7장. 머신러닝

숙명여자대학교
SOOKMYUNG WOMEN'S UNIVERSITY

Contents

I. 머신러닝의 유형

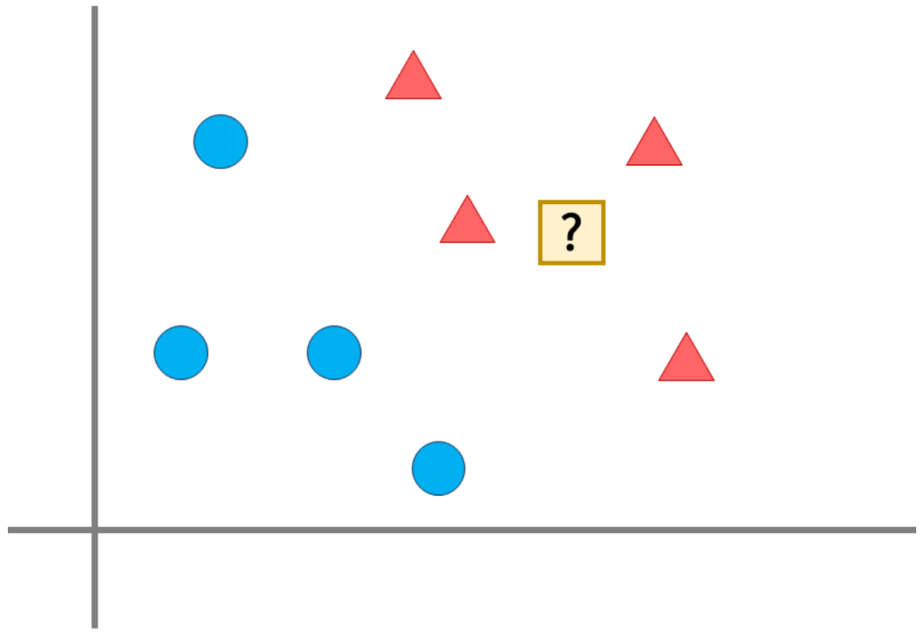
II. 머신러닝의 과정

III. 최근접 이웃 분류기 (The Nearest Neighbor Classifier)

IV. 회귀(Regression)

최근접 이웃법 (1/5)

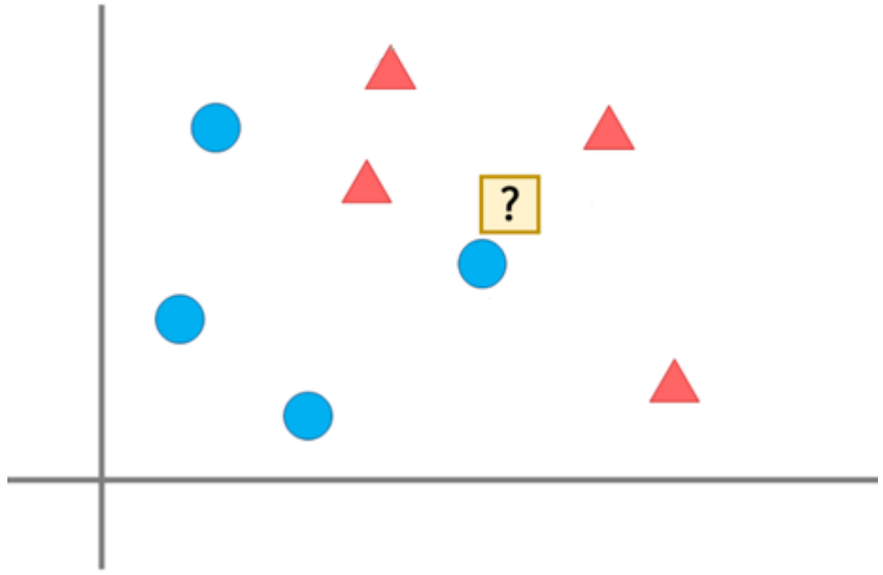
- 새로운 데이터를 입력 받았을 때 가장 가까이 있는 것이 무엇이냐를 중심으로 새로운 데이터를 분류

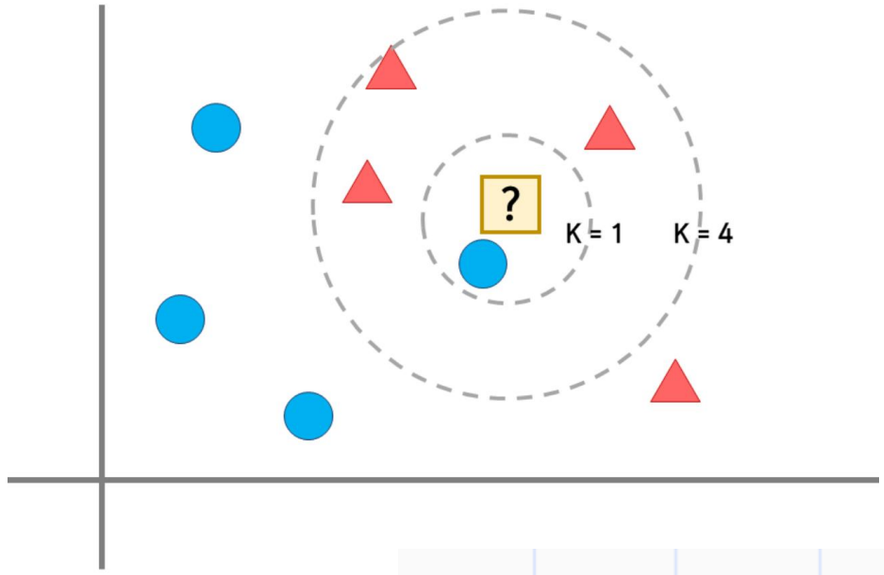


물음표 부분에 들어갈 것은 ? 세모

최근접 이웃법 (2/5)

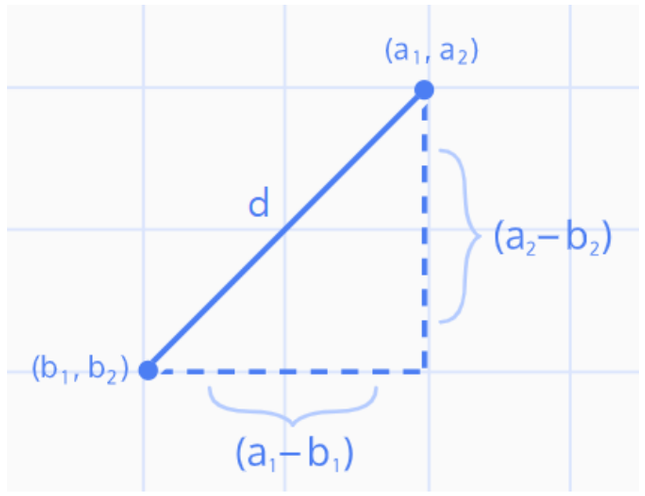
- “가장 가까이”라는 기준 ?





