



6주차: 데이터 분석 방법/데이터 마이닝의 이해

ChulSoo Park

School of Computer Engineering & Information Technology

Korea National University of Transportation



학습목표 (6주차)

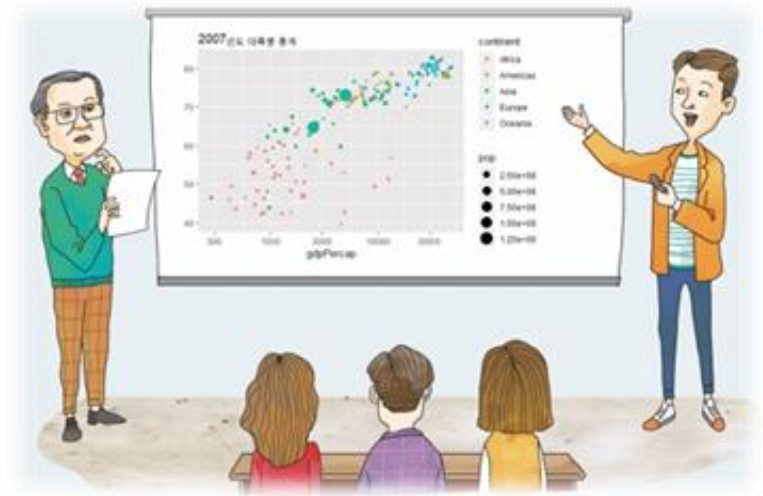
- ❖ 데이터 과학, 데이터 마이닝, 기계 학습 개념 이해
- ❖ 데이터 분석 유형의 이해
- ❖ 데이터 분석 모델 학습
- ❖ 데이터 분석 기법의 이해
- ❖ 데이터 분석 도구 파악
- ❖ 데이터 분석 사례 고찰



07

CHAPTER

데이터 분석 방법 과 데이터 마이닝



데이터 사이언스 개론(김화중,홍릉과학출판사), 데이터 과학 입문(최대우외2명,한국방송통신대학교 출판부)

CONTENTS

개론 7.1 데이터 분석 유형
개론 7.2 데이터 분석 모델
개론 7.3 기계 학습
개론 7.4 분석 모델
개론 7.5 데이터 분석 도구

입문 7.1 데이터 과학에서 마이닝의 역할
입문 7.2 데이터 마이닝의 개념
입문 7.3 데이터 마이닝 관련 분야
입문 7.4 데이터 마이닝 기법 및 도구
입문 7.5 데이터 마이닝 적용 사례



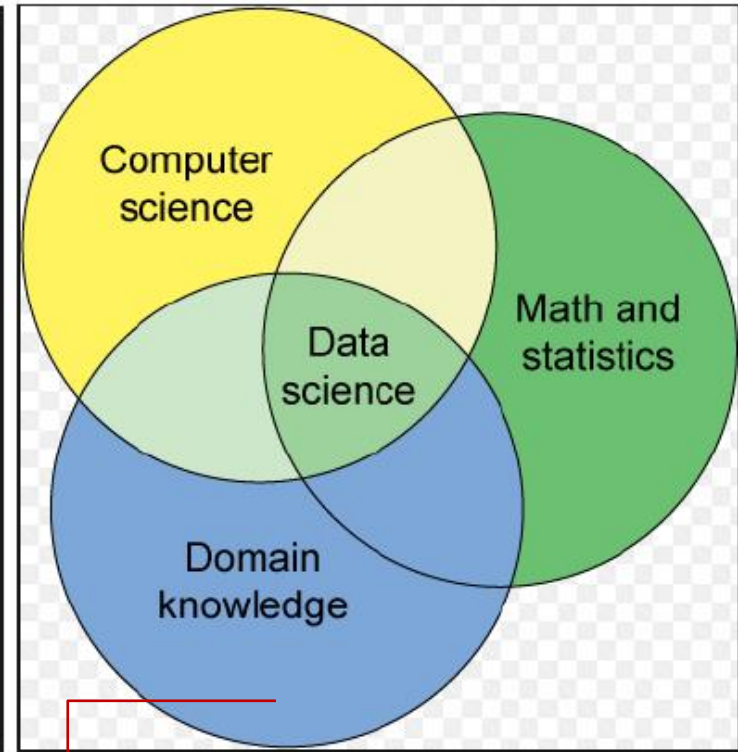
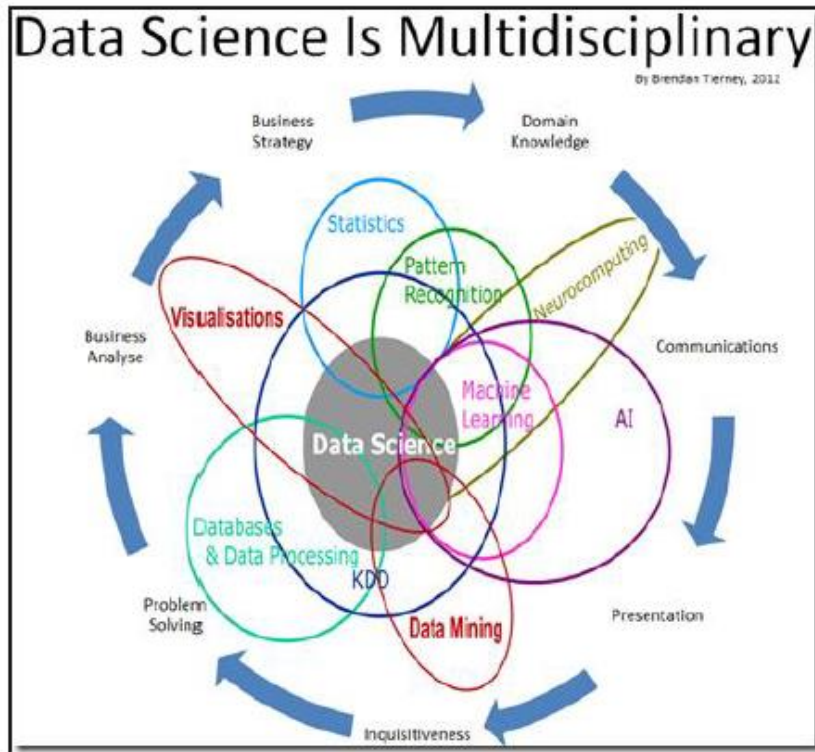
개론7.1 데이터 분석 유형

