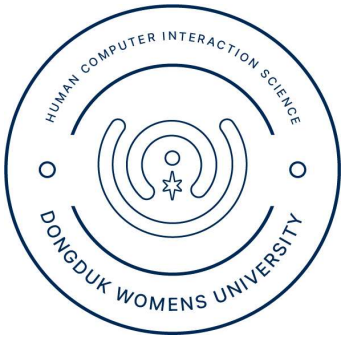
데이터마이닝
- 강의 오리엔테이션 -



2023.03.06 (MON)

동덕여자대학교 HCI사이언스

강의 오리엔테이션

- 담당 교수 : 김원준
- 연구실 및 연락처 : 인문관 621호; 02-940-4766; wjkim@dongduk.ac.kr

- 스마트 캠퍼스 상시 확인 요망

- 강의 교재 : 강의 PPT

- 평가 항목
  - 출석 : 10 %, 팀 과제 : 20 %, 개인 과제 : 40 %, 중간고사 : 30 %
  - 중간고사 미 응시자는 자동 F
  - 모든 과제는 스마트 캠퍼스에 업로드 (딜레이 시 30%감점)
  - 지각 3회 = 결석 1회 (출석은 수업 시간 중 불시에 확인)
  - 학점 구간은 학교 방침에 따름 (zero-coke)
  - **강의 계획서 반드시 확인할 것**

강의 오리엔테이션

- 강의 구성 및 흐름
  - 아래의 주제를 이론과 실습을 병행하여 학습함
    - 데이터 마이닝의 총론
    - 데이터 탐색과 차원 축소
    - 성능 평가
    - 지도 학습 방법
    - 비지도 학습 방법
    - 기타 응용 방법론
  - 수업 운용 방안
    - 수강생들은 팀을 이뤄 수업시간에 배운 이론과 관련된 내용을 실습
    - 기말고사 대체로 개인 프로젝트를 수행
    - 10주차 정도에 중간고사 시행 (일정 변경 가능)

비즈니스 애널리틱스(VS.비즈니스 인텔리전스)

비즈니스 인텔리전스

- 과거 상황과 현재 상황을 이해하기 위한 데이터 시각화 및 보고를 위해 차트, 표, 대시보드를 사용해 데이터를 표현, 검사, 탐색하는 방식으로 수행

비즈니스 애널리틱스

- 비즈니스 인텔리전스뿐만 아니라 복잡한 데이터 분석 방법을 포괄하는 용어
- 복잡한 데이터 분석 방법의 예가 통계 모델과 데이터 마이닝 알고리즘으로, 데이터를 탐색하고, 변수 관계를 측정하거나 설명하며, 변수 값을 예측
- 회귀 모델은 '평균적인' 관계(예: 광고와 매출액)를 기술하거나 정량화하고, 새로운 사건(예: 신규 환자가 치료약에 양성 반응할지 여부)을 예측할 수 있으며, 미래 변수 값(예:다음 주의 웹 트래픽 양) 예상 가능

데이터 마이닝

- **계산, 기술 통계, 보고, 비즈니스 규칙을 기반으로 하는 방법을 뛰어넘는 비즈니스 분석 방법**

- **통계학과 머신러닝(또는 인공지능)의 결합**
  - 통계는 '평균 효과'를 근거로 표본 데이터를 이용해 모집단에 대한 추론에 집중
  - 머신러닝은 '1달러 가격 인상에 대해 A라는 사람의 예상 수요는 한 박스이고, B라는 사람의 예상 수요는 세 박스다'와 같이 개개인의 값을 예측

빅데이터와 데이터 사이언스

빅데이터의 특징

- 볼륨(volume) : 데이터양
- 속도(velocity) : 유동률. 데이터가 생성되거나 변경되는 속도
- 다양성(variety) : 생성되는 다양한 유형의 데이터(화폐, 날짜, 숫자, 텍스트)
- 정확성(veracity) : 데이터가 유기적인 분산 프로세스(수백만 명의 서비스 가입 또는 무료 다운로드)에 의해 생성되며, 어떤 연구를 위해 수집된 데이터에 적용되는 통제 또는 품질 검사의 대상이 아니라는 사실

