천재교육

머신러닝(Machine Learning) 이해와 실습

KDT Genia Academy 빅데이터 과정

머신러닝 이해와 실습 강의 일정

https://codingalzi.github.io/handson-ml2/
https://codingalzi.github.io/handson-ml3/intro.html

1일차 > 머신러닝

- > 머신러닝의 개요와 데이터 전처리
- > 교차검증과 데이터 샘플링 등 머신러닝 준비

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

회귀

01 회귀란?

02 경사하강법

머신
러닝

회귀

회귀

- 회귀는 현대 통계학을 이루는 큰 축
- 회귀 분석은 유전적 특성을 연구하던 영국의 통계학자 갈톤(Galton)이 수행한 연구에서 유래했다는 것이 일반론

"부모의 키가 크더라도 자식의 키가 대를 이어 무한정 커지지 않으며,
부모의 키가 작더라도 대를 이어 자식의 키가 무한정 작아지지 않는다."

- 회귀분석은 이처럼 데이터 값이 평균과 같은 일정한 값으로 돌아가려는 경향을 이용한 통계학 기법

회귀

회귀는 여러 개의 독립변수와 한 개의 종속변수 간의 상관관계를 모델링하는 기법을 통칭

Y	X1	X2	X3
학생의 수능성적은?	9월 모의고사 성적	중간고사 성적	인강 수강 시간

$$Y = W_1 * X_1 + W_2 * X_2 + W_3 * X_3 + \ldots + W_n * X_n$$

각 독립변수의 값에 영향을 미치는 회귀 계수(Regression coefficients)

머신러닝 회귀 예측의 핵심은 주어진 피쳐와 결정 값 데이터 기반에서 학습을 통해 최적의 회귀 계수를 찾아내는 것

회귀의 유형

종속변수와 독립변수간의 선형 관계를 모델링 하는 것이기 때문에 선형 회귀임. (다중선형회귀)

$$Y = W_1 * X_1 + W_2 * X_2 + W_3 * X_3 + \ldots + W_n * X_n$$

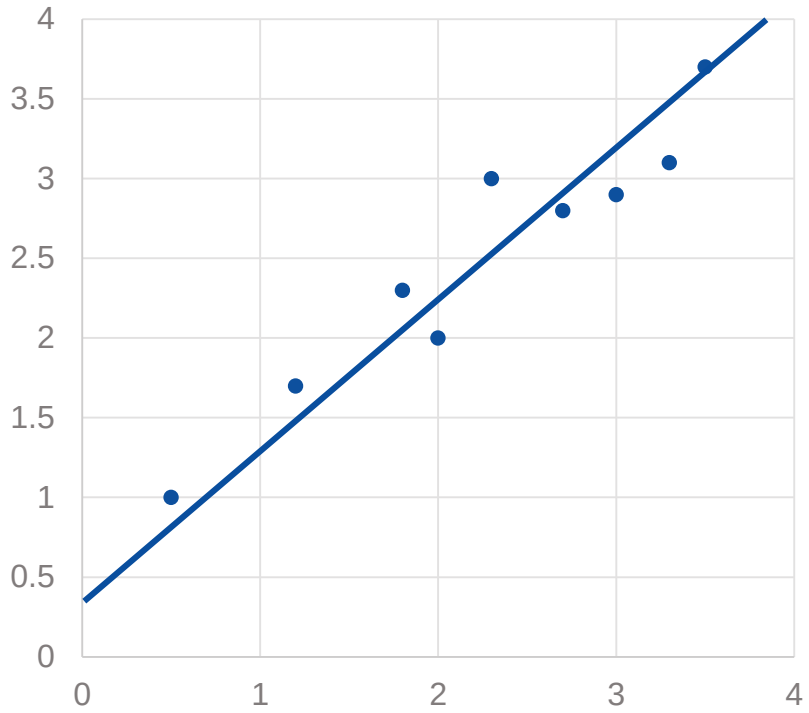
독립변수 개수	회귀 계수의 결합
1개 : 단일 회귀	선형 : 선형회귀
2개 이상 : 다중 회귀	비선형 : 비선형회귀

$$y = 3 + 2x$$

$$y = 3 + 2x_1 + 3x_2$$

$$y = 3 + 2x^{\,2}$$

단순선형회귀



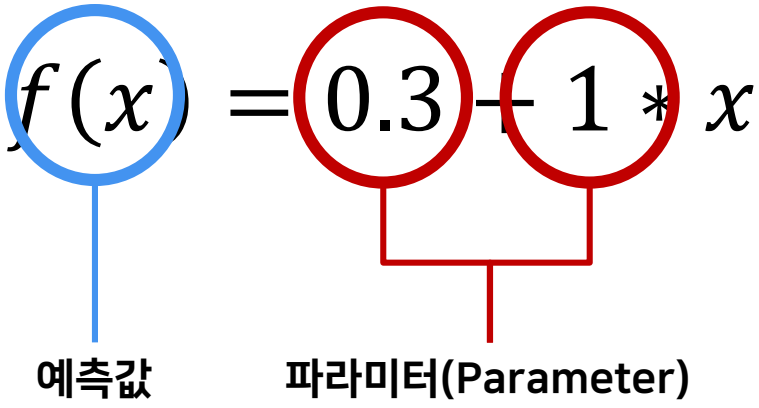
기울기 : 1
Y절편 : 0.3

$$f(x) = w0 + w1 * x$$

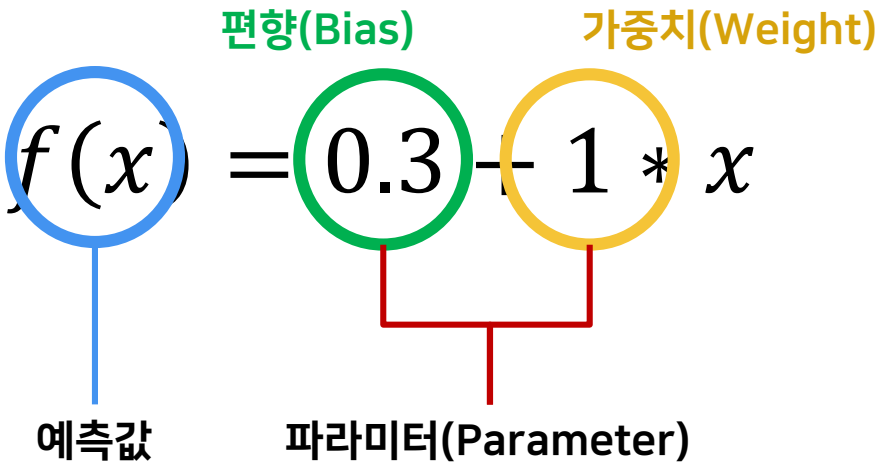
$$f(x) = 0.3 + 1 * x$$

최적의 회귀모델을 만든다는 것 : 전체 데이터의 잔차(오류값)합이 최소가 되는 모델을 만든다는 의미

선형 회귀



선형 회귀



$$\hat{y} = \theta_0 + \theta_1 \cdot x_1 + \cdots + \theta_n \cdot x_n$$

- $\hat{y}$: 예측값
- n: 특성 수
- x_i: 구역의 i 번째 특성값
- θ_0: 편향
- θ_j: j 번째 특성에 대한 (가중치) 파라미터(단, $1 \leq j \leq n$)

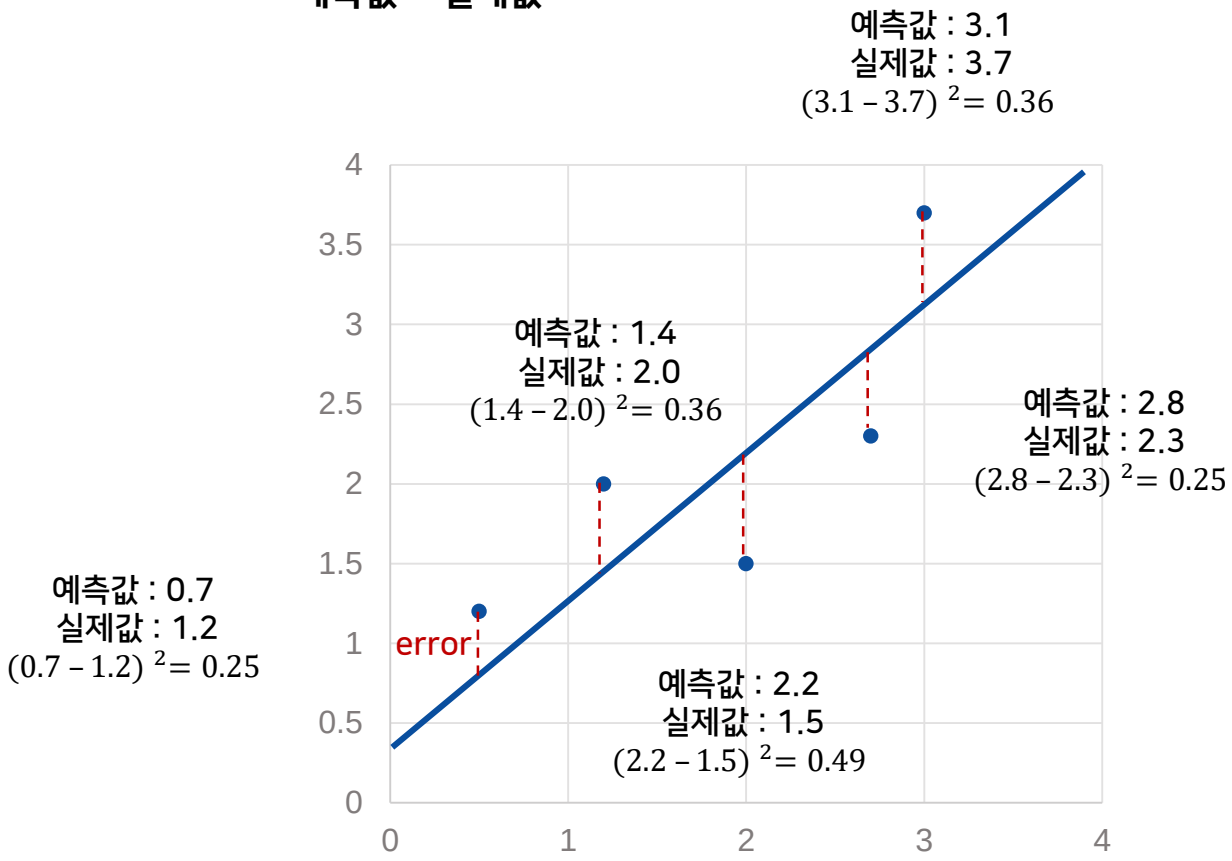
MSE(평균 제곱 오차)

$$\mathrm{MSE}(\theta) := \mathrm{MSE}(\mathbf{X}, h_\theta) = \frac{1}{m} \sum_{i=1}^{m} \left(\theta^T \mathbf{x}^{(i)} - y^{(i)} \right)^2$$

