Foundation of Machine Learning 6주차

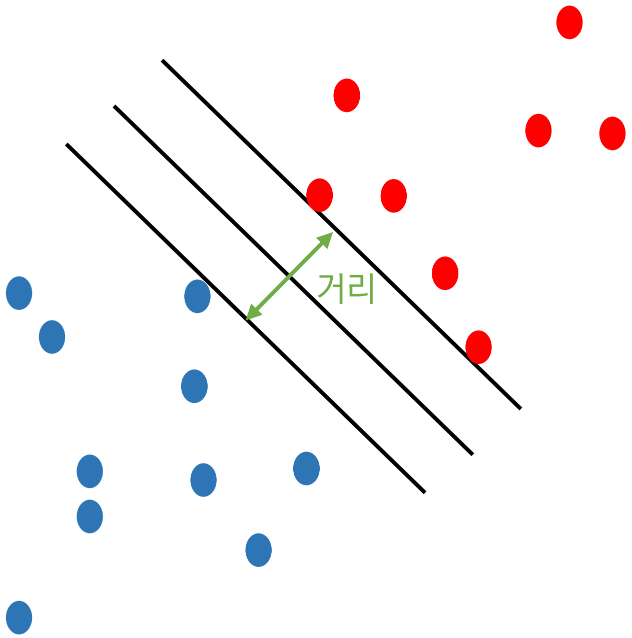
정재헌, 우지수 / 2023.02.24

Computational Data Science LAB

CONTENTS

01 What is SVM (Support Vector Machine) ?

Classification Algorithm



- ✓ 빨간색, 파란색 점을 분류할 수 있는 다양한 선형 분류 함수가 존재
- ✓ 그 중, 빨간색 점과 파란색 점의 집단 사이의 간격이 최대가 되도록 하는 최적의 선형 분류 함수를 탐색 = SVM

Decision Boundary

- Hyper Plane (초평면)

= SVM에서 두 개의 클래스를 나누는 (Decision Boundary) 경계면

bias

$$\underline{w}^T x + \overline{b} = 0$$

초평면의 법선 벡터

(기울기)

✓ 관측치가 초평면의 어디에 존재하느냐에 따라 관측치의 클래스 (y_i) 결정

01 What is SVM ?

Decision Rule

- Hyper Plane (초평면)

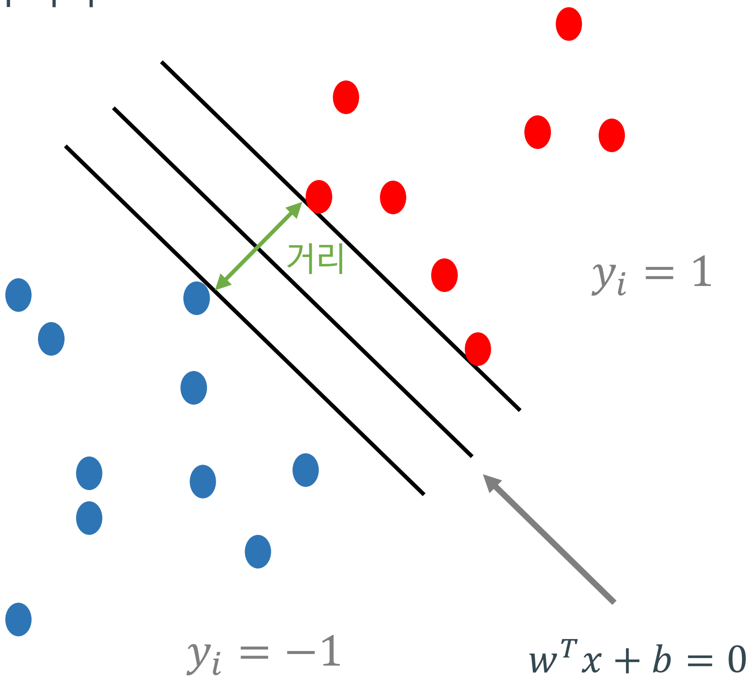
✓ Decision Boundary (결정경계) 와 가장 가까운 관측치와의 거리를 최대화하는 것이 목적

제약조건 : $y_i(w^T x_i + b) \geq 1$

$y_i = 1,\ i$번째 관측치는 양성 클래스

$y_i = -1,\ i$번째 관측치는 음성 클래스

✓ SVM은 위의 제약 조건을 만족하며 두 초평면 사이의 거리를 최대로 만드는 기울기(w)와 b를 구해야 함.



