AI융합비즈니스포럼

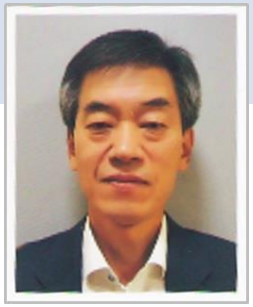
AI 융합전문가 7차시

데이터분석 II

2024.08.18.

강사 홍용기 컨설팅학 박사

현재 하고 있는 일



소속	주요 업무
㈜제타데이터 대표이사	데이터 분석, 전략 및 컨설팅, 데이터 가치평가 ODA 컨설팅 (Official Development Assistance)
㈜지구파트너스 감사	창업보육, 투자, 기업·기술가치평가, 사업타당성 분석
㈜메타로직 컨설팅 수석	ISP 컨설팅 (Information Strategy Planning) ISMP 컨설팅 (Information System Master Plan)

◎ 자격증

1. 경영지도사31기 (인적자원, 2016)
2. 창업보육매니저 (BI협회, 2018)
3. 기업·기술가치평가사 (기업·기술가치평가협회, 2018)
4. 기업재난관리사 실무과정 (행정안전부, 2019)
5. 데이터분석 준전문가 ADsP (데이터산업진흥원 K-Data, 2021)
6. 빅데이터 분석기사 (과학기술정보통신부・통계청, 2021)
7. 국제공인컨설턴트 CMC (ICMCI, 2023)
8. 인공지능(AI) 활용마스터1급 (뉴미디어교육연구소, 2024)

데이터분석 관련 비즈니스

◎ 정보화전략계획수립(ISP) 컨설팅 수행

20.05~20.08. 창업진흥원
20.10~20.12. 한국연구재단
21.01~21.04. 소상공인시장진흥공단
21.11~22.06. 서울특별시
22.08~22.12. 경찰대학교
23.08~24.05 ODA (요르단 경찰청 PSD)

◎ 2022년 AI학습용데이터 구축사업 평가

1차 08 방송 콘텐츠 대화체 음성인식 데이터
09 방송 콘텐츠 한국어·영어 통번역 데이터
43 갑각류 종자생산 데이터
48 식생 탄소 포집량 식별 데이터
2차 74 축산 기자재(소, 돼지) 3D 데이터
75 소(한우, 젖소) 및 돼지 발정행동 데이터
3차 06 인공지능 신기술 선도(자유 공모)

◎ 데이터 가치평가 컨설팅

23.09~23.11 중소벤처기업진흥공단

발급번호:00KI-I83K-W6YQ-OA64-CG1Z

소프트웨어기술자 경력증명서

성명	홍용기		생년월일	1964.08.25
현 근무처	회사명		사업자등록번호	
	전화번호		업종	
	소재지			

근무경력	확인여부	근무기간	회사명	담당업무	부서/직위

기술자격	종목 및 등급	등록번호	취득일	발급기관
	빅데이터분석기사	BAE-002000023	2021.07.16	한국데이터산업진흥원
	ADsP(데이터 분석 준전문가)	ADsP-028000961	2021.04.09	한국데이터산업진흥원

학력	학교명	학과(전공)	수학기간	학위

교육	기간	과정	수료번호	교육기관

상훈	수여일	종류	상훈기관	근거

기술경력	확인여부	참여사업명	참여기간	발주자	소속사	직위	담당업무
	확인	국민 체감형 치안 안심 플랫폼 구축 정보화전략계획 사업	2022.08.04 ~ 2022.12.31	경찰대학교	(주)메타로직컨설팅	프리랜서	IT컨설팅 > 정보기술기획
	확인	서울시 차세대 지방세입정보시스템 통합구축 변화관리 컨설팅	2021.11.01 ~ 2022.06.30	서울특별시	(주)메타로직컨설팅	프리랜서	IT컨설팅 > 정보기술컨설팅
	확인	소상공인지원사업 디지털전환 정보화전략계획(ISP) 용역	2021.01.04 ~ 2021.05.03	소상공인시장진흥공단	(주)메타로직컨설팅	수석컨설턴트	IT컨설팅 > 정보기술컨설팅
	확인	한국연구재단 중장기 정보화전략계획(ISP) 수립	2020.10.05 ~ 2021.01.04	한국연구재단	(주)메타로직컨설팅	수석컨설턴트	IT컨설팅 > 정보기술컨설팅
	확인	창업기업확인시스템 구축을 위한 정보화전략계획(ISP) 수립	2020.05.18 ~ 2020.08.17	창업진흥원	(주)메타로직컨설팅	수석컨설턴트	IT컨설팅 > 정보기술컨설팅

「소프트웨어 진흥법」 제24조제3항 및 같은 법 시행규칙 제13조제3항에 따라 소프트웨어기술자의 경력 사항을 증명합니다.

2023년 01월 25일

강사 홍용기 컨설팅학 박사

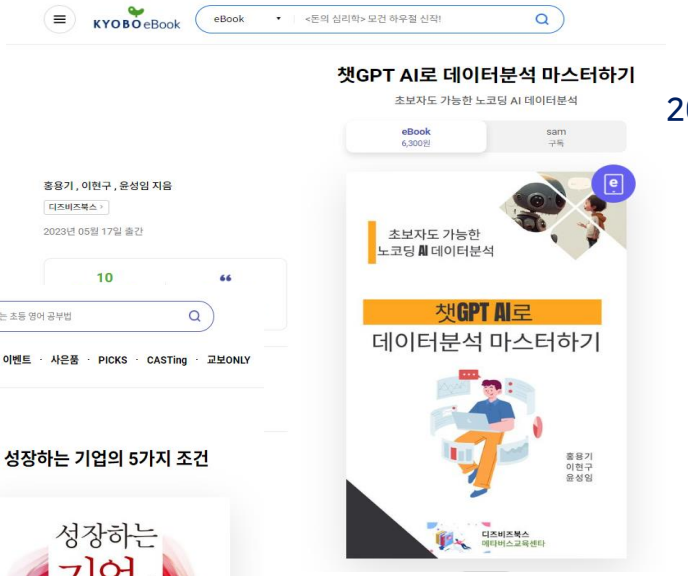
데이터분석 관련 강의

- 데이터분석 및 실전 R코딩 (경영기술지도사회, 빅데이터 분석기사 자격증 취득 과정)
- 데이터분석 Python 심화과정 (서울 여성능력개발원 강동 여성인력개발센터 / 용산 여성인력개발센터)
- 파이썬 코딩을 통한 크롤링 자동화 인텐시브 과정 (경영지도사 및 컨설턴트)
- AI & ChatGPT 활용 및 데이터분석 컨설팅 방법론 (경영기술지도사회, 국제공인컨설턴트 CMC 양성과정)
- AI & 데이터분석 (매경아카데미, 동북아 ICT 포럼)

강사 홍용기 컨설팅학 박사

책 쓰기 프로젝트

2018년 1월



2023년 5월

2024년 6월

2024년 4월

“놀랍도록 똑똑하고 충격적으로 어리석다”

최예진 교수

(미 워싱턴대 컴퓨터과학과 교수, 옥스퍼드 대학교 AI윤리 연구소 선임연구원)

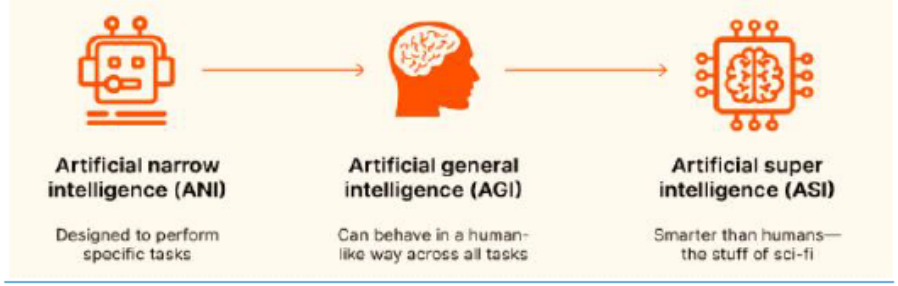
https://v.daum.net/v/20240612211300429

Road to AGI (범용인공지능, Artificial General Intelligence)

