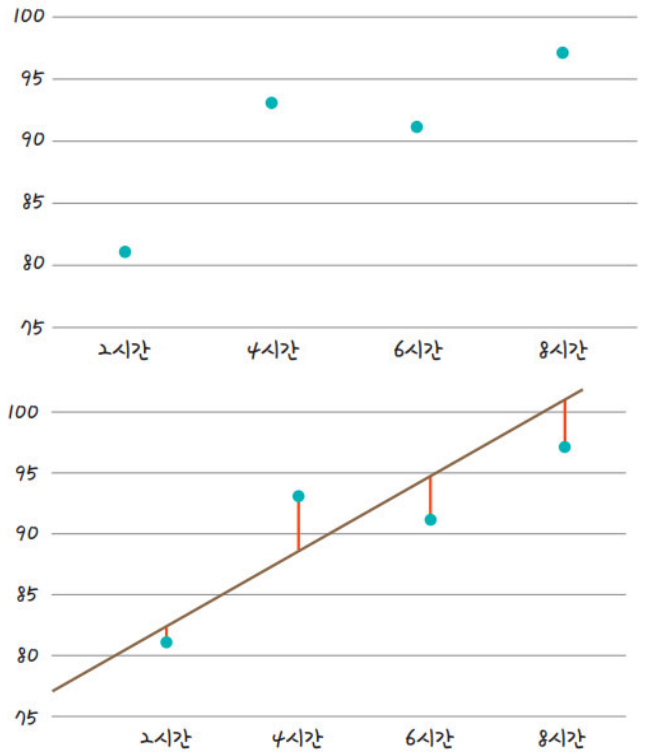
둘째마당
예측 모델의 기본 원리

4장 가장 훌륭한 예측선

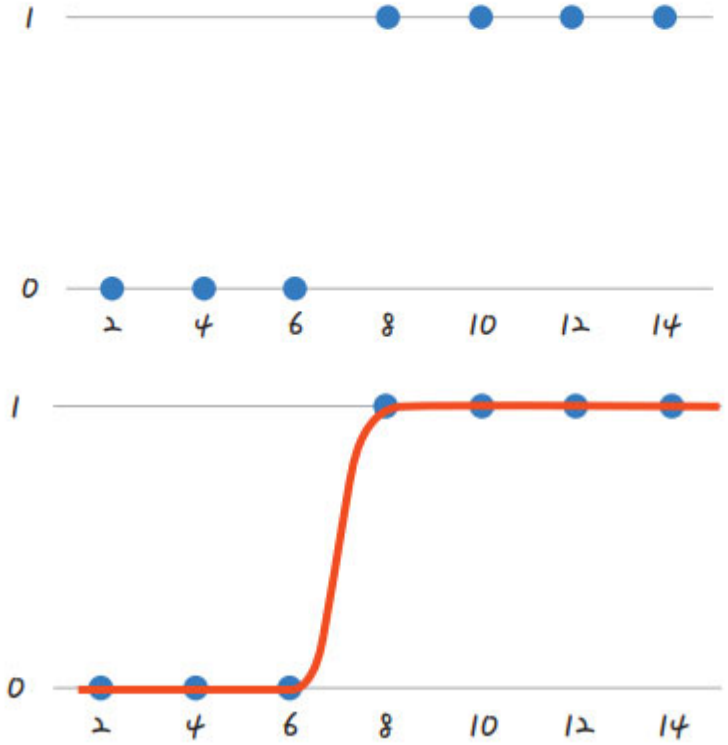
1 선형 회귀의 정의

가장 훌륭한 예측선

- 우리의 목표는 아래 그래프와 같이 데이터 분포가 있을 때 점들의 특징을 가장 잘 나타내는 함수식을 구하는 것이다.
- 왼쪽 그래프와 같이 데이터가 있을 때에는 점들의 분포에 맞게 선형회귀 선이 그어주면 된다.
- 오른쪽 그래프와 같이 데이터가 있을 때에는 S모양으로 선이 그어지게 되어 클래스가 정해져 있어서 0아니면 1과 같은 경우에 해당할 것이다.
- 그래서 딥러닝이라는 것은 이렇게 적절한 선을 긋는 것을 말한다.