02 | Margin and Error

Maximize Margin

- Margin (마진)

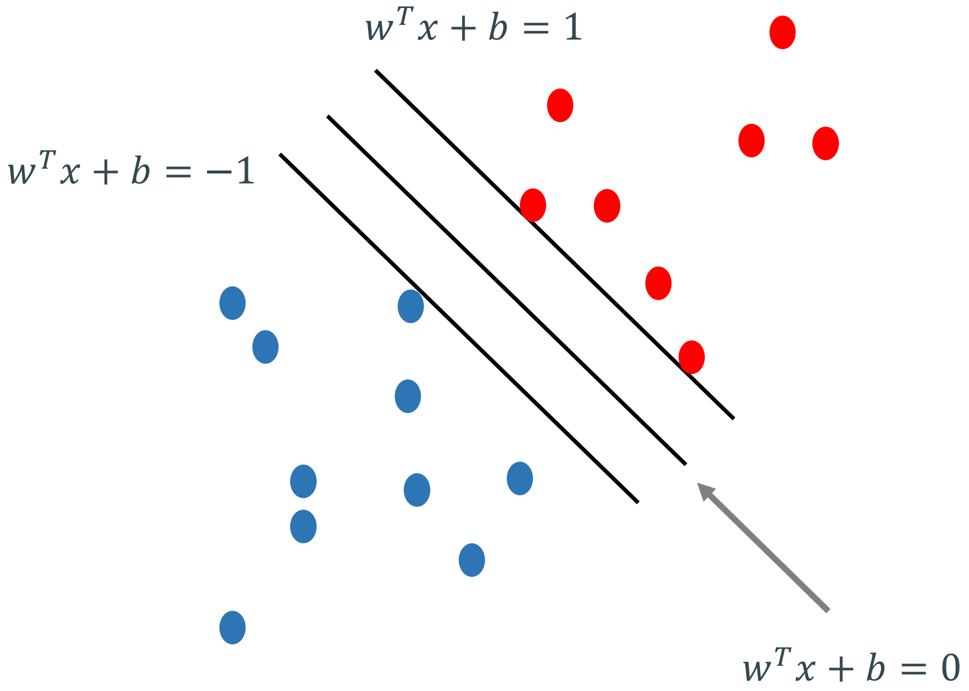
= Decision Boundary (결정경계) 를 만드는 두 직선 사이의 거리

$$\mathrm{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix},\ x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$Margin = \frac{|-(b+1)+(b-1)|}{\sqrt{w^T w}} = \underline{\frac{2}{||w||}}$$

초평면의 법선 벡터 (기울기)

w 의 크기



Maximize Margin

- Margin (마진)

$$Maximize \ \frac{2}{||w||}$$

$$subject\ to \quad y_i(w^T x_i + b) \ \geq \ 1$$

역수 $\longrightarrow$

$$Minimize \ \frac{||w||^2}{2}$$

$$subject\ to \quad y_i(w^T x_i + b) \ \geq \ 1$$

✓ 제약 조건이 있는 상황에서 최적화 문제를 해결해야 하므로 라그랑주 승수를 사용해야함.

✓ 이때, 마진의 역수를 취함으로써 라그랑주 승수 계산이 간단해지고 초평면의 기울기를 구하는 과정이 효율적이게 됨.

Minimize Error

- Error (ξ)

