구건모

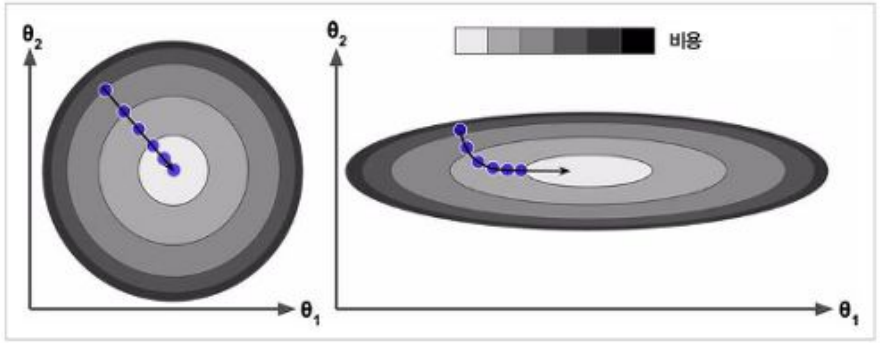
스케일을 맞춰줘야 하는 이유
- 왠만하면 성능이 좋아진다.

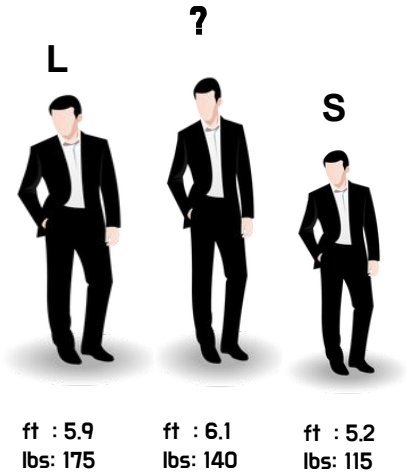
+zoom
표준화를 통하여 무엇이 더 중요한지 컴퓨터가 인식하게 도와줍니다.
Keyword : 안정화 최적화

대개 스케일링을 하면 성능이 좋아집니다. 당연하지만 당연하지 않은 **왜?** 에 대해서 생각해봤습니다.

1. 각 특성간의 벡터 공간상의 거리를 맞춰 준다.
2. 최적화 알고리즘이 목적값을 찾기위한 경로를 쉽게 만들어준다.

그림 4-7 특성 스케일에 따른 경사 하강법





1) https://box-world.tistory.com/10
2) https://bkshin.tistory.com/entry/%EB%A8%B8%EC%8B%A0%EB%9F%AC%EB%8B%9D-8-Feature-Scaling-Feature-Selection

+zoom
정렬된 데이터라면 한 쪽으로 편향된다.

