



빅데이터분석기사 필기특강 2024

강사 이래중

빅데이터 분석기사 소개

■ 빅데이터 분석기사

빅데이터 이해를 기반으로 빅데이터 분석 기획, 빅데이터 수집·저장·처리, 빅데이터 분석 및 시각화를 수행하는 실무자
#국가기술자격 #빅데이터_분석_전문가 #실무_능력

■ 빅데이터분석기사 출제내용

필기

총 80문항, 객관식 OMR 제출

- 빅데이터 분석기획
- 빅데이터 탐색
- 빅데이터 모델링
- 빅데이터 결과해석

실기

3개 유형, *CBT를 통해 제출(파이썬/R)

- 데이터 전처리
- 머신러닝
- 통계

* CBT: Computer Based Test

빅데이터 분석기사 소개

■ 강의 개요 및 순서

- ① 빅데이터 분석 개요
- ② 빅분기 필기 개념 정리

- ③ 파이썬 문법 정리
- ④ 실기1: 데이터 전처리
- ⑤ 실기2: 머신러닝 모델 - 예측
- ⑥ 실기3: 통계 검정

- ⑦ 빅분기 실기 모의고사, 기출문제 풀이

강의 목차

1. 빅데이터 분석 기획

- 1) 인공지능 - 머신러닝 - 딥러닝
- 2) 지도학습 / 비지도학습
- 3) 강화학습
- 4) 전이학습
- 5) 데이터분석, 머신러닝, 통계 비교

2. 빅데이터 탐색

- 1) 데이터 형태
- 2) 종속/독립변수
- 3) 데이터 종류
- 4) 결측치와 이상치
- 5) 차원축소
- 6) 스케일링/인코딩
- 7) 불균형 데이터
- 8) 상관관계
- 9) 통계기법

3. 빅데이터 모델링

- 1) 모델 분류 - 작업
- 2) 모델 분류 - 알고리즘
- 3) 모델 종류

4. 빅데이터 결과 해석

- 1) k-fold 교차 검증
- 2) 분류 모델 평가
- 3) 회귀 모델 평가
- 4) 과대적합 방지



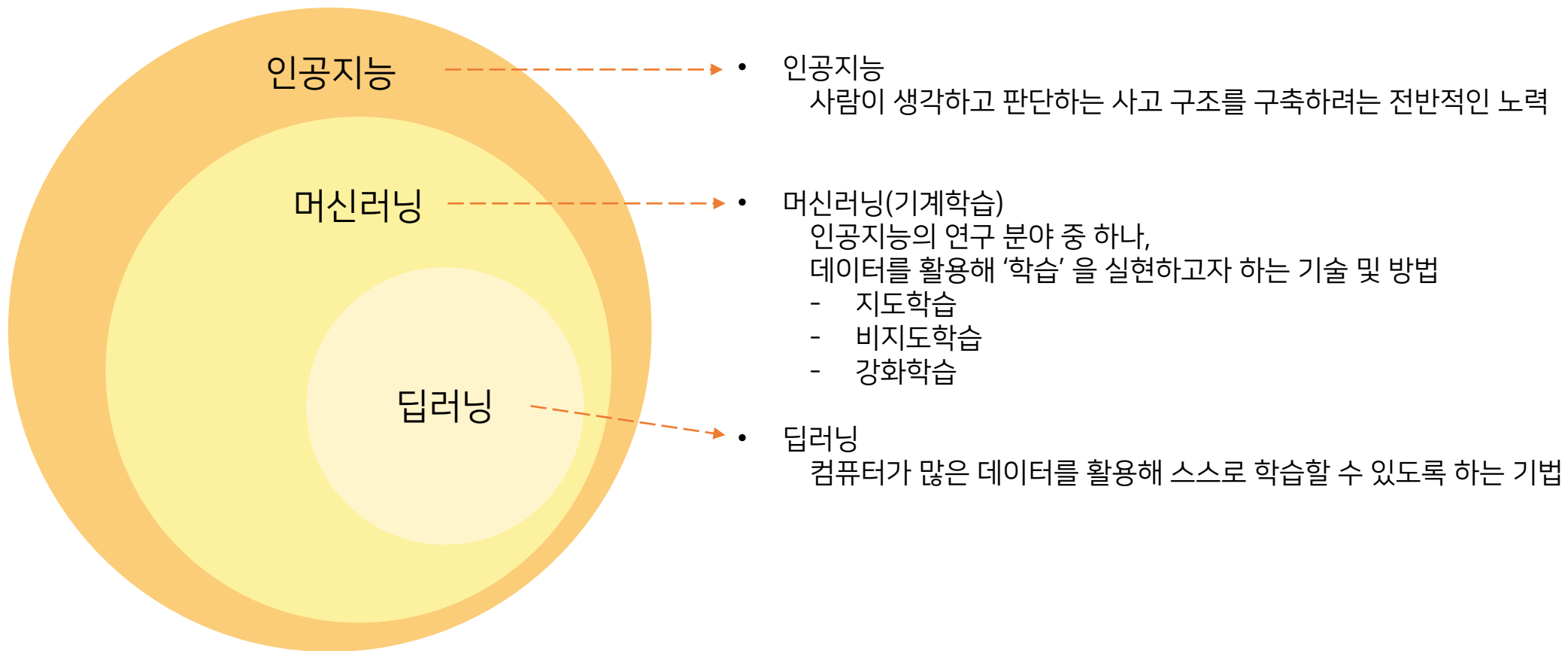
빅데이터분석기사 필기특강

빅데이터 분석 기획

머신러닝 | 지도/비지도학습 | 강화학습 | 전이학습 | 데이터분석, 머신러닝, 통계

1.1. 인공지능 - 머신러닝 - 딥러닝

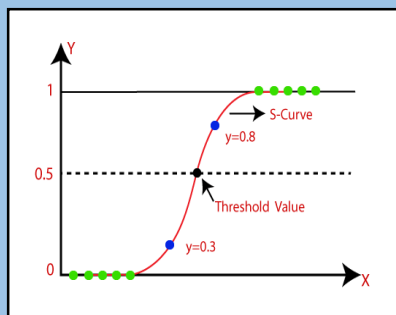
- 인공지능 ⊃ 머신러닝 ⊃ 딥러닝



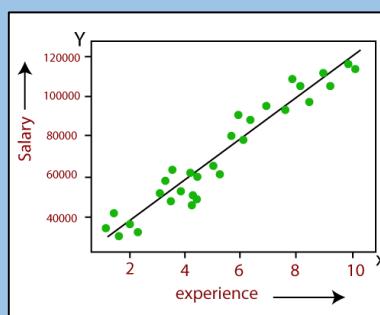
1.2. 지도학습 / 비지도학습

지도학습

분류모형

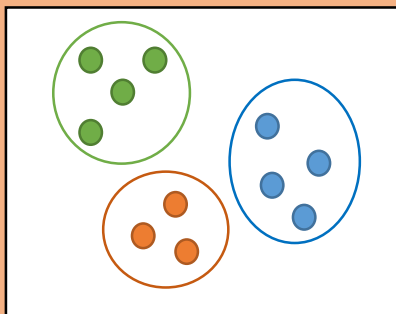


회귀모형

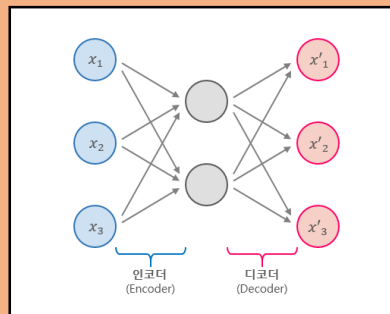


비지도학습 (자율학습)

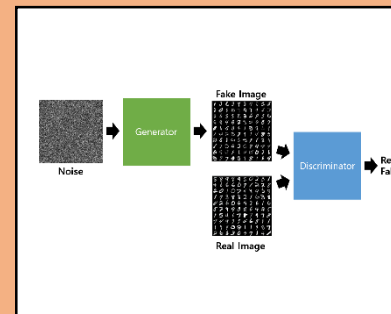
군집분석



오토인코더

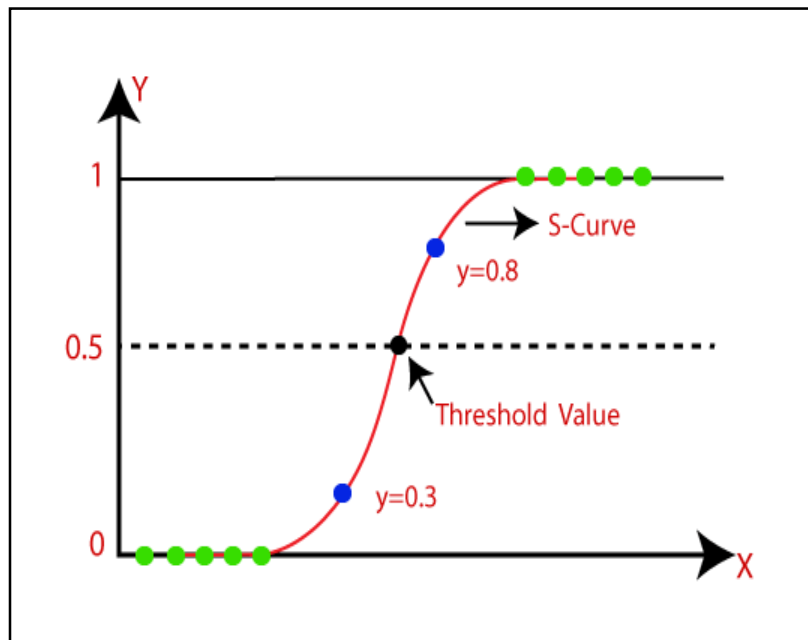


GAN(생성적 적대 신경망)