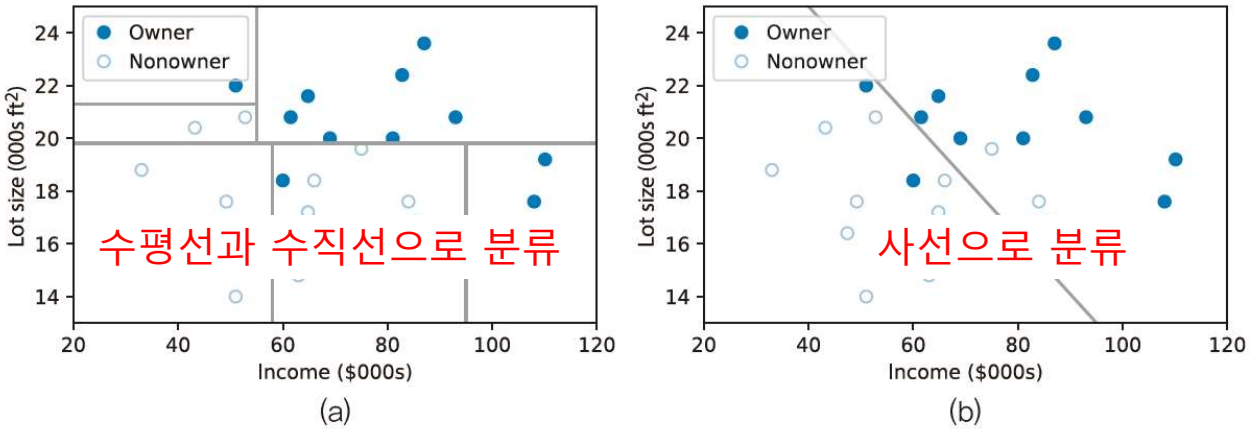
데이터 사이언스

- 통계는 '평균 효과'를 근거로 표본 데이터를 이용해 모집단에 대한 추론에 집중
- 머신러닝은 '1달러 가격 인상에 대해 A라는 사람의 예상 수요는 한 박스이고, B라는 사람의 예상 수요는 세 박스다'와 같이 개개인의 값을 예측

다양한 예측과 분류 방법 존재

- 데이터셋의 크기, 데이터에 존재하는 패턴의 유형,
  데이터가 해당 방법의 일부 기본 가정을 충족하는지 여부,
  데이터의 노이즈 양, 분석의 특정 목표와 같은 요소에 따라 달라질 수 있음

- 예) 승차식 잔디깎이 기계에 대해 비구매자로부터 구매자를 분류하는
  세대 소득 수준과 세대 대지 면적의 조합을 찾기



용어와 표기법

- **P(A|B)**
  - 사건 B가 발생했을 때, 사건 A가 발생할 조건부 확률
- **검증**
  - 데이터 모델이 얼마나 적합한지를 테스트하고, 모델을 조정해 보면서 최적 모델을 결정하는 데 사용한 일부 데이터
- **결과 변수 ‘ 반응’ 참조**
- **관측**
  - 측정된 것들(고객, 거래 등)의 분석 단위. 인스턴스, 표본, 예제, 케이스, 레코드, 패턴 또는 행이라고도 부른다. 스프레드시트에서 각 행은 레코드 하나를 나타내고, 각 열은 변수를 나타낸다. 이 책에서는 '표본'이라는 용어를 사용하는데, 통계학에서 사용하는 의미와 사뭇 다르다. 통계에서는 관측의 집합을 뜻한다.
- **레코드 ‘관측’ 참조**

