# 워크북

교과목명 : 머신 러닝

차시명: 2차시

◈ 담당교수: **장 필 훈**

⊙ 세부목차

- 제곱오류함수
- 라그랑주 승수법
- 편향분산분해

학습에 앞서

■ 학습개요

패턴인식의 가장 기본적인 선형회귀를 배운다. 먼저 선형함수를 정의하고 목표를 인식한다. 선형함수를 정의하는 데 기본인 기저함수에 관해 익히고 다양한 기저함수에 대해 배운다.
그 다음 선형회귀에서 손실함수를 더 자세히 살펴보고 최대가능도와 최소제곱오차를 자연스럽게 유도해내는 과정을 익힌다. 그 외에, 순차학습과 정규화된 최소제곱법에 관해 익힌다. 정규화된 최소제곱법에서는, 정규화항이 가지는 기하학적 의미를 배운다.
마지막으로 편향분산분해를 유도하는 배경에 관해 익히고, 다음시간에 편향분산분해과정을 자세히 익히도록 한다.

■ 학습목표

| | |
|---|---|
| 1 | 선형회귀의 정의와 목표를 이해한다. |
| 2 | 기저함수에 관해 배우고 대표적인 함수 몇가지를 이해한다. |

| | |
|---|---|
| 3 | 손실함수를 정의하고 최소제곱오차를 이용해서 더 자세히 살펴본다. |
| 4 | 정규화항의 기하학적 의미를 이해한다. |
| 5 | 편향분산분해의 배경을 이해한다. |

■ 주요용어

- 3개 정도의 용어와 간략한 설명(2줄 이내)을 적어주세요.

| 용어 | 해설 |
|---|---|
| 선형회귀 | 관측값 x_n과 이에 해당하는 표적값 t_n이 훈련집합으로 주어졌을 때 회귀모델의 목표는 새 변수 x의 표적값 t를 예측하는 것인데, 이때 회귀모델로 선형함수를 사용하면 선형회귀라고 한다. |
| 기저함수 | 선형회귀에서 입력변수에 다음과 같은 고정 비선형 함수들의 선형결합을 사용할 수 있다.<br>$y(x,w) = w_0 + \sum_{j=1}^{m-1} w_j \phi_j(x)$<br>이 때 $\phi_j(x)$를 기저함수(basis function)라고 한다. |
| 정규화항 | regularization term을 말함. normalization(정규화)와 혼동하기 쉽기 때문에 원서 그대로 표현하는 것이 좋다. 많은 수의 매개변수를 가진 모델들의 과적합문제를 어느 정도 조절하기 위해 매개변수의 크기 자체를 penalty로 주는 항을 오차함수에 추가한다. 이 때 추가된 항을 regularization term이라고 한다. |
| 편향-분산 트레이드오프 | 기대오류는 편향, 분산, 노이즈로 분해되는데, 유연한 모델은 낮은 편향값과 높은 분산값을 가지며, 엄격한 모델은 높은 편향값과 낮은 분산값을 가진다. 이 둘은 트레이드 오프 관계에 있다. |

## 학습하기

<선형회귀>

선형회귀는 입력이 주어졌을 때 그에 해당하는 타겟변수를 예측하는 것을 말합니다. 예를들어 다차원 입력이 주어졌을 때, 목표로 하는 타겟 스칼라 변수 하나를 예측해내는 함수 y를 추정해 내는 것입니다. 이때 함수 y는 어떤 형태라도 가능하지만, 특별히 선형함수만으로 제한을 두고 살펴보도록 하겠습니다.

선형회귀모델은 그래서 회귀함수를 다음과 같이 나타낼 수 있습니다.

