2026 이기적 빅데이터분석기사 필기
핵심 포인트 정리

1_1_1 빅데이터 개요 및 활용 ★

- 정량적 데이터(quantitative) : 주로 숫자로 이루어진 데이터(2025 년, 100km/h 등)
- 정성적 데이터(qualitative) : 문자와 같은 텍스트로 구성되며 함축적 의미를 지니고 있는 데이터(철수가 시험에 합격하였다.)

정형 데이터	정해진 형식과 구조에 맞게 저장되도록 구성된 데이터, 연산이 가능
반정형 데이터	데이터의 형식과 구조가 비교적 유연, 스키마 정보를 데이터와 함께 제공하는 파일 형식의 데이터, 연산이 불가능
비정형 데이터	구조가 정해지지 않은 대부분의 데이터, 연산이 불가능

- 암묵지 : 학습과 경험을 통하여 개인에게 체화되어 있지만 겉으로 드러나지 않는 지식
- 데이터베이스 : 동시에 복수의 적용 업무를 지원할 수 있도록 복수 이용자의 요구에 대응해서 데이터를 받아들이고 저장, 공급하기 위하여 일정한 구조에 따라서 편성된 데이터의 집합
- 빅데이터 : 기존보다 방대한 규모의 데이터, 새로운 통찰이나 가치를 추출할 수 있음

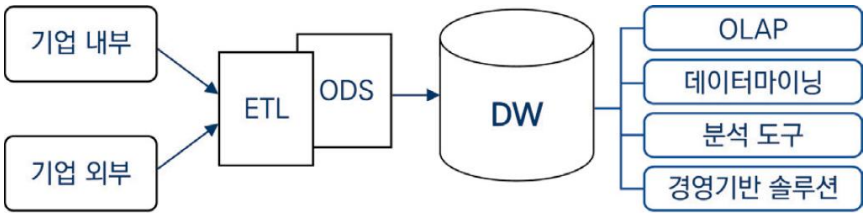
  → 규모(Volume), 유형(Variety), 속도(Velocity), 품질(Veracity), 가치(Value)의 3V+2V

  ▶ 빅데이터 활용을 위한 3 요소

자원	• 정형, 반정형, 비정형 데이터를 실시간으로 수집 • 수집된 데이터를 전처리 과정을 통해 품질 향상
기술	• 분산파일시스템을 통해 대용량 데이터를 분산 처리 • 데이터마이닝 등을 통해 데이터를 분석 및 시각화 • 데이터를 스스로 학습, 처리할 수 있는 AI 기술을 활용
인력	• 통계학, 수학, 컴퓨터공학, 경영학 등 전문지식 • 도메인 지식을 습득하여 데이터 분석 및 결과를 해석

- 데이터 웨어하우스(DW) : 의사결정에 도움을 주기 위해 기관시스템의 데이터베이스에 축적된 데이터를 공통의 형식으로 변환해서 관리하는 데이터베이스

  →주제지향성, 통합성, 시계열성, 비휘발성



- ETL(Extract, Transform, Load) : 기업의 내외부로부터 데이터를 추출, 정제 및 가공하여 DW 에 적재
- ODS(Operational Data Store) : 다양한 DBMS 에서 추출한 데이터를 통합적 관리
- 빅데이터 조직 구성 : 집중형(별도의 조직), 기능형(직접 수행), 분산형(전문 인력을 부서 배치)

Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

1_1_2 빅데이터 기술 및 제도 ★★★

▶ 빅데이터 처리과정

수집

- 크롤링(crawling) : 분산 저장되어 있는 문서를 수집하여 검색 대상의 색인으로 포함시키는 기술
- 로그 수집기, 센서 네트워크, RSS Reader/Open API, ETL 프로세스

저장

- NoSQL : 데이터 모델을 단순화하여 설계된 비관계형 데이터베이스

  → Hbase, MongoDB, Cassandra, Cloudata 등
- 공유 데이터 시스템, 병렬 데이터베이스 관리 시스템, 분산 파일 시스템, 네트워크 저장 시스템

처리

- 분산 병렬 컴퓨팅 : 다수의 독립된 컴퓨팅 자원을 네트워크상에 연결하여 이를 제어하는 미들웨어를 이용해 하나의 시스템으로 동작하게 하는 기술
- 맵리듀스(MapReduce) : 구글에서 개발한 프로그래밍 모델, 효과적 병렬 및 분산처리 지원
- 하둡 : 분산 처리 환경에서 대용량 데이터 처리 및 분석을 지원하는 오픈 소스 프레임워크

분석

- 탐구 요인 분석 : 데이터 간 상호 관계를 파악하여 데이터를 분석
- 확인 요인 분석 : 관찰된 변수들의 집합 요소 구조를 파악하기 위한 통계적 기법 활용

딥러닝

- 컴퓨터가 많은 데이터를 이용해 사람처럼 스스로 학습할 수 있도록 인공신경망 등의 기술을 이용
- 지도학습 : 학습 데이터로부터 하나의 함수를 유추, 분류와 회귀
- 비지도학습 : 입력값에 대한 목표치가 없음, 주요 특징을 발견하고 요약

전이학습

- 기존의 학습된 모델의 지식을 새로운 문제에 적용하여 학습을 빠르고 효율적으로 수행

개인정보

- 살아 있는 개인에 관한 정보, 다른 정보와 쉽게 결합하여 특정 개인을 알아볼 수 있는 정보
- 개인정보의 처리 위탁 : 개인정보 처리위탁을 받는 자, 처리위탁을 하는 업무의 내용을 알리고 동의를 받아야 한다.

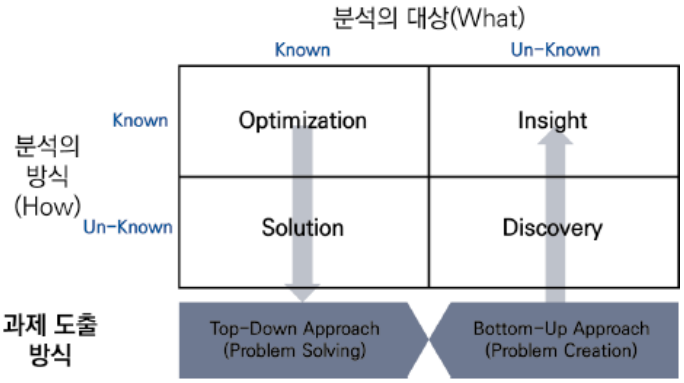
  → 단, 정보통신서비스 제공에 관한 계약을 이행하고 이용자 편의 증진 등을 위한 경우 고지절차와 동의절차를 거치지 않고, 이용자에게 이에 관해 알리거나 개인정보 처리방침 등에 공개할 수 있다.
- 개인정보 비식별화 : 개인을 식별할 수 있는 요소를 삭제하거나 대체 등의 방법으로 개인을 알아볼 수 없도록 하는 것(가명 처리, 총계 처리, 데이터 삭제, 데이터 범주화, 데이터 마스킹)
- GDPR : 유럽 의회에서 유럽 시민들의 개인정보 보호를 강화하기 위해 만든 통합 규정

Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

1_2_1 분석 방안 수립 ★★★

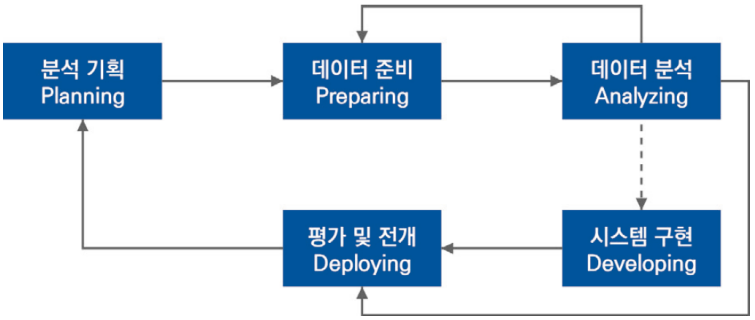
- 데이터 분석 : 데이터 집합으로부터 유용한 정보를 찾고 결과를 예측하기 위해 목적에 따라 분석기술과 방법론을 기반으로 대용량 데이터를 구축, 탐색, 분석하고 시각화를 수행
- 하향식 접근 : 문제가 주어지고 이에 대한 해법을 찾음
- 상향식 접근 : 문제의 정의 자체가 어려운 경우 데이터를 기반으로 문제의 재정의 및 해결방안 탐색

▶ 데이터 분석 주제 유형과 분석 과제 도출 방법



- 분석 방법론 : 데이터 분석의 효과적 수행과 품질 확보를 위해 분석 절차를 체계적으로 정리한 방법

▶ 빅데이터 분석 방법론의 개발절차



- KDD : 데이터셋 선택, 데이터 전처리, 데이터 변환, 데이터 마이닝, 데이터 마이닝 결과 평가
- CRISP-DM : 업무 이해, 데이터 이해, 데이터 준비, 모델링, 평가, 전개
- SEMMA : 추출, 탐색, 수정, 모델링, 평가
- 모델링

  → 기계학습 등을 이용한 데이터 모델링은 훈련용 데이터를 활용하여 분류, 예측, 군집 등의 모형을 만들어 가동중인 운영시스템에 적용 가능

  → 필요한 경우 비정형 데이터 분석결과를 활용하여 프로젝트 목적에 맞는 통합 모델링을 수행
- 데이터 거버넌스 : 전사 차원의 모든 데이터에 대하여 정책 및 지침, 운영조직과 책임 등의 표준화된 관리 체계를 수립하고 운영하기 위한 프레임워크와 저장소 구축

  → 데이터의 가용성, 유용성, 통합성, 보안성, 안전성을 확보
- 메타 데이터 : 다른 데이터를 설명하기 위해 사용되는 데이터
- 분석 성숙도 모델 : 비즈니스, 조직 및 역량, IT 부문 대상 실시