용어와 표기법

- **모델**
  - 데이터셋에 적용되는 알고리즘
- **반응**
  - 일반적으로 Y로 표기하며 지도 학습에서 예측되는 변수. 종속 변수, 출력 변수, 목표 변수, 결과 변수라고도 부름
- **범주형 변수**
  - 여러 고정값 중 하나를 취하는 변수. 예를 들어 항공편은 정시, 연착, 취소 중 하나를 취함
- **변수**
  - 기록들을 측정하는 척도. 입력 변수(X)와 출력 변수(Y)를 모두 포함

용어와 표기법

- **비지도 학습**
  - 분석 대상의 결과값을 예측하는 것 외에 데이터에 대해 뭔가를 알아보려는 분석
- **성공 클래스**
  - 이진법 결과 내의 관심 클래스(예: 결과 변수 구매/비구매의 구매자)
- **신뢰**
  - 반대의 표본을 선택했을 때의 추정 오차
- **신뢰도**
  - "A와 B를 구매하면, C도 구매한다"라는 형태의 연관 규칙에 관한 측정치. 신뢰는 만약 A와 B를 구매했다면 C도 구매한 것이라는 조건부 확률

용어와 표기법

알고리즘 분류

- 트리, 판별 분석 등 특정 데이터 마이닝 기술을 실행하기 위한 세부 과정

예측

- 연속적인 출력 변수의 예측값. 추정이라고도 함

예측 변수

- 일반적으로 x로 표기하며, 예측 모델에서 입력 변수로 사용된다. 특성, 입력 변수, 독립 변수, 데이터베이스 측면의 필드라고도 함

점수

- 예측된 값이나 계층. 새로운 데이터를 점수화한다는 것은 학습(훈련) 데이터로 개발된 모델을 사용해 새로운 데이터의 결과 값을 예측한다는 의미

용어와 표기법

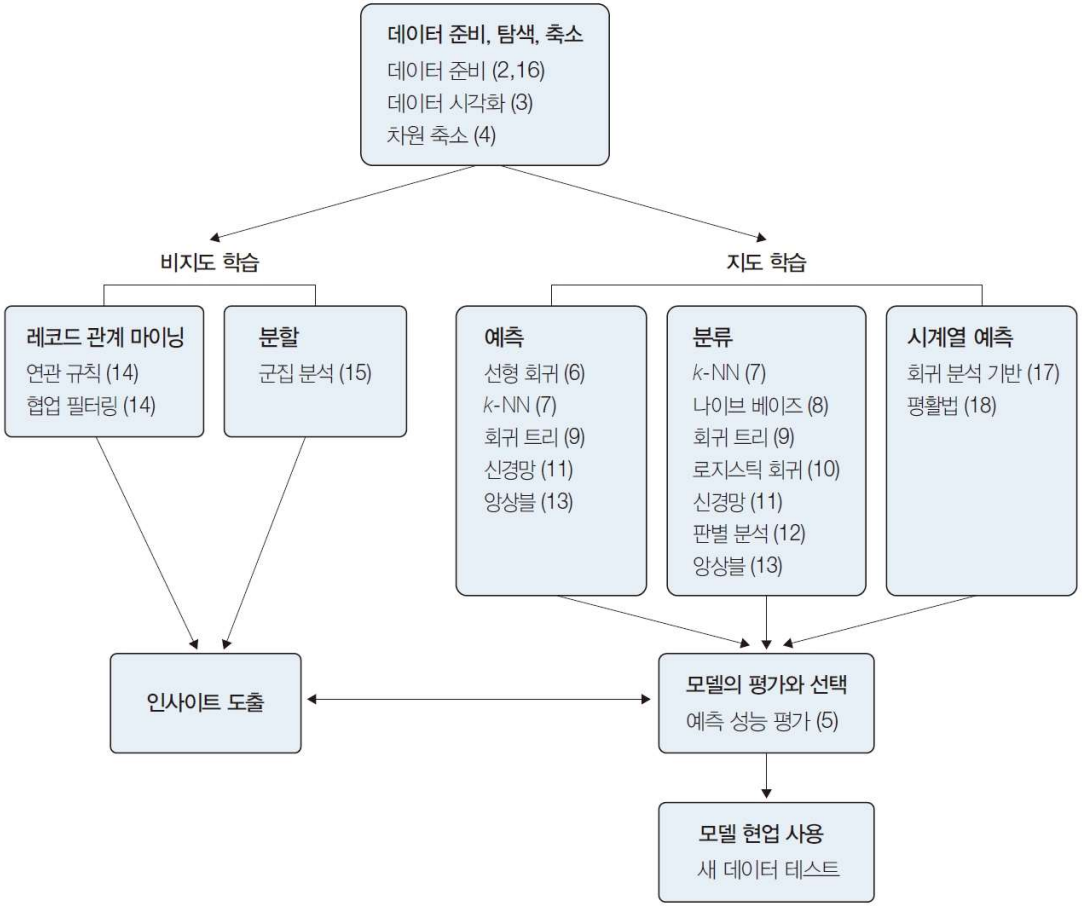
지도 학습

- 결과 변수를 아는 레코드를 알고리즘(로지스틱 회귀, 회귀 트리 등)에 제공하는 과정. 알고리즘은 결과 변수를 모르는 상태에서 새로운 레코드들을 이용해 이 값을 예측하는 방법을 배움

테스트 데이터

