로지스틱회귀
(Logistic Regression)

한국공학대학교
전자공학부
채승호 교수

분류(Classification)

- 분류
  - 입력 데이터들을 주어진 항목(class)으로 나누는 방법
  - 이진 분류 & 다중 분류
    - Ex) 내일의 날씨가 좋을지 안 좋을지 예측
    - Ex) 수박/참외/사과 분류
  - 분류 문제는 클래스 Y를 직접 예측하기 보다는 Y가 특정 클래스일 확률 $P[Y = k|\mathbf{x}]$를 예측하고자 함

  - v.s. 회귀(Regression)
    - 입력된 데이터에 대해 연속된 값으로 예측
    - 패턴이나 경향성을 예측할 때 사용

지도학습의 학습 모델의 종류

	Regression	Classification
Linear Model	Linear Regression	Logistic Regression
Discriminant Analysis	-	LDA/QDA
Nonparametric	KNN	KNN, Naïve Bayesian
Tree	Regression Tree	Classification Tree
Ensemble	Bagging, Boosting	-
Support Vector	Support Vector Regression	Support Vector Machine
Neural Networks	Multi-layer perceptron and Deep learning	-

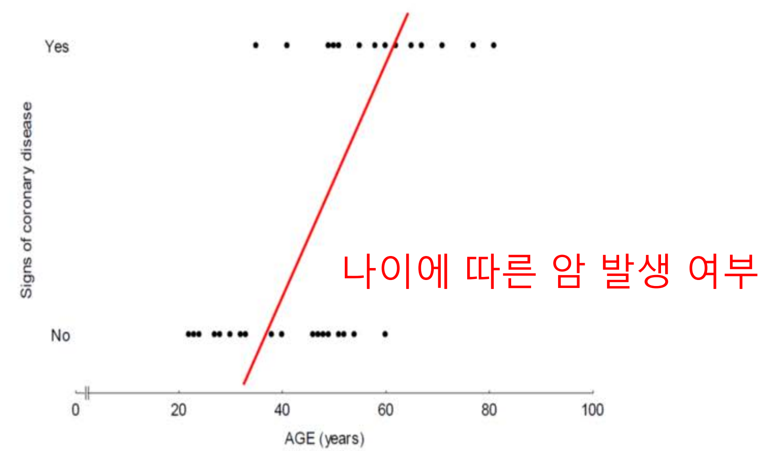
이진분류(Binary Classification)

- 이진 분류
  - 범주/클래스가 2개인 경우
  - 예제:
    - 청소년의 체중에 따른 비만도 (비만 / 정상)
    - 수박 껍질 무늬에 따른 숙성도 (잘 익음 / 덜 익음)
    - 공부 시간에 따른 시험합격 여부 (합격 / 불합격)
    - 나이에 따른 암 발생 여부 (암 발생 / 정상)

Age	CD
22	0
23	0
24	0
27	0
28	0
30	0
30	0
32	0
33	0
35	1
38	0

Age	CD
40	0
41	1
46	0
47	0
48	0
49	1
49	0
50	1
51	0
51	1
52	0

Age	CD
54	0
55	1
58	1
60	1
60	0
62	1
65	1
67	1
71	1
77	1
81	1

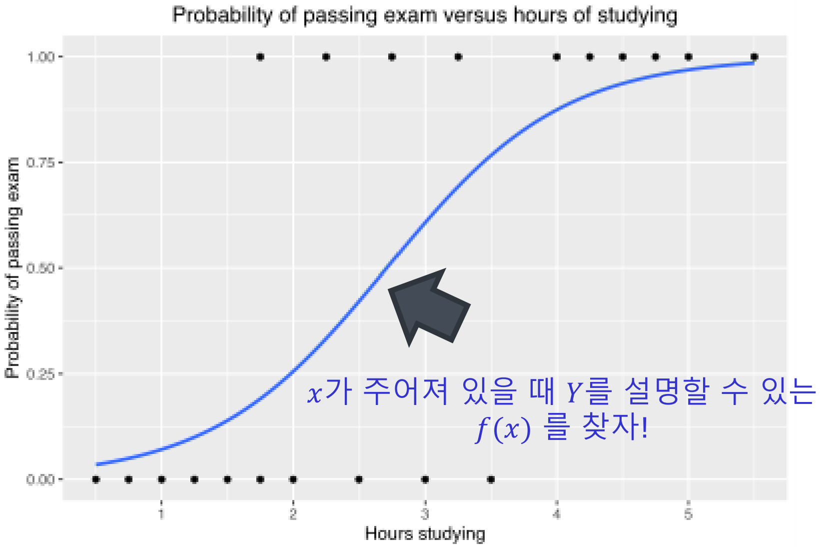


범주형 데이터에 선형 회귀를 적용할 수 없음!

이진분류(Binary Classification)

공부시간에 따른 시험합격 여부 (합격 또는 불합격)

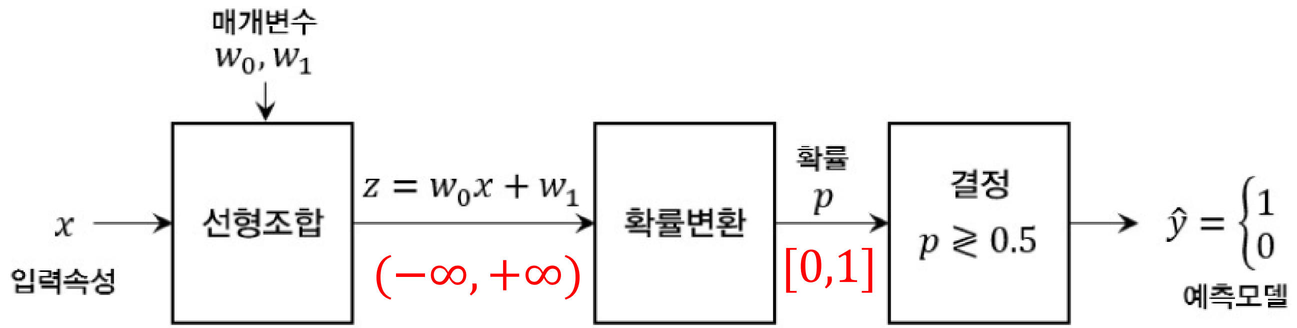
	공부시간	합격여부	$P[Y = k\|X]$
1	0.50	불합격	0
2	0.75	불합격	0
3	1.75	합격	1
4	2.00	불합격	0
5	2.25	합격	1



https://en.wikipedia.org/wiki/Logistic_regression

이진분류(Binary Classification)