현재 널리 사용되는 AI는 '좁은 의미의 인공지능(ANI)'이며, AGI, ASI로 진화

- (ANI, Artificial Narrow Intelligence, 좁은 의미의 인공지능) 언어 번역, 이미지 인식, 게임 등 특정 영역에서만 인간을 능가하는 성능을 보유한 최적화된 인공지능 의미(유아 수준)
  - 얼굴 인식, 체스 게임, 언어 번역과 질병 진단 등 인간이 할 수 있는 특정 작업을 수행할 수 있지만 해당 영역을 넘어서 학습하거나 일반화할 수는 없음
    - ANI는 좁은 의미의 인공지능(때때로 약한 AI로 간주되지만 AI 연구의 의미에 실제로 적용)은 특정 작업을 수행하도록 설계된 AI 시스템으로 ChatGPT 및 Bard와 같은 기타 AI 챗봇은 여전히 좁은 의미의 AI
- (AGI, Artificial General Intelligence, 범용인공지능) 특정 분야에만 특화되어 있는 것이 아닌, 기본적으로 기본적인 이해 능력, 추론, 문제 해결, 창의적 사고 등을 갖춘 인간과 유사한 또는 높은 지능 수준을 가진 인공지능을 의미(성인 수준)
  - 경험을 통해 배우고, 추론하고, 이해하고, 소통하고, 일반 지식과 상식이 필요한 문제를 해결
  - AGI는 그것이 무엇을 의미하든 일반적으로 지능적(generally intelligent)이어야 함(G가 중요)
    - AGI, 즉 강력한 AI는 인간과 유사한 지능(또는 "일반적으로 인간보다 똑똑함") 을 나타내는 AI
- (ASI, Artificial Super Intelligence, 초인공지능) 범용인공지능보다 한 단계 발전한 개념으로 모든 면에서 인간의 지능과 능력을 닮은 혹은 인간보다 뛰어난 지능 수준(신과 동등한 수준*)
  - AGI는 특정 조건에서만 쓰이는 현재 AI 기술과 달리 모든 상황에 두루 적용할 수 있는 차세대 AI 모델이며, 물질과 에너지 제어 등 인류와 우주의 미래에 심오한 영향을 미치게 됨
    - The Terminator의 Skynet 또는 Avengers: Endgame의 Jarvis와 같은 일부 가상 캐릭터
    - ASI는 인간의 지능을 훨씬 능가하는 AI 시스템으로, 지각 있는 공상과학 슈퍼컴퓨터

그림7 ANI vs AGI vs ASI



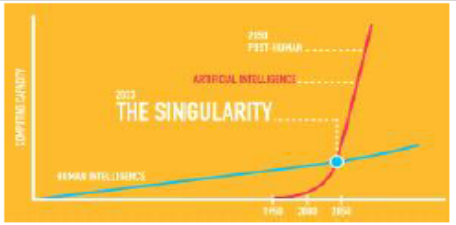
자료 : Zapier

[멀티모달 AI 경쟁과 다가오는 AGI (정보통신기획평가원 IITP, 2024.05)]

인간과 같은 수준의 AGI(범용인공지능)는 언제 나올 수 있는가

- 레이 커즈와일(Ray Kurzweil, 미래학자)은 '특이점이 온다(The Singularity is Near)'에서 '기술적 특이점' 즉, 싱귤래리티 도래를 주장(2005년)
  - 앞으로 30년 후인 2045년 경 인류는 인간보다 뛰어난 기계가 출현하는 특이점에 도달
  - 최근, 레이 커즈와일은 인간의 두뇌를 닮은 인공지능 AGI가 5년 후, 2029년까지 현실이 될 것이고 주장(SXSW 2024)

그림8 특이점(Singularity)



- 싱귤래리티라(특이점)는 원래 수학계에서 쓰던 용어로 일반적 이론이 통하지 않는 지점을 뜻함. 이후 물리학에서 중력장이 비정상적으로 무한히 뻗어 나가는 점을 가리킬 때 쓰임
- 특이점 단계는 AI 기술의 급격한 발전으로 인류의 능력이 비약적으로 향상되고, 궁극적으로는 인간과 기계의 융합을 통해 초월적 존재가 될 가능성을 시사
- 특이점은 인간 본연의 생물학적 조건을 뛰어넘는 새로운 문명을 낳는 시점

자료 : 언론 보도자료 정리 등

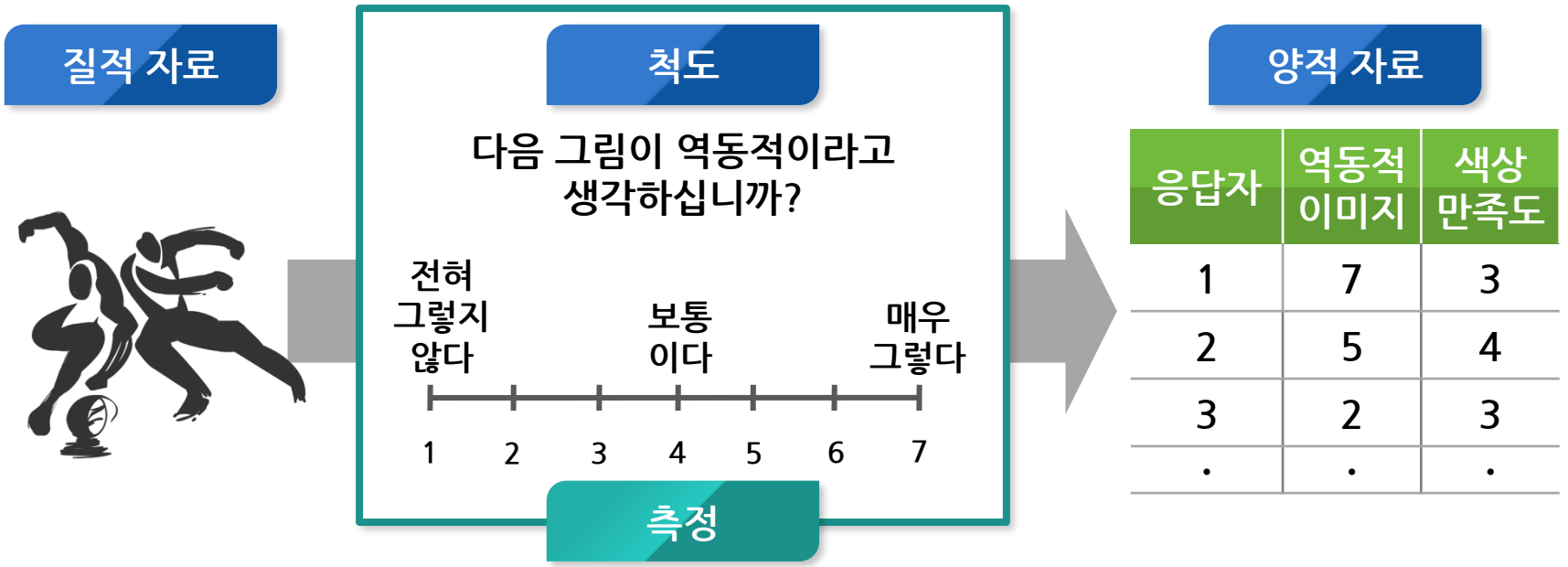
- 젠슨 황(Jensen Huang, 엔비디아 CEO)은 인간과 같은 수준의 범용인공지능(AGI)이 5년 이내 등장할 것이라고 전망(GTC 2024)
  - '인간이 치르는 모든 종류의 시험을 통과할 수 있는 AI'를 AGI라고 전제
- 일론 머스크(Elon Musk, 테슬라 CEO)는 AGI를 가장 똑똑한 인간보다 똑똑한 AI로 정의한다면 내년 혹은 후년에 가능하고, 2029년에는 모든 인간지능을 합친 것보다 더 똑똑해질 것임(2024년, 노르웨이 국부펀드 니콜라이 탕겐 CEO 인터뷰)
  - 머스크는 데이터센터 장비와 전력망 공급이 관건이라고 지적, 작년의 칩 제약에서 올해는 변압기 공급장치로 이동
- 샘 알트만(Sam Altman, OpenAI CEO), AGI는 인류가 발명한 가장 강력한 기술이 될 것이라고 생각(2023년)하며, 오픈AI 역시 '인간과 같은 추론 능력'의 AGI 개발이 핵심
  - 향후 4~5년 이내에 AGI가 구축될 것으로 전망
    - AGI는 매우 다른 세상이고 그것은 공상과학이 오랫동안 우리에게 약속해 온 세상이고, 처음으로 그것이 어떤 모습일지 볼 수 있게 된 것 같음
    - 오픈AI는 '23년 말 '인간처럼 생각하는' 능력으로 정답이 확실한 수학 문제 등을 풀어내는 'Q*(큐스타)'라는 모델을 개발, AGI 개발로 가는 돌파구를 찾은 것으로 알려짐
- 데미스 하사비스(Demis Hassabis, 딥마인드 CEO)는 AGI가 10년 내 실현하더라도 놀라지 않을 거다. 하지만 꼭 그렇게 될 거라는 말은 아니다. AGI가 10년내 실현될 확률은 50%이며, 이 시간표는 딥마인드 설립 이래 결코 변하지 않았다'고 말함(파이낸셜타임즈 인터뷰)
  - 이를 위해 '성능과 규모가 기하급수적으로 증가하는 문제 해결'과 '발열과 전기' 해결이 선행조건

