

합격에 영양 만점

ADsP 요약노트

반복 출제되는
최빈출 개념

기출분석을 기반으로
시험에 나온! 나올! 것만 모았다

데이터의 이해

DIKW 피라미드

1 DIKW 피라미드

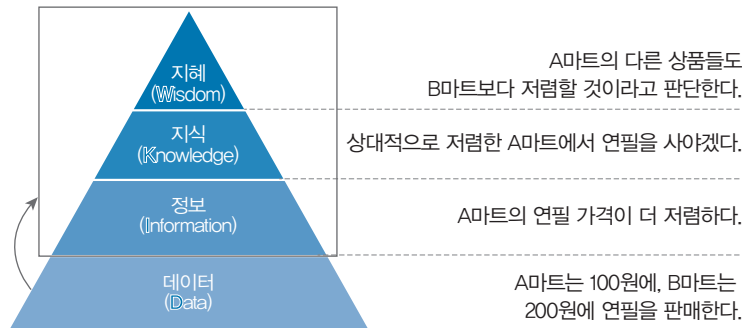
데이터, 정보, 지식을 통해 최종적으로 지혜를 얻어내는 과정을 계층구조로 설명하는 것

- 데이터를 가공 처리하여 얻을 수 있는 것 : 정보, 지식, 지혜
- Data → Information → Knowledge → Wisdom 계층구조

2 DIKW의 의미

- 데이터(Data): 타 데이터와의 상관관계가 없는 **가공하기 전**의 순수한 수치나 기호
- 정보(Information): 데이터의 가공 및 상관관계 간의 이해를 통해 패턴을 인식하고, 그 의미를 부여한 데이터
- 지식(Knowledge): 상호 연결된 정보 패턴을 이해하여 이를 토대로 예측한 결과물
- 지혜(Wisdom): 근본 원리에 대한 깊은 이해를 바탕으로 도출되는 아이디어

DIKW 피라미드



빅데이터(Big Data)

1 빅데이터의 3V, 4V

- 빅데이터의 3V: Volume, Variety, Velocity
- 빅데이터의 4V: 3V+Value, (ROI; Return On Investment, 투자자본수익률 관점에서 보는 빅데이터)

Volume	<ul style="list-style-type: none"> • 데이터의 크기, 생성되는 모든 데이터를 수집 • 구글 실시간 자동 번역 시스템에 도움을 준 빅데이터의 특징
Variety	<ul style="list-style-type: none"> • 데이터의 다양성을 의미함, 정형화된 데이터를 넘어 텍스트, 오디오, 비디오와 같은 비정형 데이터 및 웹 문서와 같은 반정형 데이터를 대상으로 함
Velocity	<ul style="list-style-type: none"> • 데이터의 속도를 의미함, 사용자가 원하는 시간 내 데이터 분석 결과 제공하며, 업데이트 속도가 빠름
Value	<ul style="list-style-type: none"> • Value: '비즈니스 효과 요소' • Volume, Variety, Velocity : '투자비용 요소'

2 데이터 크기 단위

단위	KB	MB	GB	TB	PB	EB	ZB	YB
접두어	Kilo	Mega	Giga	Tera	Peta	Exa	Zetta	Yotta
10^n	10^3	10^6	10^9	10^{12}	10^{15}	10^{18}	10^{21}	10^{24}
2^n	2^{10}	2^{20}	2^{30}	2^{40}	2^{50}	2^{60}	2^{70}	2^{80}

정비출판물

015

빅데이터의 가치와 영향

1 빅데이터 가치 산정이 어려운 이유

- 데이터의 활용 방식: 재사용이나 재조합, 다목적용 데이터 개발 등이 일반화되면서 특정 데이터를 언제, 어디서, 누가 활용할지 알 수 없음
- 새로운 가치 창출: 데이터가 기존에 없던 가치를 창출함에 따라 그 가치를 측정하기 어려움
- 분석 기술의 발달: 분석 기술의 발달로 지금은 가치 없는 데이터도 새로운 분석 기법의 등장으로 거대한 가치를 만들어내는 재료가 될 가능성이 있음

2 빅데이터가 만들어내는 본질적인 변화

사전처리	사후처리	<ul style="list-style-type: none"> • 사전처리 → 표준화된 문서 포맷 • 사후처리 → 데이터를 모은 뒤 그 안에서 숨은 정보를 찾아냄 • 구글의 자동 번역 시스템 구축 과정은 데이터의 양이 질보다 중요함을 보여주는 대표적인 사례임
표본조사	전수조사	
질(Quality)	양(Quantity)	
인과관계(Causation)	상관관계(Correlation)	

정비출판물

017

빅데이터의 위기 요인과 통제 방안

1 빅데이터의 위기 요인

빅데이터의 위기 요인에는 **사생활 침해**, **책임 원칙의 훼손**, **데이터의 오용** 등이 있음

빅데이터의 위기 요인 통제 방안

위기 요인	통제 방안
사생활 침해	동의제에서 정보 사용자의 책임제로 전환
책임 원칙의 훼손	기존 책임 원칙의 강화
데이터 오용	알고리즘에 대한 접근권 및 객관적 인증 방안 도입

2 사생활 침해

(1) 위기 요인

- 우리를 둘러싼 정보 수집 센서들의 수가 점점 늘어나고 있고, 이를 통해 수집된 특정 데이터가 본래 목적 외에 가공 처리되어 2차, 3차적 목적으로 활용될 가능성이 증가하고 있음
- 사생활 침해를 방지하기 위해 **익명화 기술**이 발전하고 있으나 아직 충분하지 않음

(2) 통제 방안

- **동의제에서 책임제로 전환함**
- 개인정보의 활용에 대해 개인이 매번 동의하는 것은 경제적으로도 매우 비효율적임
- 사생활 침해 문제를 개인정보 제공자의 동의를 통해 해결하기 보다는 **개인정보 사용자에게 책임을 지움**으로써, 개인정보 사용 주체가 더 적극적인 보호 장치를 마련하게 하는 효과가 발생할 것으로 기대됨

더 보기

- **익명화(Anonymization)**: 사생활 침해를 방지하기 위해 데이터에 포함된 개인 식별 정보를 삭제하거나 알아볼 수 없는 형태로 변환 하는 포괄적인 기술
- **가명(Pseudonym)**: 다른 추가 정보가 있으면 특정 개인 유추가 가능하며, 개인 정보 비식별화의 한 가지 방법

3 책임 원칙의 훼손

• 위기 요인

- 빅데이터 기반의 분석과 예측 기술이 발전하면서 정확도가 증가한 만큼, 분석 대상이 되는 사람들은 예측 알고리즘의 희생양이 될 가능성이 증가함
- 잠재적 위험 사항에 대해서도 책임을 추궁하는 사회로 변질될 가능성이 높아 민주주의 사회 원칙을 크게 훼손할 수 있음

예시 **범죄 예측 프로그램**을 통해 **범죄가 발생하기 전 체포**, 회사의 직원 해고, 의사의 환자 수술 거절, 어떤 사람이 특정한 사회·경제적 특성을 가진 집단에 속한다는 이유로 신용도와 무관하게 대출이 거절되는 상황 등

• 통제 방안

- **기존의 책임 원칙 보강 및 강화**
- 결과 기반의 책임 원칙 고수
- 예측 자료에 의해 불이익을 당할 가능성을 최소화하는 장치 마련

4 데이터의 오용

• 위기 요인

- 빅데이터는 일어난 일에 대한 데이터에 의존함
- 그것을 바탕으로 미래를 예측하는 것은 적지 않은 정확도를 가질 수 있지만 항상 맞을 수는 없음
- 주어진 데이터로부터 잘못된 인사이트를 얻어 비즈니스에 직접 손실을 불러올 수 있음

• 통제 방안

- **데이터 알고리즘에 대한 접근권을 허용해야 함**
- **객관적인 인증 방안을 도입해야 한다는 필요성이 제기되고 있음**

더 보기

알고리즘리스트(Algorithmist)

데이터의 오용으로 인한 부당한 피해를 보는 사람을 방지하기 위해 생겨난 직업으로, 데이터 분석 알고리즘으로 인해 불이익을 당한 사람을 구제하는 전문가를 의미합니다. 법률 전문가인 변호인, 금전 거래에 정통한 회계사처럼 컴퓨터와 수학, 나아가 통계학이나 비즈니스에 두루 깊은 지식을 갖춘 사람이 이 직업을 담당하게 됩니다.