- 선형모델을 이용한 이진분류 설계

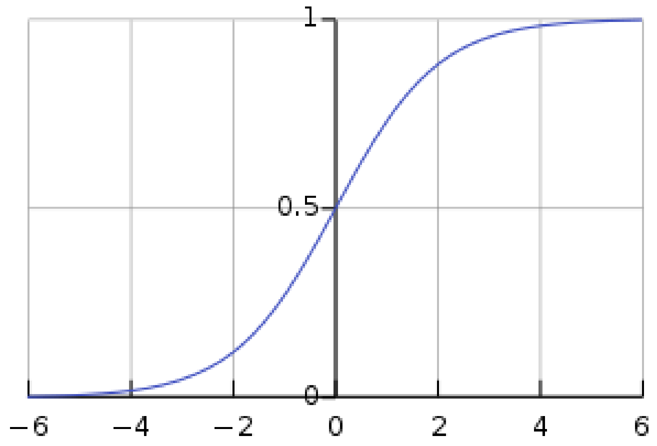


로지스틱회귀(Logistic Regression)

- Logistic Regression
  - 클래스에 대한 확률을 sigmoid 함수를 이용하여 모델링

$$f(z) = \frac{1}{1+e^{-z}}$$

    - 임의의 값을 $[0,1]$의 값으로 변환 → 입력 값이 클수록 1로 수렴, 작을수록 0로 수렴
    - 우수한 미분 특성 : $\frac{df(z)}{dz} = f(z)(1-f(z))$
  - 이진분류기 → 추정 확률이 50%가 넘으면 해당 클래스에 속한다고 예측



로지스틱회귀(Logistic Regression)

- 이진분류 ($Y = 1$ / $Y = 0$)에 대하여

$$P[Y = 1|x] = \frac{e^{w_0x+w_1}}{1+e^{w_0x+w_1}} = \frac{1}{1+e^{-(w_0x+w_1)}}$$

$$P[Y = 0|x] = \frac{1}{1+e^{w_0x+w_1}} = \frac{e^{-(w_0x+w_1)}}{1+e^{-(w_0x+w_1)}}$$

$$\underbrace{\log\left(\underbrace{\boxed{\frac{P[Y = 1|x]}{1 - P[Y = 1|x]}}}_{\text{odd}}\right)}_{\text{logit}} = \underbrace{w_0x + w_1}_{\text{linear regression}}$$

파라미터 추정

- 로지스틱 회귀에서의 우도
  - 우도(likelihood; 가능성): 어떤 모델로부터 현재 데이터를 관측할 확률
  - 주어진 x에 대해 $Y = 1$ / $Y = 0$일 확률

$$p(x) = P[Y=1|x] = \frac{e^{w_0 x + w_1}}{1+e^{w_0 x + w_1}}, \quad 1 - p(x) = P[Y=0|x] = \frac{1}{1+e^{w_0 x + w_1}}$$

  - 우도

$$l = \prod_{n:y_n=1} p(x_n) \prod_{n:y_n=0} 1 - p(x_n) = \prod_n p(x_n)^{y_n} (1 - p(x_n))^{1-y_n}$$

	공부시간	합격여부	$P[Y=1\|x]$
1	0.50	불합격	0
2	0.75	불합격	0
3	1.75	합격	1
4	2.00	불합격	0
5	2.25	합격	1

관측할 확률: $\frac{1}{1+e^{w_0(0.5)+w_1}}$

$\times$

관측할 확률: $\frac{1}{1+e^{w_0(0.75)+w_1}}$

$\times$

관측할 확률: $\frac{e^{w_0(1.75)+w_1}}{1+e^{w_0(1.75)+w_1}}$

$\times$

⋮

해당 data set을 만들어 낼 가능성= 우도

파라미터 추정

- 최대 우도 추정 (Maximum Likelihood Estimation)
  - 우도를 최대화는 최적 매개변수 **w** 파라미터를 찾아야 함!

$$\mathbf{w}^* = \arg\max l = \arg\max \prod_{n=0}^{N-1} p(x_n)^{y_n}(1 - p(x_n))^{1-y_n}$$

(AB)' = A'B+AB'

곱셈에 대한 미분 → 연산의 복잡성

  - Log-likelihood 함수를 최대화 시키는 문제로 변환
    - $f(x)$를 최대화 하는 것은 $\log(f(x))$를 최대화 하는 것과 동일

$$\mathbf{w}^* = \arg\max \mathcal{L} = \arg\max \sum_{n=0}^{N-1} y_n \log p(x_n) + (1 - y_n)\log(1 - p(x_n))$$

곱셈 → 덧셈

$$\mathbf{w}^* = \arg\min L = \arg\min - \underbrace{\sum_{n=0}^{N-1} y_n \log p(x_n) + (1 - y_n)\log(1 - p(x_n))}_{\text{cross-entropy loss}}$$

파라미터 추정

- Cross-entropy loss 최소화 문제의 최적해

$$\mathbf{w}^* = \arg\min L = \arg\min \underbrace{-\sum_{n=0}^{N-1} y_n \log p(x_n) + (1-y_n)\log(1-\ p(x_n))}_{\text{cross-entropy loss}}$$

$$w_0^*, w_1^* = \arg\min L(w_0, w_1)$$

$$\frac{\partial L(w_0,w_1)}{\partial w_0} = 0,\ \frac{\partial L(w_0,w_1)}{\partial w_1} = 0 \quad \rightarrow \quad w_0^*, w_1^*$$

파라미터 추정

- 경사하강법

cross-entropy loss $L = -\sum_{n=0}^{N-1} y_n \log p(x_n) + (1-y_n)\log(1-\ p(x_n))$

$$p(x_n) = \frac{1}{1+e^{-z_n}} = \frac{1}{1+e^{-(w_0 x_{0,n} + w_1 x_{1,n} + w_2)}}$$

  - 업데이트 규칙 (2 입력에 대해)
    - $w_0[t+1] = w_0[t] - \alpha \frac{\partial}{\partial w_0} L(w_0, w_1, w_2)$
    - $w_1[t+1] = w_1[t] - \alpha \frac{\partial}{\partial w_1} L(w_0, w_1, w_2)$
    - $w_2[t+1] = w_2[t] - \alpha \frac{\partial}{\partial w_2} L(w_0, w_1, w_2)$
  - 도함수
    - $\frac{\partial}{\partial w_0} L(w_0, w_1, w_2) = \frac{1}{N}\sum_{n=0}^{N-1}(p_n - y_n)x_{0,n}$
    - $\frac{\partial}{\partial w_1} L(w_0, w_1, w_2) = \frac{1}{N}\sum_{n=0}^{N-1}(p_n - y_n)x_{1,n}$
    - $\frac{\partial}{\partial w_2} L(w_0, w_1, w_2) = \frac{1}{N}\sum_{n=0}^{N-1}(p_n - y_n)$