MSE(평균 제곱 오차)

$$\mathrm{MSE}(\theta) := \mathrm{MSE}(\mathbf{X}, h_\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(\theta^T \mathbf{x}^{(i)} - y^{(i)}\right)^2$$

예측값 실제값



$$\frac{(0.25 + 0.36 + 0.49 + 0.36 + 0.25)}{5} = \mathbf{0.342}$$

MSE(평균 제곱 오차)

MSE가 최소가 되도록 하는 파라미터를 찾는 것이 최종 목표!

HOW ?

A. **정규방정식 또는 특이값 분해 (SVD)**

B. **경사하강법**

정규 방정식 / SVD(특이값 분해)

정규방정식

$$\hat{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

SVD(특이값 분해)

$$\hat{\theta} = \mathbf{X}^{+} \mathbf{y}$$

정규 방정식

$$\hat{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\theta^T x^{(i)} - y^{(i)})^2$$

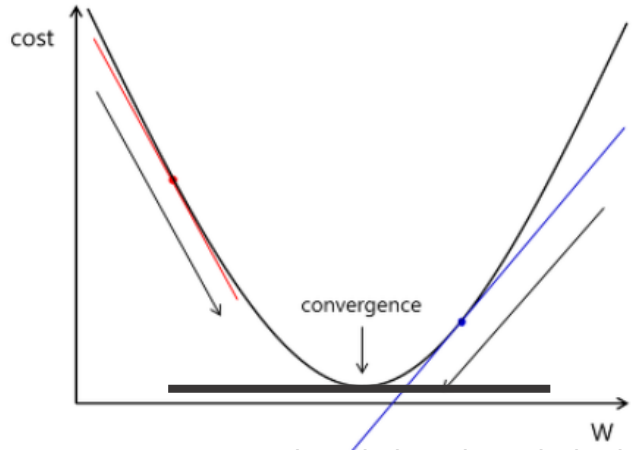
* MSE 에서 시그마를 제거하고, 행렬식으로 만들어보자.

$$MSE = \frac{1}{n}(x\hat{\theta} - y)^2$$

* $\hat{\theta}$ 에 대해 풀어본 해보자.

$$0 = \frac{2}{n} X^T (x\hat{\theta} - y) \rightarrow 0 = X^T x\hat{\theta} - X^T y$$

$$\hookrightarrow X^T x\hat{\theta} = X^T y \rightarrow \hat{\theta} = (X^T X)^{-1} X^T y$$



➔ 기울기가 0인 순간이 최소값

SVD(특이값 분해)

정규방정식 : $\hat{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$

유사역행렬을 통한 빠른 계산 : $\hat{\theta} = \mathbf{X}^+ \mathbf{y}$

$$\mathbf{y} = \mathbf{X}\,\hat{\theta}$$

$$\hat{\theta} = \mathbf{X}^- \mathbf{y}$$

선형대수의 특이값 분해

$$A = U\Sigma V^T$$

$$A = U \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \\ 0 & & \end{bmatrix} V^T \quad \rightarrow \quad A^+ = V \underbrace{\begin{bmatrix} 1/\sigma_1 & & & \\ & \ddots & & \\ & & 1/\sigma_k & 0 \end{bmatrix}}_{\Sigma^+} U^T$$

$$A^+ = V\Sigma^+ U^T$$

Linear Regression

Sklearn(사이킷런) 에서는 최적의 $\hat{\theta}$ 를 계산하는 LinearRegression 모델 제공

```
>>> import numpy as np
>>> from sklearn.linear_model import LinearRegression
>>> X = np.array([[1, 1], [1, 2], [2, 2], [2, 3]])
>>> # y = 1 * x_0 + 2 * x_1 + 3
>>> y = np.dot(X, np.array([1, 2])) + 3
>>> reg = LinearRegression().fit(X, y)
```

경사하강법

training set을 이용한 학습 과정 중, 가중치 파라미터를 조금씩 반복적으로 조정

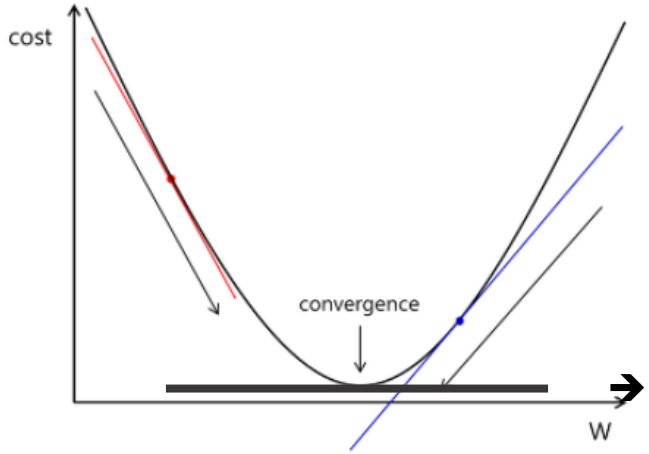
이때 비용함수의 크기를 줄이는 방향으로 조정

Cost function(비용함수)

- MSE 처럼 모델이 얼마나 나쁜가를 측정하는 함수.

$$\mathrm{MSE}(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(\theta^T \mathbf{x}^{(i)} - y^{(i)}\right)^2$$

▷선형회귀 모델의 파라미터를 조정하는 과정을 예시로, MSE 비용함수를 줄여나가 보자!



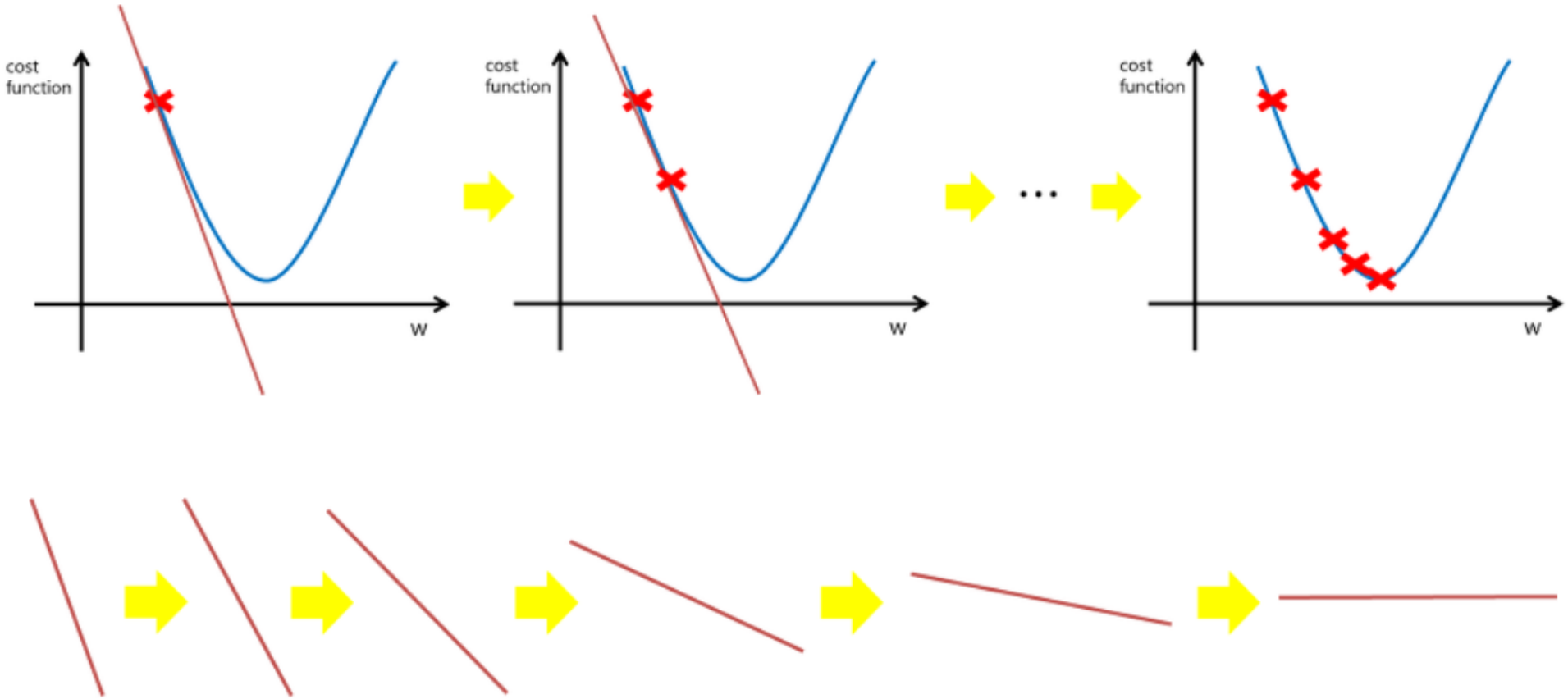
$$\nabla_\theta \mathrm{MSE}(\theta) = \frac{2}{m}\mathbf{X}^T\left(\mathbf{X}\,\theta^T - \mathbf{y}\right)$$

