01

비정형데이터분석

데이터 개념 및 환경변화

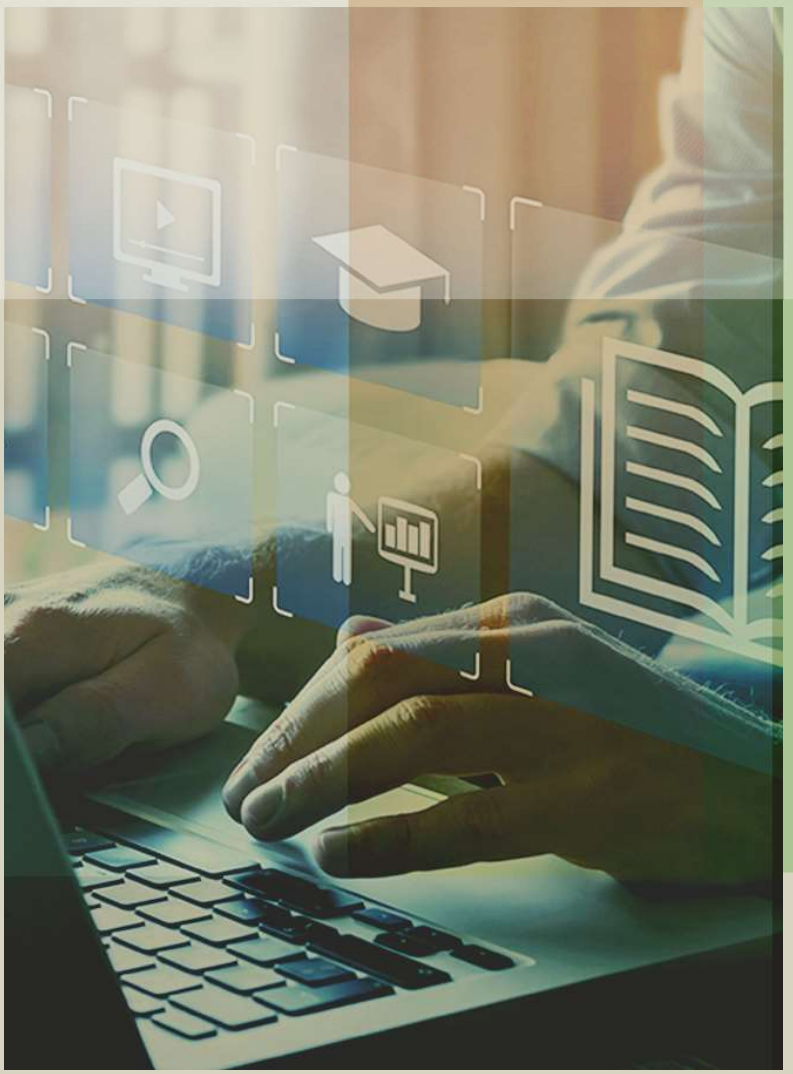
통계·데이터과학과 장영재교수

학습목차

1. 데이터의 개념
2. 데이터 환경의 변화

01

데이터의 개념



1. 데이터의 개념

데이터의 사전적 정의

데이터의 사전적 정의는 다음과 같이 정리할 수 있음

- 이론을 세우는 데 기초가 되는 사실 또는 바탕이 되는 자료
- 관찰이나 실험, 조사로 얻은 사실이나 정보. '자료'의 순화
- (컴퓨터) 컴퓨터가 처리할 수 있는 문자, 숫자, 소리, 그림 따위의 형태로 된 정보

1. 데이터의 개념

데이터의 정의는 역사적 배경과 밀접한 관계

- 과학적 방법의 주요 특징인 경험주의(Empiricism)가 중요시 되면서 구체화
- 감각기관을 통하여 얻은 직접적, 간접적 경험을 바탕으로 지식을 창출하는 방식
- 과학적 방법은 지식 창출의 기초 자료로서 데이터의 개념을 강조
- 컴퓨터가 처리할 수 있는 정보 (기술발전과 환경변화 등으로 인한 새로운 개념)

1. 데이터의 개념

데이터의 어원

데이터의 역사는 인간의 역사와 함께 해 왔음

데이터의 어원적 정의

- 'dare'는 '주다'라는 뜻의 영어 동사 'give'의 의미
- 'dare'라는 동사의 변형이 'datum'이며 이는 'thing given'
- 'data'는 'datum'의 복수형이므로 '주어진 것들' 정도의 의미로 이해

1. 데이터의 개념

역사속의데이터

- 역사 속의 기록에 남은 데이터의 흔적을 찾아 볼 수 있음

  - 선사시대의 많은 기록들을 통해 전달 방법이나 매체가 제한된 상황에서 정보의 시각화가 이루어졌음을 알 수 있음
  - 신석기 시대에 사용되었던 빗살무늬토기는 농경생활 및 정착생활의 흔적
    - 곡식의 저장이 시작되면서 수량적 기록의 필요성

    → 수량 데이터의 기록의 근거
  - 문자가 발명되면서 기록의 흔적이 확대됨
    - 종이의 발명과 기록 보존 방법의 발달은 문자적 기록의 큰 폭 증가를 초래