파라미터 추정

- 로지스틱 (logistic) 손실 함수
  - Y를 0/1로 표현하는 대신 Y를 -1/1로 표현
  - $P[Y=1|x] = \frac{e^{w_0x+w_1}}{1+e^{w_0x+w_1}} = \frac{1}{1+e^{-(w_0x+w_1)}} = \frac{1}{1+e^{-y(w_0x+w_1)}}$
  - $P[Y=-1|x] = \frac{1}{1+e^{w_0x+w_1}} = \frac{1}{1+e^{-y(w_0x+w_1)}}$

$$l = \prod_n \frac{1}{1+e^{-y_n(w_0x_n+w_1)}}$$

$$\text{Logistic loss} = -\log l = -\sum_n \log\left(1+e^{-y_n(w_0x_n+w_1)}\right)$$

일반적인 로지스틱 회귀

- 하나 이상의 독립 변수에 대한 로지스틱 회귀

$$\log\left(\frac{P[Y=1|X]}{1-P[Y=1|X]}\right) = w_0 x + w_1$$

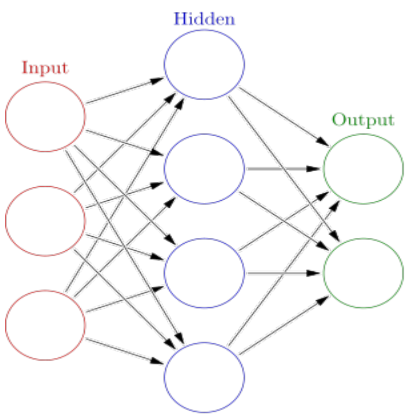
$$\log\left(\frac{P[Y=1|X]}{1-P[Y=1|X]}\right) = w_0 x_0 + w_1 x_1 + w_2 x_2 + \cdots + w_p$$

$$P[Y=1|X] = \frac{e^{w_0 x_0 + w_1 x_1 + w_2 x_2 + \cdots + w_p}}{1 + e^{w_0 x_0 + w_1 x_1 + w_2 x_2 + \cdots + w_p}} = \frac{1}{1 + e^{-(w_0 x_0 + w_1 x_1 + w_2 x_2 + \cdots + w_p)}}$$

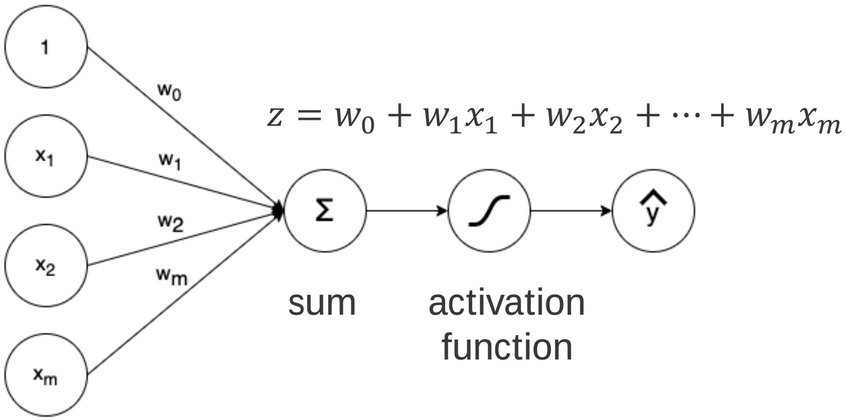
$$P[Y=0|X] = \frac{1}{1 + e^{w_0 x_0 + w_1 x_1 + w_2 x_2 + \cdots + w_p}} = \frac{e^{-(w_0 x_0 + w_1 x_1 + w_2 x_2 + \cdots + w_p)}}{1 + e^{-(w_0 x_0 + w_1 x_1 + w_2 x_2 + \cdots + w_p)}}$$

로지스틱 회귀와 퍼셉트론

- 퍼셉트론: 하나 이상의 독립 변수에 대한 선형 회귀와 같이 확장



인공신경망



퍼셉트론

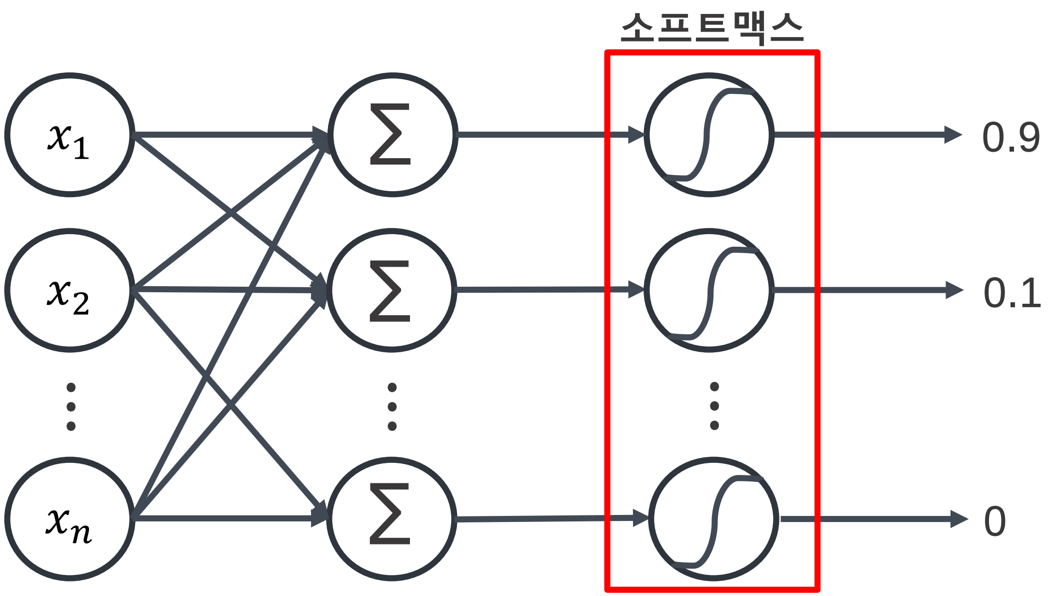
  - 다양한 활성화 함수가 사용될 수 있으나 로지스틱이 일반적으로 사용

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

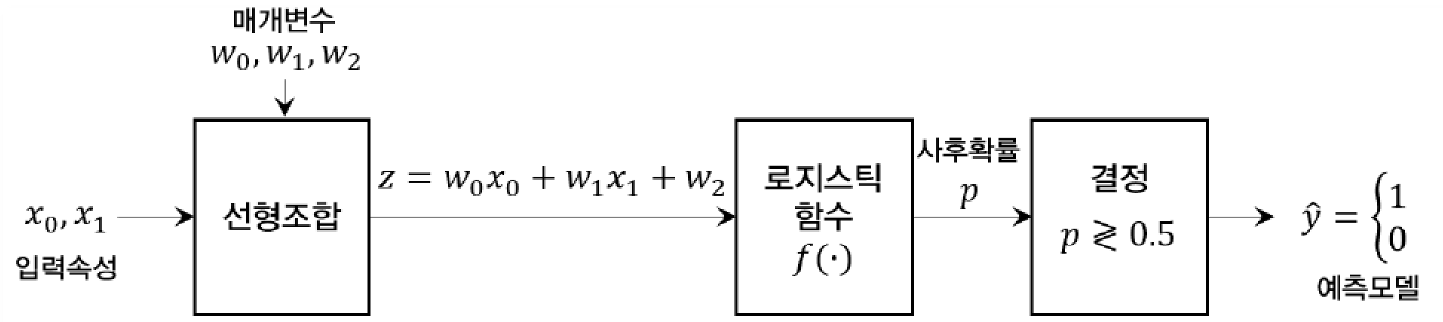
  - $Y = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-(w_0+w_1x_1+w_2x_2+\cdots+w_mx_m)}} = \frac{e^{w_0+w_1x_1+w_2x_2+\cdots+w_mx_m}}{1+e^{w_0+w_1x_1+w_2x_2+\cdots+w_mx_m}}$
  - 로지스틱 회귀와 퍼셉트론은 기본적으로 동일한 모델