OpenAI의 figure01 (https://www.youtube.com/watch?v=2RsDtRhHnOg&t=15s)

Like every great presentation, I've divided my talk into three subjects. Steve Jobs -

Ⅰ.	Ⅱ.	Ⅲ.
Data Analysis Basic Theory	Regression Analysis	Classification Analysis

- 척도의 개념
  - 특정 속성을 측정하여 그 정도를 숫자로 나타내는 규칙
  - 질적 자료를 양적 자료로 전환시켜 주는 도구



- 척도의 종류
- 어떤 척도를 사용하는지에 따라 측정된 숫자에 내재된 정보량이 달라지며, 적용 가능한 통계분석 기법이 달라짐

척도의 종류	내용
명목척도	응답보기들을 상호 배타적으로 구분하기 위해 임의의 숫자를 부여하는 척도
서열척도	응답보기들을 구분하고, 구분한 응답보기들의 순서까지 측정하는 척도
등간척도	서열 척도에 포함된 정보(분류, 서열정보)외에 거리(간격)정보까지 가지는 척도
비율척도	절대 영점을 가지고 있어서 속성의 상대적 크기 뿐만 아니라, 절대적 크기의 비교도 가능한 척도

1. 명목 척도(Nominal scale)

- 응답보기들을 상호 배타적으로 구분하기 위해 임의의 숫자를 부여하는 척도
- 선택한 응답을 기준으로 응답자들을 특정 집단으로 분류하기 위해 사용(=분류정보)

> 예
>
> **귀하는 다음 중 어떤 훈련과정에 입학을 원하십니까?**
>
> **1) A과정 2) B과정 3) C 과정 4) D 과정 5) 기타**

- 숫자는 '크기'의 의미가 없는 명칭에 해당하기 때문에 사칙연산은 무의미함
- 대표치는 최빈치(Mode) : 응답보기 중 가장 많이 선택된 응답보기의 선택된 수
- 4가지 척도 중 정보량이 가장 적은 척도 : 분류 정보만 보유

2. 서열 척도(Ordinal scale)

- 응답보기들을 구분하고, 구분한 응답보기들의 순서까지 측정하는 척도
- 응답보기들의 속성을 서열로 나타내는 척도(=분류정보 + 순서정보)
- 응답보기 간의 간격은 측정하지 않고 순서만 측정함

  - 응답 보기들 간의 순위만 나타낼 뿐, 얼마나 더 선호되는지는 측정이 불가능함

> 예 **다음 교육과정 중 귀사에서 가장 중요하다고 생각하는 대로 순서를 기입해 주십시오.**
> **A과정 ( ), B 과정 ( ), C 과정 ( ), D 과정 ( )**

- 사칙연산은 무의미

  - 순위 간 간격이 서로 달라 숫자 차이에 절대적 의미가 없기 때문
  - 1, 2순위의 차이보다 3, 6순위의 차이가 3배 크다고 할 수 없음

- 대표치로서 중앙값(Median)를 사용함
- 명목 척도 다음으로 적은 정보를 보유함 : 분류 정보 + 순서 정보

3. 등간 척도(Interval scale)

- 서열 척도에 포함된 정보(분류, 서열정보)외에 거리(간격)정보까지 가지는 척도
- 간격이 동일한 서열척도
- 속성의 상대적 크기를 측정하기 위해 균일한 간격으로 분할한 길이를 이용하여 측정

  - 예) 온도계, IQ 등

  - 온도계로 측정한 1도와 2도 간의 차이는 2도와 3도 간의 차이와 동일함

- 간격 척도의 숫자 자체는 절대적 의미를 가지지 않음
- 절대 영점이 없기 때문에 숫자 간 비율개념 없음
- 간격 척도에서 숫자 간의 차이는 절대적 의미를 가짐(차이 값 간 비율개념있음)
- 대표치로서 산술평균을 사용
- 정보량 : 분류 정보 + 순서 정보 + 상대적 크기 정보

3. 등간 척도(Interval scale)

예 **지난 6개월간 참여하신 교육과정이 취업역량 확보에 도움이 되셨습니까?**

- 5점 응답자와 3점 응답자의 만족도 차이가 5점과 4점 응답자의 만족도 차이 보다 2배 크다고 할 수 없음(응답보기(척도점) 간 간격이 동일하다고 볼 수 없기 때문)
- 따라서 간격 척도라기 보다 서열척도에 가까움
- 하지만, 사회과학연구의 특성을 고려하여 척도점 간 간격이 동일하고,각 척도점의 의미를 응답자들이 동일하게 이해한다는 전제 하에 간격 척도로 인정함

4. 비율 척도(Ratio scale)

- 절대 영점을 가지고 있어서 속성의 상대적 크기 뿐만 아니라, 절대적 크기의 비교도 가능한 척도

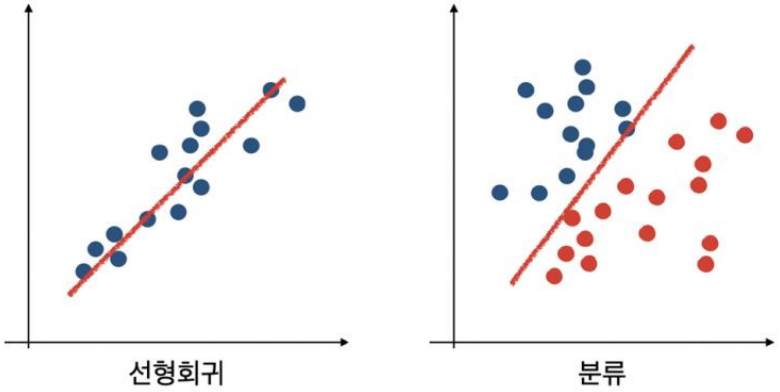
예 나이 ( )세, 근무기간( )년, 연봉( )원

예 하루에 보통 몇시간 정도 훈련이 가능하십니까? ( )시간

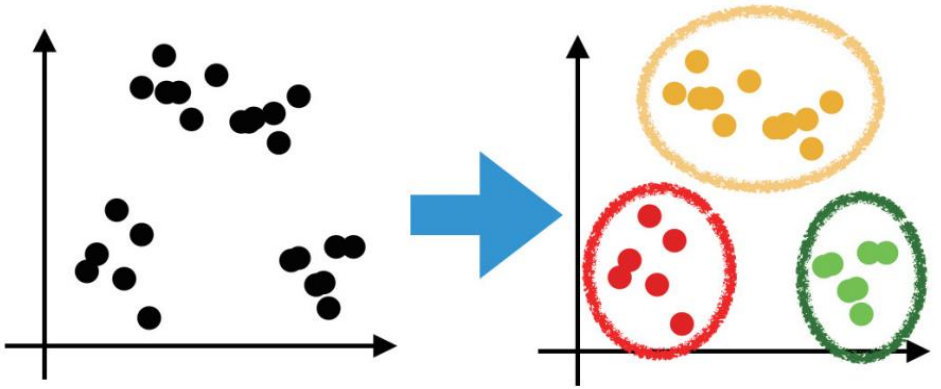
- 만족도, 선호도, 인지도 등 절대 영점이 존재하기 어려운 소비자의 사고나 인지수준에 대한 측정은 한계가 있음
- 직접 관찰할 수 있는 물리적 사건이나 현상을 측정하는데 주로 사용함
- 사칙연산이 가능하며, 대표치는 평균값
- 4가지 척도 중 가장 정보량이 많은 척도 : 분류정보 + 순서정보 + 상대적 크기 정보 + 절대적 크기 정보

