천재교육

머신러닝 (Machine Learning) 이해 및 실습

K - 디지털 아카데미

강사 소개

박나연

천재교육 AI센터 개발운영팀

- 밀크T 등 서비스를 위한 AI모델적용 PoC 진행

머신러닝 교육과정 강의 일정

1일차 > 머신러닝

- > 머신러닝의 개요
- > [실습] 데이터 분석의 시작 Numpy, Pandas 활용하기

2일차 > Classification (분류)

- > 분류를 평가하는 지표 알아보기
- > 분류 알고리즘 (결정트리, 앙상블, 랜덤포레스트 등) 익히기
- > [실습] 분류를 통한 밀크T 만료및탈퇴회원 예측(이탈 회원 예측)

3일차 > Regression (회귀)

- > 회귀와 경사 하강법
- > 로지스틱 회귀와 소프트맥스 회귀
- > [실습] 로지스틱 회귀를 통한 문항별 정오답 예측

4일차 > 차원 축소와 Clustering(군집화)

- > PCA, LDA
- > K-means, DBSCAN 등 다양한 클러스터링 기법 알아보기
- > [실습] 밀크T중학 회원수준 군집화(GMM)

5일차 > 추천시스템과 최종 프로젝트

- > 추천시스템
- > 최종 프로젝트

머신러닝의 개요

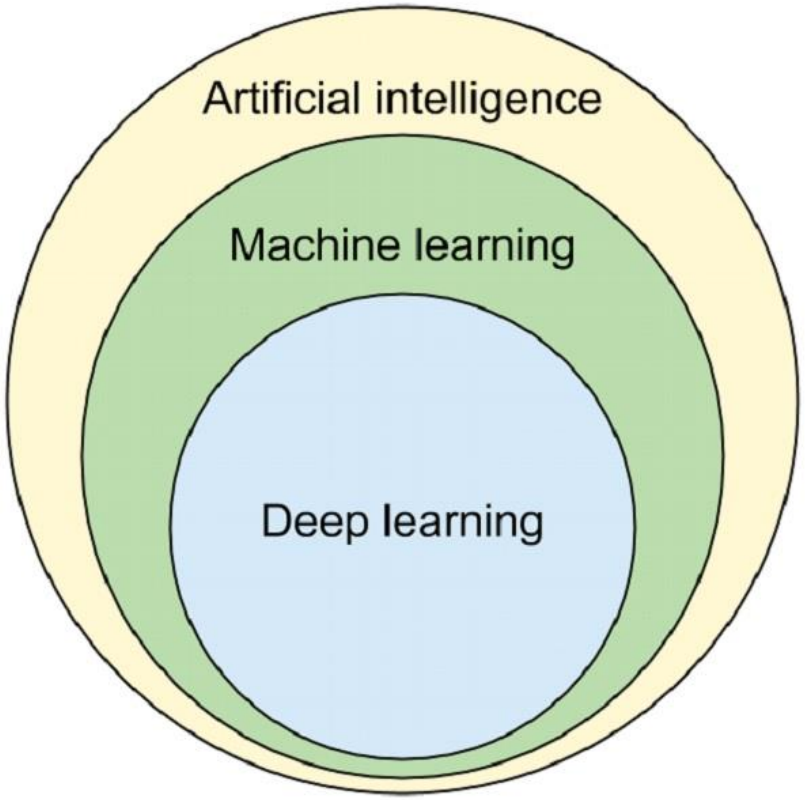
01 머신러닝 이란?

02 머신러닝의 적용

03 머신러닝의 주의사항

머신
러닝.

머신러닝(Machine Learning)의 개념



머신러닝(Machine Learning)의 개념

아서 사무엘(Arthur samuel) (1959)

컴퓨터 프로그램을 명시적으로 구현하는 대신 컴퓨터 스스로 학습하는 능력을 갖도록 하는 연구 분야

톰 미첼(Tom Mitchell) (1977)

과제 T에 대한 프로그램의 성능 P가 경험 E를 통해 향상되면 해당 “프로그램이 경험 E를 통해 학습한다” 라고 말한다.

위키 피디아

기계학습 또는 머신러닝은 경험을 통해 자동으로 개선하는 컴퓨터 알고리즘의 연구로, 컴퓨터가 학습할수 있도록 하는 알고리즘과 기술을 개발하는 분야.

머신러닝(Machine Learning)의 개념

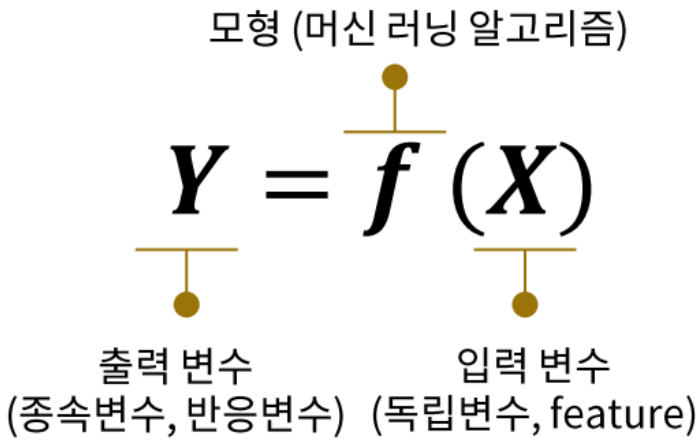
무엇(X)으로 무엇(Y)을 예측하고 싶다.

