I 팀 1조

AI 공부하조입니다!

01

하나,

인공지능이란

인공지능이란

02

두울,

머신러닝이란

지도학습
모델 훈련 방법
비지도학습

03

세엣,

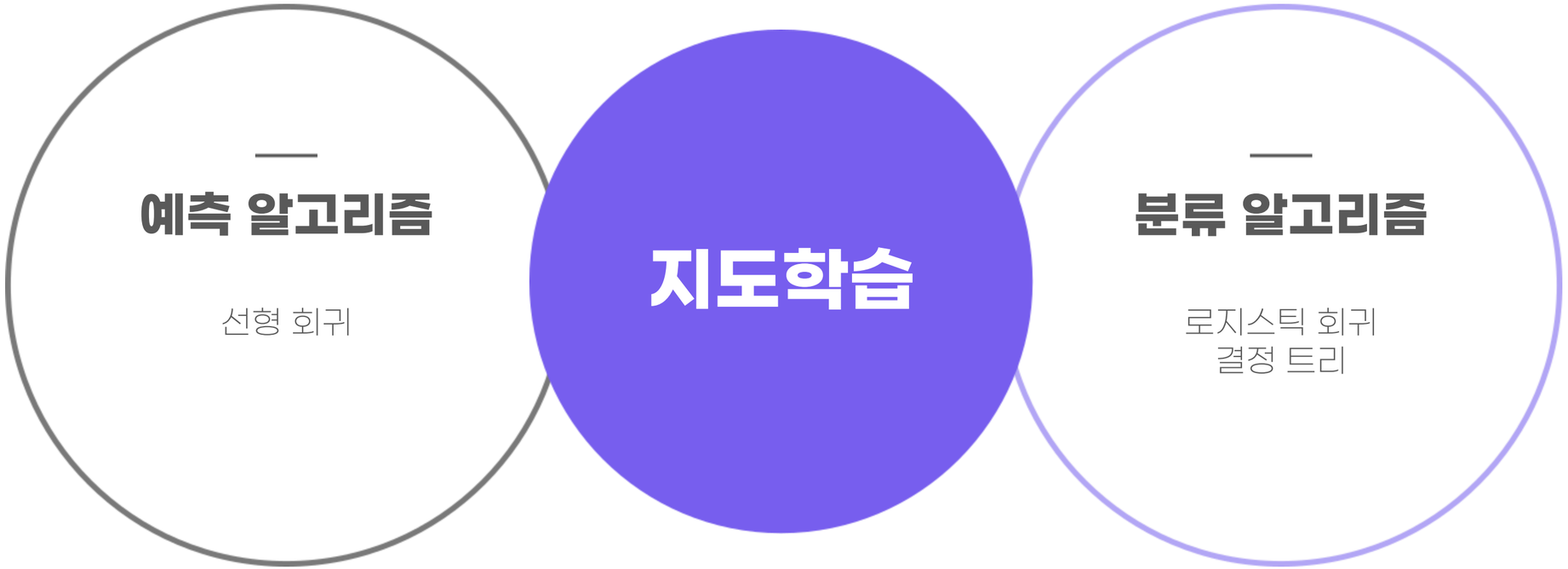
딥러닝이란

신경망
+ 공부 방법

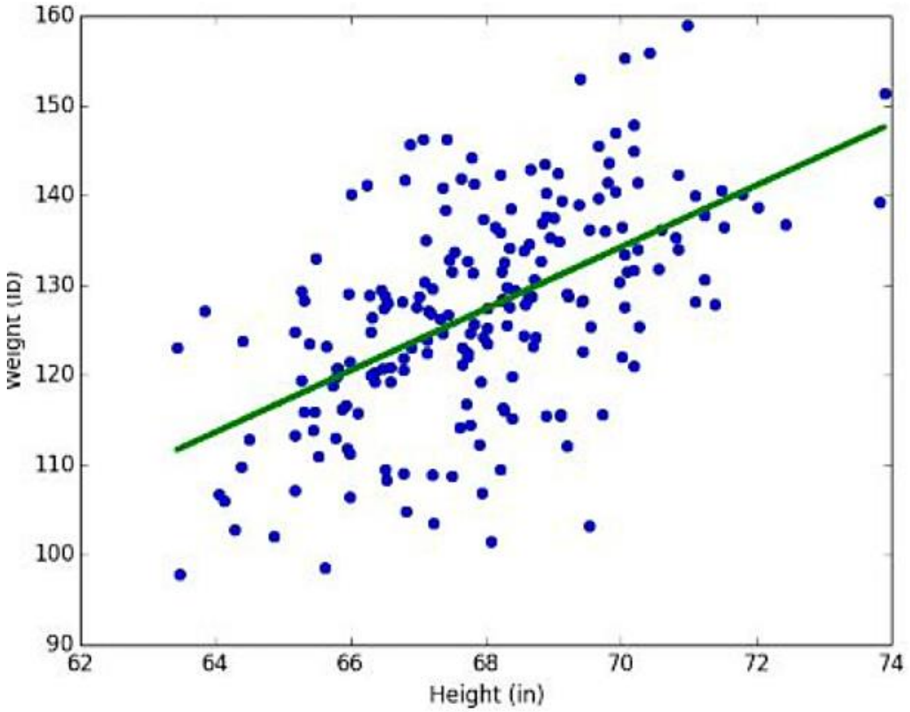
인공지능이란

인공지능 구현방법

지도학습



선형 회귀 데이터를 가장 잘 설명할 수 있는 선을 찾고 분석하는 방법



01 단순 선형 회귀

$$y = \omega x + b$$

ω = 가중치, b = 편향

02 다중 선형 회귀

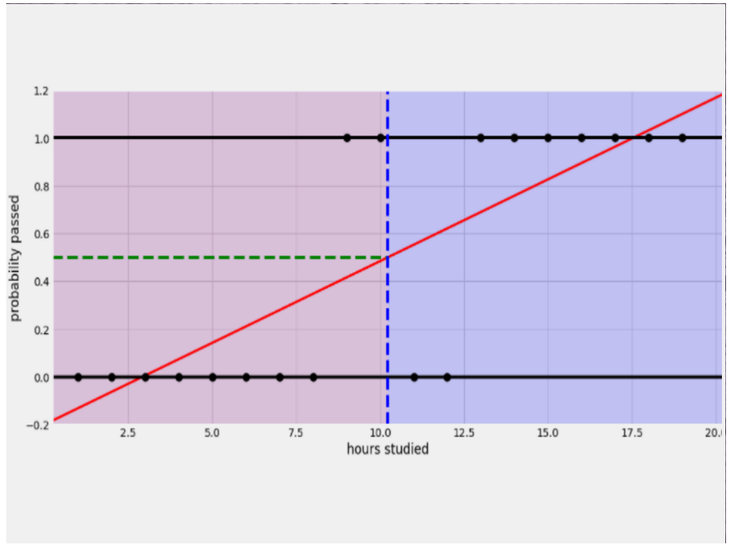
$$y = \omega_1 x_1 + \omega_2 x_2 + \cdots \omega_n x_n + b$$

