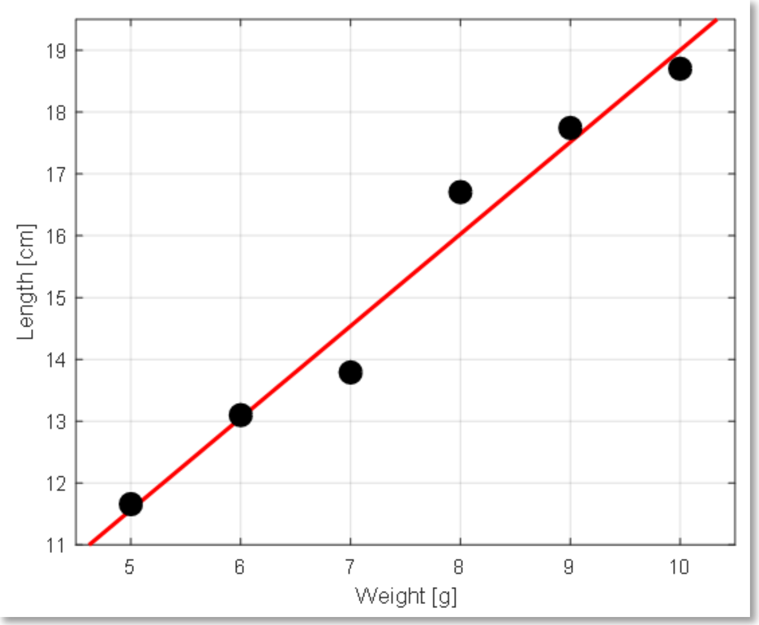
선형회귀 #2
(Linear Regression)

한국공학대학교
전자공학부
채승호 교수

단항 선형회귀

- 단항 선형회귀
  - 입력속성이 한 개인 경우의 선형회귀
  - 1개의 입력 속성 & 1개의 출력 속성으로 이루어진 2차원 속성 공간
  - 1차 직선으로 모델링
    - 2개의 매개변수 최적화 문제 : $\hat{y} = w_0 x + w_1$

데이터번호	무게(g)	늘어난길이(cm)
1	5	11.66
2	6	13.10
3	7	13.79
4	8	16.71
5	9	17.74
6	10	18.70



다항 선형회귀

- 다항 선형회귀
  - 입력 속성이 1개 이상인 경우의 선형회귀
  - 입력 속성의 증가로 다차원 속성공간 생성
  - 입력 속성이 M개인 훈련 데이터 집합의 표현

데이터 번호	입력	출력
0	$x_{0,0}, x_{1,0}, \dots, x_{m,0}, \dots, x_{M-1,0}$	y_0
1	$x_{0,1}, x_{1,1}, \dots, x_{m,1}, \dots, x_{M-1,1}$	y_1
$\vdots$	$\vdots$	$\vdots$
n	$x_{0,n}, x_{1,n}, \dots, x_{m,n}, \dots, x_{M-1,n}$	y_n
$\vdots$	$\vdots$	$\vdots$
$N-1$	$x_{0,N-1}, x_{1,N-1}, \dots, x_{m,N-1}, \dots, x_{M-1,N-1}$	y_{N-1}

  - 다항 선형모델
    - $\hat{y} = w_0 x_0 + w_1 x_1 + \cdots + w_{M-1} x_{M-1} + w_M$
    - 학습을 통해 $M+1$개의 최적 매개변수를 구함

다항 선형회귀

- Cost Function
  - 단일 변수에 대한 평균제곱오차

$$\epsilon_{MSE} = \frac{1}{N}\sum_{n=0}^{N-1}(\hat{y}_n - y_n)^2 = \frac{1}{N}\sum_{n=0}^{N-1}(w_0 \boxed{x_n} + w_1 - y_n)^2$$

단일변수(용수철무게)

  - 다항 변수에 대한 평균제곱오차

$$\epsilon_{MSE}(w_0, w_1, \cdots, w_M) = \frac{1}{N}\sum_{n=0}^{N-1}(\hat{y}_n - y_n)^2 = \frac{1}{N}\sum_{n=0}^{N-1}\Big(w_0 \underbrace{\overbrace{\boxed{x_{0,n}}}^{\text{input}}}_{\text{속도}} + w_1 \underbrace{\overbrace{\boxed{x_{1,n}}}^{\text{input}}}_{\text{질량}} + \cdots + w_M - \underbrace{\overbrace{\boxed{y_n}}^{\text{output}}}_{\text{거리}}\Big)^2$$

n: 측정한 데이터 수에 대한 index
M: 입력 데이터 Type 수 (데이터 차원 수)

다항 선형회귀

- 다항 선형회귀 모델의 해석해를 위한 표현
  - 매개변수 벡터 $\mathbf{w} = [w_0 \ \ w_1 \ \ \cdots \ \ w_{M-1} \ \ w_M]^T$
  - 출력(라벨) 벡터 $\mathbf{y} = [y_0 \ \ y_1 \ \ \cdots \ \ y_{N-1}]^T$
  - 입력 데이터 행렬 $\mathbf{X}$ ($N \times M$ 행렬)
    - M개의 입력 속성에 대해 N개의 데이터가 있음
    - 각 행은 개별 데이터를, 각 열은 개별 속성을 나타냄

데이터 번호	입력	출력
0	$x_{0,0}, x_{1,0}, \dots, x_{m,0}, \dots, x_{M-1,0}$	y_0
1	$x_{0,1}, x_{1,1}, \dots, x_{m,1}, \dots, x_{M-1,1}$	y_1
$\vdots$	$\vdots$	$\vdots$
n	$x_{0,n}, x_{1,n}, \dots, x_{m,n}, \dots, x_{M-1,n}$	y_n
$\vdots$	$\vdots$	$\vdots$
$N-1$	$x_{0,N-1}, x_{1,N-1}, \dots, x_{m,N-1}, \dots, x_{M-1,N-1}$	y_{N-1}

