동국대학교

dongguk UNIVERSITY

2020년 - 데이터 청년 캠퍼스

데이터사이언스 기반 지능소프트웨어 과정

데이터 사이언스 개론

1) 데이터 사이언스를 시작하기 전 고려해야 할 사항

강사 소개

석민수

- 동국대학교 컴퓨터공학과 박사과정
- 동국대학교 융합소프트웨어 교육원 소속 강사

- aksen123dgu@gmail.com

- 010 – 6854 – 7573

본 강의 소개

- 빅데이터 시대에서 데이터 분석의 필요성을 소개
- 데이터 분석과 활용을 위한 데이터 사이언스의 정의
- 다양한 데이터 사이언스 기법과 데이터 특성별 적절한 기법 소개

데이터 과학의 필요성

데이터 과학의 위치

DATA SCIENCE

비정형 혹은 정형 데이터를
정제, 준비, 분석하는 활동

BIG DATA

엄청난 양의 데이터를 말하며
다양한 형태의 정보 자산

DATA ANALYTICS

알고리즘과 수학적 처리 과정을
적용하여 데이터를 다루는 활동

데이터 사이언스의 절차

■ 데이터 사이언스의 4단계

1) 데이터를 처리(가공)하기 위한 준비

2) 데이터를 모델링 (Modeling) 하기 위한 적절한 알고리즘을 선정

3) 모델을 최적화하기 위한 알고리즘 튜닝 (Turning)

4) 모델의 정확도 평가

1) 데이터를 처리(가공)하기 위한 준비

- 모여있는 데이터를 '데이터 셋' (Dataset) 이라고도 함
- 데이터 셋을 표현하는 가장 흔한 방법은 "표" 형태

변 수

데이터 포인트

거래 ID	고객 종	날짜	구매한 과일 갯수	생선 구매여부	구매액 (달러)
1	펭귄	1월 1일	1	예	5.30
2	곰	1월 1일	4	예	9.70
3	토끼	1월 1일	6	아니요	6.50
4	말	1월 2일	6	아니요	5.50
5	펭귄	1월 2일	2	예	6.00
6	기린	1월 3일	6	아니요	4.80
7	토끼	1월 3일	8	아니요	7.80
8	고양이	1월 3일	?	예	7.80

1) 데이터를 처리(가공)하기 위한 준비

■ 표의 행 (Row)

- 데이터 포인트 (Data Point)
- 한 번의 관측으로 얻어진 데이터들

■ 표의 열 (Column)

- 표의 행을 설명하는 변수 (Variable)
- 변수 : 속성 (Attribute), 특징 (Feature), 차원 (Dimension) 등

■ 변수 타입 (Variable Type)

- 이진 (Binary) : 가장 기본적인 방법, 0 / 1 혹은 예 / 아니오 등
- 범주 (Categorical) : 변수 내용이 제한적이고 두 가지 이상일 때
- 정수 (Integer) : 소수점이 없는 숫자만을 나타낼 때
- 연속 (Continuous) : 소수점을 포함한 숫자를 나타낼 때

1) 데이터를 처리(가공)하기 위한 준비

■ 변수 선택

- 데이터 셋의 변수가 너무 많을 경우,
  - 처리시간이 오래 걸림
  - 과도한 노이즈로 정상적인 예측이 불가능
- 이를 방지하기 위해 필요 없는 변수를 생략 (<- **상관관계 분석**)

■ 특징 엔지니어링 (Feature Engineering)

- 임의 변수를 생성하여, 데이터 셋 처리 속도를 증가시킬 수 있음
- 토끼, 말, 기린 -> 초식동물 변수로 합침 (<- **차원 축소**)

■ 누락 데이터

- 근사 / 계산 / 제거를 통해 누락 데이터를 임의로 생성하여 채우거나 해당 행을 제거할 수 있음

2) 적절한 알고리즘을 선정

■ 비지도 학습 (Unsupervised Learning)

- 데이터에 어떤 패턴이 존재하는지 탐색
- K-평균 군집화 (k-means Clustering)
- 주성분 분석 (Principal Component Analysis) 등

■ 지도 학습 (Supervised Learning)

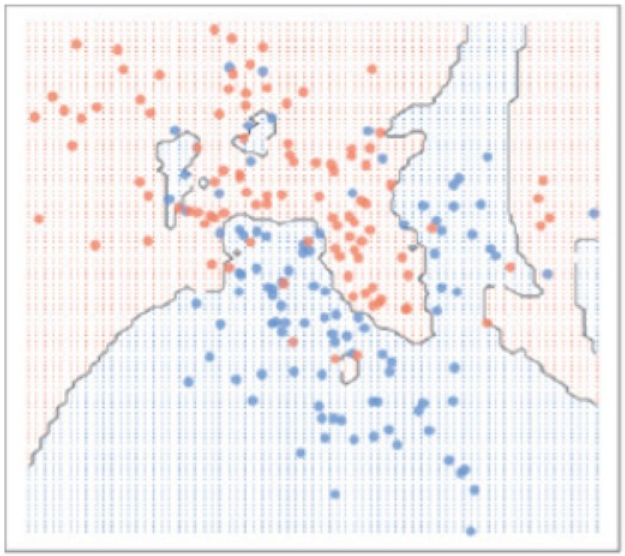
- 데이터에 존재하는 패턴을 바탕으로 예측
- 회귀 분석 (Regression Analysis)
- K-최근접 이웃 (k-Nearest Neighbors) 등