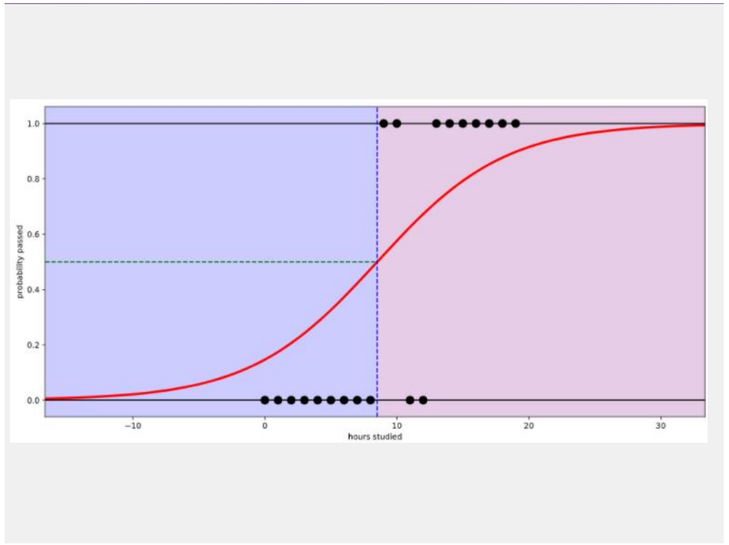
선형 회귀 예시

날씨에 따른
아이스크림 판매량 예측

01 **일별 기온, 강수량을 공변량으로 아이스크림 판매량을 종속변수로 사용하여 모델을 훈련**

02 아이스크림 판매량을 예측





로지스틱 회귀

이진분류 문제에
사용되는 분류모델

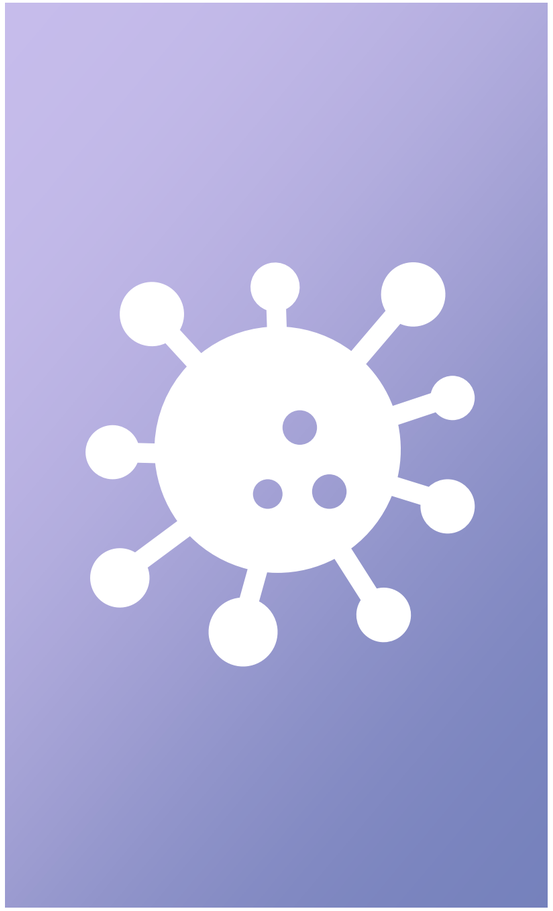
선형 회귀 ⟶ 시그모이드 함수 ⟶ 로지스틱 회귀

$$S(x) = \frac{1}{1 + e^{-x}}$$

x값을 0~1 사이의 값으로 변환

로지스틱 회귀 예시

암 발생 여부 구분

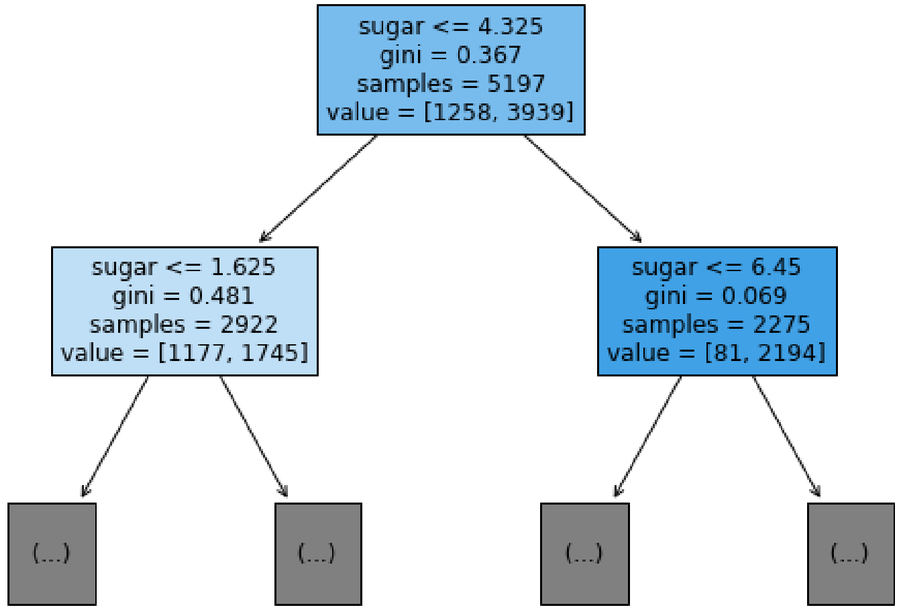


01 **종양의 크기를 독립변수로 사용하고 모델을 훈련**

02 암 발생 유무를 분류

09

결정 트리 Yes / No에 대한 질문을 이어가며 정답을 찾아 학습하는 알고리즘



01 장점

- 분류 과정을 이해하기 쉬움
- 전처리 필요 없음

02 단점

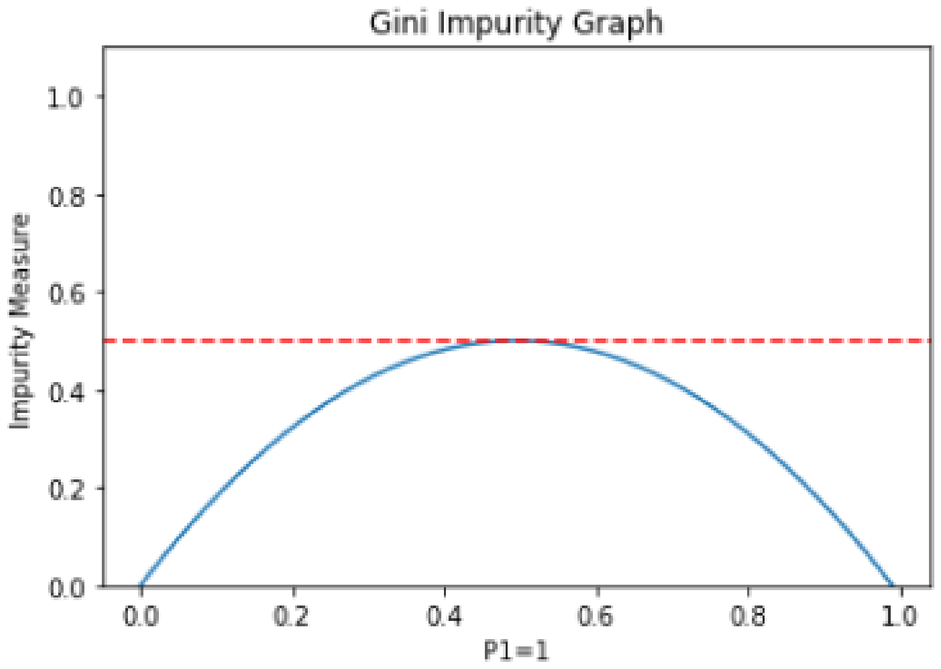
- 과적합이 발생하기 쉬움

03 학습

정보이득이 크도록 트리 발전
정보이득 : 자식과 부모의 불순도 차이

10

결정 트리 Yes / No에 대한 질문을 이어가며 정답을 찾아 학습하는 알고리즘



04 지니 불순

- 결정 트리의 각 단계에서 데이터를 분할하는 기준

지니 불순도 = 1 − (음성 클래스 비율^2 + 양성 클래스 비율^2)

