1강. 데이터 마이닝이란

통계·데이터과학과 장영재 교수

한국방송통신대학교

01 데이터 마이닝의 개념
02 데이터 마이닝 사례
03 R을 이용한 실습

1. 데이터 마이닝의 개념

01 데이터 마이닝의 개념

1 데이터 마이닝

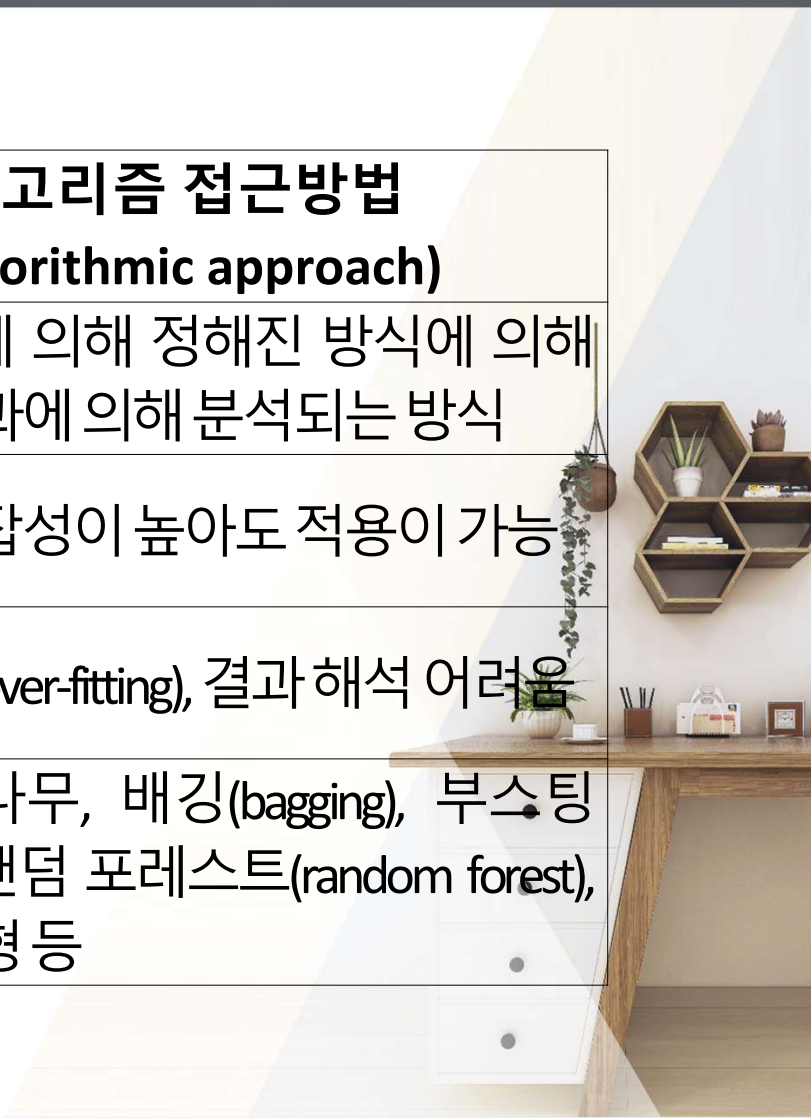
- 데이터 마이닝이란 대용량의 데이터로부터 이들 데이터 내에 존재하는 관계, 패턴, 규칙 등을 탐색하고 모형화함으로써 유용한 지식을 추출하는 일련의 과정

- 기술의 발달로 대량의 데이터 집적이 가능해지면서 활용범위 확대 (빅데이터 환경)
- 거대한 데이터 집적 뿐만 아니라 실시간 분석으로 가치 창출
- 데이터 마이닝을 이해하기 위해 모수적 모형과 알고리즘 접근 방법을 비교해 볼 필요