모두 1

$$\mathrm{X} = \begin{bmatrix} x_{0,0} & x_{1,0} & \cdots & x_{M-1,0} & \boxed{x_{M,0}} \\ x_{0,1} & x_{1,1} & \cdots & x_{M-1,1} & x_{M,1} \\ \vdots & \vdots & & \vdots & \vdots \\ x_{0,N-1} & x_{1,N-1} & \cdots & x_{M-1,N-1} & x_{M,N-1} \end{bmatrix} \begin{array}{c} \mathbf{x}_0 \\ \mathbf{x}_1 \\ \\ \mathbf{x}_{N-1} \end{array}$$

다항 선형회귀

- 선형회귀 모델
  - 기존 식

$$\hat{y} = w_0 x_0 + w_1 x_1 + \cdots + w_{M-1} x_{M-1} + w_M$$

  - 편의상 수정

$$\hat{y} = w_0 x_0 + w_1 x_1 + \cdots + w_{M-1} x_{M-1} + w_M x_M, \quad x_M = 1$$

$$\Rightarrow \quad \hat{y} = [w_0, w_1, \dots, w_{M-1}, w_M] \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{M-1} \\ x_M \end{bmatrix} = \mathbf{w}^T \mathbf{x}$$

- 비용함수(평균제곱오차)

$$\epsilon_{MSE} = \frac{1}{N} \sum_{n=0}^{N-1} (\hat{y}_n - y_n)^2 = \frac{1}{N} \sum_{n=0}^{N-1} (\mathbf{w}^T \mathbf{x}_n - y_n)^2$$

$$\hat{y}_n = [w_0, w_1, \dots, w_{M-1}, w_M] \begin{bmatrix} x_{0,n} \\ x_{1,n} \\ \vdots \\ x_{M-1,n} \\ x_{M,n} \end{bmatrix} = \mathbf{w}^T \mathbf{x}_n \qquad x_{M,n} = 1$$

다차원 입력데이터에 대한 Cost Function

→ Cost Function이 최소가 되는 $\mathbf{w} = [\mathrm{w}_0\ \mathrm{w}_1\ \mathrm{w}_2\ \cdots \mathrm{w}_M]$ 가 Optimal Solution

다항 선형회귀

- 다항 선형회귀 모델의 해석해
  - 비용 함수를 최소화하는 $M+1$개의 매개변수를 구함

$$\frac{\partial}{\partial w_m}\epsilon_{MSE} = 0, \qquad m = 0, 1, \cdots, M-1, M$$

$$\rightarrow \frac{\partial}{\partial w_m}\epsilon_{MSE} = \frac{2}{N}\sum_{n=0}^{N-1}(\mathbf{w}^T\mathbf{x}_n - y_n)\, x_{m,n} = 0 \text{ 이 되는 해}$$

- $m = 0$일 때, $\sum_{n=0}^{N-1}(\mathbf{w}^T\mathbf{x}_n - y_n)\, x_{0,n} = 0$
- $m = 1$일 때, $\sum_{n=0}^{N-1}(\mathbf{w}^T\mathbf{x}_n - y_n)\, x_{1,n} = 0$

…

- $m = M$일 때, $\sum_{n=0}^{N-1}(\mathbf{w}^T\mathbf{x}_n - y_n)\, x_{M,n} = 0$

$$\mathbf{x}_n^T = [x_{0,n}, x_{1,n}, \dots, x_{M-1,n}, x_{M,n}]$$

$$\sum_{n=0}^{N-1}\underbrace{\boxed{(\mathbf{w}^T\mathbf{x}_n - y_n)}}_{\text{scalar}}\mathbf{x}_n^T = [0,0,\dots,0,0]$$

  - 결과값

$$\mathbf{w} = \boxed{(\mathbf{X}^T\mathbf{X})^{-1}}\mathbf{X}^T\mathbf{y}$$

역행렬이 항상 존재한다는 것을 보장할 수 없음
즉, 수치적 접근을 통한 최적해를 구하는 것이 일반적

다항 선형회귀

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

다항 선형회귀 모델의 해석해 (예)

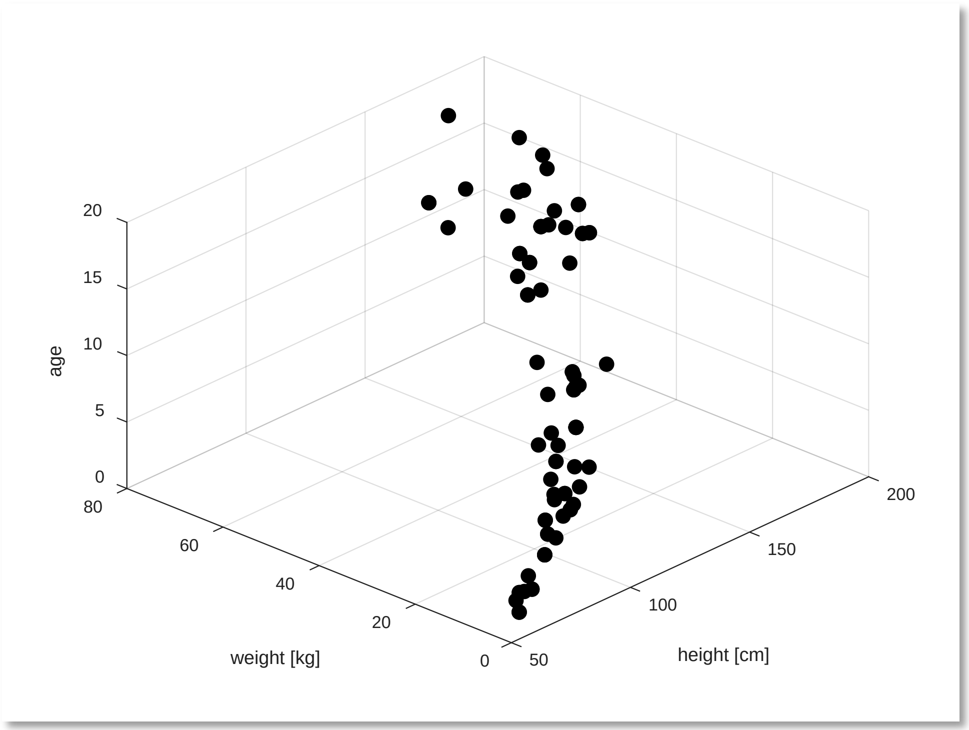
- 키와 몸무게로 나이를 추정하는 데이터 집합
  - 입력속성: 키(x_0), 몸무게(x_1) / 출력라벨: 나이(y)
- 선형회귀 모델
  - $\hat{y} = w_0x_0 + w_1x_1 + w_2x_2$, $x_2 = 1$
  - 3차원 공간에서의 평면의 방정식 (M=2)
- 80개의 데이터 (N=80)