- 유클리드 거리(Euclidean distance)

$$d = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$

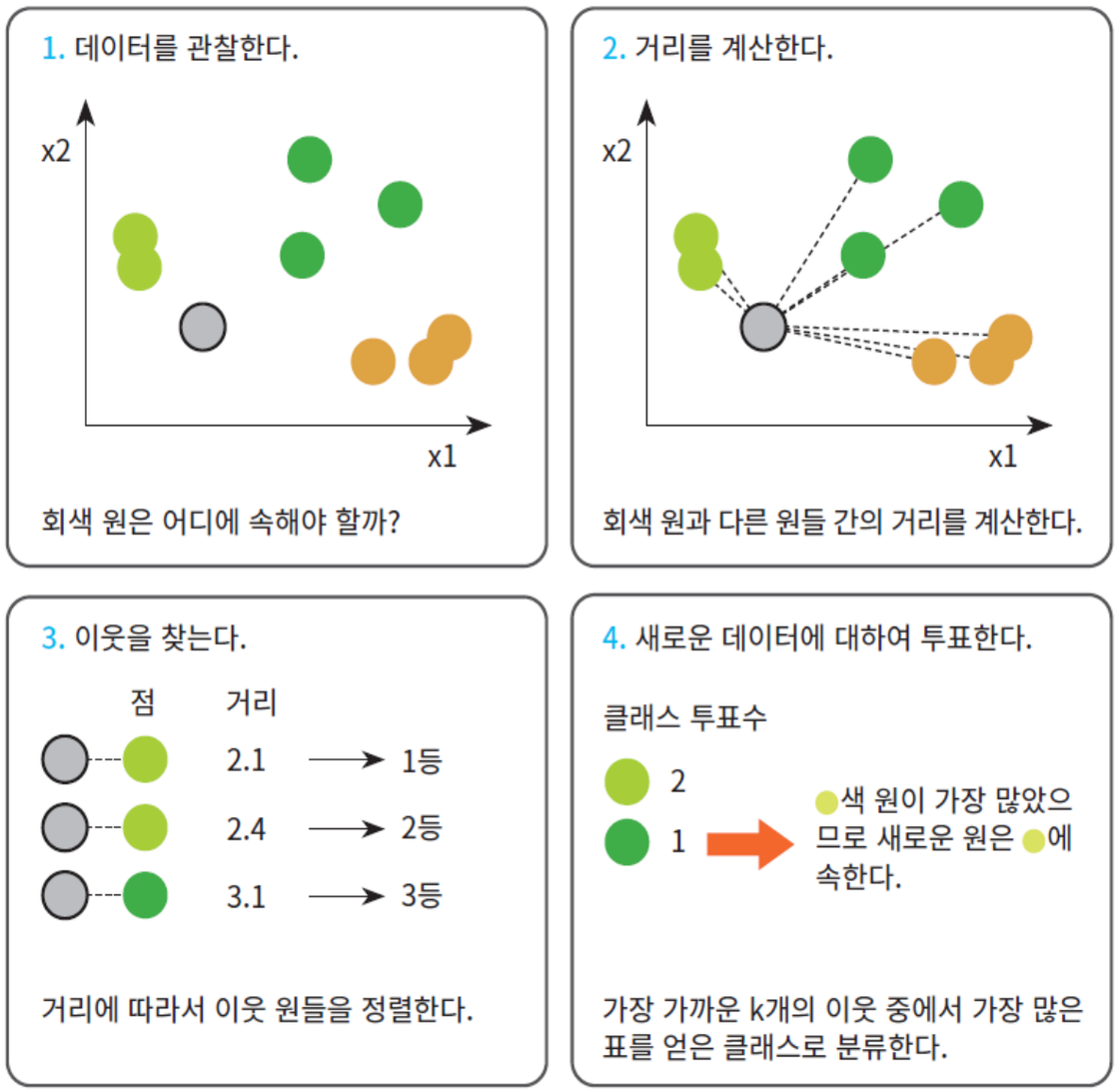


최근접 이웃법 (3/5)

- kNN(k-Nearest Neighbor)
  - 주변에 몇 개의 것들을 같이 봐서 가까운 것을 골라내는 방식
  - 1950년대에 개발된 지도 학습 모델의 분류 기법
  - 간단한 분류 기법으로 최근접 이웃 분류라고 불림
  - 가장 가까운 것들과의 거리 계산으로 클래스를 분류
  - 새로운 입력 데이터와 가장 가까운 k개의 이웃 데이터 선택
  - 이웃 데이터들의 클래스 중 다수결로 데이터의 클래스 결정
  - 다수결에서 결과가 나오기 위해 k는 반드시 홀수여야 함

최근접 이웃법 (4/5)

- kNN(k-Nearest Neighbor) 절차



최근접 이웃법 (5/5)

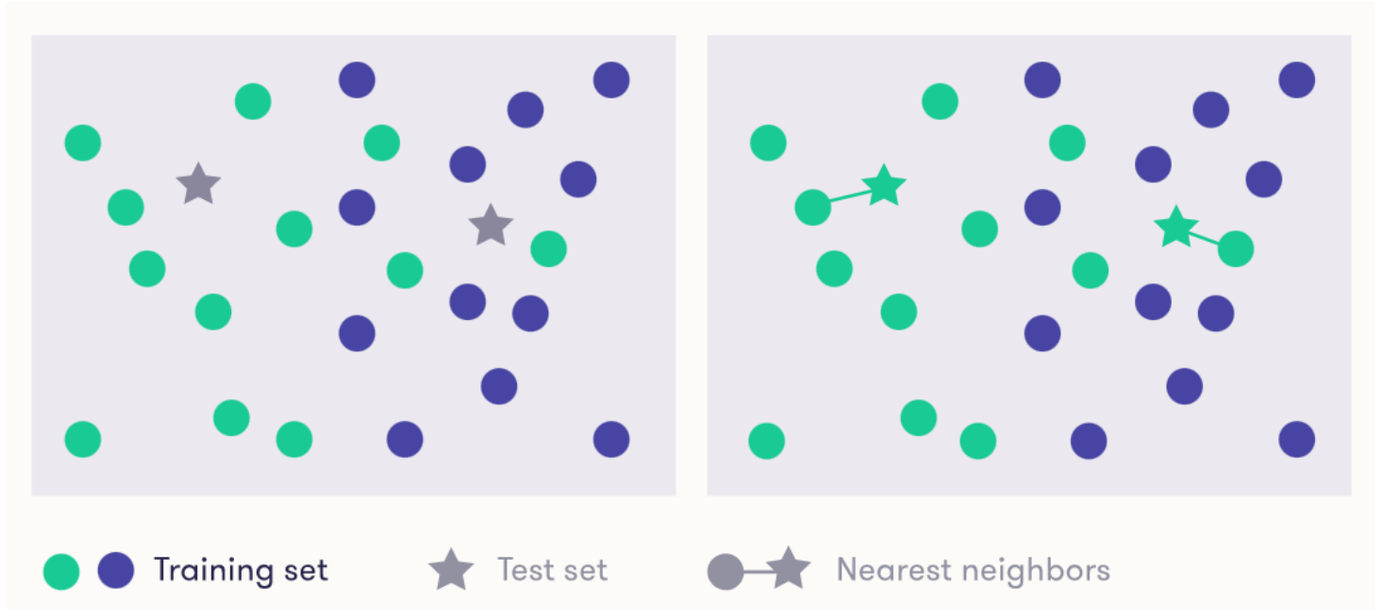
- k값은 얼마가 좋을까 ?