데이터(행렬)

X Y

성별	키	몸무게	체지방	BMI지수	폐활량	흡연여부
남	182	78	18	15	86	Y
여	156	52	25	17	95	N
여	165	58	21	19	98	N
...	...	...	...	...		...

머신러닝(Machine Learning)의 개념



주어진 데이터를 통해서 입력과 출력 간의 관계를 만드는 함수 f를 만드는 것

주어진 데이터 속에서 데이터의 특징을 찾아내는 함수 f를 만드는 것

머신러닝의 예시

X : 고객들의 개인정보 및 금융 관련 정보 → Y : 대출 연체 여부

X : 고객의 상품 구매 내역 → Y : 고객의 취향

X : 학생의 인강 수강기록, 연습문제 풀이이력 → Y : 중간고사 시험 점수

… 또 어떤 것들이 있을까요?

머신러닝의 종류

데이터의 종류에 따라, 적용할 수 있는 머신러닝의 종류는 다양하다!

분류

Yes OR No

회귀

연속적인 숫자 .. (키, 점수 등)

군집화 비슷한 X끼리 묶어보자

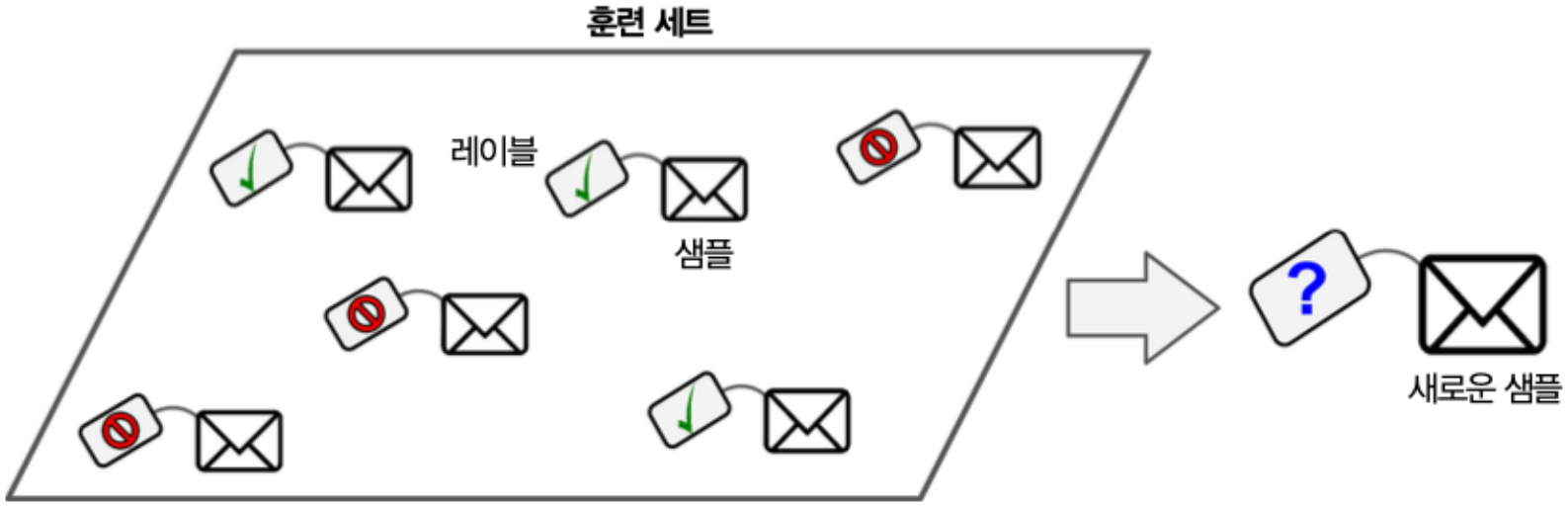
Y가 없다면?

학습 데이터? 테스트 데이터?

학습 데이터? 테스트 데이터?

지도 학습 vs 비지도 학습

지도학습은 학습 데이터에 레이블이라는 답을 표기하여 레이블을 맞추도록 유도하는 학습을 가리킨다.



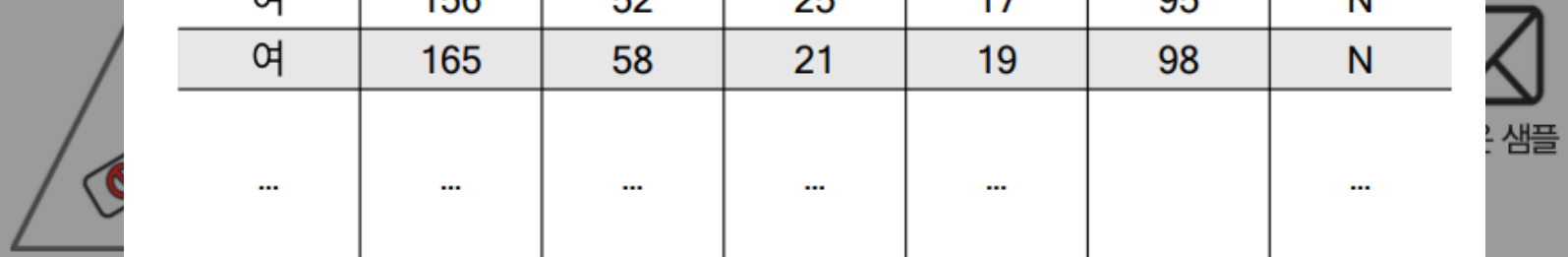
정상메일 vs 스팸메일

지도 학습 vs 비지도 학습

지도학습은 훈련 데이터에 레이블이라는 답을 표기하여 레이블을 맞추도록 유도하는 학습을 가리킨다.