로지스틱 회귀

- 다항 로지스틱 회귀(softmax)
  - 클래스가 3개 이상일 때 사용
  - 전체 확률의 합 = 1
  - 가장 높은 확률을 가진 항목이 분류값이 됨



2입력 이진분류 모델의 설계 및 학습

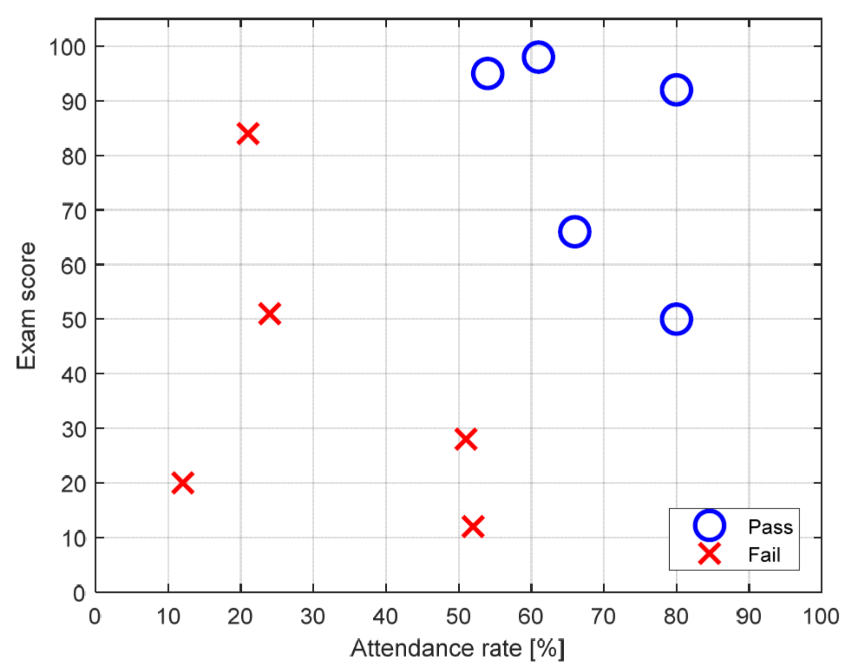


데이터 번호	입력		출력	선형조합	사후확률	예측값
	x_0	x_1	y	$z = w_0x_0 + w_1x_1 + w_2$	$p = f(z)$	$\hat{y}$
0	$x_{0,0}$	$x_{1,0}$	y_0	z_0	p_0	$\hat{y}_0$
1	$x_{0,1}$	$x_{1,1}$	y_1	z_1	p_1	$\hat{y}_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$x_{0,n}$	$x_{1,n}$	y_n	z_n	p_n	$\hat{y}_n$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$N-1$	$x_{0,N-1}$	$x_{1,N-1}$	y_{N-1}	z_{N-1}	p_{N-1}	$\hat{y}_{N-1}$

2입력 이진분류 모델의 설계 및 학습

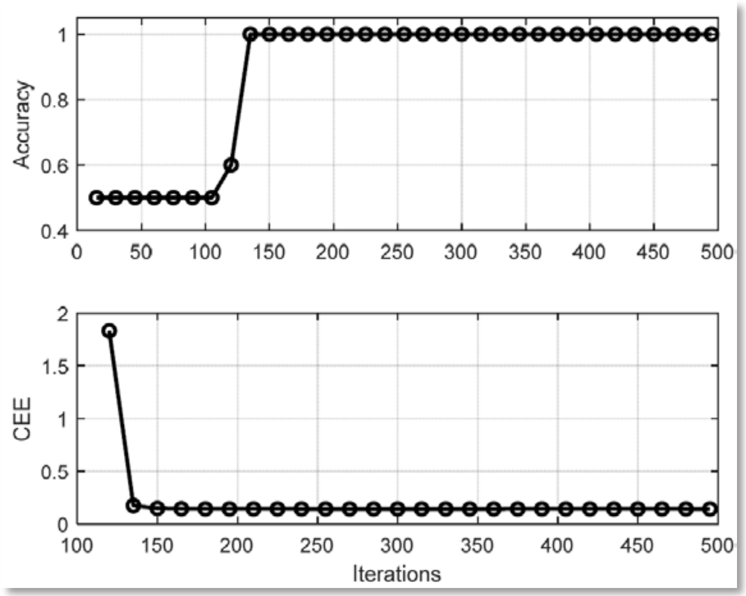
- 예제
  - 입력: 출석률, 시험성적
  - 출력: Pass 또는 Fail

데이터 번호	입력		출력 y (이수 여부, 1=pass, 0=fail)
	x_0 (출석률)	x_1 (시험성적)	
0	21	84	0
1	54	95	1
2	80	50	1
3	51	28	0
4	66	66	1
5	80	92	1
6	24	51	0
7	61	98	1
8	12	20	0
9	52	12	0

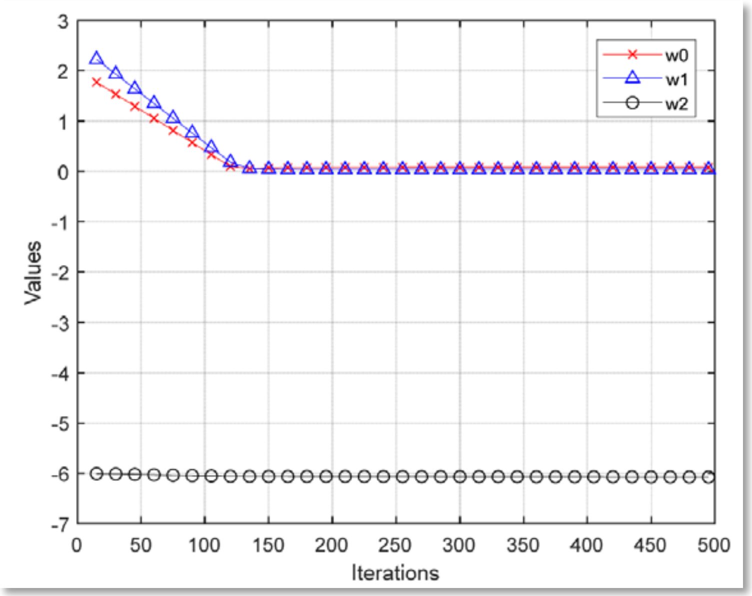


2입력 이진분류 모델의 설계 및 학습

- 경사하강법 적용
  - 초기값
    - $w_0[0] = 2, w_1[0] = 2.5, w_2[0] = -6$
  - 학습률
    - $\alpha = 0.001$