선형회귀



로지스틱 회귀

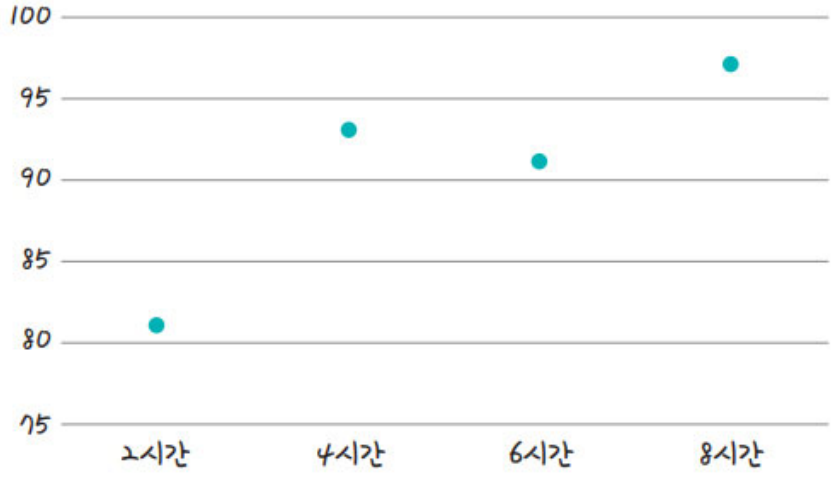
2 가장 훌륭한 예측선이란?

2 가장 훌륭한 예측선이란?

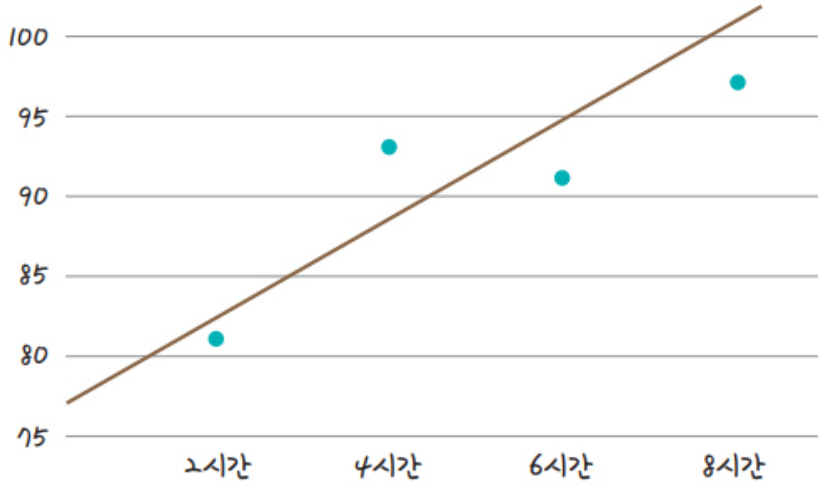
● 가장 훌륭한 예측선이란?

- 다음과 같이 중간고사를 본 4명의 학생에게 각각 공부한 시간을 물어보고 이들의 중간고사 성적을 다음과 같이 정리했다
- 공부 시간만 보고 시험 점수를 예측하는 것이 우리의 목표이다.
- 공부 시간과 점수간에는 어떠한 상관관계가 있는 지 확인하기 위해 좌표평면상에 펼쳐 놓은 것이다.
- 분포를 보면 시간이 적으면 점수가 낮고 높으면 점수가 높은 걸 알 수 있다.
- 점들이 가지는 특징을 선으로 잘 그리는 과정이 필요하다. 평면에 없는 시간에 대한 점수를 예측할 수 있다. 적절한 a값과 b값을 찾아서 적절한 선을 그어주는 것이 목표이다.

공부한 시간	2시간	4시간	6시간	8시간
성적	81점	93점	91점	97점



$$y = ax + b$$



3 최소 제곱법

3 최소 제곱법

● 최소 제곱법

● 구해야 하는 a값과 b값은 최소제곱법을 이용하면 쉽게 구할 수 있다.

$$a = \frac{(x - x\text{ 평균})(y - y\text{ 평균})\text{의 합}}{(x - x\text{ 평균})^2\text{의 합}} \qquad (\text{식 } 4.1)$$