데이터(행렬)

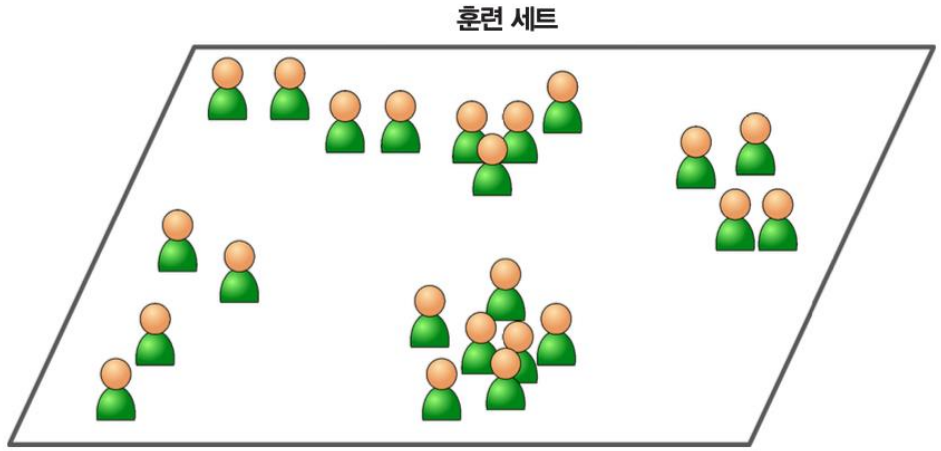
	X					Y
성별	키	몸무게	체지방	BMI지수	폐활량	흡연여부
남	182	78	18	15	86	Y
여	156	52	25	17	95	N
여	165	58	21	19	98	N
...	...	...	...	...		...

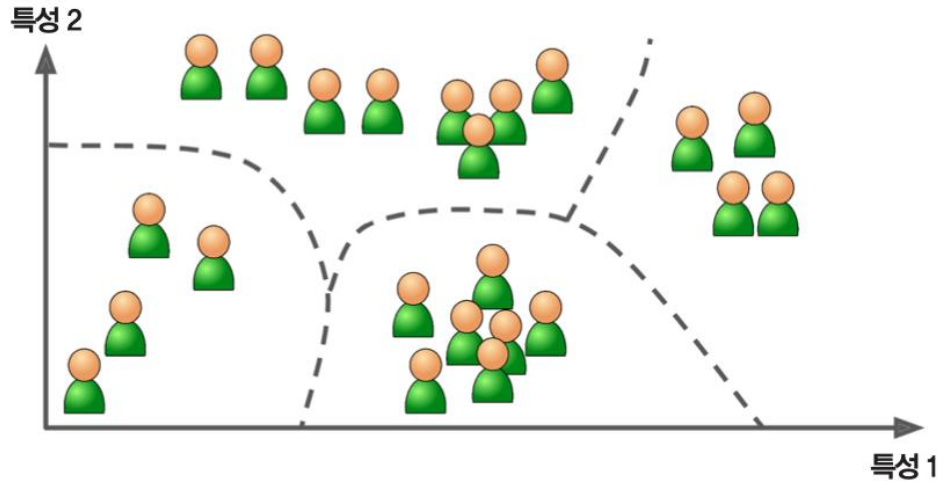


정상메일 vs 스팸메일

지도 학습 vs 비지도 학습

비지도 학습은 학습에 레이블이 없는 학습데이터를 이용하여 데이터의 상관관계와 패턴 등을 기계가 찾아내야 하는 학습이다.





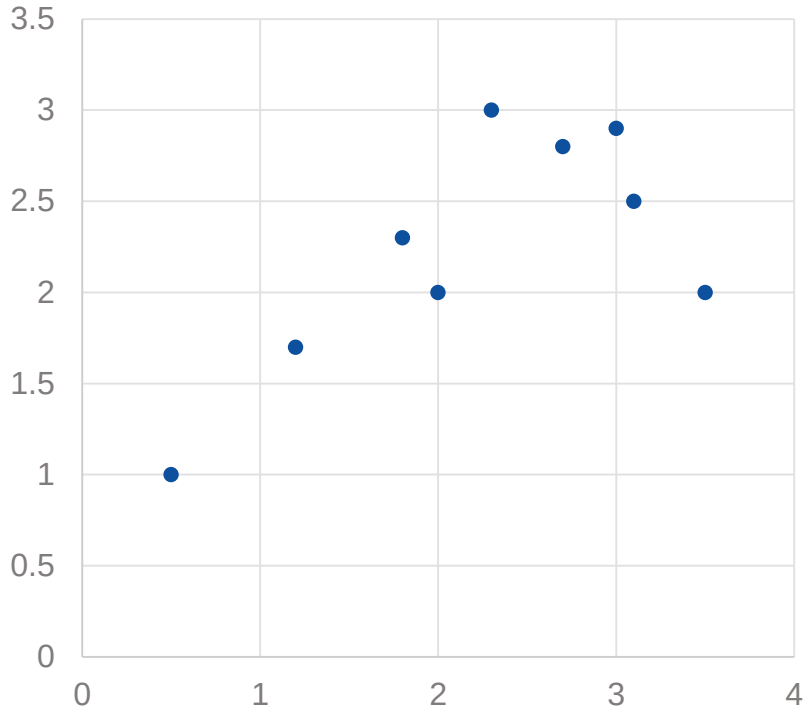
비지도 학습의 활용?