n	x0	x1	y	n	x0	x1	y	n	x0	x1	y	n	x0	x1	y
1	152.6	44.7	11.5	21	64.4	5.5	0.3	41	72.2	8.3	0.8	60	119.9	21.2	7.3
2	160.4	39.9	15.8	22	67.5	7.7	0.6	42	124.6	29.4	8.2	61	160.1	55.3	14.9
3	112.2	21.1	6.6	23	70.7	8.6	0.8	43	134.5	27.8	8.3	62	131.9	27.6	8.2
4	159.3	46.4	15.6	24	162.1	49.4	14.8	44	171.3	53.6	18.8	63	103.7	15.5	4.5
5	150.9	46.6	11	25	167.3	55.6	16.2	45	162.1	49.4	14.8	64	131.9	27.6	8.2
6	91.7	13.1	2.8	26	155	35.8	16.9	46	155	35.8	16.9	65	111.1	21	5.5
7	74.8	8	0.8	27	70.7	8.6	0.8	47	76.7	9.7	1.4	66	86.2	11	2
8	111.1	16.1	4.3	28	74.8	8	0.8	48	159.6	71.5	13.6	67	152	48.8	13.7
9	162.9	43.8	12.8	29	130.5	26.5	8.7	49	150.9	46.6	11	68	103.4	14.2	3.5
10	131.3	27.6	9.6	30	91.7	13.1	2.8	50	159.4	42.9	15.9	69	105.9	14.8	3.6
11	105.3	19.2	4.9	31	108.6	23.4	6.6	51	138.2	30.7	8.3	70	162.1	49.4	14.8
12	177.8	61.7	18.4	32	160.1	40.6	17.9	52	86.2	11	2	71	160.4	39.9	15.8
13	101.8	16.8	4.4	33	94.7	15.1	3.3	53	94.7	15.1	3.3	72	103.7	15.5	4.5
14	151.5	49	12	34	95.3	13.2	2.2	54	154.7	48.1	12.9	73	111.1	21	5.5
15	86.2	11	2	35	162.7	46.9	16.3	55	174.7	74.9	18.4	74	119.9	21.2	7.3
16	137.8	23.7	10.2	36	131.6	35.1	9.2	56	131.3	27.6	9.6	75	164.7	55.5	16.3
17	160.1	40.6	17.9	37	154.7	61.4	16.5	57	113.5	15.3	5.7	76	100.2	14.1	3.3
18	119.9	21.2	7.3	38	64.4	5.5	0.3	58	119.9	21.2	7.3	77	170.1	52.1	18.1
19	156.4	65.9	12.8	39	159.6	71.5	13.6	59	102.6	17.1	3.9	78	152	48.8	13.7
20	154.7	48.1	12.9	40	115.6	24.2	6.8	60	103.4	14.2	3.5	79	114.1	18.6	5.2

다항 선형회귀

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

- 80개의 데이터 (N=80)



다항 선형회귀

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

- 80개의 데이터 (N=80, M=2)
  - 입력 행렬 $\mathbf{X}$, 80×3
  - 출력 벡터 $\mathbf{y}$, 80×1

152.6	44.7	1
160.4	39.9	1
112.2	21.1	1
159.3	46.4	1
150.9	46.6	1
91.7	13.1	1

⋮

100.2	14.1	1
170.1	52.1	1
152	48.8	1
114.1	18.6	1

11.5
15.8
6.6
15.6
11
2.8

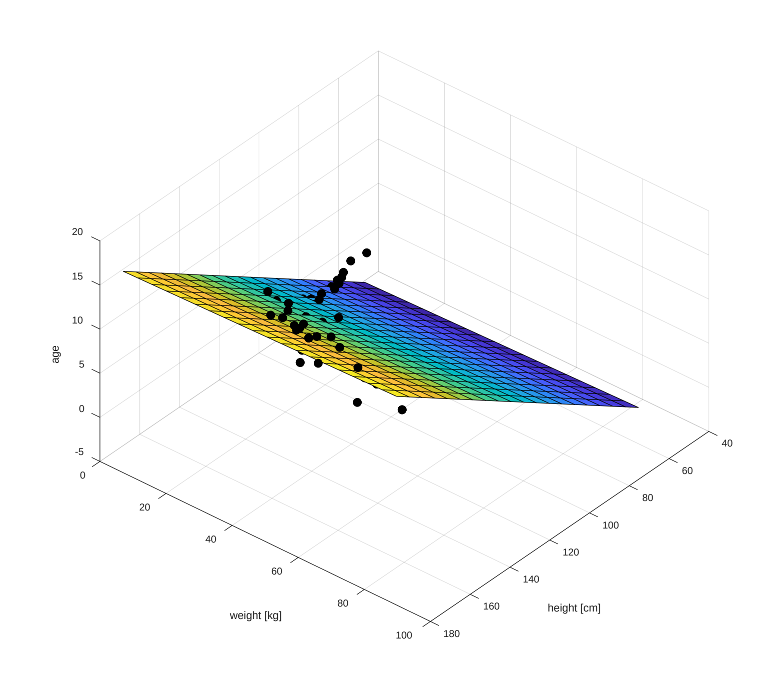
⋮

3.3
18.1
13.7
5.2

다항 선형회귀

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

- 80개의 데이터 (N=80, M=2)
  - 최적 매개변수
    - $w_0^* = 0.1625, \quad w_1^* = 0.0198, \;\; w_2^* = -12.2758$
  - 최적 선형모델
    - $\hat{y} = 0.1625x_0 + 0.0198x_1 - 12.2758$

