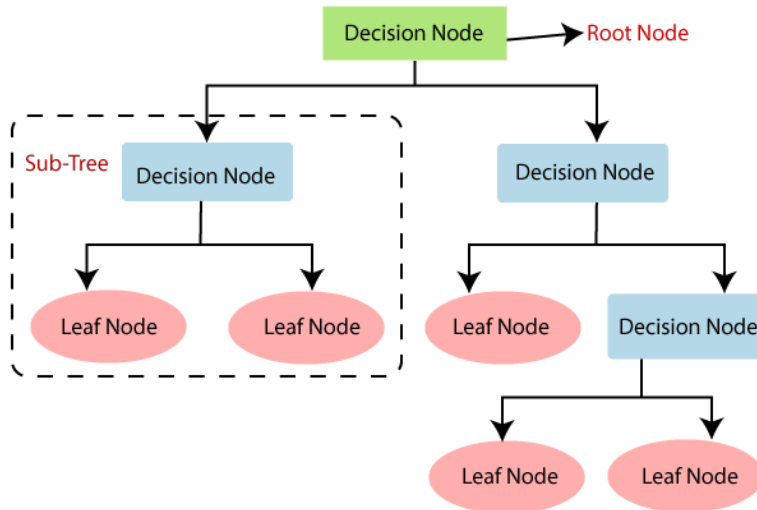
4. Machine Learning

4-5. Tree 계열 알고리즘

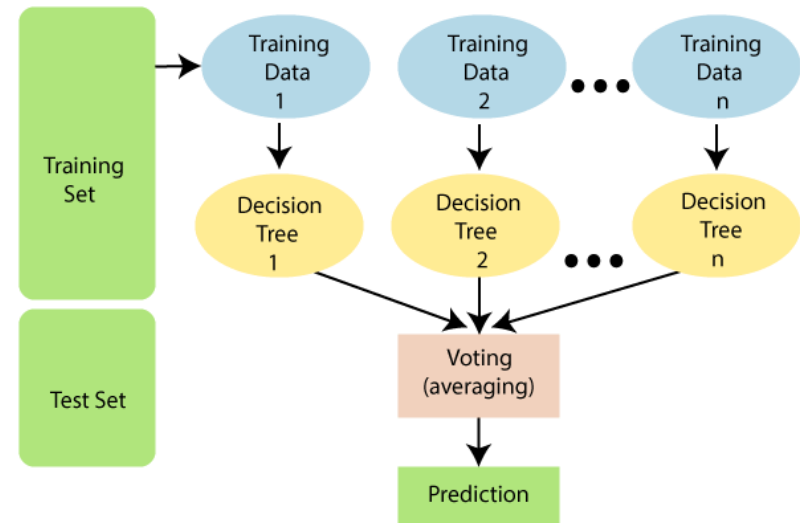
Tree 계열의 알고리즘은 지도학습 기술로 분류와 예측 문제에 모두 사용할 수 있지만 대부분 분류 문제 해결에 선호된다. 랜덤 포레스트 모델은 모델의 성능 향상을 위해 여러 Tree 모델을 결합한 앙상블 학습의 개념이다.

* 참고 : GBM / XGBoost / LightGBM

의사결정나무
(Decision Tree)



랜덤 포레스트
(Random Forest)



4. Machine Learning

4-5. Tree 계열 알고리즘

앙상블 (Ensemble)

여러 개의 분류기(Classifier)를 생성하고 그 예측을 결합함으로써 보다 정확한 최종 예측을 수행한다.