➢ 특징 공간에 있는 모든 데이터에 대한 정보가 필요 : 가장 가까운 이웃을 찾기 위해 새로운 데이터에서 모든 기존 데이터까지의 거리를 확인해야 함 ⇒ 많은 메모리 공간과 계산시간이 필요함

- kNN의 활용 분야
  - 영화나 음악 추천에 대한 개인별 선호 예측
  - 수표에 적힌 광학 숫자와 글자인식
  - 얼굴인식과 같은 컴퓨터 비전
  - 유방암 등 질병의 진단과 유전자 데이터 인식
  - 재정적인 위험성의 파악과 관리, 주식 시장 예측

최근접 이웃법 (5/5)

- “가까움(nearness)”의 척도(Metric)
  - 유클리드 거리(Euclidean distance)



- MNIST digit recognition에서는 픽셀 별 일치 수(count of pixel-by-pixel matches) ⇒ 이미지를 이동하거나 크기를 조정하는데 매우 민감

최근접 이웃법의 활용 – 사용자 행동 예측

- 사용자 행동 예측의 기본 아이디어
  - 유사한 과거 행동을 가진 사용자는 유사한 미래 행동을 하는 경향이 있음
  - 음악 추천 시스템은 청취 행동에 대한 데이터를 수집
  - 수집된 데이터에 없는 신곡은 어떻게 되나요?
  - 협업 필터링: 다른 사용자의 데이터를 사용하여 선호도를 예측

연습문제

온라인 쇼핑을 위한 추천시스템

사용자의 구매 내역이 기록되어 다음에 구매할 가능성이 있는 제품을 예측하는데 사용

User	Shopping History				Purchase
Sanni	boxing gloves	Moby Dick (novel)	headphones	sunglasses	coffee beans
Jouni	t-shirt	coffee beans	coffee maker	coffee beans	coffee beans
Janina	sunglasses	sneakers	t-shirt	sneakers	ragg wool socks
Henrik	2001: A Space Odyssey (dvd)	headphones	t-shirt	boxing gloves	flip flops
Ville	t-shirt	flip flops	sunglasses	Moby Dick (novel)	sunscreen
Teemu	Moby Dick (novel)	coffee beans	2001: A Space Odyssey (dvd)	headphones	coffee beans

User	Shopping History				Purchase
Travis	green tea	t-shirt	sunglasses	flip flops	?

유사성 (가까움)
=
두 사용자가 모두 구매한 항목의 개수

Travis의 다음 구매를 예측하기

연습문제

온라인 쇼핑을 위한 추천시스템

User	Shopping History				Purchase
Sanni	boxing gloves	Moby Dick (novel)	headphones	sunglasses	coffee beans
Jouni	t-shirt	coffee beans	coffee maker	coffee beans	coffee beans
Janina	sunglasses	sneakers	t-shirt	sneakers	ragg wool socks
Henrik	2001: A Space Odyssey (dvd)	headphones	t-shirt	boxing gloves	flip flops
Ville	t-shirt	flip flops	sunglasses	Moby Dick (novel)	sunscreen
Teemu	Moby Dick (novel)	coffee beans	2001: A Space Odyssey (dvd)	headphones	coffee beans

User	Shopping History				Purchase
Travis	green tea	t-shirt	sunglasses	flip flops	?

가장 유사성이 높은 고객은 ?
Ville : 유사도 3

Ville의 최근 구매가 sunscreen
이므로, 추천 시스템은 Travis에게도
sunscreen 추천

Contents

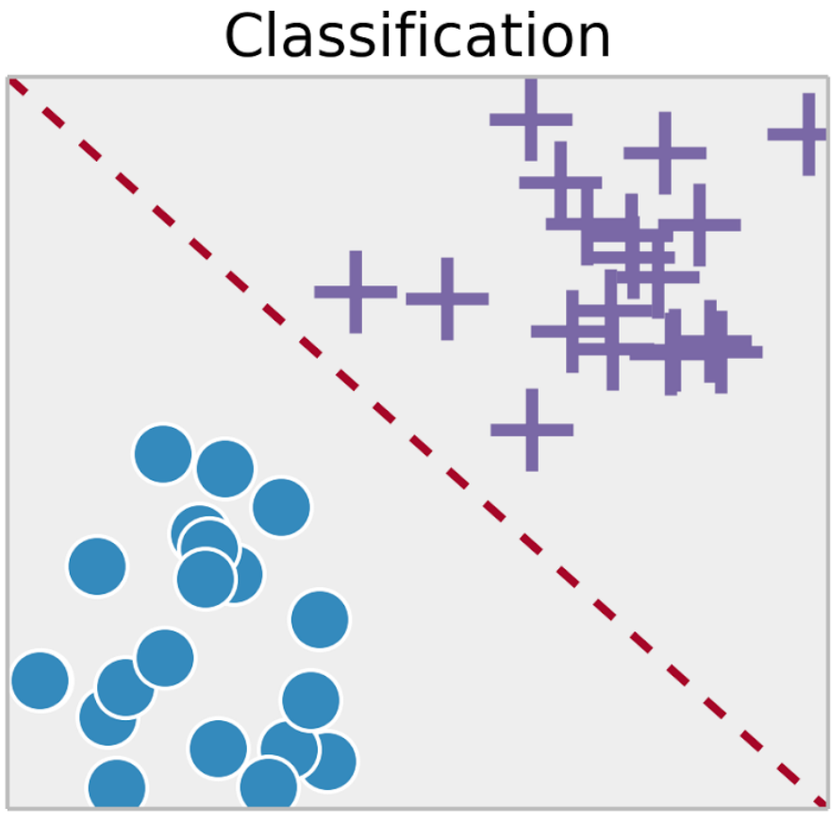
I. 머신러닝의 유형

II. 머신러닝의 절차

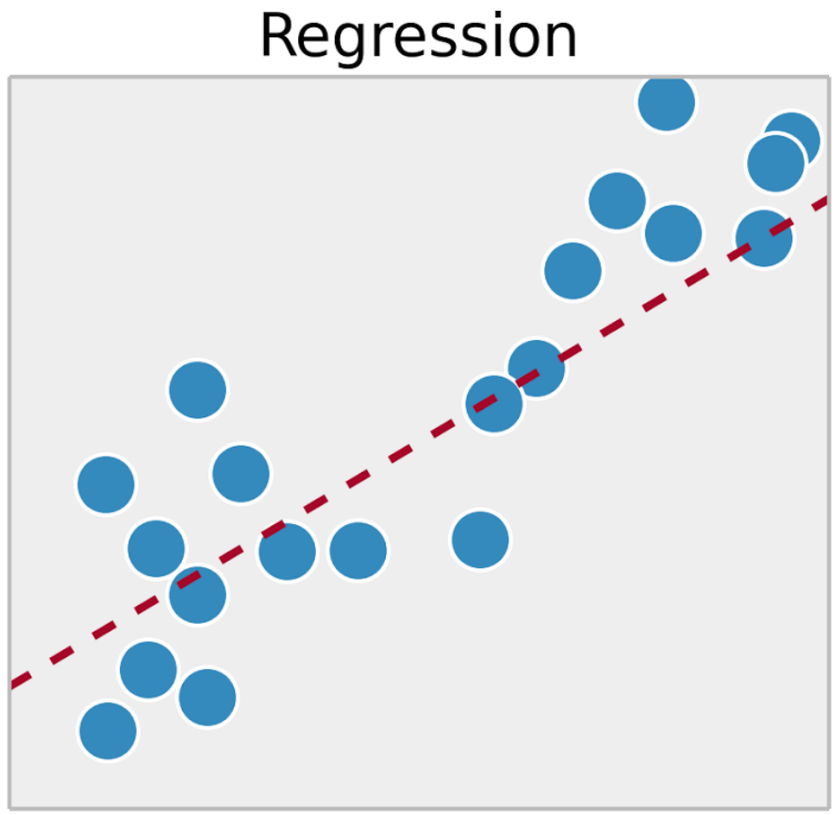
III. 최근접 이웃 분류기 (The Nearest Neighbor Classifier)