공부한 시간(x) 평균: (2 + 4 + 6 + 8) ÷ 4 = 5

성적(y) 평균: (81+ 93 + 91 + 97) ÷ 4 = 90.5

$$a = \frac{(2-5)(81-90.5)+(4-5)(93-90.5)+(6-5)(91-90.5)+(8-5)(97-90.5)}{(2-5)^2+(4-5)^2+(6-5)^2+(8-5)^2}$$

$$= \frac{46}{20}$$

$$= 2.3$$

$$b = y\text{의 평균} - (x\text{의 평균} \times \text{기울기 } a) \qquad (\text{식 } 4.2)$$

$$b = 90.5 - (2.3 \times 5)$$

$$= 79$$

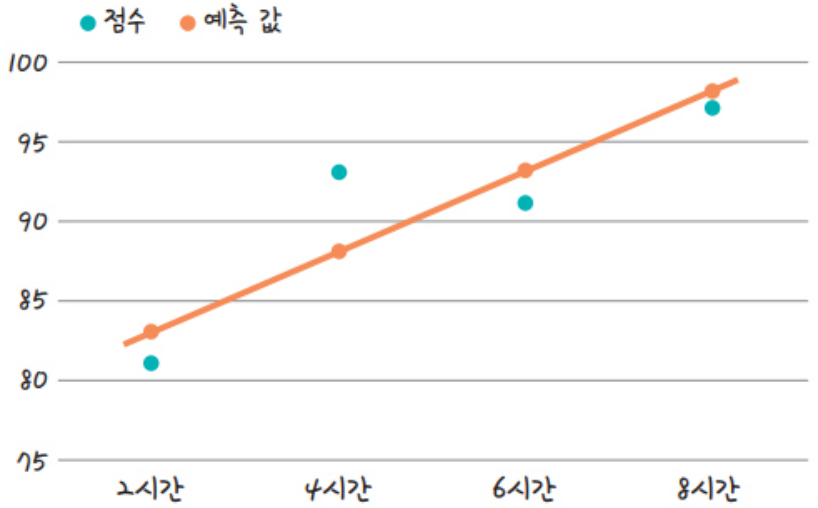
그러므로 우리가 구하려는 식은 $y = 2.3x + 79$ 가 된다.

3 최소 제곱법

● 최소 제곱법

- 최적의 선인지를 확인하기 위서 값을 대입해 보면 다음과 같이 예측 값이 찍힌다.
- $y = 2.3x + 79$에 따라 표에 정의된 것 처럼 예측 값은 주황색 점이 될 것이다.

공부한 시간	2	4	6	8
성적	81	93	91	97
예측 값	**83.6**	**88.2**	**92.8**	**97.4**



- 예측 값에 따라 선들을 이어보면 오른쪽 그래프와 같이 최적의 선이 되어 시간에 따라 예측 값을 알 수 있게 된다.

3 최소 제곱법

- **가장 훌륭한 예측선이란?**
  - 만약 x값이 두 개이면 $y = a_1x_1 + a_2x_2 + b$ 과 같은 식이 만들어 질 것이다.
  - 그리고 암환자의 데이터 같은 경우 x가 17개가 되어 x를 n개 만큼 더해야 한다.

$$y = a_1x_1 + a_2x_2 + \cdots + a_nx_n + b$$

  - 그래서 팀러닝은 항이 많이지면 최소제곱법을 사용할 수 없기 때문에 다음과 같은 과정을 반복하는데.

1. 일단 아무 값이나 넣는다

2. 오차를 구한다

3. 값을 수정한다

  - 오차가 적어질 때까지 수정을 해 나간다
  - 오차란 실제값과 예측값의 차이를 말하는 것이다.
  - 오차를 알기 위해서 사용하는 것이 평균제곱오차 이다.

4 평균 제곱 오차

5 평균 제곱 오차

- **평균 제곱 오차** (Mean Square Error, MSE)
  - 기울기 a와 y 절편 b를 임의의 수 3과 76이라고 가정할 때
  - Y = 3x + 76식이 되고 오차는 다음 표와 같게 된다.

오차 = 실제 값 – 예측 값

공부한 시간(x)	2	4	6	8
성적(실제 값, y)	81	93	91	97
예측 값	82	88	94	100
오차	1	–5	3	3

  - 이렇게 해서 구한 오차를 모두 더하면 1 + (-5) + 3 + 3 = 2가 됨
  - 그런데 여기서 +와 –의 값이 서로 같다면 0이 나오므로 오차가 얼마가 되는지 모른다.
  - 그래서 오차를 제곱해서 더하는 방법을 사용한다.

$$\text{평균 제곱 오차(MSE)} = \frac{1}{n}\sum (y_i - \hat{y}_i)^2$$

  - 이 식으로 오차의 합을 다시 계산하면 1 + 25 + 9 + 9 = 44
  - 평균제곱오차 MSE = 44/4 = 11
  - 평균제곱오차가 가장 작아질 때까지 a값과 b값을 바꿔주게 된다.