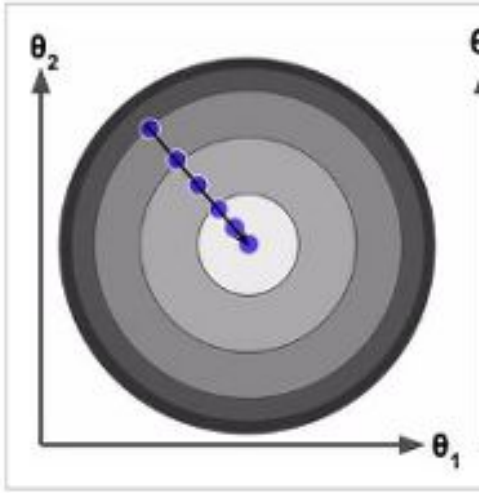
구건모

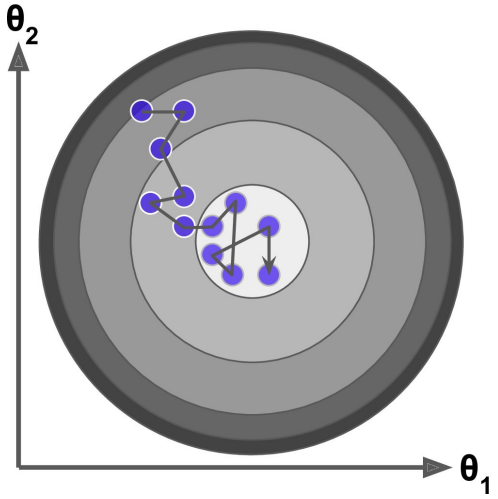
확률적 경사 하강법, Epoch마다 훈련 샘플을 섞어야 하는 이유

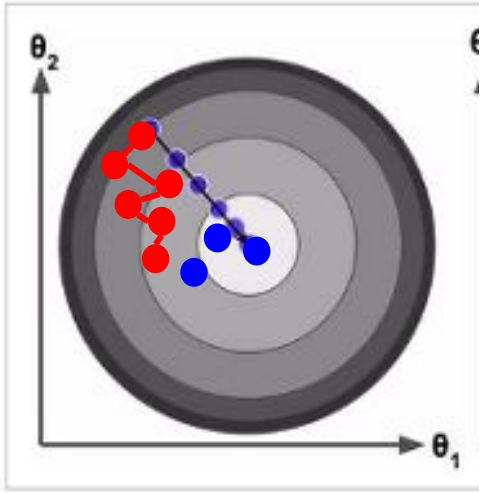
부제: 미니배치 경사 하강법이 아니다.

P174- 확률적 경사 하강법을 사용할 때 훈련 샘플이 IID(independent and identically distributed)를 만족해야 평균적으로 파라미터가 전역 최적점을 향해 진행한다고 보장할 수 있습니다.

[
[img1, 'cat'],
[img2, 'cat'],
[img3, 'cat'],
[img4, 'cat'],
[img5, 'dog'],
[img6, 'dog'],
[img7, 'dog'],
[img8, 'dog']
]







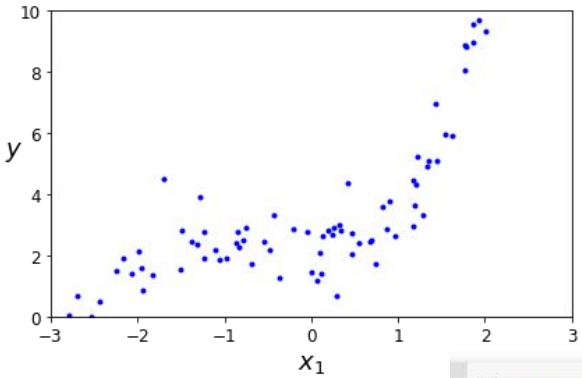
1)

Polynomial Feature를 사용하면 왜 잘 될까?

Polynomial Feature의 역할은 중요한 특성을 만들어 주는 것이고

실제로 잘 작동하는 이유는 학습이 되면서 덜 중요한 특성이 사라졌기 때문이다.