IV. 회귀(Regression)

분류와 회귀



- 스팸/햄 메일
- 0, 1, 2, …, 9



- 제품의 가격
- 장애물의 거리
- 다음 스타워즈 영화의 흥행 수익

연습문제

기대수명 예측하기

non-smoking human who don’t eat vegetables	life expectancy
women	80 years
men	75 years

	weight
cigarette smoking 1/day	- 0.5 year
vegetable consumption handful/day	+1 year

Gender	Smoking (cigarettes per day)	Vegetables (handfuls per day)	Life expectancy (years)
male	8	2	73
male	0	6	**A**
female	16	1	**B**
female	0	4	**C**

연습문제

기대수명 예측하기

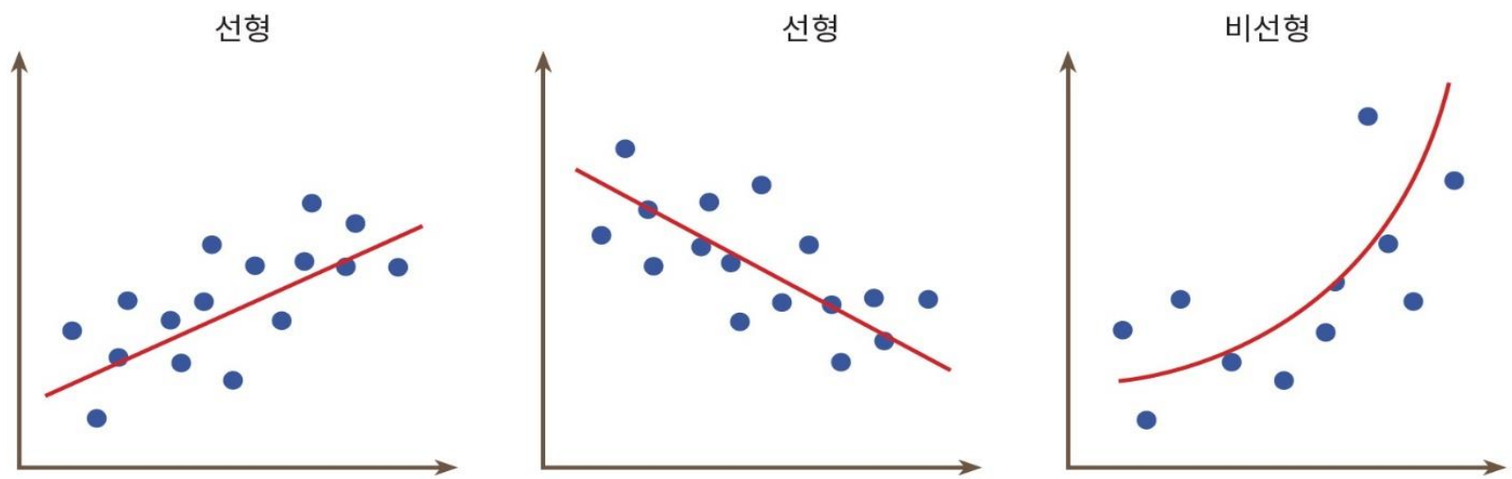
non-smoking human who don't eat vegetables	life expectancy
women	80 years
men	75 years

	weight
cigarette smoking 1/day	- 0.5 year
vegetable consumption handful/day	+1 year

Gender	Smoking (cigarettes per day)	Vegetables (handfuls per day)	Life expectancy (years)
male	8	2	73
male	0	6	81
female	16	1	73
female	0	4	84

선형회귀 (1/5)

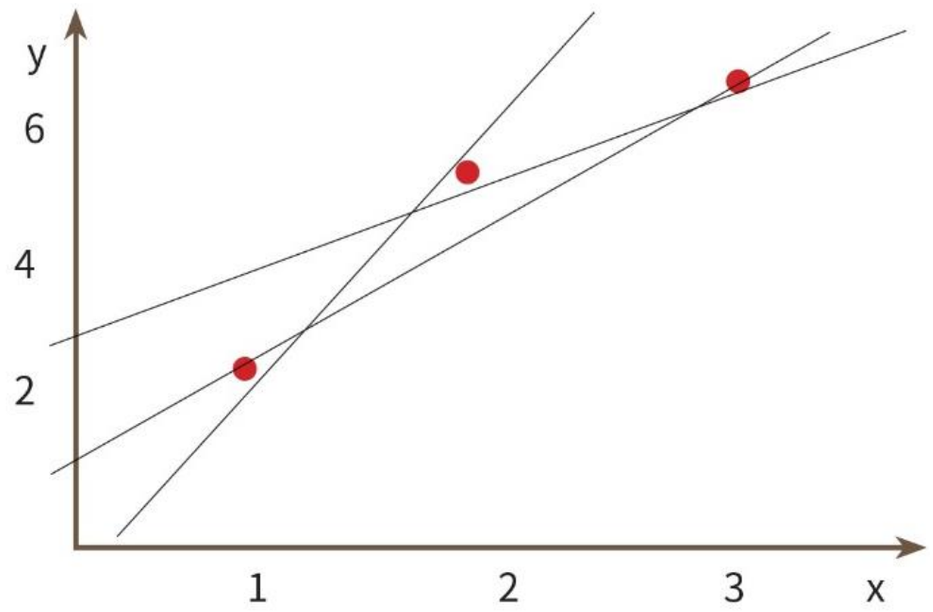
- 선형회귀(Linear Regression)
  - 데이터들을 가장 잘 설명하는 직선을 찾는 문제
  - 정수로 제한되지 않는 수치 예측
  - 각 특징변수의 효과를 더하여 예측 값을 생성 ⇒ 선형조합 (Linear combination)



회귀의 종류

선형회귀 (2/5)

- 직선의 방정식 : $f(x) = mx + b$
- 선형회귀는 **입력 데이터를 가장 잘 설명하는 기울기와 절편값을 찾는 문제**
- 선형회귀의 기본식 $f(x) = wx + b$
  - 기울기 ⇒ 가중치
  - 절편 ⇒ 바이어스



선형회귀 (3/5)

- 선형회귀 예제

키(단위: cm)	몸무게(단위: kg)
174	71
152	55
138	46
128	38
186	88

학습 데이터

학습

키(x)=165이면
몸무게는?

선형 회귀 모델

몸무게(y)=63으로
예측됩니다.

선형회귀 (4/5)