- 데이터 분석의 목적 : 데이터 사이언스의 시작은 목적을 정확히 정의하는 것에서 시작 된다.
- 축산 ICT : 축산 산업 생산성 20% 향상
- 현대 중공업(digital transformation) : 보이는 공장 + α
- IBM 왓슨(Watson)의 빅데이터 : 수백만 건의 진단서, 환자 기록, 의료서적 등의 방대한 데이터를 바탕으로 왓슨 컴퓨터는 순식간에 분석을 통해 확률이 가장 높은 병명과 성공 가능성이 높은 치료법을 동시에 의사에게 조언을 한다. 그럼으로써 의사마다 다른 판단을 내릴 수 있는 경우 판단의 오차를 크게 줄이고 객관성을 높일 수 있다.



개론7.1 데이터 분석 유형

■ 데이터 사이언스



출처 : www.oralalytics.com

- 축산 ICT : 축산 산업 생산성 20% 향상
- 현대 중공업(digital transformation) : 보이는 공장 + α



개론7.1 데이터 분석 유형

■ 데이터 분석의 유형

- 데이터 사이언스는 데이터 분석(데이터에 숨어 있는 패턴 등 데이터 특징 파악)을 통해 문제를 해결하는 과정
 - 사회 현상 분석, 과학적 연구, 정책 결정, 테러 대비 등등

특정 방법에 의존하지 않고 종합적으로 접근

통계적 모델링

- 샘플 데이터로부터 전체 집단의 의미를 추정하는 통계적 추론
- 평균, 표준편차, 히스토그램 검토
- 가설 검증, 신뢰수준 및 오차, 상관관계 분석
- 회귀 분석, 분산 분석 등 통계 분석 모델 사용
- 모델의 적합성, 가정의 합리성, 신뢰도 등에 관심

기계학습 모델링

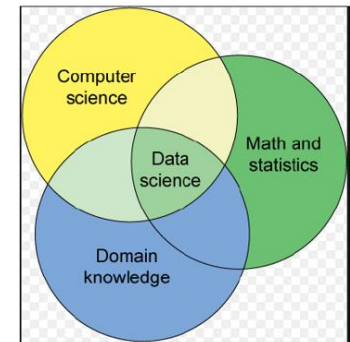
- 인공지능(AI)의 한 분야, 사람의 지능 처럼 추론이 목적은 아님
- 사람이 학습하듯이 점차 성능을 개선 하는 것
- 사람이 정답과 오답을 계속 가르쳐 주면 모델의 정확도가 높아짐
- 모델의 수학적 완성도보다 소프트웨어적 구현에 관심
- 우선 동작하는 시스템 개발, 학습을 통해 성능 개선