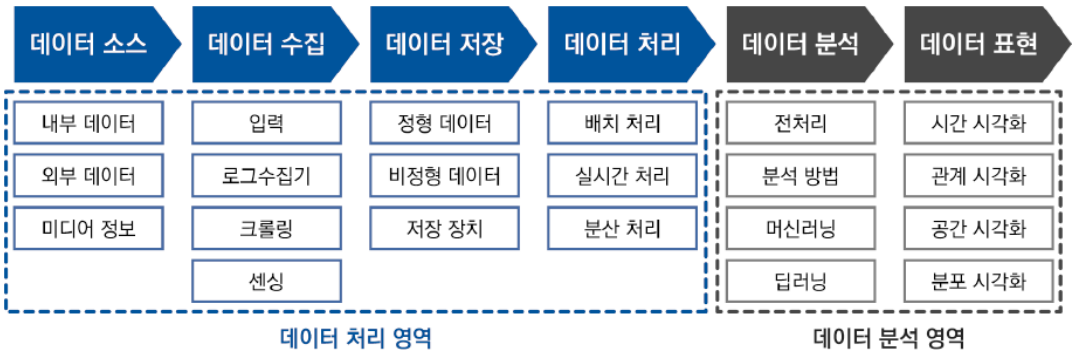
  → 도입, 활용, 확산, 최적화 단계로 구분

Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

1_2_2 분석 작업 계획 ★

- 데이터 처리 영역 : 데이터 분석을 위한 기초 데이터를 정의하고 수집 및 저장, 분석하기 수월하도록 물리적 환경을 제공
- 데이터 분석 영역 : 데이터를 추출하여 분석 목적과 방법에 맞게 가공한 후 데이터 분석을 직접 수행하고 결과를 표현

▶ 데이터 처리 프로세스



- 데이터 확보를 위한 사전 검토 : 필요 데이터의 정의, 보유 데이터의 현황파악, 분석 데이터의 유형, 편향되지 않고 충분한 양의 데이터 규모, 내부 데이터의 사용, 외부 데이터의 수집
- 데이터 전처리 수행 : 정제, 통합, 축소, 변환
- 데이터 품질 지표 : 정확성, 완전성, 적시성, 일관성

분석 절차

▶ 일반적 분석 절차

문제 인식	문제를 인식하고 분석 목적을 명확히 정의
연구조사	문제에 대한 해결방안, 중요한 요인이나 변수 파악
모형화	복잡한 문제를 논리적으로 단순화, 문제를 변수들 간 관계로 정의
데이터 수집	데이터 수집 또는 변수 측정
데이터 분석	수집된 데이터로부터 인사이트 발굴
분석 결과 제시	변수들 간 인과/상관 관계를 포함한 분석 결과 제시, 공유, 시각화

분석 프로젝트

▶ 중점 관리 영역

데이터 크기	데이터의 지속적 생성을 고려
데이터 복잡도	데이터 종류 고려, 다양한 시스템에 산재되어 있는 원천 데이터 통합 진행
속도	분석 모형의 성능과 속도를 고려한 개발과 테스트 수행
분석 모형 복잡도	분석 모형이 복잡할수록 정확도는 상승하나 해석이 어려워질 수 있음
정확도와 정밀도	분석 결과 활용에서는 정확도, 분석 모형 안정성 측면에서는 정밀도가 중요

Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

1_3_1 데이터 수집 및 전환 ★

▶ 데이터 수집 시스템 구축 절차

- 원천 데이터 정보 : 데이터의 수집 가능성, 보안, 정확성, 수집 난이도, 수집 비용

▶ 수집 데이터 구분

내부 데이터	• 조직 내부의 서비스 시스템, 네트워크 및 서버 장비, 마케팅 관련 시스템 등으로부터 생성 • 분석에 적합한 정형화된 형식으로 수집
외부 데이터	• 다양한 소셜 데이터, 특정 기관 데이터, M2M 데이터, LOD 등 • 분석 목표에 맞게 수집 데이터를 변환하는 노력이 필요

- 아파치 스쿱 : 관계형 데이터스토어 간 대량 데이터를 효과적으로 전송하는 도구
- 아파치 플럼 : 대용량의 로그 데이터를 효과적으로 수집, 집계, 이동시키는 분산 서비스 제공 솔루션
- 스크래피 : 웹사이트를 크롤링하고 구조화된 데이터를 수집하는 도구, 파이썬으로 작성됨
- 데이터 확보 비용 산정 요소 : 데이터의 종류, 크기 및 보관 주기, 수집 주기와 방식·기술, 가치성
- 데이터 저장 방식 : 파일 시스템, 관계형 데이터베이스, 분산처리 데이터베이스

▶ 프라이버시 보호 모델

k-익명성	특정인임을 추론할 수 있는지 검토, 일정 수준 이상 비식별 되도록 함	동일한 값을 가진 레코드를 k개 이상으로 하며, 특정 개인을 식별할 확률은 $1/k$
l-다양성	특정인 추론이 안된다고 해도 민감한 정보의 다양성을 높여 추론 가능성을 낮춤	각 레코드는 최소 l개 이상의 다양성을 가지도록 하여 동질성 또는 배경지식 등에 의한 추론 방지
t-근접성	민감한 정보의 분포를 낮추어 추론 가능성을 더욱 낮춤	전체 데이터 집합의 정보 분포와 특정 정보의 분포 차이를 t 이하로 하여 추론 방지

- 식별자 : 개인을 고유하게 식별할 수 있는 정보
- 속성자 : 개인에 대한 추가적인 정보
- 비식별 조치는 가명처리, 총계처리, 데이터 삭제, 데이터 범주화, 데이터 마스킹 등 여러 기법을 단독 또는 복합적으로 활용

Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

- 데이터 품질 관리 : 비즈니스 목표에 부합한 데이터 분석을 위해 가치성, 정확성, 유용성 확보

▶ 정형 데이터 품질 기준

완전성	필수항목에 누락이 없어야 함
유일성	데이터 항목은 유일해야 하며 중복 불가
유효성	정해진 데이터 유효범위 및 도메인을 충족
일관성	구조, 값, 표현되는 형태가 일관되게 정의
정확성	현실에 존재하는 객체의 표현 값이 정확히 반영

▶ 비정형 데이터 품질 기준

기능성	해당 컨텐츠가 특정 조건에서 사용될 때 요구를 만족하는 기능 제공 정도
신뢰성	규정된 신뢰 수준을 유지, 사용자로 하여금 오류를 방지할 수 있는 정도
사용성	사용자에 의해 이해되고 선호되는 정도
효율성	사용되는 자원의 양에 따라 요구된 성능을 제공하는 정도
이식성	다양한 환경과 상황에서 실행될 가능성

1_3_2 데이터 적재 및 저장 ★

- 데이터 적재 : 데이터의 유형과 실시간 처리 여부에 따라 구분

  → 데이터 수집 도구 이용, NoSQL DBMS 가 제공하는 도구를 이용, 관계형 DBMS 의 데이터를 NoSQL DBMS 에서 적재

- 데이터 저장 : 파일 시스템 저장방식, 데이터베이스 저장방식

▶ 데이터 모델에 따른 NoSQL 데이터베이스 분류

key-value DB	• 데이터를 키와 그에 해당하는 값의 쌍으로 저장하는 모델에 기반 • 관계형 데이터베이스보다 확장성이 뛰어나고 질의 응답 시간이 빠름
column-oriented DB	• 데이터를 로우가 아닌 칼럼 기반으로 저장하고 처리
document DB	• 문서 형식의 정보를 저장, 검색, 관리하기 위한 데이터베이스 • key-value 데이터베이스보다 문서의 내부구조에 기반을 둔 복잡한 형태의 데이터 저장을 지원

- 빅데이터 저장시스템 선정을 위한 분석 : 기능성 비교분석, 분석방식 및 환경, 분석대상 데이터 유형, 기존 시스템과의 연계
- 스트리밍 데이터 : 빠르고 연속적, 대용량, 다양한 장소에서 발생

  → 네트워크 모니터링 데이터, IoT 에서 발생하는 센서 데이터, 통신 데이터, 웹 로그 등

  로그 : 컴퓨터의 처리 내용이나 이용 상황을 시간의 흐름에 따라 기록한 것

Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

2_1_1 데이터 정제 ★

- 데이터 정제 과정 : 다양한 매체로부터 데이터 수집, 원하는 형태로 변환, 원하는 장소에 저장, 활용가능성을 타진하기 위한 품질확인, 사용이 원활하도록 관리

  → 비정형 데이터의 경우 구조화된 정형 데이터로 변환을 수행, 결측치와 오류 수정 과정 필요

전처리	데이터 저장 전 처리, 대상 데이터와 입수 방법 결정, 저장 방식·장소 선정
후처리	저장 후의 처리, 저장 데이터의 품질관리 등의 과정 포함

- 결측치(missing data, 손실 데이터) : 어떠한 자료값도 관측 대상 변수에 저장되지 않을 때 발생

완전 무작위 결측	어떤 변수상에서 결측 데이터가 다른 변수와 아무런 연관이 없는 경우
무작위 결측	결측 데이터가 관측된 다른 변수와 연관이 있지만 비관측값들과는 연관되지 않는 경우, 결측이 완전히 설명될 수 있음
비 무작위 결측	결측 데이터가 다른 변수와 연관있는 경우

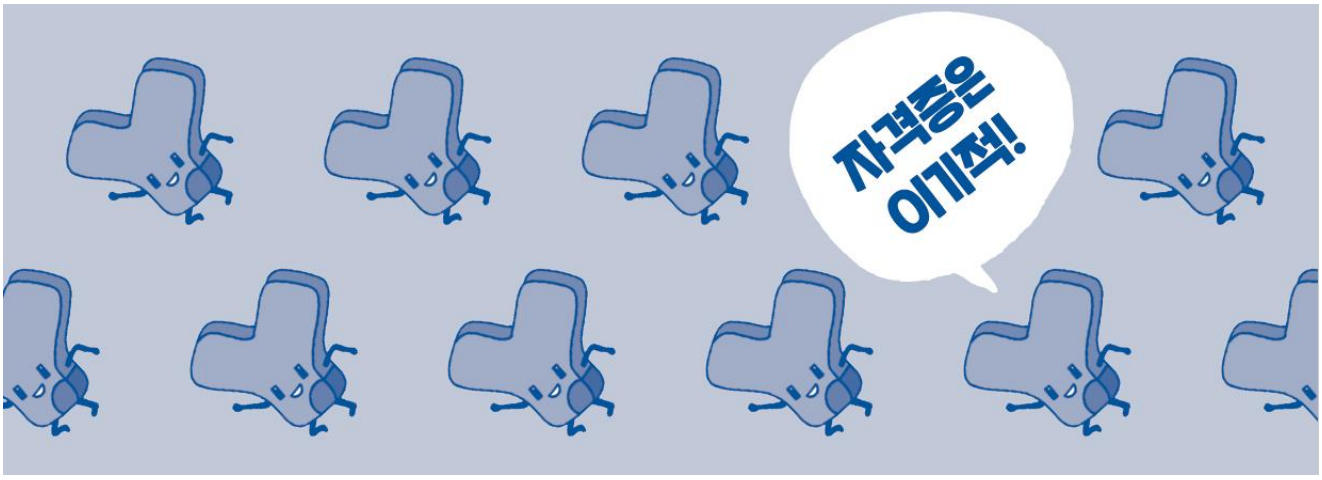
- 단순 대치법 : 결측치를 완전 무작위 결측 또는 무작위 결측으로 판단하고 처리