■ 강화 학습 (Reinforcement Learning)

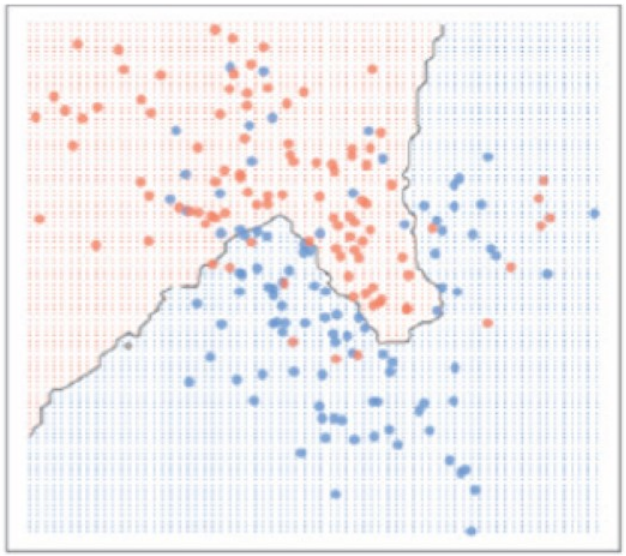
- 데이터에 존재하는 패턴을 바탕으로 예측하고, 추가 데이터를 통해 정확도까지 개선
- 멀티-암드 밴딧 (Multi-Armed Bandits)

3) 모델을 최적화하기 위한 알고리즘 튜닝

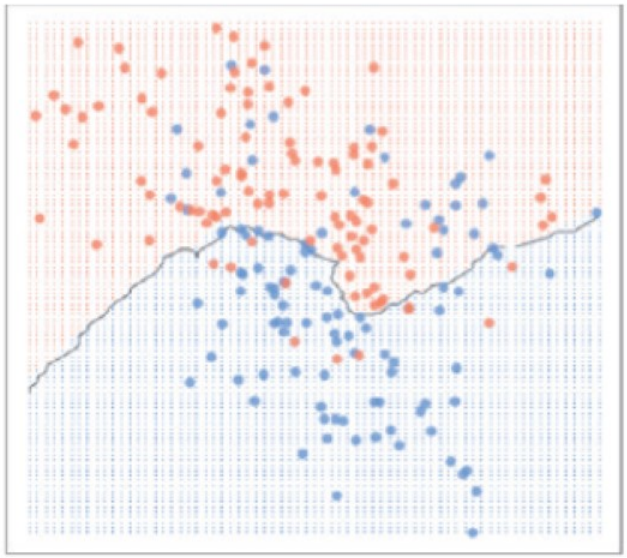
- 적당하지 않은 파라미터 (Parameter) 는 오히려 정확도가 떨어짐 (*과유불급 : 過猶不及*)



a) 과적합(Overfit)



b) 최적합(Ideal fit)



c) 부적합(Underfit)

4) 모델의 정확도 평가

■ 가장 좋은 결과란?

		예측: 살 것이다	예측: 사지 않을 것이다
실제	샀다	1 (TP)	5 (FN)
제	사지 않았다	5 (FP)	89 (TN)

- TP / FP : True Positive / False Positive
- TN / FN : True Negative / False Negative

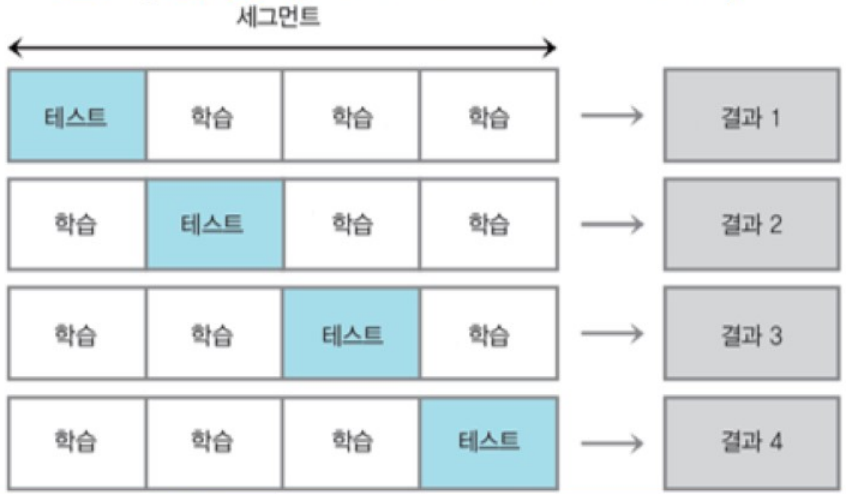
4) 모델의 정확도 평가

■ 회귀 지표

- 실제값과 예측값의 차이를 계산하여, 정도에 따라 불이익을 줌
- 평균 제곱근 오차 (RMSE : Root Mean Squared Error) 등

■ 검증

- 수집 데이터는 정확하지만, 실제 데이터는 정확하지 않을 수 있음
- 학습 / 테스트 데이터 셋 (Training / Test Dataset)
- 교차검증 (Cross-validation)




Q & A