- 모델 구축과 선택 과정이라는 마지막 단계에서 최종 모델이 새로 추가되는 데이터를 얼마나 잘 예측하는지를 테스트하는 데 사용

로드맵



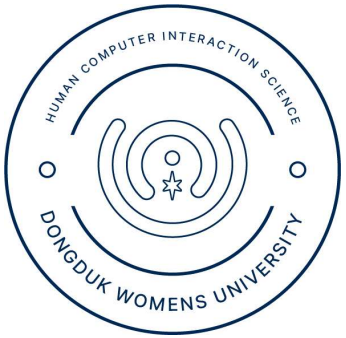
- 데이터 마이닝의 총론과 구성 요소
- 데이터 탐색과 차원 축소라는 초기 단계
- 성능 평가 설명. 예측 수행 척도부터 오분류 비용까지 광범위한 주제를 다룸
- 인기 있는 지도 학습 방법들(분류, 예측)을 다룸. 각 주제는 일반적으로 알고리즘의 난이도, 인기도, 이해도에 따라 구성되고 13장에서는 앙상블과 여러기법의 조합 소개
- 관계를 비지도 학습 방법으로 분석 연관성 규칙과 협업 필터링(14장), 군집 분석(15장) 설명
- 시계열 예측에 중점을 둔 3개 장으로 구성됨. 16장은 시계열의 취급과 이해에 대한 전반적인 이슈를 다룸. 17~18장은 가장 많이 쓰이는 예측 기법 두 가지(회귀 기반 예측과 평활법)를 다룸
- 두 가지 데이터 애널리틱스 주제를 다루는데, 소셜 네트워크 분석과 텍스트 마이닝. 특이한 데이터 구조(소셜 네트워크와 텍스트)에 데이터 마이닝을 적용한 경우
- 관련 사례 모음

로드맵

데이터 속성	지도		비지도
	연속형 변수	범주형 변수	무변수
연속형 예측 변수	선형 회귀 (6) *k*-NN (7) 신경망 (11) 앙상블 (13)	*k*-NN (7) 로지스틱 회귀 (10) 신경망 (11) 판별 분석 (12) 앙상블 (13)	주성분 분석 (4) 협업 필터링 (14) 군집 분석 (15)
범주형 예측 변수	선형 회귀 (6) 회귀 트리 (9) 신경망 (11) 앙상블 (13)	나이브 베이즈 (8) 회귀 트리 (9) 로지스틱 회귀 (10) 신경망 (11) 앙상블 (13)	연관 규칙 (14) 협업 필터링 (14)

Q & A


동덕여자대학교
DONGDUK WOMEN'S UNIVERSITY


HUMAN COMPUTER INTERACTION SCIENCE
DONGDUK WOMENS UNIVERSITY