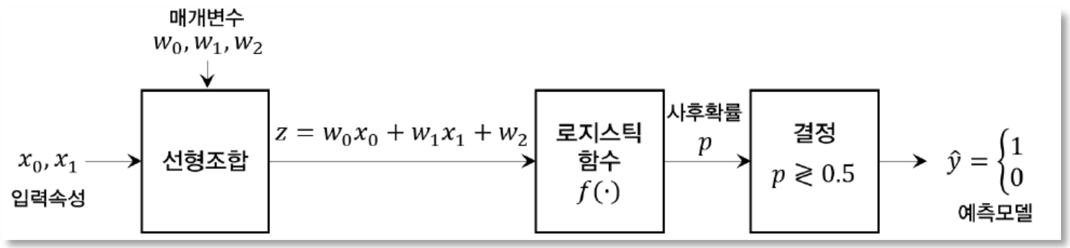


weight 변화



2입력 이진분류 모델의 설계 및 학습

- 결정경계 (Decision boundary)
  - 클래스 간의 경계
  - 2입력 이진분류 모델의 결정 경계

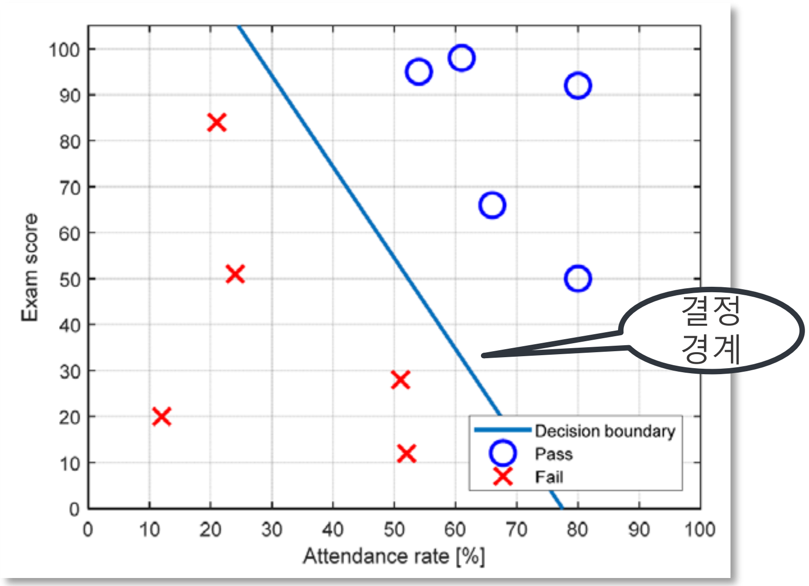


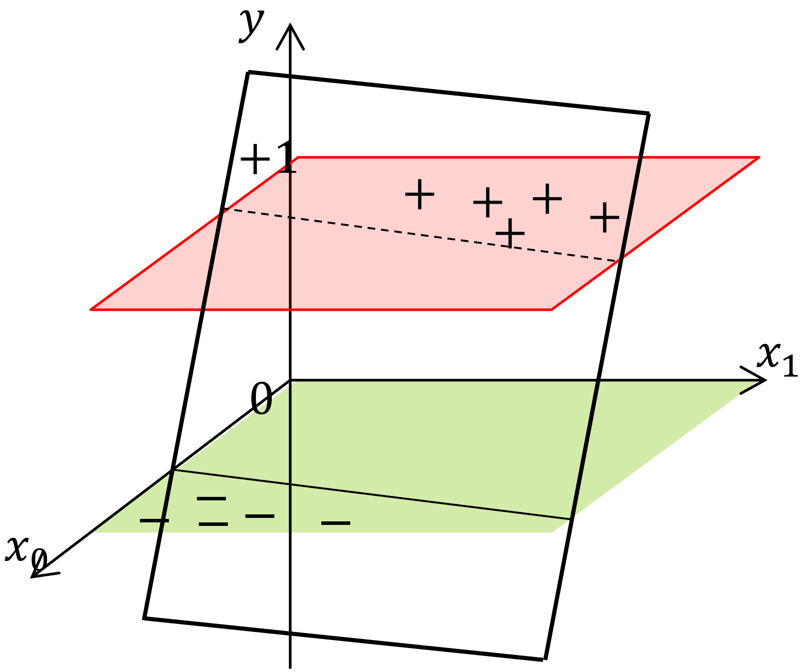
    - 확률 p=0.5를 기준으로 클래스를 결정
    - Sigmoid 함수(로지스틱 함수)을 통과한 값이 0.5를 갖는 입력은 z=0인 지점
    - 즉, 결정경계는 z=0인 지점에 형성

$$z = w_0x_0 + w_1x_1 + w_2 = 0 \rightarrow x_1 = -\frac{w_0}{w_1}x_0 - \frac{w_2}{w_1}$$

    - 입력 속성 공간을 선형 분할
    - 비선형 분할이 필요한 데이터에 적용할 수 없음

2입력 이진분류 모델의 설계 및 학습



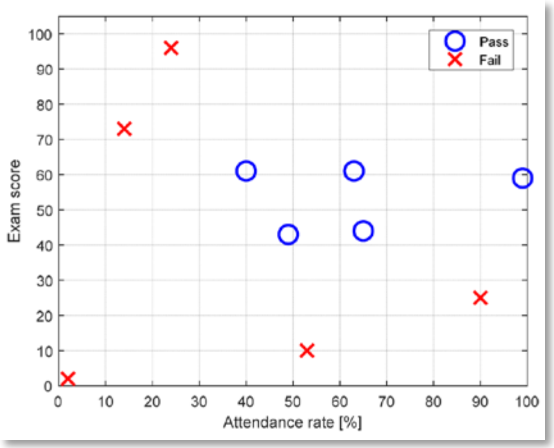


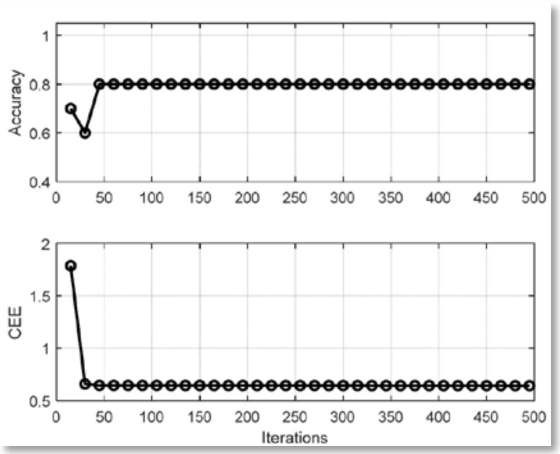
[한계점]