개론7.1 데이터 분석 유형

■ 통계적 분석

- 모든 데이터 분석은 통계 분석(statistical analysis)에서 출발한다. 통계의 본질이 “데이터로부터 의미를 찾는 것”이다.
- 샘플에서 전체 집단의 의미 탐색, 즉 수학적 모델을 중요시
- 가설 검증 시 신뢰도와 오차범위 설명 필요
 - 축구 경기의 이길 확률을 70%로 예측
 - 신뢰도가 95%, 오차범위 5%이면
 - 경기에서 이길 가능성은 65~75%내에 발생할 확률이 95%라는 의미임
- 분산 분석의 중요성
 - 마을 A: 168, 169, 170, 171, 172cm (평균 170)
 - 마을 B: 150, 160, 170, 175, 180cm (평균 170)



개론7.1 데이터 분석 유형

■ 대푯값과 분산

구분	마을A	마을B	비 고
사람1	168.00	150.00	
사람2	169.00	165.00	data error
사람3	170.00	170.00	
사람4	171.00	175.00	
사람5	172.00	190.00	data error
합계	850.00	850.00	
평균	170.00	170.00	
분산	2.50	212.50	
표준편차	1.58	14.58	

구분	마을A	마을B	비 고
사람1	165.00	150.00	
사람2	167.00	161.00	
사람3	170.00	164.00	
사람4	171.00	165.00	중앙값
사람5	172.00	176.00	
사람6	172.00	176.00	
사람7	173.00	198.00	
합계	1190.00	1190.00	
평균	170.00	170.00	
중앙값	171.00	165.00	
최빈값	172.00	176.00	
분산	8.67	233.00	
표준편차	2.94	15.26	

- 대푯값 : 평균, 중앙값, 최빈값
- 분산, 표준편차



개론7.1 데이터 분석 유형

■ 심슨 패러독스 (p.169)

- 각 그룹 데이터에서 나타나는 특징이 그룹들이 결합되었을 때 사라지는 현상으로, 이와 반대도 마찬가지로
- 각 부분에서 성립한 특성이 부분들의 합인 전체에서는 성립하지 않는 모순적인 경우를 의미
- 이러한 모순에 빠지지 않으려면
 - ✓ 전체와 부분을 따로 분리해서 검토
 - ✓ 가중 평균의 적용 검토
 - ✓ 핵심 변수 파악
 - ✓ 공정하게 비교하기



개론7.1 데이터 분석 유형

- 심슨 패러독스
- 제품의 불량률 비교

도시	A사	B사
서울	판매량 90 불량품 10 (불량률 10%)	판매량 920 불량품 80 (불량률 8%)
춘천	판매량 980 불량품 20 (불량률 2%)	판매량 99 불량품 1 (불량률 1%)
전체	A사 총 불량률 30/1,100 = 3%	B사 총 불량률 81/1,100 = 8%