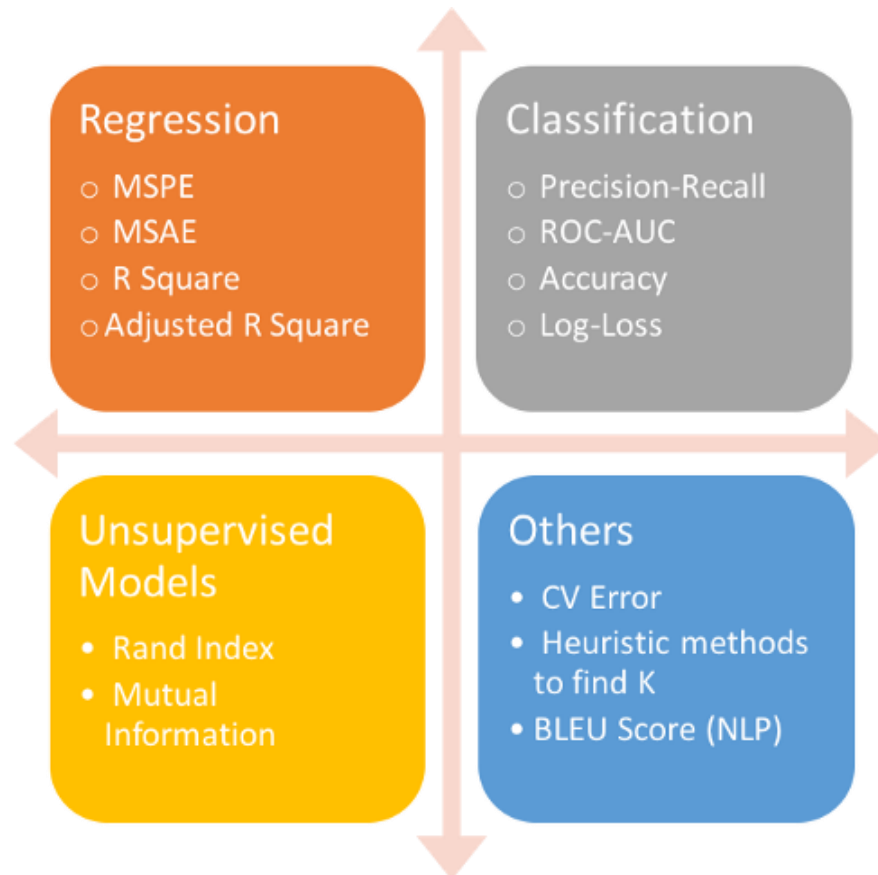
앙상블 학습 유형:

구분	지도학습	비고
보팅 Voting	서로 다른 알고리즘이 같은 데이터 세트에 대해 학습하고 예측한 결과를 보팅 (Hard Voting / Soft Voting)	랜덤 포레스트
배깅 Bagging	단일 결정 트리로 데이터 샘플링을 서로 다르게 가져가면서 학습을 수행해 보팅	
부스팅 Boosting	여러 개의 분류기가 순차적으로 학습하면서 앞에서 학습한 분류기가 틀린 데이터에 대해서는 가중치를 부여하면서 학습과 예측을 진행	GBM / XGBoost
스태킹 Stacking	스태킹은 여러가지 다른 모델의 예측 결과값을 다시 학습데이터로 만들어 다른 모델로 재학습시켜 결과를 예측하는 방법	-

4. Machine Learning

4-6. 모델 평가 (Model Evaluation)

기계 학습에는 모델의 성능을 평가하기 위한 다양한 지표들이 있으며, 각각의 알고리즘의 특성에 따라 적절한 평가지표를 사용한다.



4. Machine Learning

4-6. 모델 평가 (Model Evaluation)

회귀 모델 (Regression Model) 성능 평가지표

평가 지표	설명	수식
MAE	Mean Absolute Error이며 실제 값과 예측 값의 차이를 절대값으로 변환해 평균한 것	$\frac{1}{n} \sum_{i=1}^n y_i - \hat{y}_i $
MSE	Mean Squared Error이며 실제 값과 예측 값의 차이를 제곱해 평균한 것 *MAE값이 같은데 MSE가 클 경우 편차가 더 큼을 나타낸다.	$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$
RMSE	MSE 값은 오류의 제곱을 구하므로 실제 오류 평균보다 더 커지는 특성이 있으므로 MSE에 루트를 씌운 것이 RMSE(Root Mean Squared Error)다.	$\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$
R ²	분산 기반으로 예측 성능을 평가합니다. 실제 값의 분산 대비 예측값의 분산비율을 지표로 하며, 1에 가까울수록 예측 정확도가 높다. *R ² = 0.91인 경우, 전체 데이터 변동성의 91%를 선형회귀 모델이 설명	$\frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$