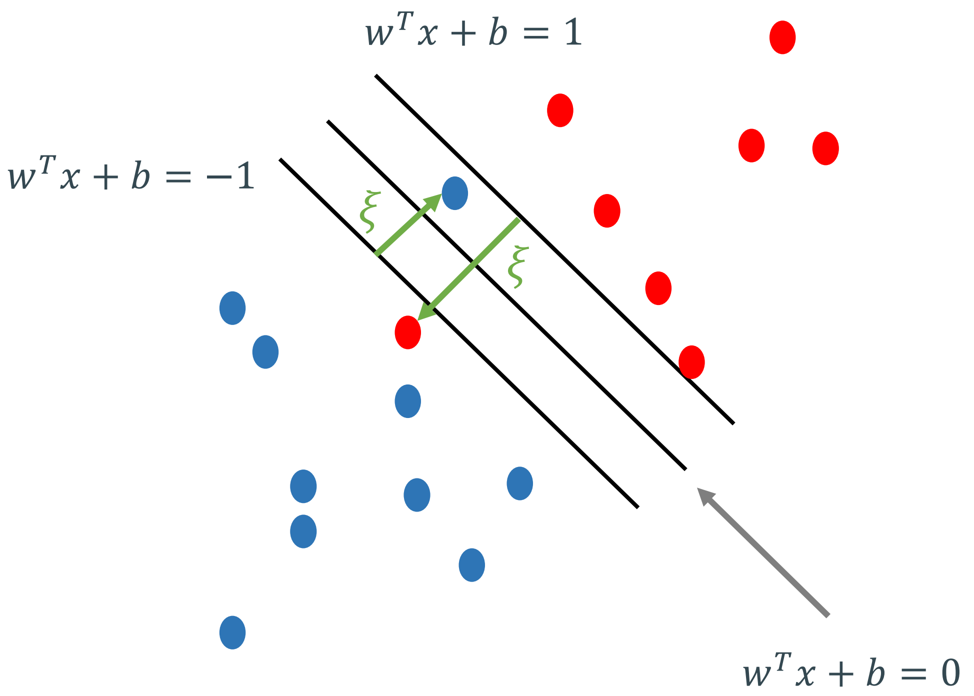
= 분리불가한 선형 SVM 학습을 위해 사용

$$Minimize\ \frac{||w||^2}{2} + C\sum_{i=1}^{n}\xi_i$$

$$subject\ to\quad y_i(w^T x_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0\ ,\ \sum \xi_i \leq constant$$

✓ ξ가 1보다 크게 되면 관측치가 반대 클래스로 넘어가 있다는 의미



02 Margin and Error

Minimize Error

- Error (ξ)

= 분리불가한 선형 SVM 학습을 위해 사용

$$Minimize\ \frac{||w||^2}{2} + C\sum_{i=1}^{n} \xi_i$$

ξ 허용 정도를 정해주는 하이퍼파라미터

$$subject\ to\ \ y_i(w^T x_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0\ ,\ \sum \xi_i \leq constant$$

✓ C가 큰 경우 : Hard SVM (C가 커질수록 전체 식을 줄여야 하므로 그만큼 ξ 를 허용해주지 않는다는 의미)

ex) C가 무한대이면 모든 관측치가 마진 바깥에 위치 = Overfitting

✓ C가 작은 경우 : Soft SVM (ξ를 어느정도 허용해준다는 의미)

03 | SVM Primal and Dual

To Minimize Cost Function

- Primal

$$Minimize\ \frac{||w||^2}{2} + \mathrm{C}\sum_{i=1}^{n}\xi_i$$

$$subject\ to\ \ y_i(w^T x_i + b)\ \geq\ 1 - \xi_i$$

$$\xi_i\ \geq 0$$

✓ 마진을 최대화 시키면서 ($||w||^2$를 최소화 시키면서) ξ_i를 최소화 하는 초평면을 구하는 것이 목적

✓ 이때, 우리가 찾아야하는 것은 초평면의 기울기 w와 b ⟶ Dual

03 | SVM Primal and Dual

To Solve Hyper Plane

- Dual

$$L(w,b,\alpha) = \frac{||w||^2}{2} - C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i(y_i(w^T x_i + b) - 1 + \xi_i) + \sum_{i=1}^{n}\mu_i\xi_i$$

✓ 라그랑주 승수법을 통해 Primal 식에 존재했던 등식제약을 함수의 제약으로 옮겨서 제약이 없는 문제로 표현

✓ 위 식을 최소화하는 w, b, ξ_i를 찾기 위해 라그랑주 승수 α_i, μ_i 도입

각각을 미분하여 0이 되는 해 찾기

1. w에 대해 미분 : $\partial_w L = 0 \rightarrow w - \sum_{i=1}^{n}\alpha_i y_i x_i = 0 \rightarrow w = \sum_{i=1}^{n}\alpha_i y_i x_i$
2. b에 대해 미분 : $\partial_b L = 0 \rightarrow -\sum_{i=1}^{n}\alpha_i y_i = 0 \rightarrow \sum_{i=1}^{n}\alpha_i y_i = 0$
3. ξ_i에 대해 미분 : $\partial_{\xi_i} L = 0 \rightarrow C - \alpha_i - \mu_i = 0 \rightarrow \alpha_i + \mu_i = C$

