# 데이터마이닝
## (Data Mining)

한국방송통신대학교
정보통계학과 장영재 교수

1 강

# 데이터과학과 데이터 마이닝

# 목차

# 1. 데이터 과학

# 데이터과학

## 1) 데이터과학의 개념

- 빅데이터 시대에 접어들면서
  데이터과학의 중요성이 날로 부각되고 있음

- 빅데이터 시대 이후의 데이터 과학은 데이터 분석에 국한하지 않고 IT기술의 접목을 통해 데이터의 크기나 형태에 상관없이 데이터의 가치를 찾는 능력을 의미

✓ 통계학적 사고와 이해는 데이터과학의 발전에 필수적인 요소로 작용

# 데이터과학

## 1) 데이터과학의 발전과정

### (1) 데이터 수집 및 분석의 역사

- 현대적 의미의 데이터 분석은 영국의 로널드 피셔경(Sir Ronald Fisher)으로부터 비롯됨
  - 로담스테드 연구소에서 수행한 수년간의 곡물 관련 실험결과를 연구하며 실험계획법의 토대를 마련
- 조지 갤럽(George Gallup)은 사람들의 생각을 질문이나 문항을 통해 데이터로 수집하기 시작
- 공학자였던 윌리엄 페어(William Fair)와 수학자인 얼 아이작(Earl Issac)은 현대적 의미의 고객 행동을 예측하여 의사결정과 연계한 개념을 최초로 소개
- 공학자이자 통계학자인 다구찌 박사는 품질관리의 새 지평을 열었음

# 데이터과학

## 1) 데이터과학의 발전과정

### (2) 데이터과학의 발전 배경

#### ▮ 데이터과학의 발전은 IT 기술이 발달이 밑거름이 됨

- 데이터 분석을 기업에 본격적으로 도입하게 된 계기는 데이터웨어하우스(data warehouse)의 보급과 도입
  - 데이터과학의 발전을 위해서는 데이터웨어하우스 구축과 더불어 능력 있는 데이터 분석가와 다양하며 복잡한 형태의 데이터를 다를 수 있는 새로운 데이터의 분석 방법이 필요

  → **데이터마이닝의 등장 배경**

# 데이터과학

## 1) 데이터과학의 발전과정

### (3) 데이터과학과 데이터 과학자

- 코펜하겐 대학의 전산학 교수인 피터 나우어 교수는 전산과학(computer science)을 대신하여 데이터과학이란 용어를 사용(1974년)
- 통계학자인 제프 우 교수(Jeff Wu)의 "통계=데이터과학? "이라는 제목의 강연(1997년)
- 퍼듀대학교 윌리엄 클리브랜드 교수는 데이터과학은 더 향상된 데이터 분석을 위해 통계학이 전산학과 융합하며 학습의 영역을 확장해 나가는 과정이라고 소개

→ 진정한 데이터 과학자가 되기 위해서는 데이터마이닝과 같은 통계학이나 전산지식의 기술적인 측면뿐만 아니라 다양한 경험과 함께 스토리텔링 능력, 문제 해결의 의지, 시각적 전달에 필요한 디자인 감각 등 종합적인 능력이 요구

# 2. 데이터마이닝

# 2. 데이터마이닝

## 1) 데이터마이닝의 개념

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

- 데이터마이닝을 이해하기 위해 모수적 모형과 알고리즘 접근 방법을 비교해 볼 필요

# 2. 데이터마이닝

## 1) 데이터마이닝의 특징 및 관련 분야

① 대용량의 관측 가능한 자료를 다룸

② 컴퓨터 중심의 기법으로서 경험적 방법이 중시

③ 통계학과 인공지능을 위주로 한 컴퓨터공학 분야에서 방법론을 개발하고 이를 경영, 경제, 정보기술(IT) 등 다양한 분야의 업무에 활용하여 의사결정을 도움

# 2. 데이터마이닝

## 1) 데이터마이닝 관련 분야

① **KDD (Knowledge Discovery in Database)**
데이터베이스 안에서의 지식발견 과정 :
데이터 웨어하우징 (data warehousing),
OLAP (On-Line Analytical Process-ing) 등도
넓은 의미에서 KDD의 한 과정 이라고 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. 데이터마이닝

## 1) 데이터마이닝 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**
데이터마이닝을 한마디로 데이터 분석 및 예측모형 적합이라고 할 수 있으므로 기존의 통계학 틀에서 크게 벗어난 것이 없다고 할 수 있으며 데이터마이닝에서 활용되는 모형은 이미 통계학의 유연한 함수추정 분야에서 다루고 있는 내용