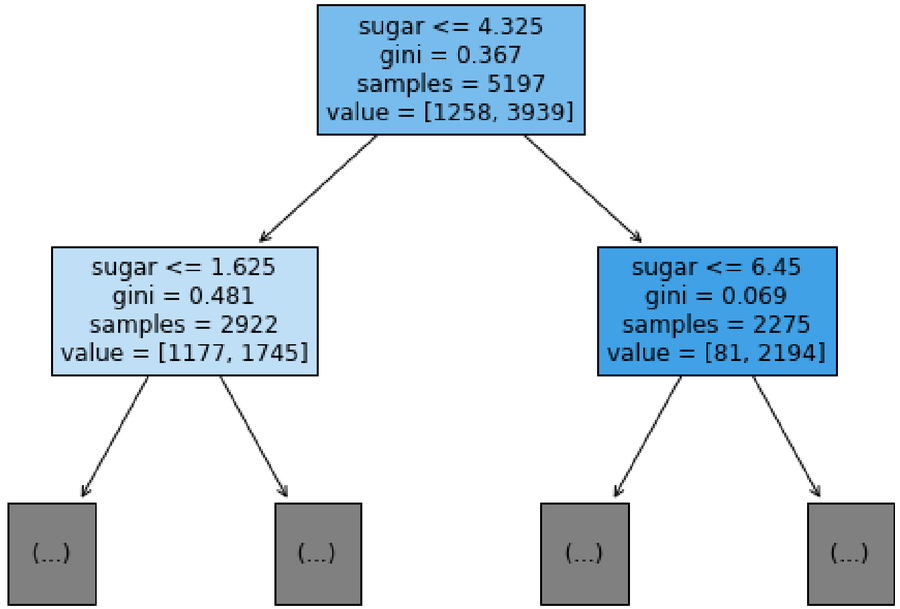
05 정보이득

- 자식과 부모의 불순도 차이

$$\text{부모 불순도} - \left(\frac{\text{왼쪽 샘플수}}{\text{부모 샘플수}}\right) \times \text{왼쪽 불순도} - \left(\frac{\text{오른쪽 샘플수}}{\text{부모 샘플수}}\right) \times \text{오른쪽 불순도}$$

- 지니 불순도 이외에 '엔트로피 불순도'를 사용할 수 있다

결정 트리 Yes / No에 대한 질문을 이어가며 정답을 찾아 학습하는 알고리즘



06 가지치기

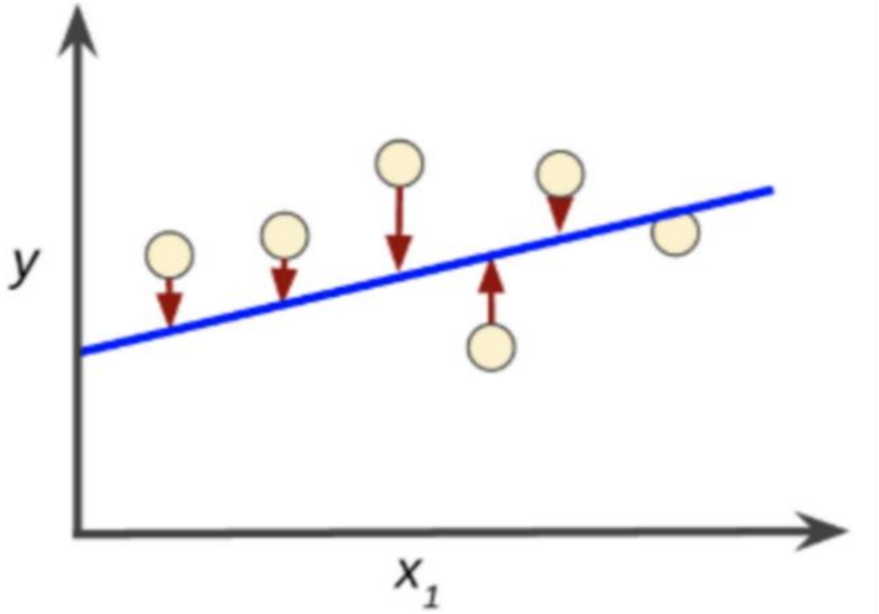
- 트리를 제한없이 훈련하면 과대적합 되기 쉬움
- 트리의 성장을 제한하는 것