1 데이터 분석 관련 직무별 필요 역량

• 데이터 분석가

- 데이터 분석 보고서 및 시각화 자료를 통해 비즈니스 결정에서 '추측'에 의한 결정을 없앨 수 있도록 해주고, 서로 다른 팀 간의 중재자 역할을 함
- 조직의 성장에 대한 정확한 지표를 확인하고, 데이터 기반 의사결정을 위해 통계적 데이터 분석을 하며, 분석 결과를 시각화함
- 필요 역량으로 문맥과 의미, 통찰력, 이론적 지식, 비즈니스/도메인 지식, 데이터 시각화 역량, 데이터 분석을 위한 통계적 지식, SQL 지식 등이 있음

• 데이터 사이언티스트

- 통찰력 있는 분석과 설득력 있는 전달을 할 수 있어야 하고, 다분야 간의 협력을 통해 빅데이터의 가치를 실현하는 역할을 함
- 머신러닝, AI에 대한 지식, 머신러닝 모델 구축을 위한 기본적인 언어를 사용한 코딩 스킬, 데이터 분석을 위한 통계적 지식 등의 능력이 필요함

• 데이터 분석가와 데이터 사이언티스트의 필요 역량

- 데이터 사이언티스트와 데이터 분석가의 필요 능력은 비슷한 부분이 많음
- 데이터 사이언티스트는 데이터 분석가보다 머신러닝, AI에 대한 많은 지식을 바탕으로 모델을 구축하고 데이터를 분석하는 능력이 필요함
- 데이터 분석가는 보고서 작성, 시각화, 통찰력, 비즈니스/도메인 지식 등의 능력이 필요함

2 데이터 사이언티스트의 세부 역량

• 가트너(Gartner)가 본 데이터 사이언티스트의 역량

- 데이터 관리, 분석 모델링, 비즈니스 분석, 소프트 스킬 등이 있음
- 공통점은 호기심에서 시작하는 것이며, 하드 스킬은 포함되어 있지 않음

• 일반적으로 언급되는 데이터 사이언티스트의 역량

- 데이터 해커, 애널리스트, 커뮤니케이션, 신뢰받는 어드바이저 등의 조합이라 할 수 있음
- 하드 스킬과 소프트 스킬 능력을 동시에 갖추고 있어야 함
- 데이터 처리 기술 이외에 사고방식, 비즈니스 이슈에 대한 감각, 고객들에 대한 공감 능력이 필요함

3 데이터 사이언티스트가 효과적인 분석 모델 개발을 위해 고려해야 하는 사항

- 분석 모델이 예측할 수 없는 위험을 살피기 위해 현실 세계를 돌아보고 분석을 경험과 세상에 대한 통찰력과 함께 활용해야 함
- 가정들과 현실의 불일치에 대해 끊임없이 고찰하고 모델의 능력에 대해 항상 의구심을 갖는 것이 필요함
- 분석의 객관성에 의문을 제기하고 분석 모델에 포함된 가정과 해석의 개입 등의 한계를 고려해야 함
- 모델 범위 바깥의 요인은 판단하지 말아야 함

3 데이터 사이언티스트가 갖추어야 할 스킬

• 하드 스킬

- Machine Learning, Modeling, Data Technical Skill
- 빅데이터에 대한 이론적 지식: 관련 기법에 대한 이해와 방법론 습득
- 분석 기술에 대한 숙련: 최적의 분석 설계 및 노하우 축적

- 소프트 스킬

- 통찰력 있는 분석: 창의적 사고, 호기심, 논리적 비판
- 설득력 있는 전달: **Storytelling**, Visualization
- 다분야 간 협력: Communication

데이터 분석 기획

분석 프로젝트의 특징 및 특성 관리

1 분석 프로젝트의 특징

- 분석가의 목표는 분석의 정확도를 높이는 것이지만 프로젝트의 관점에서는 도출된 분석 과제를 잘 구현하여 원하는 결과를 얻고 사용자가 원활하게 활용할 수 있도록 전체적인 과정을 고려해야 하므로 개별적인 분석 업무 수행뿐만 아니라 전반적인 프로젝트 관리도 중요함
- 분석 프로젝트에서는 데이터 영역과 비즈니스 영역의 현황을 이해하고 프로젝트의 목표인 분석의 정확도 달성과 결과에 대한 이해를 전달하는 조정자로서의 분석가 역할이 중요함
- 분석 프로젝트는 도출된 결과의 재해석을 통한 지속적인 반복 및 정교화가 수행되는 경우가 대부분이므로 프로토타이핑 방식의 어자일(Agile) 프로젝트 관리 방식도 고려해야 함
- 분석 과제 정의서를 기반으로 분석 프로젝트를 시작하되 지속적인 개선 및 변경을 염두해 두고 시간 내에 최선의 결과를 도출할 수 있도록 프로젝트 구성원들과 협업하여 진행해야 함
- 다양한 데이터에 기반한 분석 기법을 적용하는 특성 때문에 5가지 주요 특성을 고려하여 추가적인 관리가 필요함
- 분석 과제 주요 특성 관리 영역: Data Size, Data Complexity, Speed, Analytic Complexity, Accuracy & Precision 등
- 분석 프로젝트는 다른 프로젝트 유형처럼 범위, 일정, 품질, 리스크, 의사소통 등 영역별 관리가 수행되어야 함

더 보기

Agile 프로젝트 관리 방식

- 소프트웨어 개발 및 다른 프로젝트 영역에서 사용되는 반복적이고 유연한 접근 방식
- 변화에 빠르게 대응하고 고객 요구 사항을 우선적으로 고려하는 민첩한 프로젝트 관리 방법론, 불확실한 환경에서 효과적

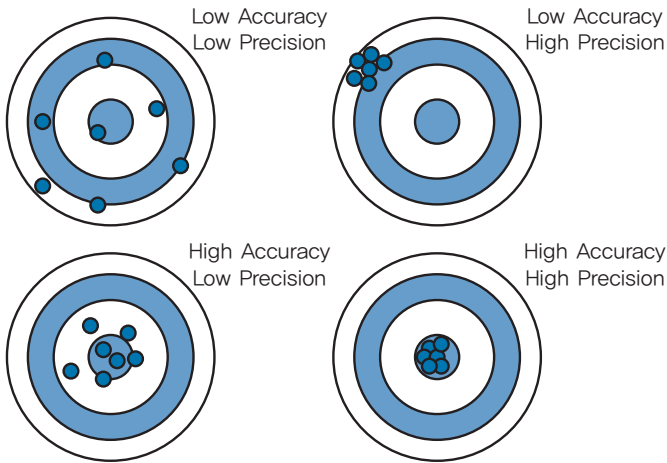
2 분석 과제 정의서

- 다양한 분석 과제 도출 방법을 통해 도출된 분석 과제를 분석 과제 정의서로 정리함
- 필요한 소스 데이터, 분석 방법, 데이터 입수 난이도, 데이터 입수 사유, 분석 수행 주기, 분석 결과에 대한 검증, 분석 과정 상세 등을 작성함
- 프로젝트 수행 계획의 입력물로 사용됨
- 이해관계자가 프로젝트의 방향을 설정하고, 성공 여부를 판별할 수 있는 중요한 자료로 명확하게 작성해야 함

3 분석 과제의 주요 5가지 특성 관리 영역

Data Size	분석하고자 하는 데이터의 양을 고려하는 관리방안 수립이 필요함
Data Complexity	비정형데이터 및 다양한 시스템에 산재되어 있는 데이터들을 통합해서 분석 프로젝트를 진행할 때는 해당 데이터에 잘 적용될 수 있는 분석 모델 선정에 대한 고려가 필요함
Speed	<ul style="list-style-type: none"> • 분석 결과 도출 후, 활용하는 시나리오 측면에서 일, 주 단위 실적은 배치 형태 작업, 사기 탐지, 서비스 추천은 실시간 수행되어야 함 • 분석 모델의 성능 및 속도를 고려한 개발 및 테스트가 수행되어야 함

Analytic Complexity	<ul style="list-style-type: none"> 정확도(Accuracy)와 복잡도(Complexity)는 트레이드 오프 관계가 존재 분석 모델이 복잡할수록 정확도는 올라가지만 해석이 어려워짐 기준점을 사전에 정의해 두어야 함
Accuracy & Precision	<ul style="list-style-type: none"> Accuracy: 분석의 활용적인 측면 (모델과 실제 값의 차이) Precision: 분석의 안정성 측면 (모델을 반복했을 때의 편차) Accuracy, Precision은 트레이드 오프인 경우가 많음 모델의 해석 및 적용 시 사전에 고려해야 함