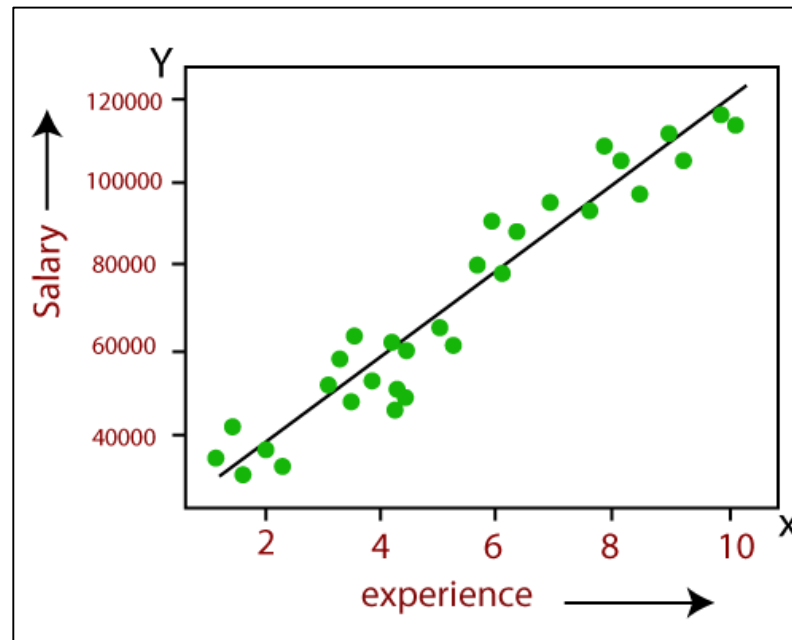


1.2.1. 지도학습

분류모형

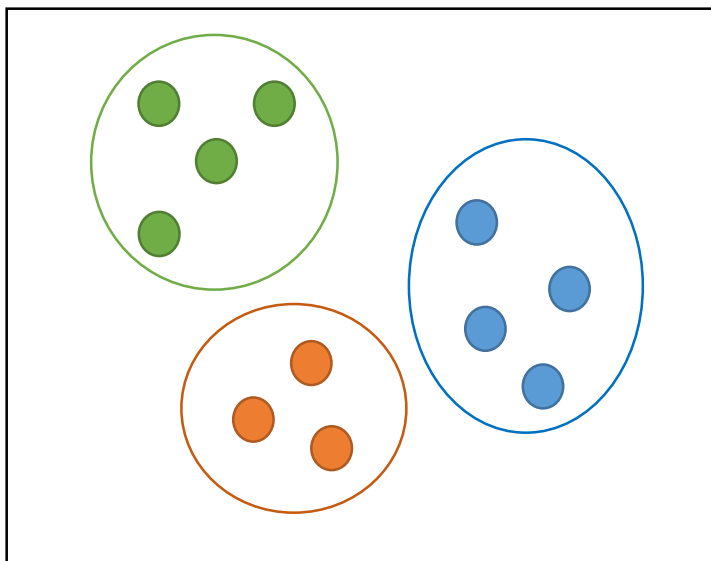


회귀모형

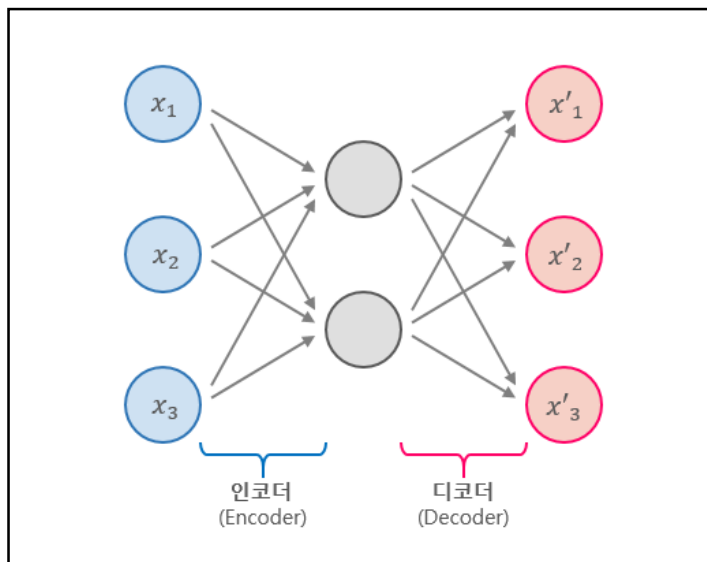


1.2.2. 비지도학습

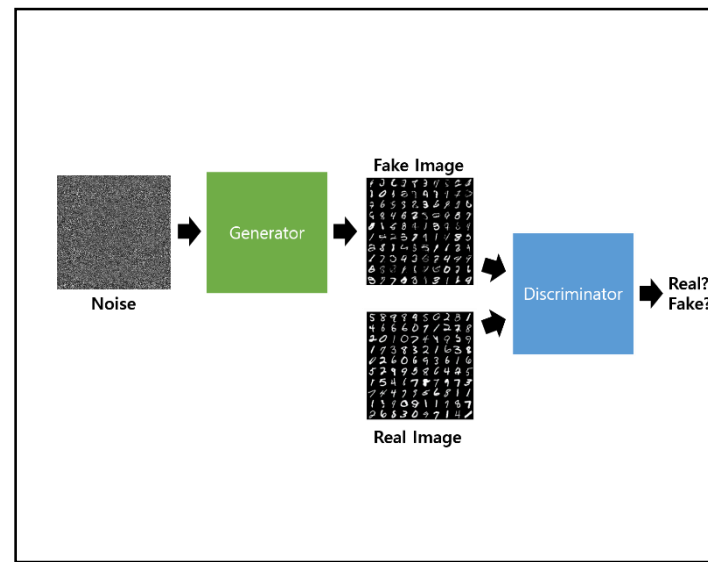
군집분석



오토인코더

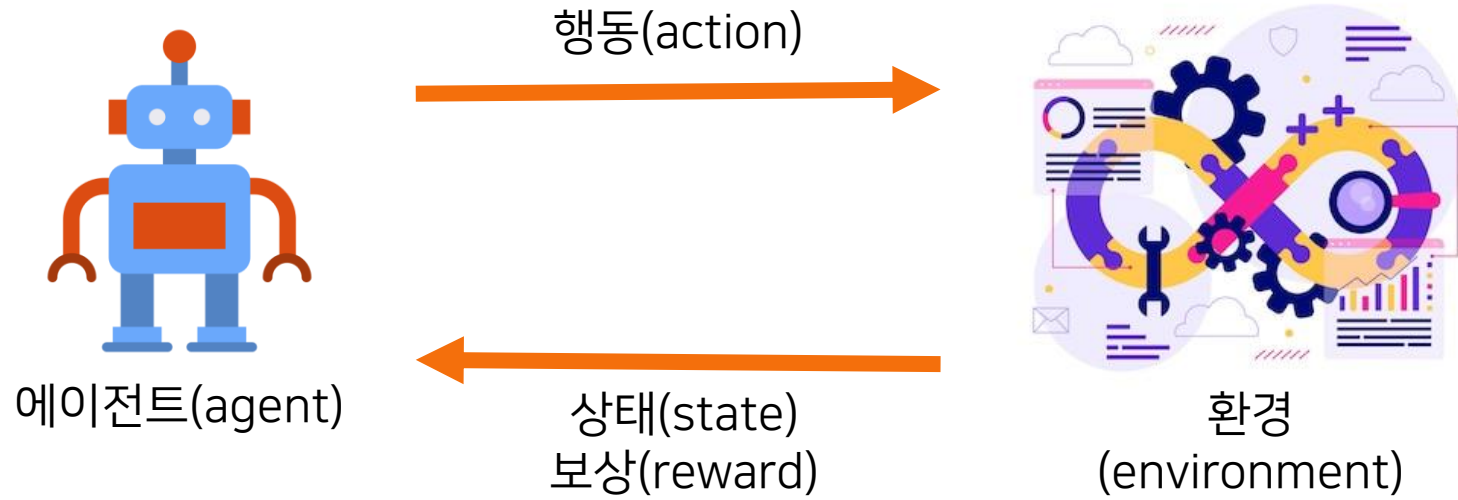


GAN(생성적 적대 신경망)



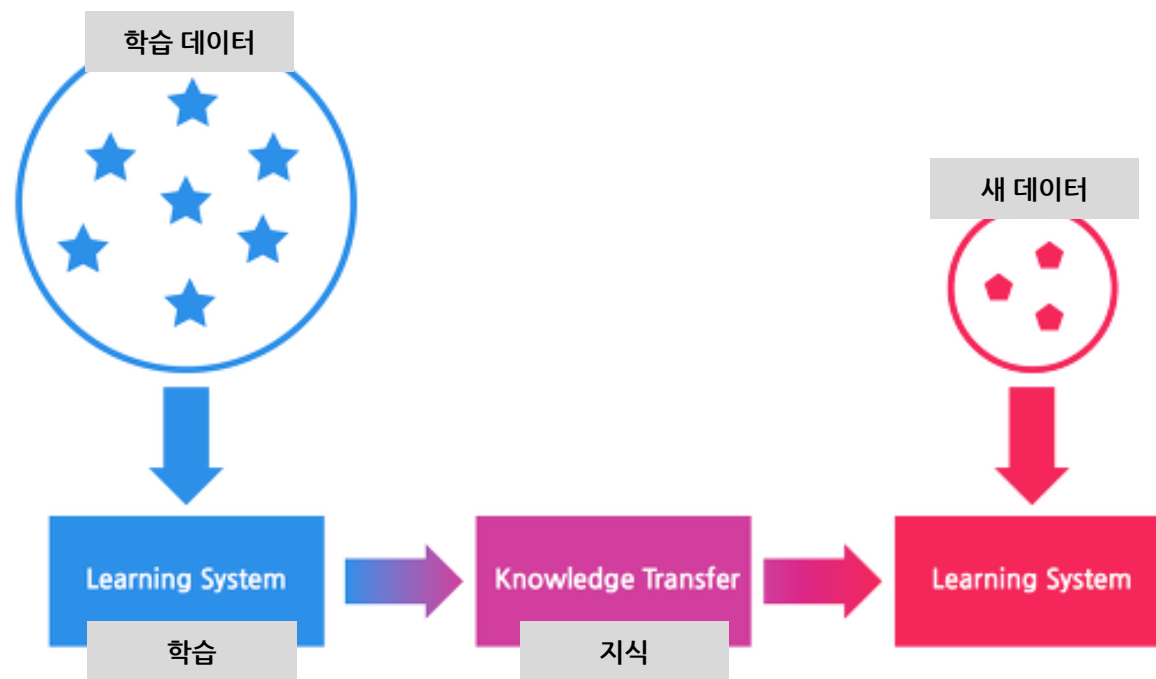
1.3. 강화학습

- 선택 가능한 행동들 중 보상을 최대화하는 행동을 선택하는 방법



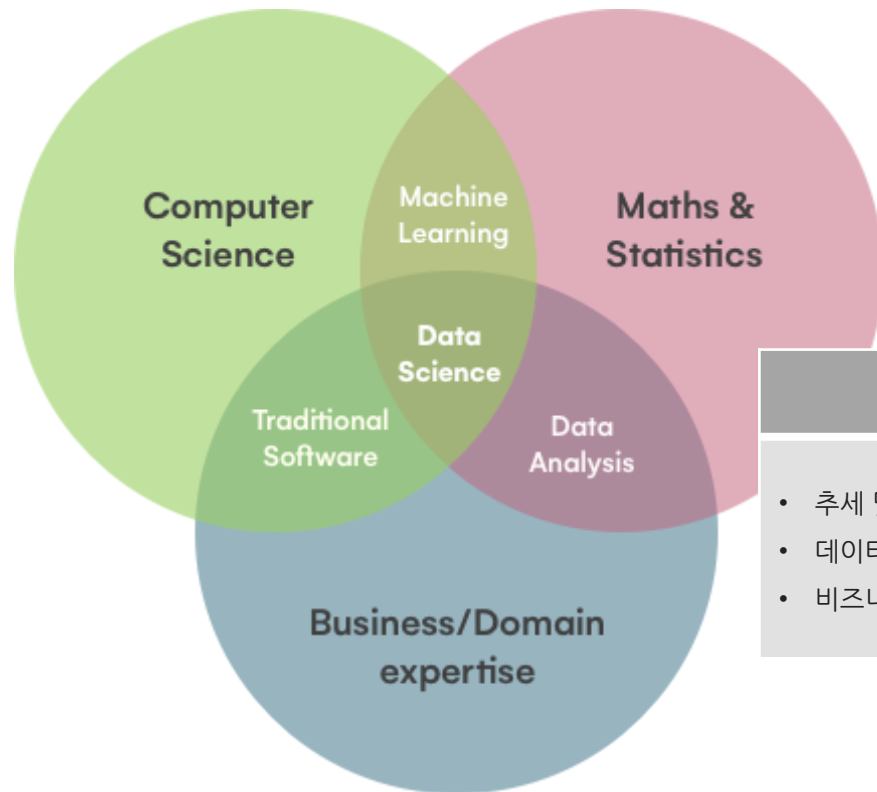
1.4. 전이학습

- 기존에 학습된 모델의 지식을 새 문제에 적용해 학습을 빠르고 효율적으로 수행하는 머신러닝 기법



1.5. 데이터분석 / 머신러닝 / 통계

- 데이터 분석은 데이터 탐색 및 가공을 통하여 의미 있는 정보를 발굴한다.
- 머신 러닝의 역할은 알고리즘을 통해 기계가 의사 결정을 수행하도록 구성한다.
- 통계는 데이터를 이해하고 해석하고 제시하기 위한 기초 지식이다.



데이터 분석	데이터 과학
<ul style="list-style-type: none">• 추세 및 메트릭 탐색• 데이터 시각화• 비즈니스 지식 및 의사 결정 능력	<ul style="list-style-type: none">• 수학적 알고리즘 설계 및 구현• 기계 학습에 대한 지식• 비정형 데이터로 작업하는 경향



빅데이터분석기사 필기특강

빅데이터 탐색

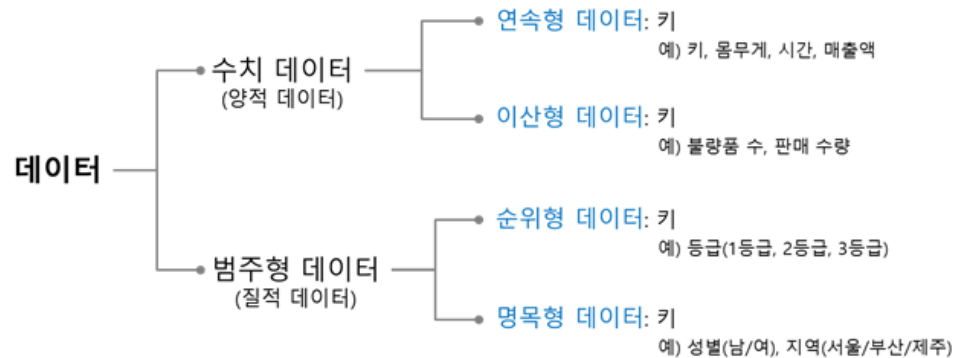
데이터 형태 | 종속/독립변수 | 데이터 종류 | 결측치/이상치 | 차원축소 |
스케일링/인코딩 | 불균형 데이터 | 상관관계 | 통계기법

2.1. 데이터 형태

데이터란?

이론을 세우는 기초가 되는 사실 또는 자료

컴퓨터와 연관되어 프로그램을 운용할 수 있는 형태로 기호화, 수치화한 자료



Wide Format

Team	Points	Assists	Rebounds
A	88	12	22
B	91	17	28
C	99	24	30
D	94	28	31


Long Format

Team	Variable	Value
A	Points	88
A	Assists	12
A	Rebounds	22
B	Points	91
B	Assists	17
B	Rebounds	28
C	Points	99
C	Assists	24
C	Rebounds	30
D	Points	94
D	Assists	28
D	Rebounds	31

2.2. 종속변수 / 독립변수

- 독립 변수 (X)
- 종속 변수 (y)

- Predictor variables(예측변수)
- Input variables(입력변수)
- Independent(독립변수)
- Target variables(타겟변수)
- Output variables(출력변수)
- Dependent variables(종속변수)




id	X_1	X_2	...	X_p	Y
1	x_{11}	x_{12}	...	$x_{1,p}$	y_1
2	x_{21}	x_{22}	...	$x_{2,p}$	y_2
...
n	$x_{n,1}$	$x_{n,2}$...	$x_{n,p}$	y_n

2.3. 데이터 종류

- 명목 변수와 서열 변수는 **범주형** 이고 등간 변수와 비율 변수는 **수치형**이다.
- **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

2.4. 결측치와 이상치

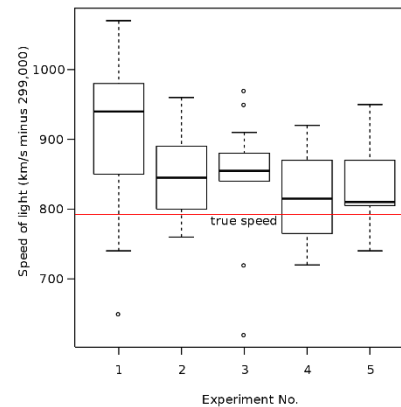
결측치 (Missing Data)

- 결측치는 Missing data, 즉 데이터가 없음을 의미한다.

x1	x2	x3	...	y
1	여	39		yes
2	남			no
3		27		no
4	남	41		yes

이상치 (Outlier)

- 정상의 범주(데이터의 전체적 패턴)에서 벗어난 값을 의미한다.