→ 기울기가 0인 순간이 최소값

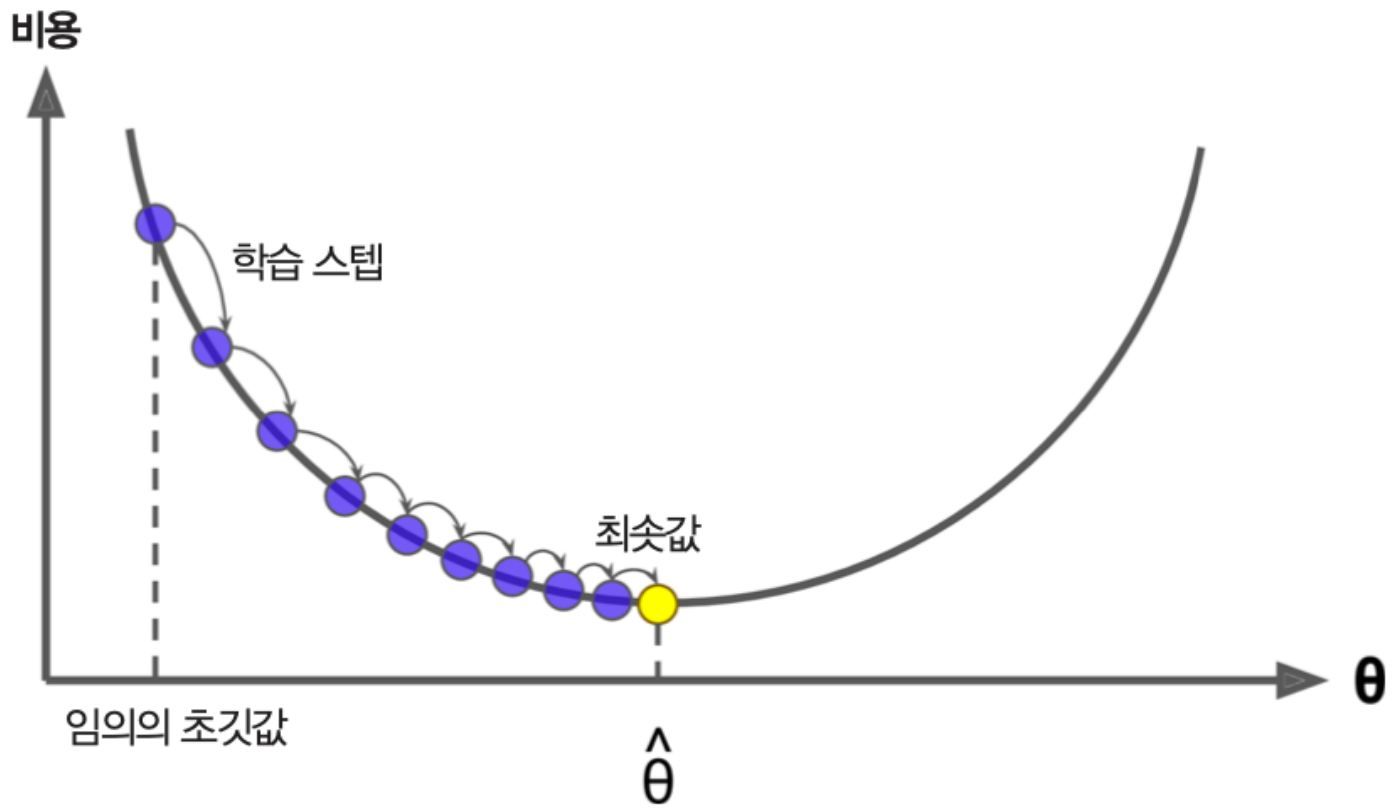
MSE는 **파라미터(θ) 값에 따른 오차(비용)**을 의미함

https://sdolnote.tistory.com/entry/Gradient

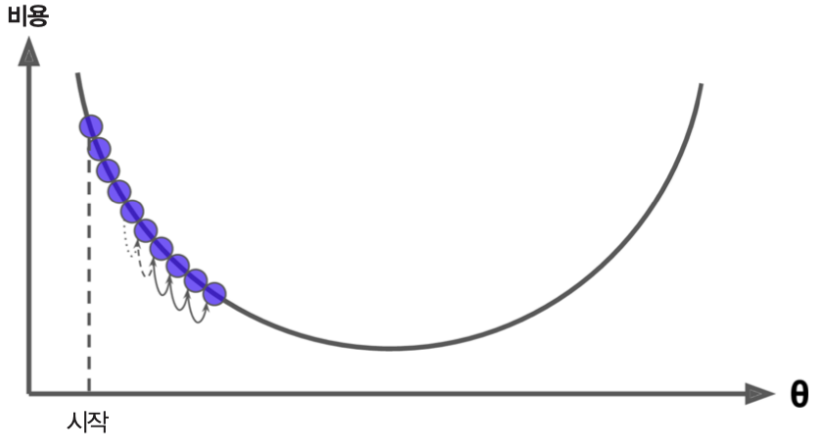
경사하강법

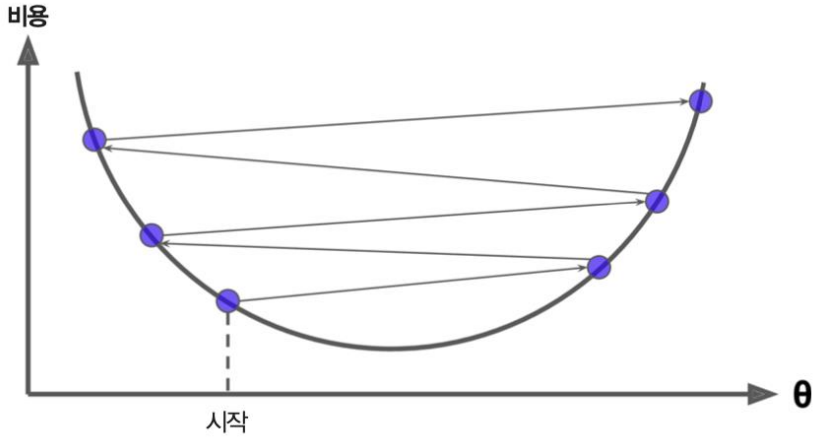


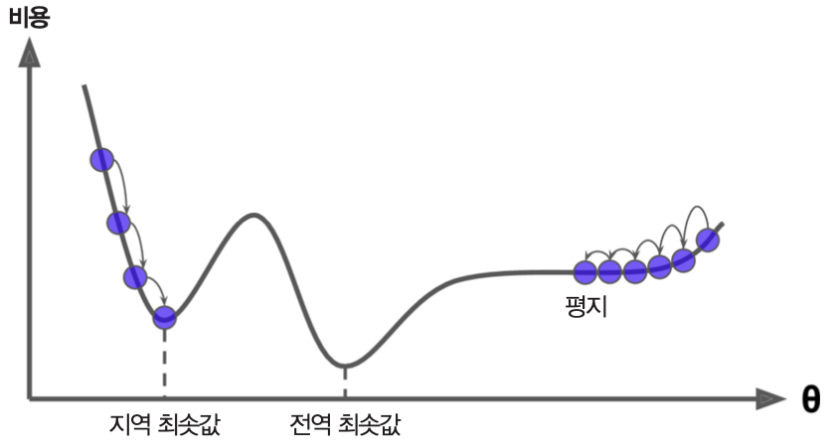
경사하강법의 진행과정



경사하강법의 주의사항







경사하강법의 종류

모델에 지정하는 배치 크기에 따라 경사하강법은 세가지로 나뉜다.

Batch Size 배치 크기란?