01 데이터 마이닝의 개념

1 데이터 마이닝

접근 방법	모수적 모형 접근방법 (Parametric modeling approach)	알고리즘 접근방법 (Algorithmic approach)
특징	단순 선형 회귀분석 Y=a+bx와 같이 모수 a와 b를 과거 데이터로부터 적합(fitting)	알고리즘에 의해 정해진 방식에 의해 계산된 결과에 의해 분석되는 방식
장점	결과의 해석이 대체로 용이하며, 결과가 복잡하지 않음	데이터 복잡성이 높아도 적용이 가능
단점	가정이나 설정한 식에 부합하지 않는 데이터의 경우, 정확도 등의 성능이 낮음	과도적합(over-fitting), 결과 해석 어려움
해당 방법	선형 회귀분석, 로지스틱 회귀모형 등	의사결정나무, 배깅(bagging), 부스팅(boosting), 랜덤 포레스트(random forest), 신경망 모형 등



01 데이터 마이닝의 개념

2 데이터 마이닝의 특징 및 관련 분야

(1) 데이터 마이닝의 특징

- 대용량의 관측 가능한 자료
- 컴퓨터 중심의 기법
- 경험적 방법이 중시되는 특징(일반화와 관련)
- 일반화(generalization)의 특징
- 다양한 분야의 업무에 활용하여 의사결정에 도움

2 데이터 마이닝의 특징 및 관련 분야

(2) 데이터 마이닝 관련 분야

① **KDD(Knowledge Discovery in Database)**
데이터베이스 안에서의 지식발견 과정 :
데이터 웨어하우징(data warehousing),
OLAP(On-Line Analytical Processing) 등도
넓은 의미에서 KDD의 한 과정이라고 할 수 있음

② **기계학습(Machine Learning)**
인공지능(Articial intelligence)의 한 분야로서
입력되는 자료를 바탕으로 기계(컴퓨터)가 판단을
할 수 있는 방법에 대한 연구가 진행

2 데이터 마이닝의 특징 및 관련 분야

(2) 데이터 마이닝 관련 분야

③ **패턴인식(Pattern Recognition)**
거대한 자료로부터 일정한 패턴을 찾아가는 과정으로
이미지 분류와 깊은 관련이 있다. 통계학의 판별 및 분류
분석과 유사

④ **통계학**
데이터 마이닝을 한마디로 데이터 분석 및 예측모형 적합이라고 할
수 있으므로 기존의 통계학 틀에서 크게 벗어난
것이 없다고 할 수 있으며 데이터 마이닝에서 활용되는
모형은 이미 통계학의 유연한 함수추정 분야에서 다루고 있는 내용

3 데이터 마이닝 기법의 구분

- 데이터 마이닝에서 사용되는 기법은 크게 지도학습(supervised learning)과 자율학습(unsupervised learning)으로 나눌 수 있음

- 지도학습의 목표는 입출력 간의 관계를 결정하는 시스템에 대한 유용한 근사 시스템를 구하는 것으로 정의할 수 있음

- 자율학습에서는 '교사'의 역할에 해당하는 실제 출력값이 존재하지 않음

  - 데이터에 존재하는 여러 가지 형태의 특징을 찾는데 그 목표를 둠

01 데이터 마이닝의 개념

3 데이터 마이닝 기법의 구분

지도학습(감독학습)

<분류>
판별분석
로지스틱 회귀분석
의사결정나무
신경망
앙상블 기법
서포트벡터머신

<회귀>
회귀분석
회귀나무
신경망
앙상블 기법

자율학습

<군집분석>
계층적 군집분석
- 응집분석
- 분할분석

비계층적 군집분석
- K-평균 군집분석

<연관성 분석>
장바구니 분석
순차적 장바구니 분석

<가중치 결정>
신경망

비정형분석

텍스트마이닝
사회연결망 분석

<그림1> 데이터 마이닝 기법의 구분

01 데이터 마이닝의 개념

4 데이터 마이닝의 수행 단계

- 일반적으로 데이터 마이닝의 수행 단계는 <그림 2>와 같음

<그림 2> 데이터 마이닝의 수행단계
출처: Shmueli 등((2010).『Data Mining for Business Intelligence』

2. 데이터 마이닝 사례

02 데이터 마이닝의 사례

1 데이터 마이닝 활용 분야

- 데이터 마이닝 기법은 범용 방법론을 제공하고 있으므로 그 활용분야도 매우 다양하고 제한이 없음

(1) 고객관계관리(customer relationship management; CRM)

- 데이터베이스의 고객정보를 토대로 마케팅 도구로 활발하게 이용

  목표마케팅(target marketing), 고객세분화(segmentation), 고객성향 변동분석(churn analysis), 교차판매(cross selling), 장바구니 분석(market basket analysis) 등