Supervised learning VS Unsupervised learning

지도 학습supervised learning



비지도 학습unsupervised learning



※출처: 으뜸 데이터 분석과 머신러닝(박동규·강영민, 2021)

Data split

id	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	y

Data split

Training data

id	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	y
					X_train								y_train

y_predict

Test data

						X_test							y_test

Data split

Training data

id	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	y
					X_train								y_train

Validation data

y_predict

					X_val								y_val

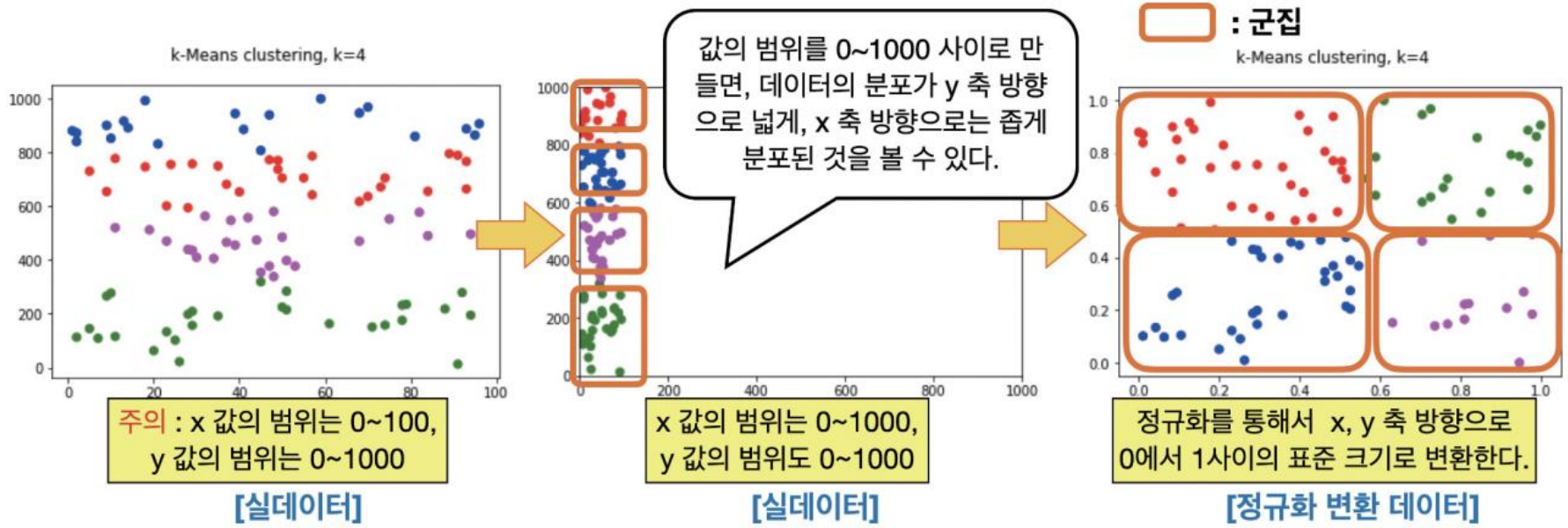
Test data

					X_test								y_test

Missing value : NA, NAN, Null

➔ 결측치 제거 또는 대체 (평균, 중위수, 최빈값)

정규화normalization(min-max scale), **표준화**standardization(평균=0, 분산=1로 만듦)



※출처: 으뜸 데이터 분석과 머신러닝(박동규·강영민, 2021)

1단계 : 데이터 확인

- 분석할 데이터의 특성을 확인하는 단계
- 변수의 특성(독립변수/입력변수)과 타겟(종속변수/반응변수)의 존재 여부 파악
- 적용가능한 분석모델 확인(ex. 타겟 연속된 수치형이라면 회귀분석, 범주형이라면 분류분석)
- 타겟이 없는 데이터라면 비지도학습 적용

- 독립변수, 종속변수 확인
- 연속형 vs 범주형 확인
- 범주형 독립변수 여부확인
- 적용가능한 분석모델 확인
  (회귀, 분류, 비지도 학습)

2단계 : 데이터 분할

- 학습용 데이터와 평가용 데이터를 분할하는 단계
- 데이터는 학습데이터(60~80%), 검증데이터(10~20%), 평가데이터(10~20%)로 분할
- 예측을 수행하는 데이터 세트는 학습용 데이터 세트가 아니라 평가 전용 데이터세트여야 함
- 단순 학습데이터 + 복잡한 평가데이터의 경우 평가데이터의 특징을 반영하지 못할 수 있음
- 데이터 크기가 작은 경우나, 검증 결과를 일반화하기 위해 교차검증방법을 적용

- 독립변수, 종속변수 확인
- 연속형 vs 범주형 확인
- 범주형 독립변수 여부확인
- 적용가능한 분석모델 확인
  (회귀, 분류, 비지도 학습)

- 학습데이터: 60~80%
- 검증데이터: 10~20%
- 평가데이터: 10~20%
- 교차검증방법 적용 가능

3단계 : 전처리

- 데이터의 특성에 따라 분석이 가능한 형태로 변형하는 단계
- 독립변수에 범주형 변수가 있을 경우 데이터 분할 전 One-hot Encoding으로 데이터를 변형
- 변수마다 단위 특성에 차이가 클 때 분석결과에 영향을 줄 수 있으므로, 정규화나 표준화 실시
- 결측치와 이상치는 분석가의 판단과 도메인 상황에 따라 적절한 방법으로 처리

STEP 1 데이터 확인	STEP 2 데이터 분할	STEP 3 전처리	STEP 4 모델학습	STEP 5 성능평가
▪ 독립변수, 종속변수 확인 ▪ 연속형 vs 범주형 확인 ▪ 범주형 독립변수 여부확인 ▪ 적용가능한 분석모델 확인 (회귀, 분류, 비지도 학습)	▪ 학습데이터: 60~80% ▪ 검증데이터: 10~20% ▪ 평가데이터: 10~20% ▪ 교차검증방법 적용 가능	▪ 표준화(평균 0, 표준편차 1) ▪ 정규화(Min-Max Scaling) ▪ 범주형 독립변수 OHE ▪ 결측치 확인 후 처리 ▪ 이상치 확인 후 처리		

4단계 : 모델학습

- 머신러닝 알고리즘을 학습데이터에 적용하는 단계
- 1단계에서 파악한 분석방법에 따라 적합한 라이브러리를 사용해 머신러닝 알고리즘을 적용
- 머신러닝 분석방법은 지도학습과 비지도학습으로 구분되며, 지도학습은 회귀와 분류로 나뉨
- 학습데이터로 학습을 수행, 검증데이터로 학습결과 확인 후 하이퍼파라미터 탐색 및 조절

STEP 1 데이터 확인	STEP 2 데이터 분할	STEP 3 전처리	STEP 4 모델학습	STEP 5 성능평가
▪ 독립변수, 종속변수 확인 ▪ 연속형 vs 범주형 확인 ▪ 범주형 독립변수 여부확인 ▪ 적용가능한 분석모델 확인 (회귀, 분류, 비지도 학습)	▪ 학습데이터: 60~80% ▪ 검증데이터: 10~20% ▪ 평가데이터: 10~20% ▪ 교차검증방법 적용 가능	▪ 표준화(평균 0, 표준편차 1) ▪ 정규화(Min-Max Scaling) ▪ 범주형 독립변수 OHE ▪ 결측치 확인 후 처리 ▪ 이상치 확인 후 처리	▪ 머신러닝 알고리즘 적용 ▪ 회귀, 분류, 비지도학습 ▪ 하이퍼파라미터 탐색·조절 ▪ 최적의 하이퍼파라미터 결정	

5단계 : 성능평가

- 최적의 하이퍼파라미터 및 최종모델 결정 단계
- 최종모델에 평가데이터를 적용하여 머신러닝 알고리즘의 예측성능을 평가
- 평가데이터는 반드시 학습 과정이나 검증 과정에서 사용되지 않은 데이터로 사용해야 함
- 평가데이터에 대한 평가지표를 머신러닝 분석에 대한 최종성능으로 제시