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \cdots + w_D x_D$$

$$y(\mathbf{x}, \mathbf{w}) = w_o + \sum_{J=1}^{M-1} w_j \phi_j(\mathbf{x})$$

$$y(\mathbf{x}, \mathbf{w}) = \sum_{J=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})$$

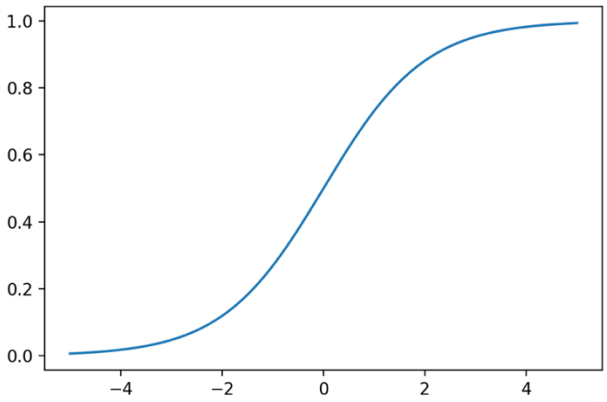
맨 위부터 동일한 식을 다른 형태로 나타낸 것입니다. 마지막 식은 $\phi_0(x)$를 1로 정의하고 w_0(=bias)까지 ∑에 포함시킨 형태입니다. $\phi$는 기저함수라고 하고 어떤 형태의 함수든 가능합니다.(하지만 보통 해석과 추정의 난이도까지 고려하여 미분가능한 함수로 제한합니다) 이 기저함수에 다항함수만을 사용하면 선형회귀가 됩니다. 물론, 이 기저함수로 비선형함수도 가능하고, 그렇게 사용하면 선형회귀형식을 사용해서 비선형적인 복잡한 데이터도 회귀가 가능하게 됩니다. 예를들어 다음과 같은 비선형함수도 기저함수로 가능합니다.

$$\phi_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{2s^2}\right)$$

$$\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right)$$

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

여기서 예로 든 함수 σ는 '시그모이드'라는 이름을 가진 자주 사용되는 함수입니다. x값에 대해 그래프를 그려보면 다음과 같습니다. x가 음의 무한대로 가면 0으로, 양의 무한대로 가면 1로 수렴합니다.




그러면, 회귀함수로 선형함수를 사용하기로 하고 목표값을 최대한 정확하게 추정해 내기로 했을 때, '얼마나 정확히 추정했느냐'의 척도가 필요합니다. 그래야 여러가지 회귀모델중에 어떤 모델이 가장 훌륭한지(=정확한지) 알 수 있으니까요. 원래 예측해내야 하는 값(정답)과 우리의 모델이 예측해내는 값과의 차이를 '손실'이라고 보통 부릅니다. 그리고 손실로는 보통 제곱손실을 사용합니다.

$$L(t, y(x)) = \{y(x) - t\}^2$$

여기서 y(x)가 우리의 모델, t가 목표값(target)입니다. 손실로 다른 형태도 물론 가능합니다.

제곱손실을 사용한다고 했을 때, 오차의 기댓값은 다음과 같습니다

$$E[L] = \iint \{y(x) - t\}^2 p(x,t) dx dt$$

오차를 최소화하는 y를 구해보겠습니다. 우선 위 함수를 y(x)에 대해 미분하고 극점을 찾겠습니다. 그 점이 오차의 기댓값을 최소화하겠죠.

$$\frac{\delta E[L]}{\delta y(x)} = 2\int \{y(x) - t\} p(x,t) dt = 0$$

$$\int y(x)p(x,t)dt \; - \int tp(x,t)dt = 0$$

$$y(x)\int p(x,t)dt = \int tp(x,t)dt = E_t[t|x]$$

좌변에서 p의 적분값은 1(확률이므로)이고 우측은 기댓값을 나타냅니다. 즉 평균을 뜻합니다.

따라서 '제곱오차함수의 손실을 최소화하려면 조건부평균값을 출력으로 주는 함수를 디자인하면 된다'는 뜻입니다.

이제 우리가 가질 수 있는 오차에 대해 분포를 가정해보겠습니다. 즉 다음을 가정합니다.

$$t = y(x,w) + \epsilon$$

$$\epsilon \sim N(0, \beta^{-1})$$

가정에 따라, t값은 y(x,w)를 평균으로 하는 가우시안 분포를 따르게 됩니다.