- 유사한 구매패턴의 고객 그룹화
- 관심사가 유사한 사용자 분류

머신러닝의 주요 도전 과제

과대적합(Overfitting)

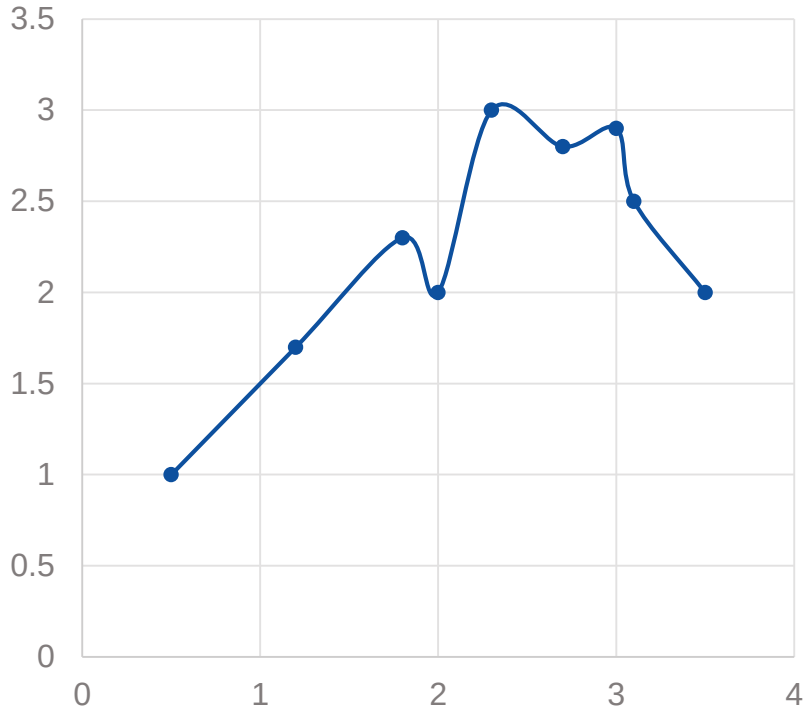
모델이 학습과정에서 학습 데이터에 특화되어 일반화 성능이 떨어지는 현상



머신러닝의 주요 도전 과제

과대적합(Overfitting)

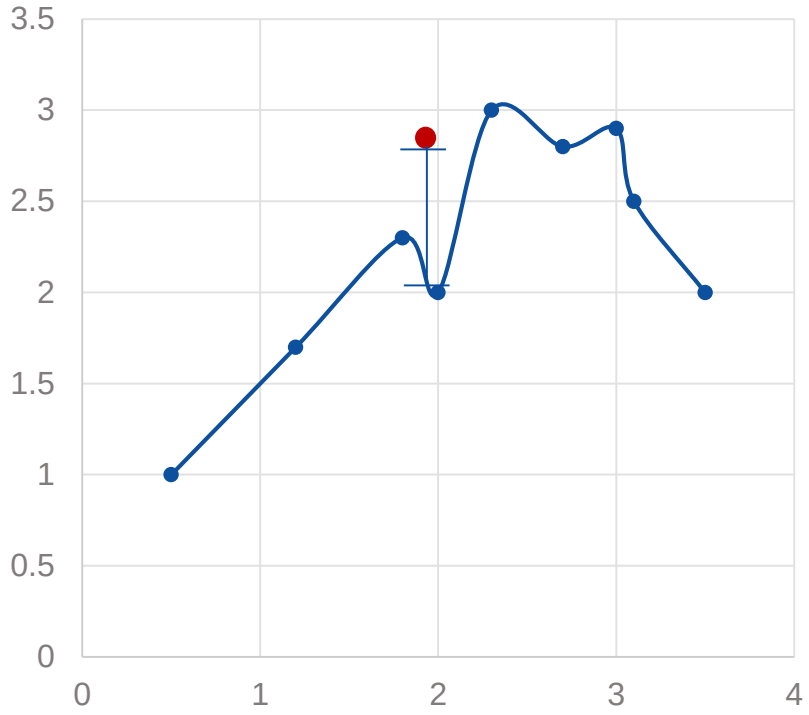
모델이 학습과정에서 학습 데이터에 특화되어 일반화 성능이 떨어지는 현상



머신러닝의 주요 도전 과제

과대적합(Overfitting)