- 파라미터를 업데이트 하기 위해 사용되는 학습 데이터 개수

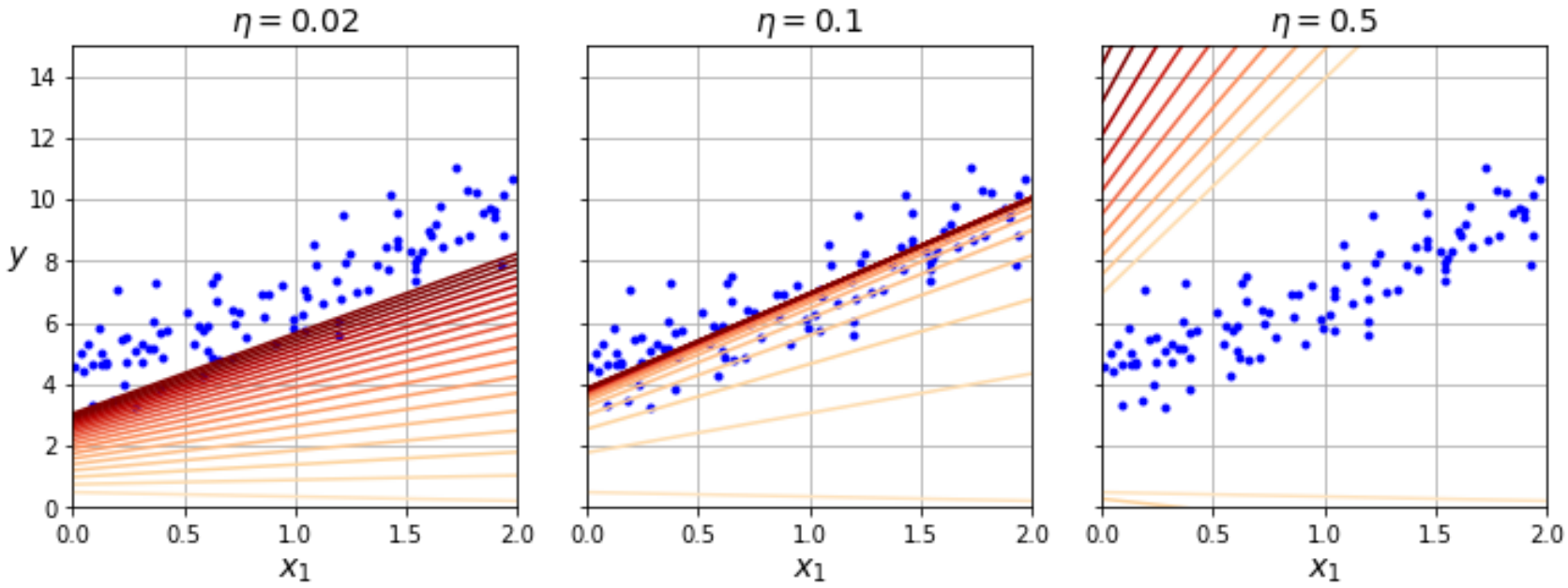
ex. 학습데이터가 1000개이고, 배치크기가 10이라면 총 100번의 스텝이 실행된다.

Q. 학습데이터가 20000개이고, 총 10번의 스텝이 실행되었다면 배치크기는 얼마?
Q. 학습데이터가 10000개이고, 배치크기가 300이라면 총 몇번의 스텝 진행?

배치 경사하강법

배치크기(Batch size)가 전체 학습데이터셋 크기와 같은, 즉 스텝이 1번 발생하는 경사하강법

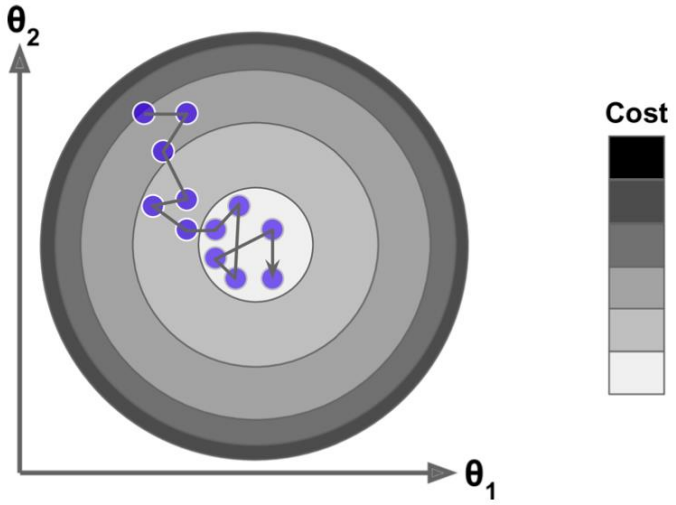
- 학습데이터 셋이 크면 그래디언트를 계산하는데에 많은 시간과 메모리가 필요하다는 단점 존재 → 사이킷런 지원x

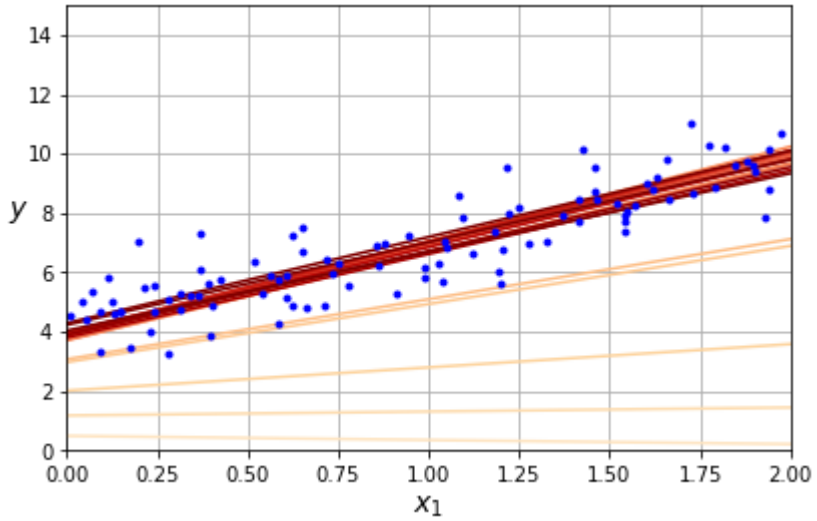


△ 학습률에 따른 수렴 속도를 나타낸 그래프

확률적 경사하강법(SGD)

한번의 스텝에 하나의 데이터에 대한 예측값을 실행한 후에 그 결과를 이용하여 그래디언트를 계산하고 파라미터를 조정한다.