STEP 1 데이터 확인	STEP 2 데이터 분할	STEP 3 전처리	STEP 4 모델학습	STEP 5 성능평가
▪ 독립변수, 종속변수 확인 ▪ 연속형 vs 범주형 확인 ▪ 범주형 독립변수 여부확인 ▪ 적용가능한 분석모델 확인 (회귀, 분류, 비지도 학습)	▪ 학습데이터: 60~80% ▪ 검증데이터: 10~20% ▪ 평가데이터: 10~20% ▪ 교차검증방법 적용 가능	▪ 표준화(평균 0, 표준편차 1) ▪ 정규화(Min-Max Scaling) ▪ 범주형 독립변수 OHE ▪ 결측치 확인 후 처리 ▪ 이상치 확인 후 처리	▪ 머신러닝 알고리즘 적용 ▪ 회귀, 분류, 비지도학습 ▪ 하이퍼파라미터 탐색·조절 ▪ 최적의 하이퍼파라미터 결정	▪ 평가데이터 최종모델 적용 ▪ 평가데이터에 대한 평가지표를 머신러닝 분석에 대한 성능으로 제시

Like every great presentation, I've divided my talk into three subjects. Steve Jobs -

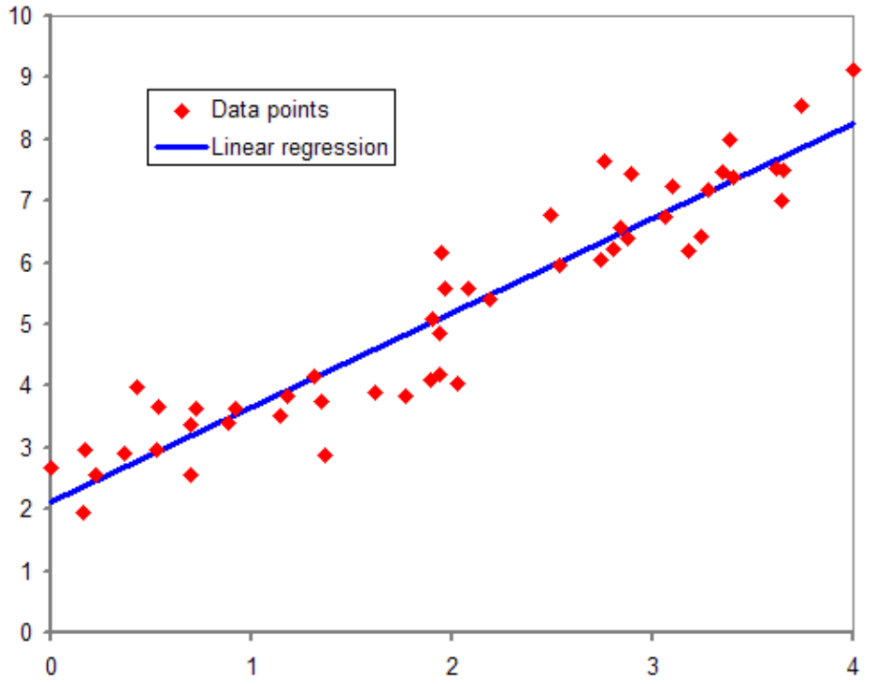
Ⅰ. Data Analysis Basic Theory

Ⅱ. **Regression Analysis**

Ⅲ. Classification Analysis

위키백과 : '회귀분석'

- 회귀(regress)의 원래 의미는 옛날 상태로 돌아가는 것을 의미. 영국의 유전학자 프랜시스 골턴은 부모의 키와 아이들의 키 사이의 연관관계를 연구하면서 부모와 자녀의 키 사이에는 선형적인 관계가 있고, 키가 커지거나 작아지는 것보다는 전체 키 평균으로 돌아가려는 경향이 있다는 가설을 세웠으며, 이를 분석하는 방법을 '회귀분석'이라고 함
- 이후 칼 피어슨은 아버지와 아들의 키를 조사한 결과를 바탕으로 함수 관계를 도출하여 회귀분석 이론을 수학적으로 정립



Regression line for 50 random points in a en:Gaussian distribution around the line y=1.5x+2 (not shown). The regression line (shown) that best fits these points is actually y=1.533858x+2.129333.

- 단순회귀분석 목적

1 하나의 변수(독립변수, 예측변수)를 이용해서 다른 변수(종속변수, 결과변수)를 예측함

예 영업사원의 수나 판촉행사 횟수, 매장의 면적 등 어떤 특정한 하나의 변수를 이용해서 매출액을 예측함

2 하나의 변수(독립변수, 설명변수)를 이용해서 다른 변수(종속변수, 결과변수)를 설명함

예 가격만족도, 품질만족도 등 어떤 특정한 하나의 변수를 이용해서 전반적인 만족도를 설명함

- 단순회귀분석 회귀식

$$Y = \beta_0 + \beta_1 \cdot X$$

Y : 종속변수　　X : 독립변수　　β_1 : 회귀계수　　β_0 : 상수

예 우리회사 내년도 매출액 규모(Y)를 영업사원 수(X)로 예측

➡ **매출액 = β_0 + β_1·(영업사원 수)**

Multiple linear regression

- 다중회귀분석 목적

1

2개 이상 변수(독립변수, 예측변수)를 이용해서 다른 변수(종속변수, 결과변수)를 예측함

예 영업사원의 수, 판촉행사 횟수, 매장의 면적 등 3가지 변수를 이용해서 매출액을 예측함

2

2개 이상 변수(독립변수, 예측변수)를 이용해서 다른 변수(종속변수, 결과변수)를 설명함

예 가격만족도, 품질만족도, 디자인만족도, 무게만족도 등 4가지 변수를 이용해서 전반적인 만족도를 설명함

- 다중회귀분석 회귀식

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \cdots\cdots + \beta_i \cdot X_i$$

Y : 종속변수 **β_i : X_i의 회귀계수**

X_i : 독립변수 **β_0 : 상수**

예 우리회사 내년도 매출액 규모(Y)를 '영업사원 수(X_1), 프로모션 횟수(X_2), 광고비 규모(X_3)'를 이용해 예측하는 다중 회귀식

➡ **매출액 = β_0 + β_1(영업사원 수) + β_2(프로모션 횟수) + β_3(광고비)**

회귀식의 설명력 R^2

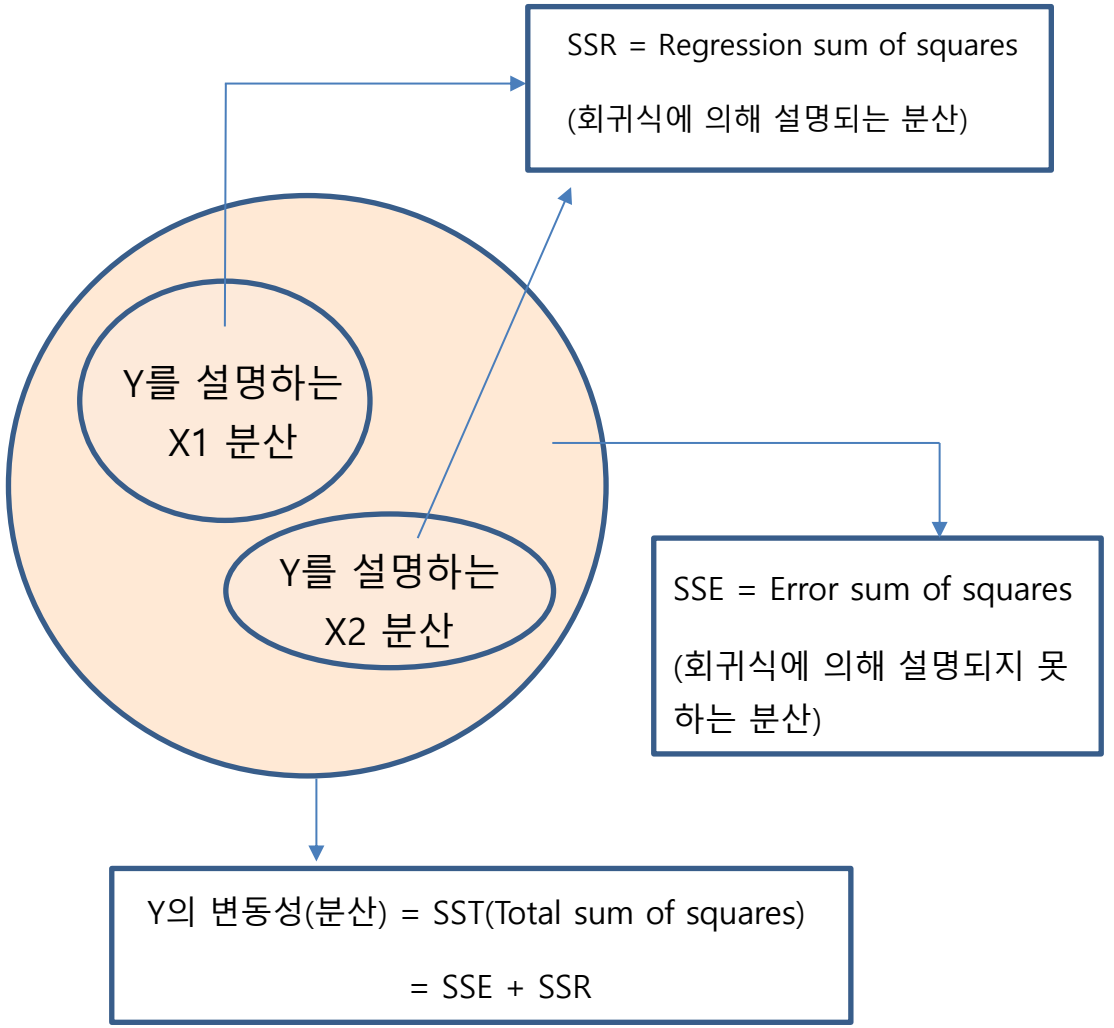
- 회귀식이 종속변수를 설명하고 예측하는데 유용한가를 판단
- 판단지표 : R^2 = (결정계수, 기여율, 설명력), $0 < R^2 < 1$
- R^2 은 종속변수의 분산 중 독립변수에 의해 설명되는 비율을 의미

