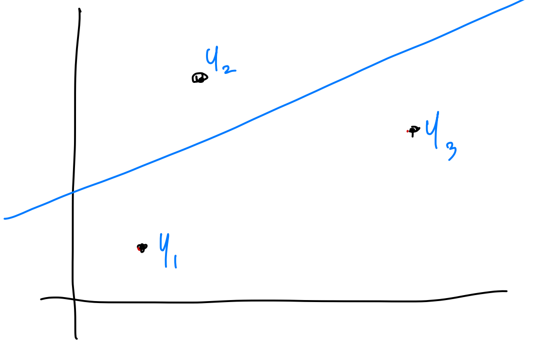
Linear Regression

$y = \theta_1 x + \theta_0$



$$\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_0 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

$$X \theta = y$$

$$\boxed{\theta = X^{-1} y}$$

X^{-1} 존재 X.

normal equation

$$\boxed{(X^T X)\theta = X^T y}$$

psedo-inverse of X

$$\boxed{\theta = X^{\dagger} y}$$

Normal Equation?

$$\Rightarrow \quad MSE = \frac{1}{m}\sum_{i=1}^{m}\underbrace{\left(\theta^T x^{(i)} - y^{(i)}\right)^2}_{l_2\text{-norm}}$$

MSE를 최소화 시키는 θ를 찾는 공식.

$$(X^T X)\hat{\theta} = X^T y$$

$$\boxed{\hat{\theta} = (X^T X)^{-1} X^T y}$$

psedo-inverse?

definition of A^{-1}) $A: n \times n$.

$$A^{-1}A = I = AA^{-1}.$$

$$Ax = y \Rightarrow \underbrace{x = A^{-1}y}_{\text{unique solution}}.$$

but psedo-inverse.

$$\boxed{Ax = y \Rightarrow x = A^{+}y.}$$

(If $A: 5 \times 4$, 무수히 많은 해)

psedo-inverse 구하는 방법?

SVD 이용

$(A = U\Sigma V^{T})$.

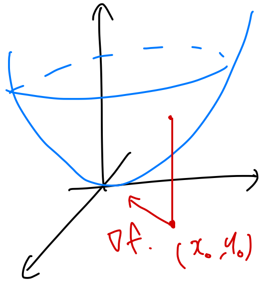
$$A^{+} = V\Sigma^{-1}U^{T}.$$

SVD의 장점은 차원 축소

+

psedo-inverse 구하기도 좋음.

Gradient Descent.



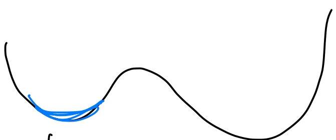
∇f (
방향 : 가장 크게 변화하는 방향.

크기 : 변화율.

minimum에 가까워질수록
알아서 줄어듦.

(but 추가로 step size 조절하는 경우도 많음).

Gradient descent의 문제점?



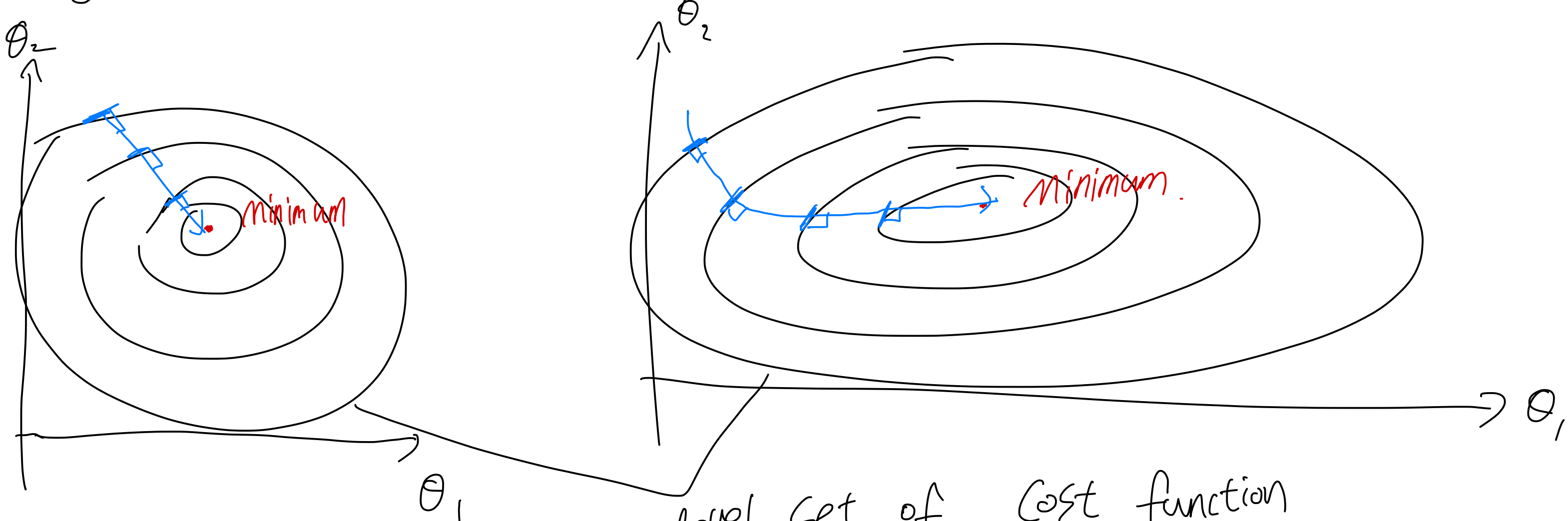
local minimum에
갇힐 수 있음.