02

데이터 환경의 변화

2. 데이터 환경의 변화

빅데이터의 출현

현대의 IT 기술의 발전은 데이터 집적과 생성에 매우 큰 변화를 초래

- 인터넷은 정보의 교환과 파급 속도를 증폭
- 사용자들 간의 연결망은 더욱 촘촘해짐
- 새로운 형태의 데이터 생성도 가속화

데이터의 양적 팽창 현상 심화

- 기술의 발전으로 이전의 데이터 생성 규모나 속도와 비교할 수 없을 정도의 대량의 데이터 생성 가능
- 데이터웨어하우스(data warehouse) 기술의 발전으로 대규모 데이터의 집적도 가능

2. 데이터 환경의 변화

데이터의 형태가 매우 다양해짐

- 과거에는 생성 및 수집이 어려웠던 다양한 형태의 데이터 생성과 저장이 쉬워짐
- 모바일 기기 등 사용자가 쉽게 생성할 수 있는 환경
  - 음성이나 이미지, 영상 등의 데이터와 같이 형태가 매우 다양한 데이터 생성
- 데이터를 생성하는 플랫폼도 다양해지고 사용자의 편의성도 개선되면서 더 많은 다양한 데이터의 생성이 가능

2. 데이터 환경의 변화

데이터 생성 주기의 변화

- 휴대용 모바일 기기처럼 개개인이 손쉽게 사용할 수 있는 도구와 인터넷의 결합은 데이터 생성 속도를 급격하게 증가시킴

→ 생성 시간의 단축 (짧은 시간에 대량의 데이터를 생성)

ex) 소셜네트워크서비스

2. 데이터 환경의 변화

- 소셜데이터, 신용카드데이터, 스캐너데이터 등 짧은시간내에생성되는 데이터의종류가증가
  - 이상과 같은 특징을 지니고 생성되는 데이터를 '빅데이터(big data)'라고 정의하고 세 가지 요건을 의미하는 'V'로 시작하는 영단어 Volume(양), Variety(다양성), Velocity(속도) 등을 빅데이터의 3요소 3V라고 지칭

  - 3V에 'Veracity'를 더하여 4V로 빅데이터의 특징을 나타내기도 하는데, Veracity는 정합성을 의미하는 단어로서 이는 빅데이터 분석에 있어서 정확성을 기하고 분석할 만한 가치가 있는지 판단할 필요가 있다는 점을 강조한 것임

2. 데이터 환경의 변화

비정형 데이터

- 빅데이터는 다양한 형태로 나타나지만 이를 크게 분류해 보면 정형 데이터와 비정형 데이터로 나눌 수 있음

• 빅데이터 중 비정형 데이터의 사례

→ 페이스북, 트위터와 같은 소셜 미디어나 웹 공간의 블로그나 게시판, 또는 인스타그램이나 유튜브와 같은 매체를 통한 사진, 동영상 데이터

• 빅데이터 중 정형 데이터의 사례

→ 신용카드 사용 데이터나 공공기관의 축적된 공공데이터, 기업 내부 데이터, 컴퓨터 시스템이나 센서를 통해 축적되는 소위 사물 인터넷(Internet of Things; IoT)으로 인한 데이터

2. 데이터 환경의 변화

- 시장조사기관IDC(International Data Corporation)에 따르면 빅데이터 시대에 접어들면서 전체데이터의 80% 이상을 비정형 데이터가 차지
  - 소셜 미디어 사용량도 증가하면서 비정형 데이터는 더욱 증가
  - 데이터의 연결성, 내재한 중요 정보 등 정형 데이터가 갖지 못한 숨겨진 가치가 있음

→ 비정형 데이터 분석 방법들이 개발되고 또 지속적으로 개선(기존의 정형화된 틀로 분석하기 상당히 어렵기 때문)

2. 데이터 환경의 변화

- 머신러닝이나 데이터마이닝 분류에 있어서도 비정형 분석방법을 별도 범주로 규정
  - 전통적으로는 컴퓨터가 학습하는 데이터의 형태에 따라 관리학습 (지도학습 또는 감독학습)과 자율학습(비지도학습)으로 분류
    - → 학습의 교사역할을 하는 출력변수(종속변수)의 존재 유무로
  - 비정형 데이터의 분석은 명확하게 관리학습이나 자율학습의 범주로 나누기 어려우므로 별도로 구분
    - → 대표적인 비정형 데이터 분석기법은 텍스트 마이닝, 연결망 분석 등

관리학습

분류분석

- 판별분석
- 로지스틱회귀분류
- 최근접이웃기법
- 의사결정나무
- 나이브베이즈분류
- 신경망
- 지지도벡터기계

예측분석

- 회귀분석
- 최근접이웃기법
- 신경망
- 평활법

자율학습

군집분석

- K-평균
- 계층적 군집분석
- 유한혼합모형
- 이중군집법

연관분석

- 장바구니분석
- 서열분석
- 트랜잭션데이터분석

비정형분석

- 텍스트마이닝, 사회연결망분석

<그림> 데이터마이닝 기법의 분류

다음시간안내

02

데이터 활용의 제도적 장치 및 유의점