다양한 분류 알고리즘
'로지스틱 회귀'

4조
노지예, 부도현, 임청수, 한세림

목차

① K-최근접 이웃 분류기 Review

② 분류를 위한 회귀

③ 시그모이드 함수(Sigmoid)

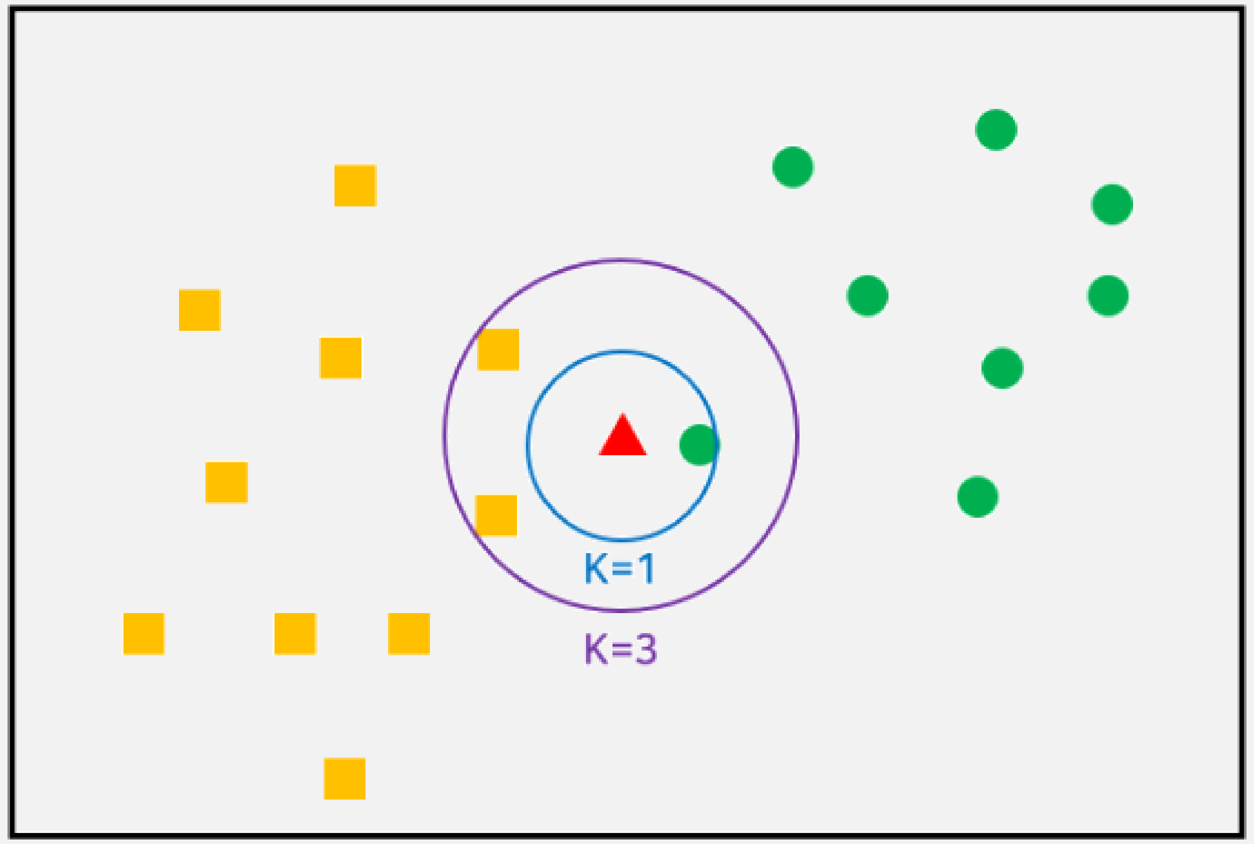
④ 교차 엔트로피 오차(Cross Entropy Error)

⑤ 로지스틱 회귀 분석의 장단점

K-최근접 이웃 분류기 Review

이전에 배운 분류 기법

K-최근접 이웃 분류기 (K-NN) 란?