- ✓ 서울과 춘천 모두에서는 B사의 품질이 우수
- ✓ 전체를 보면 A사의 품질이 우수



개론7.1 데이터 분석 유형

■ 기계학습(machine learning) 분석

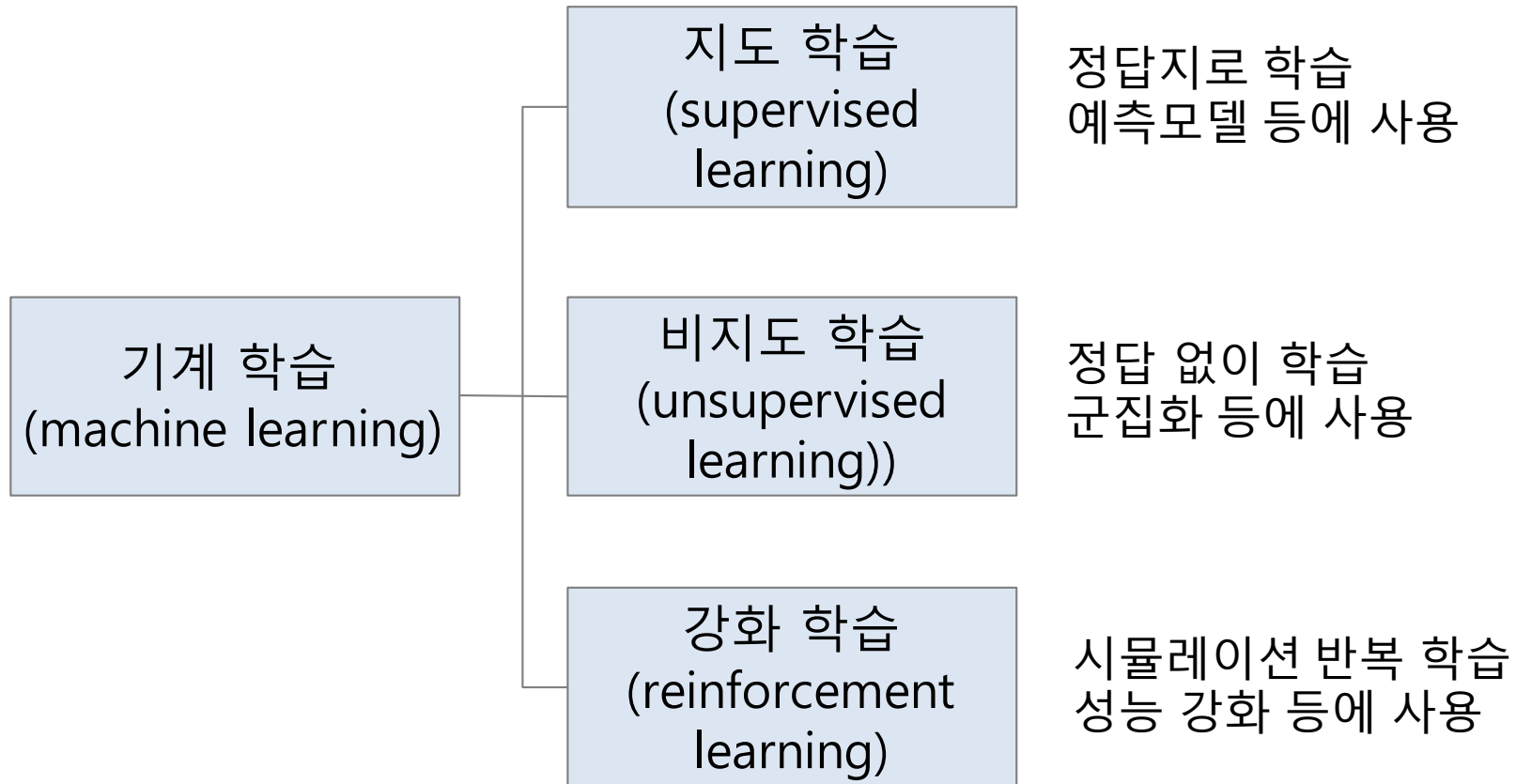
기계 학습 또는 머신 러닝(machine learning)은 경험을 통해 자동으로 개선하는 컴퓨터 알고리즘의 연구이다. 인공지능의 한 분야로 간주된다. 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야이다. 가령, 기계 학습을 통해서 수신한 이메일이 스팸인지 아닌지를 구분할 수 있도록 훈련할 수 있다.

예를 들어 사람이 손으로 쓴 글씨를 컴퓨터가 인식하는 경우에, 글자 인식이 맞으면 다음 단계로 넘어가지만 인식이 틀렸으면 사람이 글자를 다시 고쳐 씀으로써 정답을 가르쳐 주게 된다. 이렇게 사람이 정답과 오답을 계속 가르쳐 주면 컴퓨터의 인식률이 높아지도록 하는 기술이 기계학습이다.



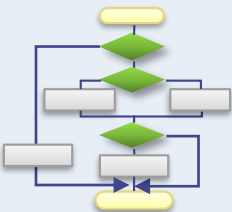
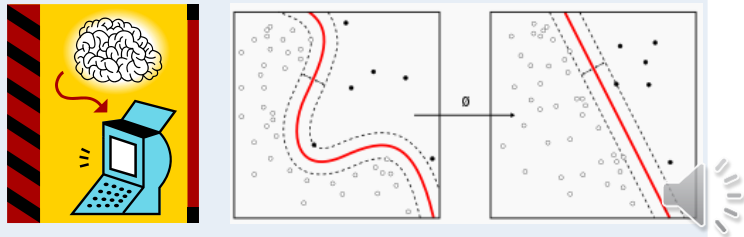
개론7.1 데이터 분석 유형

■ 기계학습(machine learning)



개론7.2 데이터 분석 모델

- **데이터 분석 결과**는 연구 대상에 대한 특징 설명 또는 어떤 값의 예측 형태로 나타남. 이를 위해 **모델링 방식을 사용함**
- 모델링이란 데이터를 발생시킨 원래 시스템을 설명하기 위해 설정한 구조
- 모델 표현 방법은 수식, 다이어그램, 알고리즘 등

수식	다이어그램	알고리즘
가장 명쾌하나 현실적으로 많은 대상이 수식으로 설명하기 어려움	순서도 같은 형태의 처리 흐름도로 알고리즘에 비해 비교적 단순함	패턴 분석과 예측 모델 작성 등 보다 복잡한 논리적 흐름
$F = G \frac{m_1 m_2}{r^2}$ <p>중력 이론</p>		

개론7.2 데이터 분석 모델

■ 모델의 특징 (1/2)