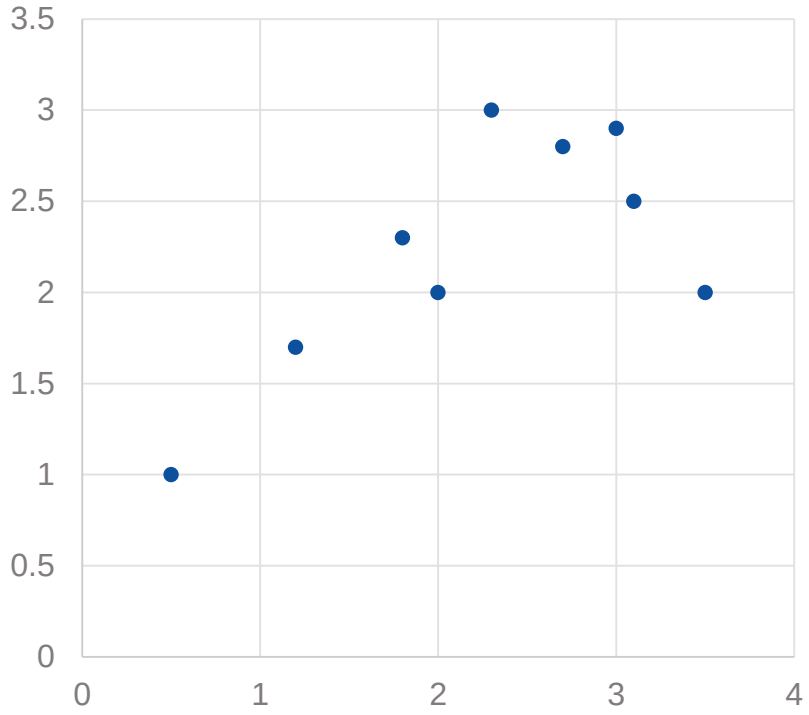
모델이 학습과정에서 학습 데이터에 특화되어 일반화 성능이 떨어지는 현상



머신러닝의 주요 도전 과제

과소적합(Underfitting)

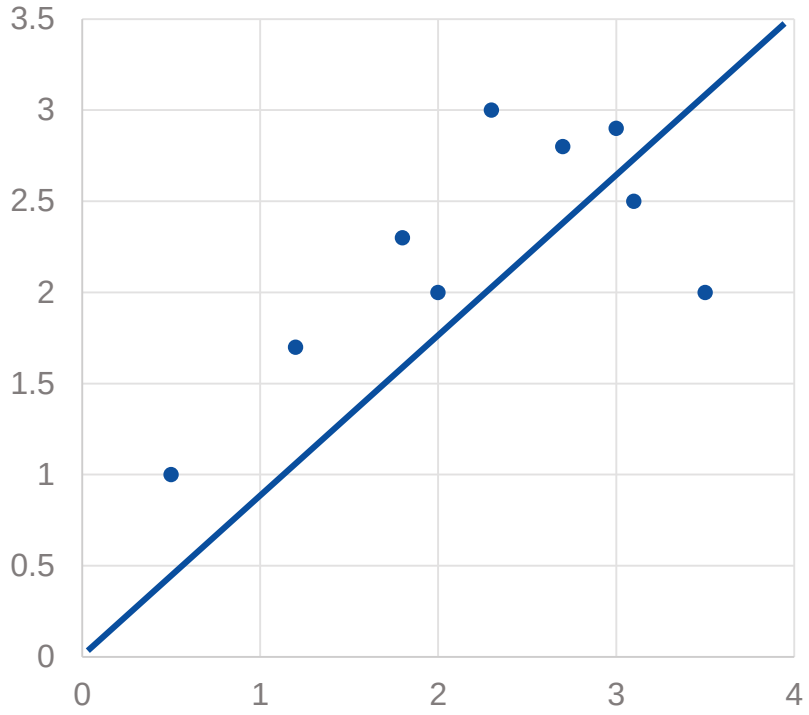
모델이 너무 단순해서 학습 데이터를 제대로 대변하지 못하는 경우



머신러닝의 주요 도전 과제

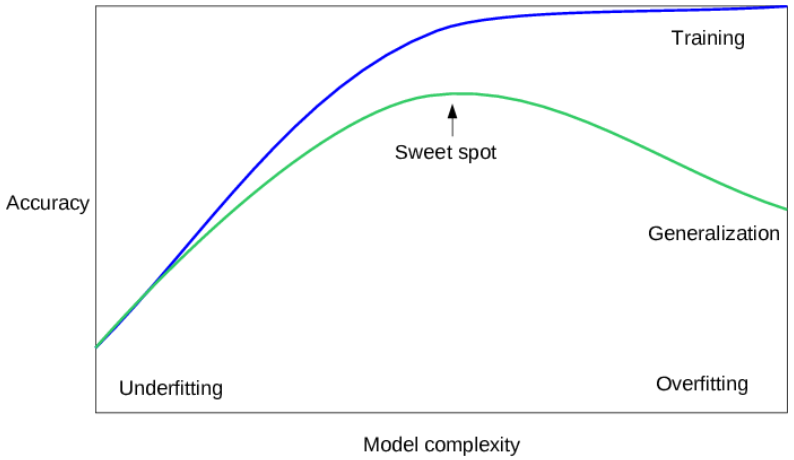
과소적합(Underfitting)