4. Machine Learning

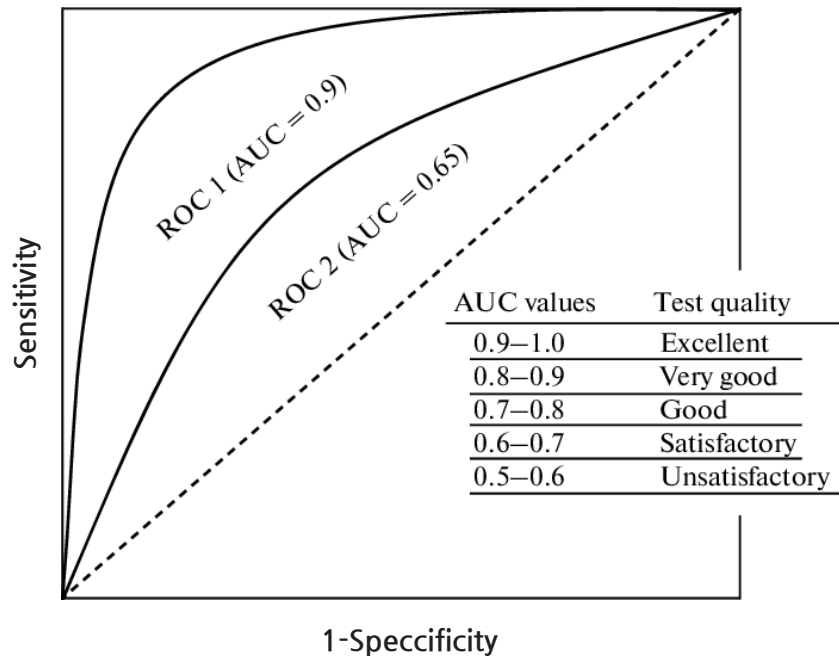
4-6. 모델 평가 (Model Evaluation)

ROC (Receiver Operating Characteristic) / AUC (Area Under Curve)

ROC/AUC는 분류 모델의 성능을 평가하기 위한 지표이다.

ROC/AUC는 오차 행렬의 민감도와 특이성을 그래프화 한 것으로 AUC 값이 1에 가까울 수록 분류를 잘 하는 모델이다.

ROC는 확률 곡선이고 AUC value는 ROC 곡선의 면적을 구한 값이다.



4. Machine Learning

4-6. 모델 평가 (Model Evaluation)

오차 행렬 (Confusion Matrix)

오차 행렬(또는 혼동 행렬)은 분류 모델의 성능을 평가하기 위해 실제 값(Actual Values)과 예측 값(Predictive Value)을 비교 하는 표이다.

		PREDICTIVE VALUES		
		POSITIVE (1)	NEGATIVE (0)	
ACTUAL VALUES	POSITIVE (1)	TP = 3	FN = 1	4
	NEGATIVE (0)	FP = 2	TN = 4	6
		5	5	

PRECISION (green box around TP and FP)

RECALL (red box around TP and FN)