  → 완전 분석법, 평균 대치법, 회귀 대치법, 단순확률 대치법, 최근접 대치법
- 다중 대치법 : 단순 대치법을 복수로 시행, 통계적 효율성 및 일치성 문제를 보완

- 이상치(outlier) : 정상의 범주(데이터의 전체적 패턴)에서 벗어난 값

  → 이상치가 비무작위로 나타나면 데이터의 정상성 감소를 초래하고 이는 신뢰성 저하로 연결
- 이상치 탐지 : 시각화(상자그림, 줄기-잎 그림, 산점도 그림), Z-Score, 밀도기반 클러스터링, 고립 의사나무 방법

- 모수(parameter) : 모집단(전체 집단)의 모평균, 모표준편차, 모분산 등
- 모수적 방법 : 정규분포를 따른다는 가정으로 모수적 특성을 이용하는 통계적 방법
- 비모수적 방법 : 정규분포임을 가정할 수 없을 때 사용하는 방법



Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

2_1_2 분석 변수 처리 ★★

- 회귀(regression) : 변수 x 와 y 의 관계를 함수식으로 설명하는 통계적 방법
- 변수 선택 : 변수는 기본적으로 많을수록 신뢰성이 높아지나 더 작은 변수를 사용 시 동일한 설명력이 나온다면 효율성이 증가

전진 선택법	종속변수와 단순상관계수의 절댓값이 가장 큰 변수를 분석모형에 포함시키는 것, 한번 추가된 변수는 제거하지 않음
후진 소거법	종속변수와 단순상관계수의 절댓값이 가장 작은 변수를 분석모형에서 제외시키는 것, 한번 제거된 변수는 추가하지 않음
단계적 선택법	전진 선택법을 통해 가장 유의한 변수를 모형에 포함 후 나머지 변수들에 대해 후진 선택법을 적용하여 새롭게 유의하지 않은 변수들 제거

- 차원 축소 : 변수(데이터의 종류)의 양을 줄이는 것, 복잡도를 축소하여 분석시간과 저장변수의 양을 효율적으로 줄임, 과적합 발생 가능성을 줄여 정확도 저하 방지, 이해와 해석 용이
- 차원의 저주 : 데이터 분석 및 알고리즘을 통한 학습을 위해 차원이 증가하면서 학습데이터의 수가 차원의 수보다 적어져 성능이 저하되는 현상

  → 차원을 줄이거나 데이터 수를 늘려 해결
- 요인 분석 : 다수의 변수들 간 관계를 분석하여 공통 차원을 축약, 독립변수/종속변수 개념이 없음

주성분 분석	분포된 데이터들의 특성을 설명할 수 있는 하나 또는 복수 개의 특징을 찾는 것
특이값 분해	적당한 특이값을 이용해 원래 데이터와 비슷한 정보력을 가지는 차원을 만들어 냄
음수 미포함 행렬분해	음수를 포함하지 않은 행렬 V를 행렬 W와 H의 곱으로 분해, 행렬 곱셈에서 V보다 W, H가 적은 차원을 가짐, 정확한 해가 없으므로 대략적 해를 구함

▶ 분석 목표에 적합한 데이터 형태로 보완

파생변수	특정 조건을 만족하거나 특정 함수에 의해 값을 만들어 의미를 부여
요약변수	수집된 정보를 분석에 맞게 종합한 변수

- 변수 변환 : 데이터를 분석하기 좋은 형태로 바꾸는 작업, 어떤 변수로 나타낸 식을 다른 변수로 바꿔 나타냄, 데이터 전처리 과정 중 하나로 간주

  → 범주형 변환, 정규화(일반, 최소-최대, Z-Score), 로그 변환, 역수 변환. 지수 변환, 제곱근 변환 등

- 클래스 불균형 : 각 클래스가 갖고 있는 데이터의 양에 차이가 큰 경우

▶ 비대칭 데이터의 정밀도 향상

언더샘플링	대표클래스의 일부만을 선택하고, 소수클래스는 최대한 많은 데이터를 사용
오버샘플링	소수클래스의 복사본을 만들어 대표클래스만큼 데이터를 만드는 방법

Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

2_2_1 데이터 탐색의 기초 ★★

- 탐색적 데이터 분석(EDA) : 본격적 데이터 분석 전에 자료를 직관적인 방법으로 통찰하는 과정
  → 내재된 잠재적 문제에 대해 인식하고 해결안을 도출, 새로운 양상·패턴 발견 가능
- 상관분석 : 2 개 이상의 양적 변수 간의 관계가 유의한지 확인

▶ 상관분석의 기본가정

선형성	두 변인 X와 Y의 관계가 직선적인지 알아보는 것
동변량성	X 값에 관계없이 Y의 흩어진 정도가 같은 것 ↔ 이분산성
정규분포성	두 변인의 측정치 분포가 모집단에서 모두 정규분포를 이루는 것
무선독립표본	모집단에서 표본을 뽑을 때 표본 대상이 확률적으로 선정되는 것

▶ 상관분석 방법

피어슨 상관계수	+1과 -1 사이의 값, +1은 완벽한 양의 선형 상관관계, 0은 선형 상관관계 없음, -1은 완벽한 음의 선형 상관관계
스피어만 상관계수	자료의 값 대신 순위를 매기는 경우의 상관계수

기초통계량

산술평균	모든 자료들을 합한 후 전체 자료수로 나누어 계산하는 일반적 평균 • 모평균 μ – 모집단 전체의 산술평균 • 표본평균 $\bar{X}$ – 모집단의 부분집합인 추출된 표본의 산술평균
기하평균	n개의 자료에 대해 관측치를 곱한 후 제곱근 표현 $=\sqrt[n]{x_1 \times x_2 \times x_3 \cdots \times x_n}$
중앙값(median)	자료를 크기 순으로 나열할 때 가운데에 위치한 값
최빈값(mode)	가장 노출 빈도가 높은 자료, 좌로 치우친 그래프에서 제일 작음
분위수(quantile)	자료의 위치를 표현하는 수치

▶ 분산과 표준편차

분산	• 평균을 중심으로 밀집되거나 퍼짐 정도를 나타내는 척도 • 각각의 자료값과 평균과의 편차의 제곱을 이용하여 표현
표준편차	분산의 제곱근, 분산으로 얻은 수치를 해석하기 곤란하다는 단점 보완

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \left(\text{모분산}\right), \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \,(\text{표본분산})$$

- 평균 절대 편차 : 관측값에서 평균을 빼고, 절댓값을 취하여 산술평균
- 사분위범위 : 자료를 크기 순으로 배열 후 3사분위수(Q3) – 1사분위수(Q1)로 정의
- 왜도 : 분포가 어느 한쪽으로 치우친 정도를 나타내는 통계적 척도
- 첨도 : 분포의 뾰족한 정도를 나타내는 통계적 척도
- 상자수염그림(Box Plot) : 수치적 자료 표현, 자료로부터 얻어 낸 통계량(최솟값, Q1, Q2, Q3, 최댓값)을 가지고 그림, 이상치는 파악 가능하나 분산과 같은 퍼짐정도는 파악 어려움

Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

2_2_2 고급 데이터 탐색 ★

- 시공간 데이터 : 공간적 정보에 시간의 흐름이 결합된 다차원 데이터
- 시간 데이터 : 데이터에 유효 시간, 거래 시간, 사용자 정의 시간과 같은 연관된 시간 표현 정의
- 공간 데이터 : 래스터, 벡터 공간, 기하학, 위상적 타입 등 정의

▶ 공간 데이터 모델

관계형 모델	정적 모델, 표현이 유연하지 못해 실세계 공간 객체의 특징 표현에 한계
객체지향 모델	비 구조적, 자연스런 표현, 연산과 함수 확장이 쉬움, 무결성 검사 용이

▶ 시공간자료 질의어

시공간자료 정의언어	• 시공간 테이블 인덱스 및 뷰의 정의문, 변경문 등 • 공간적, 시간적 속성을 동시 포함
시공간자료 조작언어	• 객체의 삽입, 삭제, 변경 등의 검색문 • 시간지원, 공간 연산자를 포함, 공간관리와 이력정보 제공

- 다변량 데이터 탐색 : 변수들 간 인과관계의 규명과 분석

종속변수와 독립변수 간 인과관계	변수축약	개체유도
다중회귀, 로지스틱 회귀, 분산분석	주성분분석, 요인분석, 정준상관분석	군집분석, 다차원 척도법, 판별분석

- 변수축약 : 변수들 간 상관관계를 이용하여 변수를 줄이는 방법, 변수유도기법
- 개체유도 : 개체들의 특성을 측정한 변수들의 상관관계를 이용하여 유사한 개체를 분류하는 방법

비정형 데이터의 분석

데이터 마이닝	• 체계적이고 자동적으로 통계적 규칙이나 패턴을 분석하여 가치 있는 정보를 추출하는 과정 • 자료에 의존하여 현상을 해석하므로 자료가 현실을 충분히 반영하지 못한 상태인 경우 잘못된 모형을 구축하는 오류를 범할 수 있음
텍스트 마이닝	인간의 언어로 이루어진 비정형 텍스트 데이터들을 자연어 처리방식을 이용하여 숨겨진 의미를 발견하는 기법
오피니언 마이닝	• 텍스트 마이닝의 한 분류, 특정 주제에 대한 사람들의 주관적 의견을 통계·수치화해 객관적 정보로 바꾸는 기술 • 텍스트 마이닝은 문장 내 주제 파악, 오피니언 마이닝은 감정·태도 판별
웹 마이닝	웹으로 통한 모든 것(기기 내 쌓이는 로그, 사용자 행동 및 작성 콘텐츠 등)을 분석하여 유용한 정보를 추출하는 것