"가까운 K개"의 데이터 중 "다수의 그룹"으로 분류

- (K = 1 일때) ● 으로 분류 (● : 1개, ■ : 0개)
- (K = 3 일때) ■ 으로 분류 (● : 1개, ■ : 2개)

※ Feature Scaling ☆☆☆ : Standardization / Normalization

- 데이터를 표현하는 특성(feature)들의 단위 통일
  (각 특성(feature)들을 **동일**하게 고려)

K-최근접 이웃 분류기의 한계

① 데이터의 특성들을 동일하게 고려하여 거리 측정

→ 특성 간의 중요도 차이를 반영하지 못함

② 인접한 K개의 데이터로만 분류

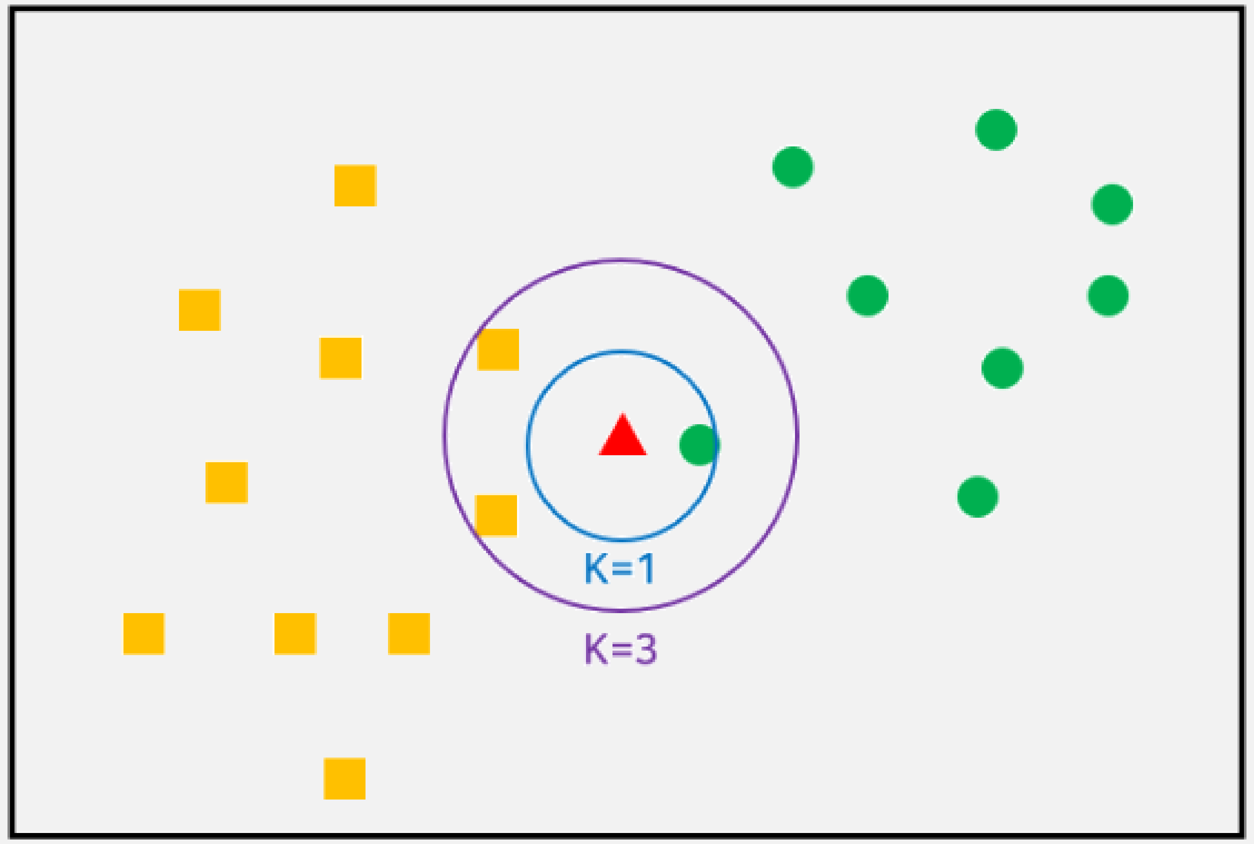
→ 특성이 늘어날수록 가까운 데이터 존재 X

K-최근접 이웃 분류기 Review

분류될 확률을 안다면?

from sklearn.neighbors import KNeighborsClassifer

KNeighborsClassifier.predict_proba(X)