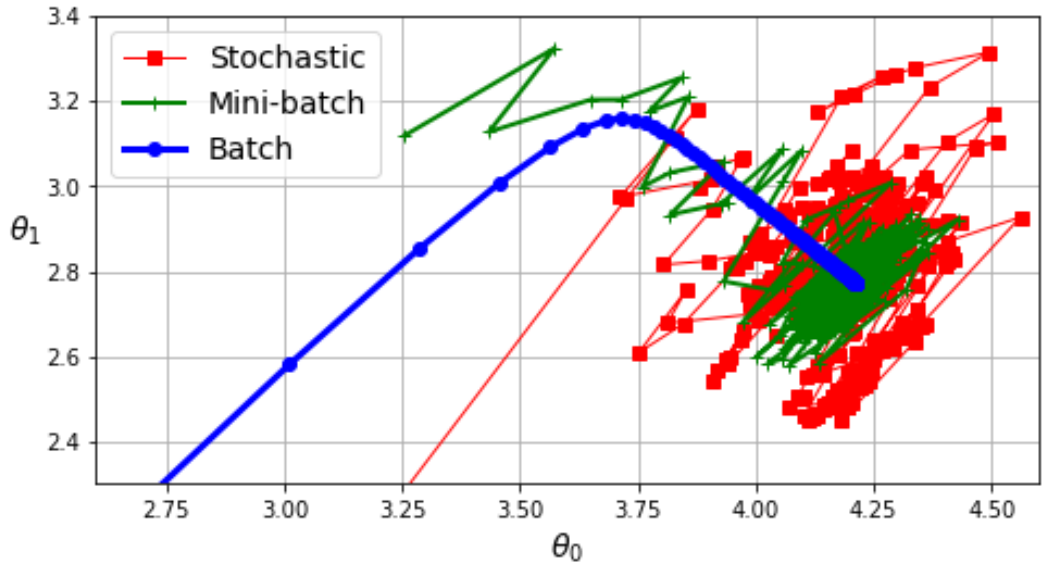




사이킷런 SGDRegressor 제공

미니 배치 경사하강법

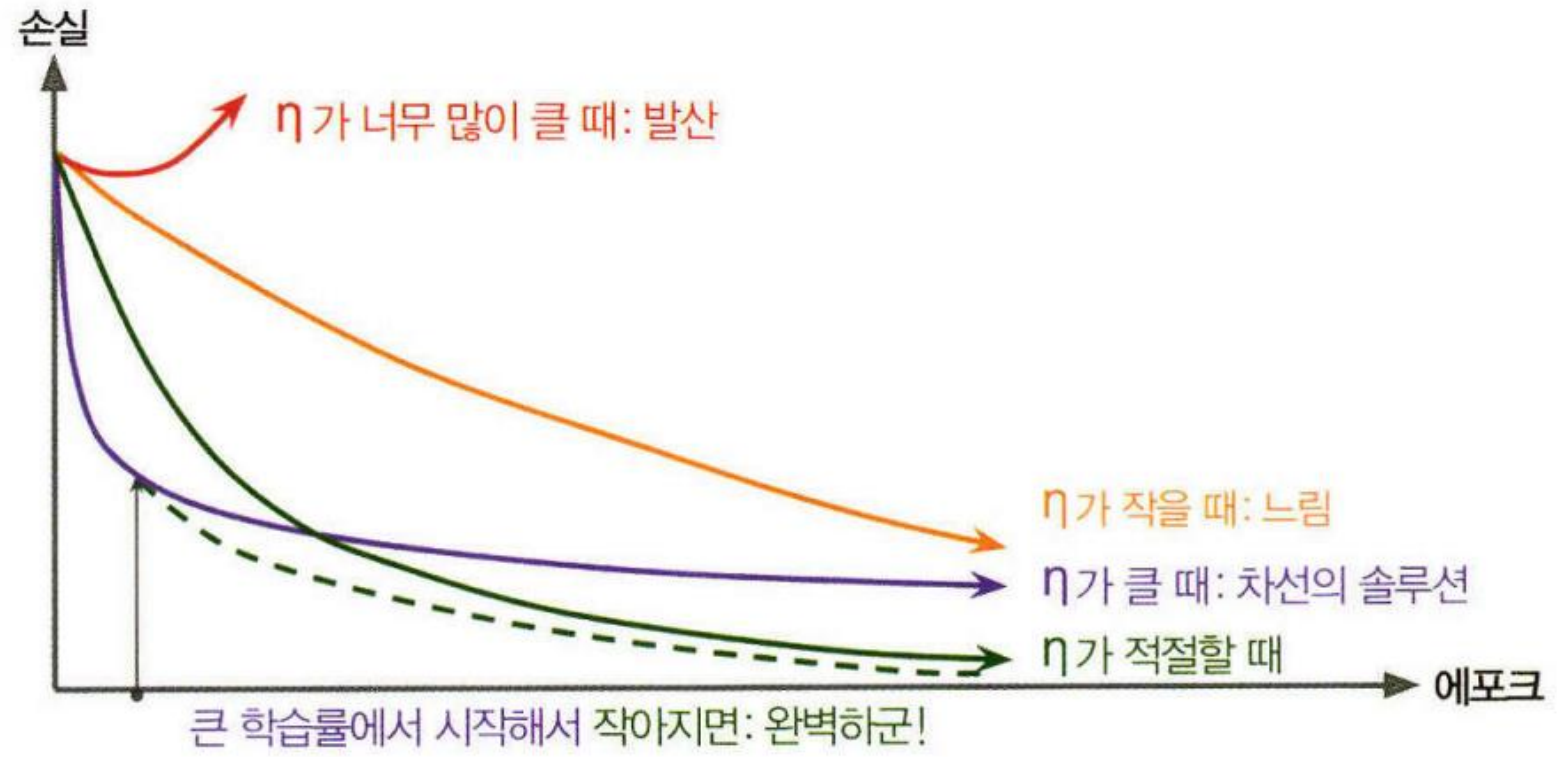
배치크기가 2부터 수백사이로 정해지고, 최적의 배치크기는 경우에 따라 모두 다르다.
배치크기를 어느정도 크게 하면 SGD 보다 파라미터의 움직임이 덜 불규칙적이 되며, 배치 경사하강법보다 빠르게 학습한다.
반면에 SGD에 비해 지역 최솟값에 수렴할 위험도가 보다 커진다.



Learning schedule

학습 스케줄? (Learning Schedule)

- 큰 학습률에서 시작하고 학습속도가 느려질때 학습률을 낮추면 최적의 고정 학습률보다 더 좋은 솔루션을 빨리 발견할 수 있다.



다항 회귀

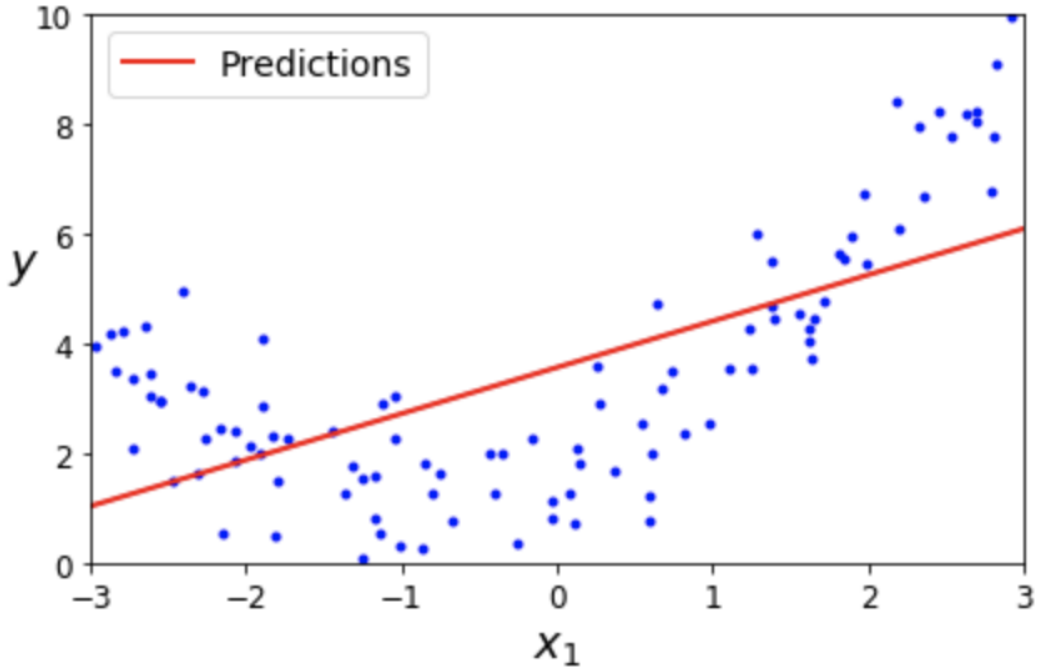
01 다항 회귀

02 학습 곡선
(과대적합, 과소적합)

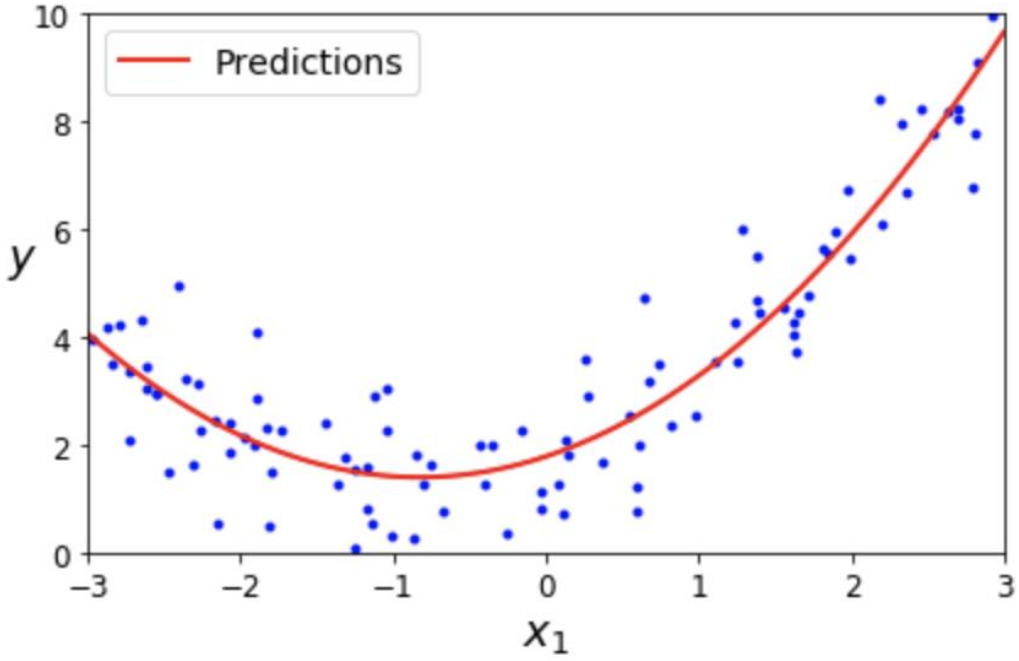
머신
러닝.

다항 회귀

비선형 데이터를 선형회귀를 이용하여 학습하는 기법



$$\hat{y} = \theta_0 + \theta_1 x_1$$

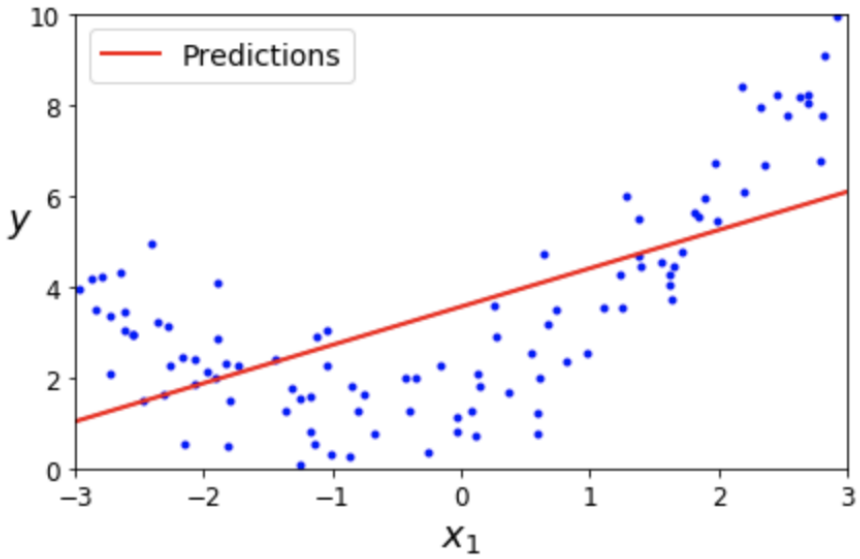


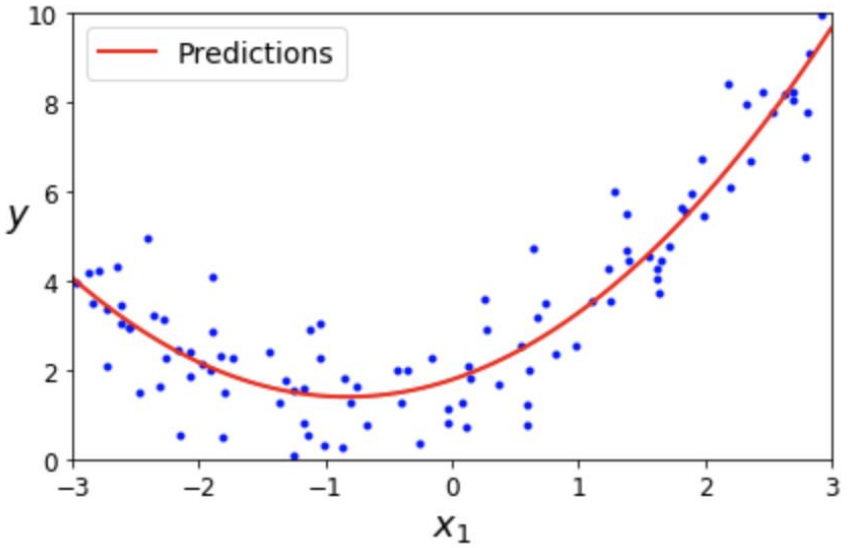
$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 {x_1}^2$$

다항 회귀

선형모델을 사용하여 비선형 데이터를 학습하는 방법
- 기존 데이터에 특성을 제곱하여 특성을 확장한다.

$$\theta = (\theta_0, \theta_1, \theta_2)$$





$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \quad (= {x_1}^2)$$

x_1	x_2
2	4
3	9
4	16

$$\hat{y} = \theta_0 + \theta_1 x + \theta_2 x^2$$

다항 회귀

$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2$ ⟹ x^2 : 2차 다항회귀 ???