- 자연어 처리 : 인간의 언어 현상을 컴퓨터를 이용하여 모사할 수 있도록 연구하고 구현하는 인공지능 분야

Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

2_3_1 기술통계 ★★

▶ 확률 표본추출 기법

단순무작위 추출	가장 기본이 되는 표본추출
계통 추출	모집단에서 추출간격을 설정하여 간격 사이에서 무작위 추출
층화 추출	모집단을 서로 겹치지 않게 분할된 층별로 임의 추출
군집 추출	모집단을 차이가 없는 여러 개 군집으로 나누어 군집의 단위의 일부 또는 전체에 대한 분석을 시행

- 조건부 확률 : 사건 B 가 일어났다는 조건하에 다른 사건 A 가 일어날 확률 $P(A|B) = P(A \cap B)/P(B)$
- 결합 확률 : 사건 A 와 B 가 동시에 발생하는 확률, 확률의 곱셈 법칙 $P(A) \times P(B) = P(A \cap B)$
- 베이지안 정리 : 사전에 사건 A 에 대한 사전확률이 부여된 상태에서 사건 B 에 관한 정보를 종합하여 사건 A 에 대한 사후확률을 정리

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{k} P(B|A_i)P(A_i)}$$

- 확률변수 : 사건 시행의 결과(확률)를 하나의 수치로 대응시킬 때의 값, 확률값
- 확률분포 : 수치로 대응된 확률변수의 개별 값들이 가지는 확률값의 분포, 확률변수가 취할 수 있는 값의 수가 유한하면 이산확률분포, 무한하면 연속확률분포

▶ 이산확률분포

베르누이분포	결과가 성공 아니면 실패, 두 가지로 귀결 $f(x) = p^x q^{1-x}$
이항분포	베르누이시행을 n번 독립적으로 시행할 때 성공횟수를 X로 정의 $f(x) = \binom{n}{x} p^x q^{n-x}$
다항분포	여러 번의 독립적 시행에서 각각의 값이 특정 횟수가 나타날 확률을 정의 $f(x_1, x_2, \dots x_k;\ n, p_1, p_2, \dots, p_k) = \frac{n!}{x_1!\, x_2! \dots x_k!}\ {p_1}^{x_1}\, {p_2}^{x_2} \dots {p_k}^{x_k}$
포아송분포	단위 시간 안에 어떤 사건이 몇 번 발생할 것인지를 표현 $f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$
기하분포	베르누이 시행에서 처음 성공까지 시도한 횟수를 분포화 $f(x) = pq^{x-1}, (q = 1 - p)$
음이항분포	x번의 베르누이 시행에서 k번째 성공할 때까지 계속 시행에서 확률 $f(x) = \binom{x-1}{k-1} q^{x-k} p^k,\ x = k, k+1, \cdots$
초기하분포	비복원 추출에서 n개를 추출했을 때, 원하는 것 k개가 뽑힐 확률을 표현 $f_X(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$

Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

▶ 연속확률분포

연속균등분포	분포가 특정 범위 내에서 균등하게 나타나 있을 경우 $f(x) = \frac{1}{b-a}$
지수분포	포아송과정에서 한 개의 사건이 발생할 때까지 대기 시간 $f(x) = \frac{1}{\beta} e^{-x/\beta}$
정규분포	평균을 중심으로 대칭, 종모양, 모양과 위치는 평균과 표준편차에 의해 결정, 정규곡선과 X축 사이의 전체 면적은 1 $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
표준정규분포	평균 $\mu = 0$, 표준편차 $\sigma = 1$이 되도록 한 정규분포 $Z = \frac{X-\mu}{\sigma}$
감마분포	포아송과정에서 k개의 사건이 발생할 때까지의 대기시간 $f(x,k,\theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta}$
카이제곱분포	k개의 서로 독립인 표준정규확률변수를 각각 제곱 후 합해서 얻어지는 분포 $f(x;k) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-x/2}$
스튜던트 t 분포	정규분포의 평균 측정 시 주로 사용, 종모양으로 t=0에 대하여 대칭 $f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)}$
F 분포	두 개의 확률 변수 V_1, V_2의 자유도가 각 k_1, k_2이고 카이제곱분포를 따를 때 $F = \frac{V_1/k_1}{V_2/k_2} \sim F(k_1, k_2)$ $f(x) = \frac{1}{B(d_1/2, d_2/2)} \left(\frac{d_1 x}{d_1 x + d_2}\right)^{\frac{d_1}{2}} \left(1 - \frac{d_1 x}{d_1 x + d_2}\right)^{\frac{d_2}{2}} x^{-1}$ 여기서 B는 베타함수로 $B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1} dt$

- 자유도 : 자료집단의 변수 중에서 자유롭게 선택될 수 있는 변수의 수
- 표본분포 : 크기 n 의 확률표본(모집단에서 동등한 확률로 추출된 개체들의 집합)의 확률변수의 분포
- 표준오차 : 표본평균의 표준편차

  → 모집단의 크기가 무한 : $\sigma/\sqrt{n}$, 모집단의 크기가 유한 : $\sqrt{\frac{N-n}{N-1}} \cdot \frac{\sigma}{\sqrt{n}}$ (N: 모집단크기, n: 표본크기)
- 중심극한정리 : 모집단의 분포에 상관없이 표본의 수가 큰 표본분포들의 표본평균의 분포는 정규분포를 이룸
- 표본비율 : 표본을 구성하는 n 개의 개체 중에서 성공으로 나타나는 개체 수의 비율

Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

2_3_2 추론통계 ★★

- 점추정 : 모수 즉 모평균이나 모표준편차 등과 같은 추정치를 이에 대응하는 통계량으로 추정

▶ 모수와 추정량 정리

모수	추정량
모평균(μ)에 대한 점추정	표본집단의 표본평균 $\bar{x} = \frac{1}{k}\sum_{i=1}^{k} x_i$
모분산(σ^2)에 대한 점추정	표본집단의 표본분산 $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
모비율에 대한 점추정	$\hat{P} = \frac{X}{n}$ X: 표본 중에 성공으로 나타난 개체수, n: 표본의 개체수

- 편향 : 기대하는 추정량과 모수의 차이, 편향이 0 이 되면 불편추정량
- 최대우도추정량 : 표본을 얻을 확률이 가장 높은, 즉 주어진 관찰값을 가장 잘 설명해 주는 $\hat{\theta}$

- 구간추정 : 점추정에 오차의 개념을 도입하여 모수가 포함되는 확률변수구간을 어떤 신뢰성 아래 추정하는 것

▶ 모평균 신뢰구간 정리

구분	신뢰구간
모집단의 분산을 아는 경우	$\bar{X} - Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$
모집단의 분산을 모르는 경우 (표본크기가 작음)	$\bar{X} - t_{\frac{\alpha}{2},n-1} \cdot \frac{S}{\sqrt{n}} \le \mu \le \bar{X} + t_{\frac{\alpha}{2},n-1} \cdot \frac{S}{\sqrt{n}}$
모집단의 분산을 모르는 경우 (표본크기가 큼)	$\bar{X} - Z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}} \le \mu \le \bar{X} + Z_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}$

▶ 모분산, 모비율 신뢰구간

모분산 신뢰구간	모비율 신뢰구간
$\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}$	$\hat{p} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

- 가설검정 : 검정통계량의 표본분포에 따라 채택여부를 결정짓는 통계적 분석과정

Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

- 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

실제상황 / 검정결과	H_0	H_1
H_0 채택	success	Type 2 Error
H_0 기각	Type 1 Error	success

- 유의수준 : 귀무가설이 맞는데 틀렸다 결론 내리게 될 확률, 제 1 종 오류를 범할 확률의 최대 허용한계
- *p*-value : 귀무가설을 기각하려고 할 때 필요한 최소의 유의수준

▶ 집단크기에 따른 검정통계량 설정

대표본 또는 모집단이 정규분포	정규분포 따르면서 소표본
$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$	$T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$

▶ 표본에 따른 검정통계량 설정

두 독립표본의 평균차이 검정	대응표본의 평균차이 검정
$T = \frac{\bar{X}-\bar{Y}}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}}$	$T = \frac{D}{S_D/\sqrt{n}}$
$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$	$S_D^2 = \frac{\sum(D_i-\bar{D})^2}{(n-1)}$

단일표본 모분산에 대한 가설검정(χ^2 검정)	두 모분산비에 대한 가설검정(F 검정)
$\chi^2 = \frac{\sum(x_i-\bar{x})^2}{\sigma_0^2} = \frac{\phi s^2}{\sigma_0^2}$	$F = s_1^2/s_2^2$

Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

3_1_1 분석 절차 수립 ★

- 분석 모형 : 분석 목표에 따라 데이터 특성을 도출, 가설 수립에 다라 전체적 분석 방향을 정의
  → 예측 분석 모형, 현황 진단 모형, 최적화 분석 모형 등
- 분석 모형 선정 프로세스 : 문제요건 정의 – 데이터 수집·정리 – 데이터 전처리 – 분석 모형 선정

- 상향식 접근 : 특정 영역을 지정하여 의사결정 지점으로 진행하는 과정에서 분석 기회를 발굴
- 분석 유스케이스 기반 접근 : 분석 테마 후보 그룹(pool)을 활용하고 연관된 분석 기회를 발굴