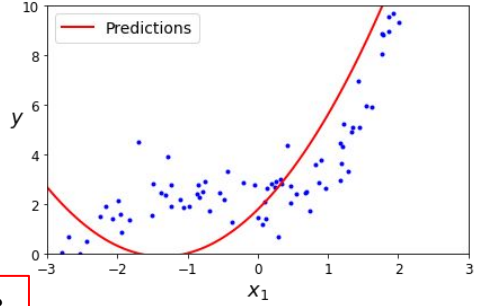
데이터: $X^3 + X^2 + 2$



주어진 특성:X

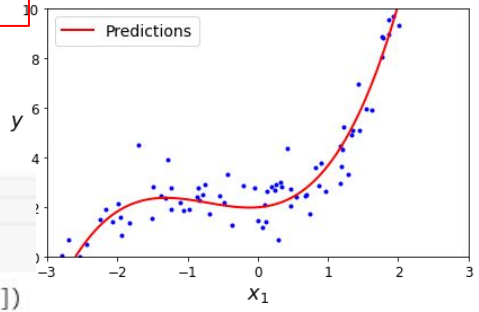
X, X^2

degree = 2



X, X^2, X^3

degree = 3



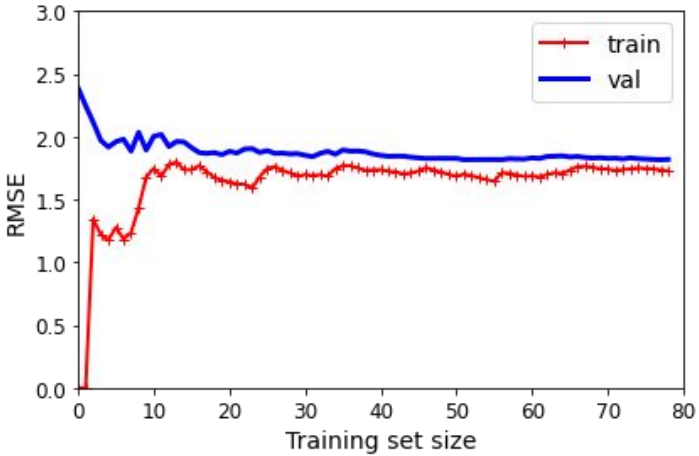

```
lin_reg.coef
--NORMAL--
array([[0.21770159, 1.00127472, 0.46749475]])
```

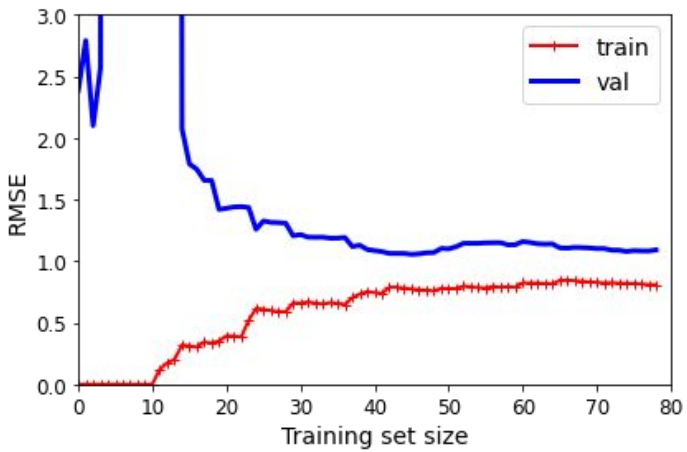
X, X^2, X^3

1)

실제로 저러한 X^2의 특성들을 추출할 수 있을까?

사실 학습전에는 판단하기에 불가능하다. 학습 곡선을 보면서 과대 적합일 경우, 과소 적합임을 판단해서 적절하게 특성 추출을 하는 것이 중요하다.





1)

 # 구건모

MSE를 사용하면 무조건 Cost Function이 Convex한가?

+zoom
무조건 Convex한 Cost Function은
linear regression에 대한 것이다.

P167 - 두 점을 이은 선분이 두 점 사이에서 항상 곡선 위에 위치할 경우를 볼록 함수, 아래에 위치할 경우 오목 함수라고 합니다.
그렇다면 왜 볼록함수인 MSE를 학습 시킬 때 local minima에 수렴 하는 경우가 있을까?

아래 링크 답변)
사람들은 종종 신경망의 Loss 함수가 Convex하지 않다고 말할 때 그것은 MSE를 말하는 것이 아닙니다. MSE는 항상 Convex합니다.

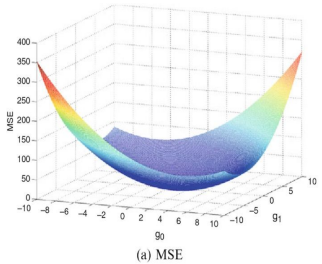
우리가 모델을 구현할 때 컨트롤 하는 것이 무엇인지 생각을 해봐야 합니다. 우리는 모델 파라미터를 컨트롤 합니다.
즉, 우리가 궁금해야 하는 것은 모델 파라미터와 손실 함수의 볼록성입니다.

y^ = f(w, x) # **w**:model parameter, **x**: input example

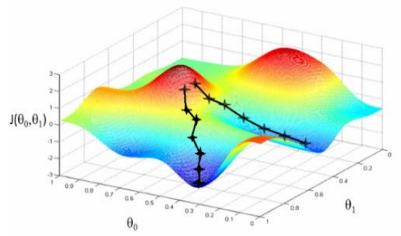
g(x, y, w) = L(y^, y) = L(f(w, x), y) **# MSE = L(y^, y)**

일반적으로 **f**(신경망)는 볼록하지 않고 그러므로 **g**도 볼록하지 않다.