02 데이터 마이닝의 사례

1 데이터 마이닝 활용 분야

(2) 신용평가

- 특정인의 과거 거래 내역을 바탕으로 신용거래 대출한도를 결정

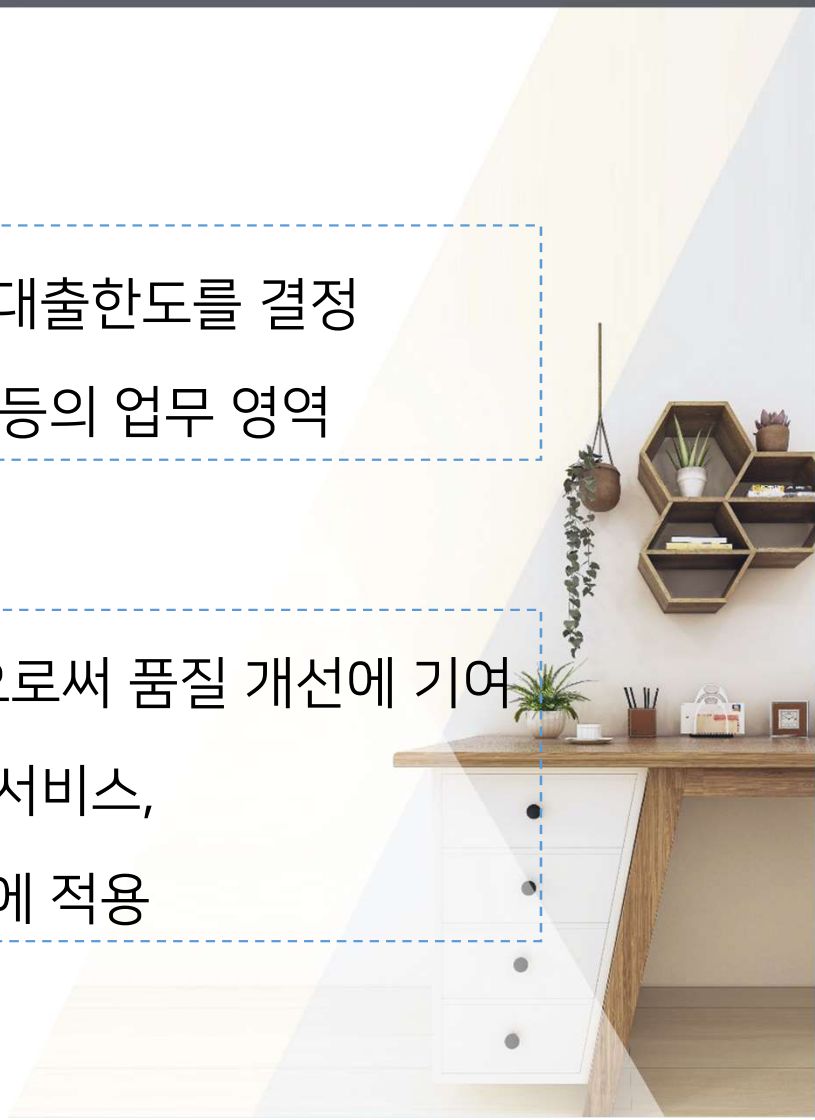
  신용카드, 주택할부금융, 소비자대출, 상업대출 등의 업무 영역

(3) 비즈니스 프로세스의 혁신

- 이상치에 해당하는 불량품을 적절하게 판별함으로써 품질 개선에 기여

  제조업 제품 생산 활동, 음원 및 영상 등 미디어 서비스,

  대중교통 서비스, 세관에서의 통관 프로세스 등에 적용



02 데이터 마이닝의 사례

1 데이터 마이닝 활용 분야

(4) 부정행위 적발

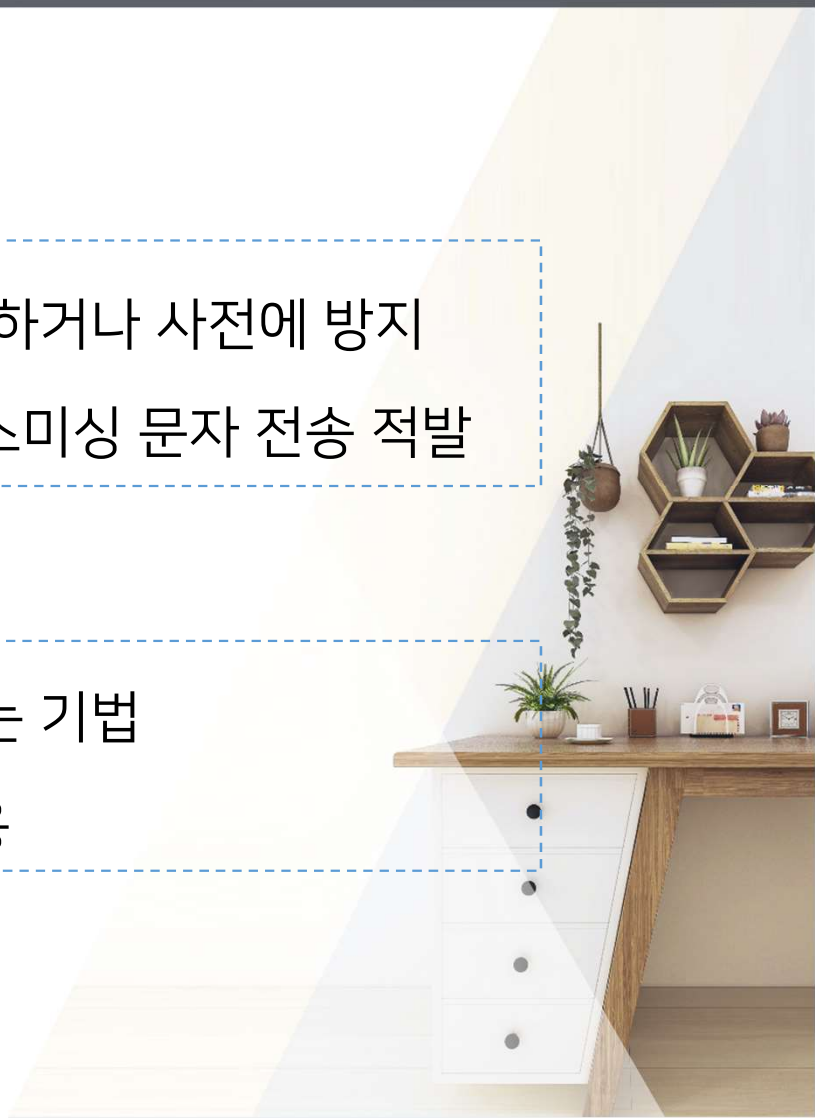
- 사기행위를 발견할 수 있는 패턴을 파악해 적발하거나 사전에 방지

  신용카드 거래사기, 보험금의 허위·과다 청구, 스미싱 문자 전송 적발

(5) 이미지 분석

- 디지털화된 영상, 사진으로부터 패턴을 추출하는 기법

  천문학, 문자인식, 의료진단, 방위산업 등에 활용



02 데이터 마이닝의 사례

1 데이터 마이닝 활용 분야

(6) 생명정보학(Bioinformatics)

- IT 및 생명공학 기술 발전으로 인체로부터 비롯된 인간유전체 등 대량의 데이터를 축적하고 분석할 수 있는 토대를 마련
- 유전자 서열 데이터를 분석해 유전체 각 부분의 기능을 판단하고 예측

3. R을 이용한 실습

03 R을 이용한 실습

1 R의 기초

(1) R이란

- R이란 데이터 분석과 그래프 작성 등을 위하여 개발된 오픈 소스 데이터 분석용 프로그램

  - CRAN(Comprehensive R Archive Network) 사이트를 통해 최신 버전을 다운로드 할 수 있음(http://www.r-project.org)

(2) 패키지(Package)

- 데이터 마이닝과 같은 특화된 분석을 실시하기 위해서는 R에서 제공하는 패키지를 설치가 필요

  - 패키지란 특정 분석을 수행할 수 있는 함수, 객체, 도움말, 데이터 등의 집합을 의미

03 R을 이용한 실습

2 실습 데이터

- 본 교재에서는 데이터 분석 예제 구성과 모형평가를 위해 2장에서 6장까지 공통되는 데이터를 사용

- 목표변수가 연속형인 회귀모형 예제
  : Rahim et al.(2021)의 의류생산성데이터(Productivity Prediction of Garment Employees Data Set)
- 목표변수가 범주형인 분류의 예제 : 와인품질데이터(Wine Quality Data)