# 2. 데이터마이닝

## 1) 데이터마이닝 기법의 구분

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

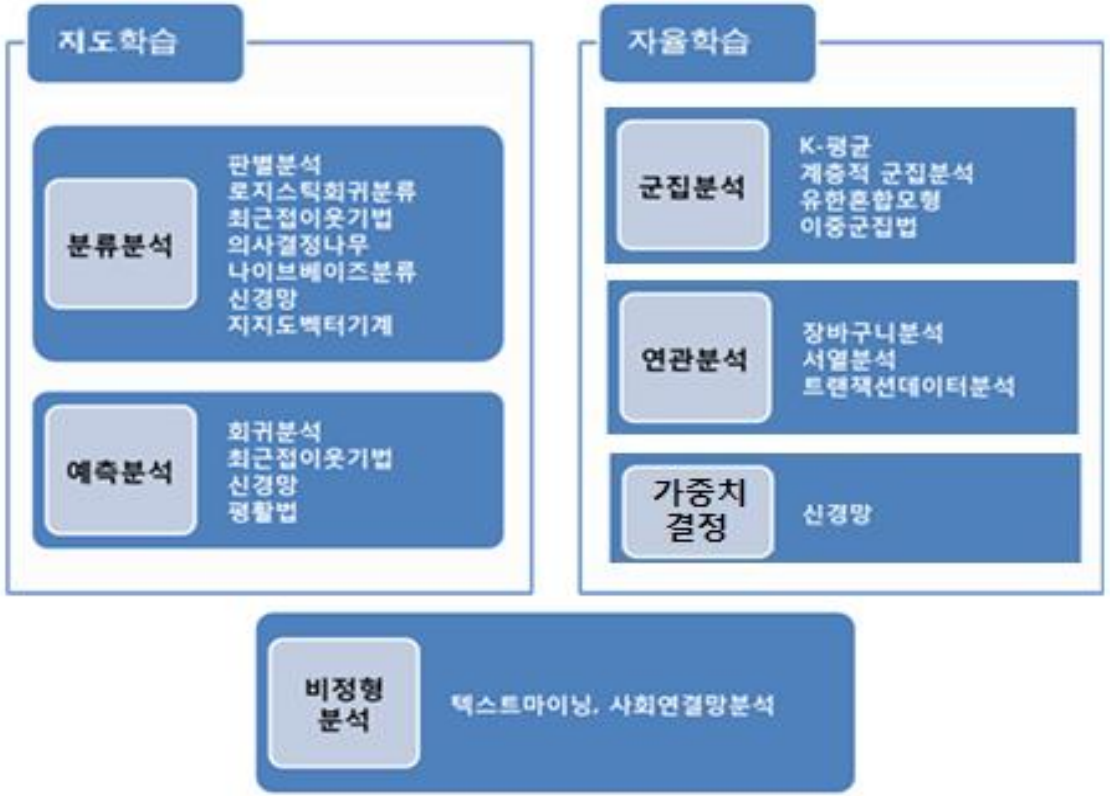
- 지도학습의 목표는 입출력 간의 관계를 결정하는 시스템에 대한 유용한 근사 시스템을 구하는 것으로 정의할 수 있음

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

  - 데이터에 존재하는 여러 가지 형태의 특징을 찾는데 그 목표를 둠

# 2. 데이터마이닝

## 3) 데이터마이닝 기법의 구분




〈그림1〉 데이터마이닝 기법의 구분

# 2. 데이터마이닝

## 4) 데이터마이닝의 수행 단계

- 일반적으로 데이터마이닝의 수행 단계는 〈그림 2〉와 같음


〈그림 2〉 데이터마이닝의 수행단계
출처: Shmueli 등((2010). 『Data Mining for Business Intelligence』

# 2. 데이터마이닝

## 5) 데이터마이닝 활용 분야

- 데이터마이닝 기법은 범용 방법론을 제공하고 있으므로 그 활용분야도 매우 다양하고 제한이 없음. 몇 가지 사례는 다음과 같음

① 고객관계관리
② 신용평가
③ 품질개선
④ 부정행위 적발
⑤ 이미지분석
⑥ 생명정보학
⑦ 인터넷 비즈니스

# 3. R을 이용한 실습

# 1. 데이터과학

## 1) 데이터과학의 발전과정

### (1) R이란

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

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

### (2) 패키지 (Package)

- 데이터마이닝과 같은 특화된 분석을 실시하기 위해서는 R에서 제공하는 패키지를 설치가 필요

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

# 1. 데이터과학

## 2) 실습 데이터

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

  - 목표변수가 연속형인 회귀모형 예제를 위해서는 R에도 내장되어 있는 보스턴하우징데이터(보스턴 주택 가격데이터)가 사용되었으며 목표변수가 범주형인 분류의 예제를 위해 독일신용평가데이터(Germa Credit Data) 사용

## 2) 실습 데이터

### (1) R이란

- 보스턴하우징데이터는 R에 기본적으로 설치되어 있는 MASS 패키지에 포함된 데이터

- 데이터의 속성을 파악하기 위해서는 아래와 같은 명령을 수행하면 간략한 설명과 함께 변수별 설명을 찾아볼 수 있음

```
> library(MASS)
> ?Boston
```

〈그림3〉 보스톤하우징데이터의 호출

# 1. 데이터과학

## 2) 실습 데이터

### (2) 독일신용평가데이터 (Germa Credit Data)

- 독일신용평가데이터는 UCI (University of California, Irvine) 머신러닝 저장소 (Machine learning repository)에 탑재되어 있는 데이터

  - 분류의 예제에 많이 활용
    https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)

# 1. 데이터과학

## 2) 실습 데이터

### (3) 데이터 정보 요약 및 변환

- 각 데이터의 변수별 주요 기술통계량을 요약하여 출력하기 위해서는 summary함수를 사용
- 범주형 변수의 경우 factor함수를 이용하여 생성하는 것이 일반적이나 경우에 따라 이러한 범주형 변수를 인식하지 못할 수 있는데, 가변수(dummy variable)을 생성하여 범주형 변수를 적절하게 변환해 주는 dummy라는 패키지가 매우 유용

# 강의를 마쳤습니다.

다음시간에는...