Local Minima에 빠진 것은 MSE가 아니라 가중치다.
> 책에서도 선형 회귀에 대해서만 볼록 함수라고 한다.



(a) MSE

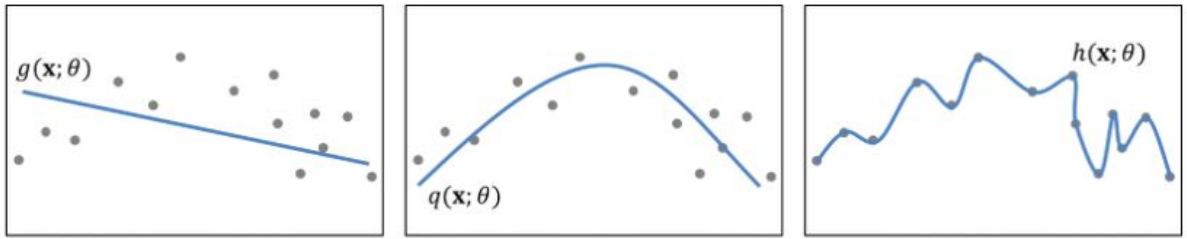


1) https://www.quora.com/MSE-is-a-convex-loss-func-for-linear-and-logistic-regression-how-come-it-isnt-for-NN-and-backprop-i-e-local-optimums

조성진

편향 분산 트레이드 오프

+zoom
모델을 잘못 선택하여 편향이 커지면
분산값이 줄어들고 그만큼 과소적합된다.