2.4.1. 결측치

결측치 (Missing Data)

- 결측치는 Missing data, 즉 데이터가 없음을 의미한다.

x1	x2	x3	...	y
1	여	39		yes
2	남			no
3		27		no
4	남	41		yes

결측치 제거

x1	x2	x3	...	y
1	여	39		yes
4	남	41		yes

결측치 대체

x1	x2	x3	...	y
1	여	39		yes
2	남	38		no
3	여	27		no
4	남	41		yes

① 단순 대체법

- 완전 분석
- 평균 대체법
- 회귀 대체법
- 단순확률 대체법
- 최근접 대체법

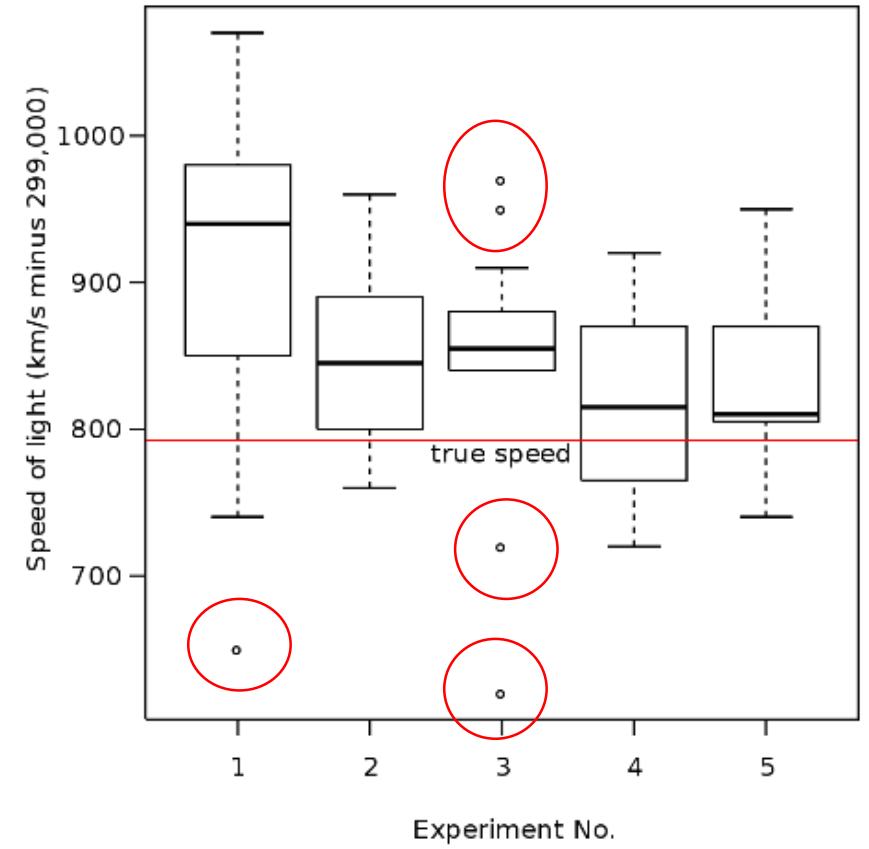
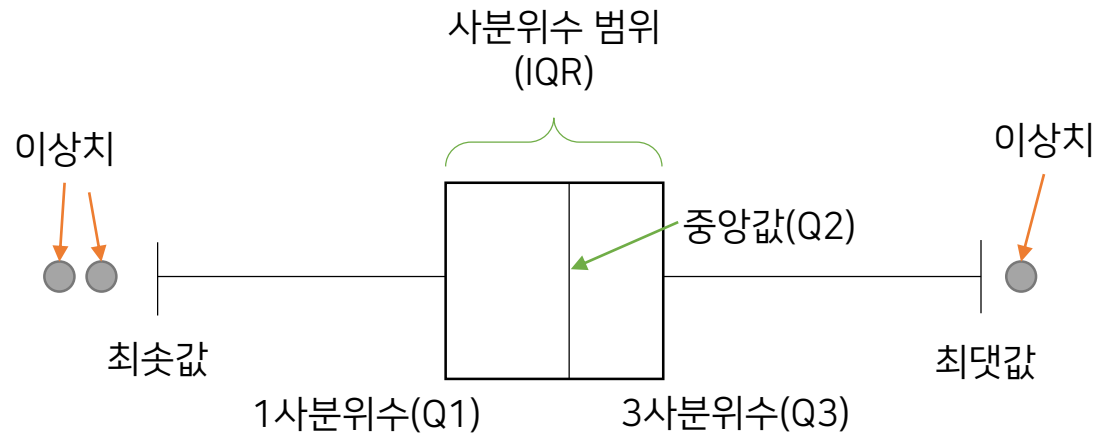
② 다중 대체법

- 대체 → 분석 → 결합

2.4.2. 이상치

이상치 (Outlier)

- 정상의 범주(데이터의 전체적 패턴)에서 벗어난 값을 의미한다.



2.5. 차원 축소

차원 축소

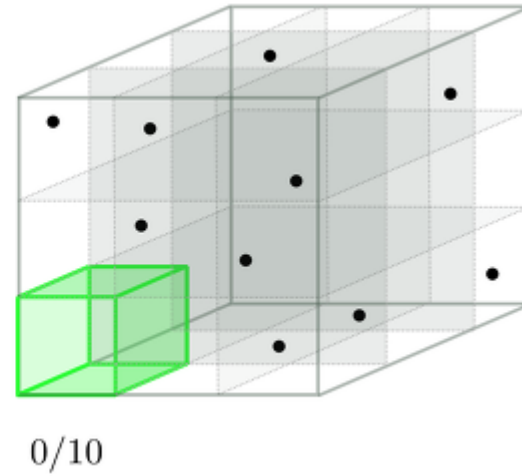
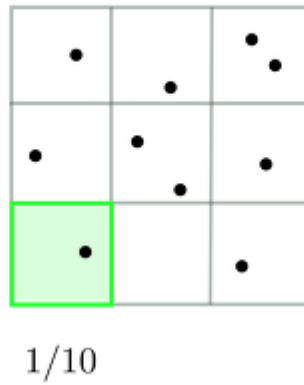
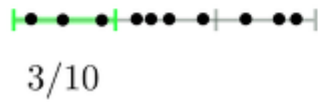
- 어떤 목적에 따라 변수(데이터의 종류)의 양을 줄이는 것

자료의 차원

- 분석하는 데이터의 종류 수

차원 축소의 방법

- ① 요인 분석
- ② 주성분 분석
- ③ 특이값 분해



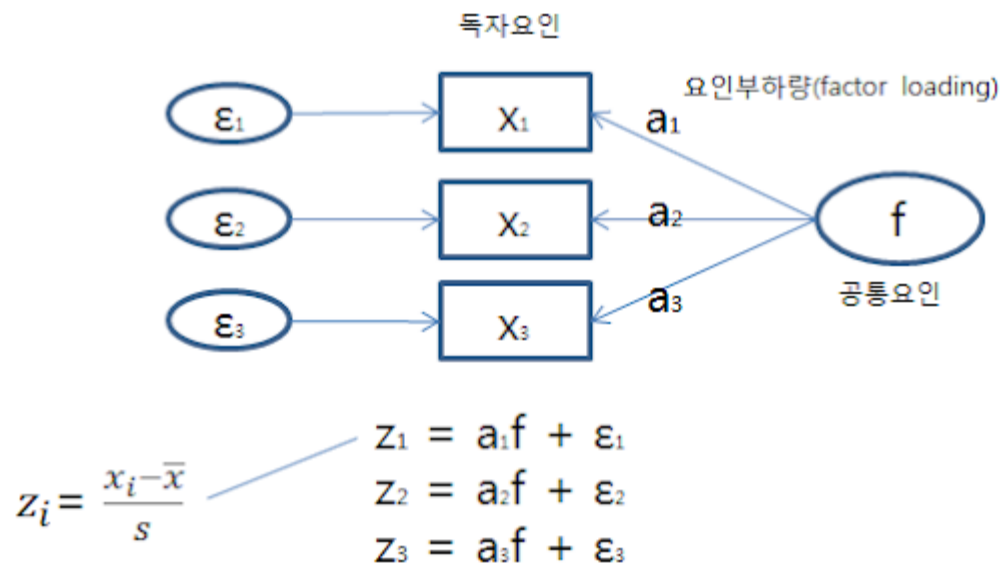
2.5.1. 차원 축소 - ① 요인 분석

요인 분석 (Factor Analysis)

- 변수들 간의 관계(상관관계)를 분석해 공통차원을 축약하는 통계분석 과정
- 주로 기술 통계에 의한 방법을 이용한다.
- 독립변수, 종속변수 개념이 없다.

요인 분석의 목적

- 변수 축소
- 변수 제거
- 변수특성 파악
- 타당성 평가
- 파생변수 생성



2.5.2. 차원 축소 - ② 주성분 분석

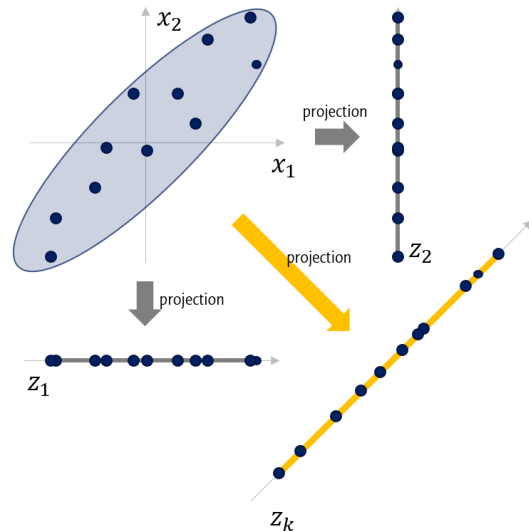
주성분 분석 (PCA: Principal Component Analysis)

- 분포된 데이터들의 특성을 설명할 수 있는 **하나 또는 복수 개의 특징(주성분)**을 찾는 것

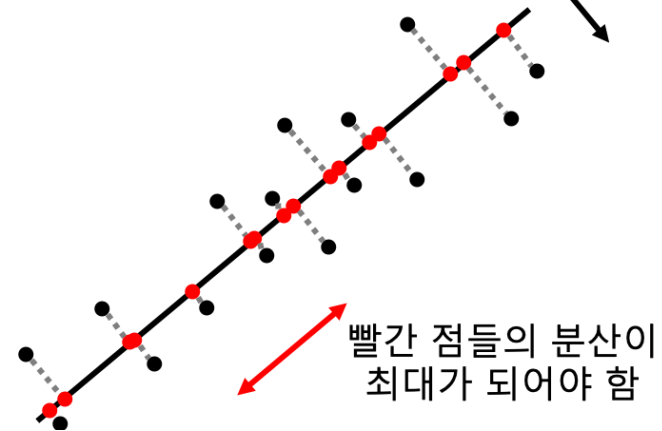
주성분 분석의 특징

- 차원 축소에 폭넓게 사용
- 가장 큰 분산의 방향들이 주요 중심 관심으로 가정한다.
- 본래 변수들의 선형결합으로만 고려한다.
- 스케일에 대한 영향이 크다. (PCA를 하기 위해서는 변수들 간의 스케일링이 필수)

Find space z_k which maximize variance of projected data



검은 점에서 빨간 점으로
투영된 거리가 최소가 되어야
함



2.5.3. 차원 축소 - ③ 특이값 분해

특이값 분해(SVD: Singular Value Decomposition)

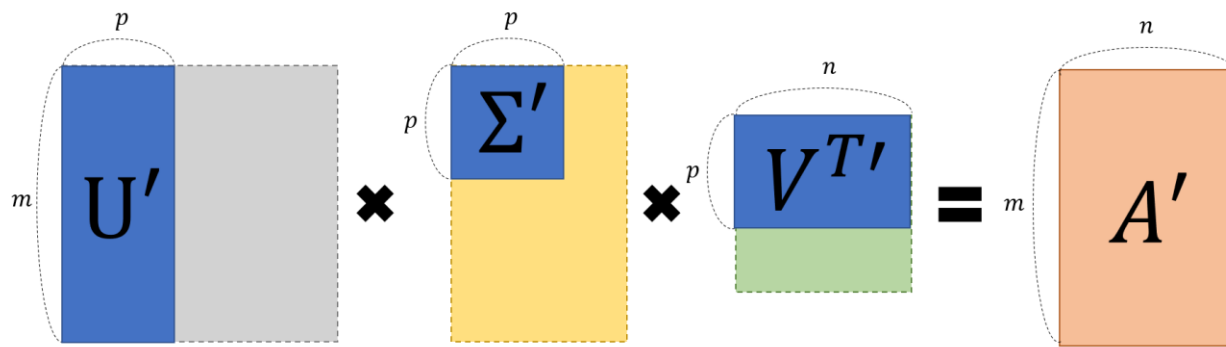
- 데이터 공간을 나타내는 $m \times n$ 크기의 행렬 M 에 대해, 다음과 같이 분해 가능하다.

$$M = U\Sigma V^t$$

- U 는 $m \times m$ 크기의 직교행렬
- Σ 는 $m \times n$ 크기의 대각행렬
- V^t 는 $n \times n$ 크기의 직교행렬

특이값 분해의 차원 축소 원리

- 큰 몇 개의 특이값을 가지고도 충분히 유용한 정보를 유지할 수 있는 차원을 생성할 수 있다.