분석 모형 구축 절차

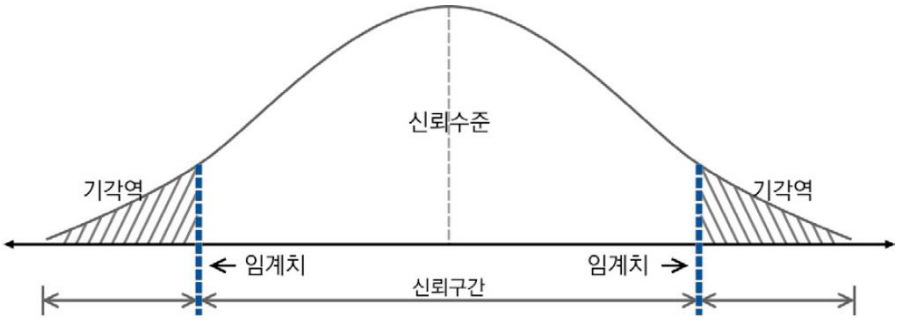
- 분석 시나리오 작성 : 데이터 분석 대상 및 범위를 요구사항에 맞게 정의, 해결할 수 있는 문제와 목표, 목표별 구현 모델과 예상 결과 작성
- 분석 모형 설계 : 분석 대상 및 범위를 정하여 분석 목적 구현을 위한 분석 방법론 설계

분석 모델링 설계와 검정 – 분석 목적에 기반한 가설검정 방법

① 유의수준 결정, 귀무가설과 대립가설 설정

② 검정통계량(가설을 검정하기 위한 기준으로 사용하는 값)의 설정

③ 기각역 설정



④ 검정통계량 계산

$$\text{검정통계량} = \frac{(\textit{표본평균} - \textit{모평균})}{(\textit{표본 표준편차})}$$

⑤ 통계적인 의사결정(가설검정)

양측검정	• 귀무가설을 기각하는 영역이 양쪽에 있는 검정 • 대립가설이 ~가 아니다(크거나 작다)인 경우 사용
단측검정	• 귀무가설을 기각하는 영역이 한쪽 끝에 있는 검정 • 대립가설이 ~보다 작다 혹은 크다인 경우 사용

분석 모델링 설계와 검정 – 추정 방법에 대한 기술 검토

- 데이터 전처리 과정을 거치며 모형에 활용될 후보 변수와 후보 분석 모형에 사용할 알고리즘 파악

- 분석 모형 선정 문제 : 비즈니스 환경 여건, 종속 변수 유무에 따라 달라짐
  → 종속 변수가 없으면 사용가능 알고리즘이 군집과 원인 분석, 이상치, 연관 법칙 등으로 제한

Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

3_1_2 분석 환경 구축 ★

▶ R 과 Python

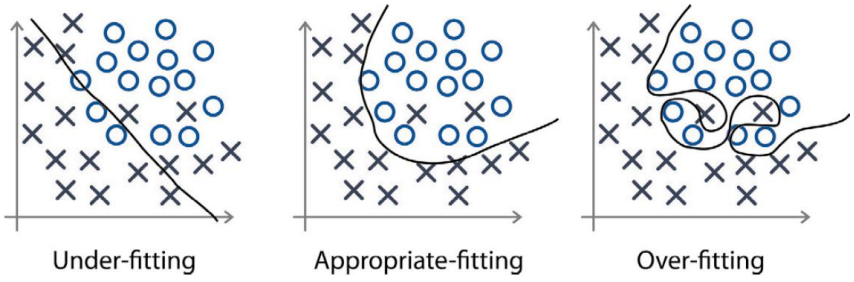
R	• 객체지향 언어, 고속메모리 처리, 다양한 최신 패키지 제공 • 벡터, 배열, 행렬, 데이터프레임, 리스트 등 다양한 자료구조와 연산기능 • 데이터 분석과 표현을 위한 다양한 그래픽 도구 제공, 시각화 특화 단점: 대용량 메모리 처리가 어렵고 보안에 취약, 별도의 모듈 연동이 아니면 웹 브라우저에서 사용할 수 없음
Python	• 인터프리터 언어, 컴파일, 실행, 테스트 용이 • 동적으로 데이터타입 결정 • 플랫폼 독립적, 컴파일 없이 동작 실행 • 리스트, 사전, 튜플 등 유연한 내장 객체 자료형 지원, 메모리 자동할당 뒤 종료 시 자동해지하는 메모리 청소 기능 제공 단점: 인터프리터 방식은 비교적 실행속도가 느린 단점을 가짐

- 인터프리터 : 프로그래밍 소스 코드를 바로 실행하는 환경, 원시 코드를 기계어로 번역하는 컴파일러와 대비

- 데이터 분할 : 분석용 데이터로 모형을 구축하여 평가 및 검증하기 위해 전체 데이터를 분할

학습 데이터	데이터를 학습하여 분석 모형을 만드는 데에 직접 사용되는 데이터
평가 데이터	추정한 분석모델이 과대·과소적합인지 모형의 성능을 평가하기 위한 데이터
테스트 데이터	최종적으로 일반화된 분석 모형을 검증하는 테스트를 위한 데이터

- 과대적합(과적합) : 학습 데이터에 대해서는 높은 정확도를 나타내지만 테스트 데이터나 새로운 데이터에 대해서는 예측을 잘 하지 못하는 것
- 과소적합 : 모형이 단순하여 데이터 내부의 패턴 또는 규칙을 잘 학습하지 못하는 것
- 일반화 : 학습 데이터를 통해 생성된 모델이 평가 데이터를 통한 성능 평가 외에도 테스트 데이터를 통해 정확하게 예측하는 것



Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

3_2_1 분석기법 ★★★

▶ 학습 유형에 따른 데이터 분석 모델

지도학습	주어진 데이터에 대해 정답을 부여하고 동일한 정답이 나오도록 분류 또는 새로운 데이터의 정답을 예측하도록 학습 • 분류 : 의사결정나무, 랜덤 포레스트, 인공신경망, SVM, 로지스틱 회귀분석 • 회귀(예측) : 의사결정나무, 선형 회귀분석, 다중 회귀분석
비지도학습	정답없이 컴퓨터 스스로 입력 데이터의 패턴을 찾아내고 구조화 • 군집 분석, 연관성 분석, 인공신경망, 오토인코더
준지도학습	효율적 학습을 위해 목표값이 표시된 데이터와 그렇지 않은 데이터를 모두 학습에 사용함으로써 주어진 데이터 특징을 표현하는 잠재변수를 찾음 • 셀프 트레이닝, GAN
강화학습	주어진 환경에서 보상을 최대화하도록 에이전트를 학습 • Q-Learning, 정책경사(PG)

- 회귀분석 : 원인과 결과의 연관을 분석, 예측이나 분류에 사용

선형 회귀분석	통계적 의미로 종속변수 y와 한 개 이상의 독립변수 x와의 선형상관성 파악
로지스틱 회귀분석	• 종속변수가 이항형(유효한 범주의 개수가 두 개)일 때 사용 • 신용 평가에 많이 사용

- 의사결정나무 : 의사결정 규칙을 나무 모양으로 조합하여 목표 변수에 대한 분류 또는 예측을 수행

  → 부모마디보다 자식마디의 순수도(purity) 증가, 불확실성은 감소하도록 분리 진행(정보 획득)

의사결정나무의 분석 과정

① 변수 선택 : 목표변수와 관련된 설명(독립) 변수들을 선택

② 의사결정나무 형성 : 분석목적에 따라 적절히 훈련데이터를 활용

③ 가지치기 : 부적절한 추론규칙을 가지거나 불필요 또는 분류오류를 크게 할 위험이 있는 마디 제거

④ 타당성 평가 : 이익, 비용, 위험 등을 고려하여 모형을 평가

⑤ 해석 및 예측 : 최종 모형에 대한 해석으로 분류 및 예측 모델을 결정

- 랜덤 포레스트 : 부트스트래핑 기반 샘플링을 활용한 의사결정나무 생성 이후 배깅 기반 나무들을 모아 앙상블 학습하여 숲을 형성
- 부트스트래핑 : 단순 복원 임의추출법(랜덤 샘플링)으로 크기가 동일한 여러 개의 표본자료 생성
- 배깅 : 여러 부트스트랩 자료를 생성하여 학습하는 모델링
- 부스팅 : 가중치를 활용하여 약분류기를 강분류기로 만드는 방법
- 앙상블 학습 : 여러 모델을 학습시켜 결합

인공신경망

- 가중치 : 노드와의 연결계수
- 학습 : 가중치와 편향을 훈련 데이터에 적응하도록 조정하는 과정

  → 1 단계: 미니배치 – 2 단계: 가중치 매개변수 기울기 산출 – 3 단계: 매개변수 갱신

Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

- 오차역전파 : 가중치 매개변수 기울기를 미분을 통해 진행하지 않고 오차를 출력층에서 입력층으로 전달, 연쇄법칙을 활용한 역전파를 통해 가중치와 편향을 계산
- 과대적합 해결방안

  1. 가중치 감소 : 가중치가 클수록 일종의 패널티를 부과하여 가중치 매개변수 절대값을 감소시킴

  → 패널티 역할로 L1 규제(라쏘), L2 규제(릿지)

  2. 드롭아웃 : 은닉층의 뉴런을 임의로 삭제하면서 학습하는 방법, 적은 뉴런만으로 훈련한 뒤 테스트 시 전체 뉴런을 사용하면 정답을 더 잘 찾음

  3. 초매개변수 최적화 : 수동으로 변수들을 설정하여 과적합 방지

▶ 활성 함수

Sigmoid	참에 가까워지면 0.5~1, 거짓에 가까워지면 0~0.5 사이의 값으로 출력
Relu	0보다 크면 입력값을 그대로 출력, 0이하의 값만 0으로 출력

▶ 딥러닝 모델 종류

합성곱 신경망 (CNN)	• 사람의 시신경 구조 모방, 모든 입력 데이터들을 동등한 뉴런으로 처리 • 데이터의 특징, 차원을 추출하여 패턴을 이해하는 방식, 이미지의 특징을 추출하는 과정(합성곱 계층, 풀링 계층)과 클래스를 분류하는 과정으로 진행
순환 신경망 (RNN)	• 순서를 가진 데이터를 입력하여 단위 간 연결이 시퀀스를 따라 방향성 그래프를 형성하는 모델 • 필기나 음성 인식과 같이 시변적 특징을 지니는 데이터 처리에 적용
LSTM	• 점차 데이터가 소멸되는 RNN의 단점을 보완 • 보통 신경망 대비 4배 이상 파라미터를 보유, 여러 단계를 거쳐도 오랜 시간동안 데이터를 잘 기억
오토인코더	입력으로 들어온 다차원 데이터를 저차원으로 바꾸고, 다시 고차원으로 바꾸면서 특징점 탐색
생성적 적대 신경망 (GAN)	학습 데이터 패턴과 유사하게 만드는 생성자 네트워크와 패턴의 진위 여부를 판별하는 판별자 네트워크가 서로의 목적을 달성하도록 학습 반복