정확도 $Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$

재현도 $Recall / Sensitivity = \frac{TP}{TP + FN}$

특이성 $Specificity = \frac{TN}{TN + FP}$

정밀도 $Precision = \frac{TP}{TP + FP}$

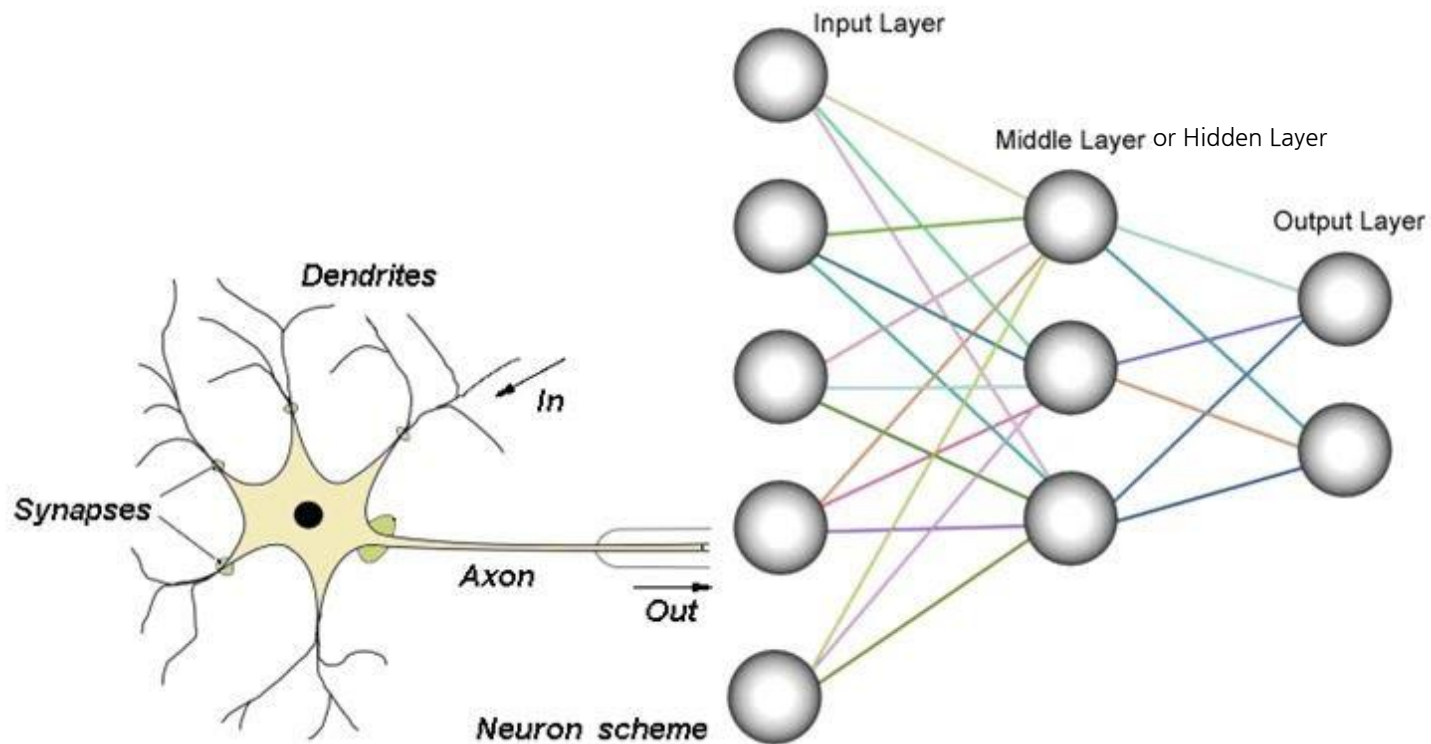
$$F1-Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

5. Deep Learning

5. Deep Learning

5-1. 인공 신경망 (Artificial Neural Network, ANN)

인공 신경망이란 인간이 배우는 방식을 복제하기 위해 인간 두뇌의 신경망 작용을 모방하는 시스템이다.
신경망은 뇌와 마찬가지로 시행 착오를 통해 패턴을 학습함으로써 훈련된다.



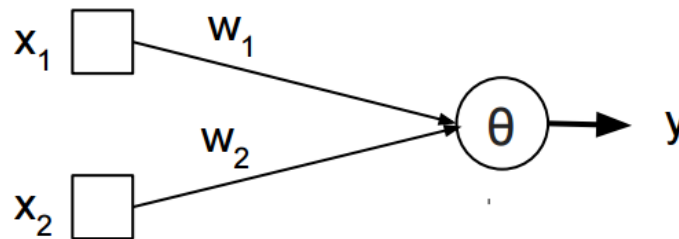
5. Deep Learning

5-2. XOR 문제

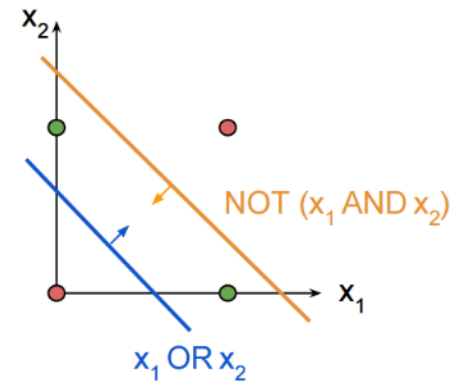
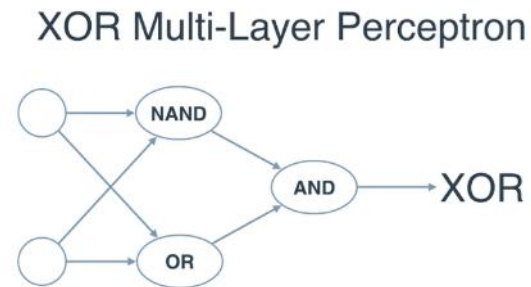
퍼셉트론 (Perception + Neuron)은 인공 신경망의 한 종류이다.