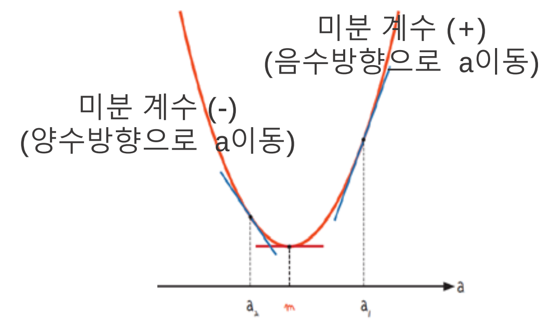


경사하강법(Gradient Descent Method)

- 위치 업데이트 규칙(다음 시점, $t+1$)

$\alpha \in [0,1]$

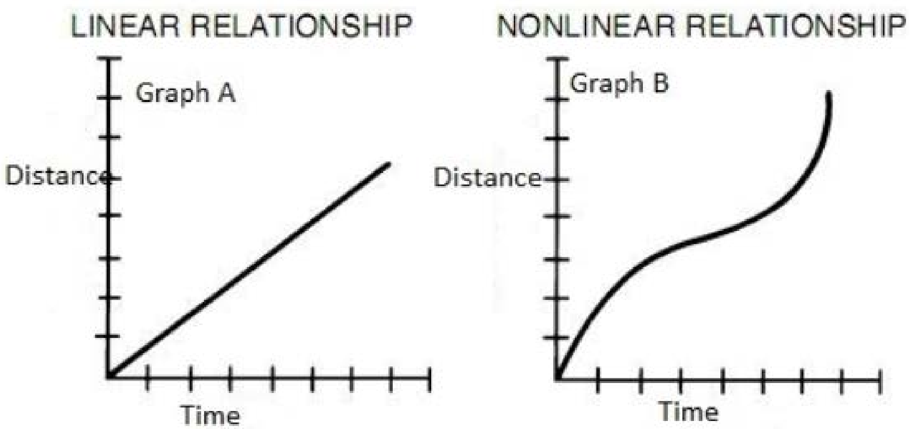
  - $w_0[t+1] = w_0[t] - \alpha \frac{\partial}{\partial w_0} \epsilon_{MSE}(w_0, w_1, w_2)\Big|_{w_0=w_0[t],\ w_1=w_1[t],\ w_2=w_2[t]}$
  - $w_1[t+1] = w_1[t] - \alpha \frac{\partial}{\partial w_1} \epsilon_{MSE}(w_0, w_1, w_2)\Big|_{w_0=w_0[t],\ w_1=w_1[t],\ w_2=w_2[t]}$
  - $w_2[t+1] = w_2[t] - \alpha \frac{\partial}{\partial w_2} \epsilon_{MSE}(w_0, w_1, w_2)\Big|_{w_0=w_0[t],\ w_1=w_1[t],\ w_2=w_2[t]}$



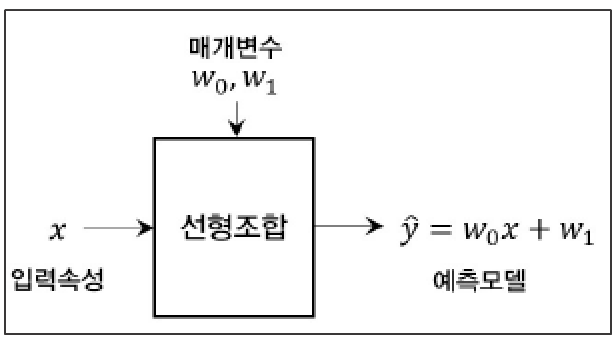
  - $w_0[t+1] = w_0[t] - \alpha \frac{2}{N} \sum_{n=0}^{N-1} x_{0,n}\left(w_0[t]x_{0,n} + w_1[t]x_{1,n} + w_2[t] - y_n\right)$
  - $w_1[t+1] = w_1[t] - \alpha \frac{2}{N} \sum_{n=0}^{N-1} x_{1,n}\left(w_0[t]x_{0,n} + w_1[t]x_{1,n} + w_2[t] - y_n\right)$
  - $w_2[t+1] = w_2[t] - \alpha \frac{2}{N} \sum_{n=0}^{N-1} \left(w_0[t]x_{0,n} + w_1[t]x_{1,n} + w_2[t] - y_n\right)$

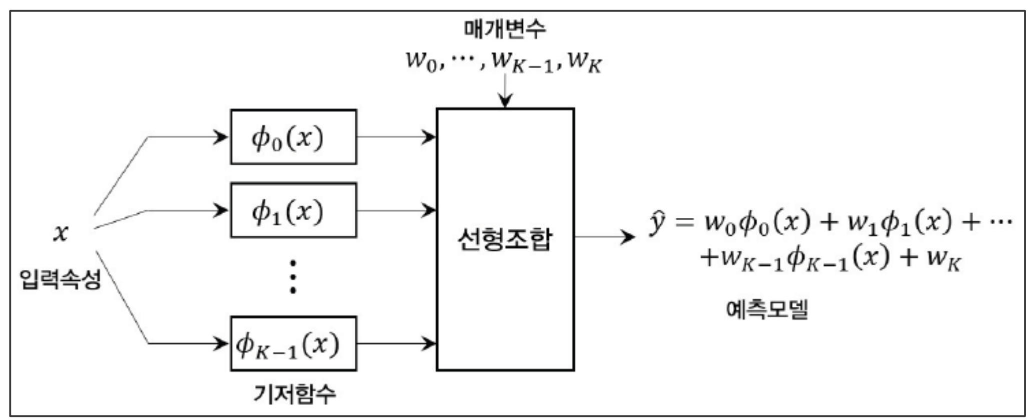
Check 필요!!

선형기저함수 모델



Non-Linear한 특성을 가지는 함수에 대한 표현을
선형함수의 조합으로 표현이 가능한가?





Non-Linear 함수들의 조합으로 표현하자!