두개의 항의 관계를 2차까지 모두 포함한 특성을 추가해야 함.

$$y = ax^2 + bx + c$$

$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2$

⇓ 2차 다항 회귀

$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n^2$

↳ x의 2차 결합

$x_1^2, x_1 \times x_2, x_2^2$

$\therefore \hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 \underbrace{x_1^2}_{x_3} + \theta_4 \underbrace{x_1 x_2}_{x_4} + \theta_5 \underbrace{x_2^2}_{x_5}$

다항 회귀

$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2$ ⟹ x^3 : 3차 다항회귀 ???

두개의 항의 관계를 3차까지 모두 포함한 특성을 추가해야 함.

$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2$

⇓ 3차 다항회귀

$$y = ax^3 + bx^2 + cx + d$$

$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_{n-1} x_{n-1}^2 + \theta_n x_n^3$

2차 3차 모두포함!

2차 : $x_1^2,\ x_1 x_2,\ x_2^2$

3차 : $x_1^3,\ x_1^2 x_2,\ x_1 x_2^2,\ x_2^3$

$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_1 x_2 + \theta_5 x_2^2$ (x_3, x_4, x_5)

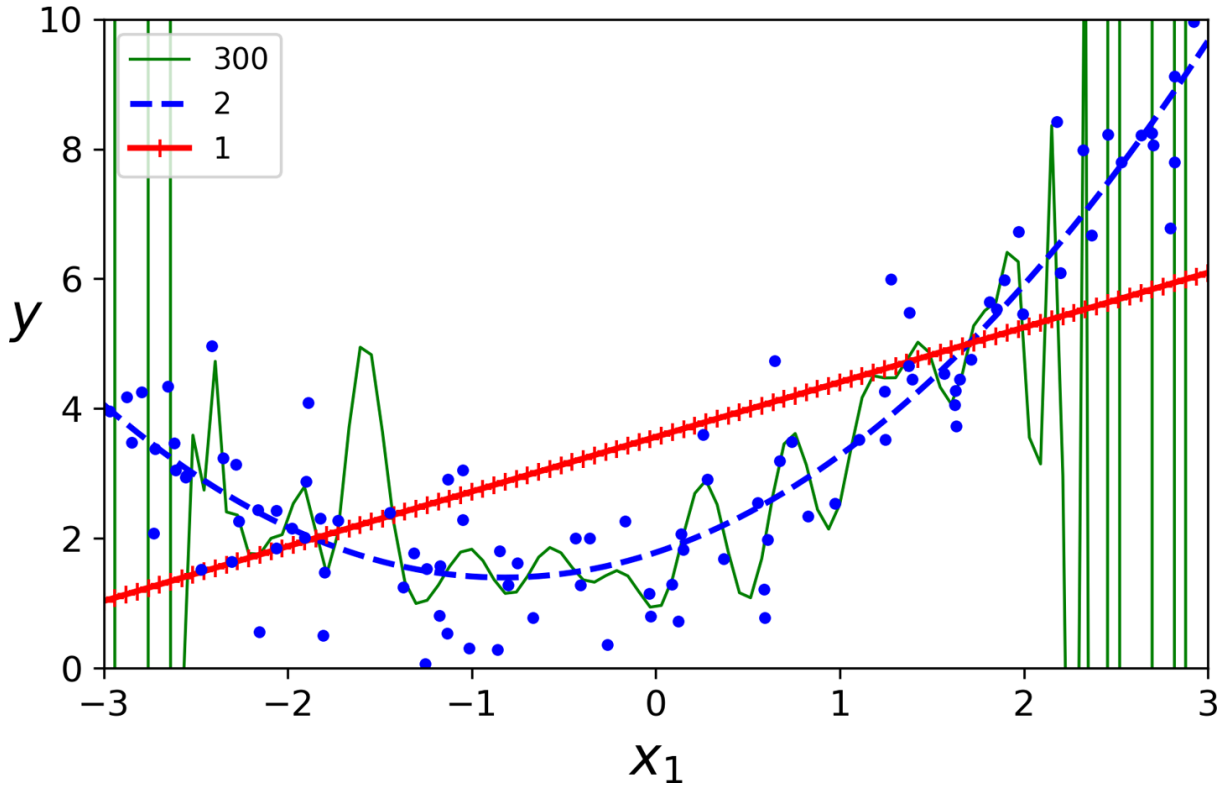
$+ \theta_6 x_1^3 + \theta_7 x_1^2 x_2 + \theta_8 x_1 x_2^2 + \theta_9 x_2^3$ (x_6, x_7, x_8, x_9)

$x_1^2,\ x_1 x_2,\ x_2^2,\ x_1^3,\ x_1^2 x_2,\ x_1 x_2^2,\ x_2^3$;

다항 회귀

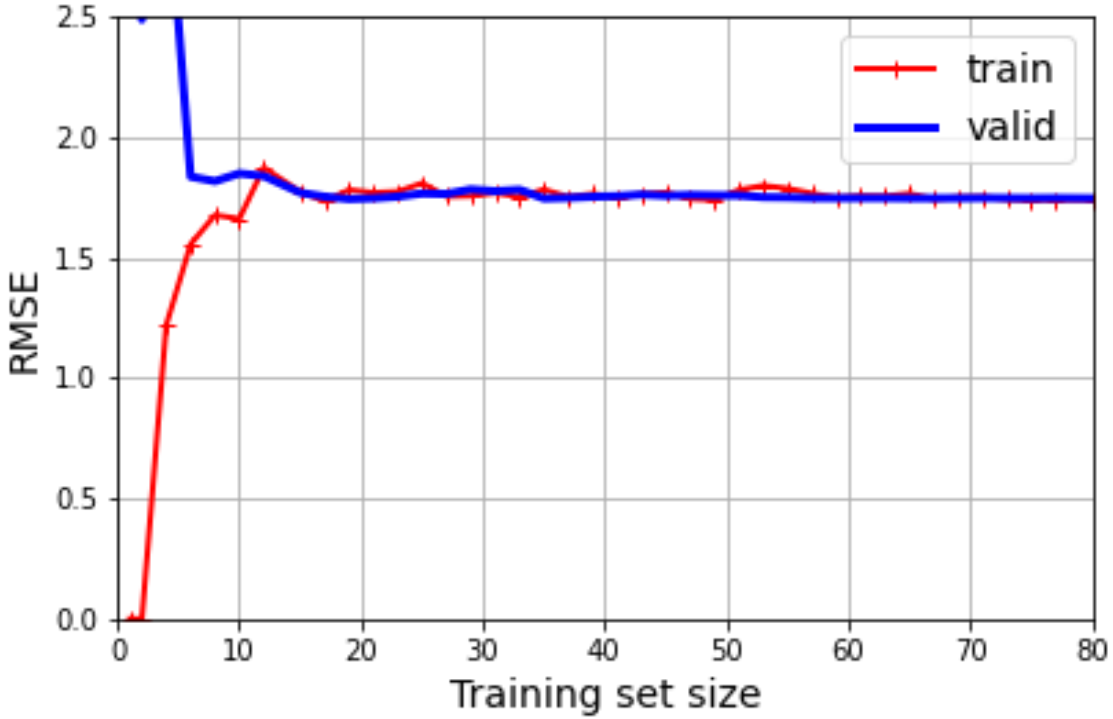
```
from sklearn.preprocessing import PolynomialFeatures

PolynomialFeatures(degree=d, include_bias=False)
```

학습 곡선(Learning Curve)