- 학습(training)을 통한 성능 개선 가능
- 모델의 일반화(generalization) 가능
- 잡음(noise)이란 모델링을 할 때 방해가 되는 데이터로 잡음을 제외하고 모델링 해야 함
- 잡음이 포함되는 경우, 과도한 적용(overfitting)이 발생할 수 있으며 일반화가 어려움
- 실제 현상을 단순화하기 위해 수용할 수 있는 가정 필요
 - ✓ 사람의 성격 유형 4, 5, 16, 32가지 등



개론7.2 데이터 분석 모델

■ 모델의 특징 (2/2)

- 적절하고 유용한 모델의 선택
- 타당한 가정 세우기
- 데이터 분석의 성공 여부는 모델의 정확성과 가정의 타당성에 따라 좌우됨
 - ✓ 분석 대상에 대한 이해는 필수적
 - ✓ 사람의 생각과 행동에 대한 이해(인문학, 심리학)
- 기계학습에 기반한 알고리즘을 활용하여 우선 동작하는 모델부터 찾고, 이를 차츰 개선해 나감
- 경험적 heuristic 접근법 채택



개론7.3 기계학습

■ 기계학습의 유형 (1/2) p.174

용어	설명
서술형 모델	<ul style="list-style-type: none"> - 데이터를 분석하여 어떤 현상을 설명하는 모델 - 주어진 데이터의 속성 파악 가능
예측형 모델	<ul style="list-style-type: none"> - 미래에 발생할 데이터 값의 예측 - 분류 예측과 수치 예측이 있음
비지도 학습	<ul style="list-style-type: none"> - 정답이 없는 모델로 정답을 맞추어 볼 수 없음
지도 학습	<ul style="list-style-type: none"> - 예측하려는 변수의 정답을 알 수 있는 경우 - 정답과 비교 후 기계학습 모델의 개선 가능



개론7.3 기계학습

■ 기계학습의 유형 (2/2) p.175

구분	기계학습 유형		대표 알고리즘
비지도 학습 (unsupervised learning)	서술형 (descriptive)	클러스터링	k-means
		연관 분석	패턴 분석
지도 학습 (supervised learning)	예측형 (prediction)	분류 예측	k-NN, 베이어스, 의사결정 트리
		수치 예측	선형 회귀 분석, 회귀 트리, SVM

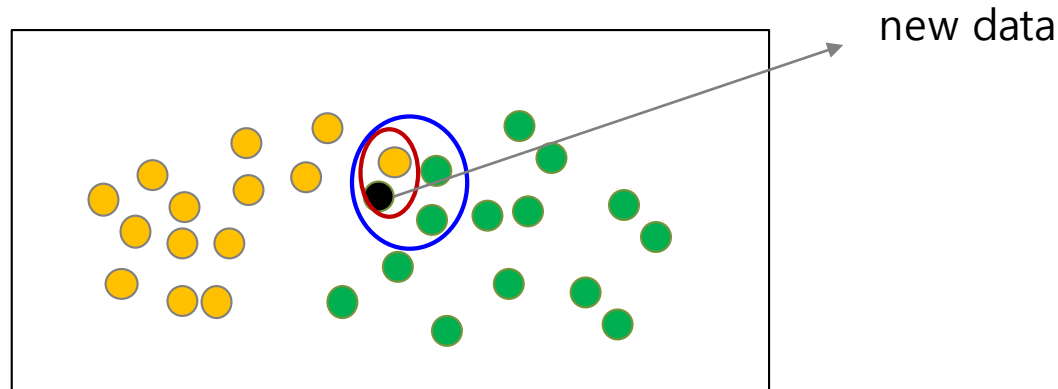
서포트 벡터 머신(SVM, Support Vector Machine)은 기계 학습의 분야 중 하나로 패턴 인식, 자료 분석을 위한 지도 학습 모델이며, 주로 분류와 회귀 분석을 위해 사용