▲ Accuracy와 Precision의 관계

3 10개 주제별 프로젝트 관리 체계

분석 프로젝트는 시간, 범위, 품질, 통합, 이해 관계자, 자원, 원가, 리스크, 조달, 의사소통과 같은 관리 영역에서 일반 프로젝트와 다르게 유의해야 할 요소가 존재함

시간	<ul style="list-style-type: none"> 초기에 의도했던 결과가 나오기 쉽지 않아 지속적으로 반복되어 많은 시간이 소요될 수 있음 분석 결과에 대한 품질이 보장된다는 것을 전제하여 Time Boxing 기법으로 일정 관리를 진행하는 것이 필요함
범위	<ul style="list-style-type: none"> 프로젝트 범위가 분석을 진행하면서 데이터의 형태와 양 또는 적용되는 모델의 알고리즘에 따라 범위가 빈번하게 변경됨 분석의 최종 결과물이 분석 보고서 형태인지, 시스템인지에 따라 투입되는 자원 및 범위가 크게 변경되므로 사전에 충분한 고려가 필요함
품질	<ul style="list-style-type: none"> 품질 보증과 품질 통제를 계획하고 확립하는 데 요구되는 프로세스
통합	<ul style="list-style-type: none"> 프로젝트와 관련된 다양한 활동과 프로세스를 도출, 정의, 결합, 단일화, 조정, 통제, 종료에 필요한 프로세스
이해관계자	<ul style="list-style-type: none"> 프로젝트 스폰서, 고객사, 기타 이해관계자 식별 및 관리에 필요한 프로세스
자원	<ul style="list-style-type: none"> 인력, 시설, 장비, 자재, 기반 시설, 도구와 같은 적절한 프로젝트 자원을 식별하고 확보하는 데 필요한 프로세스
원가	<ul style="list-style-type: none"> 개발 예산과 원가 통제의 진척 상황을 관찰하는 데 요구되는 프로세스
리스크	<ul style="list-style-type: none"> 위험과 기회를 식별하고 관리하는 프로세스
조달	<ul style="list-style-type: none"> 계획에 요구된 프로세스를 포함하며, 제품 및 서비스 또는 인도물을 인수하고 공급자와의 관계를 관리하는 데 요구되는 프로세스
의사소통	<ul style="list-style-type: none"> 프로젝트와 관련된 정보를 계획, 관리, 배포하는 데 요구되는 프로세스

더 보기

Time Boxing 기법

- 프로젝트 관리 및 작업 관리에서 사용되는 접근 방식 중 하나
- 특정 작업이나 활동에 대한 시간 제한을 설정하여 그 시간 동안 해당 작업을 완료하도록 장려하는 방법
- Agile 방법론과 같은 프로젝트 관리 접근 방식에서 자주 사용됨
- 프로젝트를 작은 단위로 나누고 각 단위에 대한 시간 제한을 설정하여 반복적이고 효율적인 작업 촉진에 도움이 됨

1 데이터 분석 수준 진단 개요

- 데이터 분석 수준 진단으로 데이터 분석 기법을 구현하기 위해 무엇을 준비하고 보완해야 하는지 알 수 있음
- 분석의 유형 및 분석의 방향성 결정에 도움을 줌
- 분석 준비도와 분석 성숙도를 함께 평가함으로써 수행됨

2 데이터 분석 준비도(Readiness)

- 기업의 데이터 분석 도입의 수준을 파악하기 위한 진단 방법으로 6가지 영역을 대상으로 현 수준을 파악함
- 분석 준비도의 6가지 영역: **분석 업무 파악**, **인력 및 조직**, **분석 기법**, **분석 데이터**, **분석 문화**, **IT 인프라(=분석 인프라)**

분석 업무 파악	인력 및 조직	분석 기법
<ul style="list-style-type: none"> • 발생한 사실 분석 업무 • 예측 분석 업무 • 시뮬레이션 분석 업무 • 최적화 분석 업무 • 분석 업무 정기적 개선 	<ul style="list-style-type: none"> • 분석 전문가 직무 존재 • 분석 전문가 교육 훈련 프로그램 • 관리자의 기본 분석 능력 • 전사 분석 업무 총괄 조직 존재 • 경영진 분석 업무 이해 능력 	<ul style="list-style-type: none"> • 업무별 적합한 분석 기법 사용 • 분석 업무 도입 방법론 • 분석 기법 라이브러리 • 분석 기법의 효과성 평가 • 분석 기법의 정기적 개선
분석 데이터	분석 문화	IT 인프라(=분석 인프라)
<ul style="list-style-type: none"> • 분석 업무를 위한 데이터 충실성 • 분석 업무를 위한 데이터 신뢰성 • 분석 업무를 위한 데이터 적시성 • 비구조적 데이터 관리 • 외부 데이터 활용 체계 • 기존 데이터 관리 	<ul style="list-style-type: none"> • 사실에 근거한 의사결정 • 관리자의 데이터 중시 • 회의 등에서 데이터 활용 • 경영진의 직관보다 데이터 활용 • 데이터 공유 및 협업 문화 	<ul style="list-style-type: none"> • 운영 시스템 데이터 통합 • EAI, ETL 등 데이터 유통 체계 • 분석 전용 서버 및 스토리지 • 빅데이터 분석 환경 • 비주얼 분석 환경

3 분석 성숙도(Maturity)

- 시스템 개발 업무 능력과 조직의 성숙도 파악을 위해 CMMI 모델을 기반으로 분석 성숙도를 평가함
- **비즈니스**, **조직/역량**, **IT 부문**을 대상으로 성숙도 수준에 따라 **도입**, **활용**, **확산**, **최적화** 단계로 구분해 살펴볼 수 있음
- 데이터 분석 성숙도 수준에 따른 단계

	도입	→	활용	→	확산	→	최적화
단계	도입		활용		확산		최적화
설명	분석을 시작하여 환경과 시스템 구축		분석 결과를 실제 업무에 적용		전사 차원에서 분석을 관리하고 공유		분석을 진화시켜 혁신 및 성과 향상에 기여
비즈니스 부문	<ul style="list-style-type: none"> • 실적 분석 및 통계 • 정기 보고 수행 • 운영 데이터 기반 		<ul style="list-style-type: none"> • 미래 결과 예측 • 시뮬레이션 • 운영 데이터 기반 		<ul style="list-style-type: none"> • 전사 성과 실시간 분석 • 프로세스 혁신 3.0 • 분석 규칙 관리, • 이벤트 관리 		<ul style="list-style-type: none"> • 외부환경 분석 활용 • 최적화 업무 적용 • 실시간 분석 • 비즈니스 모델 진화
조직 역량 부문	<ul style="list-style-type: none"> • 일부 부서에서 수행 • 담당자 역량에 의존 		<ul style="list-style-type: none"> • 전문 담당 부서에서 수행 • 분석 기법 도입 • 관리자가 분석 수행 		<ul style="list-style-type: none"> • 전사 모든 부서 수행 • 분석 CoE 조직 운영 • 데이터 사이언티스트 확보 		<ul style="list-style-type: none"> • 데이터 사이언스 그룹 • 경영진 분석 활용 • 전략 연계
IT 부문	<ul style="list-style-type: none"> • 데이터 웨어하우스 • 데이터 마트 • ETL / EAI, OLAP 		<ul style="list-style-type: none"> • 실시간 대시보드 • 통계 분석 환경 		<ul style="list-style-type: none"> • 빅데이터 관리 환경 • 시뮬레이션/최적화 • 비주얼 분석 • 분석 전용 서버 		<ul style="list-style-type: none"> • 분석 협업 환경 • 분석 Sandbox • 프로세스 내재화 • 빅데이터 분석

능력 성숙도 통합 모델(Capability Maturity Model Integration, MMI)

소프트웨어 개발 및 전산장비 운영 업체들의 업무 능력 및 조직의 성숙도를 평가하기 위한 모델

CoE(Center of Excellence)

구성원들이 비즈니스 역량, IT 역량 및 분석 역량을 고루 갖추어야 하며, 협업 부서 및 IT 부서와의 지속적인 커뮤니케이션을 수행하는 조직내 분석 전문조직