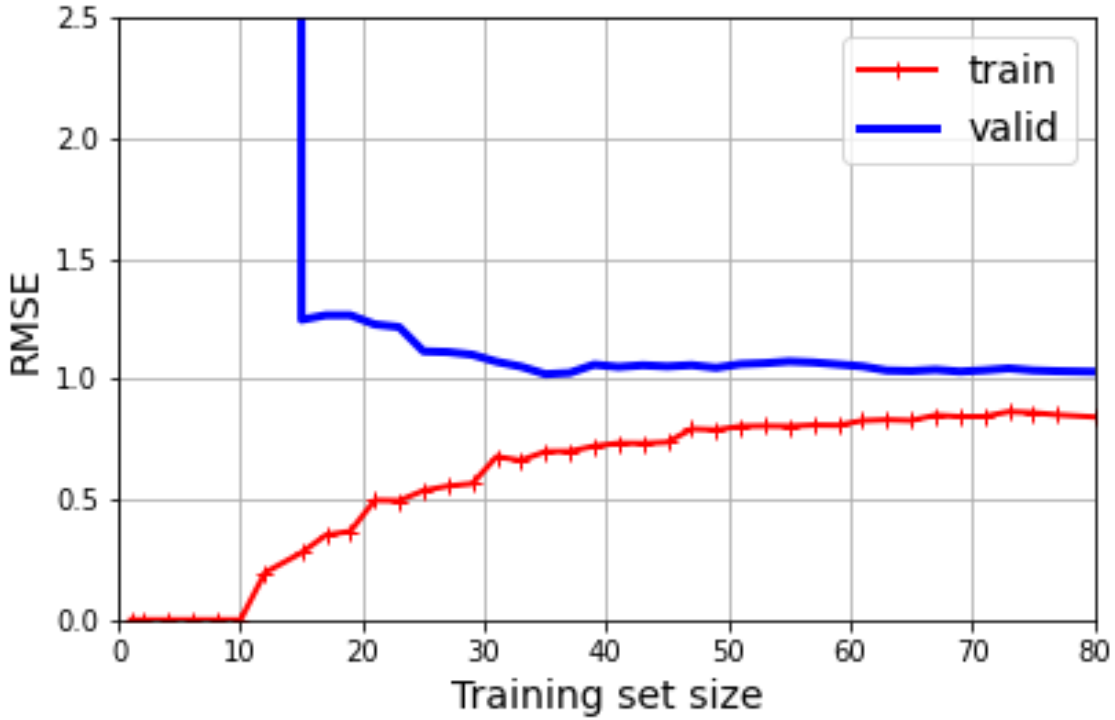


학습 곡선(Learning Curve) – 과소적합(Underfitting)

RMSE
- 평균 제곱근 오차
- MSE에 루트를 씌운 값



학습 곡선(Learning Curve) – 과대적합(Overfitting)



회귀

01 로지스틱 회귀

02 소프트맥스 회귀

03 실습

머신
러닝

로지스틱 회귀

회귀모델을 분류모델로 활용할 수 있다 ?!

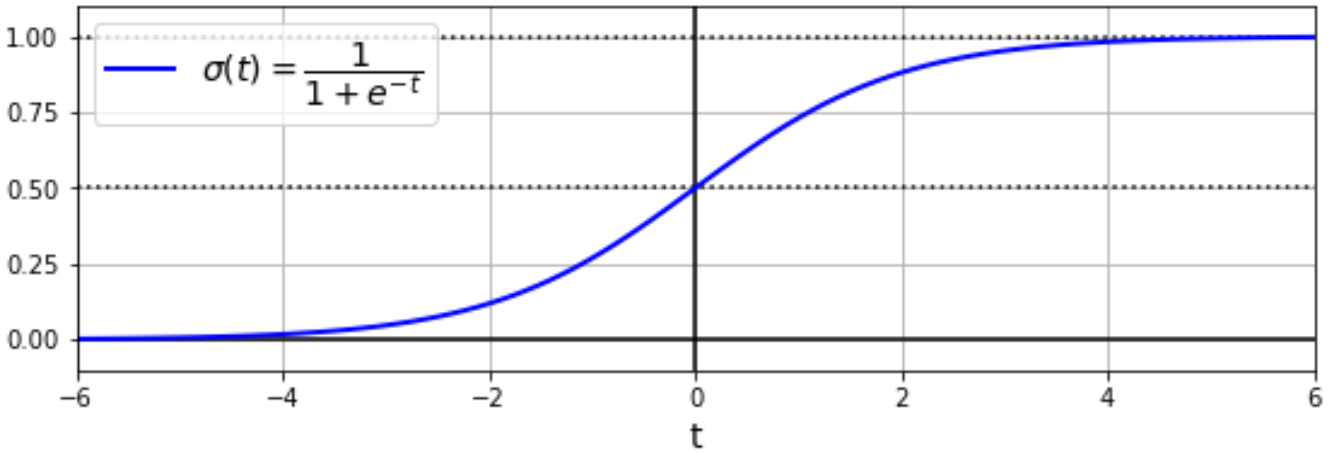
이진분류 : 로지스틱 회귀 (Logistic Regression)

다중 클래스 분류 : 소프트맥스 회귀 (Softmax Regression)

로지스틱 회귀

선형회귀 모델이 예측한 값에 시그모이드(Sigmoid)함수를 적용하여 0과 1사이의 값, 즉 양성일 확률 로 지정한다.

$$\hat{p} = h_\theta(\mathbf{x}) = \sigma(\theta^T \mathbf{x}) = \sigma(\theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n)$$



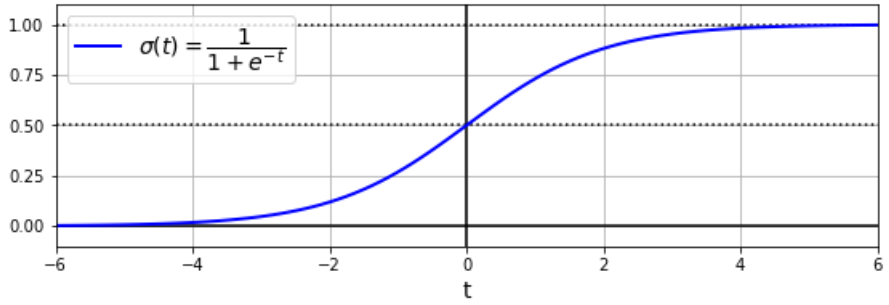
$$\hat{y} = \begin{cases} 0 & \text{if } \hat{p} < 0.5 \\ 1 & \text{if } \hat{p} \geq 0.5 \end{cases}$$

$$\sigma(t) = \frac{1}{1 + \exp(-t)}$$

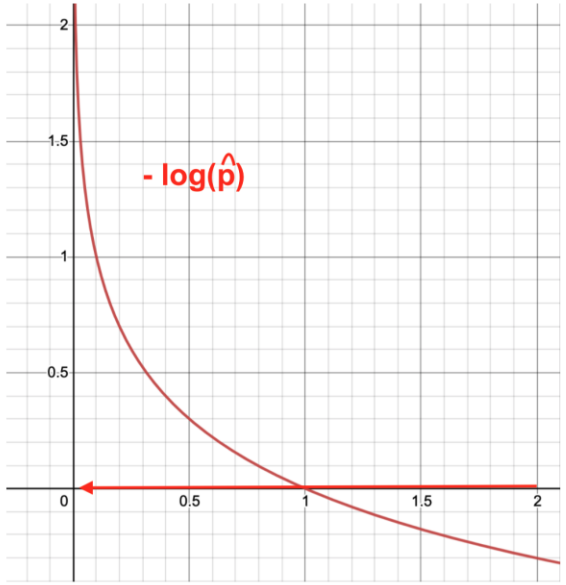
- 양성: $\theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n \geq 0$
- 음성: $\theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n < 0$

https://ywpop.tistory.com/7744

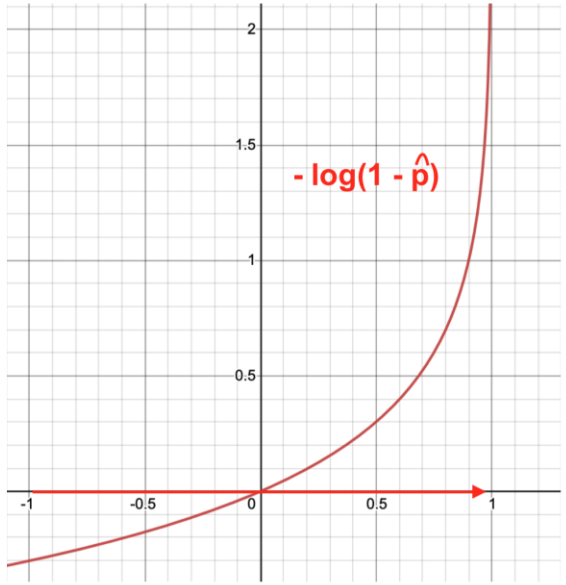
학습과 비용함수(Cost Function)



y는 1인데 $\hat{p}$는 0에 가까워지는 경우



y는 0인데 $\hat{p}$는 1에 가까워지는 경우



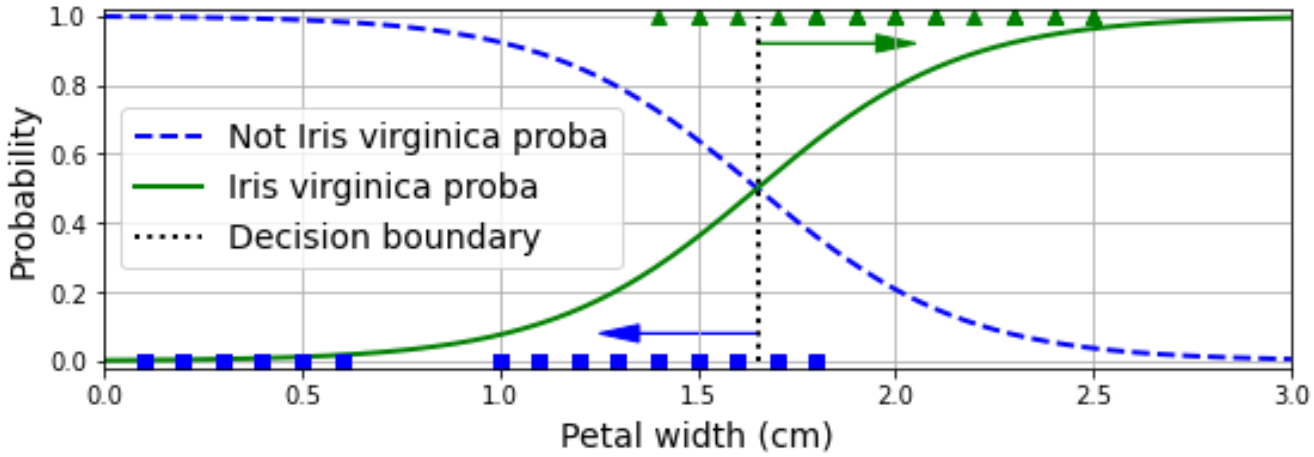
$$c(\theta) = \begin{cases} -\log(\hat{p}) & y = 1 \text{ 인 경우} \\ -\log(1-\hat{p}) & y = 0 \text{ 인 경우} \end{cases}$$