- 선형회귀의 종류
  - 단순선형회귀 : 독립 변수(x)가 하나인 선형 회귀

$$f(x) = wx + b$$

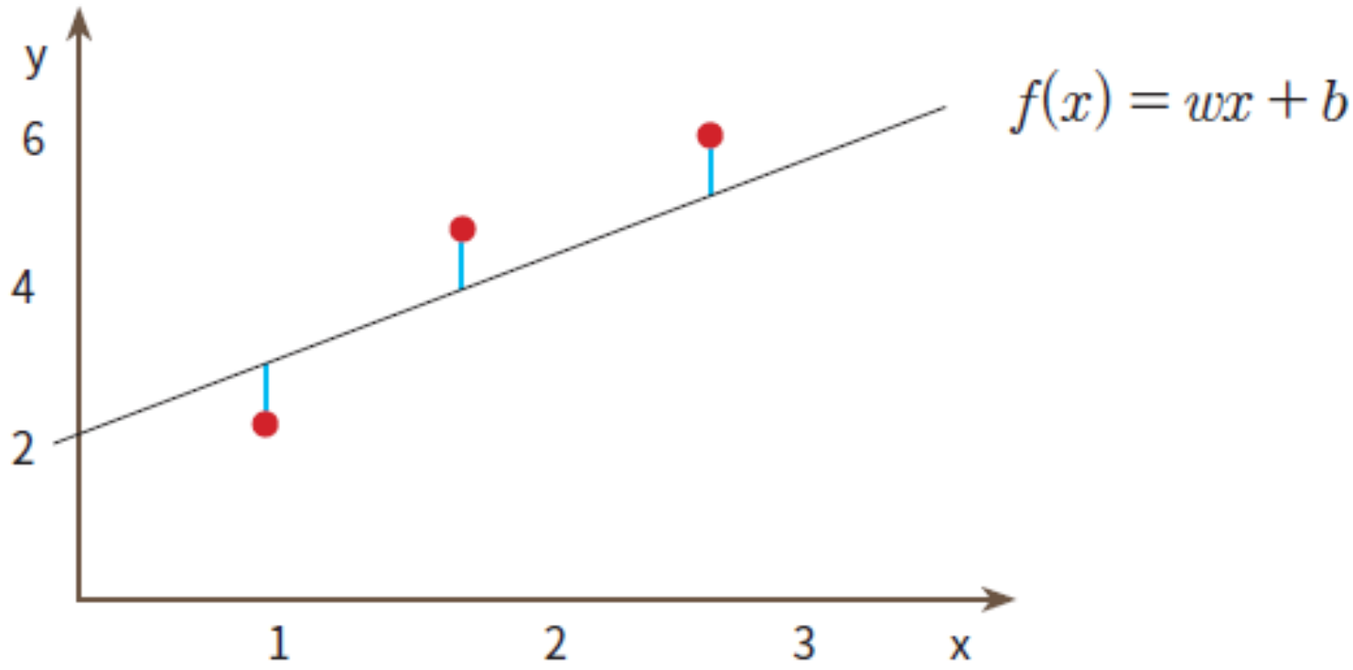
    - w, b : 정확한 예측을 생성하기 위해 알고리즘이 '학습'하려고 시도하는 매개 변수
    - x, y는 학습 데이터, f(x)는 우리의 예측

  - 다중선형회귀

$$f(x) = w_0 + w_1 x + w_2 y + w_3 z$$

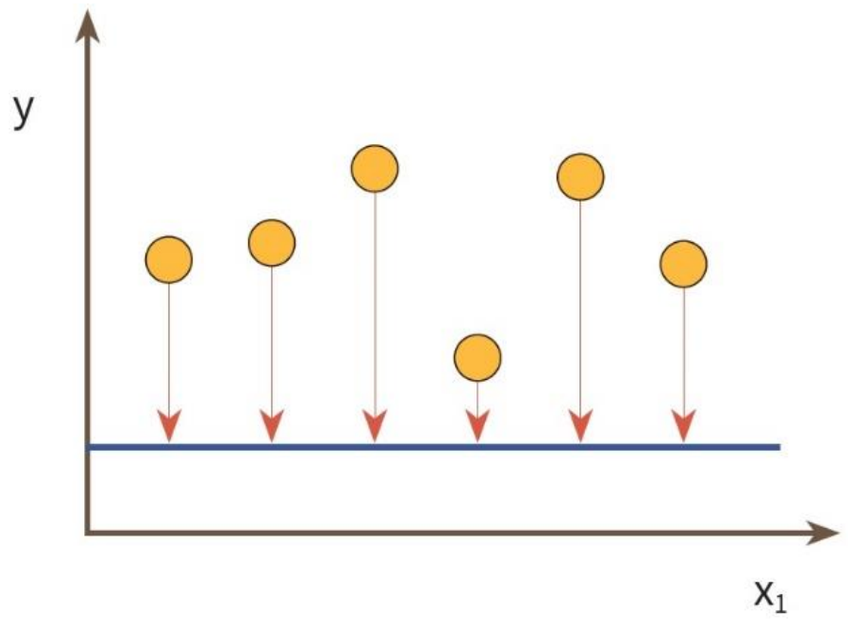
선형회귀 (5/5)

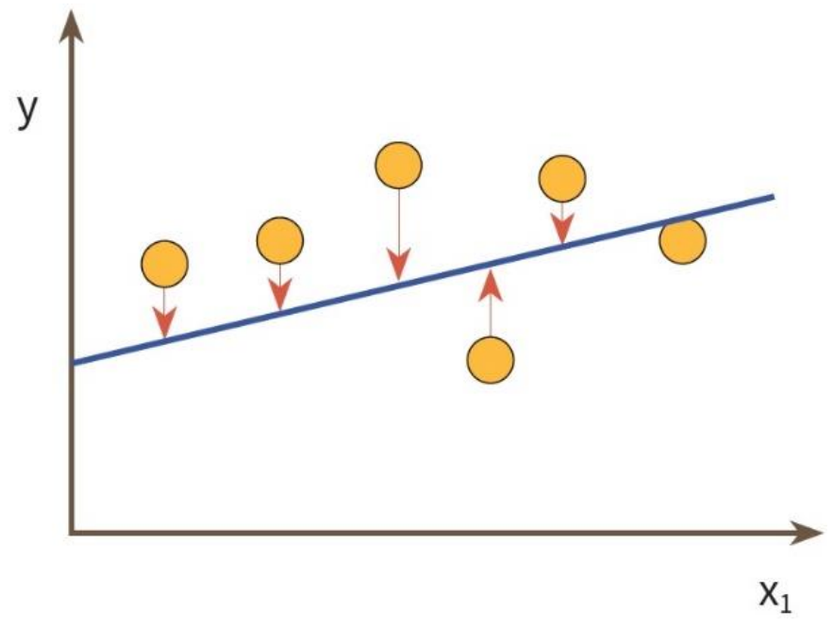
- 문제는 어떤 직선이 우리가 제공한 데이터와 가장 잘 맞느냐



손실함수 (1/2)

- 손실 함수(loss function) 또는 비용 함수(cost function)
  - 예측값과 실제값(레이블)의 차이를 구하는 기준
  - 학습 중에 알고리즘이 잘못 예측하는 정도를 확인하기 위한 함수로써 최적화를 위해 최소화하는 것이 목적
  - 머신러닝 모델의 학습에서 필수적





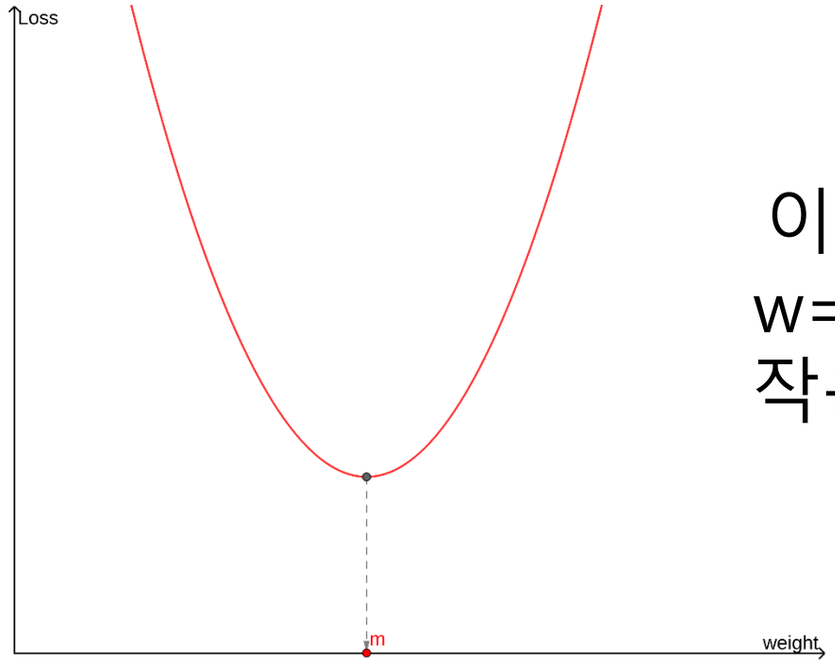
손실함수 (2/2)