선형기저함수 모델

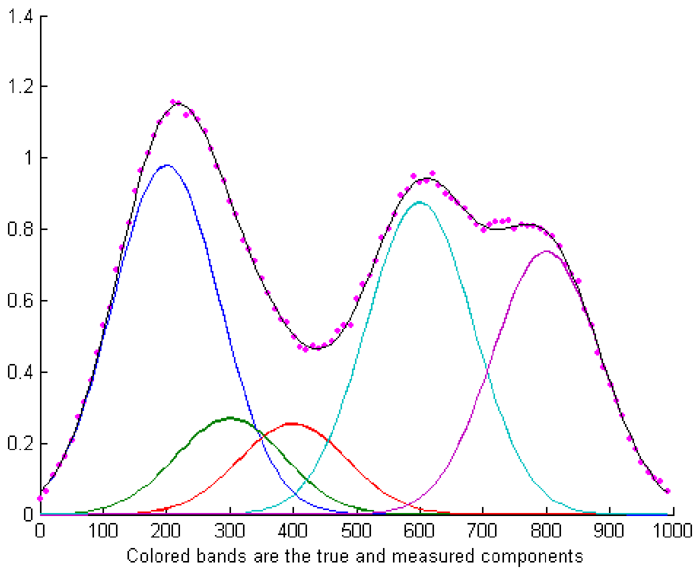
- 대표적 기저함수 모델
  - Polynomial basis function (다항식 기저함수)

$$\phi_k(x) = x^k$$

  - Gaussian basis function (가우시안 기저함수)

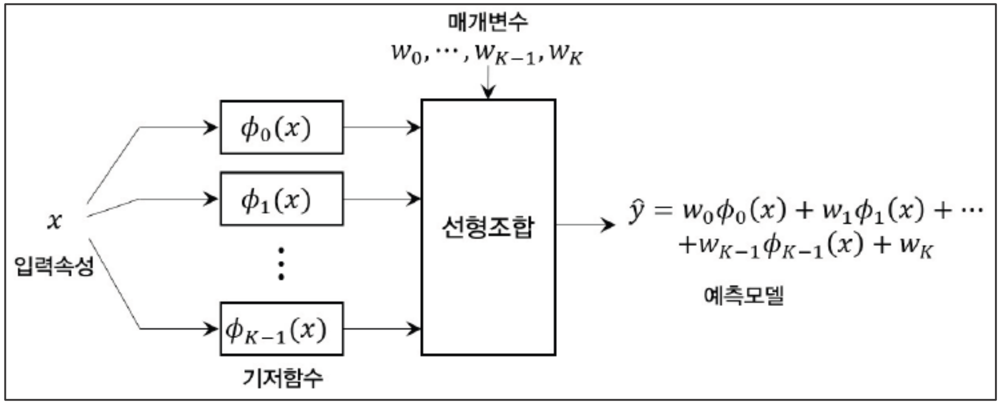
$$\phi_k(x) = e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}$$

가우시안 함수들의 조합으로
데이터의 특성 함수를 표현



선형기저함수 모델

- 선형기저함수 모델의 해석해
  - 입력이 K개인 선형회귀 모델과 동일한 방법 사용 가능



  - 예측 출력: 기저함수들의 선형 조합
    - $\phi_K = 1$: 행렬 표현의 편의를 위해 추가한 변수

$$\hat{y} = [w_0, w_1, \ldots, w_{K-1}, w_K]\begin{bmatrix}\phi_0(x)\\ \phi_1(x)\\ \vdots \\ \phi_{K-1}(x)\\ \phi_K(x)\end{bmatrix} = \mathbf{w}^T\boldsymbol{\phi}(x) \qquad \mathbf{w} = \begin{bmatrix}w_0\\ w_1\\ \vdots\\ w_{K-1}\\ w_K\end{bmatrix} \qquad \boldsymbol{\phi}(x) = \begin{bmatrix}\phi_0(x)\\ \phi_1(x)\\ \vdots \\ \phi_{K-1}(x)\\ \phi_K(x)\end{bmatrix}$$

선형기저함수 모델

- 선형기저함수 모델의 해석해
  - 비용함수: 평균제곱오차

$$\epsilon_{MSE} = \frac{1}{N}\sum_{n=0}^{N-1}(\hat{y}_n - y_n)^2 = \frac{1}{N}\sum_{n=0}^{N-1}(\boldsymbol{\phi}(x_n) - y_n)^2$$

$$\boldsymbol{\phi}(x_n) = \begin{bmatrix} \phi_0(x_n) \\ \phi_1(x_n) \\ \vdots \\ \phi_{K-1}(x_n) \\ \phi_K(x_n) \end{bmatrix}$$

  - N개 훈련 데이터 입력에 대한 행렬
    - (데이터 개수가 N개임, 각 개수마다 기저는 K개, 마지막 열은 모두 1 값임)

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_0(x_0) & \phi_1(x_0) & \cdots & \phi_{K-1}(x_0) & \phi_K(x_0) \\ \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_{K-1}(x_1) & \phi_K(x_1) \\ \vdots & \vdots & & \vdots & \vdots \\ \phi_0(x_{N-1}) & \phi_1(x_{N-1}) & \cdots & \phi_{K-1}(x_{N-1}) & \phi_K(x_{N-1}) \end{bmatrix}$$

  - N개 훈련 데이터 출력에 대한 벡터

$$\mathbf{y} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{N-1} \end{bmatrix}$$

선형기저함수 모델

- 선형기저함수 모델의 해석해
  - 최적 매개변수

$$\mathbf{w} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{y}$$

  - (cf.) 선형회귀 모델의 해석해

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

선형기저함수 모델

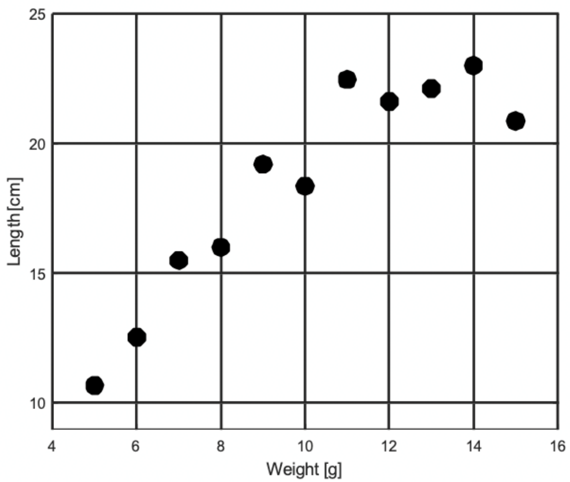
- 기저함수 개수 K와 각 기저함수 파라미터 설정
  - k번째 가우스 함수의 평균

$$\mu_k = x_{\min} + \frac{x_{\max} - x_{\min}}{K-1} k, \qquad k = 0,1,\dots,K-1$$