Sandbox

보안모델, 외부 접근 및 영향을 차단하여 제한된 영역 내에서만 프로그램을 동작시키는 것

데이터 거버넌스 체계 수립**1 데이터 거버넌스**

- 전사 차원의 모든 데이터에 대해 정책 및 지침, 표준화, 운영 조직 및 책임 등의 표준화된 관리 체계를 수립하고 운영을 위한 프레임워크(Framework) 및 저장소(Repository)를 구축하는 것
- 마스터데이터(Masterdata), 메타데이터(Metadata), 데이터 사전(Data Dictionary)은 데이터 거버넌스의 중요한 관리 대상임
- 기업은 데이터 거버넌스 체계를 구축함으로써 데이터의 가용성, 유용성, 통합성, 보안성, 안정성을 확보할 수 있음
- 빅데이터 프로젝트를 성공으로 이끄는 기반이 됨
- 데이터 거버넌스는 독자적으로 수행될 수도 있지만 전사 차원의 IT 거버넌스나 EA(Enterprise Architecture)의 구성 요소로써 구축되는 경우도 있음

2 데이터 거버넌스 구성 요소

- 원칙(Principle): 데이터를 유지 관리하기 위한 지침과 가이드 **예시** 보안, 품질 기준, 변경 관리
- 조직(Organization): 데이터를 관리할 조직의 역할과 책임 **예시** 데이터 관리자, 데이터베이스 관리자, 데이터 아키텍트
- 프로세스(Process): 데이터 관리를 위한 활동과 체계 **예시** 작업 절차, 모니터링 활동, 측정 활동

3 데이터 거버넌스 체계 요소

- 데이터 거버넌스 체계 요소 이해

데이터 표준화	데이터 표준용어 설정, 명명 규칙 수립, 메타데이터 구축, 데이터 사전 구축
데이터 관리 체계	메타데이터와 데이터 사전(Data Dictionary)의 관리 원칙 수립
데이터 저장소 관리	메타데이터 및 표준 데이터를 관리하기 위한 전사 차원의 저장소를 구성
표준화 활동	데이터 거버넌스 체계 구축 후, 표준 준수 여부를 주기적으로 점검, 모니터링

- 데이터 거버넌스의 데이터 저장소 관리
 - 메타데이터 및 표준 데이터를 관리하기 위한 전사 차원의 저장소를 구성
 - 저장소는 데이터 관리 체계 지원을 위한 워크프로우 및 관리용 응용 소프트웨어를 지원하고 관리 대상 시스템과의 인터페이스를 통한 통제가 이루어져야 함
 - 데이터 구조 변경에 따른 사전영향평가도 수행되어야 효율적인 활용이 가능함

4 빅데이터 거버넌스의 특징

- 데이터 거버넌스의 체계에 더해 빅데이터의 효율적인 관리, 다양한 데이터의 관리 체계, 데이터 최적화, 정보 보호, 데이터 생명 주기 관리, 데이터 카테고리별 관리 책임자 지정 등을 포함
- 기업이 가진 과거 및 현재의 모든 데이터를 분석하여 비즈니스 인사이트를 찾는 노력은 비용면에서 효율적이지 못함
→ 분석 대상 및 목적을 명확히 정의하고, 필요한 데이터를 수집, 분석하여 점진적으로 확대해 나가는 것이 좋음
- 빅데이터 분석에서 품질관리도 중요하지만, 데이터 수명주기 관리방안을 수립하지 않으면 데이터 가용성 및 관리 비용 증대 문제에 직면할 수 있음
- ERD는 운영 중인 데이터베이스와 일치하기 위해 계속해서 변경사항을 관리하여야 함
- 산업 분야별, 데이터 유형별, 정보 거버넌스 요소별로 구분하여 작성함
- 적합한 분석 업무를 도출하고 가치를 높여줄 수 있도록 분석 조직 및 인력에 대해 지속적인 교육과 훈련을 실시함
- 개인정보보호 및 보안에 대한 방법을 마련해야 함

최민호 논자
046

데이터 분석을 위한 조직 구조

1 집중형 조직 구조

- 조직 내 별도 독립적인 분석 전담 조직을 구성하여 분석 전담 조직에서 회사의 모든 분석 업무를 담당함
- 전사 분석 과제의 전략적 중요도에 따라 우선순위를 정해 추진함
- 일부 협업 부서와 분석 업무가 중복 또는 이원화될 가능성이 있음

2 기능 중심 조직 구조

- 일반적인 분석 수행 구조로, 별도 분석 조직을 구성하지 않고 각 해당 업무 부서에서 직접 분석함
- 전사적 관점에서 핵심 분석이 어려움
- 특정 업무 부서에 국한된 분석 수행 가능성이 높거나 일부 중복된 분석 업무를 수행할 수 있음

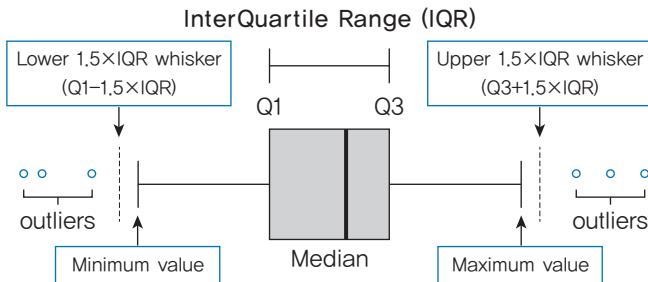
3 분산 조직 구조

- 분석 조직의 인력들이 협업 부서에 배치되어 업무를 수행함
- 전사 차원에서 분석 과제의 우선순위를 선정해 수행이 가능함
- 분석 결과를 신속하게 실무 적용 가능함
- 부서 분석 업무와 역할 분담을 명확히 해야 함



데이터 분석

상자그림(Boxplot)



- Min, Q1, Median, Q3, Max 값과 이상값(Outlier)을 확인할 수 있음
- Min, Max는 일반적인 범위(이상치 범위 안쪽)에서 가장 작은 값과 큰 값을 의미하므로 Min, Max 값까지 연결된 선(=수염)의 길이는 대칭이 아닐 수 있음
- Q1, Q3는 각각 제1사분위수, 제3사분위수를 의미하며 25%, 75% 위치를 나타냄
- IQR은 사분위수 범위라고 하여 $Q3 - Q1$ 으로 구함
- 이상값의 기준이 되는 하한은 $Q1 - 1.5 \times IQR$, 상한은 $Q3 + 1.5 \times IQR$
- 하한보다 작은 값이나 상한보다 큰 값은 이상값으로, 그래프에서 동그라미로 표시됨
- Median은 중앙값을 의미하고, 50% 위치를 나타냄
- 상자그림에 평균, 분산, 데이터의 개수 등의 정보는 들어있지 않음

회귀모형 해석

1 표본 회귀선의 유의성 검정

- 두 변수 사이에 선형관계가 성립하는지 검정함
- 귀무가설과 대립가설
 - 귀무가설: 회귀식의 기울기 계수 β_1 은 0과 같다.
 - 대립가설: 회귀식의 기울기 계수 β_1 은 0과 같지 않다.