모델 훈련 방법

손실함수
경사 하강법
과대적합과 과소적합
검증세트

손실함수

머신러닝 모델이 잘 훈련되었는지 여부를 보여주는 지표

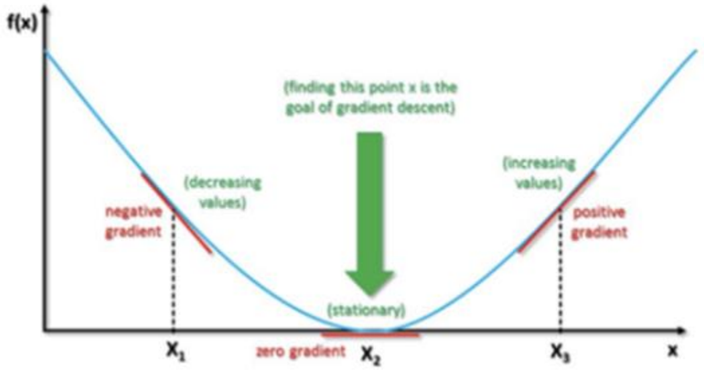


ex) 평균제곱오차

- 모델은 손실함수의 값이 작아지는 방향으로 훈련
- 경사 하강법을 사용하여 손실함수를 최소화

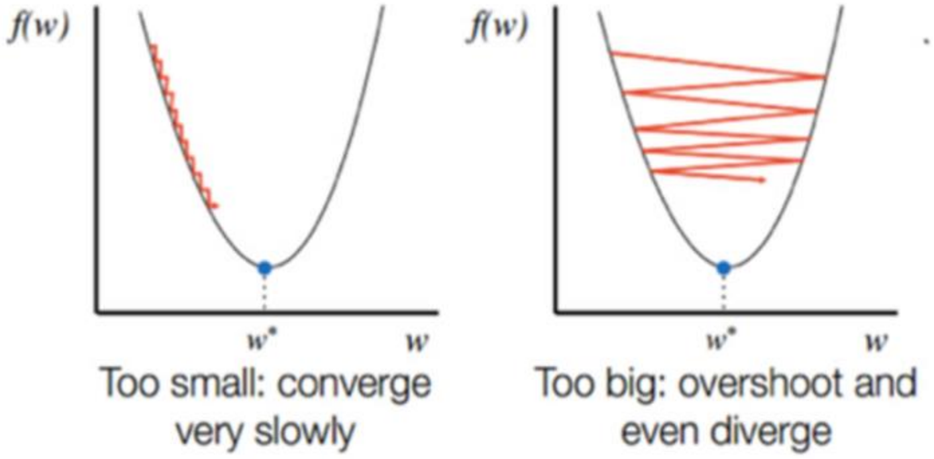
14

경사 하강법의개념



경사 하강법

- 손실을 줄이는 알고리즘
- 함수의 기울기(경사)를 구해서 기울기가 낮은 쪽으로 이동시켜서 극값(최적값)에 이를 때까지 반복



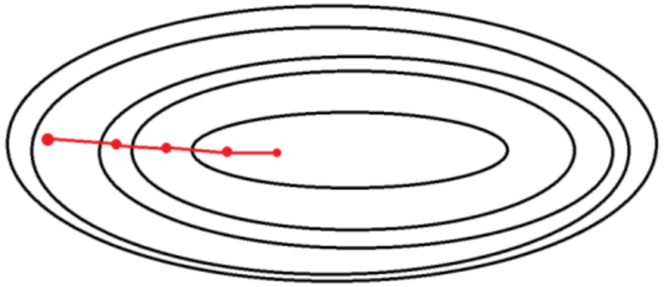
경사 하강법의 Step Size

Step size가 너무 큰 경우:
- 전역 최소값(global minimum)을 지나칠 수 있다

Step size가 너무 작은 경우:
- 학습 시간이 오래 걸린다
- 지역 최소값에 수렴할 수 있다

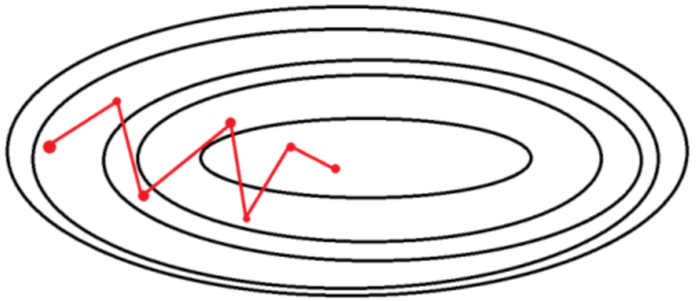
여러가지 경사 하강법

배치 경사 하강법



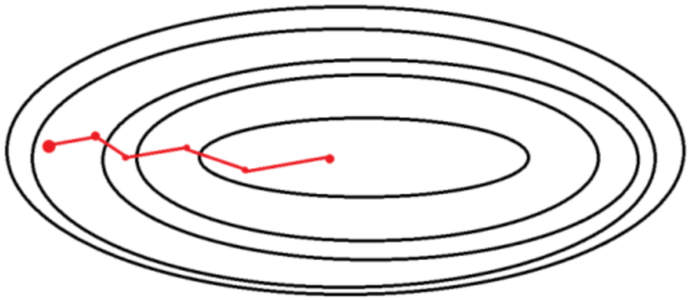
- 한번에 전체 데이터를 모두 사용
- 학습하는데 많은 시간 소요

확률적 경사 하강법



- 한번에 한 개의 데이터를 무작위로 선택하여 사용

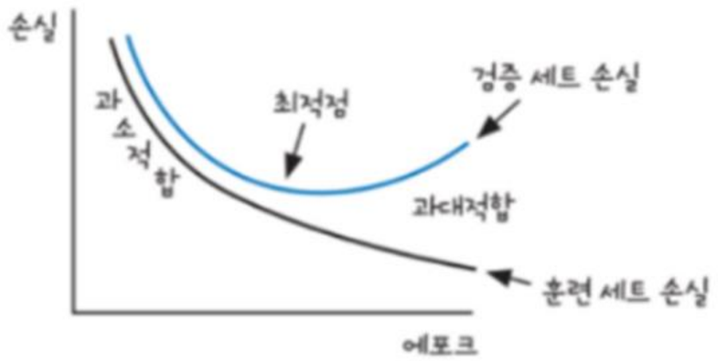
미니배치 경사 하강법

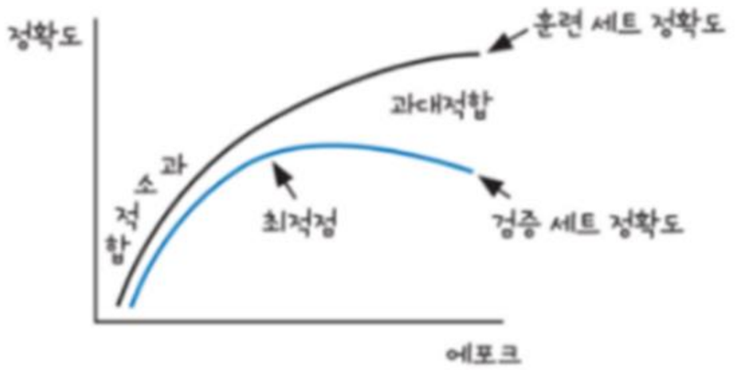


- 전체 데이터 중에서 일부의 데이터를 무작위로 선택하여 사용
- 배치 경사 하강법 보다 계산량이 적다
- 확률적 경사 하강법보다 빠르게 된다

16

과대적합과 과소적합





과대적합

모델이 훈련데이터에서만 좋은 결과를 보여주는 현상

해결방법

- 데이터 수 증가
- 교차검증
- 에포크 증가

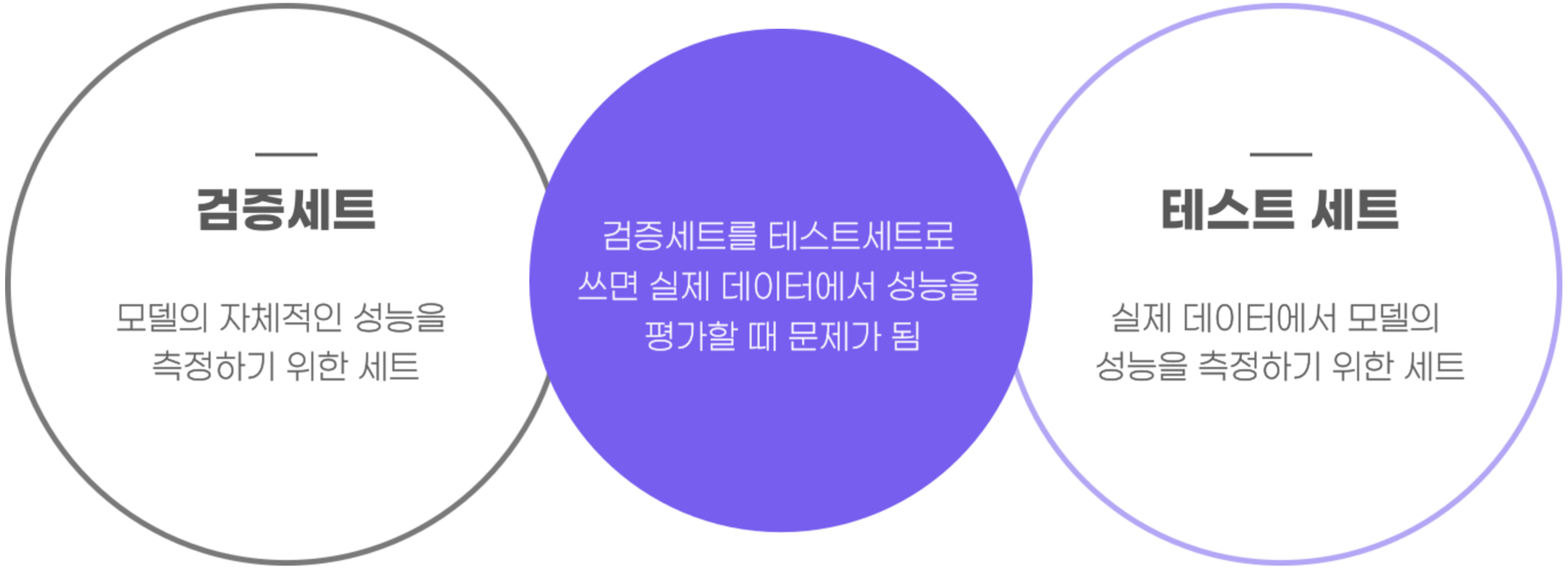
과소적합

모델이 전반적으로 좋지 않은 결과를 보여주는 현상

해결방법

- 규제
- 에포크 감소

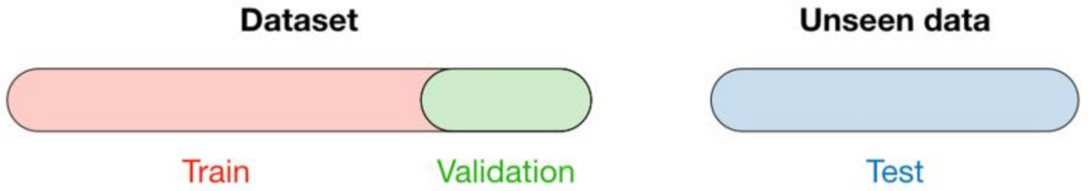
검증세트


검증세트
모델의 자체적인 성능을
측정하기 위한 세트
검증세트를 테스트세트로
쓰면 실제 데이터에서 성능을
평가할 때 문제가 됨
테스트 세트
실제 데이터에서 모델의
성능을 측정하기 위한 세트