- 합성곱 신경망 모델(CNN)

  1. 필터(커널) : 이미지 특징을 찾기 위한 정사각형 행렬로 정의된 파라미터

  2. 스트라이드 : 필터는 입력 데이터를 일정한 간격인 스트라이드로 순회하면서 특징을 추출하며 결과로 특징지도(feature map)가 생성

  3. 패딩 : 생성된 특징지도는 입력데이터 크기보다 작은데, 해당 출력데이터 크기가 줄어드는 것을 방지하고자 입력데이터 주변을 특정값으로 채우는 것

$$OH = \frac{H + 2P - FH}{S} + 1, \quad OW = \frac{W + 2P - FW}{S} + 1$$

P: 패딩, S: 스트라이드, (H, W) : 입력크기, (FH, FW) : 필터크기, (OH, OW) : 출력크기

Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

서포트벡터머신(SVM)

- 벡터 : 점들 간 클래스
- 초평면 : 서로 다른 분류에 속한 데이터들 간 거리를 가장 크게 하는 분류 선
- 서포트벡터 : 두 클래스를 구분하는 경계
- 마진 : 서포트벡터를 지나는 초평면 사이의 거리

연관성분석

▶ 기준이 되는 규칙

지지도(support)	데이터 전체에서 해당 사건이 나타나는 확률
신뢰도(confidence)	어떠한 사건이 다른 사건에 대하여 나타나는 확률
향상도(lift)	두 규칙의 상관관계, 독립인지 판단하는 개념 $lift(A \rightarrow B) = \frac{P(A,B)}{P(A) \cdot P(B)}$

- 아프리오리(Apriori) 알고리즘 : 최소 지지도 이상의 빈발항목집합만 찾아내서 연관규칙을 계산

군집분석

- 비지도학습, 각 개체들의 유사성을 분석해서 높은 대상끼리 일반화된 그룹으로 분류, 이상치에 민감하여 신뢰성과 타당성 검증이 어려움, 사전 정보 없이 특정 패턴·속성 파악에 효과적

▶ 군집분석의 척도

유클리드 거리	2차원 공간에서 두 점 간의 가장 짧은 거리 개념, 피타고라스 정리
맨해튼 거리	택시 거리, 시가지 거리, 가로지르지 않고 도착하는 최단거리
민코프스키 거리	$d(A,B) = \sqrt[m]{\sum_{i=1}^{p}(x_i - y_i)^m} = [\sum_{i=1}^{p}(x_i - y_i)^m]^{\frac{1}{m}}$ ※ m=1일 때 맨해튼 거리, m=2일 때 유클리드 거리와 같음
마할라노비스 거리	특정 값이 얼마나 평균에서 멀리 있는지를 나타냄, 변수 간 상관관계 고려
자카드 거리	두 집합 간 비유사성을 측정하는 지표

- 계층적 군집분석 : 계층화된 구조로 군집을 형성, 군집 수 명시 불필요, 덴드로그램 통해 결과 표현
  → 최단, 최장, 평균, Ward 연결법, 계층적 병합 군집화
- 비계층적 군집분석 : 사전 군집 수로 표본을 나누며 레코드들을 정해진 군집에 할당, 적은 계산량으로 대규모 DB 에서 처리가 유용
  → K-평균 군집 분석

Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

3_2_2 고급 분석기법 ★★

범주형 자료분석

- 빈도분석 : 질적자료를 대상으로 빈도와 비율을 계산할 때 쓰임
- 카이제곱검정 : 두 범주형 변수가 서로 상관이 있는지 독립인지를 판단하는 통계적 검정 방법
- 로지스틱 회귀분석 : 분석하고자 하는 대상들이 두 집단 또는 그 이상의 집단으로 나누어진 경우 개별 관측치들이 어느 집단으로 분류될 수 있는지 분석할 때 사용
- T 검정 : 독립변수가 범주형(두 개의 집단)이고 종속변수가 연속형인 경우 사용되는 검정 방법
- 분산분석 : 독립변수가 범주형(두 개 이상 집단)이고 종속변수가 연속형인 경우 사용되는 검정 방법

다변량분석

- 다중회귀분석 : 다수의 독립변수 변화에 따른 종속변수의 변화를 예측
- 다변량분산분석 : 두 개 이상의 범주형 종속변수와 다수의 계량적 독립변수 간 관련성을 동시에 알아볼 때 이용되는 통계적 방법
- 다변량공분산분석 : 실험에서 통제되지 않은 독립변수들의 종속변수들에 대한 효과를 제거하기 위해 이용되는 방법
- 정준상관분석 : 종속변수군과 독립변수군 간의 상관을 최대화하는 각 변수군의 선형조합을 찾음
- 요인분석 : 많은 변수들 간 상호관련성을 분석하고 어떤 공통 요인들로 설명하고자 할 때 이용
- 군집분석 : 집단에 관한 사전정보가 전혀 없는 각 표본에 대하여 그 분류체계를 찾음
- 다중판별분석 : 종속변수가 비계량적 변수인 경우, 집단 간 차이를 판별하며 어떤 사례가 여러 개의 계량적 독립변수에 기초하여 특정 집단에 속할 가능성을 예측하는 것이 주목적
- 다차원척도법 : 개체들을 원래의 차원보다 낮은 차원의 공간상에 위치시켜 개체들 사이의 구조 또는 관계를 쉽게 파악하는 목적

시계열분석

- 추세성분, 계절성분, 순환성분, 복합성분, 자기상관성, 백색잡음
- 정상성(stationarity) : 시계열 데이터가 평균과 분산이 일정한 경우, 분석이 용이한 형태

  → 모든 시점의 평균과 분산이 일정, 공분산이 시차에만 의존, 정상시계열은 평균회귀 경향을 가짐

▶ 시계열 자료의 대표 분석 방법

단순 방법	• 이동평균법 – 일정기간을 시계열을 이동하며 평균을 계산 • 지수평활법 – 관찰기간 제한 없이 모든 시계열 데이터를 사용, 최근 시계열에 더 많은 가중치를 줌 • 분해법 – 시계열 자료의 성분 분류대로 분해하는 방법
모형기반 방법	• 자기회귀모형 – 과거의 패턴이 현재자료에 영향을 준다는 가정 • 자기회귀이동평균모형 – AR(p) 모형과 MA(q) 모형의 결합형태 • 자기회귀누적이동평균모형 – 비정상성을 가지는 시계열 데이터 분석

베이즈 기법

- 회귀분석모델 적용 : 추정치와 실제의 차이를 최소화하는 것이 목표
- 나이브 베이즈 분류 : 분류에 필요한 파라미터를 추정하기 위한 학습 데이터의 양이 매우 적음, 간단한 디자인, 지도학습 환경에서 효율적

Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

딥러닝 분석

- 인공신경망 : 시냅스의 결합으로 네트워크를 형성한 인공 뉴런(노드)이 학습을 통해 시냅스의 결합 세기를 변화시켜 문제 해결 능력을 가지는 모델 전반
- 심층 신경망(DNN) : 입력층과 출력층 사이에 여러 개의 은닉층들로 이루어진 인공 신경망
- 합성곱 신경망(CNN) : 최소한의 전처리를 사용하도록 설계된 다계층 퍼셉트론의 한 종류
- 순환 신경망(RNN) : 인공 신경망을 구성하는 유닛 사이의 연결이 directed cycle 을 구성
- 심층 신뢰 신경망(DBN) : 잠재변수의 다중계층으로 이루어진 심층 신경망

비정형 데이터 분석

- 비정형 데이터 분석 기본 원리 : 비정형 데이터의 내용 파악과 패턴 발견을 위해 다양한 기법 활용, 정련 과정을 통해 정형 데이터로 만든 후 데이터 마이닝을 통해 의미있는 정보 발굴
- 데이터 마이닝 : 데이터 안에서 통계적 규칙이나 패턴을 분석하여 가치 있는 정보 추출

  → 텍스트 마이닝, 자연어 처리, 웹 마이닝, 오피니언 마이닝, 리얼리티 마이닝

앙상블 분석

- 주어진 자료로부터 여러 개의 학습 모형을 만든 후 조합하여 하나의 최종 모형을 만드는 개념, 약학습기를 통해 강학습기를 만들어내는 과정

- 약학습기 : 무작위 선정이 아닌 성공률이 높은 학습 규칙

▶ 앙상블 분석의 종류

보팅(voting)	서로 다른 알고리즘 모델을 조합해서 사용, 결과물에 대해 투표로 결정
부스팅(boosting)	가중치를 활용해 연속적인 약학습기를 생성하고 이를 통해 강학습기를 만듦, 순차적인 학습을 하며 가중치를 부여해서 오차를 보완, 병렬처리 어려움
배깅(bagging)	같은 알고리즘 내에서 다른 표본 데이터 조합을 사용, 샘플을 여러 번 뽑아 각 모델을 학습시켜 결과물을 집계 → 랜덤 포레스트
스태킹(stacking)	다양한 모델들의 예측 결과를 결합

비모수 통계

- 통계학에서 모수에 대한 가정을 전제로 하지 않고 모집단의 형태에 관계없이 주어진 데이터에서 직접 확률을 계산하여 통계학적 검정을 하는 분석

  → 모집단의 형상이 정규분포가 아닐 때, 표본이 적을 때, 자료들이 서로 독립적일 때

  → 질적척도로 측정된 자료도 분석 가능, 비교적 신속하고 쉽게 통계량 구할 수 있음

  → 부호검정, 윌콕슨 부호순위 검정, 만 휘트니 검정, 크루스칼-왈리스 검정

Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

4_1_1 분석모형 평가 ★★★

(지도학습) 분류모델 평가 지표

▶ 오차행렬

		실제 답	
		Positive	Negative
예측 결과	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

정확도(accuracy)	정밀도(precision)	재현율(recall)
$\frac{TP+TN}{TP+FP+TN+FN}$	$\frac{TP}{TP+FP}$	$\frac{TP}{TP+FN}$