2 회귀 모형의 해석 방법

- ‘모형이 통계적으로 유의미한가?’ → F 통계량의 유의확률(p-value)로 확인함
- ‘회귀 계수들이 통계적으로 유의미한가?’ → 회귀 계수의 t 값에 대한 유의 확률로 확인함
- ‘모형이 얼마나 설명력을 갖는가?’ → 결정계수(R^2)의 크기로 확인함
- ‘모형이 데이터를 잘 적합하고 있는가?’ → 잔차를 그래프로 그려 회귀 진단을 해야 함

- ‘데이터가 모형 가정을 만족시키는가?’를 확인 → 모형 가정: 선형성, 독립성, 등분산성, 비상관성, 정상성

3 F 통계량

$$F \text{ 통계량} = \frac{\text{회귀제곱평균(MSR)}}{\text{잔차제곱평균(MSE)}}$$

- 모형의 통계적 유의성을 검정하기 위한 검정통계량
- F 통계량이 클수록 회귀 모형은 통계적으로 유의함
- F 통계량에 대한 $p\text{-value} < 0.05$ 일 때 통계적으로 유의함

4 t 값

$$t \text{ 값} = \frac{\text{Estimate(회귀계수)}}{\text{STD. Error(표준오차)}}$$

- t 통계량이 크다는 것은 표준오차가 작다는 의미함
- t 통계량이 클수록 회귀계수가 유의함
- t 값에 대한 $p\text{-value} < 0.05$ 일 때 통계적으로 유의함

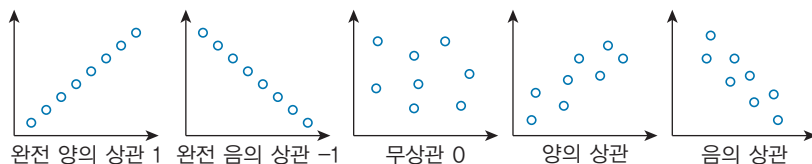
최빈도 노트

081

상관 분석

1 상관계수의 이해

- 상관계수는 두 변수의 **관련성의 정도를 의미함**(-1 ~ 1의 값으로 나타냄)
- 두 변수의 상관관계가 존재하지 않을 경우 상관계수는 '0'임



- 상관관계가 높다고 인과관계가 있다고 할 수는 없음
- 피어슨 상관계수와 스피어만 상관계수가 있음
- 피어슨 상관계수는 두 변수 간의 **선형적인 크기만 측정 가능하며** 스피어만 상관계수는 두 변수 간의 **비선형적인 관계도 나타낼 수 있음**
- R의 `cor.test()` 함수를 사용해 상관계수 검정을 수행하고, 유의성검정을 판단할 수 있음
- 이때 귀무가설은 ‘상관계수는 0이다.’이고, 대립가설은 ‘상관계수는 0이 아니다.’임

2 공분산

- 2개 확률변수의 선형 관계를 나타내는 값
- 모집단의 공분산 $cov(x, y) = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{N}$, 표본의 공분산 $cov(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$
- 하나의 변수가 상승하는 경향을 보일 때 다른 값도 상승하는 선형 상관성이 있다면 양의 공분산을 갖음

- 두 확률변수 x, y 가 독립이면 공분산 $cov(x, y)=0$ 이며, 관측값들이 4면에 균일하게 분포되어 있다고 추정할 수 있음
- $cov(x, y)=0$ 이라고 해서 항상 두 확률변수 x, y 가 독립인 것은 아님

3 피어슨 상관계수와 스피어만 상관계수

- 피어슨 상관계수(Pearson correlation)
 - x, y 의 공분산을 x, y 의 표준편차의 곱으로 나눈 값
 - 모집단의 경우: $corr(x, y) = \frac{cov(x, y)}{\sigma_x \sigma_y}$
 - 표본의 경우: $corr(x, y) = \frac{cov(x, y)}{s_x s_y}$
 - 대상자료는 등간척도, 비율척도를 사용함
 - 두 변수 간의 선형적인 크기만 측정 가능함
- 스피어만 상관계수(Spearman correlation)
 - 서열척도인 자료에서 사용 가능함
 - 두변수 간의 비선형적인 관계를 나타낼 수 있음
 - 두 변수의 순위 사이의 통계적 의존성을 측정하는 비모수적인 척도, 연속형 외에 이산형도 가능함
 - 스피어만 상관계수는 원시 데이터가 아니라 각 변수에 대해 순위를 매긴 값을 기반으로 함
 - 두 변수 안의 순위가 완전 일치하면 1, 완전 반대이면 -1

예시 수학 잘하는 학생이 영어도 잘하는 것과 상관있는지 알아보는데 사용될 수 있음

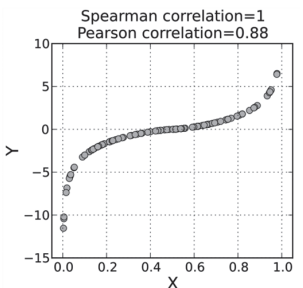
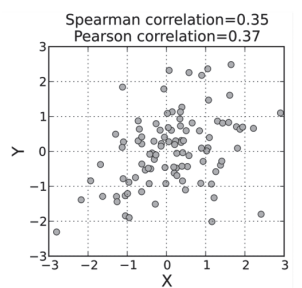
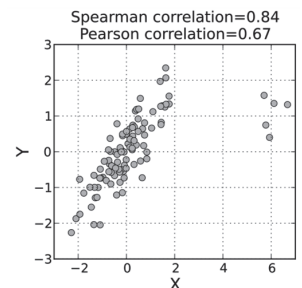
4 피어슨 상관계수 구하기

- 응답자1의 표준편차 2, 응답자2의 표준편차 2, 두 응답자의 공분산 값 4이면 피어슨 상관계수는?
 피어슨 상관계수(p): $corr(x, y) = \frac{cov(x, y)}{\sigma_x \sigma_y} = \frac{cov(x, y)}{s_x s_y} = \frac{4}{4} = 1$
- 다음 [표]는 응답자의 키와 몸무게를 나타낸 것이다. 키와 몸무게의 피어슨 상관계수는 얼마인가?(단, 모집단에 대한 값을 구한다.)

응답자 ID	키	몸무게
1	165	65
2	170	70
3	175	75
4	180	80
5	185	85

- 피어슨 상관계수(p): $corr(x, y) = \frac{cov(x, y)}{\sigma_x \sigma_y}$
- 키의 평균은 175, 표준편차는 $\sqrt{5}$ 이고, 몸무게의 평균은 75, 표준편차는 $\sqrt{75}$
- 이때 키와 몸무게의 공분산은 50
- 이를 피어슨 상관계수 식에 대입하면 $\frac{50}{\sqrt{50} \times \sqrt{50}} = 1$

5 스피어만 상관계수 그래프 해석

		
<p>두 변수 X, Y가 선형 관계가 아니더라도 스피어만 상관계수는 1이 될 수 있음</p>	<p>데이터가 뚜렷한 경향성을 보이지 않을 경우 스피어만 상관계수와 피어슨 상관계수는 비슷한 값을 가짐</p>	<p>스피어만 상관계수는 피어슨 상관계수에 비해 이 상치에 덜 민감함. 이는 스피어만 상관계수가 이 상치를 그 값이 아닌 순위로써만 고려하기 때문임</p>

6 귀무가설 예시 1

```
> cor.test(c(1,3,5,7,9), c(1,2,4,6,8), method='pearson')

Pearson's product-moment correlation

data: c(1, 3, 5, 7, 9) and c(1, 2, 4, 6, 8)
t = 15.588, df = 3, p-value = 0.0005737
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9065015 0.9996163
sample estimates:
      cor
0.9938837
```

- 귀무가설: 상관계수가 0이다.
- 대립가설: 상관계수가 0이 아니다.
- p-value가 0.05보다 작은 값(p-value=0.0005737)이므로 귀무가설을 기각하고 '상관계수가 0이 아니다'라는 대립가설을 채택함

7 귀무가설 예시 2

```
> Carseats <- read.csv('data/Carseats_dataset.csv', stringsAsFactors=TRUE)
> rcorr(as.matrix(Carseats[, c(1:6, 8)]), type="pearson")
```

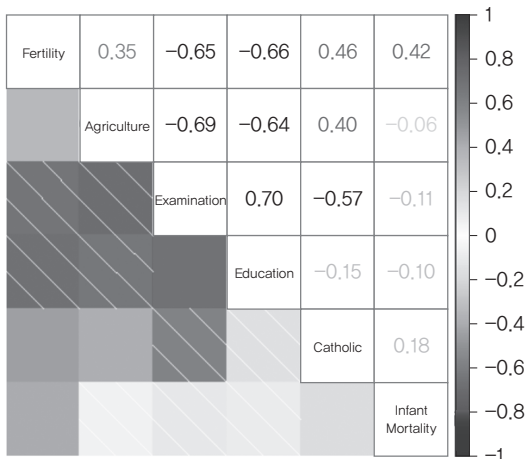
	Sales	CompPrice	Income	Advertising	Population	Price	Age
Sales	1.00	0.06	0.15	0.27	0.05	-0.44	-0.23
CompPrice	0.06	1.00	-0.08	-0.02	-0.09	0.58	-0.10
Income	0.15	-0.08	1.00	0.06	-0.01	-0.06	0.00
Advertising	0.27	-0.02	0.06	1.00	0.27	0.04	0.00
Population	0.05	-0.09	-0.01	0.27	1.00	-0.01	-0.04
Price	-0.44	0.58	-0.06	0.04	-0.01	1.00	-0.10
Age	-0.23	-0.10	0.00	0.00	-0.04	-0.10	1.00

n= 400
P

	Sales	CompPrice	Income	Advertising	Population	Price	Age
Sales		0.2009	0.0023	0.0000	0.3140	0.0000	0.0000
CompPrice	0.2009		0.1073	0.6294	0.0584	0.0000	0.0451
Income	0.0023	0.1073		0.2391	0.8752	0.2579	0.9258
Advertising	0.0000	0.6294	0.2391		0.0000	0.3743	0.9276
Population	0.3140	0.0584	0.8752	0.0000		0.8087	0.3948
Price	0.0000	0.0000	0.2579	0.3743	0.8087		0.0411
Age	0.0000	0.0451	0.9258	0.9276	0.3948	0.0411	