  - 모든 가우스 함수의 분산

$$\sigma = \frac{x_{\max} - x_{\min}}{K-1}$$

  - (예)

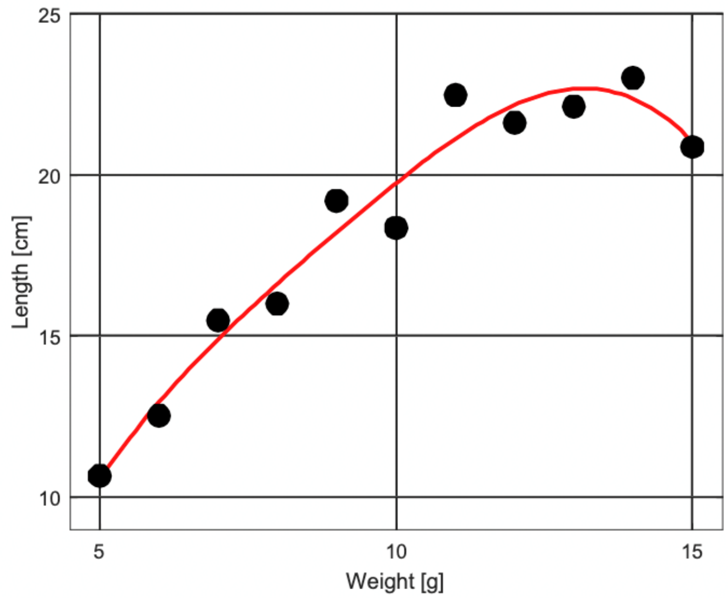


기저함수	매개변수	
	μ_k	σ
$\phi_0(x)$	5	5
$\phi_1(x)$	10	
$\phi_2(x)$	15	

선형기저함수 모델

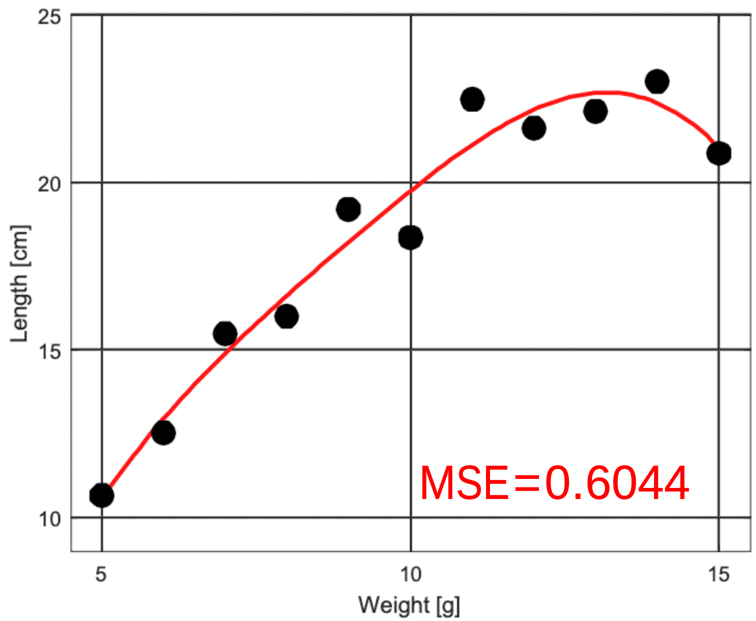
- 선형기저함수 회귀의 예
  - 최적 매개변수 및 회귀 모델, $K = 3$

$$\mathbf{w} = \begin{bmatrix} 27.02 \\ 3.46 \\ 39.08 \\ -23.82 \end{bmatrix} \qquad \hat{y} = 27.02e^{-\frac{1}{2}\left(\frac{x-5}{5}\right)^2} + 3.46e^{-\frac{1}{2}\left(\frac{x-10}{5}\right)^2} + 39.08e^{-\frac{1}{2}\left(\frac{x-15}{5}\right)^2} - 23.82$$