1) 직선(평면) 형태로만 공간을 분할 한다.
2) Binary Classification (0 or 1) 에만 적용가능

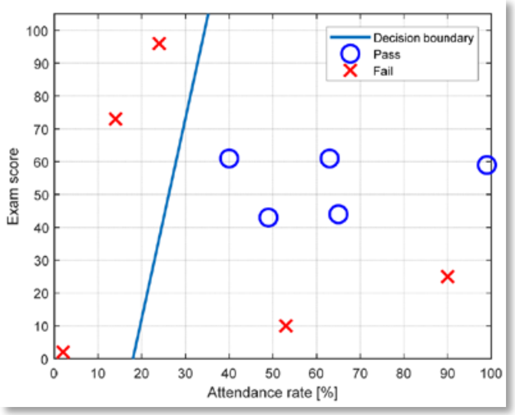
2입력 이진분류 모델의 설계 및 학습

- 속성 공간의 비선형 분할이 필요한 데이터
  - 로지스틱 회귀를 적용했을 때 (예)



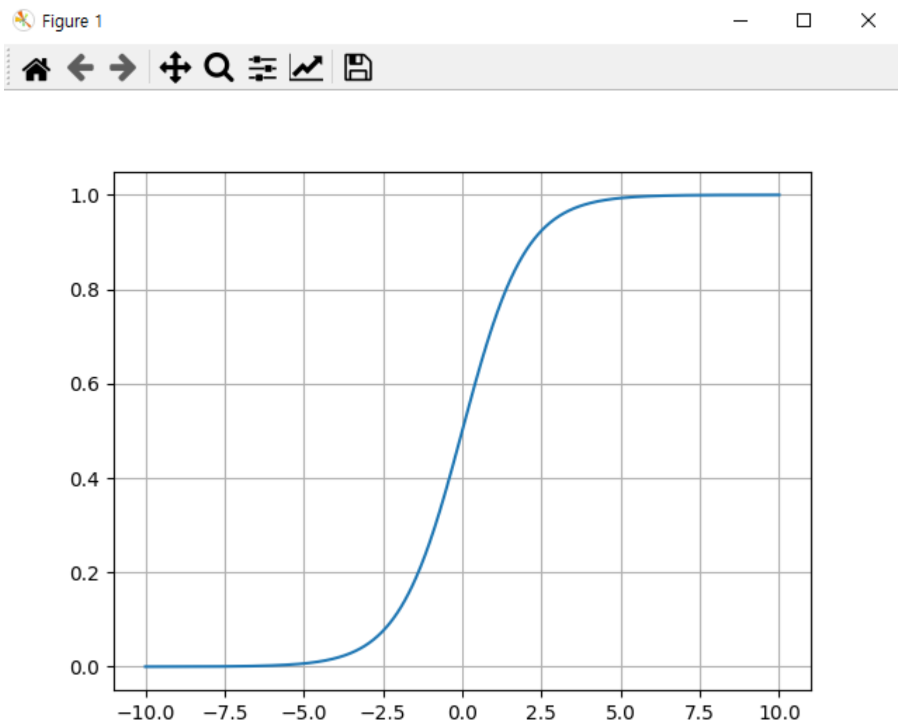


데이터의 속성을 충분히 학습하지 못함



실습 #1 (8주차)

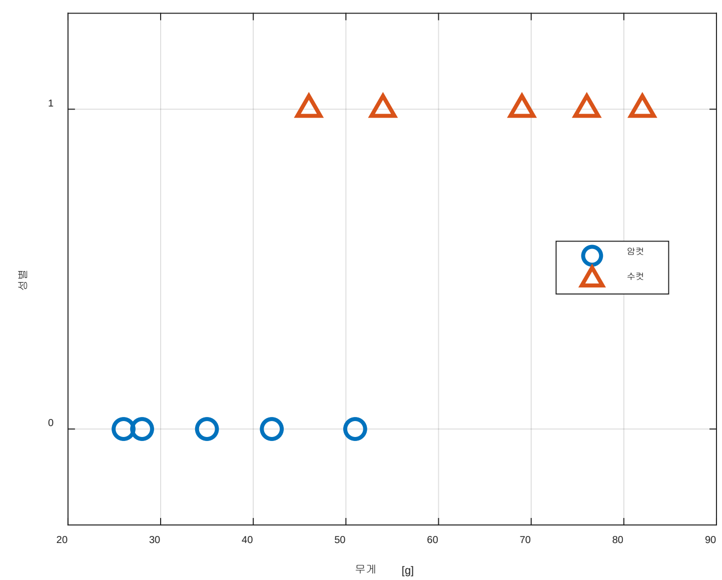
- 문제1) Logistic Regression에 사용되는 Sigmoid 함수를 plot 하시오.



실습 #1 (8주차)

- 사슴벌레 10마리 관찰(“binary_data_insect.csv”)
  - 무게와 성별 측정
  - 입력: 무게(g)   /   출력: 성별(0 = 암컷, 1 = 수컷)
  - 문제 2) 제공된 데이터 파일을 불러들여 x축은 사슴벌레의 무게(g), y축은 성별을 나타내는 2차원 평면에 각 데이터의 위치를 표시하시오.

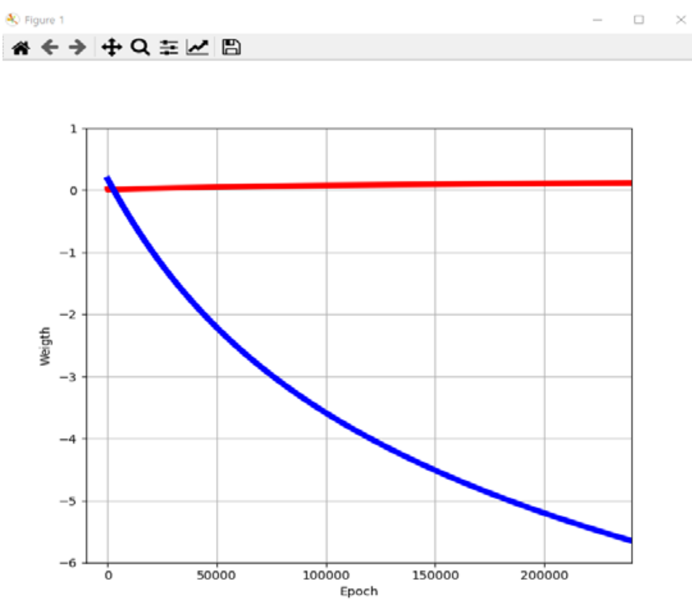
번호 (n)	무게 (x)	성별 (y)
0	26	0
1	28	0
2	35	0
3	42	0
4	51	0
5	46	1
6	54	1
7	69	1
8	76	1
9	82	1