따라서 어떤 특정 데이터가 관측될 확률은 다음과 같고,

$$p(t|x,w,\beta) = \mathrm{N}(\mathrm{t}|\mathrm{y}(\mathrm{x},\mathrm{w}),\beta^{-1}), \qquad \beta = \frac{1}{\sigma^2}$$

여러개의 데이터를 관측하므로 가능도 함수는 다음과 같이 됩니다.(모든 가능도의 곱)

$$p(t|x,w,\beta) = \prod_{n=1}^{N} N(t_n|y(x_n,w),\beta^{-1})$$

가능도함수의 최댓값을 구하기 위해 로그를 취합니다. 로그는 단조증가함수이므로 로그를 취해도 최댓값을 가지는 x값이 변하지 않습니다.

$$\ln \prod_{n=1}^{N} N(t_n|y(x_n,w),\beta^{-1})$$

$$= -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,w) - t_n\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

위 식에서 첫번째 항이 바로 제곱합 오차함수와 같음을 볼 수 있습니다. 두번째, 세번째 항은 상수값이

므로 최대값을 가지는 x위치와 무관합니다.

즉, 노이즈로 가우시안 분포를 가정했을 때, 제곱합오류함수를 사용하는 것이 가능도함수를 최대화하는 y함수를 추정해 내는데 적절함을 알 수 있습니다. y대신 기저함수를 나타내는 표현을 사용해서, 제곱합오차항을 미분한 것을 다시 쓰면 다음과 같고,

$$\beta \sum_{n=1}^{N} \{t_n - w^T \phi(x_n)\} \phi(x_n)^T = 0$$

여기서 w를 계산해내면 선형회귀를 완성한 것입니다.

<편향>

제곱합오류함수를 선형회귀모델을 사용해서 다시 적어보면 다음과 같고,

$$E_D(w) = \frac{1}{2} \sum_{n=1}^{N} \left\{ t_n - w_0 - \sum_{j=1}^{M-1} w_j \phi_j(x_n) \right\}^2$$

w를 구하기 위해, w에 대해 손실함수를 미분한 식을 0으로 둡니다. 그래야 손실함수를 최소화하는 w값을 구할 수 있습니다. 아래 식의 전개를 보세요. (E_D의 ½은 계산의 편의를 위해 곱한 상수입니다. 우리의 관심은 E_D를 최대화하는 w값이므로 E에 상수를 곱해도 w값은 변하지 않습니다.)

$$\frac{d}{dw_0} \left[ \frac{1}{2} \sum_{n=1}^{N} \left\{ t_n - w_0 - \sum_{j=1}^{M-1} w_j \phi_j(x_n) \right\}^2 \right] = 0$$

$$\sum_{n=1}^{N} \left( t_n - w_0 - \sum_{j=1}^{M-1} w_j \phi_j(x_n) \right) (-1) = 0$$

$$\sum_{n=1}^{N} t_n - \sum_{n=1}^{N} w_0 - \sum_{n=1}^{N} \sum_{j=1}^{M-1} w_j \phi_j(x_n) = 0$$

$$\sum_{n=1}^{N} t_n - N w_0 - \sum_{j=1}^{M-1} \sum_{n=1}^{N} w_j \phi_j(x_n) = 0$$

$$w_0 = \frac{1}{N} \sum_{n=1}^{N} t_n - \sum_{j=1}^{M-1} \frac{1}{N} w_j \sum_{n=1}^{N} \phi_j(x_n) = \bar{t} - \sum_{j=1}^{M-1} w_j \overline{\phi_j}$$

마지막줄은 w_0를 다른 항들로 나타낸 것입니다. 즉, '편향(w_0)이 훈련집합의 타겟변수들 평균($\bar{t}$)과 기저함숫값 평균의 가중합 차이를 보상한다'는 것을 알려줍니다.

<순차학습>

위의 식은 모든 데이터를 한번에 처리하는 것을 가정합니다. 예를 들어 ∑항을 계산하려면 데이터의 모든 항에 대한 합을 모두 계산해야 합니다.(당연한 것입니다) 그런데 데이터가 너무 크면 이런 계산을 한번에 할 수가 없습니다. 그럴때는 순차적으로 계산하는 방법을 취합니다. 이때의 조건은, 새로운 데이터가 추가되었을 때, 기존에 이미 계산된 w를 새로운 데이터만으로 업데이트할 수 있을때 가능한가이고, 결국 '각각의 데이터가 가지는 오류값의 **합**이 전체데이터의 오류값과 같을때'입니다. 비선형결합이거나 곱이 포함된 복잡한 형태라면 불가능하거나 매우 어렵다는 뜻입니다. 식으로 나타내면 다음과 같습니다

$$w^{(\tau+1)} = w^{\tau} - \mu \nabla E_n$$

$$w^{(\tau+1)} = w^{\tau} - \mu\left(t_n - w^{(\tau)T}\phi_n\right)\phi_n$$

여기서 tau는 특정 시점을 나타냅니다. 즉 w^tau 에서 새로운 데이터를 반영한 새로운 w는 w^(tau+1)로 나타낸 것입니다.