"가까운 K개"의 데이터의 그룹 분포로 분류될 확률 표현

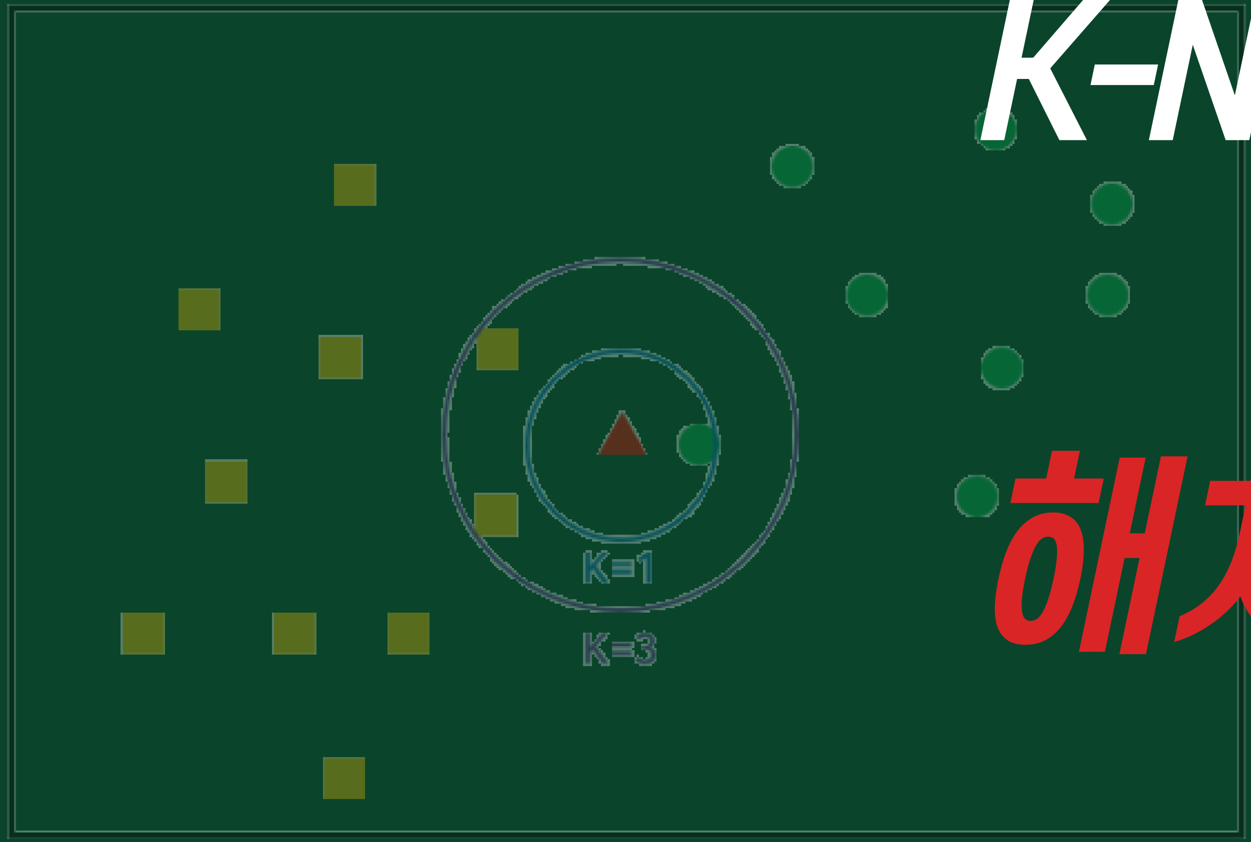
- (K = 1 일때) ● 일 확률 1/1 = 1, ■ 일 확률 0/1 = 0
- (K = 3 일때) ● 일 확률 1/3 = 0.33, ■ 일 확률 2/3 = 0.66

K-최근접 이웃 분류기 Review

분류될 확률을 안다면?

```
from sklearn.neighbors import KNeighborsClassifer

KNeighborsClassifier.predict_proba( X )
```



데이터의 그룹 분포로 분류될 확률 표현

- (K = 1 일때) ● 일 확률 1/1 = 1, ■ 일 확률 0/1 = 0
- (K = 3 일때) ● 일 확률 1/3 = 0.33, ■ 일 확률 2/3 = 0.66

K-NN의 한계...

해치웠나...?

K-최근접 이웃 분류기 Review

이게 확률이라고?