일반적으로 퍼셉트론 은 선형으로 분리 가능한 함수만 정의할 수 있다. 그러나 XOR과 같은 문제를 해결 하려면 두 줄 (다층 퍼셉트론)이 필요하다.

단층 퍼셉트론
(Single-layer Perceptron)



다층 퍼셉트론
Multi-layer Perceptron)

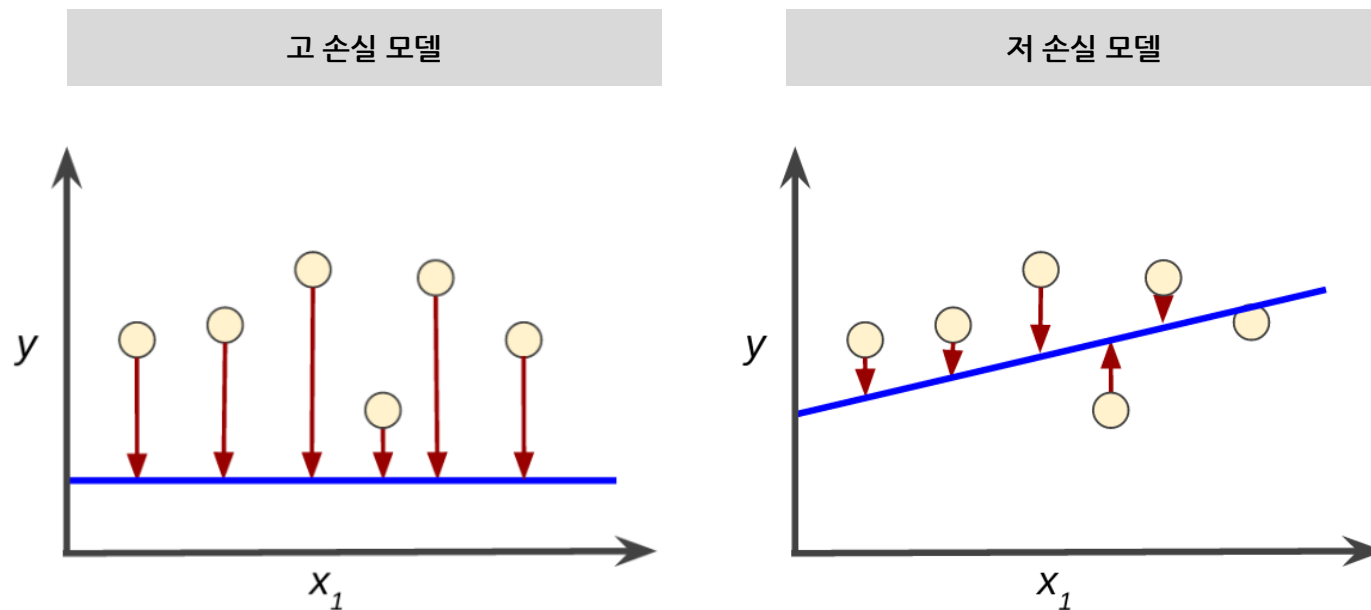


5. Deep Learning

5-3. 손실 함수 (Cost Function)

손실 함수는 모델이 예측하는 값과 실제 결과 값 사이의 오차를 측정하여 모델의 성능을 평가하는 함수이다.

손실 함수의 목적은 '손실을 최소화' 하거나 '보상을 최대화' 하는 매개 변수(가중치 W + 편향 b)의 값을 찾는 것이다.