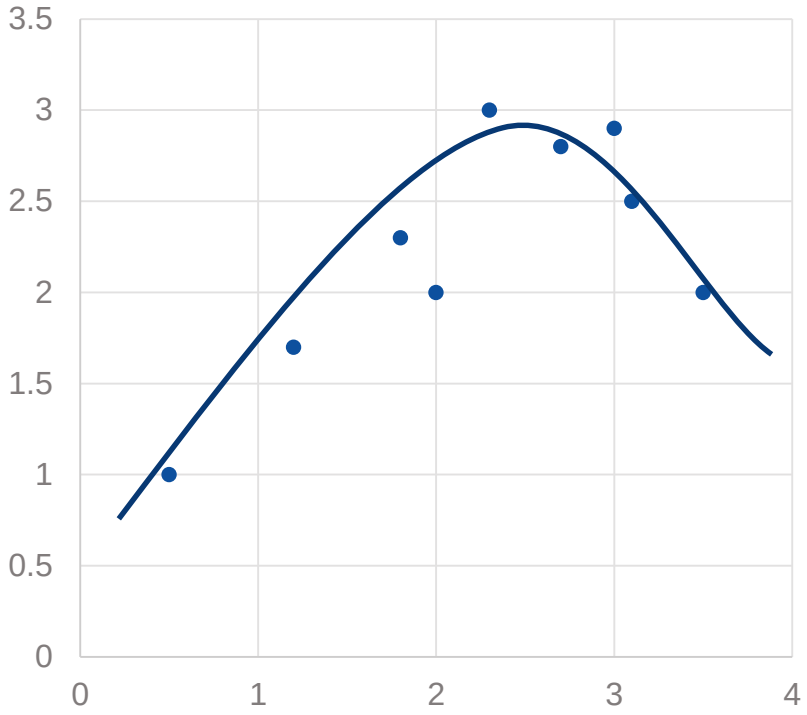
모델이 너무 단순해서 학습 데이터를 제대로 대변하지 못하는 경우



머신러닝의 주요 도전 과제

Overfitting 과 Underfitting을 방지하기 위해 일반화 필수

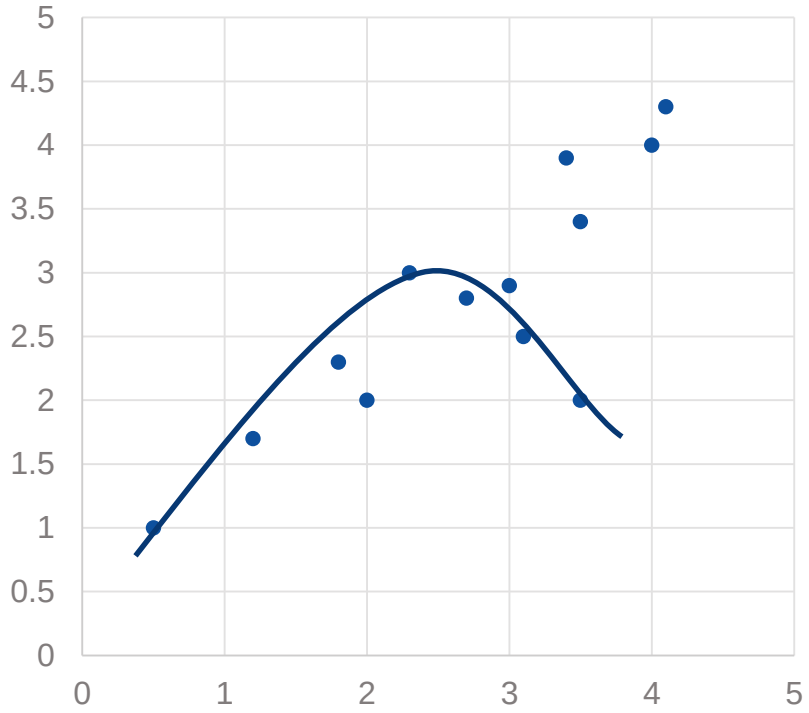




머신러닝의 주요 도전 과제

아웃라이어

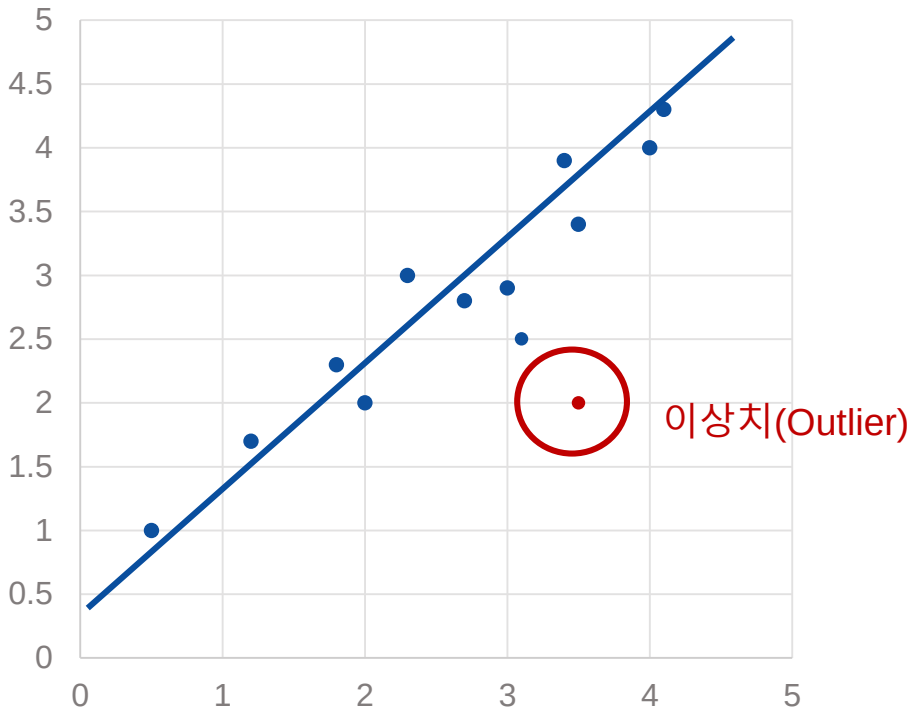
보통의 관측 데이터 범위에서 많이 벗어난 아주 작은 값이나 큰값을 말한다.