⇒ Loss Function은 가능한한 Convex 하게 정의.

convexity test 기법 존재.

L_1 loss
L_2 loss
Hinge loss
Cross-entropy
⋮
⇒ All convex.

· 전처리 (standard scaler) 해야 함.



level set of cost function.

ex). $J(\theta) = 1$

gradient 구하는 방법.

- 배치
- 미니배치
- 확률적.

모델 파라미터 최적화 (θ 구하기).

- 해석적 (공식) - normal equation, SVD
- 조금씩 바꾸면서 - gradient descent.

normal equation vs gradient descent.

SVD

- 데이터 많으면 느림.
- 데이터 추가되면 처음부터 다시 훈련해야 함.
- but 식 하나로 깔끔. hyper parameter 없음.

- 데이터 많아도 빠름
- 데이터 추가돼도 업데이트 가능.
- but 훈련 필요 hyper parameter 필요.

다항회귀 (비선형 데이터)

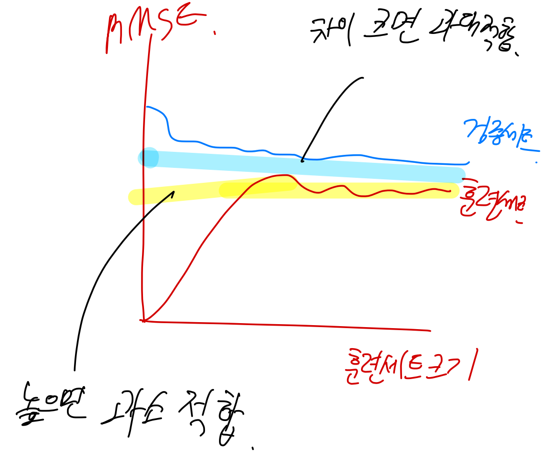
너무 낮은 차원 → 과소 적합.

필요 이상으로 높은 차원 → 과대 적합.

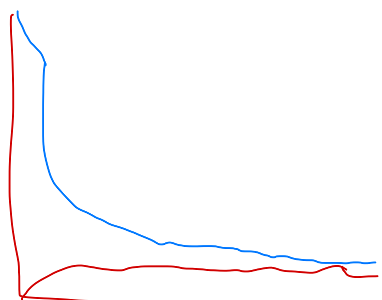
과대적합 확인하는 방법?

차수 결정하는 방법? → 학습곡선

시각화 굿.



ideal



차수는 입력해야 함.

(but 경우의 수 별로 없을듯...)

특성 n

차원 d → $\frac{(n+d)!}{n!\,d!}$ 새로운 특성.

Vector norm. $\|V\|_p$.

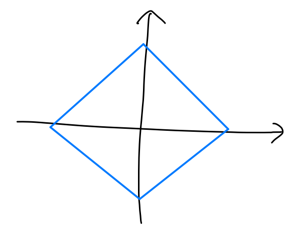
$V = (V_1, V_2, \cdots, V_n)$.

$\|V\|_p = \left(|V_1|^p + |V_2|^p + \cdots + |V_n|^p\right)^{\frac{1}{p}}$.

$\|V\|_p = 1$ in $\mathbb{R}^2$

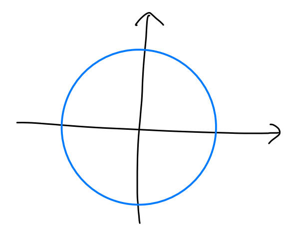
i) $p=1$.

$\|V\|_1 = |V_1| + |V_2| + \cdots + |V_n|$.



ii) $P=2$.

$\|V\|_2 = \sqrt{|V_1|^2 + |V_2|^2 + \cdots + |V_n|^2}$



iii) $p = \infty$

$$\|V\|_\infty = \left(|V_1|^p + |V_2|^p + \cdots + |V_n|^p \right)^{\frac{1}{p}}$$

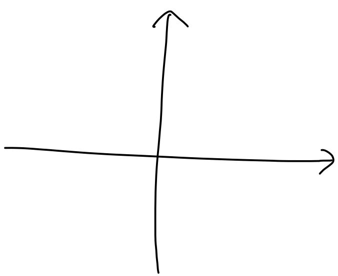
$$= \left\{ |V_k|^p \left(\cancel{\left(\frac{|V_1|}{|V_k|}\right)^p} + \cancel{\cdots} + \left(\frac{|V_k|}{|V_k|}\right)^p + \cancel{\cdots} + \cancel{\left(\frac{|V_n|}{|V_k|}\right)^p} \right) \right\}^{\frac{1}{p}}.$$