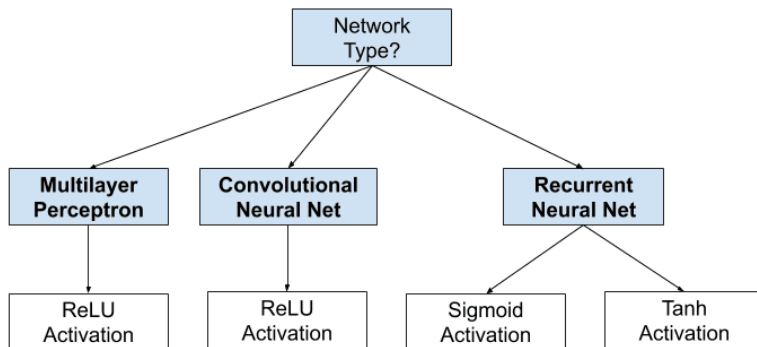
5. Deep Learning

5-4. 활성화 함수 (Activation Function)

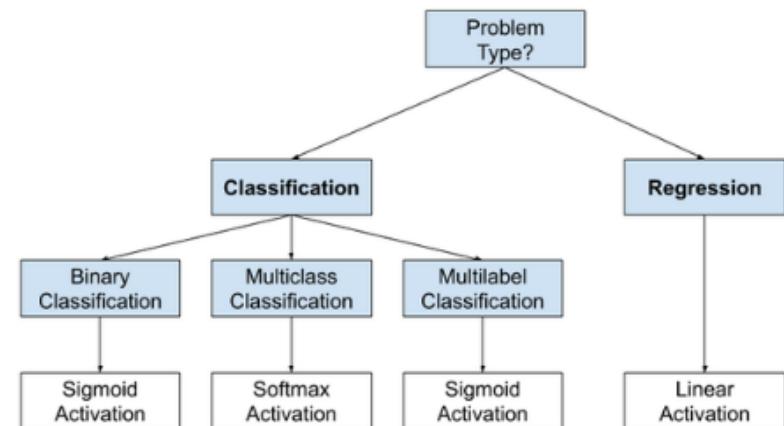
활성화 함수는 가중치 합을 계산하고 바이어스를 추가하여, 인공 신경망의 활성화 여부를 결정한다.

활성화 함수는 히든 레이어에서 사용되는 것과 출력 레이어에서 사용되는 것으로 구분할 수 있다.

히든 레이어의 활성화 함수



출력 레이어의 활성화 함수



- 일반적으로 미분 할 수 있는 비선형 활성화 함수는 신경망의 숨겨진 계층에서 사용된다. 그 이유는 선형 활성화 함수를 사용하여 훈련된 네트워크보다 모델이 더 복잡한 패턴을 학습할 수 있기 때문이다.

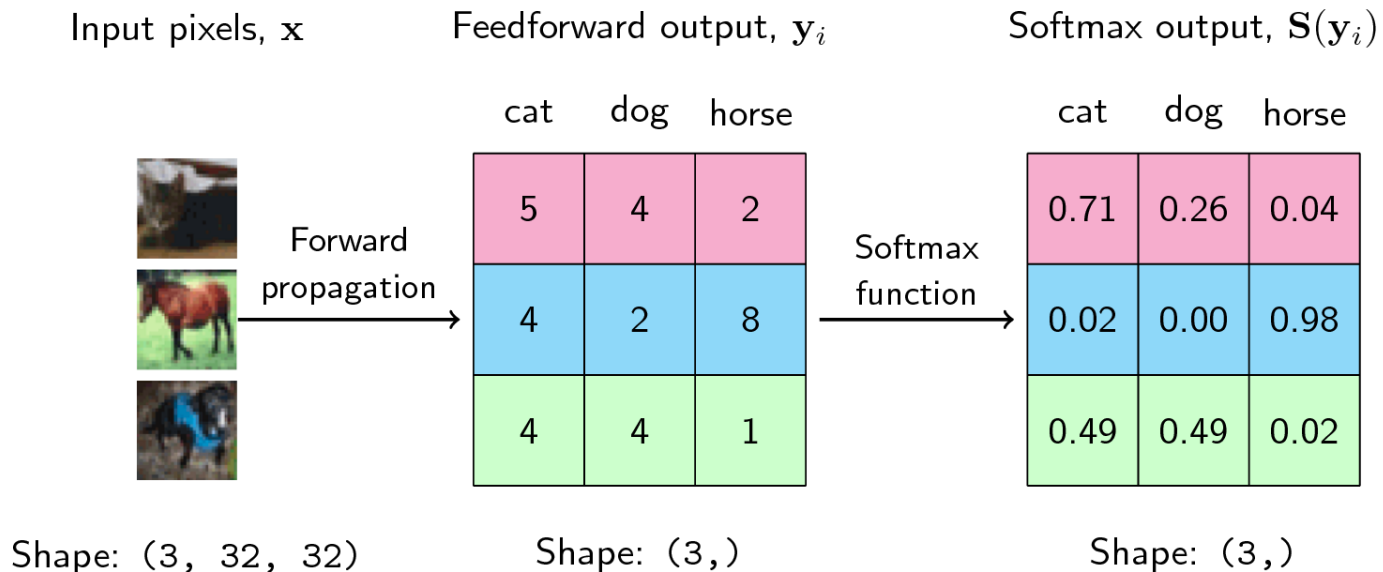
- 범주형 변수 예측 (분류)과 수치 변수 예측 (회귀)에 따라 활성화 함수를 선택한다.