개론7.3 기계학습

■ K-최근접이웃(k-NN, K-Nearest Neighbor) p.175

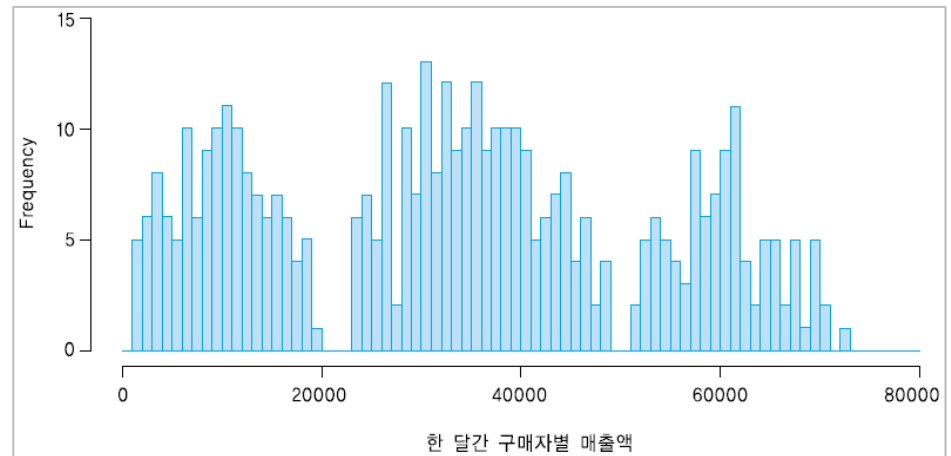
- k-NN은 새로운 데이터가 주어졌을 때 기존 데이터 가운데 가장 가까운 k 개 이웃의 정보로 새로운 데이터를 예측하는 방법론.
- 아래 그림처럼 검은색 점의 범주 정보는 주변 이웃들을 가지고 추론해낼 수 있음.
- 만약 k가 1이라면 오렌지색, k가 3이라면 녹색으로 분류(classification)하는 것.
- 만약 회귀(regression) 문제라면 이웃들 종속변수(y)의 평균이 예측값이 됨.



개론7.3 기계학습

■ 클러스터링(clustering) p.175

- 성격이 비슷한 항목들을 그룹으로 묶는 것
- 그룹화(명칭, 개수) 기준이 미리 정해져 있지 않으며, 클러스터링 결과를 “분류”에 활용할 수 있음
- k-means 알고리즘을 많이 사용함
- 문구점의 구매 매출액
 - 2만원 이하
 - 2~5만원
 - 5만원 이상
- 구매 액수에 따라 선물을 준비한다면, 선물의 가짓수 결정을 위해 먼저 고객 유형의 종류 파악이 선행되어야 함



개론7.3 기계학습

■ 연관 분석(association analysis) p.175

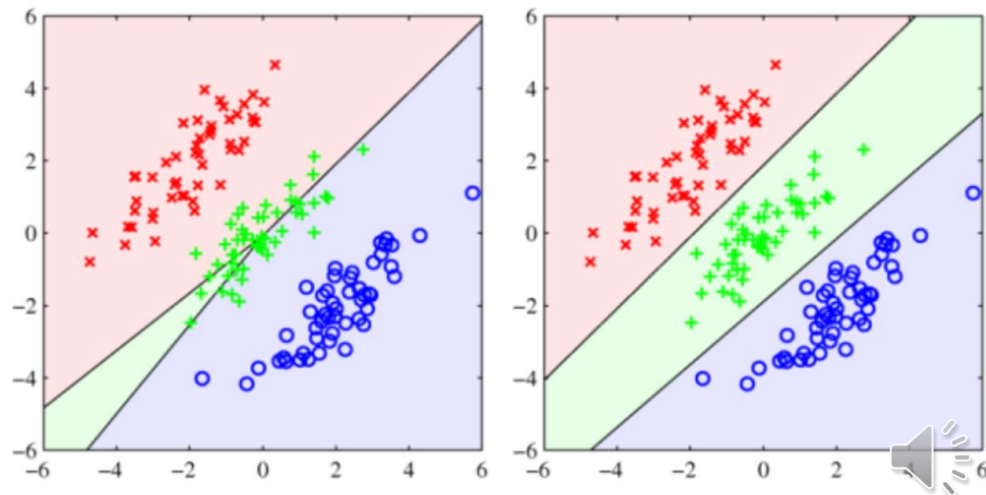
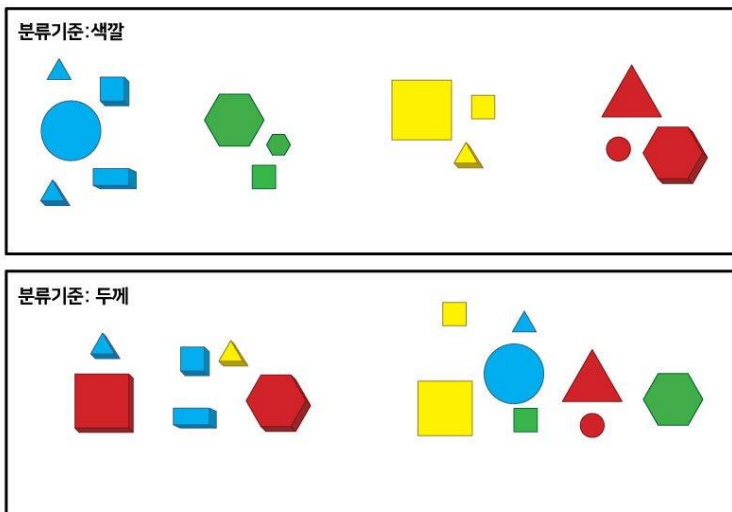
- 어떤 사건이 다른 사건과 얼마나 자주 동시에 발생하는지 파악하는 것
- 자주 발생하는 패턴 찾기(상품의 연관성, 취향의 연관성 등 분석)
- 같이 구매한 상품 분석(market basket analysis, 장바구니 분석)
- 상품의 진열 배치 및 상품 프로모션(쿠폰 발행 등)에 활용 됨