(a) Underfit (b) Ideal fit (c) Overfit

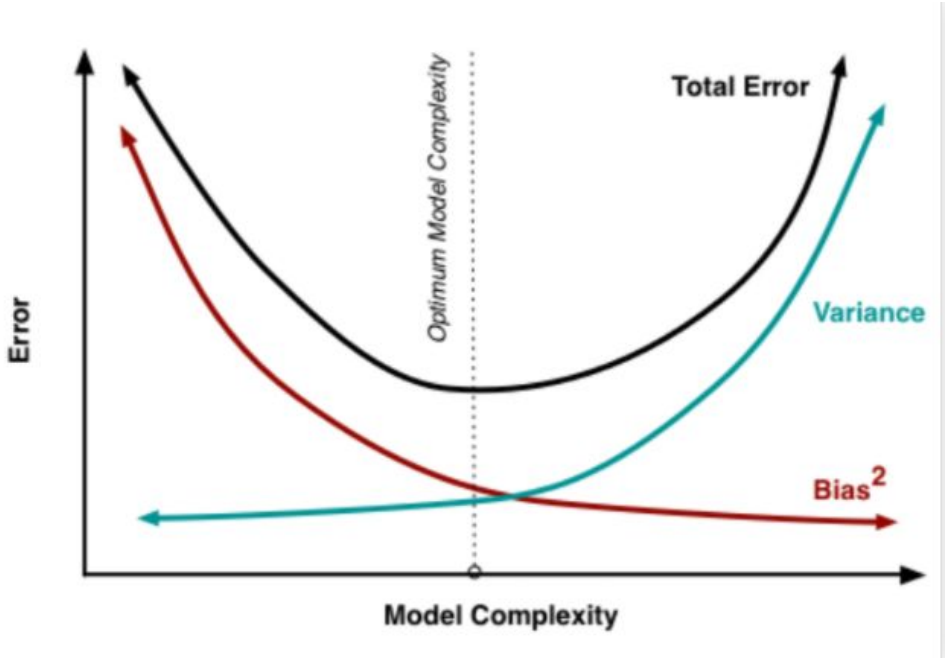
MSE = (bias)² + variance + σ^2

모델 복잡도가 높다 = 구불구불한게 많다

https://untitledtblog.tistory.com/143

조성진

편향 분산 트레이드 오프



https://untitledtblog.tistory.com/143

조성진

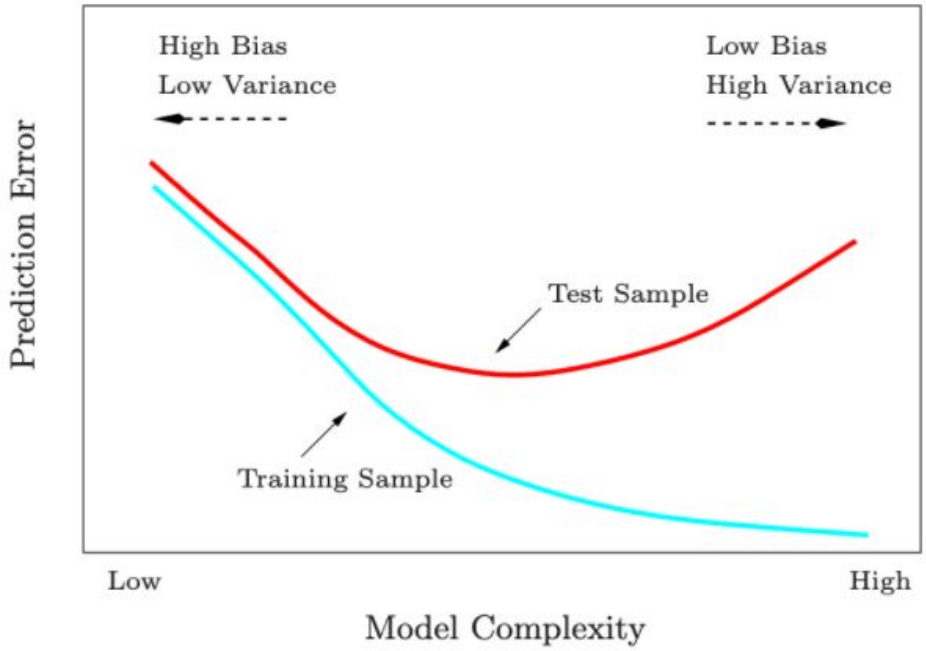
편향 분산 트레이드 오프

모델이 복잡하다? 덜 복잡한 모델을 사용

변수가 많다? 변수를 줄인다

데이터가 수상하다? 데이터를 정제한다

학습이 과하면? 중간에 멈춘다



https://bahnsville.tistory.com/1199

김유지

학습 방법이 Monte Carlo 알고리즘?

+zoom
몬테카를로 : 정답이 아닐 수 있지만 시간이 적게 걸림
라스베가스 : 항상 정답이지만 시간이 무한대로 갈 수 있음

P 171 수렴율

특정 오차 범위 c 안에서 최적의 솔루션에 도달하기 위해서는 O(1/c)의 반복이 걸릴 수 있다
즉, 학습 방법이 monte carlo 알고리즘이라고 생각할 수 있다

Monte Carlo 알고리즘이란?
특정 확률 (예를 들면 99%의 정확도)를 달성하고 싶으면, c는 1 - 0.99의 값이 된다
그러면 monte carlo 방법론에서 1 - c의 정확도를 도달하기 위해서는 아래 횟수만큼 반복 필요
따라서 O(1/c)의 반복 걸린다 (아래 델타가 여기에서의 c)

$$\binom{n}{2} log(\frac{1}{\delta})$$

1)

김유지

특성이란?

P 164 이외 다수

특성이란 정확하게 무엇인가요?
특성과 클래스의 다른 점은?

MNIST에서는 특성과 클래스가 같은 것 같고,
타이타닉에서는 embarked 같은 것들이 특성이고 생존 유무를 따지는 이진 분류

+zoom
MNIST : 28 x 28 각 pixel은 특성이고
0~9인 각 결과는 클래스이다.

개별을 특성으로 봐야하는가?
28 x 28 자체를 특성으로 봐야하는가?