<정규화항>

이제 오차함수에 정규화항을 포함해보겠습니다. 정규화항은 overfitting을 막는 대표적인 방법으로서, 오차함수에 최적화된 회귀함수가 지나치게 오차함수에 적합되어 학습데이터에 대해서는 예측력이 매우 뛰어나지만, 실제 데이터가 들어왔을 때 예측력이 형편없이 떨어지는 현상을 막습니다. 오차함수가 정의되면 이론적으로 (모델의 표현력이 충분하기만 하다면) 오차를 0으로 만들수 있기 때문에 그렇게 되는 것을 막습니다. 이것은 어떻게 보면 '오차함수' 자체가 가지는 한계라고 볼수도 있습니다. 오차함수가 실제로 완벽한 척도라면 오차 0이 나왔을 때 최고의 예측력을 가져야 하지만, 우리의 오차함수도 결국 우리가 디자인한 척도일 뿐이라 그렇게 될 수는 없습니다.

정규화항으로는 계수(w)의 제곱합을 많이 사용합니다.

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$$

$$E_w(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$$

위에서 마지막항은 정규화항이 포함된 오차함수입니다. 정규화항은 제곱이 아닌것(예: 계수의 절댓값의 합)도 가능하므로 더 일반적인 형태로는 다음과 같습니다.

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$

여기서 람다는 정규화항의 계수로서 그때그때 실험적으로 결과를 보고 조절합니다. 정답은 없고 우리가 얻고자 하는 모델의 특성을 이를 통해 정해줄 수 있습니다. q=1일때는 lasso라는 이름이 있습니다.

<라그랑주 승수법>

제약조건을 전제하고 식의 최댓값이나 최솟값을 구할 때는 라그랑주 승수법을 가장 많이 사용합니다. 우리가 특정 식의 최댓값이나 최솟값을 찾으려면 미분한 함수가 0이 되는 지점(stationary point)를 찾는데, 그 성질을 이용합니다. 좀 더 formal하게 보자면, 연속미분가능함수 f, g가 주어졌을 때 g=0을 제약조건으로 f의 최댓값이나 최솟값을 찾으려면 $\nabla f(x) = \lambda \nabla g(x),\ g(x) = 0$ 을 풉니다. 라그랑지안 함수 $\mathcal{L}(x, \lambda) = f(x) - \lambda g(x)$ 를 이용하면 식을 다루기가 쉽습니다.

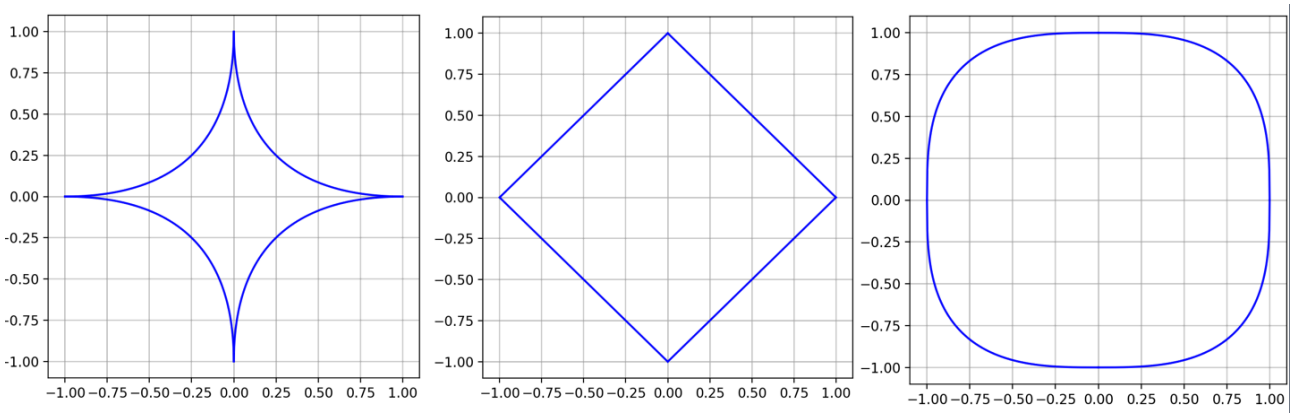
우리가 이 관점에서 정규화항이 포함된 오차 함수를 관찰하면 다음 제약 조건하에서 제곱합오차를 최소화하는 것임을 알 수 있습니다.

$$\sum_{j=1}^{M} |w_j|^q \leq \eta$$

라그랑지안으로 쓰면 다음과 같습니다.