▶ SVD를 통해 얻어진 일부만으로도 A 와 유사한 정보력을 가지는 A' 행렬을 생성할 수 있다.

2.6. 스케일링 / 인코딩

- 데이터 분석의 **안정적인 결과와 성능 향상**을 위해서 **주어진 데이터를 분석에 적합하게 가공하는 작업**이다.
- 대표적인 작업으로는 필터링, 클리닝, 결측치 처리, 이상치 처리, 데이터 형태 변경 등이 있다.
- 범주형 데이터 인코딩 : 레이블 인코딩 (Label Encoding) & 원핫 인코딩 (One-hot Encoding)
- 수치형 데이터 스케일링 : 표준화 (Normalization) & 정규화 (Standardization)
- Filtering / Cleaning / Missing Value / Outlier
- Data Shape : Long Data, Wide Data

One-hot Encoding			
Color	Red	Yellow	Green
Red	1	0	0
Red	1	0	0
Yellow	0	1	0
Green	0	0	1
Yellow	0	0	1

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

2.7. 불균형 데이터

- 각 클래스가 갖고 있는 데이터의 양에 차이가 큰 경우, **클래스 불균형**이 있다고 한다.

클래스 불균형



불균형 데이터 처리

- ① 가중치 균형방법
- ② 언더샘플링
- ③ 오버샘플링

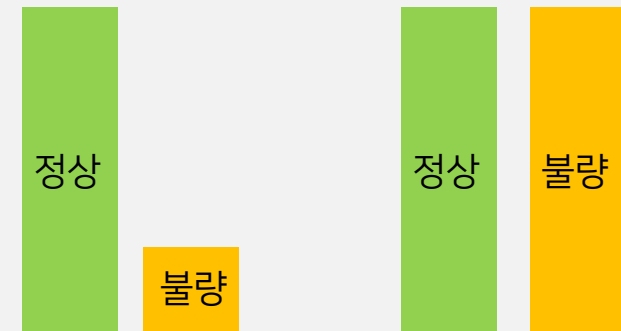
가중치 균형방법

- 각 클래스별 특정 비율로 가중치를 주어 분석하는 방식

언더샘플링



오버샘플링



2.8. 상관관계

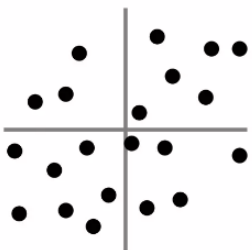
- 두 변수 간에 어떤 선형적 관계를 갖고 있는지 분석하는 방법
- 두 변수 간의 관계의 강도를 상관관계(correlation)이라고 한다.

상관분석 방법

① 피어슨 상관계수

① 스피어만 상관계수

상관관계 약함



상관관계 있음

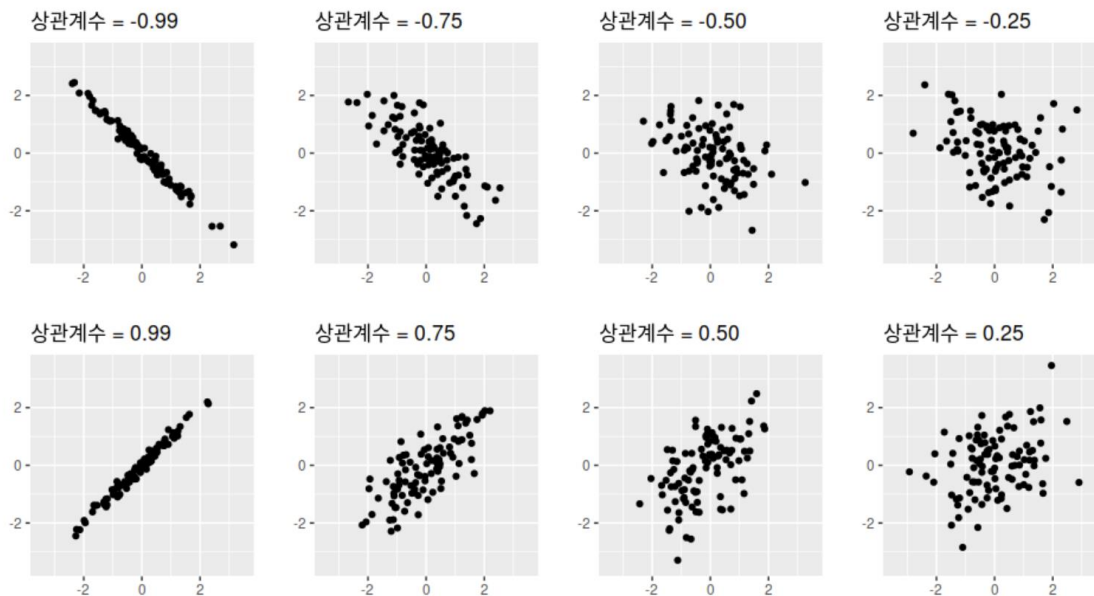
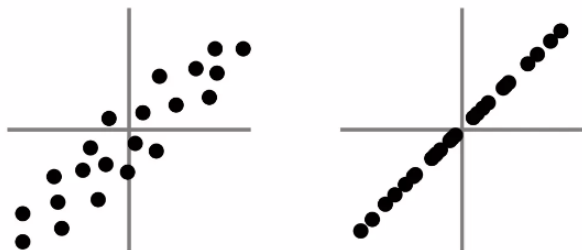


Figure 2.9: 상관계수가 다른 데이터를 나타내는 예제.

2.9. 통계 기법

모집단

- 연구, 실험의 결과가 일반화된 큰 집단, 정보를 얻고자 하는 관심 대상의 전체집합

표본

- 여러 자료를 포함하는 모집단 속에서 그 일부를 끄집어 내어 조사한 결과로 원래 집단의 성질을 추측할 수 있는 자료

표본추출

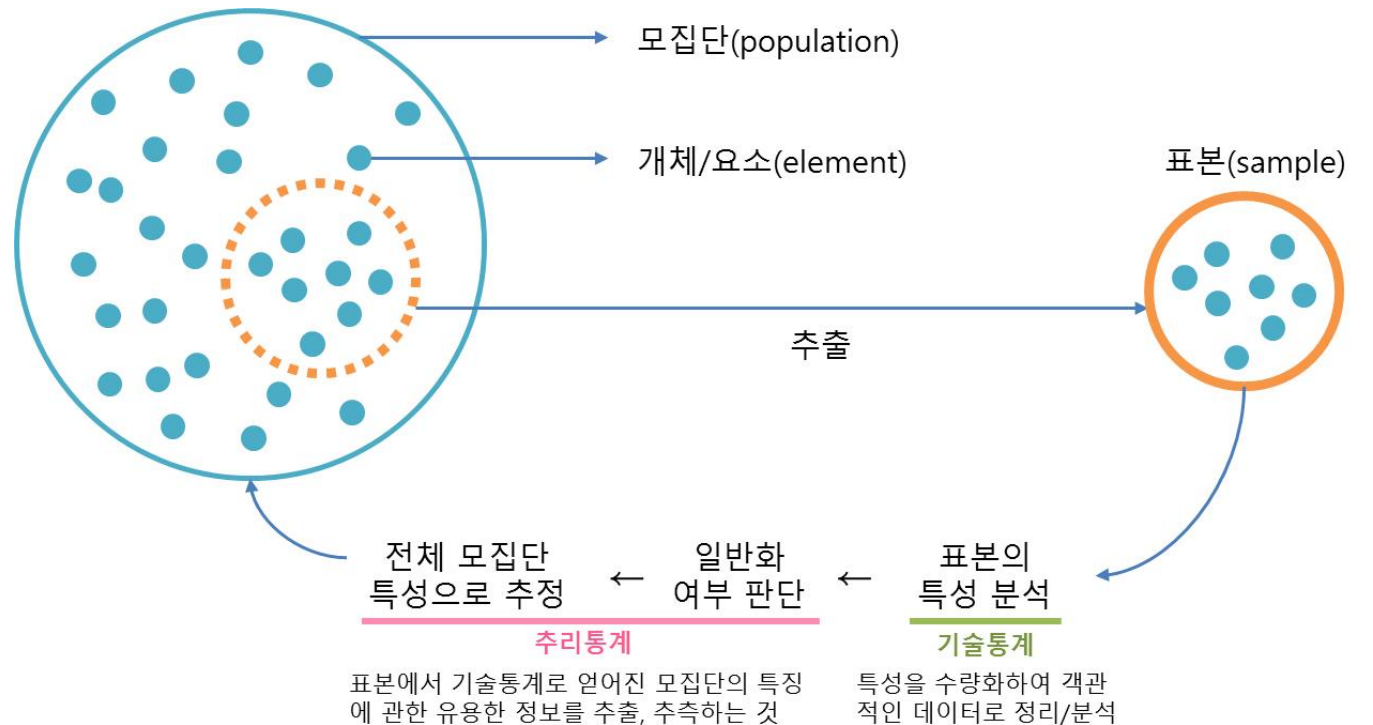
- 모집단으로부터 표본을 선택하는 행위(과정)

기술통계

- 평균(산술, 기하, 조화), 분산, 표준편차, 중앙값

추론통계

- 추정, 가설검정, t-test



2.9.1. 기술 통계

평균(Mean)

- ① 산술평균 (Arithmetic Mean)
 - 모든 자료를 합한 후 전체 자료수로 나눠 계산하는 일반적인 평균
 - 모평균 : 모집단 전체 자료의 산술 평균
 - 표본평균 : 모집단의 부분집합인 추출된 표본 전체의 산술평균
- ② 기하평균 (Geometric Mean)
 - N개의 자료에 대해 관측치를 곱한 후 n 제곱근으로 표현
 - ex. 다기간의 수익률에 대한 평균 수익률, 평균물가상승률
- ③ 조화평균 (Harmonic Mean)
 - 각 요소의 역수의 산술평균을 구한 후 다시 역수를 취하는 형태로 표현
 - ex. 변화율 등의 평균

분산(Variance)

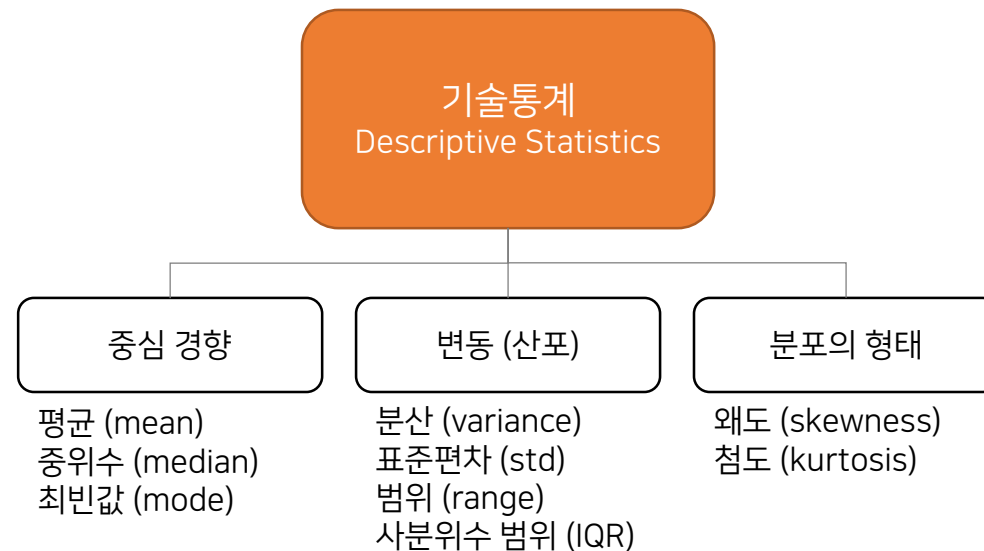
- 평균을 중심으로 밀집되거나 퍼짐 정도를 나타내는 척도

표준편차(Standard Deviation)

- 분산에 제곱근을 취한 척도
- 분산으로 얻은 수치가 해석하기 곤란하다는 단점을 보완해 만든 척도

중앙값(Median)