- F1 score : 정밀도와 재현율을 결합한 조화평균 지표, 값이 클수록 모형이 정확
- ROC curve : FP rate 가 변할 때 TP rate 가 어떻게 변화하는지 나타내는 곡선, 하단 면적=AUC

(지도학습) 회귀모델 평가 지표

- SSE : 실제값과 예측값의 차이를 제곱하여 더한 값

  → SSE 에 평균을 취하면 MSE, MSE 에 루트를 취하면 RMSE, MSE 를 퍼센트로 변환하면 MSPE
- MAE : 실제값과 예측값의 차이의 절대값을 합한 평균값
- 결정계수 R^2 : 회귀모형이 실제값에 대해 얼마나 잘 적합한지에 대한 비율
- AIC : 최대 우도에 독립변수의 개수에 대한 손실분을 반영하는 목적으로 모형과 데이터의 확률 분포 차이를 측정하는 것, AIC 값이 낮을수록 모형의 적합도가 높음

(비지도학습) 군집분석 평가 지표

- 실루엣 계수 : a(i)는 군집 내 데이터 응집도, b(i)는 군집 간 분리도, 0.5 보다 클 시 적절한 군집 모델, 0 이면 군집으로 분리가 의미 없음 $s(i) = (b(i) - a(i))/\max\{a(i), b(i)\}$
- Dunn Index : 군집 간 거리의 최소값을 분자, 군집 내 요소 간 거리의 최대값을 분모, 값이 클수록 좋음

분석모형 진단

- 정규성 가정 : 분석을 진행하기 전 데이터가 정규분포를 따르는지 검정
- 중심극한정리 : 동일한 확률분포를 가진 독립확률변수 n 개의 평균의 분포는 n 이 적당히 크다면 정규분포에 가까워진다는 이론
- 잔차 진단 : 회귀분석에서 독립변수와 종속변수의 관계를 결정하는 최적의 회귀선은 잔차(실측치와 예측치의 차이)를 가장 작게 해주는 선

  → 정규성 진단, 등분산성 진단, 독립성 진단

k-폴드 교차검증

- k개의 서브셋, k-1개의 훈련데이터, 1개의 검증데이터, 모든 데이터 셋을 평가에 활용하여 과적합 방지

- 홀드아웃 기법 : 훈련데이터, 검증데이터, 테스트데이터를 일정 비율로 지정, 데이터셋 크기가 작을수록 데이터를 나누는 방식에 따라 모델 성능 추정에 영향

적합도 검정

- 카이제곱 검정 : k 개의 범주별로 나뉘어진 관측치들과 동일한 범주의 가정된 분포 사이의 적합도 검정
- 콜모고로프 스미르노프 검정 : 관측된 표본분포와 가정된 분포 사이의 적합도를 검사하는 누적분포함수의 차이를 이용한 검정법, 연속형 데이터에도 적용 가능

Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

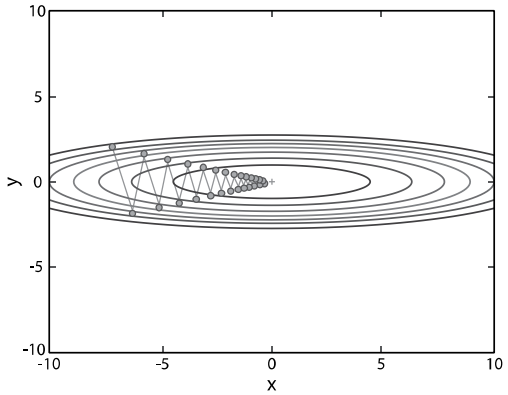
4_1_2 분석모형 개선 ★

과대적합 방지

- 드롭아웃 : 신경망 모델에서 은닉층의 뉴런을 임의로 삭제하면서 학습하는 방법
- L2 규제 : 규제란 과대적합이 되지 않도록 모델을 제한하는 의미, 손실함수에 가중치에 대한 L2 norm 의 제곱을 더한 패널티를 부여하여 가중치 값을 비용함수 모델에 비해 작게 만듦
- L1 규제 : 손실함수에 가중치의 절대값인 L1 norm 을 추가 적용하여 대부분의 특성 가중치를 0 으로 만듦

매개변수 최적화

- 확률적 경사 하강법(SGD) : 손실함수의 기울기를 따라 조금씩 아래로 내려가다 최종적으로 손실함수가 가장 작은 지점에 도달하도록 하는 알고리즘



- 모멘텀 : SGD 에 속도 개념인 기울기 방향으로 힘을 받으면 가속되는 물리법칙을 알고리즘에 적용
- AdaGrad : 개별 매개변수에 적응적으로 학습률을 조정하면서 학습을 진행하는 알고리즘
- Adam : 모멘텀과 AdaGrad 를 결합한 방법론, 모멘텀과 비슷하게 진행되나 좌우 흔들림이 덜함
- 초매개변수 최적화

학습율	기울기 방향으로 얼마나 빠르게 이동할지 결정, 작으면 학습 시간 길어짐
미니배치 크기	전체 학습 데이터를 주어진 배치 크기로 나눔, 큰 경우 병렬연산 구조를 사용할 때 효과적, 작은 경우 더 많은 가중치 업데이트 가능
훈련 반복 횟수	학습의 조기 종료를 결정하는 변수
이터레이션	하나의 미니배치를 학습할 때 1 iteration으로 1회 매개변수 업데이트 진행
은닉층 개수	• 많아질수록 특정 훈련 데이터에 더 최적화 • 모든 은닉층의 뉴런의 개수를 동일하게 하는 것이 가변적으로 하는 것보다 효과적

분석모형 융합

- 앙상블 학습 : 치우침 있는 여러 모형의 평균을 취할 시 균형적인 결과를 얻음, 과적합 여지 줄어듬
- 결합분석 모형 : 두 종류 이상의 결과변수를 동시에 분석, 결과 변수 간 유의성·관련성 설명

최종모형 선정

- 회귀모형에 대한 주요 성능평가 지표 : SSE, 결정계수 R^2, MAE, MAPE
- 분류모형에 대한 주요 성능평가 지표 : 특이도, 정밀도, 재현율, 정확도
- 비지도학습 모형에 대한 주요 성능평가 지표 : 군집분석, 연관분석

Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

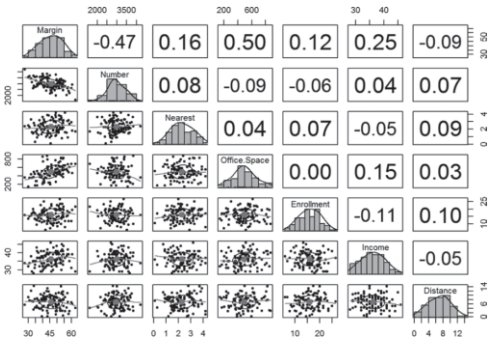
4_2_1 분석결과 해석 ★

▶ 분석 모델별 결과 해석

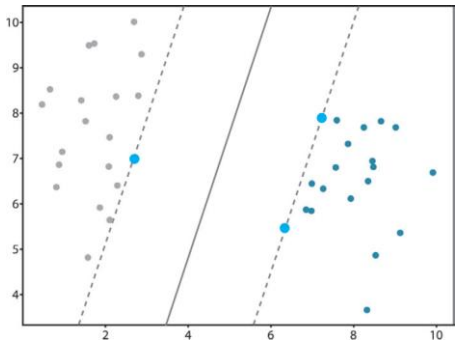
회귀 모델	잔차, 결정계수(추정된 회귀식이 변동을 얼마나 잘 설명하는지) 등을 사용
분류 모델	각각의 경우에 따라 클래스 별로 속할 확률의 정확도를 검토
딥러닝 모델	• 분류문제인 경우 정확도나 오차율을 사용 • 오차율은 상대오차나 평균 제곱근 편차를 사용
군집분석 모델	• 연속형 변수의 경우 평균 또는 중앙값을 계산 • 범주형 변수가 있는 경우 범주별로 각 군집의 분포 사용
연관분석 모델	• 두 개 또는 그 이상의 품목들 사이의 상호 관련성으로 해석 • 지지도, 신뢰도 및 향상도가 높은 규칙들을 찾되 최소 기준점을 적용

▶ 분석 모델별 시각화

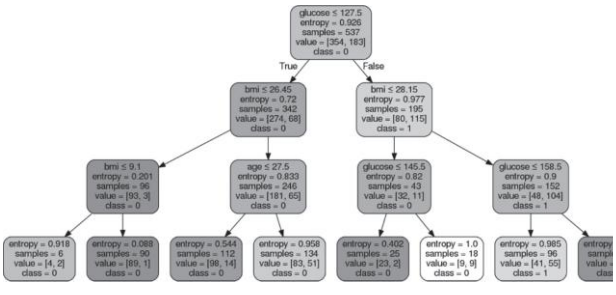
회귀 모델	변수들 간 관계 분석을 위해 히트맵과 산점도를 활용
분류 모델	• SVM : 산점도와 구분선을 통한 비교시각화 • KNN : 평행좌표계로 변수들과의 연관성 및 그룹데이터 경향성 파악 • 의사결정나무 : 트리 다이어그램으로 시각화
딥러닝 모델	모델 아키텍처에서 파라미터, 가중치 및 특징 차원감소를 통해 시각화
군집분석 모델	클러스터별 단위로 산점도로 시각화
연관분석 모델	연관성 있는 항목끼리 묶어 네트워크 그래프를 활용