검증세트 이용법

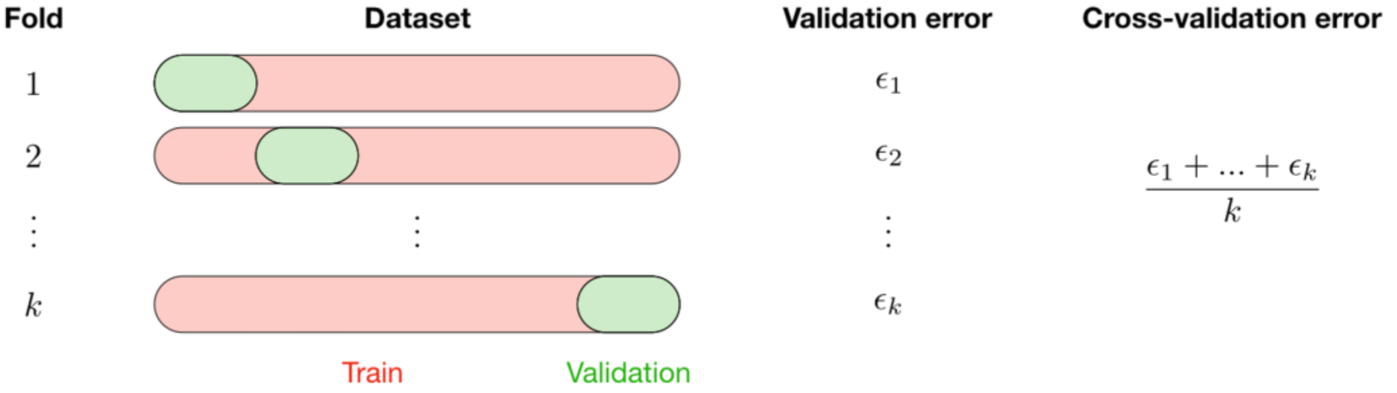
01 홀드아웃 방식

훈련데이터에서 일부(약 20%)를 떼어 사용



02 K-폴드 방식

구간을 K개로 나누어 각각 검증 세트로 사용
적은 데이터로 안정적인 오차를 구할 수 있다



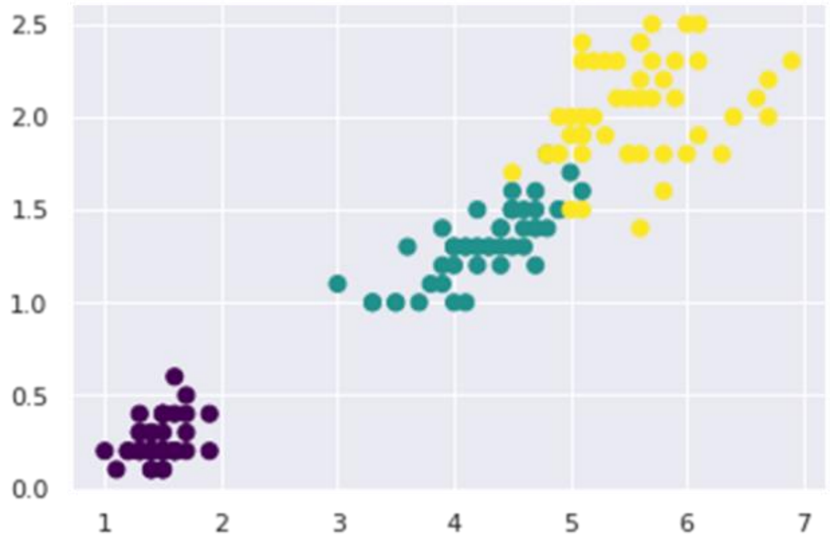
19

비지도 학습

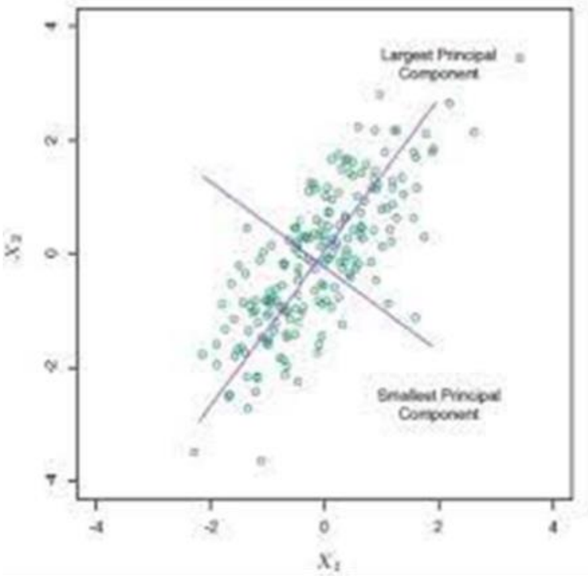
비지도학습의 개념
K 평균 알고리즘
클러스터링 사례

비지도 학습

- 모델을 훈련할 때 해결하려는 문제의 정답을 주지 않고 진행
- 주로 “분류" 모델

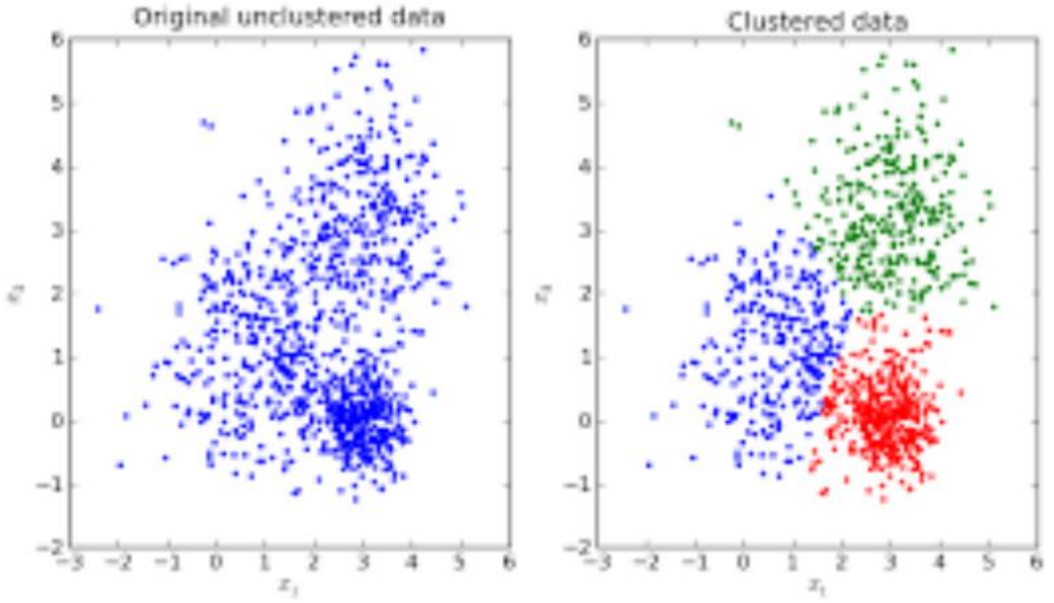


clustering



PCA

K-평균 알고리즘



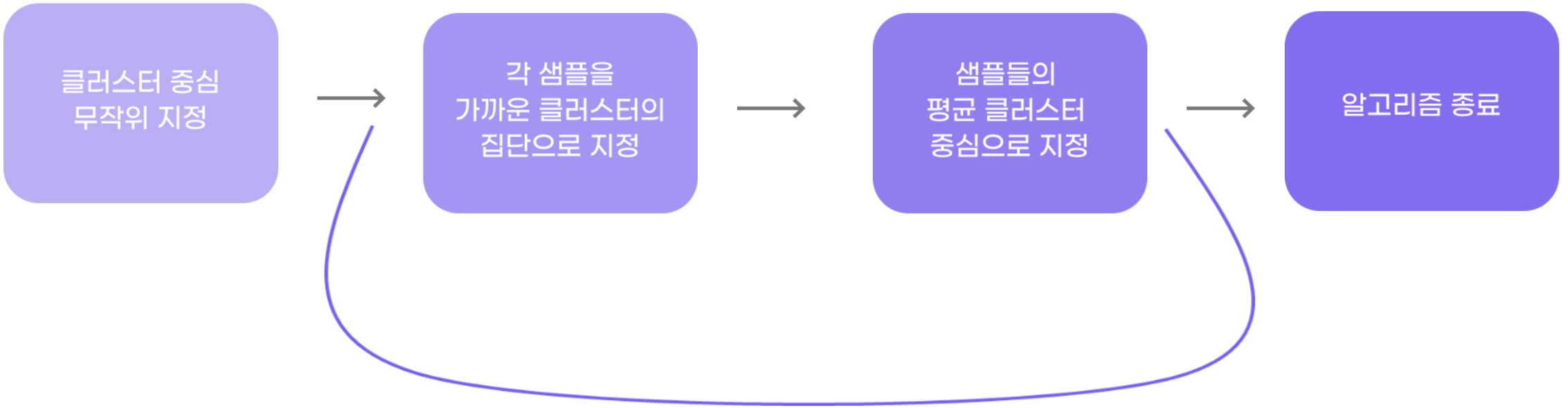
"클러스터의 중심을 찾아 분류하는 알고리즘"

클러스터란?