- 자료를 크기 순으로 나열할 때 가운데에 위치한 값



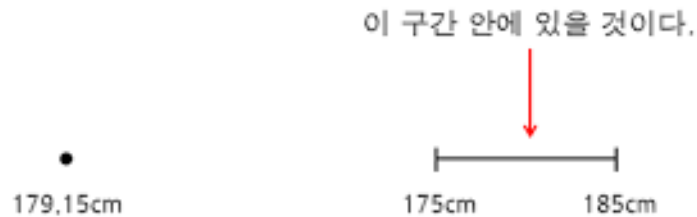
2.9.2. 추론 통계 - 추정

점추정

- 모수에 대한 모평균이나 모표준편차 등과 같은 추정치를 이에 대응하는 통계량으로 추정하는 것
- 한 개의 값을 이용해 모수 추정

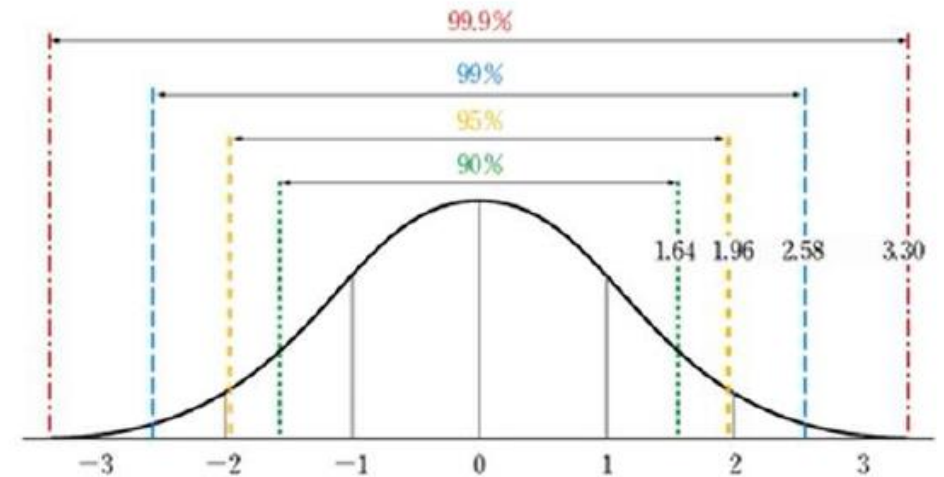
구간추정

- 점추정에 오차(error)의 개념을 도입해 모수가 포함되는 확률변수구간을 어떤 신뢰성 아래 추정하는 것
- 모수가 있을 것으로 예상되는 구간을 정해 그 구간에 실제 모수가 있다고 예상되는 확률을 기반으로 수행
- 구간의 값을 이용해 모수 추정



신뢰구간

- 모수 추정치 주변에 구간을 형성해 모수의 값이 해당 구간에 속할 확률



모집단 평균에 대한 신뢰구간

$$\bar{X} - z \cdot SE \leq \mu \leq \bar{X} + z \cdot SE$$

2.9.2. 추론 통계 - 가설검정

가설검정

- 모집단에 대해 어떤 가설을 설정하고 그 모집단으로부터 추출된 표본을 분석함으로써 그 가설이 틀리는지 맞는지 타당성 여부를 결정하는 통계적 기법

검정통계량

귀무가설이 참이라는 가정 아래 얻은 표본통계량

귀무가설 (H_0)

현재 통념적으로 믿어지고 있는 모수에 대한 주장 또는 원래의 기준이 되는 가설

대립가설 (H_1)

연구자가 모수에 대해 새로운 통계적 입증을 이루어 내고자 하는 가설

제 1종 오류

귀무가설이 참일 때 귀무가설을 기각하도록 결정하는 오류 (무죄인데 유죄라고 할 오류)

제 2종 오류

귀무가설이 거짓인데 귀무가설을 채택할 오류 (유죄인데 무죄라고 할 오류)

유의수준 α

가설검정의 결과로 가설의 채택여부를 결정하게 될 때, 제1종 오류를 범할 확률의 최대 허용한계를 의미한다.

유의확률 p-value

관찰된 데이터의 검정통계량이 귀무가설을 지지하는 정도를 확률로 표현한 것

검정결과 실제상황	H_0	H_1
귀무가설 H_0 채택	성공	2종 오류
귀무가설 H_0 기각	1종 오류	성공

$p\text{-value} < \alpha$: 귀무가설을 기각

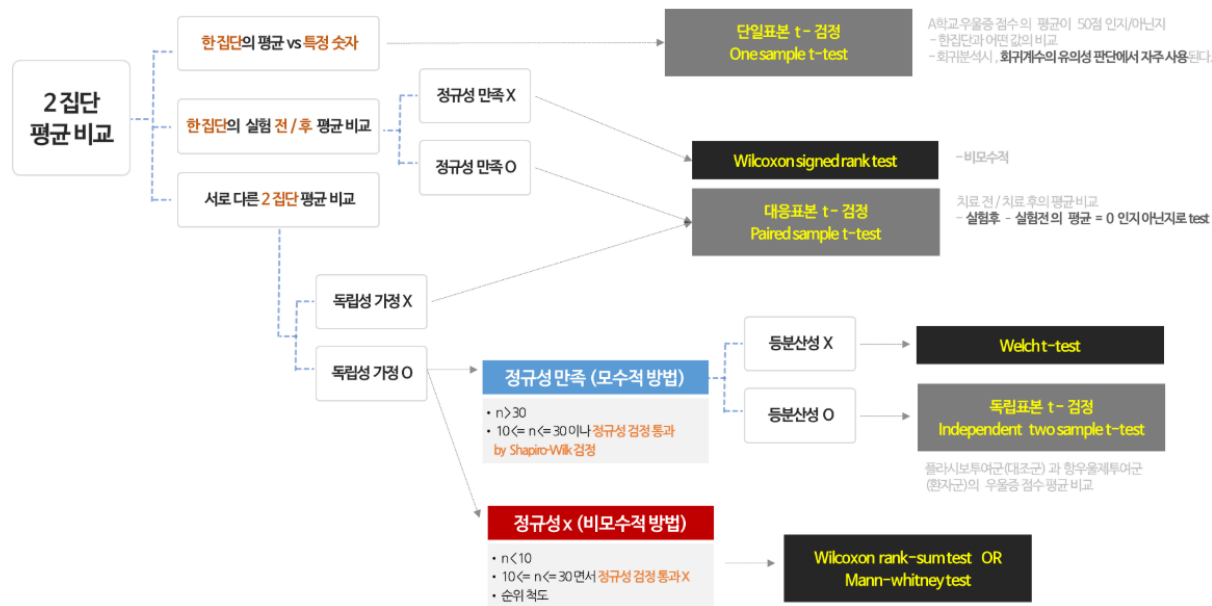
$p\text{-value} > \alpha$: 귀무가설을 채택

2.9.2. 추론 통계 - t-test

t-검정

두 집단 간의 평균을 비교하는 모수적 통계방법으로, 표본이 **정규성 / 등분산성 / 독립성** 등을 만족하면 적용할 수 있다.

- ① 단일표본의 평균 검정
- ② 두 독립표본의 평균차이 검정
 - 서로 다른 두 모집단으로부터 데이터를 추출하는 경우
- ③ 대응표본의 평균차이 검정
 - 하나의 모집단으로부터 데이터를 반복 추출하는 경우 (한 집단 내 비교)



2.9.2. 추론 통계 - ANOVA

ANOVA (ANalysis Of VAriance)

분산을 활용한 검정

- ① One-way ANOVA (일원 배치 분산분석)
 - 요인이 1개이고, 그룹이 3개 이상인 경우 각 그룹 간 유의한 차이가 있는지 검정
 - ② Repeated Measures ANOVA (반복측정 분산분석)
 - 요인이 1개이고, 2번 이상 반복 측정된 샘플의 변화량에 유의미한 차이가 있는지 검정
 - ③ Two-way ANOVA (이원 배치 분산분석)
 - 요인이 2개인 경우 사용
 - ④ Two-way Repeated Measures ANOVA (이원 반복측정 분산분석)
 - 요인이 2개이고 2번 이상 반복 측정된 샘플의 변화량에 유의미한 차이가 있는지 검정
- 사후검정 : 어떤 그룹 간에 차이가 있는지 확인하는 과정
 - Duncan Test
 - Tukey's Test



2.9.3. 확률

확률의 개념

- 통계적 현상: 불확정 현상을 반복하여 관찰하거나, 집단 안에서 대량으로 관찰하여 고유의 법칙성을 찾아내는 것이 가능한 현상
- 확률 실험 : 같은 조건 아래에서 반복할 수 있다.

확률

- 통계적 현상의 확실함의 정도를 나타내는 척도
- 랜덤 시행에서 어떠한 사건이 일어날 정도를 나타내는 사건에 할당된 수
- 관측값 또는 관측 구간이 주어진 확률분포 안에서 얼마만큼 나타날 수 있는가에 대한 값

① 수학적 확률

표본공간 S 의 각 사건이 일어날 가능성이 동등할 때, 사건 A 에 대하여 $n(A)/n(S)$ 를 사건 A 의 수학적 확률이라고 한다.

$$P(A) = \frac{n(A)}{n(S)}$$

② 통계적 확률

사건이 일어나는 확률을 상대도수에 의해 추정한다.

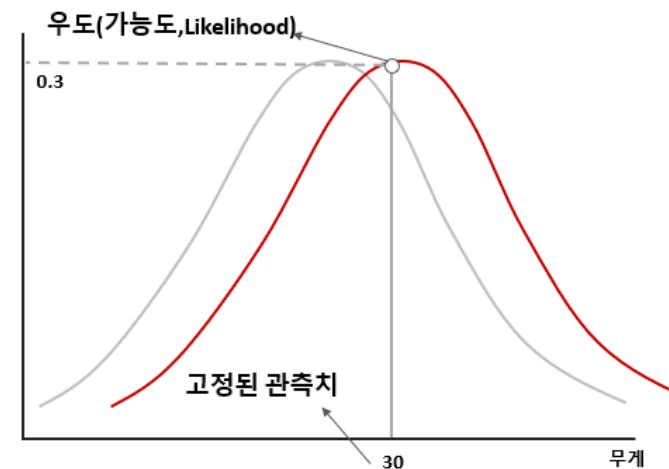
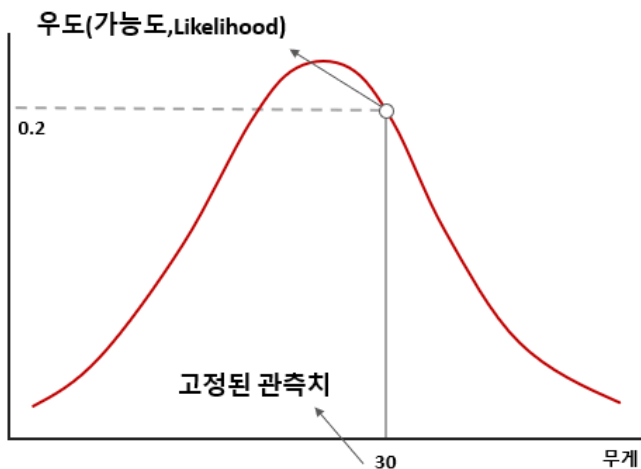
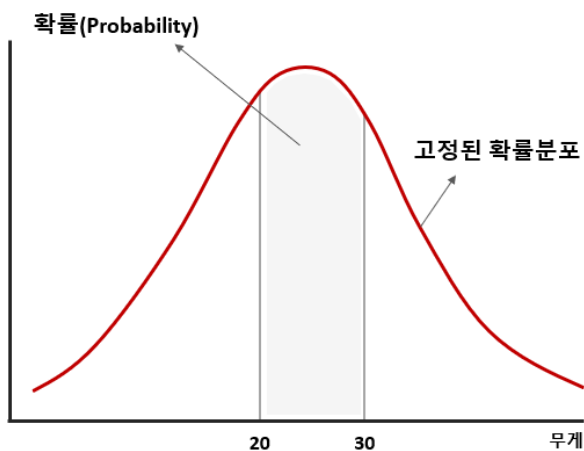
$$p = \frac{r_n}{n}$$

2.9.4. 확률과 우도

우도 (가능도, Likelihood)

어떤 특정한 값을 관측할 때, 이 관측치가 어떠한 확률분포에서 나왔는가에 관한 값

- 확률(Probability)
고정된 확률분포에서 어떠한 관측값이 나타나는지에 대한 확률
- 우도(가능도, Likelihood)
고정된 관측값이 어떠한 확률분포에서 어느정도의 확률로 나타나는지에 대한 확률