머신러닝의 주요 도전 과제

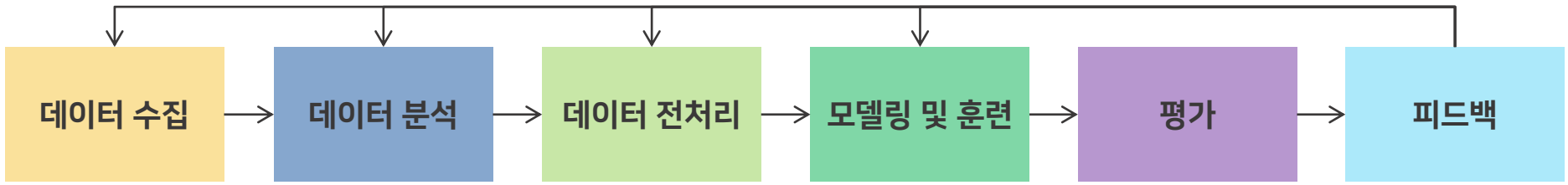
아웃라이어

보통의 관측 데이터 범위에서 많이 벗어난 아주 작은 값이나 큰값을 말한다.



이상치(Outlier)의 경우에 해당 데이터를 수정하거나 무시
→ 어떻게 제거할까 ? (추후에…)

머신러닝의 과정



데이터 전처리

01 라이브러리

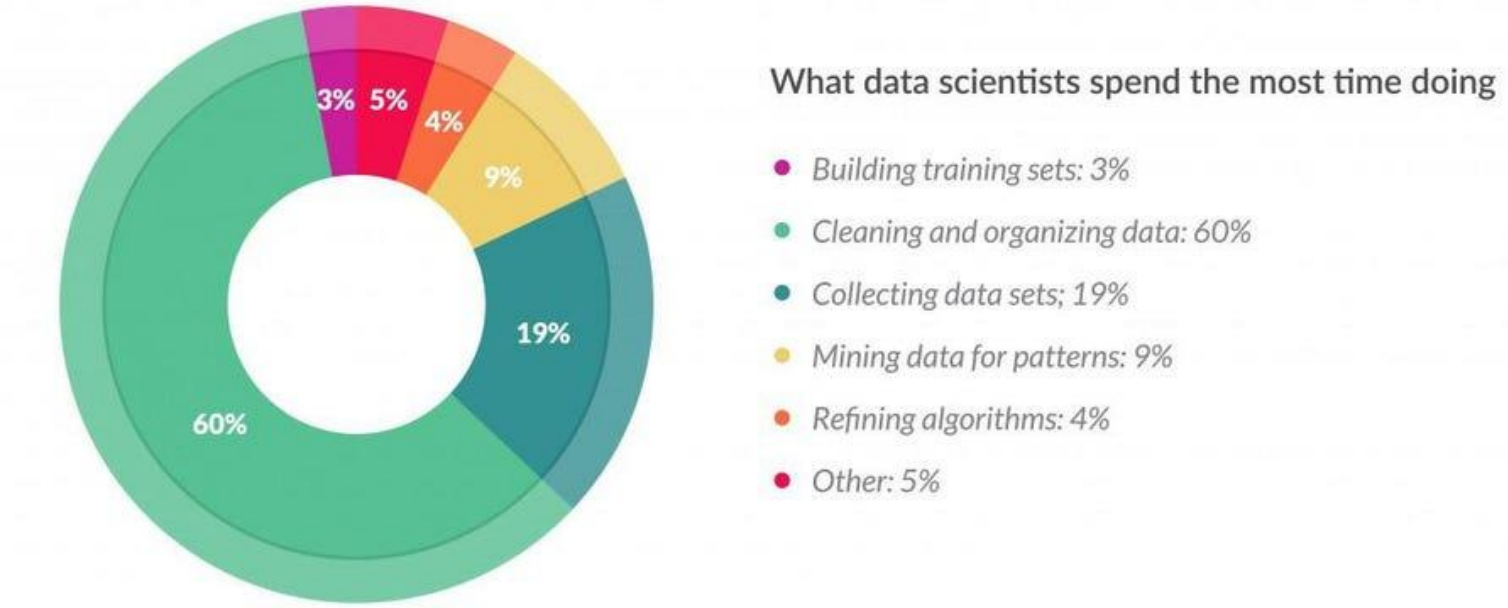
02 인코딩(Encoding)

03 스케일링(Scaling)

머신
러닝.

데이터 핸들링을 위한 라이브러리

데이터 전처리



https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=123b5f906f63

데이터 전처리

데이터를 분석 및 처리에 적합한 형태로 만드는 과정을 총칭

- 인코딩
- 스케일링

인코딩

인코딩이란? 사람이 인지할 수 있는 문자(언어)를 규칙에 다라 컴퓨터가 이해할 수 있는 언어로 바꾸는 것을 말한다.

라벨 인코딩

과목
국어
수학
영어
사회
과학

과목
0
1
2
3
4

원핫 인코딩

과목
국어
수학
영어
사회
과학

과목				
1	0	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	0	1

스케일링

넌 90점이고 난 800점이니까
내가 더 공부를 잘해!