EX) 수능 점수로 합격/불합격 예측 사례

국어 점수
수학 점수
영어 등급
한국사 등급
화학1 점수
생명과학1 점수

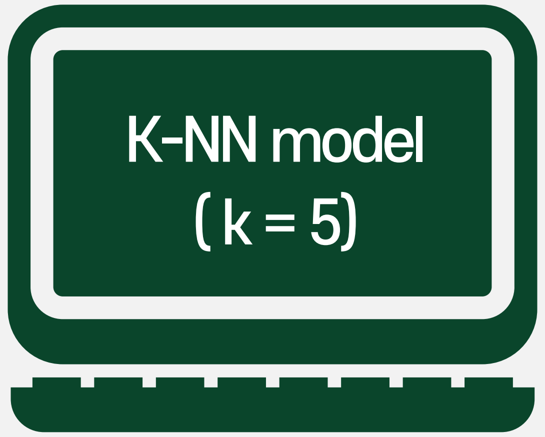
합격 / 불합격

표준화

<표준화 ver>
국어 점수
수학 점수
영어 등급
한국사 등급
화학1 점수
생명과학1 점수

합격 / 불합격

학습



예측

비타민대학교
머신러닝학과
"불합격"

확률 = 1

???

왜 이상하다고 느낄까?

① 확률이 1이라는 것이 아니다

→ 가장 가까운 5개의 데이터가 모두 '불합격'일 뿐임

② 인접한 데이터가 생각보다 멀리 떨어져 있을 수도 있다

→ 국수영탐 전부 비슷한 표준화 점수가 존재할까?

③ 비타민대학교의 과목별 비중이 고려되지 않았다

→ 한국사 등급과 수학 점수의 중요도가 동일할까?

분류를 위한 회귀

사람이 한다면?

EX) 입시업체들이 하는 합격 예측

국어 점수
수학 점수
영어 등급
한국사 등급
화학1 점수
생명과학1 점수

합격 / 불합격

<대학 변환 점수>
비타민대학교 기준
변환 점수

합격 / 불합격

합격 가능성
계산

비타민대학교
머신러닝학과

"합격" / "불합격"
분류

분류를 위한 회귀

비타민대학교의 성적산출식은 *비공개*

국어 점수
수학 점수
영어 등급
한국사 등급
화학1 점수
생명과학1 점수

합격 / 불합격

<표준화 ver>
국어 점수
수학 점수
영어 등급
한국사 등급
화학1 점수
생명과학1 점수

합격 / 불합격

1

합격 가능성
점수 계산

연속적인 값 예측
"회귀"!!!

2

비타민대학교
머신러닝학과

"합격" / "불합격"
분류

분류를 위한 회귀

이전에 배운 회귀 기법 - 선형 회귀 활용

<합격 가능성 점수> 선형 회귀 모델

국어 점수
수학 점수
영어 등급
한국사 등급
화학1 점수
생명과학1 점수

합격 / 불합격

표준화
수치화

<표준화 ver>
국어 점수
수학 점수
영어 등급
한국사 등급
화학1 점수
생명과학1 점수

+ ∞ / - ∞

학습

y : 합격 가능성 점수
x : 국수영탐 표준화 점수
A : 과목별 계수(과목별 중요도)
b : 상수

예측

합격 가능성 점수 y

y > 0 : 합격
y ≤ 0 : 불합격

선형 회귀 모델 알고리즘

선형 회귀 모델의 손실함수 : 평균오차제곱(MSE)

→ 평균오차제곱(MSE)을 최소화 하는 방향으로 계수 업데이트

<표준화 ver>
국어 점수
수학 점수
영어 등급
한국사 등급
화학1 점수
생명과학1 점수
x

+ ∞ / - ∞
y

$$minimize\ MSE$$

$$MSE = [y - (Ax+b)]^2$$

$$But, \quad y = \begin{cases} \infty & (if, True) \\ -\infty & (otherwise) \end{cases}$$

$$MSE = \begin{cases} [\infty - (Ax+b)]^2 & (if, True) \\ [-\infty - (Ax+b)]^2 & (otherwise) \end{cases}$$

평균오차제곱(MSE)이
항상 부정형(∞)으로
나와서 최소화 불가능

시그모이드 함수(Sigmoid)

로지스틱 회귀의 등장

Problem : 가능성 점수를 통해 선형회귀 학습이 불가능함(∵평균오차제곱(MSE) 사용 불가)

Solution : y를 가능성 점수가 아닌, 확률값(0 ~ 1)으로 두고 학습하자

Q1. 어떻게 확률값을 구할 것인가?