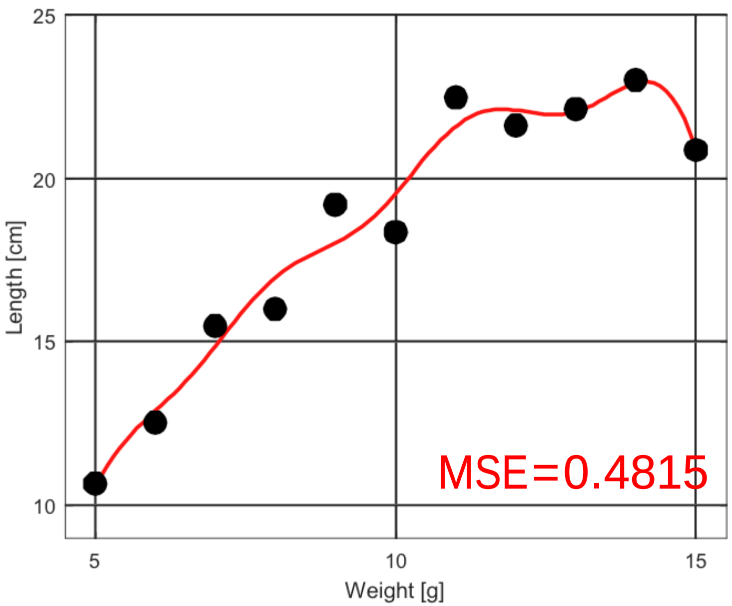


선형기저함수 모델

- K값에 따른 성능 비교



K=3



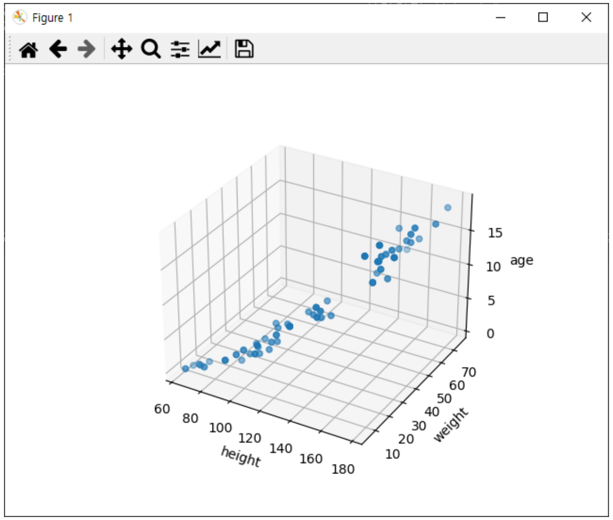
K=8

MSE가 작다고 무조건 좋은 것일까?

실습

- 실습데이터 (파일명: multiple_linear_regression_data.csv)
  - 데이터는 학생들의 키(cm)와 몸무게(kg)를 측정하고 나이를 표시한 것
  - 파일에서 첫번째 열은 키(cm), 두번째 열은 몸무게(kg), 세번째 열은 나이를 의미

- 실습 #1
  - 제공된 데이터 파일을 불러들여 x축은 키(cm), y축은 몸무게(kg), z축은 나이를 나타내는 3차원 공간에 각 데이터의 위치를 점으로 표시하시오.
  - 결과물: 코드, 매개변수
  - Hint) scatter 함수 사용



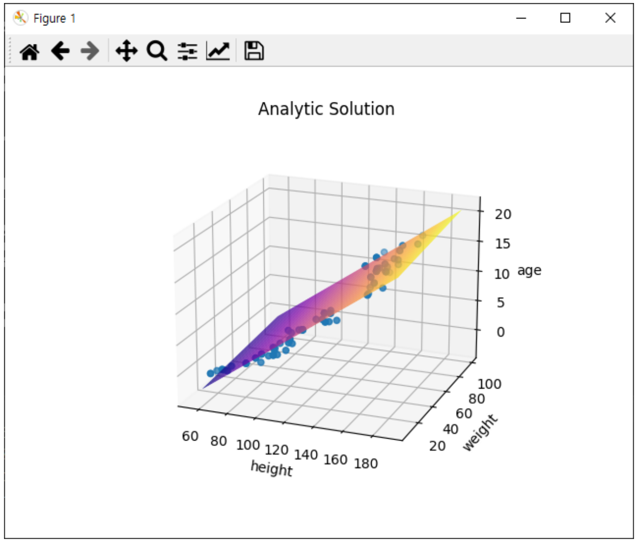
실습

- 실습 #2
  - 해석해로 구한 선형모델과 데이터를 한 그래프에 표시하라.
  - 필수요소: x축, y축, z축 이름, grid, legend
  - 결과물: 그래프
  - Hint) pseudo inverse matrix → numpy.linalg.pinv()

    평면 생성: linspace()로 height, weight 데이터 임의 생성 → meshgrid() 사용

    3D plot:

    ax = fig.add.subplot(projection=‘3d’), ax.plot_surface()



실습

- 실습 #3
  - 해석해로 구한 선형모델의 평균제곱오차를 구하라.

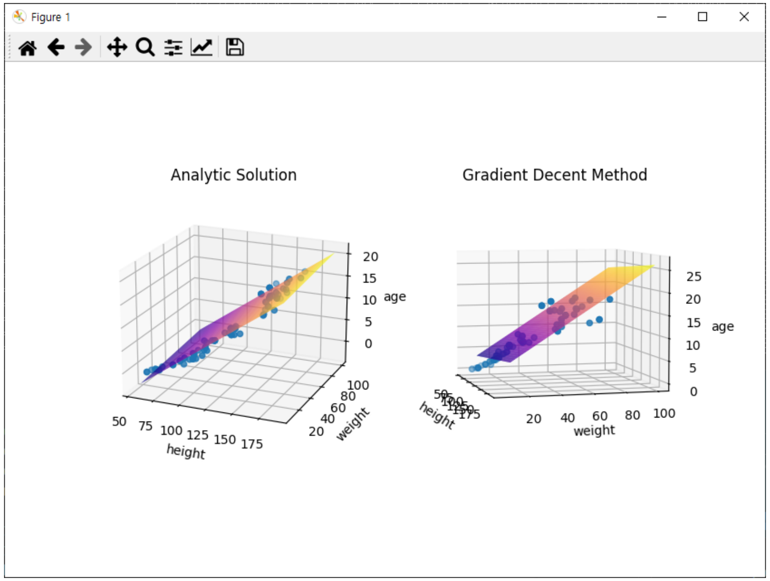
$$\epsilon_{MSE} = \frac{1}{N}\sum_{n=0}^{N-1}(\hat{y}_n - y_n)^2 = \frac{1}{N}\sum_{n=0}^{N-1}(\mathbf{w}^T\mathbf{x}_n - y_n)^2$$

  - 결과물: 코드, 평균제곱오차

실습

- 실습 #4
  - 경사하강법 프로그램을 이용해 최적 매개변수를 구하라.
    - 단, 경사하강법 외부 함수 사용 금지.
    - 단, 학습률, 초기값, 반복 회수는 임의로 정하여 사용하라.
  - 결과물: 코드
  - 결과물: 학습률, 초기값, 반복 횟수, 최종 평균제곱오차, 최적 매개변수

```
epoch:99700 =====================> W:[ 0.01793957  0.22764151 -0.12103336], gradient:[-0.00340941  0.00488573  0.28744971], mse:5.446940361853763
epoch:99800 =====================> W:[ 0.01794298  0.22763662 -0.1213208 ], gradient:[-0.00340933  0.00488561  0.2874427 ], mse:5.44676989653737
epoch:99900 =====================> W:[ 0.01794639  0.22763174 -0.12160824], gradient:[-0.00340925  0.00488549  0.28743569], mse:5.446599439408866
```