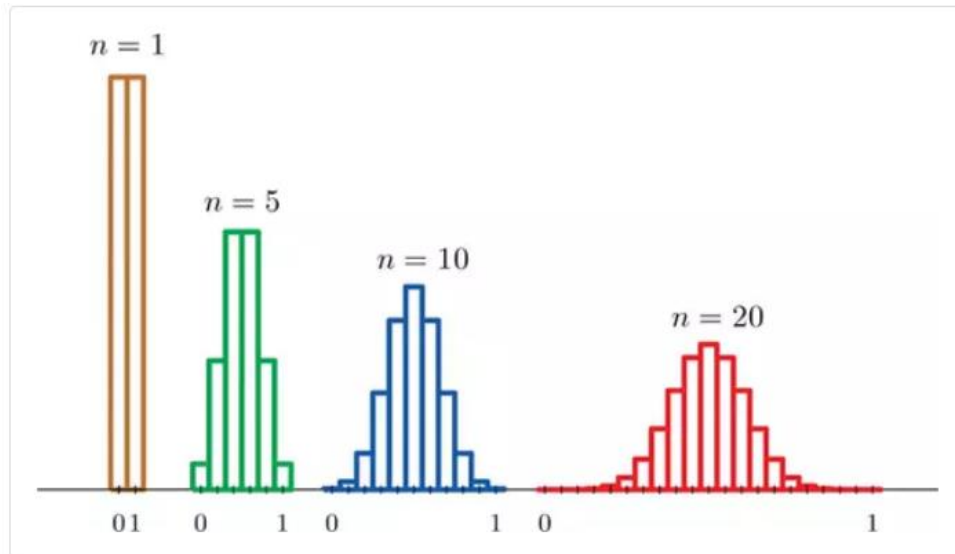


2.9.5. 중심극한정리

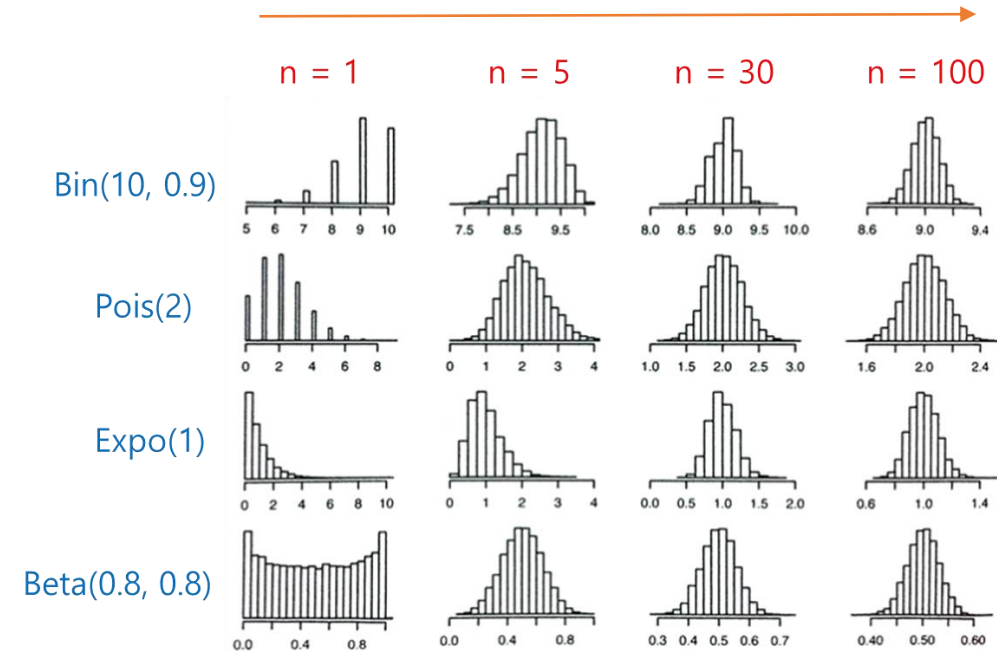
중심극한정리 (Central Limit Theorem)

동일한 확률분포를 가진 독립 확률변수 n 개의 평균의 분포는 n 이 적당히 크다면 정규분포에 가까워진다는 정리

- 린데베르그-레비 중심극한정리



표본의 크기





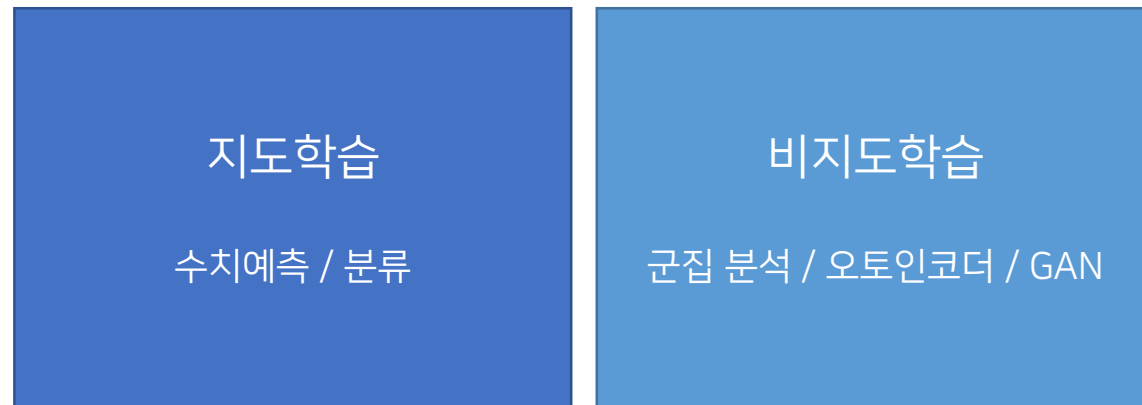
빅데이터분석기사 필기특강

빅데이터 모델링

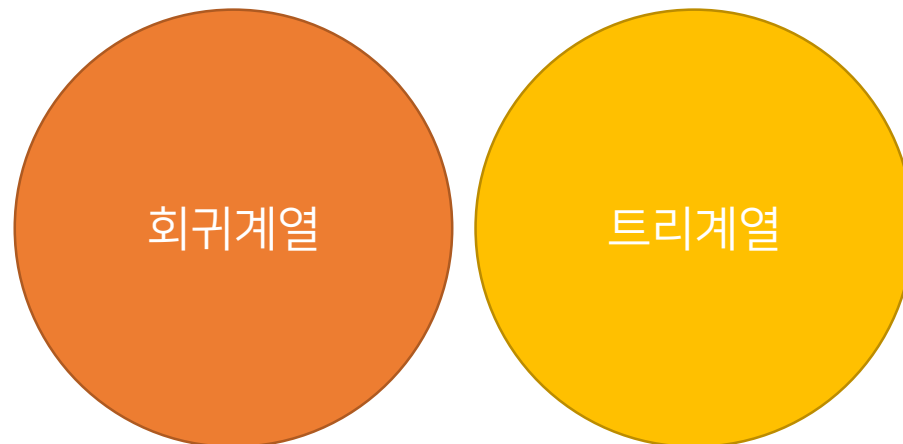
모델 분류 | 모델 종류

3.1. 모델 분류

① 하고자 하는 **작업**에 따라 모델 분류하기



② **알고리즘**에 따라 모델 분류하기



3.1. 모델 분류 - 지도학습/비지도학습

지도학습 수치예측 / 분류	분류 <ul style="list-style-type: none">- 의사결정트리- 랜덤포레스트- 인공신경망- 서포트벡터머신- 로지스틱 회귀분석	회귀(예측) <ul style="list-style-type: none">- 의사결정트리- 선형회귀분석- 다중회귀분석
비지도학습 군집 분석 / 오토인코더 / GAN	<ul style="list-style-type: none">- 군집분석- 연관성분석- 인공신경망- 오토인코더	

3.1. 모델 분류 - 회귀/트리

회귀 계열

특정 변수가 다른 변수에 어떤 영향을 미치는지를 수학적 모형으로 설명, 예측하는 기법

- 선형회귀분석
- 로지스틱 회귀분석

트리 계열

의사결정 규칙을 나무 모양으로 나타내어 전체 자료를 몇 개의 소집단으로 분류하거나 예측하는 기법

- 분류나무
- 회귀나무
- 대표 알고리즘 : CART, C4.5/C5.0, CHAID, 랜덤 포레스트

3.2. 모델의 종류

- 회귀분석
- 로지스틱 회귀분석
- SVM
- 의사결정나무 - 불순도(지니, 엔트로피), 인포메이션게인, 오버피팅 방지
- 앙상블 - 부트스트래핑, 배깅, 부스팅, 보팅, 스태킹
- 인공신경망

3.2.1. 회귀분석 (Regression)

- 특정 변수가 다른 변수에 어떤 영향을 미치는지를 수학적 모형으로 설명, 예측하는 기법
- 독립변수로 종속변수를 예측하는 기법

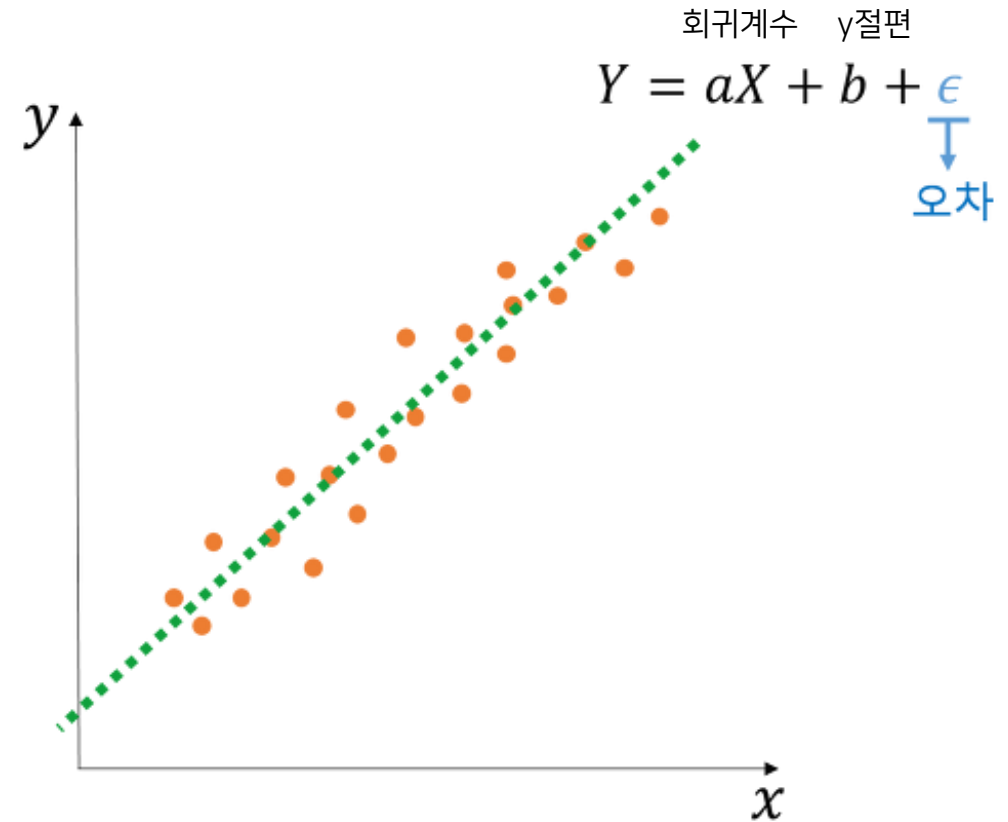
① 단순 선형회귀분석

$$y = ax + b$$

② 다중 선형회귀분석

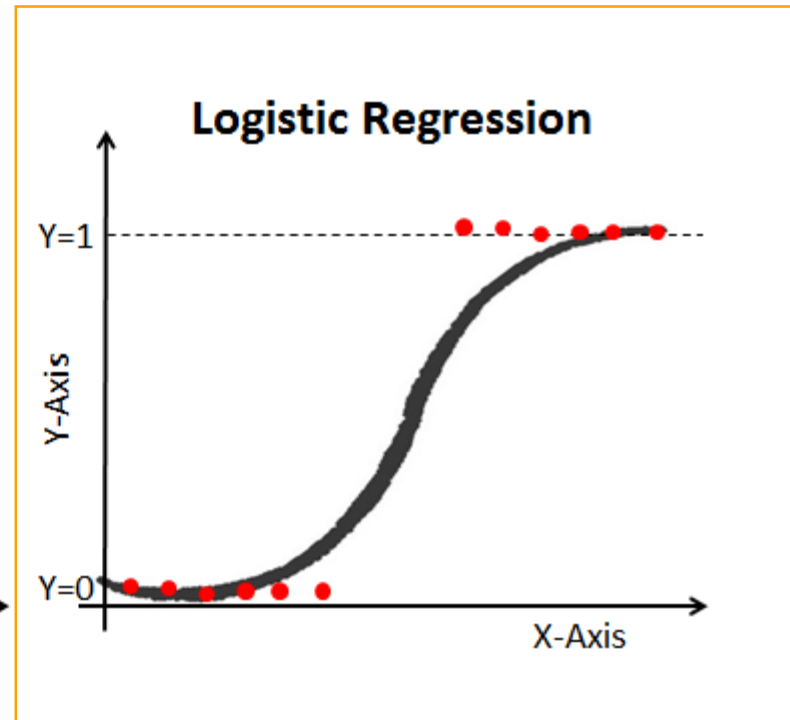
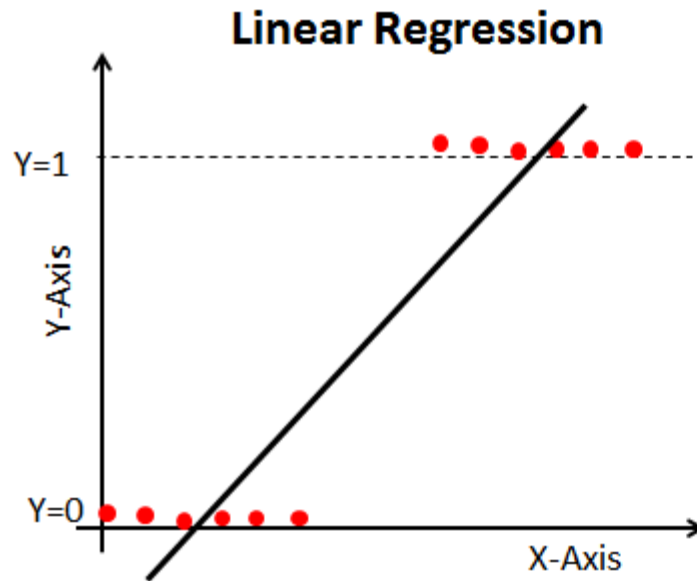
$$y = ax_1 + bx_2 + \dots + c$$