03 R을 이용한 실습

2 실습 데이터

(1) 의류생산성데이터(Productivity Prediction of Garment Employees Data Set)

- UCI(University of California, Irvine) 머신러닝 저장소 (Machine learning repository)에 탑재

- Rahim et al.(2021)에서 사용한 데이터로서 실제 의류 업체의 작업과정을 관측하고 생산성을 측정하여 정리한 데이터

➡ 범주형 변수의 변환, 이상치 제거 등 데이터 전처리 수행하여 사용하기로 함

03 R을 이용한 실습

2 실습 데이터

(1) 의류생산성 데이터(Productivity Prediction of Garment Employees Data Set)

변수명	속성	변수 설명
date	수치형	날짜
quarter	범주형	해당 월 중 날짜가 속한 주차
department	범주형	소속 부서
day	범주형	요일
team	범주형	작업에 결부된 팀 번호
target	수치형	목표생산성
smv	수치형	소요시간
wip	수치형	진행 중 작업 아이템 수
over_time	수치형	초과근무 시간(분)
incentive	수치형	인센티브
idle_time	수치형	외부요인으로 인해 생산이 중단된 시간
idle_men	수치형	생산 중단으로 인해 발생된 유휴 인력
numchange	수치형	제품의 스타일 변경 수
numworkers	수치형	각 팀에 배정된 작업 인력
productivity	수치형	실제 작업으로 인해 측정된 생산성

- 이상치 제거 후 idle_time, idle_men, numchange 등의 값은 사라지고 date 변수도 의미가 없으므로 제외

2 실습 데이터

(2) 와인품질 데이터(Wine Quality Data)

- UCI(University of California, Irvine) 머신러닝 저장소 (Machine learning repository)에 탑재

- 목표변수를 변형하여 분류의 예제에 많이 활용
- 원 데이터의 목표변수는 0점부터 12점까지 품질을 평가한 수치형 변수이지만, 이를 6점 이상은 우수, 미만은 보통 등 2개의 범주로 변형하여 범주형 변수를 생성

  ➡ 이상치 제거 등 전처리 수행하여 사용하기로 함

03 R을 이용한 실습

2 실습 데이터

(2) 와인품질 데이터(Wine Quality Data)

변수명	속성	변수 설명
fixed	수치형	고정산(fixed acidity)
volatile	수치형	휘발산(volitile acidity) : 냄새와 관련된 산
citric	수치형	구연산(citric acidity) : 감귤향과 같은 상쾌함
residsugar	수치형	잔여 당분(resid sugar) : 단맛과 관련
chlorides	수치형	염화물 : 짠맛과 관련되어 산도에 영향
freeSD	수치형	유리이산화황(free sulfur dioxide) : 산화 및 갈변 방지, 살균작용(가스형태)
totalSD	수치형	총이산화황(total sulfur dioxide) : 유리상태와 결합상태를 합친 것. 결합상태는 항미생물, 항산화작용 없음
density	수치형	밀도 : 목 넘김 시 묵직함과 관련
pH	수치형	산도
sulphates	수치형	황산염 : 유리아황산의 결합, 비가역적 산화로 생성
alcohol	수치형	알코올 향
quality	범주형	와인의 품질

- 이상치 제거 후 목표변수는 549개의 보통(0), 645개의 우수(1) 관측값을 지님

2 실습 데이터

(3) 가변수 생성

- 패키지에 따라서는 factor함수를 이용하여 생성한 범주형 변수가 적절하게 사용되지 못하는 경우도 있음
- 가변수를 생성하는 dummy라는 패키지가 유용
  - 5장 신경망 모형에서 사용하는 의류생산성데이터의 변환에 이용
  - factor 함수 등을 통해서 범주화를 거친 변수들, 또는 수치형임 문자형으로 인식된 변수들은 as.numeric 함수로 수치화 변환 과정 필요

다음시간안내

2강.
회귀모형 I