예 R^2=0.76이라는 것은 종속변수가 가지는 정보 중에서 76%를 독립변수가 설명할 수 있다는 의미

회귀식의 설명력 R^2

$$R^2 = \frac{SSR}{SST}$$

- 그러나, 변수의 수가 증가하면 SSR이 증가하면서 R^2 도 증가하는 하는 문제가 있음
- R^2에 변수의 수 만큼 penalty를 주는 지표인 *adjusted* R^2 를 주로 활용



- 회귀 분석 결과 예시

모형		비표준화 계수		표준화 계수	t	유의 확률	공선성 통계량	
		B	표준오차	베타			공차	VIF
1	(상수)	-.631	.519		-1.215	.235		
	가격만족도	.744	.114	.668	6.528	.000	.298	3.356
	구매 횟수	.302	.094	.331	3.223	.003	.295	3.387
	연령	.011	.011	.054	.962	.345	.983	1.017

a. 종속변수 : 소비자만족도

- $Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3$
- **소비자만족도 = -0.631 + 0.744 · 가격만족도 + 0.302 · 구매 횟수 + 0.011 · 연령**

- 회귀 분석 결과 예시

모형		비표준화 계수		표준화 계수	t	유의확률	공선성 통계량	
		B	표준오차	베타			공차	VIF
1	(상수)	-.631	.519		-1.215	.235		
	가격만족도	.744	.114	.668	6.528	.000	.298	3.356
	구매 횟수	.302	.094	.331	3.223	.003	.295	3.387
	연령	.011	.011	.054	.962	.345	.983	1.017

a. 종속변수 : 소비자만족도

- 가격만족도와 구매횟수의 유의확률이 유의수준보다 작으므로(p-value < 0.05), 통계적으로 유의미한 변수로 판단
- 연령은 유의확률이 유의수준보다 크므로 (p-value > 0.05), 통계적으로 유의하지 않으며 소비자만족도에는 영향을 미치지 않는 변수로 판단

- 회귀모델의 성능 지표

구 분	개 요	수식
평균절대오차 MAE (Mean Absolute Error)	실제 값과 예측한 값의 차이를 절댓값으로 변환해 평균한 값	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert Y_i - \hat{Y}_i\rvert$
평균제곱오차 MSE (Mean Squared Error)	실제 값과 예측한 값의 차이를 제곱한 후 평균한 값	$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
평균제곱근오차 RMSE (Root Mean Squared Error)	실제 값과 예측한 값의 차이를 제곱한 후 평균한 값의 제곱근	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$
평균절대비율오차 MAPE (Mean Absolute Percentage Error)	실제 값과 예측한 값의 차이를 백분율로 표현	$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left\lvert\frac{Y_i - \hat{Y}_i}{Y_i}\right\rvert$

Data : insurance.csv (미국 건강보험료 데이터셋)

✓ 1338행, 7개의 변수 (Target = charges)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column      Non-Null Count   Dtype
---  ------      --------------   -----
 0   age         1338 non-null    int64     연령
 1   sex         1338 non-null    object    성별 (male or female)
 2   bmi         1338 non-null    float64   체질량 지수(body mass index)
 3   children    1338 non-null    int64     자녀의 수(number of children)
 4   smoker      1338 non-null    object    흡연 여부(yes or no)
 5   region      1338 non-null    object    사는 지역(northeast, southeast, northwest, southwest)
 6   charges     1338 non-null    float64   건강보험에서 지출되는 개인별 의료비
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

Like every great presentation, I've divided my talk into three subjects. Steve Jobs -

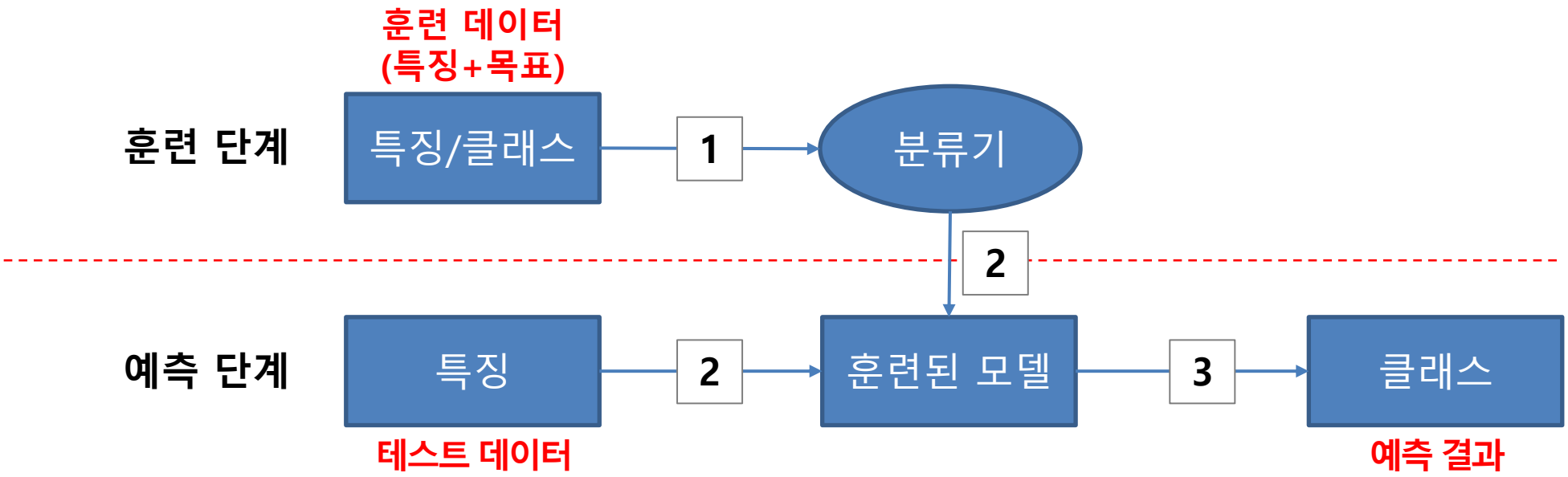
Ⅰ. Data Analysis Basic Theory

Ⅱ. Regression Analysis

Ⅲ. **Classification Analysis**

분류의 개념

- 분류(classification)는 지도학습의 하나로 관측값과 해당 관측값에 대한 범주형 출력을 포함하는 훈련데이터셋이 주어졌을 때 관측값을 목표 범주에 올바르게 매핑하는 규칙을 학습하는 것
- 관측값(observation)은 특징(feature) 또는 예측변수라고도 하며, 목표 범주(category)는 레이블(label), 클래스(class) 또는 타겟(target)이라고도 한다.