개론7.3 기계학습

■ 분류(classification) p.177

- 특정 항목이 어느 그룹에 속하는지 구분하는 예측 알고리즘 작업
- 분류할 그룹의 개수와 명칭이 미리 정해져 있음
- k-NN(k-nearest neighbors) 알고리즘, 베이시안 알고리즘(bayesian algorithm), 의사결정 트리 (decision tree)를 많이 사용함
- 예시
 - 고객 유형 분류: 구매 / 쇼핑
 - 스팸 메일 여부
 - 추천 서비스

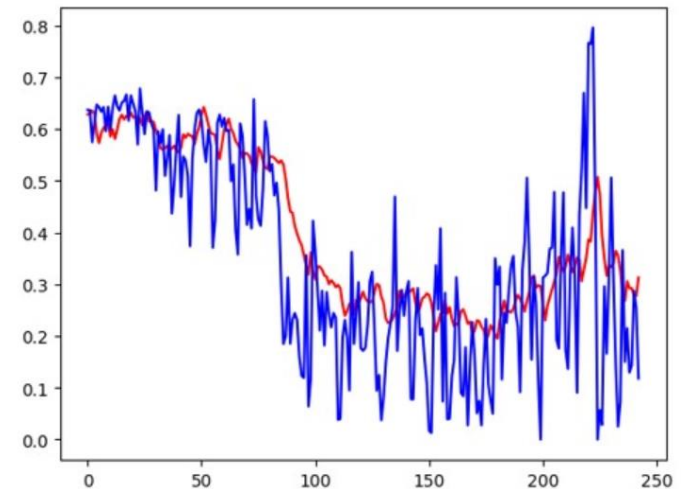


개론7.3 기계학습

■ 예측(prediction)

- 과거의 데이터를 보고 미래에 어떤 값이 나타날지 예상하는 것
- 과거 월별 구매 기록으로 다음 달 구매 예측
- 특정 광고나 프로모션을 했을 때 매출 증가 예측
- 수치 예측에 유용한 선형 회귀분석이 가장 많이 사용됨
 - ✓ 독립 변수의 기여도, 중요도 파악이 용이
- **신경망(neural network)** 모델도 널리 사용됨
 - ✓ 신경망 규모 증대로 예측도가 많이 향상 됨
 - ✓ 변수간의 영향도 파악이 어려운 것이 단점

스마트팩토리 (기계 장비 교체시기 예측 등)



한국의 SMP를 예측한 모델입니다. 빨간선은 예측값, 파란선은 실제값입니다.



개론7.4 분석 모델 선택 (p.179)