비슷한 데이터끼리 모아 만들어진 그룹

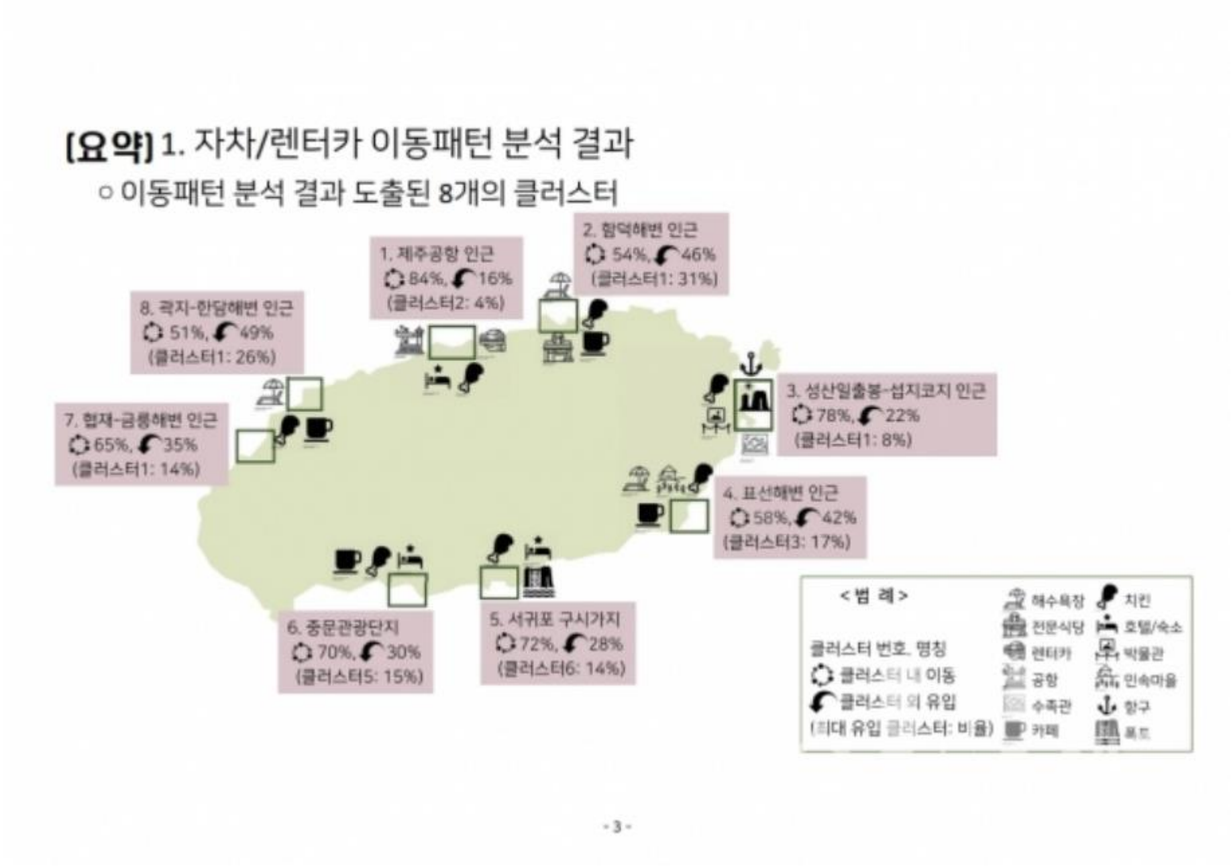
K-평균 알고리즘

알고리즘 작동 순서



23

클러스터링 사례



제주방문관광객 이동패턴 빅데이터 분석

제주도 공공데이터와 SKT 네비게이션 데이터 속에서 제주도 관광객들이 주로 방문하는 지역이 8곳으로 나뉜다는 것을 파악

24

01

하나,

인공지능이란

인공지능이란

02

두울,

머신러닝이란

지도학습
모델 훈련 방법
비지도학습

03

세엣,

딥러닝이란

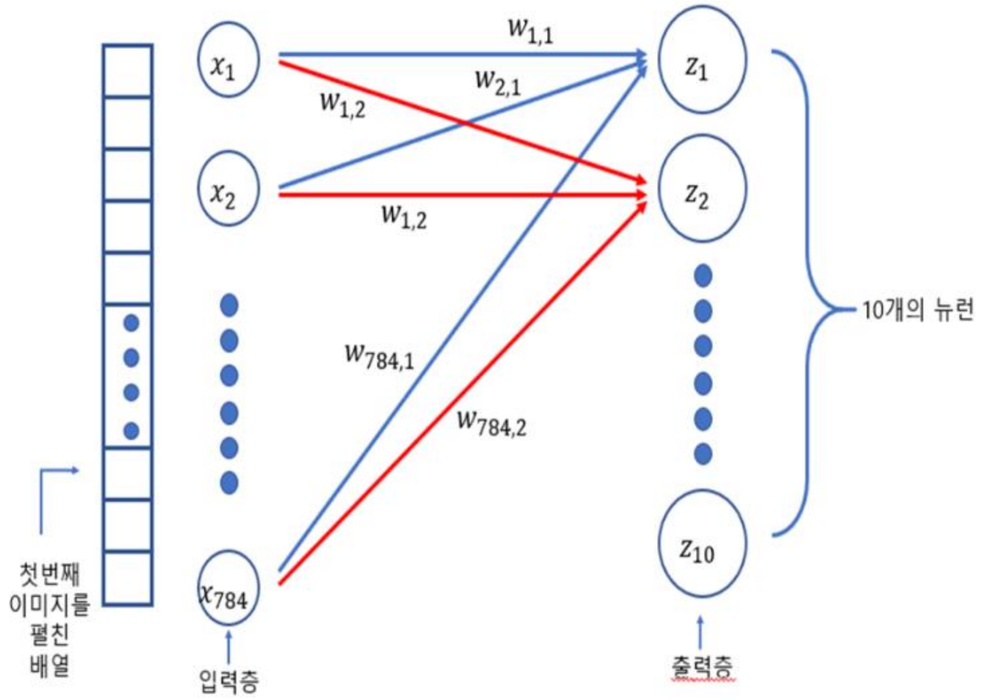
신경망
+ 공부 방법

25

신경망

인공신경망
심층신경망
드롭아웃
신경망 사례

인공 신경망



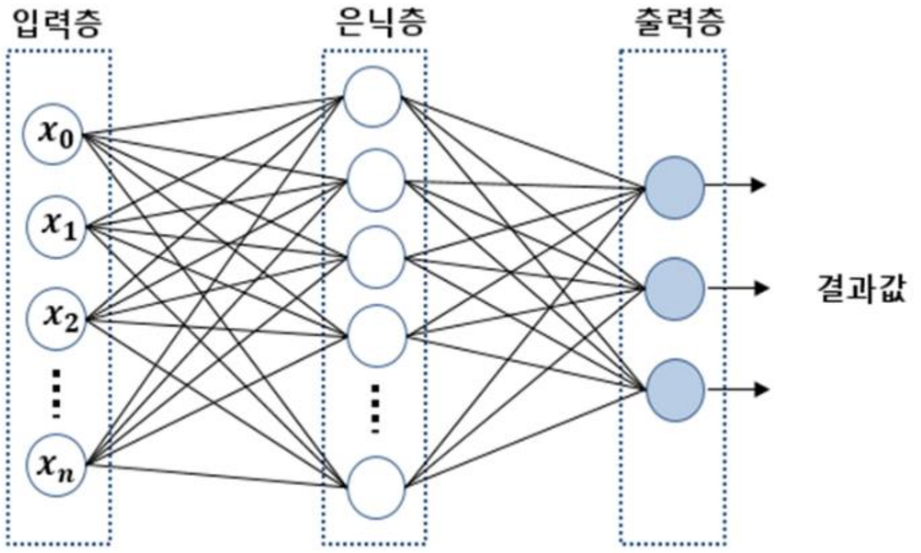
높은 성능을 발휘하고 있는 머신러닝 알고리즘 중 하나

- '출력층', '입력층', '은닉층'으로 구성
- 각 층은 1개 이상의 뉴런으로 이뤄짐
- 뉴런들은 활성화 함수를 통해 입력된 값을 다음 층으로 전달할지 여부를 결정