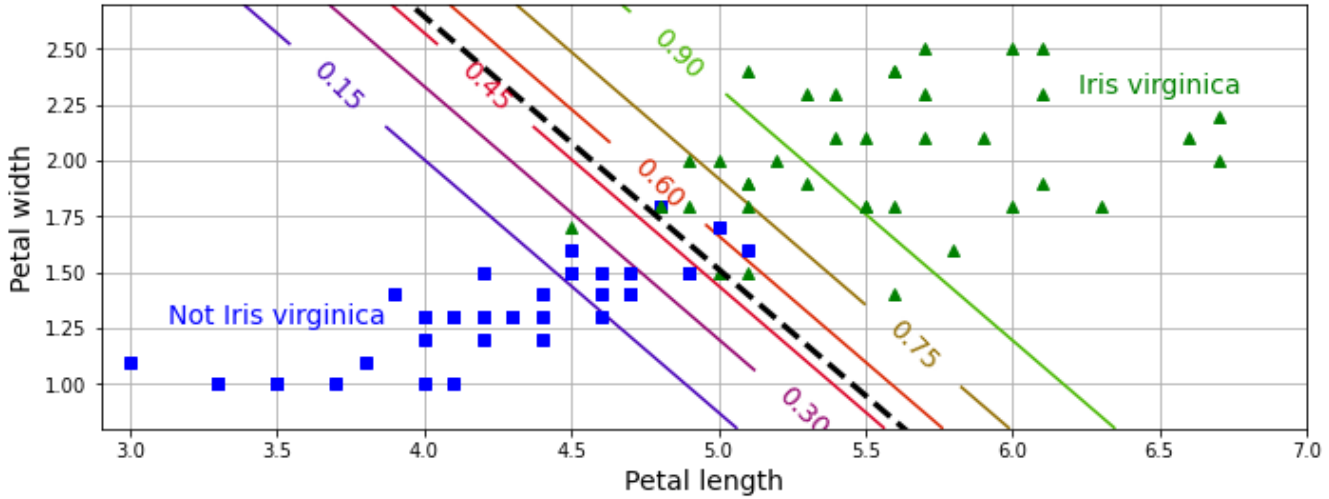
학습과 비용함수(Cost Function)

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\ \log(\hat{p}^{(i)}) + (1-y^{(i)})\ \log(1-\hat{p}^{(i)})]$$

정답을 더 높은 확률로 예측할 수록 좋은 모델이라고 평가하는 것

로지스틱 회귀

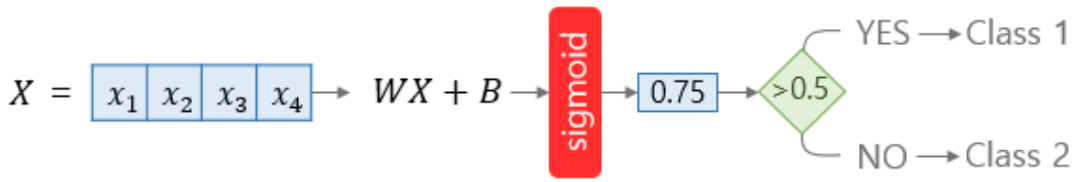

Not Iris virginica proba
Iris virginica proba
Decision boundary
Probability
Petal width (cm)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0


0.15
0.45
0.90
0.60
0.75
0.30
Iris virginica
Not Iris virginica
Petal width
Petal length
2.50
2.25
2.00
1.75
1.50
1.25
1.00
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0

소프트맥스 회귀

소프트맥스 회귀(Softmax Regression)

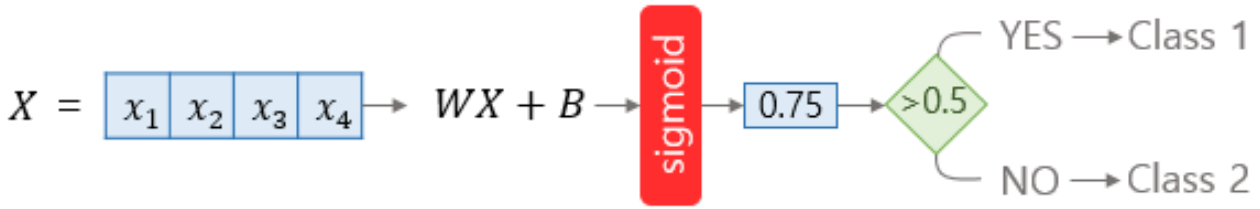
- 로지스틱 회귀모델을 일반화하여 다중 클래스 분류를 지원하도록 만든 모델



로지스틱 회귀 : 예측값이 0과 1사이의 값

소프트맥스 회귀 : 확률의 총 합이 1이 되는 다중 클래스 분류

소프트맥스 회귀



클래스 1인가, 클래스 2인가?

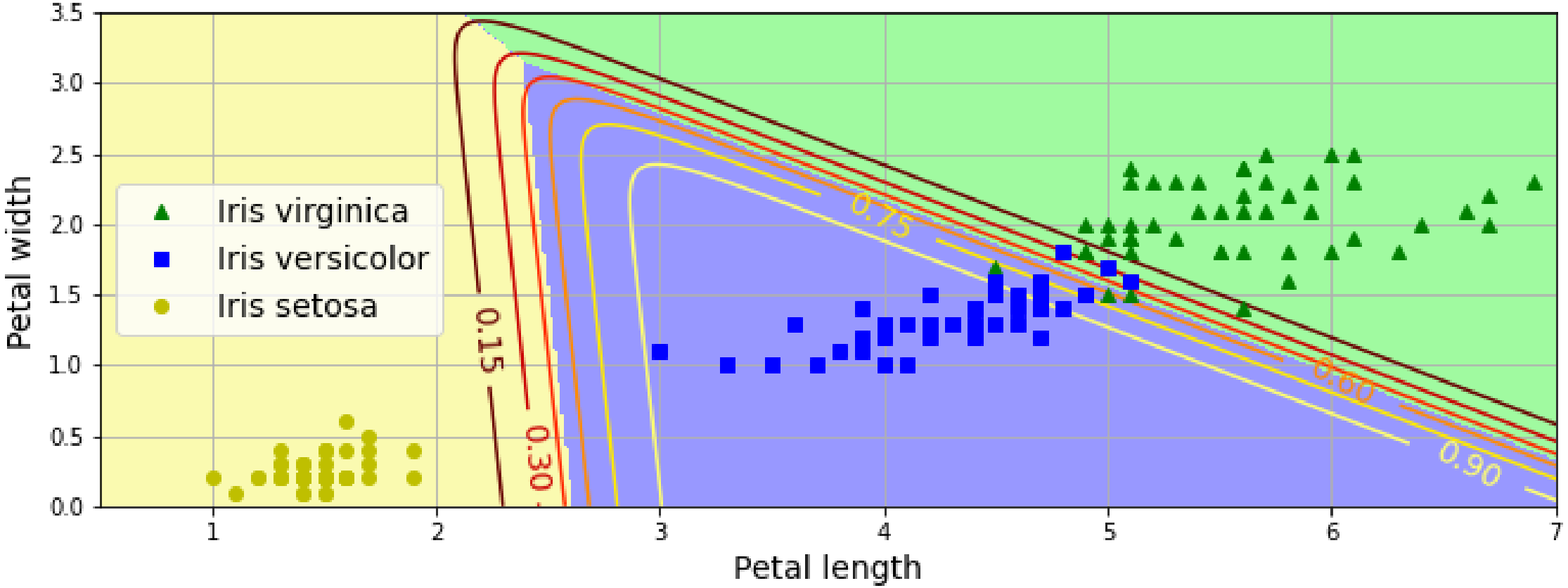
클래스 1인 확률
클래스 2인 확률
클래스 3인 확률

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \quad for\ i = 1, 2, \ldots k$$

$$softmax(z) = [\frac{e^{z_1}}{\sum_{j=1}^{3} e^{z_j}} \quad \frac{e^{z_2}}{\sum_{j=1}^{3} e^{z_j}} \quad \frac{e^{z_3}}{\sum_{j=1}^{3} e^{z_j}}] = [p_1, p_2, p_3] = \hat{y} = \text{예측값}$$

소프트맥스 함수는 클래스별 확률들의 총 합이 1이 되도록 바꾸어 주는 함수!

소프트맥스 회귀


Iris virginica
Iris versicolor
Iris setosa
Petal width
Petal length
0.15
0.30
0.75
0.60
0.90
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
1
2
3
4
5
6
7

[실습] 로지스틱 회귀를 적용한 밀크T 문항별 정오답 예측

	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

감사합니다