③ 일반화 선형모형 (GLM: Generalized Linear Model)



3.2.2. 로지스틱 회귀분석 (LogisticRegression)

- 종속변수가 연속형이 아닌 범주형으로 입력데이터가 주어졌을 때 특정 분류로 결과가 나타나는 확률 모델
- ① 단순 로지스틱 회귀분석
종속변수가 이항형 문제(범주의 개수가 2개인 경우)인 회귀분석
- ② 다중 로지스틱 회귀분석
종속변수가 두 개 이상의 범주를 가지게 될 경우의 회귀분석

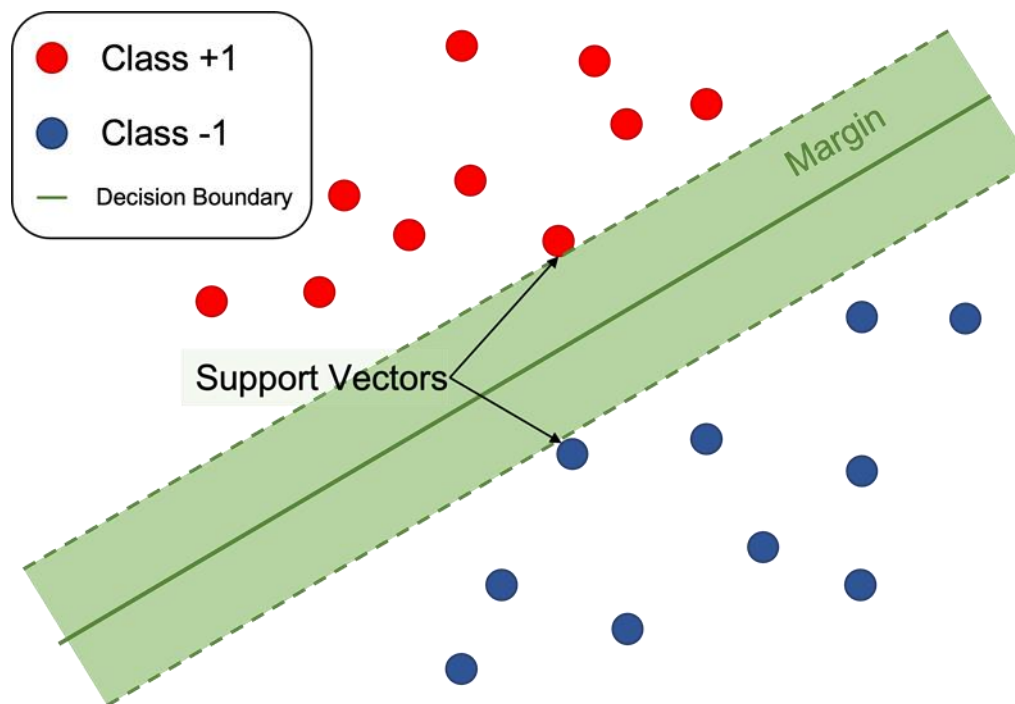


3.2.3. SVM(Support Vector Machine)

- 두 클래스 중 어느 하나에 속한 데이터의 집합이 주어졌을 때, 주어진 데이터 집합을 바탕으로, 새로운 데이터가 어느 카테고리에 속할지 판단하는 비확률적 이진 선형 분류 모델을 만드는 기법
- 가장 큰 폭(마진)을 가진 경계를 찾는 알고리즘

SVM의 특징

- ① 여백(마진) 최대화로 일반화 능력의 극대화 추구
- ② 선형 분류와 더불어 비선형 분류에서도 사용 가능



3.2.4. 의사결정나무 (Decision Tree)

- 의사결정 규칙을 **나무 모양**으로 나타내어 전체 자료를 몇 개의 소집단으로 분류 또는 예측하는 기법

① 분류나무

이산형 목표변수 : 목표변수 범주에 속하는 빈도 기반 입력 데이터가 분류되는 클래스

② 회귀나무

연속형 목표변수 : 목표변수 평균/표준편차 기반 예측된 결과로 특정 의미를 지니는 실수값 출력

불순도 함수

① 카이제곱 통계량

② 지니 지수 (Gini index)

특정 집합에서 한 항목을 뽑아 무작위로 라벨 추정 시 틀릴 확률

③ 엔트로피 지수 (Entropy index)

무질서 정도에 대한 측도

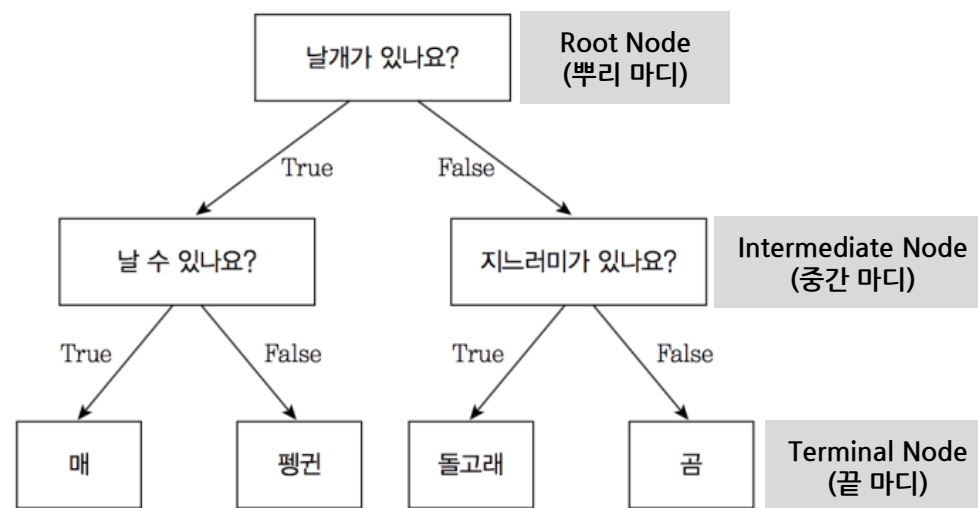
정보 획득 (Information Gain)

- 정보이론(Information Theory)에서 **순도가 증가하고 불확실성이 감소하는 것**
- 현재 노드의 불순도와 자식노드의 불순도의 차이

의사결정나무의 분석 과정

변수 선택 → 의사결정나무 형성 → ***가지치기** → 모형 평가 및 예측

*** 가지치기** : **오버피팅**을 막고 일반화 성능을 높여주는 과정



3.2.5. 앙상블 (Ensemble)

- 주어진 자료로부터 여러 개의 학습 모델을 만든 후, 학습 모델들을 조합해 하나의 최종 모델을 만드는 개념
- 다양한 약학습기를 통해 강학습기를 만들어가는 과정

- ① 약학습기(약분류기, Weak Learner)
무작위 선정이 아닌 성공확률이 높은, 즉 오차율이 일정 이하(50% 이하)인 학습 규칙
- ② 강학습기(강분류기, Strong Learner)
약학습기로부터 만들어 내는 강력한 학습 규칙

앙상블 분석의 종류

- ③ 보팅 (Voting)
- ④ 부스팅 (Boosting)
- ⑤ 배깅 (Bagging: Bootstrap Aggregation)
- ⑥ 스택킹 (Stacking)

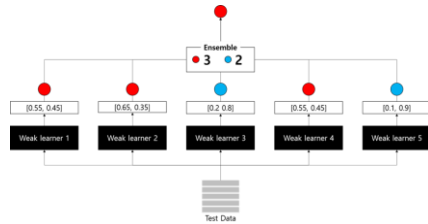


3.2.5. 앙상블 (Ensemble) - 앙상블의 종류

① 보팅 (Voting)

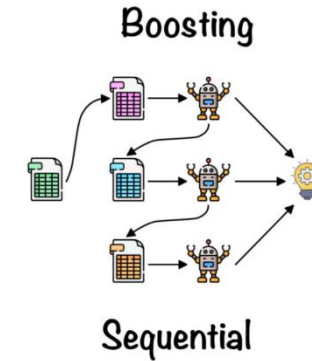
서로 다른 여러 학습 모델을 조합해서 결과물에 대해 최종 투표하는 방식

- 하드 보팅 : 결과물에 대한 최종 값을 투표해서 결정
- 소프트 보팅 : 최종 결과물이 나올 확률 값을 다 더해서 최종 결과물에 대한 각각의 확률을 구한 뒤 최종 값을 도출하는 방법



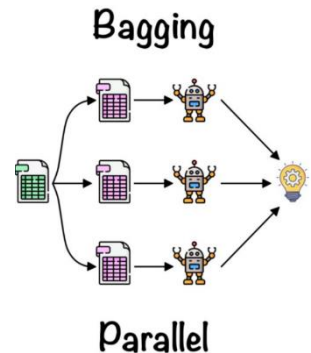
② 부스팅 (Boosting)

가중치를 활용해 **연속적인** 약학습기를 생성, 이를 통해 강학습기를 만드는 방법



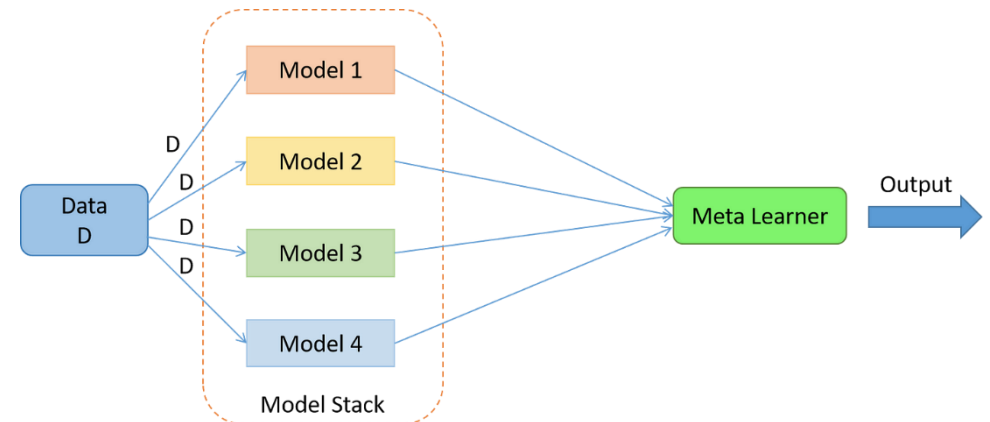
③ 배깅 (Bagging: Bootstrap Aggregation)

샘플을 여러 번 뽑아 각 모델을 학습시켜 결과물을 집계하는 방법



④ 스택킹 (Stacking)

여러 개의 다른 모델을 조합해 예측 성능을 향상시키는 방법



3.3. 인공지능망

인공신경망 (ANN: Artificial Neural Network)

- 인간의 두뇌 신경세포인 뉴런을 기본으로 한 기계학습 기법

인공신경망의 발전

① 기존 신경망 다층 퍼셉트론이 가진 문제

문제 1. 사라지는 경사도

신경망 층수를 늘릴 때 데이터가 사라져 학습이 잘 되지 않는 현상

문제 2. 과대적합(Overfitting)

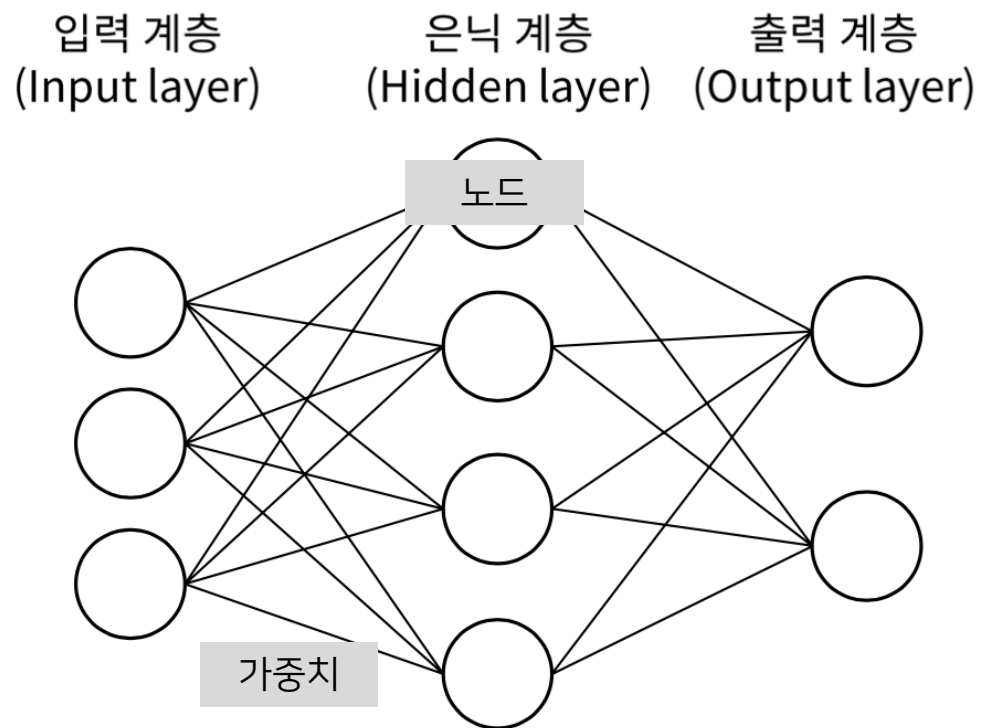
데이터가 충분하지 않은 경우, 신규 데이터에 대한 추론처리 성능이 낮아지는 문제

② 딥러닝의 등장

문제 1 해결. 사전학습(pre-training)으로 사라지는 경사도 문제 해결

문제 2 해결. 초기화 알고리즘 발전 및 드롭아웃을 사용해 과대적합 문제 해결

3.3.1. 인공신경망의 구조



노드

신경계 뉴런.
가중치와 입력값으로 활성화함수를 통해 다음 노드로 전달

가중치

신경계 시냅스
노드와의 연결계수

활성함수

임계값을 이용해 노드의 활성화 여부를 결정

입력층

학습 위한 데이터 입력

은닉층

다층 네트워크에서 입력층과 출력층 사이
데이터를 전파학습

출력층

결과값 출력

3.3.2. 인공지능망의 학습

- 신경망에는 적응 가능한 **가중치**와 **편향**이 있으며, 이를 **훈련 데이터 적응하도록 조정하는 과정**을 학습이라고 정의한다.