신경망의 훈련

- 입력층의 노드와 출력층의 노드는 서로 연결되어 있음
- 각 연결 관계에는 '가중치'가 부여되어 있으며 훈련 목적은 적절한 가중치를 찾는 것
- 대부분은 역전파 방법을 사용하여 신경망 훈련을 진행

심층 신경망



'은닉층'의 개수가 2개 이상일 때의 인공신경망

은닉층

- 입력층과 출력층 사이에 존재하는 층들을 통칭
- 입력층과 출력층 사이에서 추가적인 학습을 위해 각 은닉층마다 또 다른 활성화 함수를 사용 (ex. ReLU, Sigmoid, Softmax)

28

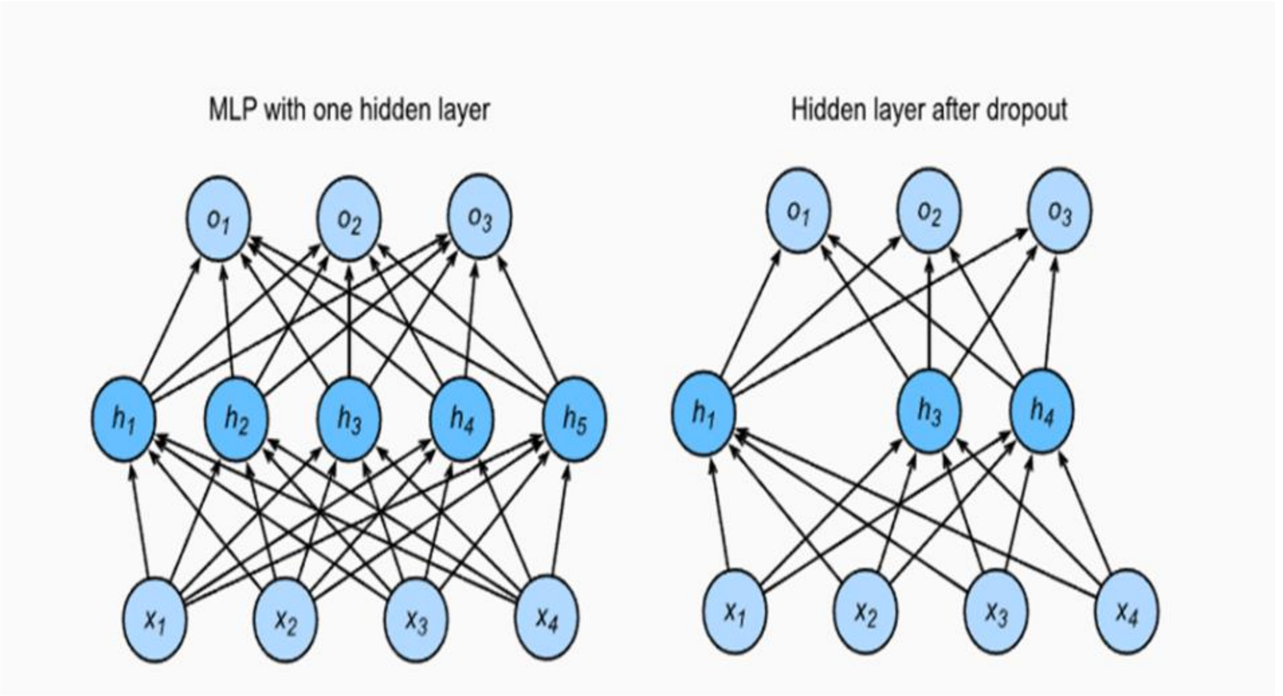
드롭아웃

개념

- 신경망의 과대적합을 해결하는 방법
- 일부 뉴런을 모델 학습에 참여하지 않게 하는 방식으로 진행

효과

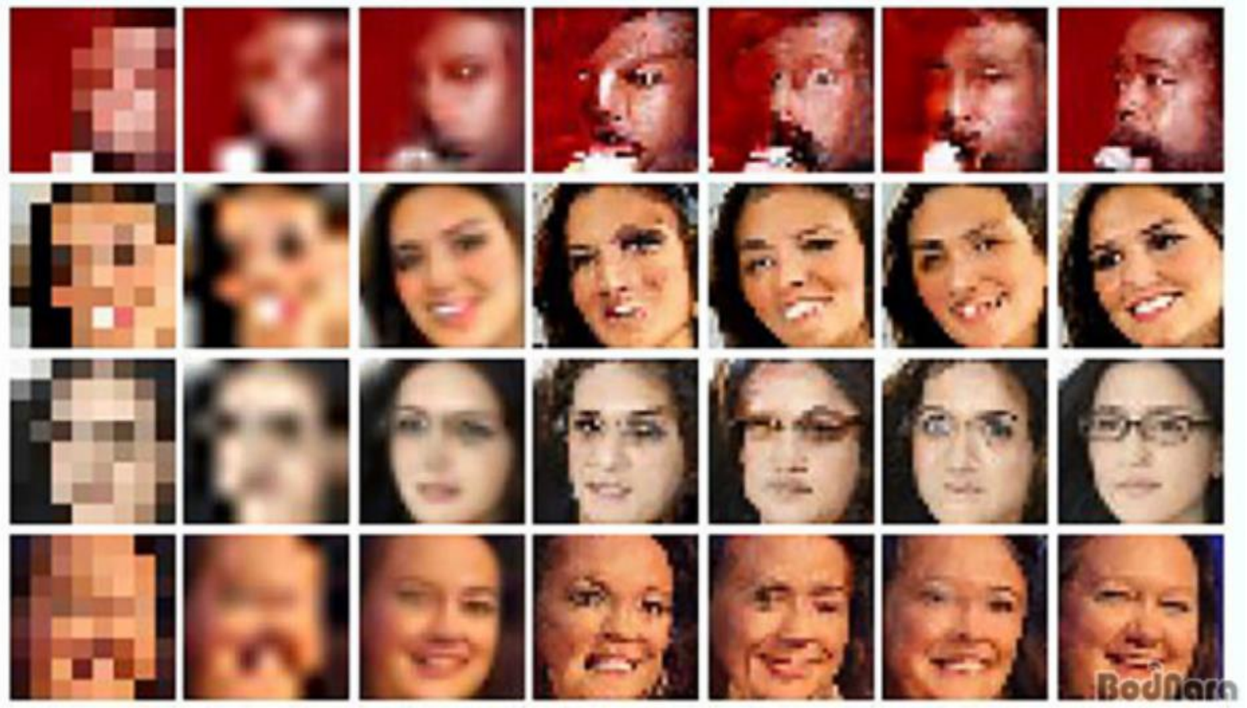
- 모델의 특정 뉴런에 대한 과대한 의존 감소



29

인공 신경망 사례

2017년 초반, 구글의 연구자들이 딥러닝 네트워크를 이용해 얼굴 이미지를 저해상도로 변환시킨 후 이미지가 무엇과 유사한 형태를 보이는지 예측하는 모델 개발



30

Notion I팀 개별 공부 페이지

📖

팀원별 공부방

일요일 15시까지 자습한 내용 업로드 해주세요~~

- 김선엽
- 송다애
- 채현우
- 정승원
- 윤이영

📈 로지스틱 회귀 (Logistic Regression)