구건모

정규방정식 유도 과정

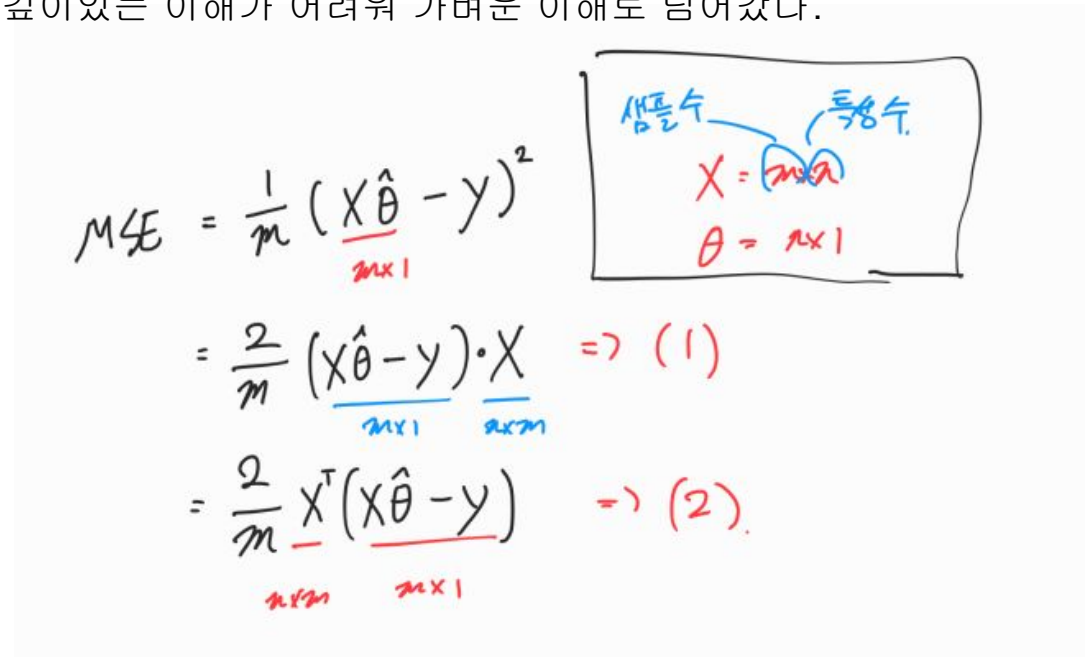
P 162, P 169 참고

(2)번 식으로 유도되는 과정이 궁금합니다.

스칼라 전치 공식을 사용해서 저렇게 되는 건가요?

+zoom

이번 장에서 나오는 여러가지 수식에 대해서는
깊이있는 이해가 어려워 가벼운 이해로 넘어갔다.



1)

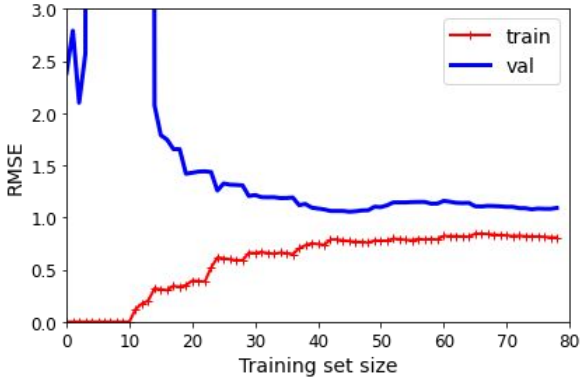
2차원 데이터셋에 10차원 모델, 학습 데이터가 늘어난다고 과대적합이 사라질까?