손실함수

- 신경망이 출력한 값과 실제 값과의 오차에 대한 함수
- 일반적으로 **평균제곱오차** 또는 **교차엔트로피 오차**를 활용한다.

평균제곱 오차 (MSE: Mean Squared Error)

인공지능망의 출력 값과 사용자가 원하는 출력 값 사이의 거리 차이를 오차로 사용
각 거리 차이를 제곱하여 합산한 후 평균을 구함

교차엔트로피 오차 (CEE: Cross Entropy Error)

분류 부문으로 t값이 원-핫 인코딩 벡터이며, 모델의 출력 값에 자연로그를 적용해 곱함

학습 알고리즘

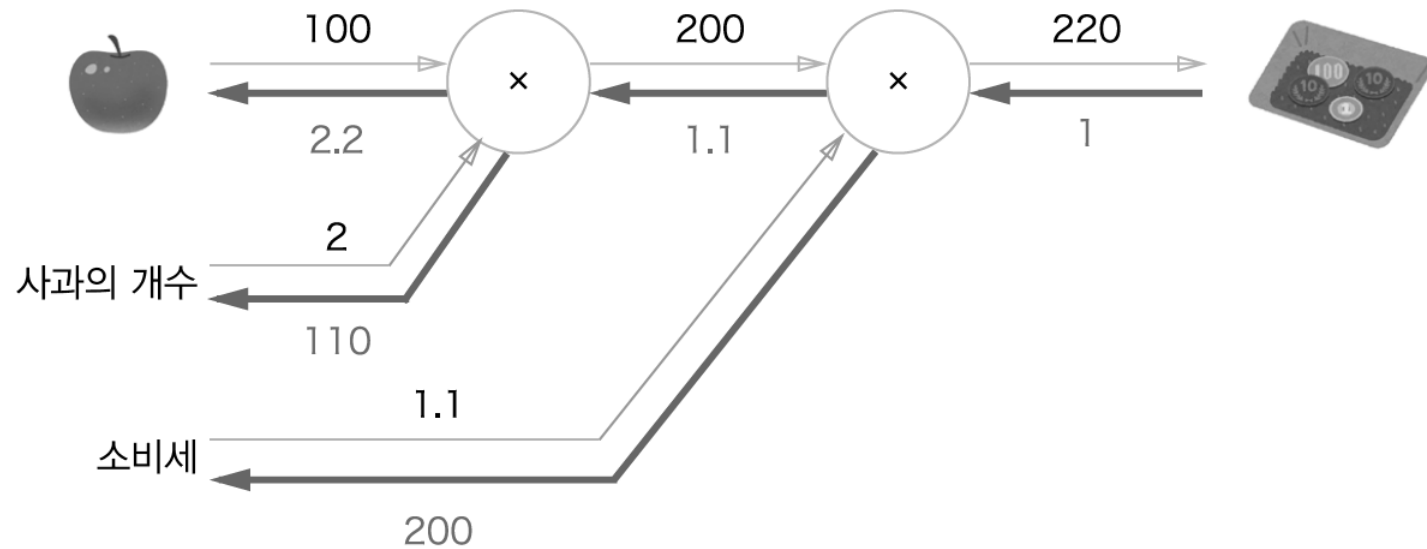
: 미니배치 → 기울기 산출 → 매개변수 갱신

- ① 미니배치
훈련 데이터 중 일부를 무작위로 선택
- ② 기울기 산출
미니배치의 손실함수 값을 최소화하기 위해 경사법으로 가중치 매개변수의 기울기를 미분을 통해 도출
경사 하강법 / 경사 상승법 / 확률적 경사 하강법을 활용함
- ③ 매개변수 갱신
가중치 매개변수를 기울기 방향으로 조금씩 업데이트하며 ①~③단계를 반복함

3.3.3. 인공지능망 - 오차역전파

오차역전파 (Back Propagation)

- 오차를 출력층에서 입력층으로 전달하며 연쇄법칙을 활용한 역전파를 통해 가중치와 편향을 계산하고 업데이트하는 방식



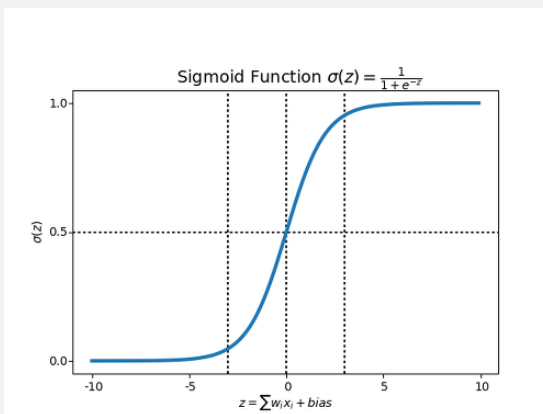
3.3.4. 인공지능망 – 활성화 함수

활성화 함수 : 시그모이드, ReLU, 소프트맥스

시그모이드

이진 분류

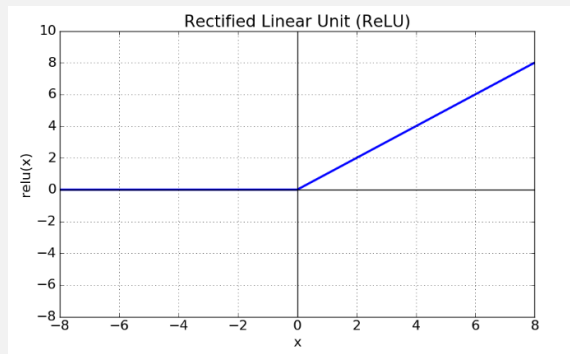
참: 0.5~1 사이 값 출력
거짓: 0~0.5 사이의 값 출력



ReLU (Rectified Linear Unit)

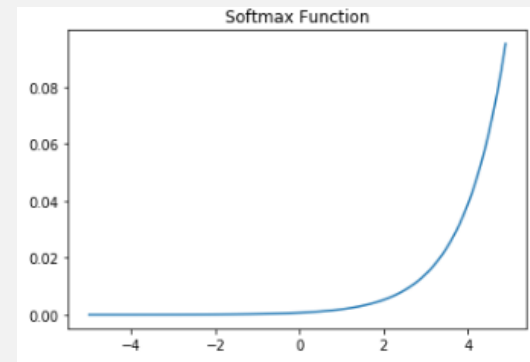
이진 분류
시그모이드의 사라지는 경사도
(Gradient Vanishing) 문제 해결

0 보다 크면 입력값 그대로 출력
0 이하의 값만 0으로 출력



소프트맥스

입력받은 값을 0~1 사이 값으로
모두 정규화해서 출력
모든 값들의 합은 항상 1이 됨



3.4. 군집 분석

- 주어진 개체들의 유사성을 분석해서 높은 대상끼리 일반화된 그룹으로 분류하는 기법
- 비지도학습의 일종

군집분석의 목적 및 활용

- 특정 변수에 대한 정의가 필요하지 않은 탐색적 기법
- 규칙 내지 결과 없이 주어진 데이터들을 가장 잘 설명하는 그룹 또는 클러스터를 찾을 수 있는 방법
- ex. 인터넷 사기/스팸 패턴 발견, 주거 그룹 판별 조사, 생물체 분류 연구

유사성 계산 방식별 척도

- 1) 거리 기반
 - ① 유클리드 거리
 - ② 맨해튼 거리
 - ③ 민코프스키 거리
 - ④ 자카드 거리
- 2) 유사성 기반
 - ① 코사인 값
 - ② 상관계수

군집 분석의 종류

- 1) K-Means
- 2) 병합군집
- 3) 확률분포 기반 클러스터링 (Gaussian Mixture Model)
- 4) 밀도 기반 클러스터링 (DBSCAN)



빅데이터분석기사 필기특강

빅데이터 결과 해석

교차 검증 | 모델 평가 | 과대적합 방지

4.1. k-fold 교차검증

- 전체 데이터셋을 k개의 서브셋으로 분리해 그 중에 k-1개를 훈련 데이터로 사용하고 1개의 서브셋은 테스트 데이터로 사용해 모델의 성능을 평가하는 교차검증 기법
- 테스트를 중복없이 병행 진행한 후 평균을 내어 최종 모델의 성능을 평가한다.



4.2. 분류 모델 평가

- 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 $Precision = \frac{TP}{TP + FP}$

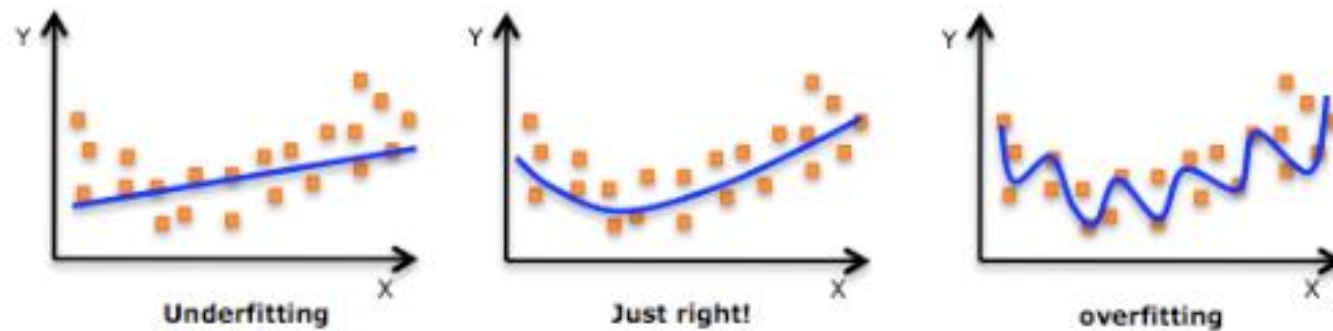
$$F1-Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

4.3. 회귀 모델 평가

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	$\frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$

4.4. 과대적합 방지

- 과대적합: 훈련 시에는 높은 성능을 보이지만, 테스트 데이터에 대해서는 낮은 성능을 보여주는 것



과대적합 방지를 위한 기법들

- ① 모델의 낮은 복잡도 → 정규화, 드롭아웃
- ② 가중치 감소 → L1 규제, L2 규제
- ③ 편향-분산 트레이드 오프
- ④ Max Depth, 가지치기 등

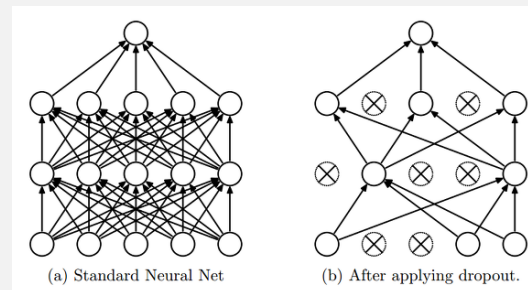
4.4.1. 과대적합 방지 기법들

과대적합 방지를 위한 기법들

- ① 모델의 낮은 복잡도 → 정규화, 드롭아웃
- ② 가중치 감소 → L1 규제, L2 규제
- ③ 편향-분산 트레이드 오프
- ④ Max Depth, 가지치기 등

드롭아웃

신경망 모델에서 은닉층의 뉴런을 임의로 삭제하면서 학습하는 방법



L1 규제, L2 규제

규제 : 과대적합이 되지 않도록 모델을 강제로 제한하는 것

- ① L2 규제 : 가중치 값을 비용함수 모델에 비해 작게 만들어 내는 방식
- ② L1 규제 : L2 규제의 가중치 제약을 절대값으로 바꾸는 개념
→ 회귀모델에서 L1규제를 적용한 것이 라쏘(Lasso) 모델



빅데이터분석기사 필기특강

필기 특강 완료



빅데이터분석기사 필기특강

파이썬 문법 정리



빅데이터분석기사 필기특강

실기1: 데이터 전처리



빅데이터분석기사 필기특강

실기2: 머신러닝 모델 - 예측



빅데이터분석기사 필기특강

실기3: 통계 검정



빅데이터분석기사 필기특강

빅분기 실기 모의고사, 기출문제 풀이