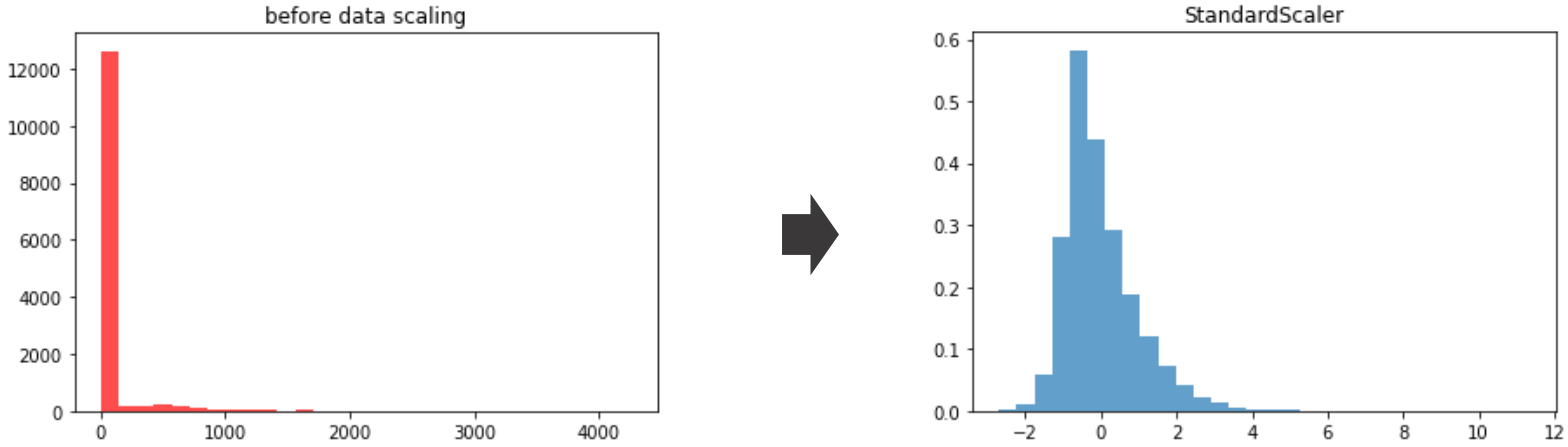
A

??

B

스케일링

StandardScaling 데이터가 평균을 0 분산이 1인 정규분포를 나타내도록 스케일링



Minmax scaling 데이터를 0과 1사이의 값으로 축소하는 것 (최소값이 0, 최댓값이 1이 되도록 함)

```
[ 0  1  2  3  4  5  6  7  8  9 10]
```

```
[0.  0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1. ]
```

인코딩과 스케일링

[실습]

실습 목록

넘파이, 판다스 익히기

0. ndarray
1. 인덱싱
2. 판다스 인덱싱
3. unique, groupby, mean 등
4. lambda, apply 활용

데이터 핸들링

0. 데이터 불러오기
1. 결측치 제거
2. 데이터 인코딩
3. 데이터 스케일링
4. 데이터 시각화
5. 학습, 테스트 데이터 split

행과 열

0번째 열 ↓ 1번째 열 ↓ 2번째 열 ↓ 3번째 열 ↓ 4번째 열 ↓ 5번째 열 ↓ 6번째 열 ↓ 7번째 열 ↓

	사용자ID	차시코드	단원코드	단원 제목	강의내용	강의명	학년	학기
0번째 행 → 0	62fbd7b7-6a32-4da2-b203-acdec95e00b2	T0ME41U43001	T0ME41U43	1학기 복습	만, 다섯 자리 수 알아보기	수학 1단원 【복습①】	4	1
1번째 행 → 1	62fbd7b7-6a32-4da2-b203-acdec95e00b2	T0ME41U43001	T0ME41U43	1학기 복습	만, 다섯 자리 수 알아보기	수학 1단원 【복습①】	4	1
2번째 행 → 2	62fbd7b7-6a32-4da2-b203-acdec95e00b2	T0ME41U43001	T0ME41U43	1학기 복습	만, 다섯 자리 수 알아보기	수학 1단원 【복습①】	4	1
3번째 행 → 3	62fbd7b7-6a32-4da2-b203-acdec95e00b2	T0ME41U43001	T0ME41U43	1학기 복습	만, 다섯 자리 수 알아보기	수학 1단원 【복습①】	4	1
4번째 행 → 4	62fbd7b7-6a32-4da2-b203-acdec95e00b2	T0ME41U43001	T0ME41U43	1학기 복습	만, 다섯 자리 수 알아보기	수학 1단원 【복습①】	4	1

밀크T 차시별 강의 확인문제 정오답 데이터

	사용자ID	차시코드	단원코드	단원제목	강의내용	강의명	학년	학기	시험구분	강의구분	강의타입	동영상재생시간	실제재생시간	확인문제점수	학습일	문항번호	문항코드	사용자입력	정오답	영역	대단원코드	대단원제목	중단원코드	중단원제목	소단원코드	소단원제목	토픽코드	토픽제목	난이도	평가영역
0	62fbd7b7-6a32-4da2-b203-acdec95e00b2	T0ME41U43001	T0ME41U43	1학기복습	만, 다섯자리수알아보기	수학1단원【복습①】	4	1	NaN	E	AAA	0	8	-1	2022-07-04 21:00:03	1.0	30016642.0	10000	O	MA	17120995.0	1. 큰수	14201237.0	다섯자리수	12233514.0	다섯자리수의이해	12234054.0	모형세보기	2.0	91.0
1	62fbd7b7-6a32-4da2-b203-acdec95e00b2	T0ME41U43001	T0ME41U43	1학기복습	만, 다섯자리수알아보기	수학1단원【복습①】	4	1	NaN	E	AAA	0	8	-1	2022-07-04 21:00:03	2.0	30016654.0	1000/^ROW^/100/^ROW^/10/^ROW^/1	O	MA	17120995.0	1. 큰수	14201237.0	다섯자리수	12233514.0	다섯자리수의이해	12233718.0	10000알아보기	2.0	91.0

실제 밀크T 회원의 문항별 정오답, 차시 수강 정보 등이 담긴 데이터를 핸들링 해보자!

감사합니다