- 결과에서 위쪽에 있는 것이 상관계수 행렬, 아래에 있는 것은 상관계수에 대한 p-value 행렬임
- 위쪽에 있는 상관계수 행렬은 상관관계에 대해 보여주는 것으로 -1 ~ 1의 값으로 표시되며 대각선은 1로 채워지며, 두 변수의 교차 지점의 숫자가 상관계수임
- 예를 들어 Price와 Sales는 -0.44이며, 음의 상관관계를 가짐
- 아래쪽에 있는 p-value 행렬을 사용해 위쪽의 상관계수가 통계적으로 유의미한지를 알 수 있음
- 예를 들어 Price와 Sales는 통계적으로 유의미하며, Sales와 CompPrice, Sales와 Population는 통계적으로 유의하지 않음
- Sales와 가장 강한 상관관계를 보이는 변수는 Price로 절댓값을 보았을 때 가장 크기 때문임
- 전체에서 가장 큰 상관관계를 보이는 두 변수는 Price와 CompPrice임

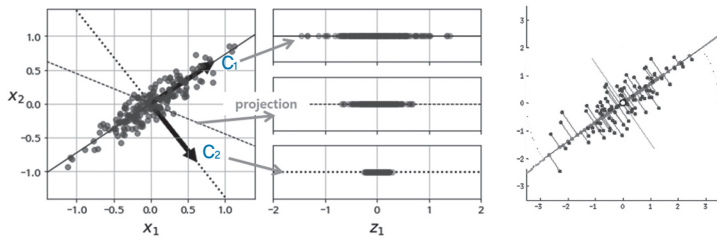
8 그래프로 표시된 상관계수 행렬 해석



- 변수 이름에서 행과 열의 방향으로 있는 숫자를 사용하여 해석함(변수 이름의 교차지점의 숫자가 두 변수의 상관계수임)
- 사각형에 해당하는 0.70은 Examination과 Education 변수의 상관계수임

1 주성분 분석(PCA, Principal Component Analysis)

- 데이터를 분석할 때 변수의 개수가 많다고 모두 활용하는 것이 꼭 좋은 것은 아님
- 오히려 변수가 '다중공선성'이 있을 경우 분석 결과에 영향을 줄 수도 있음
- 공분산 행렬 또는 상관관계수 행렬을 사용해 모든 변수들을 가장 잘 설명하는 주성분을 찾는 방법
- 상관관계가 있는 변수들을 선형 결합에 의해 상관관계가 없는 새로운 변수(주성분)를 만들고 분산을 극대화하는 변수로 축약함
- 주성분은 변수들의 선형결합으로 이루어져 있음
- 독립변수들과 주성분과의 거리인 '정보손실량'을 최소화하거나 분산을 최대화함



원본 데이터 셋과 투영(projection)된 데이터셋 간의 분산이 최대가 되는 축을 찾는다(그림에서 C1).
⇒ "PCA는 데이터의 분산이 최대가 되는 축을 찾는다" = "정보의 손실을 최소화한다"

2 주성분 분석을 할 때 고민해야 하는 것

- 공분산 행렬과 상관관계수 행렬 중 어떤 것을 선택할 것인가?
- 주성분의 개수를 몇 개로 할 것인가?
- 주성분에 영향을 미치는 변수로 어떤 변수를 선택할 것인가?

3 공분산 행렬(default) VS 상관관계수 행렬

- 공분산 행렬은 변수의 측정 단위를 그대로 반영한 것이고, 상관관계수 행렬은 모든 변수의 측정 단위를 표준화한 것임
- 공분산 행렬을 이용한 경우 측정 단위를 그대로 반영하였기 때문에 변수들의 측정 단위에 민감함
- 주성분 분석은 거리를 사용하기 때문에 척도에 영향을 받음(정규화 전후의 결과가 다름)
- 설문조사처럼 모든 변수들이 같은 수준으로 점수화 된 경우 공분산 행렬을 사용함
- 변수들의 scale이 서로 많이 다른 경우에는 상관관계수 행렬(correlation matrix)을 사용함

4 주성분 분석에서의 상관관계수 행렬 적용

- `prcomp(data, scale=TRUE)`, `princomp(data, cor=TRUE)`
- `scale`, `cor`을 FALSE로 지정하거나 생략 시 공분산 행렬이 사용됨
- `prcomp`, `princomp`의 결과는 같음

5 주성분 개수 결정 기준

- 성분들이 설명하는 분산의 비율
 - 누적 분산 비율을 확인하면 주성분들이 설명하는 전체 분산 양을 알 수 있음
 - 누적 분산 비율이 70~90% 사이가 되는 주성분 개수 선택

- 고윳값(Eigenvalue): 분산의 크기(=중요도 기준)를 나타내며, 고윳값이 1보다 큰 주성분만 사용함
- 스크리 플롯(Scree Plot): 고윳값을가장 큰 값에서 가장 작은 값을 순서로 정렬해 보여줌(1보다 큰 값 사용 또는 Elbow 기법)

```
fit <- prcomp(USArrests, scale=TRUE)
summary(fit)
```

Importance of components:

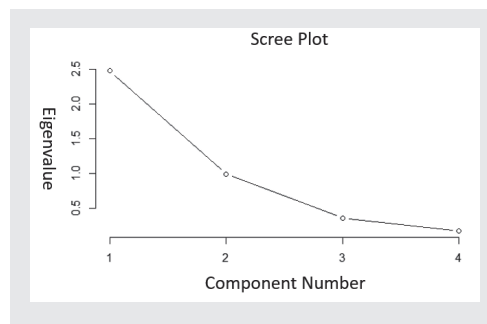
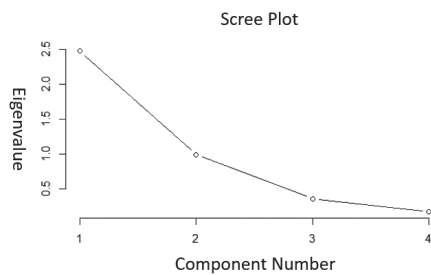
	PC1	PC2	PC3	PC4
Standard deviation	1.5749	0.9949	0.59713	0.41645
Proportion of Variance	0.6201	0.2474	0.08914	0.04336
Cumulative Proportion	0.6201	0.8675	0.95664	1.00000

fit\$rotation

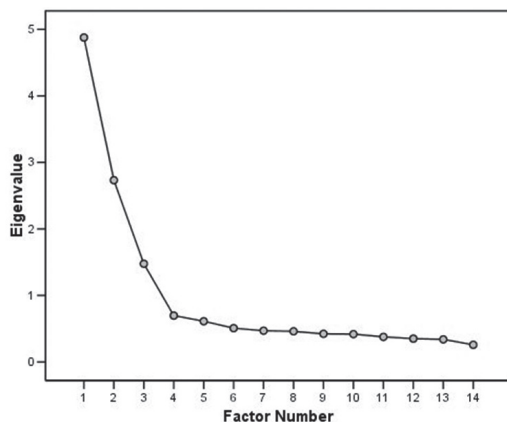
	PC1	PC2	PC3	PC4
Murder	-0.5358995	0.4181809	-0.3412327	0.64922780
Assault	-0.5831836	0.1879856	-0.2681484	-0.74340748
UrbanPop	-0.2781909	-0.8728062	-0.3780158	0.13387773
Rape	-0.5434321	-0.1673186	0.8177779	0.08902432

- 제1주성분 식: $PC1 = -0.536Murder - 0.583Assault - 0.278UrbanPop - 0.543Rape$
- 주성분 결과를 보면 PC2의 Cumulative Proportion이 0.8675로 86.75%이므로, 2개의 주성분을 사용함