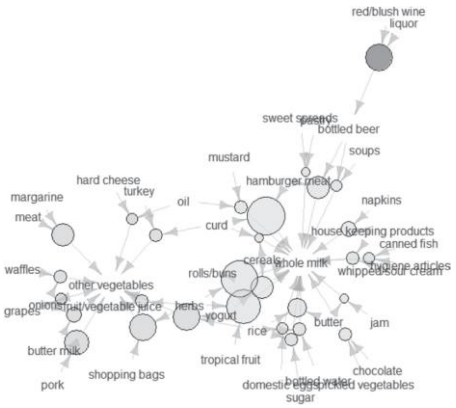
▲ 다중회귀분석 산점도 시각화



▲ SVM 시각화



▲ 의사결정나무 시각화



▲ 연관분석 시각화

Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

4_2_2 분석결과 시각화 ★★

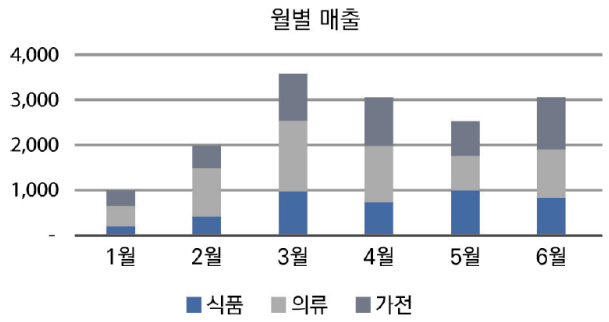
- 데이터 시각화 : 정보를 명확하고 효과적으로 전달하는 것을 목적으로 시각적 표현

  → 기능적 측면과 심미적 측면을 모두 고려

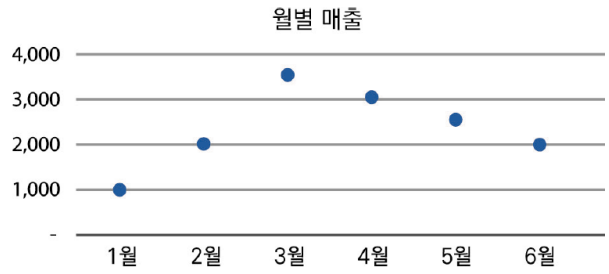
정보 시각화	추상화된 데이터를 사람이 인지하기 쉽도록 시각화하여 표현
정보 디자인	시각 디자인의 하위 영역, 정보를 구성하여 효율적 사용을 지원하는 디자인
인포그래픽	복잡한 수치, 글로 표현된 정보와 지식을 한눈에 파악하도록 시각적 표현

시간 시각화

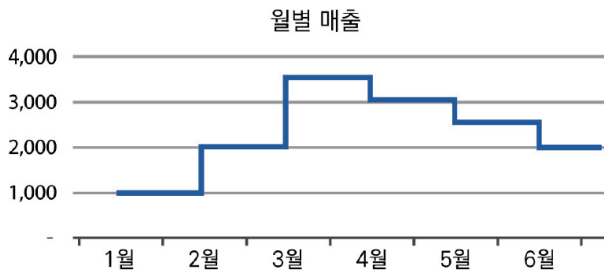
이산형	특정 시점의 값을 표현 → 막대그래프, 점그래프 등
연속형	구간의 변화하는 값을 표현 → 꺾은선그래프, 계단그래프, 추세선 등



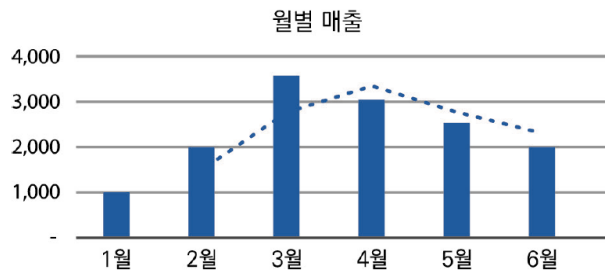
▲ 누적막대그래프



▲ 점그래프



▲ 계단그래프

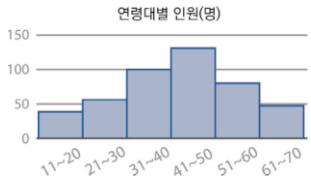


▲ 추세선

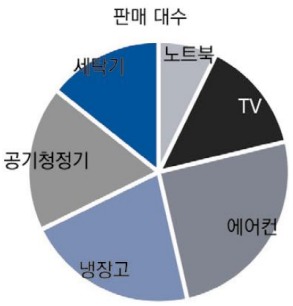
분포 시각화

- 각 영역을 모두 합치면 1 또는 100%가 됨(히스토그램, 파이차트, 도넛차트, 트리맵, 누적영역차트 등)

연령	인원(명)
11 ~ 20	35
21 ~ 30	54
31 ~ 40	98
41 ~ 50	129
51 ~ 60	77
61 ~ 70	43



▲ 도수분포표와 히스토그램

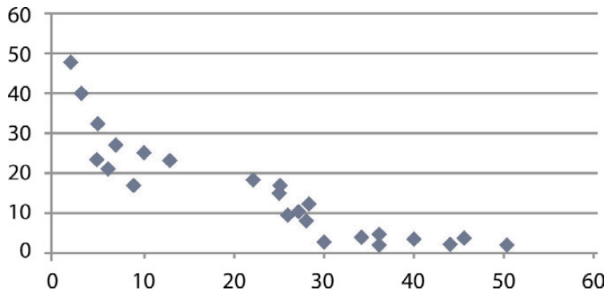


▲ 원그래프(파이차트)

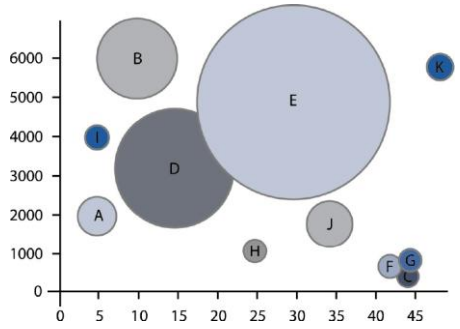
Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

관계 시각화

- 데이터셋에 변수가 두 개 이상 있을 때 상관관계(산점도, 버블차트, 히트맵 등)



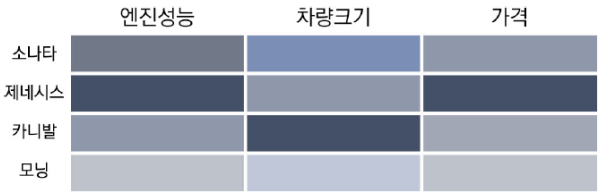
▲ 산점도(음의 상관관계)



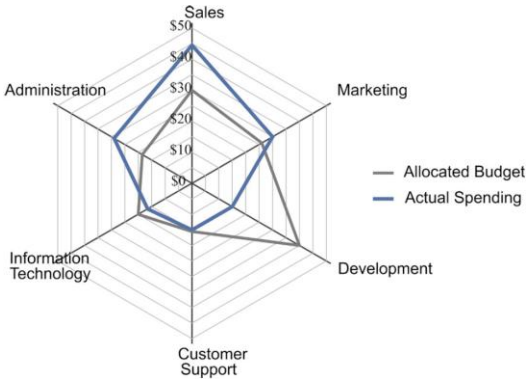
▲ 버블차트

비교 시각화

- 하나 이상의 변수에 대해 변수 사이의 차이와 유사성 등을 표현(히트맵, 체르노프 페이스, 스타차트, 평행좌표계, 다차원척도법 등)



▲ 히트맵

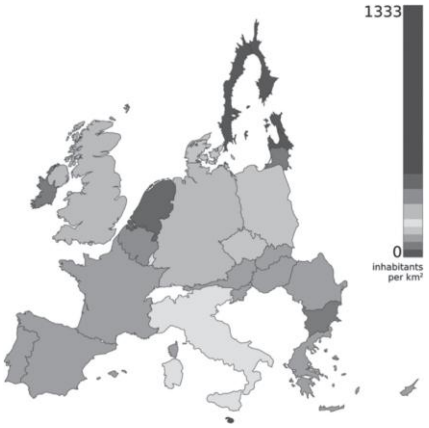


▲ 스타차트

공간 시각화

- 장소나 지역에 따른 데이터의 분포 표현(단계구분도, 카토그램 등)

▲ 단계구분도



▲ 카토그램

Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.

4_2_3 분석결과 활용 ★

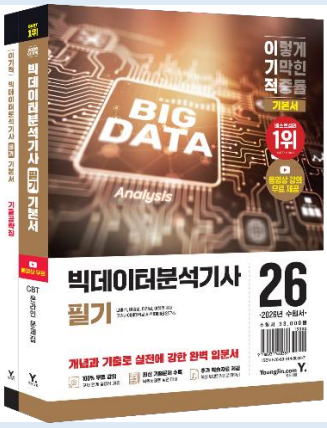
- 빅데이터 분석 방법론 참조모델(데이터산업진흥원) : 분석기획 – 데이터 준비 – 데이터 분석 – 시스템 구현 – 평가 및 전개

▶ 보편적 방법론 비교

CRISP-DM	비즈니스 이해 – 데이터 이해 – 데이터 준비 – 모델링 – 평가 – 전개
SEMMA	샘플링 – 탐색 – 전처리 – 모델링 – 평가
KDD	데이터 추출 – 전처리 – 변환 – 데이터 마이닝 – 해석/평가

- 전개 단계 : 개발된 모델을 적용하여 결과를 확인하고 계속적인 관리를 위한 방법을 제시

분석결과 활용 계획 수립	빅데이터 분석 결과를 어떻게 업무에 반영할 것인지에 대한 액션 플랜을 만들고 업무 성과를 지속적으로 모니터링할 수 있는 방안 수립
분석결과 적용과 보고서 작성	분석 모델과 결과를 업무 현장에 적용하고 업무 데이터베이스 시스템 일부로 표현, 성과 측정 지표에 따라 분석 성과 측정, 개선 계획 수립
분석모형 모니터링	이전에 수립한 활용방안이 잘 수행되고 있는지 확인하고 주변 환경과 데이터의 변화를 빅데이터 분석 모델에 지속적으로 반영하기 위함
분석모형 리모델링	분석 모형이 변화된 업무와 데이터를 지속적으로 수용할 수 있도록 데이터 품질 검토, 알고리즘 개선, 매개변수 최적화 등 과정 진행



2026 이기적 빅데이터분석기사 필기 기본서

나홍석, 배원성, 이건길, 이혜영 공저

고려사이버대학교 AI·빅데이터 연구소

- 전문 집필진 참여, 출제기준에 맞는 꼼꼼한 설명!
- 문제 풀이 동영상 강의 무료 제공!
- CBT 온라인 문제집 서비스 제공!
- 수험생을 위한 추가자료, 핵심 개념으로 단기완성!

Copyright 2025-2026 **영진닷컴**(Youngjin.com). All right reserved.