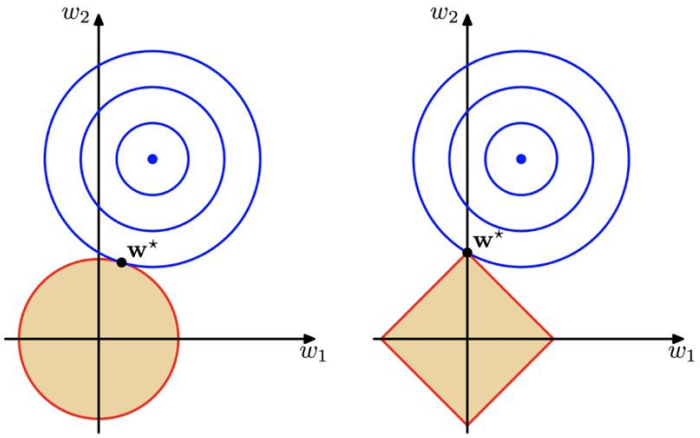
$$\mathcal{L}(w, \lambda) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\left(\sum_{j=1}^{M}|w_j|^q - \eta\right)$$

제약조건을 q값에 따라 그래프로 그려보면 다음과 같습니다.



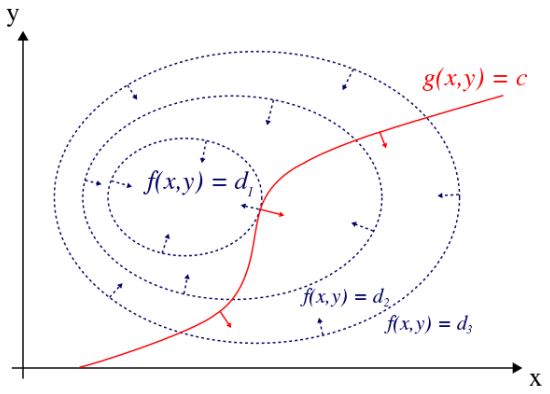
왼쪽부터 차례로 q가 $\frac{1}{2}$, 1, 2일때 입니다.

이 제약조건하에서 제곱합오차가 최소인 지점을 고르면 다음과 같은 w를 가지게 됩니다.




[Bishop. p146]

위에서 그래프가 만나는 지점(w*)이 아래 라그랑주 승수법의 개념을 설명하는 그림에서 두 함수가 서로 접하는 부분입니다.




[Lagrange multiplier © https://commons.wikimedia.org/wiki/File:LagrangeMultipliers2D.svg]

q=1인 경우는 w들이 sparse한 값을 가짐을 그림을 통해 관찰할 수 있습니다. (w가 다차원임을 잊지 마세요. 다차원을 구성하는 하나하나의 좌표들이 모두 값을 가지기보다 어떤것은 값을 가지고 어떤것은 0이 된다는 뜻입니다. 그림에서 w*의 좌표에 주목하세요)

<편향분산분해>

지금까지는 기저함수의 수와 형태를 고정하고 살펴 봤습니다. 기저함수를 너무 많이 하면 과적합, 너무 적게 하면 과소적합의 문제가 있습니다. 그래서 보통, 기저함수를 많이 써서 모델의 표현력을 높이고, 정규화항을 사용해서 과적합을 막는 방법을 사용합니다. (그때는 정규화항의 계수(lambda)를 정해야 하는 문제가 발생합니다.)