6 주성분 개수 결정의 예



- Scree Plot을 보면 Eigenvalue가 1보다 큰 것이 1, 2 주성분이므로 2개의 주성분을 사용함



- Scree Plot에서 Eigenvalue가 1보다 큰 것을 사용하려면 3개의 주성분을 사용해야 하고, 팔꿈치 부분(각도가 급격하게 완만해지는 부분)을 찾는 Elbow 기법을 사용하여 최적의 요소 수를 찾으면 4개임

7 주성분 분석 결과 해석

Importance of components:

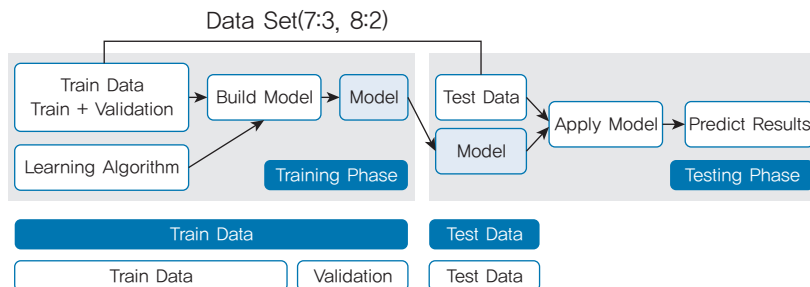
	PC1	PC2	PC3	PC4
Standard deviation	1.5749	0.9949	0.59713	0.41645
Proportion of Variance	0.6201	0.2474	0.08914	0.04336
Cumulative Proportion	0.6201	0.8675	0.95664	1.00000

- Standard deviation(표준편차): 자료의 산포도를 나타내는 수치로, 분산의 양의 제곱근, 표준편차가 작을수록 평균값에서 변량들의 거리가 가까움
- Proportion of Variance(분산 비율): 각 분산이 전체 분산에서 차지하는 비중
- Cumulative Proportion(누적 비율): 분산의 누적 비율
- 첫 번째 주성분 하나가 전체 분산의 62%를 설명함
- 두 번째는 24.7%를 설명함
- 반대로 이야기하면 첫 번째 주성분 부분만 수용했을 때 정보 손실은 $(100-62)=38\%$ 가 됨

최비공노론자
087

모형 평가

1 홀드아웃(Holdout)의 데이터셋 분리



- Training Data: 학습용 데이터
- Test Data: 학습 종료 후 성능 확인용 데이터
- Validation Data: 학습 중 성능 확인용 데이터(Overfitting 여부 확인, Early Stopping 등을 위해 사용), 반복되는 학습 기법에서 사용되며 몇 번의 반복을 하면서 과대적합이 되는지 확인하거나, 과대적합 발생이 감지되면 빠르게 학습을 멈추기 위해 사용하는 데이터

최비공노론자
093

앙상블(Ensemble)모형

1 배깅(Bagging, Bootstrap AGGREGatING)

- 서로 다른 훈련 데이터 샘플로 훈련하며 서로 같은 알고리즘을 사용하는 방법
- 원 데이터에서 중복을 허용하는 크기가 같은 표본을 여러 번 단순 임의 복원 추출하여 각 표본에 대해 모델을 생성하는 기법
- 여러 모델이 병렬로 학습하며 그 결과를 집계하는 방식으로, 같은 데이터가 여러 번 추출될 수도 있고 어떤 데이터는 추출되지 않을 수 있음

2 부스팅(Boosting)

- 이전 모델의 결과에 따라 다음 모델 표본 추출에서 분류가 잘못된 데이터에 가중치(weight)를 부여하여 표본을 추출함
- 여러 모델이 순차적으로 학습
- 맞추기 어려운 문제를 맞추는데 초점이 맞춰져 있고, 다른 앙상블 기법에 비해 이상치(Outlier)에 민감함
- 대표적 알고리즘: AdaBoost, GradientBoost(XGBoost, **Light GBM**) 등

더 보기

Light GBM: Leaf-wise-node 방법을 사용하는 알고리즘

최민준 노트

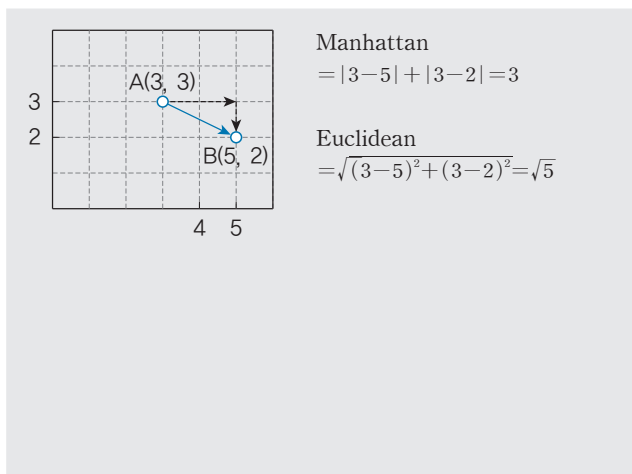
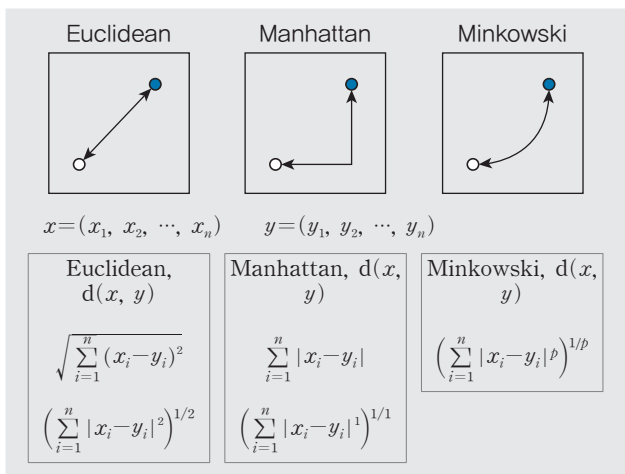
096

군집분석(Clustering Analysis)-계층적 군집

1 계층적 군집(Hierarchical Clustering)의 특징

- 가장 유사한 개체를 묶어 나가는 과정을 반복하여 원하는 개수의 군집을 형성하는 방법
- 유사도 판단은 **두 개체 간의 거리에 기반하므로 거리 측정에 대한 정의가 필요함**
- 유클리드, 맨해튼, 민코프스키, 마할라노비스 등
- **이상치에 민감함**(거리에 기반하는 경우 이상치에 민감한 특징을 갖게 됨)
- **사전에 군집 수 k를 설정할 필요가 없는 탐색적 모형**
- 군집을 형성하는 데 매 단계에서 지역적 최적화를 수행해 나가는 방법을 사용하므로 그 결과가 전역적인 최적해라고 볼 수 없음
- 병합적 방법에서 **한 번 군집이 형성되면 군집에 속한 개체는 다른 군집으로 이동할 수 없음**
- hclust() 함수, cluster 패키지의 agnes(), mclust() 함수 사용

2 수학적 거리의 종류



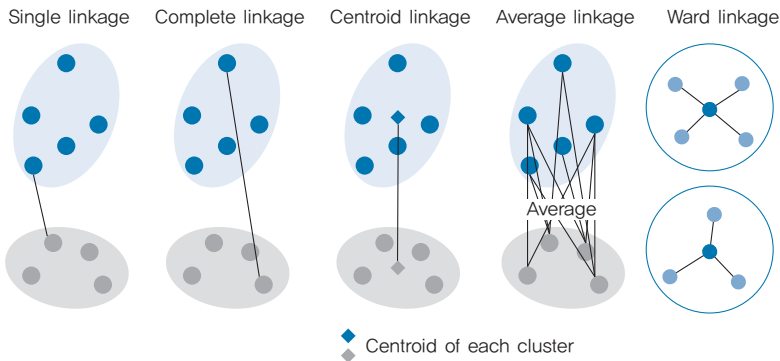
보충 학습

아래 데이터셋에서 a, b 간의 맨해튼 거리는 얼마인가?

구분	a	b
Score	90	80
Time	60	75

문제에서 a, b 라고 했기 때문에, 두 지점은 a, b 이고 Score, Time이 a, b 의 성분입니다. 따라서 a 와 b 의 두 개의 성분의 거리 절댓값을 구하면 Score의 경우 $|90-80|=10$, Time의 경우 $|60-75|=15$ 이므로 $10+15=25$ 가 됩니다.