- 최소자승법(Least Squares Method)
  - 손실함수 최소화 방법
  - 직선과 데이터 사이의 간격을 제곱하여 합한 값이 최소가 되도록 해를 구하는 방법
  - $f(x) = wx + b$ 일 경우, 다음 값이 최소화되도록 $w,\ b$를 결정

$$\sum_{i=1}^{n} (y_i - w \cdot x_i - b)^2$$

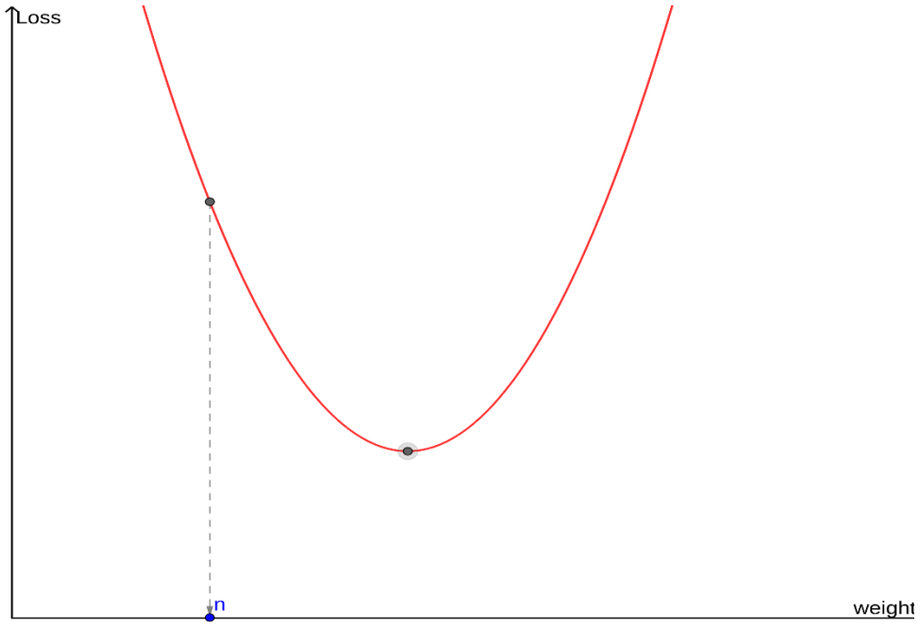
경사하강법 (1/3)

- 경사하강법(Gradient Descent)
  - 머신러닝 모델의 옵티마이저의 한 종류
  - 기울기(경사)를 이용하여 손실함수의 값을 최소화 하는 방법
  - 조정하고자 하는 값은 가중치(w)와 바이어스(b) ⇒ 손실함수를 w와 b에 관한 함수로 생각한다면,



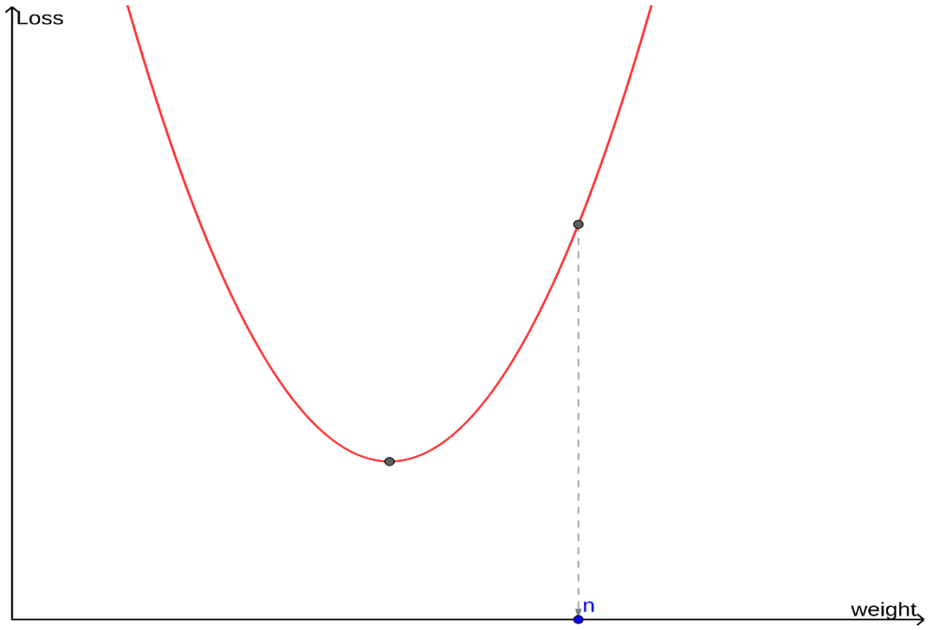
이 손실함수는 w=m에서 가장 작은 값을 가짐

경사하강법 (2/3)



이동 방향은 +가 된다.

w = n이라면,
w = w + 양수로
조정



이동 방향은 -가 된다.

w = n이라면,
w = w + 음수로
조정

경사하강법 (3/3)

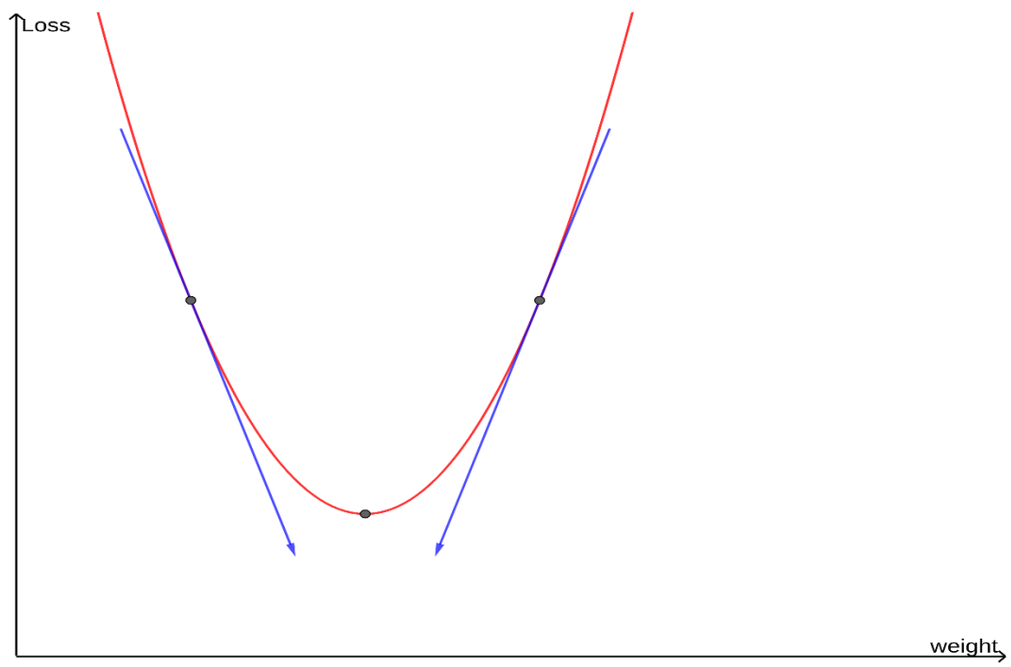
- 경사하강법은 기울기를 이용하는 방법으로
  - w 값에서 손실함수의 미분계수가 음수이면, w를 양의 방향으로, 양수이면 w를 음의 방향으로 이동
  - w를 얼만큼 이동시키는지가 관건