분류의 종류와 클래스

- 일반적으로 분류는 두개의 클래스로 분류하는 이진 분류(binary dassification)와 셋 이상의 클래스로 분류하는 다중 분류(multidass dassification)로 나눌 수 있음
- 이진 분류에서 한 클래스를 양성(positive) 클래스, 다른 하나를 음성(negative) 클래스라 함
- 양성 클래스라고 해서 좋은 값이나 장점을 나타내는 것이 아니고 학습하고자 하는 대상을 의미
- 일반 메일에서 스팸 메일을 골라내는 분석의 경우 스팸메일이 양성 클래스가 되고, 양성 종양과 악성 종양을 분별하는 분석에서는 악성 종양이 양성 클래스가 됨

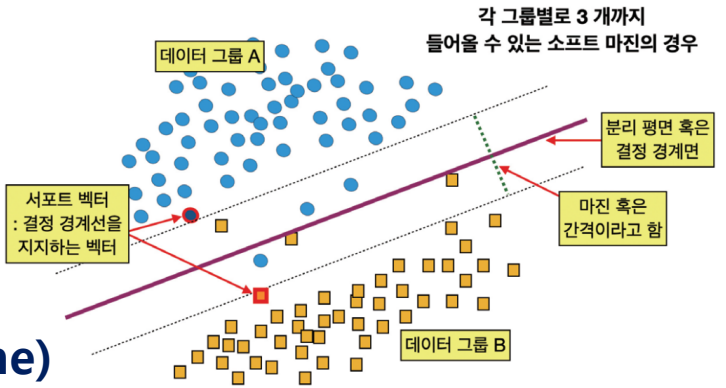
[일반화, 과대적합, 과소적합]

- ✓ 지도학습에서는 훈련데이터로 학습한 모델이 훈련데이터와 특성이 같다면 새로운 데이터가 주어져도 정확히 예측할 거라 기대
- ✓ 모델이 처음 보는 데이터에 대해 정확하게 예측할 수 있으면 이를 "훈련세트에서 데이터 세트로 일반화" 되었다고 함
- ✓ 과대적합은 모델이 훈련세트의 각 데이터에 너무 맞춰져서 새로운 데이터에 일반화되기 어려움
- ✓ 과소적합은 모델이 너무 간단하여 데이터의 면면과 다양성을 잡아내지 못하고 훈련세트에도 잘 맞지 않음

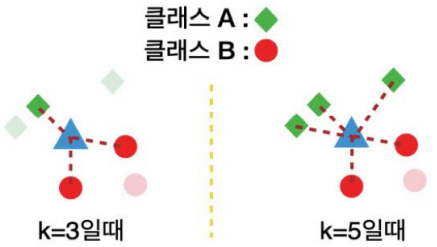
Algorithms ➔ 분류와 회귀 모두 가능한 알고리즘이 많이 있음

- **의사결정 나무 (Decision Tree)**
- **앙상블 모형 (Ensemble)**
  1. Bagging
  2. Boosting
     - AdaBoost (Adaptive Boosting)
     - GBM (Gradient Boosting Machine)
     - XGBoost
     - LightGBM
     - CatBoost
  3. Random Forest

- **서포트 벡터 머신 (SVM; Support Vector Machine)**



- **K 최근접 이웃 (K-Nearest Neighbor)**



- **소프트맥스 (Softmax) 회귀 ➔ 다항 로지스틱 회귀라고도 함**

Confusion matrix

		예측(prediction)	
		양성(Positive)	음성(Negative)
		Positive	Negative

Confusion matrix

		예측(prediction)	
		양성(Positive)	음성(Negative)
실제 (real)	양성	Positive	Negative
	음성	Positive	Negative

Confusion matrix

		예측(prediction)	
		양성(Positive)	음성(Negative)
실제 (real)	양성	True Positive	False Negative
	음성	False Positive	True Negative

Confusion matrix

		예측(prediction)	
		양성(Positive)	음성(Negative)
실제 (real)	양성	TP (True Positive)	FN (False Negative)
	음성	FP (False Positive)	TN (True Negative)

Confusion matrix

		예측(prediction)	
		양성(Positive)	음성(Negative)
실제 (real)	양성	TP	FN
	음성	FP	TN

- 정확도(Accuracy) = (제대로 예측)/(전체) = (TP+TN)/(TP+FN+FP+TN)
- 정밀도(Precision) = (실제 양성)/(양성으로 예측) = TP/(TP+FP)
- 재현률(Recall) = (양성으로 예측)/(실제 양성) = TP/(TP+FN) = 민감도(Sensitivity)
- 특이도(Specificity) = (음성으로 예측)/(실제 음성) = TN/(TN+FP)
- 거짓양성율(FPR) = 1 – 특이도
- F1 score = 2 x 정밀도 x 재현률 / (정밀도 + 재현률)

Data : heart_disease.csv (미국 심장질환 데이터셋)

✓ 303행, 14개의 변수 (Target = target)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count   Dtype
---  ------    --------------   -----
 0   age       303 non-null     int64
 1   sex       303 non-null     int64
 2   cp        303 non-null     int64
 3   trestbps  303 non-null     int64
 4   chol      303 non-null     int64
 5   fbs       303 non-null     int64
 6   restecg   303 non-null     int64
 7   thalach   303 non-null     int64
 8   exang     303 non-null     int64
 9   oldpeak   303 non-null     float64
 10  slope     303 non-null     int64
 11  ca        303 non-null     int64
 12  thal      303 non-null     int64
 13  target    303 non-null     int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

0	age	303 non-null	int64	
1	sex	303 non-null	int64	성별 (1=남성, 0=여성)
2	cp	303 non-null	int64	가슴 통증 (1=안정형 협심증, 2=불안정형 협심증, 3=협심증 이외 통증, 4=무증상)
3	trestbps	303 non-null	int64	휴식 시 혈압
4	chol	303 non-null	int64	콜레스테롤 수치
5	fbs	303 non-null	int64	공복 혈당
6	restecg	303 non-null	int64	휴식 상태의 심전도
7	thalach	303 non-null	int64	최대 심장 박동수
8	exang	303 non-null	int64	운동 유발 협심증
9	oldpeak	303 non-null	float64	운동에 의한 상대적 휴식 시 ST 하강
10	slope	303 non-null	int64	최대 운동 ST 세그먼트의 기울기
11	ca	303 non-null	int64	형광 투시로 착색된 주요 혈관 수
12	thal	303 non-null	int64	탈라세미아 유형
13	target	303 non-null	int64	심장질환의 존재 여부 (1=예, 0=아니오)

GPT-2는 매우 나빴다. GPT-3는 꽤 나빴다.

GPT-4는 나쁜 수준이다. 하지만 GPT-5는 좋을 것이다.

GPT2 was very bad. 3 was pretty bad.

4 is bad. 5 would be okay.

(2024.01, 다보스 세계경제포럼)

선형계획법에 의한 자원 최적화

가구 제조업체 A사는 매일 32kg의 원목을 제공받는다. 공장에는 노동자가 10명 있으며, 이들은 하루에 6시간 근무한다. 책상을 만들려면 3시간의 노동시간과 4kg의 원목이 필요하고, 의자를 만들려면 6시간의 노동시간과 2kg의 원목이 필요하다. 책상은 20만원, 의자는 24만원에 팔린다. **매출을 최대로 올리려면 책상과 의자를 몇 개씩 생산해야 하는가?**

[변수]
x: 생산할 책상의 수
y: 생산할 의자의 수

[제약 조건]
1. 원목 제한: $4x + 2y \leq 32$
2. 노동 시간 제한: $3x + 6y \leq 60$

[목표 함수 (매출 최대화)]
$Z = 200{,}000x + 240{,}000y$

각 제약 조건 한계점에서 x와 y 계산

원목 제한 조건에서:
$x = 0$일 때 $2y \leq 32$이므로 $y = 16$
$y = 0$일 때, $4x \leq 32$이므로 $x = 8$

노동 시간 제한 조건에서:
$x = 0$일 때, $6y \leq 60$이므로 $y = 10$
$y = 0$일 때, $3x \leq 60$이므로 $x = 20$

$x = 0$, $y = 10$ (의자만 최대로):
$Z = 240{,}000 \times 10 = 2{,}400{,}000$원

$x = 8$, $y = 0$ (책상만 최대로):
$Z = 200{,}000 \times 8 = 1{,}600{,}000$원

$x = 4$, $y = 8$
(원목제한과 노동시간제한을 동시 만족):
$Z = 200{,}000 \times 4 + 240{,}000 \times 8$
$= 800{,}000 + 1{,}920{,}000 = 2{,}720{,}000$원

※출처: 한경인터넷신문(https://www.hankyung.com/article/2016052090221)

선형계획법에 의한 자원 최적화

한국주식회사에서 생산하는 제품 A와 제품 B는 용접 공정, 연마공정을 거쳐 생산한다. 제품 A를 생산하기 위해서는 용접 4시간, 연마 3시간, 제품 B를 생산하기 위해서는 용접 2시간, 연마 5시간이 소요된다. 설비의 작업가능 시간은 용접 120시간, 연마공정 100시간이고, 제품 A, B의 단위당 판매 이익은 각각 12만원, 15만원이다. 판매이익을 최대화하기 위한 공정계획은 어떻게 되는가?