3 계층적 군집: 응집형(병합 군집) 군집의 종류



- ① 최단연결법(Single Linkage Method, **단일 연결법**): 두 군집 사이의 거리를 군집에서 하나씩 관측값을 뽑았을 때 나타날 수 있는 **거리의 최솟값**으로, 두 관측값을 연결한다.
- ② 최장연결법(Complete Linkage Method, **완전 연결법**): 두 군집 사이의 거리를 군집에서 하나씩 관측값을 뽑았을 때 나타날 수 있는 **거리의 최댓값**으로, 두 관측값을 연결한다.
- ③ 중심연결법(Centroid Linkage Method)
 - 두 군집의 **중심 간의 거리**를 측정하여, 중심끼리 연결함
 - 두 군집이 결합 될 때 새로운 군집의 평균은 가중평균을 통해 구해짐
- ④ 평균연결법(Average Linkage Method)
 - 모든 항목에 대한 거리 평균을 구하면서 군집화함
 - **계산량이 많아질 수 있음**
- ⑤ 와드연결법(Ward Linkage Method)
 - 계층적 군집 내의 **오차(편차) 제곱합(Error Sum of Square)**에 기초하여 군집을 수행하는 군집 방법
 - 크기가 비슷한 군집끼리 병합하는 경향이 있음

1 연관 규칙 측정지표

규칙 표기: $A \rightarrow B$

if A then B \rightarrow A가 팔리면 B가 같이 팔린다.

(1) 지지도(Support)

$$P(A \cap B) = \frac{A \text{와 } B \text{가 동시에 포함된 거래 수}}{\text{전체 거래 수}}$$

- 전체 거래 항목 중 차지하는 비율을 통해 해당 연관 규칙이 얼마나 의미가 있는 것인지를 확인함
- 전체 거래 항목 중 상품 A와 상품 B를 동시에 포함하여 거래하는 비율

(2) 신뢰도(Confidence)

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{A \text{와 } B \text{가 동시에 포함된 거래 수}}{A \text{가 포함된 거래 수}}$$

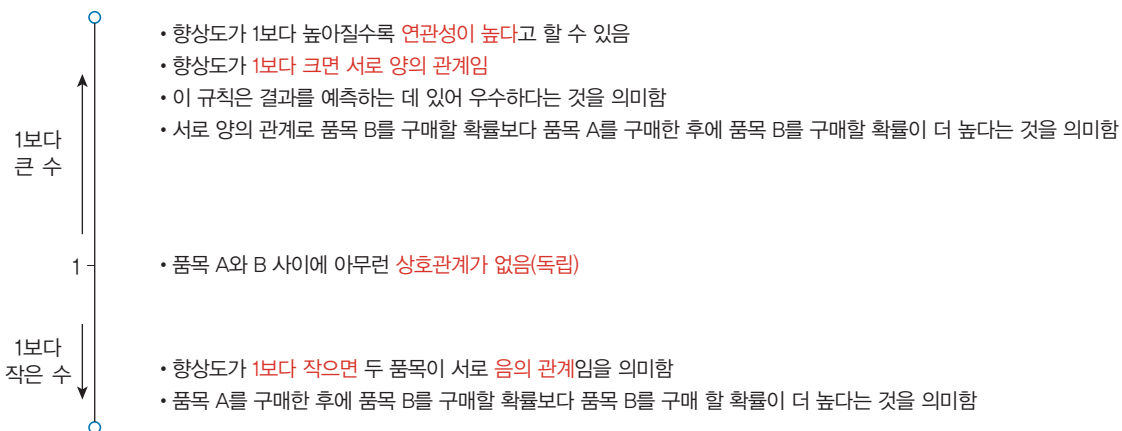
- 상품 A를 구매했을 때 상품 B를 구매할 확률이 어느 정도 되는지 확인함
- 상품 A를 포함하는 거래 중 A와 B가 동시에 거래되는 비율

(3) 향상도(Lift)

$$\begin{aligned} \frac{P(B|A)}{P(B)} &= \frac{P(A \cap B)}{P(A) \times P(B)} = \frac{\text{상품 A의 거래 중 상품 B가 포함된 거래의 비율}}{\text{전체 상품 거래 중 상품 B가 거래된 비율}} \\ &= \frac{A \text{와 } B \text{가 동시에 일어난 확률}}{A, B \text{가 독립된 사건일 때 } A, B \text{가 동시에 일어날 확률}} \end{aligned}$$

- A가 주어지지 않았을 때 B의 확률 대비, A가 주어졌을 때 B의 확률 증가 비율
- 품목 B를 구매한 고객 대비, 품목 A를 구매한 후 품목 B를 구매하는 고객에 대한 확률

2 향상도 해석



1 기계 학습의 분류

인공지능(AI)

머신 러닝(Machine Learning)

딥러닝(Deep Learning)

2 지도 학습(Supervised Learning)

- X를 사용해 Y를 예측할 때, 학습 데이터에 X, Y 데이터가 모두 존재하는 학습
- X를 독립변수, Y를 종속변수라고 하며, Y에는 실제값과 예측값이 존재함
- 회귀(Regression) 모형과 분류(Classification) 모델
 - 회귀(Regression): 예측값이 실제값보다 크거나 작거나 사이 값일 수 있음(연속형 결과), 부모키를 사용해 딸의 키를 예측, 판매량 예측, 집값 예측
 - 분류(Classification): 예측값이 실제값에서 주어진 데이터 범주(종류)로 제한됨(범주형 결과), 화물의 정시 도착 여부 예측, 생존 여부 예측, 품종 예측, 이미지 숫자 예측

3 비지도 학습(Unsupervised Learning)

- ① 학습 데이터에 X에 대한 데이터만 존재한 학습을 의미한다.
- ② 군집(Clustering) 모형과 연관(Association) 모형이 있다.
 - 군집(Clustering): 데이터를 특성에 따라 구분되는 몇 개의 그룹으로 나누는 학습, 고객을 k개의 그룹으로 나눔(그룹 내 서로 유사한 특성 범주형 결과)
 - 연관(Association): 항목들 간의 '조건-결과' 식으로 표현되는 유용한 패턴을 발견하는 것, 삼겹살 → 상추 빵 → 우유(지도도 신 회도 향상도 등으로 연속형 결과)

4 분석모형의 종류

(1) 지도 학습(종속 변수에 따라 나뉨)

회귀(연속형 종속변수)	분류(범주형 종속변수)	회귀+분류
<ul style="list-style-type: none"> • 단순 선형 회귀 • 다중 선형 회귀 • 다항 회귀 	<ul style="list-style-type: none"> • 로지스틱 회귀 • 다중 로지스틱 회귀 	<ul style="list-style-type: none"> • KNN (K-Nearest Neighbors) • Decision Tree • SVM(Support Vector Machine) • Ensemble <ul style="list-style-type: none"> – Bagging – Boosting – Voting – Stacking • 인공 신경망(ANN) <ul style="list-style-type: none"> – 다층퍼셉트론
<ul style="list-style-type: none"> • Linear Regression • Lasso Regression • Ridge Regression ElasticNet 	<ul style="list-style-type: none"> • Logistic Regression 	

(2) 비지도 학습(종속 변수 없음)

군집			연관
계층적	비계층적	비지도 신경망	<ul style="list-style-type: none"> • Apriori • FP-Growth
<ul style="list-style-type: none"> • 단일(최단)연결법 • 완전(최장)연결법 • 평균연결법 • 중심연결법 • 와드(Ward)연결법 	<ul style="list-style-type: none"> • K-means • K-medoids • DBSCAN • EM 알고리즘 • PAM 	<ul style="list-style-type: none"> • SOM 	

5 강화 학습(Reinforcement Learning)

- 인공지능이 어떤 환경에서 특정 목표를 달성하기 위해 시행착오를 통해 스스로 학습하는 방법
- 에이전트라고 불리는 인공지능 시스템이 환경과 상호작용을 함
- 에이전트는 특정 행동을 선택하여 환경에 작용하고, 보상 또는 벌점을 받게 됨

예시 로봇 제어, 자율주행 자동차, 게임 에이전트 등에 사용될 수 있음

- Q-Learning, Deep Q-Network, DQN, 정책 그레이던트(Policy Gradient) 등의 알고리즘이 있음