➢ 경사하강법에서 w의 조정 식

$$w = w - \alpha \times \frac{\partial L}{\partial w}$$

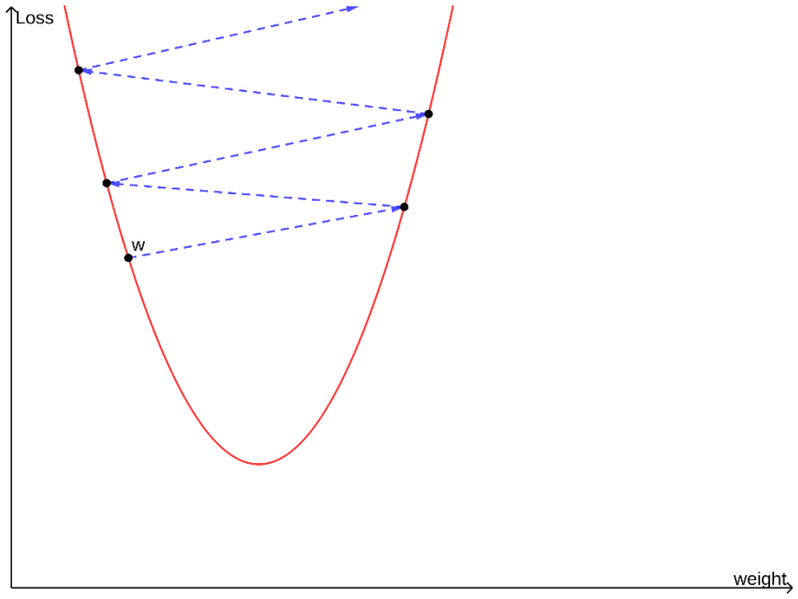
(α는 학습율)

➢ 손실함수가 최소값을 갖도록 w값을 반복적으로 조정

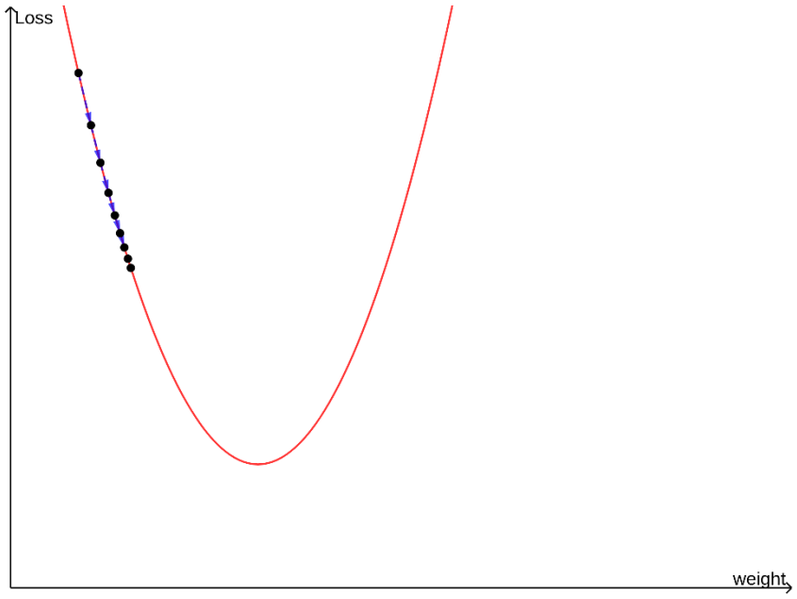


학습율

- 학습률(learning rate)은 한 번에 매개 변수를 변경하는 비율



학습율이 너무 큼



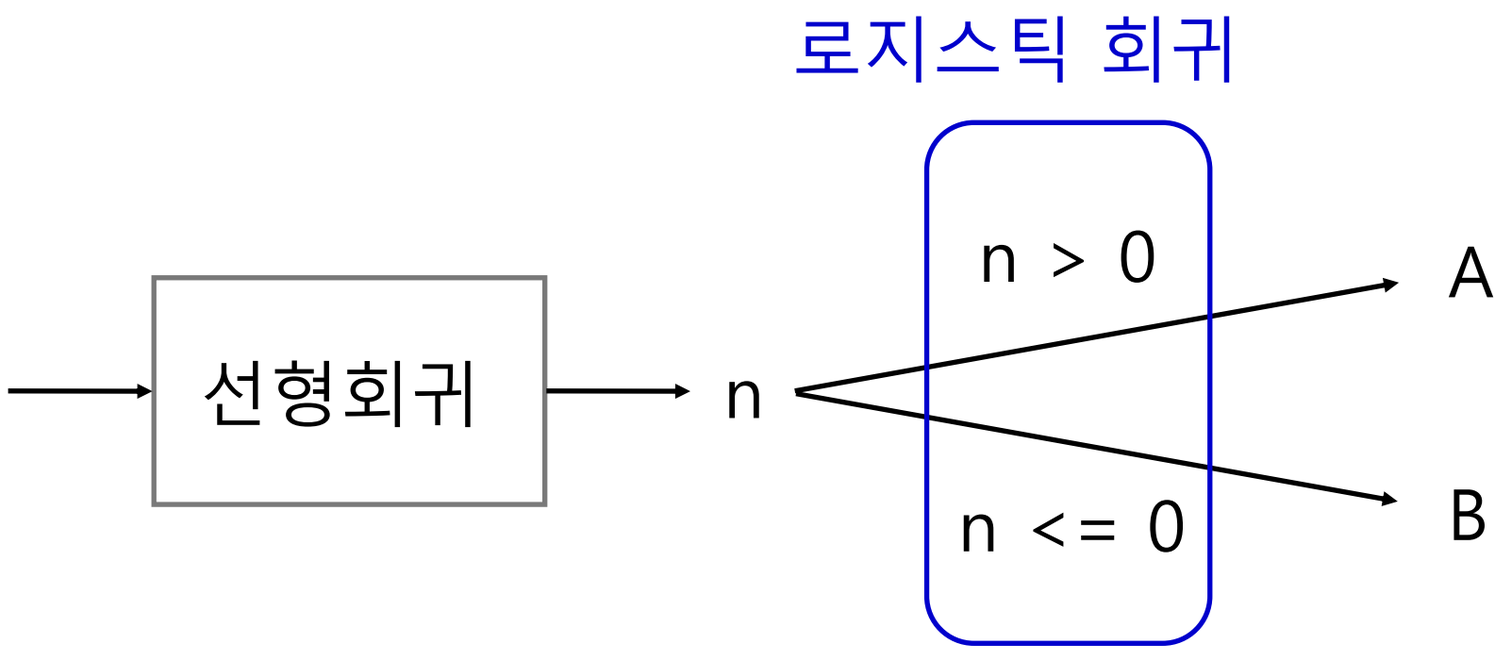
학습율이 너무 작음

선형회귀의 응용

- 온라인 광고의 클릭률 예측
- 제품에 대한 소매 수요 예측
- 헐리우드 영화 흥행 수익 예측
- 소프트웨어 비용 예측
- 보험 비용 예측
- 범죄율 예측
- 부동산 가격 예측