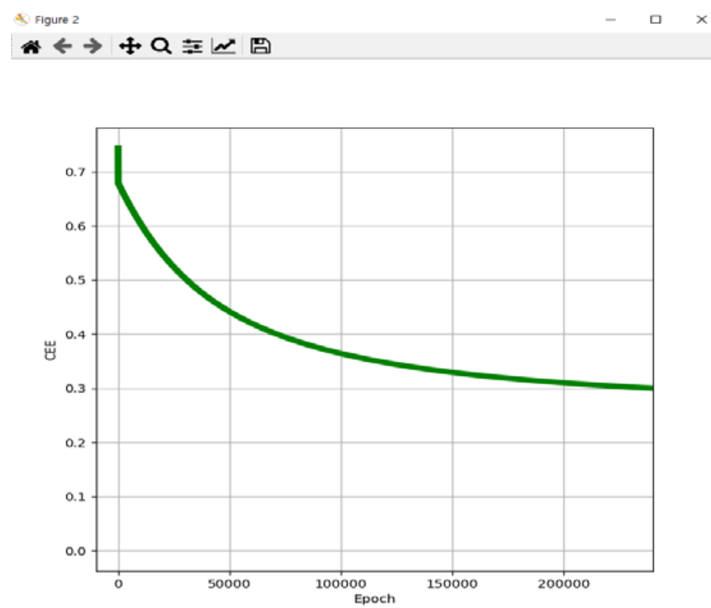


실습 #1 (8주차)

- 문제3) 사슴벌레 분류 문제를 Logistic Regression으로 판별하고자 한다. Cross Entropy Loss를 정의하고 Gradient Descent 알고리즘을 적용하여 가중치 w_0, w_1을 구하시오. 가중치의 변화와 Cross Entropy Loss의 변화과정을 plot 하시오.

```
epoch : 230000 ===========> W : [ 0.11509211 -5.54126925], cee : 0.3022621579238681
epoch : 240000 ===========> W : [ 0.11718469 -5.64566565], cee : 0.3000487512395725
GD 종료
w0 = 0.11723166996504109, w1 = -5.6480077099576045
```

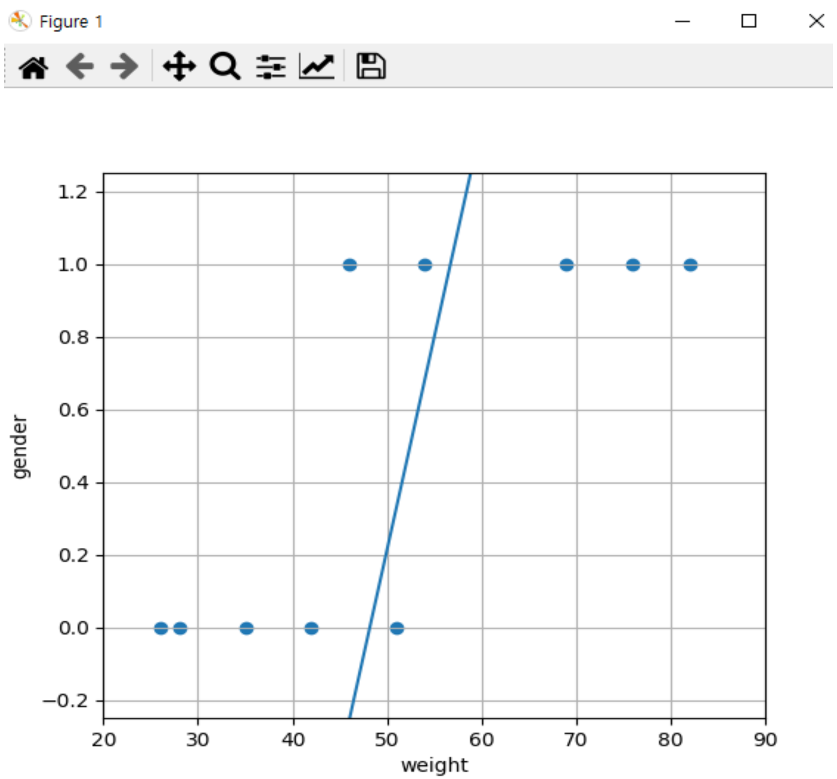




실습 #1 (8주차)

- 문제4) Logistic regression의 정확도(accuracy %)를 산출하고 Decision boundary를 그리시오. 또한, predict 함수를 만들어 임의의 값을 이용해 성별을 판별하시오.

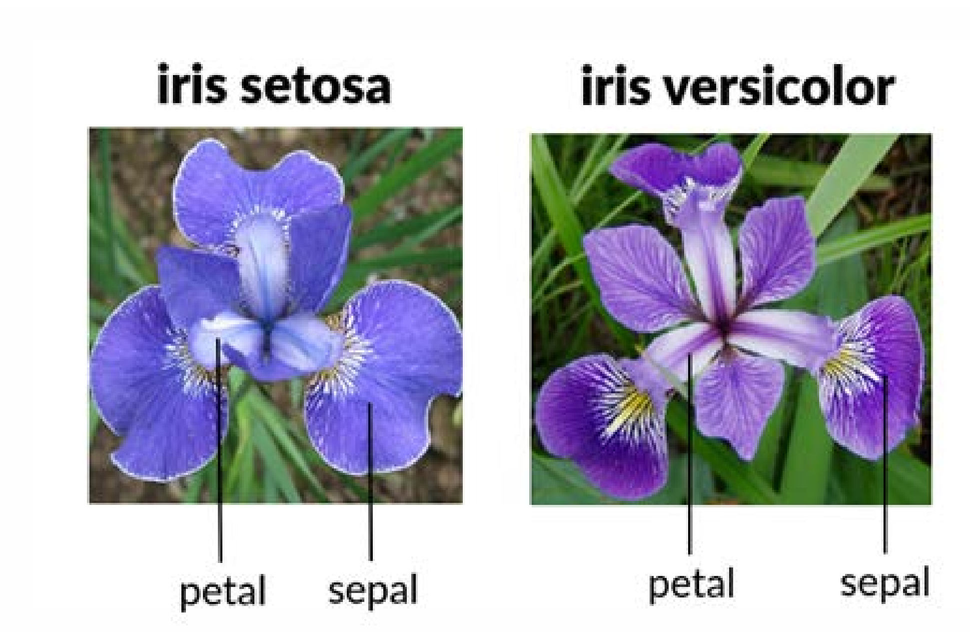
  ex) predict(20, 찾은 weight 값) = 0 또는 1을 판별