모델복잡도에 관해서는 편향–분산 트레이드오프 관계를 살펴봄으로써 더 깊은 이해를 얻을 수 있습니다. 우선 기대제곱오류는 다음과 같이 쓸 수 있습니다.

$$\mathbb{E}[L] = \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t$$

적분 안의 항을 정리하면,

$$\begin{aligned} \{y(\mathbf{x}) - t\}^2 &= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}] + \mathbb{E}[t|\mathbf{x}] - t\}^2 \\ &= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 + 2\{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\} + \{\mathbb{E}[t|\mathbf{x}] - t\}^2 \end{aligned}$$

이 식을 t에 대해 적분하면 교차항이 사라지고 다음 손실함수를 얻습니다.

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x})\, \mathrm{d}\mathbf{x} + \int \{\mathbb{E}[t|\mathbf{x}] - t\}^2 p(\mathbf{x})\, \mathrm{d}\mathbf{x}.$$

첫째항에서, y(x)가 기댓값, 즉 평균일때 손실함수가 최소가 됨을 볼 수 있습니다. 앞서 얻은 결론과 같습니다. 두번째 항은 y와 상관이 없으므로 우리가 아무리 y를 잘 추정해내도 줄일 수 없습니다. 데이터가 가지고 있는 내내적 변동성을 표현하는 것입니다. 내재적 변동성의 원인은 어떤것이라도 가능하겠지만, 단순히 노이즈라고 생각하시면 됩니다. 관측과정에서 생길수도 있고, 데이터 자체가 이미 가지고 있을수도 있지만 어차피 정복할 수 없는 오류이기 때문에 원인은 중요하지 않습니다.

무한히 많은 데이터포인트를 가진 데이터 D의 분포를 추정해내고자 할때, 무한한 데이터를 모두 관찰

할 수는 없으므로 여러번 데이터집합을 추출해서 y를 추정하는 과정을 반복합니다. 이때 제곱오류는 여러번 시행의 평균값으로 정의됩니다. (다음시간에 계속하겠습니다. )

## 연습문제

1. (OX 문제) $y = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(x)$에서 $\phi$가 선형이 아니면 선형회귀문제가 아니다.

   X 기저함수는 비선형함수일 수 있다.

2. (OX 문제) 입력공간을 여러 부분으로 나누고 각 부분에 대해 서로 다른 다항식을 이용하여 피팅한 함수를 스플라인 함수라고 한다.

   O 스플라인 함수에 대한 설명이다.

3. (OX 문제) 어떤 손실함수를 사용하든지 최적의 예측값은 조건부 평균이다.

   X 제곱손실함수를 사용할 때 그렇다.

4. (OX 문제) 정규화항의 계수 $\lambda$가 너무 커지면 편향은 작아지고 분산은 커진다.

   X 분산이 작아지고 편향이 커진다.

## 정리하기

선형회귀는 기저함수들의 선형결합함수를 통해 타겟변수를 예측하는 것이 목표이다.

2. 기저함수는 비선형 함수들을 사용할 수 있어서, 선형결합을 통해서도 복잡한 형태를 표현하는 것이 가능하다. (아래에서 phi가 기저함수)

$$y(x, w) = \sum_{j=0}^{M-1} w_j \, \phi_j(x) = w^T \phi(x)$$

3. 시그모이드함수, 가우시안함수등을 기저함수로 사용할 수 있다.

4. 제곱손실함수를 최소화 하는 예측값은 조건부 평균이다.

5. 기저함수를 어떤 형태를 선택하든지, 제곱합 오차함수를 통해 w를 알아낼 수 있다.

6. 제곱합 오차함수를 전개함으로써 $w_0$가 무엇을 보정하는지 알 수 있다.

7. 큰 데이터는 w를 바로 계산하기 힘들기 때문에 순차학습을 통해 알아낸다.

8. 오차함수에 정규화항을 넣어서 과적합을 막을 수 있다.

9. 정규화항의 차수에 따라 추정되는 w의 성질이 다르다.(lasso -> sparse )

10. 손실함수를 적절히 전개하여 편향과 분산으로 나눠볼 수 있다.

## 참고하기

Bishop, C. M. "Bishop-Pattern Recognition and Machine Learning-Springer 2006." Antimicrob. Agents Chemother (2014): 03728-14.

## 다음 차시 예고

편향분산분해를 마무리하고, 베이지안 선형회귀를 살펴보겠습니다.
선형회귀를 마치고 선형분류를 배워보겠습니다.