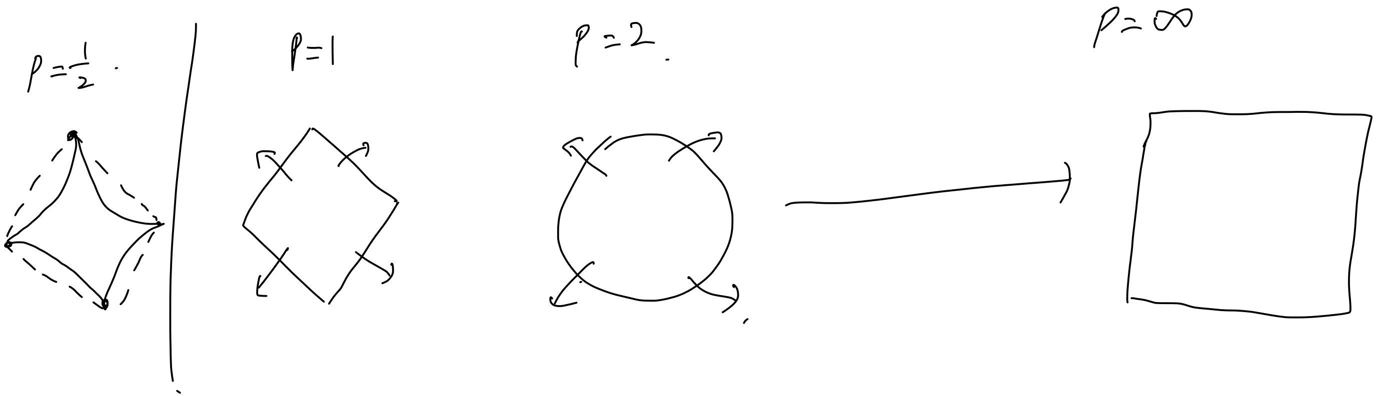
$$= |V_k|^p.$$

$$\therefore \ \|V\|_\infty = \max |V_i|$$

(max-norm

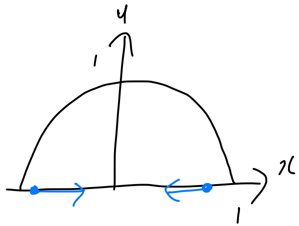
infinity-norm.





"거리의 개념이 바뀌면 level set 모양이 바뀐다!"

그래프를 바라보는 두 가지 관점.



i) graph of function.

$y = f(x)$. $\quad f(x) = \sqrt{1-x^2}$.

$$\nabla f(x) = \left((1-x^2)^{\frac{1}{2}}\right)'$$

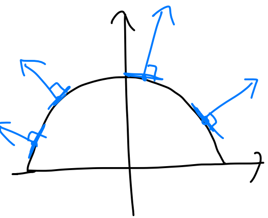
$$= \frac{1}{2}(1-x^2)^{-\frac{1}{2}} \cdot -2x.$$

$$= \frac{-x}{\sqrt{1-x^2}}.$$

$\nabla f(x) > 0$ when $x < 0$.

$\nabla f(x) < 0$ when $x > 0$.

ii) graph of <u>level set</u>.



$F(x,y) = k$. (k: constant).

$F(x,y) = x^2+y^2$, $F(x,y) = 1$.

$$\nabla F = \left(\frac{\partial F}{\partial x}, \frac{\partial F}{\partial y} \right)$$

$$= (2x, 2y).$$

f: function $\Rightarrow$ ∇f: The direction which f is increasing the fastest.

F: level set $\Rightarrow$ ∇F: normal vector of tangent plane.

규제

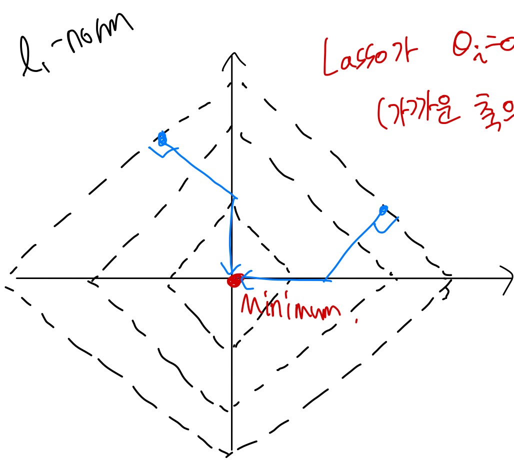
- 릿지 회귀 $$J(\theta) = MSE + \underbrace{\frac{1}{2}\lambda\sum_{i=1}^{n}\theta_i^2}_{l_2\text{-norm}}$$

특징

그래디언트 구하기 쉬움.

- 라쏘 회귀 $$J(\theta) = MSE + \underbrace{\lambda\sum_{i=1}^{n}|\theta_i|}_{l_1\text{-norm}}$$

$\theta_i = 0$ (중요하지 않은 특성들) 걸러낼 수 있음.



Lasso가 $\theta_i=0$ 많은 이유.
(가까운 축으로 먼저 떨어짐).