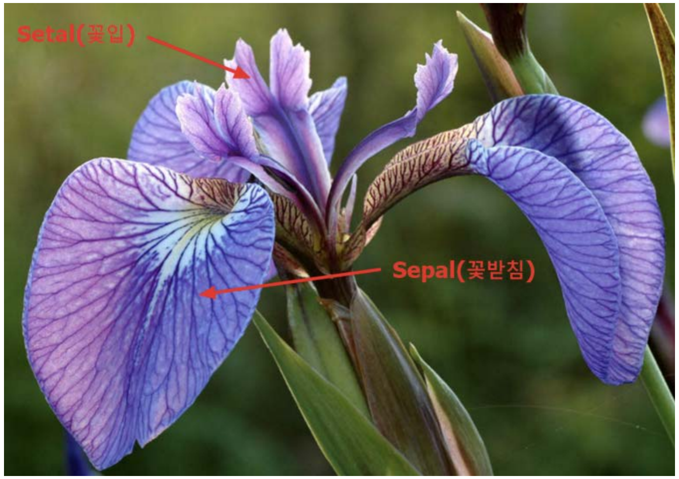



```
훈련결과, 정확도 80.0%
w0 = 0.11723162223765038, w1 = -5.6480053307224525
```

실습 #2 (9주차)

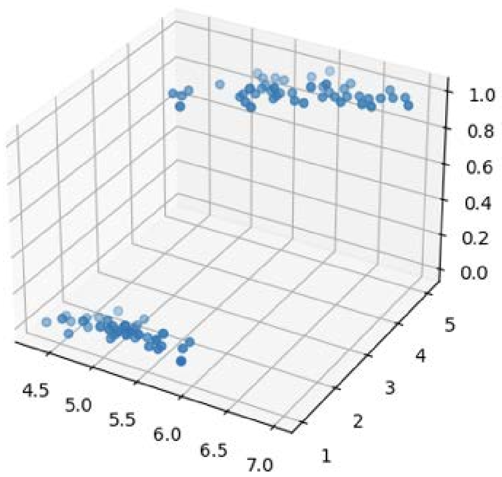
- "Iris.csv" 파일은 Iris(붓꽃)의 꽃잎 길이와 꽃받침 길이에 대해 두 가지의 Iris의 종류(Setosa, Versicolor)를 구분한 데이터





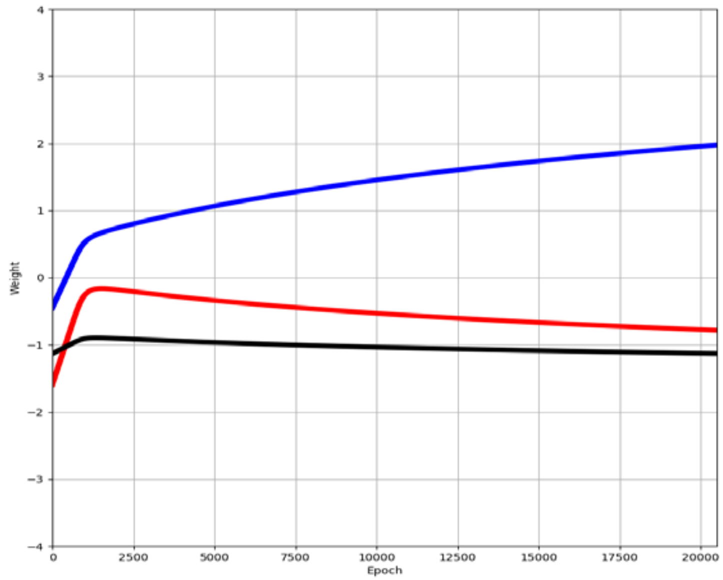
실습 #2 (9주차)

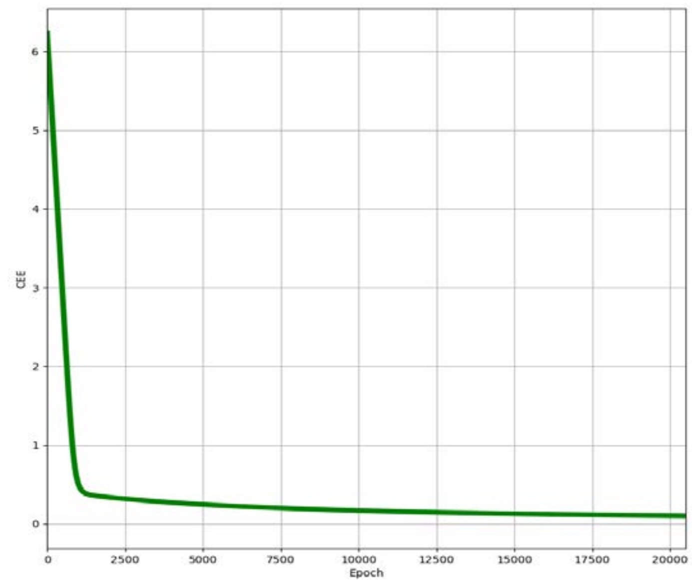
- "Iris.csv" 파일은 Iris(붓꽃)의 꽃잎 길이와 꽃받침 길이에 대해 두 가지의 Iris의 종류(Setosa, Versicolor)를 구분한 데이터
  - 문제1) 제공된 데이터 파일을 불러들여 3차원 공간에 표시하시오.
    - x축은 꽃받침 길이(cm), y축은 꽃잎 길이(cm), z축은 iris의 종류 (0 또는 1)



실습 #2 (9주차)

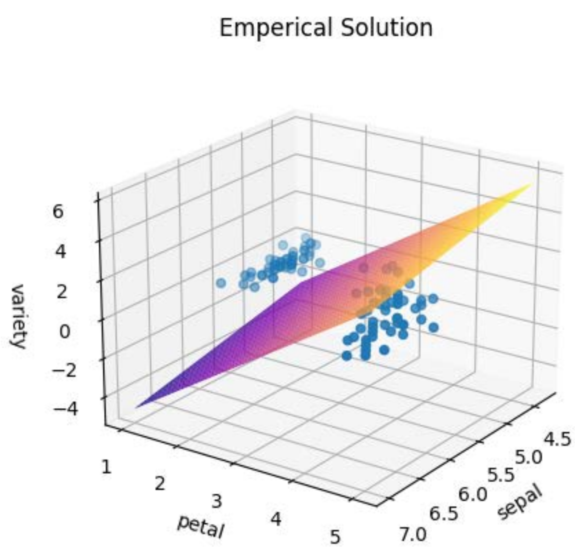
- 문제2) Cross entropy loss를 비용함수로 사용하는 경사하강법을 구현하고, 경사하강법의 반복 횟수에 따른 Cross entropy loss의 변화 및 매개변수의 변화를 그래프로 보이시오.
  - 초기값은 -3 ~ 3 사이의 랜덤값으로 설정할 것





실습 #2 (9주차)

- 문제3) 최적 매개변수의 값과 로지스틱 회귀 모델의 정확도를 산출하고, 3차원 공간에 Decision boundary를 그리시오. 또한, predict 함수를 만들어 임의의 값을 이용해 판별하시오.




```
훈련결과, 정확도 100.0%
w0 = -0.8071480165010575, w1 = 1.9843888524367115, w2 = -1.028333032360405
```