A1. 시그모이드 함수(Sigmoid) : 가능성 점수를 확률값으로 변환해주는 함수

Q2. 로지스틱 회귀란?

A2. 입력값 → 가능성 점수 → 확률값 → 확률값에 따라 그룹 **분류**

시그모이드 함수(Sigmoid)

로지스틱 회귀 모델은 분류 모델이다!

"로지스틱 회귀"는 ***"분류"***를 위한 확률값 회귀

① 회귀를 통해 해당 그룹으로 분류될 확률값을 예측하고

② 산출된 확률값을 근거로 **분류**한다

시그모이드 함수(Sigmoid)

확률값과 가능성 점수의 관계

합격으로 분류될 확률

p

불합격으로 분류될 확률 : 1 - p

$p \in (0, 1)$

$1 - p \in (0, 1)$

Odds

$\frac{p}{1-p}$

$Odds \in (0, \infty)$

Log(Odds)

$log(\frac{p}{1-p})$

$Log(Odds) \in (-\infty, \infty)$

Log(Odds)

: 가능성 점수 !

시그모이드 함수(Sigmoid) 유도

Log(Odds)

$log(\frac{p}{1-p})$

Log(Odds)
: 가능성 점수 !

가능성 점수 : $Likelihood\ Score\ \boldsymbol{LS(x)} = \boldsymbol{Ax + b}$

$$\boldsymbol{LS}(\boldsymbol{x}) = Log(\frac{p}{1-p})$$

$$e^{\boldsymbol{LS}(\boldsymbol{x})} = \frac{p}{1-p}$$

$$(1-p)e^{\boldsymbol{LS}(\boldsymbol{x})} = p$$

$$\therefore\ p = \frac{1}{1+e^{-\boldsymbol{LS}(\boldsymbol{x})}} \quad : \quad sigmoid(LS(x))$$

→ 시그모이드 함수(sigmoid) : 가능성 점수를 넣으면 확률값이 나오는 함수!

교차 엔트로피 오차(Cross Entropy Error)

확률값으로 어떻게 학습할까?

'손실함수 = 평균오차제곱(MSE)' 가능할까?

$$MSE = \begin{cases} [\,1 - sigmoid(Ax+b)\,]^2 \ (if, True) \\ [\,sigmoid(Ax+b)\,]^2 \ \ (otherwise) \end{cases}$$

$$\frac{d}{dA} MSE = \begin{cases} -2x[1-sigmoid(Ax+b)]^2[sigmoid(Ax+b)] \quad (if, True) \\ 2x[1-sigmoid(Ax+b)][sigmoid(Ax+b)] \ \ (otherwise) \end{cases}$$

$$\frac{d}{dA} MSE = \begin{cases} <0 \ (if, True\ and\ x>0) \\ >0 \ (if, True\ and\ x<0) \\ >0 \ (if, False\ and\ x>0) \\ <0 (if, False\ and\ x<0) \end{cases} \rightarrow$$

A의 값에 관계 없이 부호가 결정됨

→ 손실함수(MSE)가 최소가 되는 A 결정 ***"불가능"***

교차 엔트로피 오차(Cross Entropy Error)

통계적 이론 적용 - 데이터 1개

특정 데이터의 합격/불합격을 제대로 예측할 확률

$x_i : i$ 번째 데이터의 입력값

$p_i : i$ 번째 데이터가 합격일 확률

$1 - p_i : i$ 번째 데이터가 불합격일 확률

$$C_i = \begin{cases} 1 & (i\text{ 번째 데이터가 합격일때}) \\ 0 & (i\text{ 번째 데이터가 불합격일때}) \end{cases}$$

$$(i\text{ 번째 데이터가 제대로 분류되었을 확률}) = \begin{cases} Pr(C_i = 1 \mid x_i) = p_i & (i\text{ 번째 데이터가 합격일때}) \\ Pr(C_i = 0 \mid x_i) = 1 - p_i & (i\text{ 번째 데이터가 불합격일때}) \end{cases}$$

교차 엔트로피 오차(Cross Entropy Error)

통계적 이론 적용 - 데이터 n개

클수록 좋겠다!