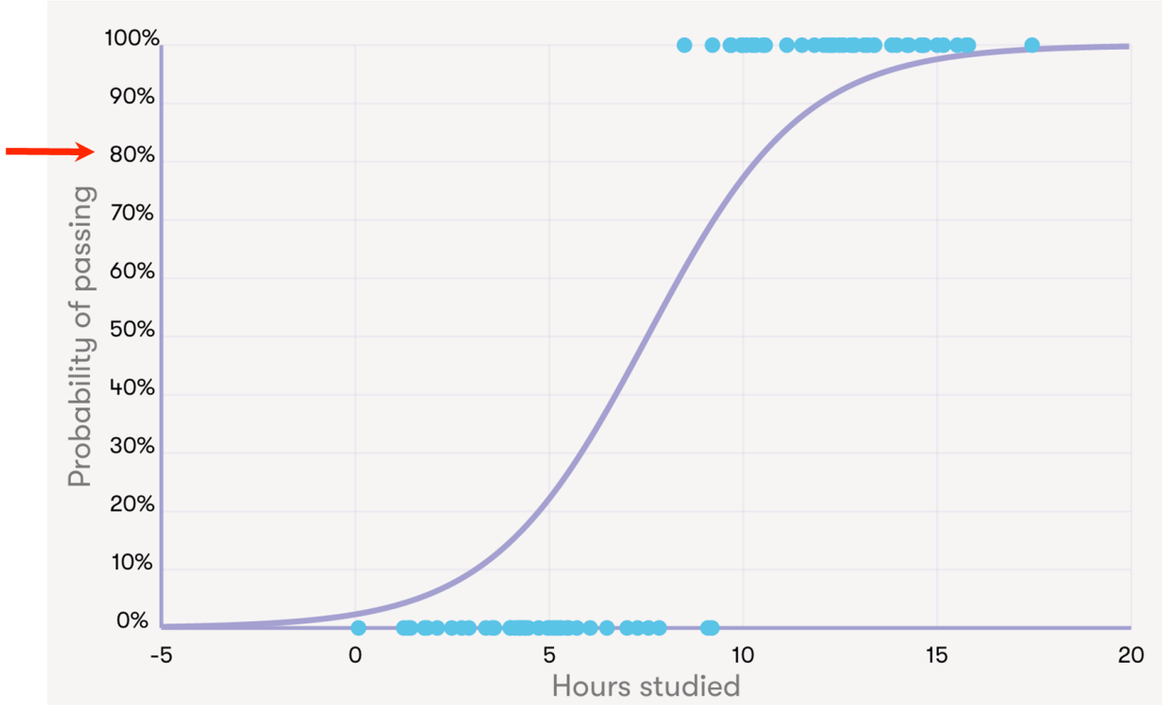
회귀를 이용한 레이블 예측 ?

- 로지스틱 회귀(Logistic regression)
  - 선형회귀의 출력을 레이블에 대한 예측으로 바꾸는 방법
  - 선형회귀를 사용하여 데이터가 어떤 범주에 속할 확률을 0~1 사이의 값으로 예측하고 그 확률에 따라 가능성이 더 높은 범주에 속하는 것으로 분류해주는 지도학습 알고리즘



연습문제 : 로지스틱 회귀

Student ID	Hours studied	Pass/fail
24	15	Pass
41	9.5	Pass
58	2	Fail
101	5	Fail
103	6.5	Fail
215	6	Pass

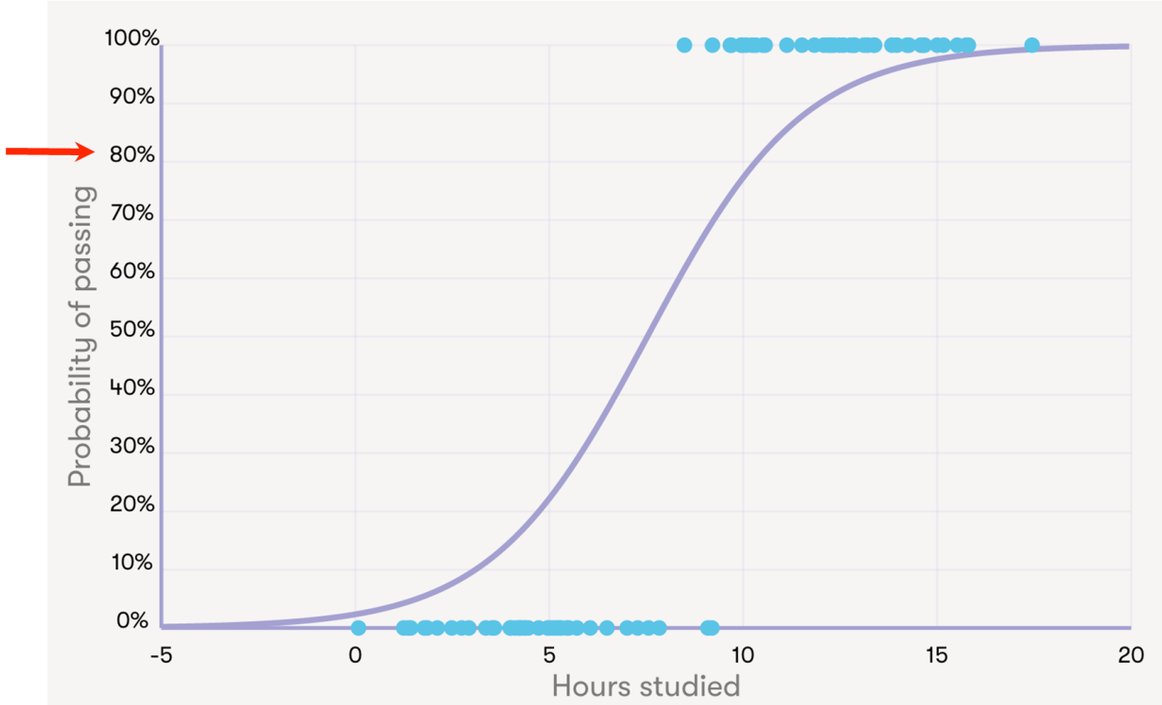


대학시험에 합격할 확률을 80%로 하고 싶다면 대략 몇 시간 정도 공부해야 할까?

6 ~ 7 시간 | 7 ~ 8 시간 | 8 ~ 9 시간 | 10 ~ 11 시간

연습문제 : 로지스틱 회귀

Student ID	Hours studied	Pass/fail
24	15	Pass
41	9.5	Pass
58	2	Fail
101	5	Fail
103	6.5	Fail
215	6	Pass



대학시험에 합격할 확률을 80%로 하고 싶다면 대략 몇 시간 정도 공부해야 할까?

6 ~ 7 시간 | 7 ~ 8 시간 | 8 ~ 9 시간 | 10 ~ 11 시간

머신러닝의 한계

- 머신 러닝은 AI 애플리케이션을 구축하기 위한 매우 강력한 도구지만 매우 어려운 문제
- ➢ 항상 올바른 레이블을 생성하는 완벽한 방법은 없음

- ML 기술(최근접 이웃 방법, 선형 회귀, 로지스틱 회귀 등 수백 가지)이 완벽하지는 않지만, 좋은 예측이 없는 것보다 낫다

예측 오류(Prediction Errors)

- 예측 품질 측정(Measuring the quality of prediction)
  - 분류 오류율
  - 예상 주택 가격과 최종 판매 가격의 차이
  - 치명적 오류! : 차 앞 보행자 감지 실패예측

➢ 정확도의 목표는 사례별로 달라짐 (The goal of how accurate the prediction should be is different case-by-case)