5. Deep Learning

5-5. 소프트맥스 함수 (Softmax Function)

소프트맥스 함수는 출력 레이어에서 사용되며, 다중 클래스 분류 문제에 활용된다.

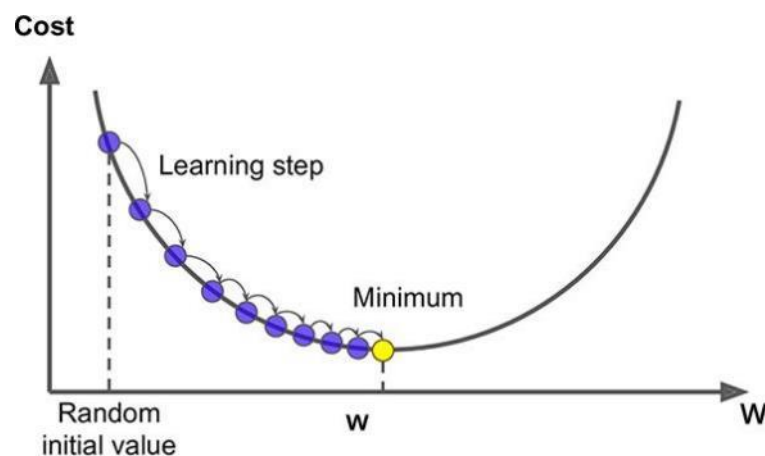
소프트맥스 함수가 반환하는 점수의 합은 1이다. 따라서, 점수가 가장 높은 항목을 예측하고자 하는 클래스로 본다.



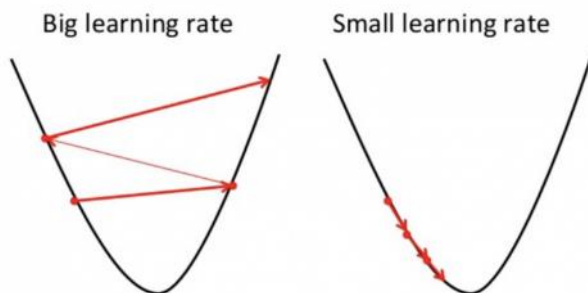
5. Deep Learning

5-6. 경사 하강법 (Gradient Descent)

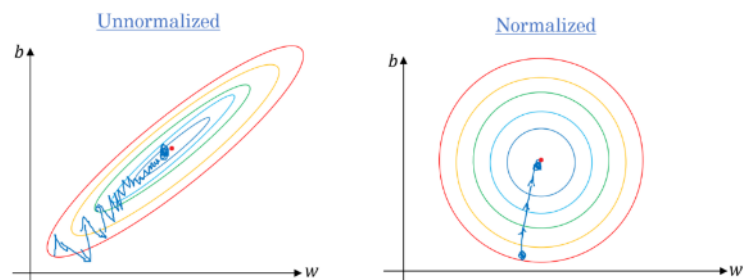
경사 하강법은 비용 함수를 최소화하기 위해 매개 변수(가중치 W + 편향 b)의 값을 반복적으로 조정하는 데 사용되는 최적화 알고리즘이다. 기울기가 0이면 최소값에 도달한다.