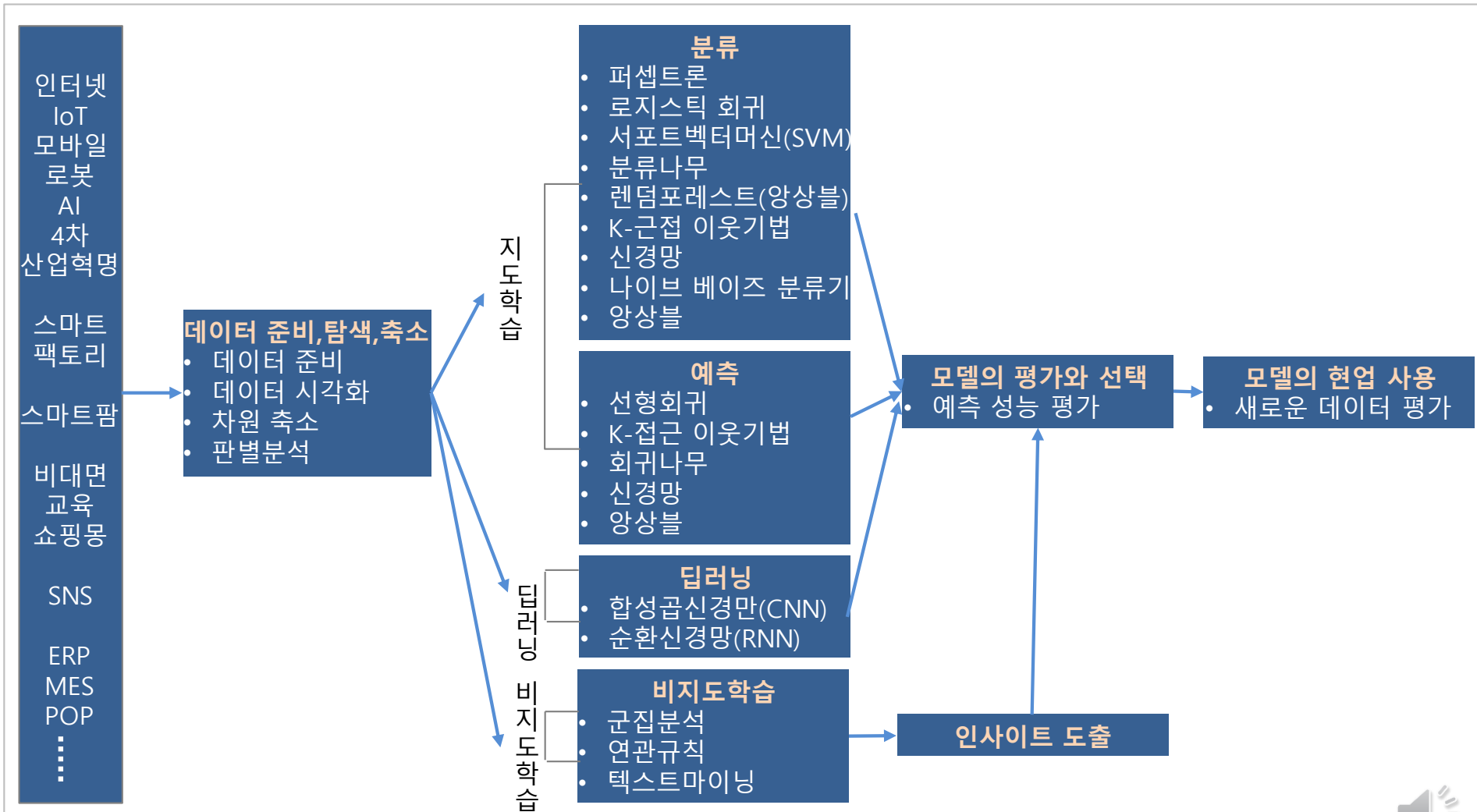
- 데이터 분석에서 가장 어려운 것은 주어진 문제를 해결하기 위해서 어떤 모델을 사용할 것인지 정하는 것이다.

분석가들은 프로젝트를 진행하면서 여러가지 분석 기법을 선정하고, 선택하며 최고의 퍼포먼스를 이끌어내기위해 많은 고민을 하게 된다. 데이터의 형태, 종속변수의 포맷 등 여러가지 사항을 고려하여 모델을 선택하는 것 또한 분석가의 역량이다.

- 분류를 통해 해결할 수 있는 문제인지
- 클러스터링을 먼저 하여 그룹을 만들어 봐야 할 것인지
- 예측을 해야 하는 문제인지를 파악해야 한다.



개론7.4 분석 모델 선택



개론7.4 분석 모델 선택

■ 모델 선택 고려사항

- 주어진 문제 해결에 적합한 모델 선택이 중요
- 선택한 모델의 성능 평가 기준의 사전 정의 필요
 - ✓ 상황에 따라 좋은 분석 알고리즘의 조건이 달라짐
- 모델의 성능 평가를 위해 수치화된 객관적인 기준 필요
- 분석 모델 성능의 최저 기준은 널(null) 모델
- 널 모델이란 어떤 모델도 적용하지 않고 측정된 평균치 등을 그대로 적용
 - ✓ 스팸 메일 분류 시 모든 메일이 무조건 스팸이 아니라고 분류하는 것
 - ✓ 스팸 메일이 1%라고 할 때, 99%는 맞고, 1%는 틀린 예측을 하게 됨
 - ✓ 최저 성능 기준이 됨



개론7.4 분석 모델 선택

■ 알고리즘의 동작 속도

- 모델의 분류 or 예측 정확도 높이는 것 중요
- 모델을 구현한 알고리즘의 처리 속도 또한 중요
- 알고리즘의 속도(시간)의 종류:
 - ✓ 훈련 데이터를 사용해서 모델을 만드는 데 걸리는 시간
 - ✓ 모델의 데이터 분류 또는 예측 시간
- 알고리즘의 속도 요구사항에 따라 선택 알고리즘이 달라짐
- 알고리즘이 정교하고 복잡할수록 정확도는 높아지지만, 실행 시간이 오래 걸리는 것이 일반적임



Thank you