우도 함수(Likelihood Function) : n개 데이터의 합격/불합격 제대로 예측할 확률

$$p_i = sigmoid(\,\boldsymbol{LS}(x_i)\,) = sigmoid(Ax_i + b)$$
$$1 - p_i = 1 - sigmoid(Ax_i + b)$$

$$Pr(C_i = c_i \mid x_i) = \{sigmoid(Ax_i + b)\}^{c_i} + \{1 - sigmoid(Ax_i + b)\}^{1-c_i} \quad , (c_i = 0 \text{ or } 1)$$

$$(n\text{개의 데이터}) : X = \{x_1,\ x_2, \dots, x_n\}, y = \{c_1,\ c_2, \dots, c_n\}$$

$$(n\text{개의 데이터를 제대로 분류할 확률} : Likelihood\ Function\ LF(A,b)\,)$$
$$= (1\,\text{번째 데이터를 제대로 분류할 확률}) \times \dots \times (n\,\text{번째 데이터를 제대로 분류할 확률})$$
$$= \prod_{i=1}^{n} Pr(C_i = c_i \mid x_i) = \prod_{i=1}^{n} [\{sigmoid(Ax_i + b)\}^{c_i} + \{1 - sigmoid(Ax_i + b)\}^{1-c_i}]$$

교차 엔트로피 오차(Cross Entropy Error)

통계적 이론 적용 - Skill

우도 함수가 곱의 형태로 되어있어서 미분이 어려움 → 미분 쉽게하기 위한 Skill

우도 함수 최대화 → 로그 우도 함수 최대화 (∵ Log(x) : 증가함수)

$$Maximize\ LF(A,b) \rightarrow Maximize\ Log(LF(A,b)) \rightarrow \underline{minimize\ -\{Log(LF(A,b))\}}$$

최소화? → 손실함수로 두면 되겠다!

- { Log(LF(A,b)) } : Cross Entropy Error

$$CEE(A,b) = -Log(LF(A,b))$$

$$= -\sum_{i=1}^{n}[c_i\ Log(sigmoid(Ax_i + b)) + (1 - c_i)\ Log(1 - sigmoid(Ax_i + b))]$$

로지스틱 회귀 알고리즘

회귀계수 A,b 학습 과정

손실함수인 Cross Entropy Error를 최소화하는 방향으로!

A와 b에 대하여 편미분한 값을 빼자

$$\frac{d}{dA}CEE(A,b) = -\sum_{i=1}^{n}[x_i\{c_i - sigmoid(Ax_i + b)\}]$$

$$\frac{d}{db}CEE(A,b) = -\sum_{i=1}^{n}\{c_i - sigmoid(Ax_i + b)\}$$

$$A_{k+1} = A_k - \alpha[\frac{d}{dA_k}CEE(A_k, b_k)] = A_k + \alpha\sum_{i=1}^{n}[x_i\{c_i - sigmoid(A_k x_i + b_k)\}]$$

$$b_{k+1} = b_k - \alpha[\frac{d}{db_k}CEE(A_k, b_k)] = b_k + \alpha\sum_{i=1}^{n}\{c_i - sigmoid(A_k x_i + b_k)\}$$

* α : 학습률(Learning rate)

로지스틱 회귀 모델의 장점

① 보다 정확한 확률값을 얻을 수 있다

② 이진 분류 성능이 뛰어나다

③ 적은 데이터에서도 뛰어난 성능을 보인다

④ 가볍고 빠르다

로지스틱 회귀 모델의 단점

① 가능성 점수를 ***선형 관계***로 만들 수 있어야 한다

ex) 특성 간의 시너지가 있는 데이터 분류 어려움 → K-NN, 결정트리에 적합

② 분류된 이유를 설명하기 어려움

- 회귀 계수가 양수/음수 : 확률 증가/감소

- 회귀 계수의 절댓값 : 클수록 확률에 큰 영향력 행사

→ 결정 트리, 랜덤 포레스트 등장 배경

로지스틱 회귀 이진분류를 마치며...

로지스틱 회귀 요약/정리

① Logistic : "논리학의"(인간의 합리적인 추론을 다루는 학문)

→ 사람처럼 생각하고 분류하는 모델

② 오즈 → 시그모이드 함수

우도함수 → 교차 엔트로피 오차

③ 반드시 분류 가능성 점수와 특성간의 선형관계가 존재해야함

THANK YOU

4조

노지예, 부도현, 임청수, 한세림