실습

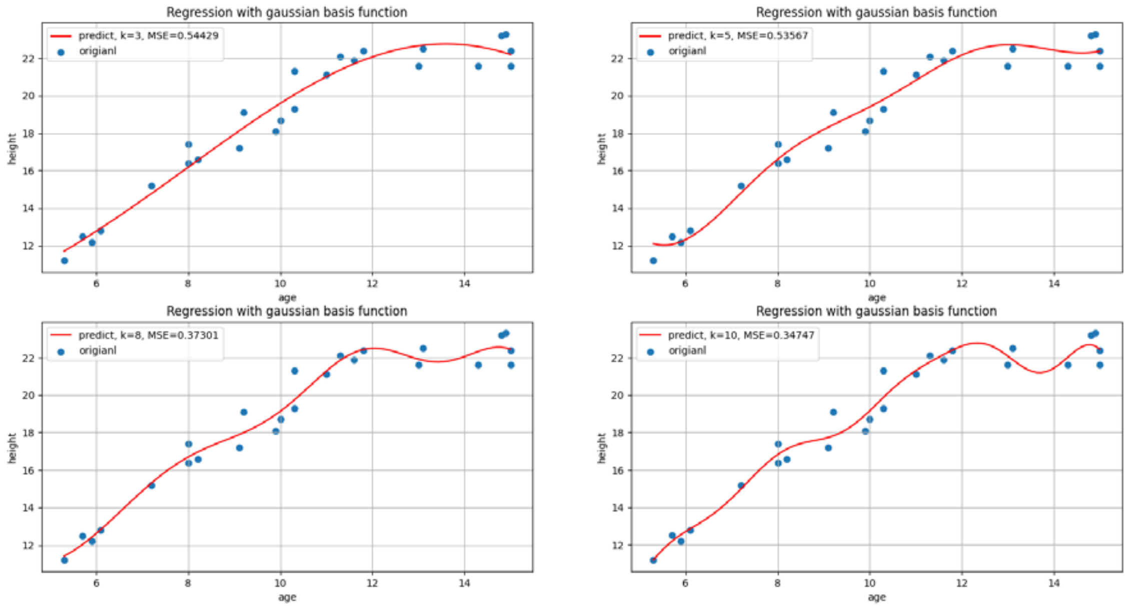
- 실습데이터 (파일명: lin_regression_data01.csv → 지난주 data)
  - 데이터는 유아들의 나이(개월)와 키(cm)를 측정한 것
  - 파일 형식은 쉼표로 구분된 데이터 파일
  - 파일에서 첫번째 열은 나이(개월), 두번째 열은 키(cm)를 의미

- 실습 #5
  - 주어진 데이터에 대해 K개의 가우스 함수를 이용한 선형 기저함수 회귀모델의 최적 매개변수(해석해)를 자동 계산하는 프로그램을 작성하고, K가 3, 5, 8일 때의 매개변수를 구하라.
  - 결과물: 코드, 매개변수

$$\mathbf{w} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\mathbf{y}$$

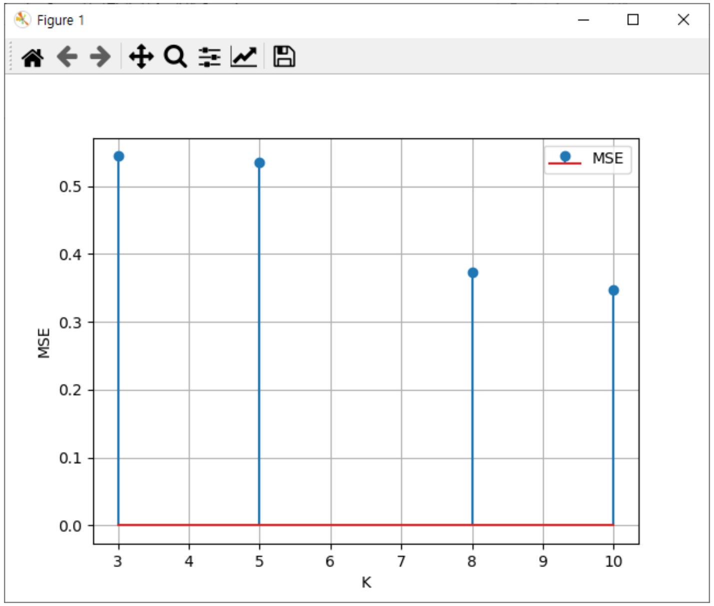
실습

- 실습 #6
  - K = 3, 5, 8, 10일 때, 훈련 데이터와 선형 기저함수 회귀 모델을 그래프에 표시하라.
  - 필수요소: x축, y축 이름, grid, legend
  - 결과물: 그래프



실습

- 실습 #7
  - 실습 #6에서 각 K개에 대한 평균제곱오차를 구하고 x축은 K값, y축은 평균제곱오차를 나타내는 2차원 그래프를 구하라.
  - 필수요소: x축, y축 이름, grid, legend
  - 결과물: 코드, 그래프



보고서 작성

- 각 주차에 해당하는 실습과제에 대해 **하나의 보고서로 작성**
  - 보고서 작성 형식은 hwp, word
  - **첫 페이지(표지)에 O주차, 실습과제 #O, 이름, 학번, 제출 날짜 기입**
  - 각 결과에 대한 해석은 **간단 명료**하게 서술
- 실행 Code는 보고서에 복사 붙여넣기 하여 넣기
  - **Code를 해석할 수 있는 주석 처리 필수**
- 결과는 해당 화면을 캡쳐하여 보고서에 포함
- 즉시 실행 가능한 .py 파일을 함께 업로드
  - (**Run만 눌러도 오류없이 돌아가는 코드**)
  - 파일명은 본인영문이니셜_학번.py로 할것

보고서 + .py파일 => 압축하여 학번_이름.zip으로 e-class에 제출