[변수]
x: 용접 시간
y: 연마 시간

[제약 조건]
1. 용접 시간 제한: $4x + 2y \leq 120$
2. 연마 시간 제한: $3x + 5y \leq 100$

[목표 함수 (이익 최대화)]
$Z = 120{,}000x + 150{,}000y$

각 제약 조건 한계점에서 x와 y 계산

용접 시간 제한 조건에서:
$x = 0$일 때 $2y \leq 120$이므로 $y = 60$
$y = 0$일 때 $4x \leq 120$이므로 $x = 30$

연마 시간 제한 조건에서:
$x = 0$일 때, $5y \leq 100$이므로 $y = 20$
$y = 0$일 때, $3x \leq 100$이므로 $x = 33.3$

파이썬 반복문으로: {x: 28, y: 3}

제품 A는 $x = 28$개,
제품 B는 $y = 3$개를 생산할 때,
용접 공정, 연마 공정의 제한시간을 만족
(용접 시간 = 118시간, 연마 시간 = 99시간)

$x = 28,\ y = 3$
$Z = 120{,}000 \times 28 + 150{,}000 \times 3$
$= 3{,}360{,}000 + 450{,}000 = 3{,}810{,}000$원

※출처: 블로그 강소기업 제조인(https://m.blog.naver.com/sigmagil/221734015065)

RFM 분석을 통한 고객 세분화와 타겟마케팅

RFM 분석은 마케팅과 고객 관계관리에서 사용되는 기법으로, 고객의 구매 행동을 기반으로 고객을 세분화하고 평가하는 방법

- Recency (최근성): 마지막으로 구매한 시점에서 얼마나 지났는지 측정. 최근 구매 고객일수록 더 가치 있는 고객
- Frequency (빈도): 일정한 기간 동안 얼마나 자주 구매했는지 측정. 구매 빈도가 높은 고객은 충성도가 높은 고객
- Monetary (금액): 일정 기간 동안 얼마나 많은 금액을 지출했는지 측정. 금액이 높은 고객은 많은 수익을 가져다 주는 고객

핵심 프롬프트

1. 파일 업로드(olist order dataset.xlsx, 출처:Kaggle) ➔"RFM 분석을 하고 싶다."
2. "R, F, M 스코어를 기준으로 고객 세그먼트를 해"
   - 고객 세그먼트를 최대한 MECE 하게 설계해
   - 각 세그먼트를 정의하고, 분류기준을 설명해
   - 고객들을 세그먼트 별로 분류해
   - 시트에 세그먼트 이름을 업데이트 해
   - 최종적으로 각 세그먼트 별로 고객 몇 명 포함되어 있는지 설명해
3. "각 액션플랜을 실행해 보기 위하여 개별적인 액션플랜 단가를 책정해"
4. "각 세그먼트별 고객 수에 액션플랜 단가를 곱해서 액션플랜별 소요 예산을 산출해"
5. "각 세그먼트별 예산 합계를 계산해서 엑셀파일로 다운로드 해"

출처: 일잘러 장피엠 유튜브 https://www.youtube.com/watch?v=Ou9X4yu0KG8

인적자원 관리 차원의 퇴직위험직원 예측

기업 경영에서 인적자원은 기업의 성공을 결정짓는 핵심 요소

- 인적자원은 창의성과 혁신의 원천으로 대체가 불가하며, 기업의 경쟁력을 구성하는 기반
- 긍정적이고 협력적인 조직 문화는 직원의 만족도와 충성도를 높이며 성과 향상으로 이어짐
- 직원을 대상으로 한 직업 만족도 조사 데이터를 활용하여 만족도에 영향을 미치는 요소를 파악
- 잠재적으로 퇴직 위험이 있는 직원을 예측

핵심 프롬프트

1. 파일 업로드(employee satisfaction survey.xlsx, 출처:Kaggle) ➔페르소나 제시
2. # Role "당신은 기업의 조직 심리학자로 직무 설계, 리더십 스타일, 조직의 행동 및 문화 등 직무 만족도에 영향을 주는 요소들에 대한 통찰력이 있다."
3. # Objective "직업 만족도 관점에서 설문결과 요약을 작성하고, 개선을 위한 액션플랜을 작성해."
   - 요약은 철저하고 상세하며 체계적으로 작성해
   - 액션플랜은 설문 결과에서 확인된 특정 문제를 해결하도록 맞춤화해야 되며, 구체적이고 즉시 실행이 가능해야 돼
   - 실행할 액션플랜을 5가지 제안해
4. 직업 만족도에 영향을 미치는 변수를 분석해 보라고 지시하니 상관분석을 실시함
5. 퇴직여부를 종속변수로 해서 예측모델을 만들어 보라고 하니 랜덤 포레스트 분류 분석
6. 퇴직 여부가 'No'인 직원들 중에서 퇴사 위험성이 높은 사람을 예측해 보라고 지시함

출처: 일잘러 장피엠 유튜브 https://www.youtube.com/watch?v=Ou9X4yu0KG8

재무제표 분석을 통한 투자판단 및 경영개선

투자자 관점에서 공시자료의 내역을 기반으로 기업의 안정성, 성장성, 활동성을 분석

- 안정성은 기업이 재정적으로 얼마나 안정적인지를 평가하는 지표
- 성장성은 기업의 매출, 영업이익, 순이익 등의 증가율을 통해 평가
- 활동성은 기업의 자산을 얼마나 효율적으로 활용하고 있는지를 평가. 주로 회전율로 나타남

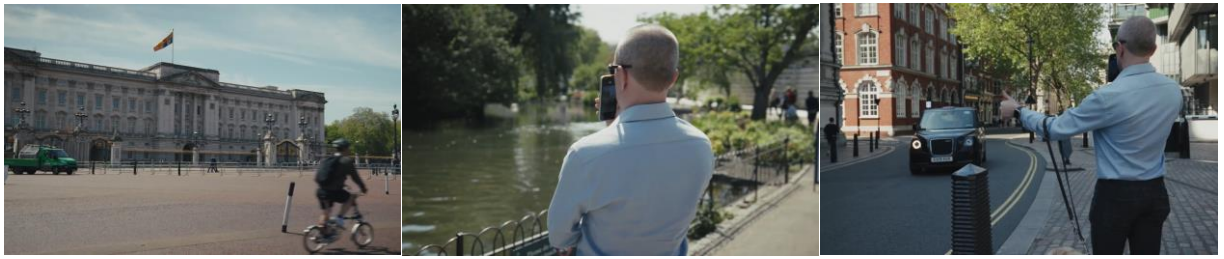
회사의 경영진 관점에서는 듀폰(DuPont) 분석을 통해 재무성과를 심층적으로 분석 가능

- 자기자본 이익률(ROE) 분해 순이익/자기자본 ➔ 순이익/매출액 * 매출액/총자산 * 총자산/자기자본

핵심 프롬프트	
	1. 전자공시시스템(https://dart.fss.or.kr/dsab007/main.do) 요약연결재무정보 복사+붙여넣기
	2. "투자자 관점에서 기업의 안정성, 성장성, 활동성을 분석해서 알려줘."
	3. "투자자 관점에서 개선의 필요성이 있는 점을 상세히 제시해"
	4. "듀폰 시스템 분석에 대해 설명해 봐 "
	5. 듀폰 시스템 분석을 통해 나타난 개선점을 영역별로 구분해 제시해 달라고 함
	6. 경영진 관점에서 개선해야 할 문제를 구체적으로 실행하기 위한 방안을 제시해 달라고 함

Appendix > 대폭 강화된 ChatGPT-4o의 음성인식기능

https://openai.com/index/hello-gpt-4o/



GPT-4o with Andy, from BeMyEyes in London

Customer service proof of concept

고생 많으셨습니다. 감사합니다.

홍용기 컨설팅학박사
010-3366-9010 / 123biz@naver.com