그러면 적은 데이터셋을 여러번 학습시키는 것으로 가중치의 변화가 불가능 할까? -> **테스트 불가능**

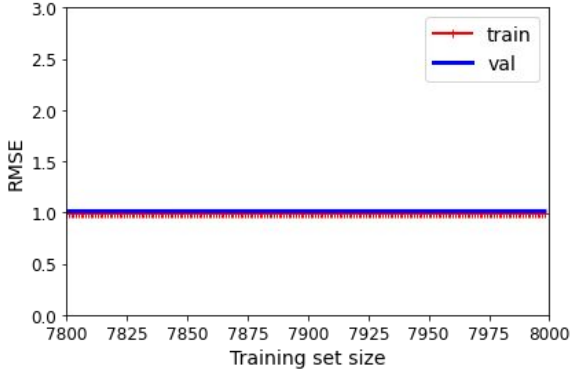
+zoom

학습 데이터가 늘어나면 좋다

같은 데이터를 반복하면 과적합이 일어난다.



m = 100



m = 10000

1)

 # 조성진

외부 메모리 학습 지원? = 온라인 학습

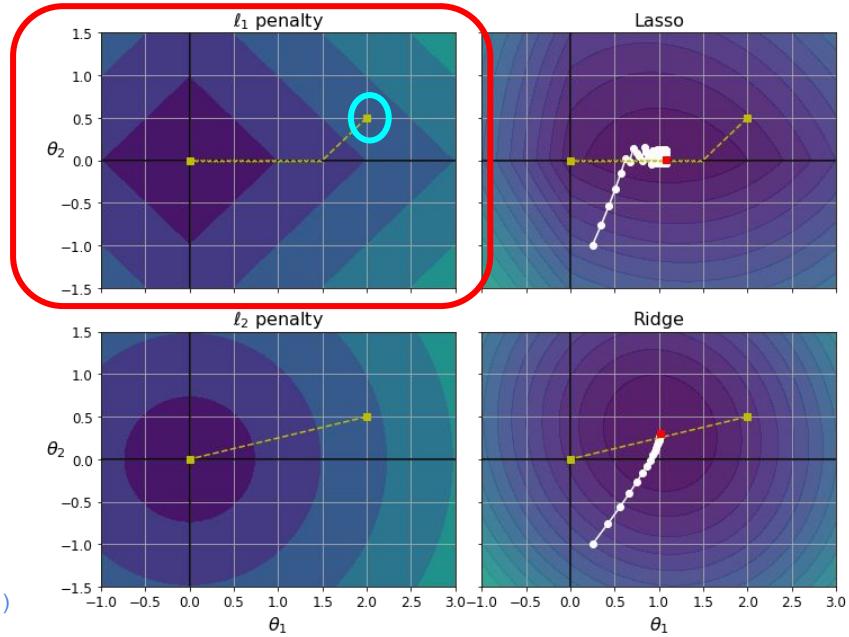
확률적 경사 하강법, 미니배치 경사 하강법이 외부 메모리 학습 지원이 가능한 이유는?

+zoom
전체가 아닌 일부를 통하여 학습하기 때문이다.

구건모

이 그래프를 여러분들이 어떻게 이해했는지 궁금합니다.

P-188 L1의 그레이디언트는 0에서 정의되지 않기 때문에 진동이 조금 있습니다.



1)

+zoom

라쏘 회귀 -> 가중치 중 덜 중요한 특성을 0으로 만든다

https://bskyvision.com/193

$$J(\mathbf{\theta}) = \mathbf{MAE}\ (\mathbf{\theta}) + \alpha \sum_{i=1}^{n} |\ \theta_i\ |$$

$$J(\mathbf{\theta}) = \mathrm{MSE}\,(\mathbf{\theta}) + \alpha \sum_{i=1}^{n} |\ \theta_i\ |$$

$$g(\mathbf{\theta}, J) = \nabla_{\mathbf{\theta}}\ \mathrm{MSE}\,(\mathbf{\theta}) + \alpha \begin{pmatrix} \mathrm{sign}\,(\theta_1) \\ \mathrm{sign}\,(\theta_2) \\ \vdots \\ \mathrm{sign}\,(\theta_n) \end{pmatrix} \quad \text{where } \mathrm{sign}\,(\theta_i) = \begin{cases} -1 & \text{if } \theta_i < 0 \\ 0 & \text{if } \theta_i = 0 \\ +1 & \text{if } \theta_i > 0 \end{cases}$$