로지스틱 회귀 이진분류

선형 방정식을 학습한 뒤, 시그모이드 함수(Sigmoid Function) = 로지스틱 함수(Logistic Function)를 사용하여 Class에 속할 확률을 구하는 모델

$$P = \frac{1}{1 + e^{-f(X)}} \qquad f(X) = \omega X + \beta$$

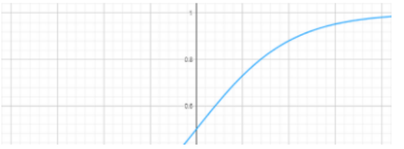
로지스틱 회귀 다중분류

각각의 feature에 대하여 선형방정식을 학습한 뒤 소프트맥스 함수(Softmax Function)를 사용하여 Class에 속할 확률을 구하는 모델

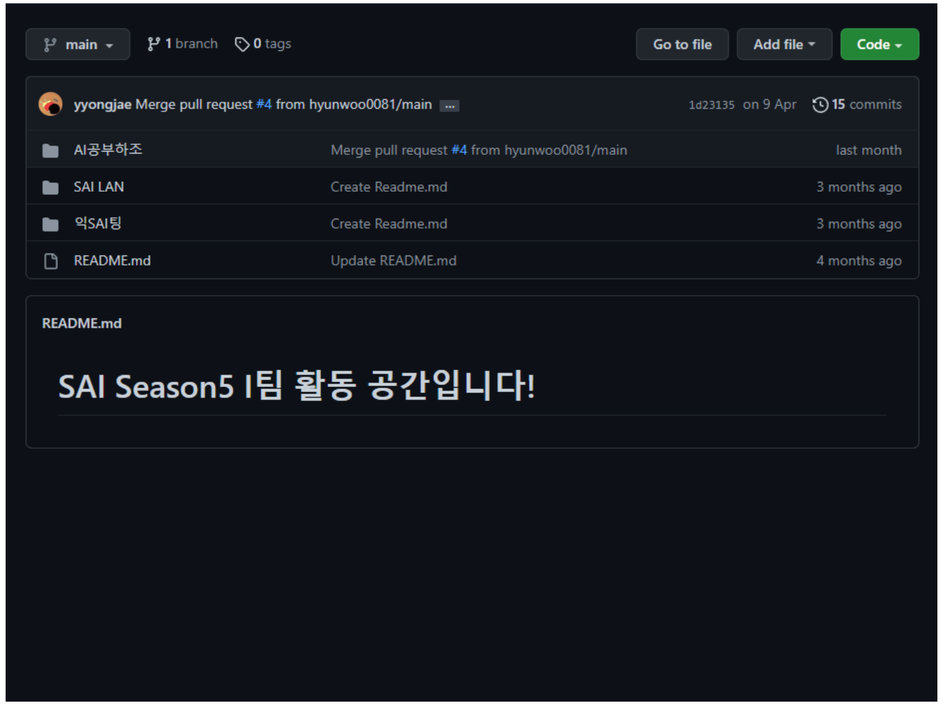
〽️ 시그모이드 함수 (Sigmoid Function = Logistic Function)

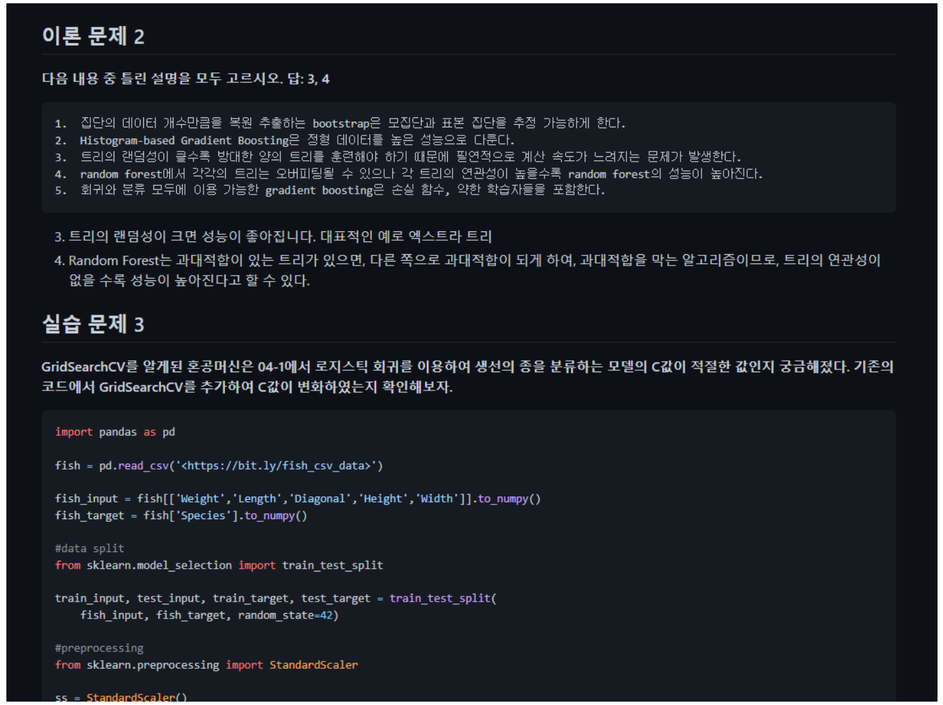
$$\phi = \frac{1}{1 + e^{-z}}$$

오즈(odds)비의 log를 씌운 값을 변형(로짓변환)하여 얻어진 함수🦁



Github I팀 문제풀이 페이지





이론 문제 2

다음 내용 중 틀린 설명을 모두 고르시오. 답: 3, 4

1. 집단의 데이터 개수만큼을 복원 추출하는 bootstrap은 모집단과 표본 집단을 추정 가능하게 한다.
2. Histogram-based Gradient Boosting은 정형 데이터를 높은 성능으로 다룬다.
3. 트리의 랜덤성이 클수록 방대한 양의 트리를 훈련해야 하기 때문에 필연적으로 계산 속도가 느려지는 문제가 발생한다.
4. random forest에서 각각의 트리는 오버피팅될 수 있으나 각 트리의 연관성이 높을수록 random forest의 성능이 높아진다.
5. 회귀와 분류 모두에 이용 가능한 gradient boosting은 손실 함수, 약한 학습자들을 포함한다.

3. 트리의 랜덤성이 크면 성능이 좋아집니다. 대표적인 예로 엑스트라 트리
4. Random Forest는 과대적합이 있는 트리가 있으면, 다른 쪽으로 과대적합이 되게 하여, 과대적합을 막는 알고리즘이므로, 트리의 연관성이 없을 수록 성능이 높아진다고 할 수 있다.

실습 문제 3

GridSearchCV를 알게된 혼공머신은 04-1에서 로지스틱 회귀를 이용하여 생선의 종을 분류하는 모델의 C값이 적절한 값인지 궁금해졌다. 기존의 코드에서 GridSearchCV를 추가하여 C값이 변화하였는지 확인해보자.

```
import pandas as pd

fish = pd.read_csv('<https://bit.ly/fish_csv_data>')

fish_input = fish[['Weight','Length','Diagonal','Height','Width']].to_numpy()
fish_target = fish['Species'].to_numpy()

#data split
from sklearn.model_selection import train_test_split

train_input, test_input, train_target, test_target = train_test_split(
    fish_input, fish_target, random_state=42)

#preprocessing
from sklearn.preprocessing import StandardScaler

ss = StandardScaler()
```

자료공유 I팀 강의자료 페이지

발표 자료

일요일에 15시까지 업로드해주세요

발표 순서

: 송다애 → 윤이영 → 채현우 → 김선엽 → 정승원

- 2주차(3/7)
- 3주차(3/14)
- 4주차(3/21)
- 5주차(3/28)
- 6주차(4/4)

참고자료

🤔Notion에서 수학 기호, 공식 작성하는 방법 (김선엽)
https://ordinary-code.tistory.com/66

🤔 SAI 깃허브에 실습코드 올리는 방법

감사합니다,
AI 공부하조였습니다 !