학습률 (Learning Rate)



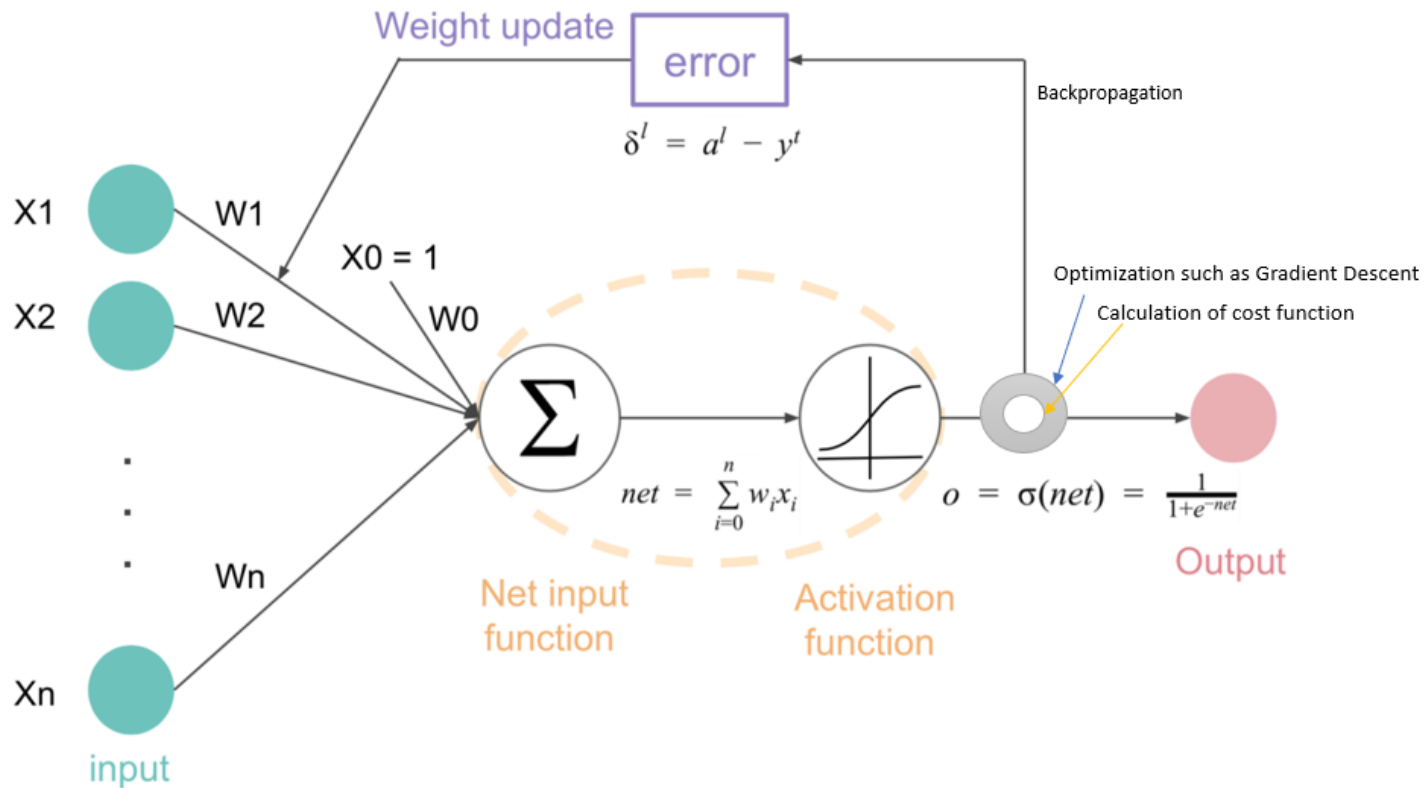
피쳐 스케일링 (Feature Scaling)



5. Deep Learning

5-7. 오차 역전파 (Error Backpropagation)

오차 역전파는 계산이 복잡한 다층 퍼셉트론에서, 비용 함수의 도함수(또는 기울기)를 계산하는 방법이다.
결과 값이 가지는 오차(Error)의 역방향으로 보내며 가중치를 조정한다.



End of Document