- Dual

1. w에 대해 미분 : $\partial_w L = 0 \quad \rightarrow \quad w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \quad \rightarrow \quad w = \sum_{i=1}^{n} \alpha_i y_i x_i$
2. b에 대해 미분 : $\partial_b L = 0 \quad \rightarrow \quad -\sum_{i=1}^{n} \alpha_i y_i = 0 \quad \rightarrow \quad \sum_{i=1}^{n} \alpha_i y_i = 0$
3. ξ_i에 대해 미분 : $\partial_{\xi_i} L = 0 \quad \rightarrow \quad C - \alpha_i - \mu_i = 0 \quad \rightarrow \quad \alpha_i + \mu_i = C$

대입해줌으로써 α_i로 모든 것을 표현 가능

$$\text{maximize} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \, y_i y_j x_i^T x_j$$

- Dual

$$maximize \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j\, y_i y_j x_i^T x_j$$

✓ 이 식을 최대화하는 α_i 를 구한 후 w, b 식에 대입해주면 구할 수 있음.

✓ 이는 KKT condition을 만족한다고 가정

[KKT condition]

1. $\alpha_i[y_i(x_i^T w + b) - (1 - \xi_i)] = 0$
2. $\mu_i \xi_i = 0$
3. $y_i\left(x_j^T w + b\right) - (1 - \xi_i) \ \geq 0$

- Dual

[KKT condition]

1. $\alpha_i[y_i(x_i^T w + b) - (1 - \xi_i)] = 0$
2. $\mu_i \xi_i = 0$
3. $y_i\left(x_i^T w + b\right) - (1 - \xi_i) \geq 0$

✓ 3번째 조건을 만족한다고 하였을 때, 1번째 조건을 만족해야 라그랑주 문제를 풀 수 있음.

✓ 1번째 조건이 0이 되려면 α_i가 0이 되거나 $y_i\left(x_i^T w + b\right) - (1 - \xi_i)$가 0이 되어야 한다는 의미임.

- $y_i\left(x_i^T w + b\right) - (1 - \xi_i) \neq 0$: 마진 바깥에 관측치가 존재하는 경우, α_i가 0이 되어야함
- $\alpha_i \neq 0$: 관측치가 마진 위에 존재한다는 의미 = support vector

Support Vector

- Dual

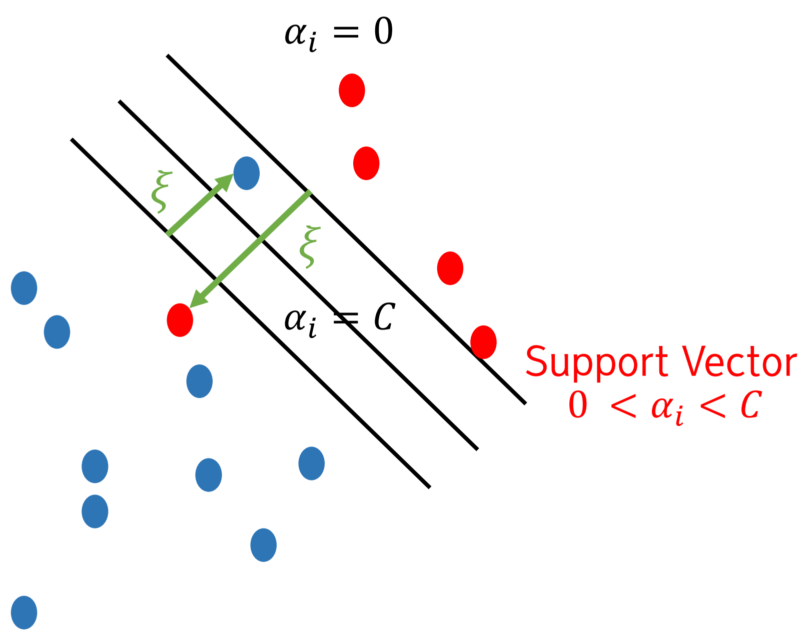
$$maximize \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j\, y_i y_j x_i^T x_j$$

✓ α_i : 관측치 i가 Support Vector인지에 대한 여부

$0 < \alpha_i < C$ 이면, Support Vector

$\alpha_i = 0$ 이면, 마진 밖에 위치

$\alpha_i = C$ 이면, 마진 내에 위치



✓ y_i : 기존 관측치 i의 y값

Support Vector and New Data

- Dual

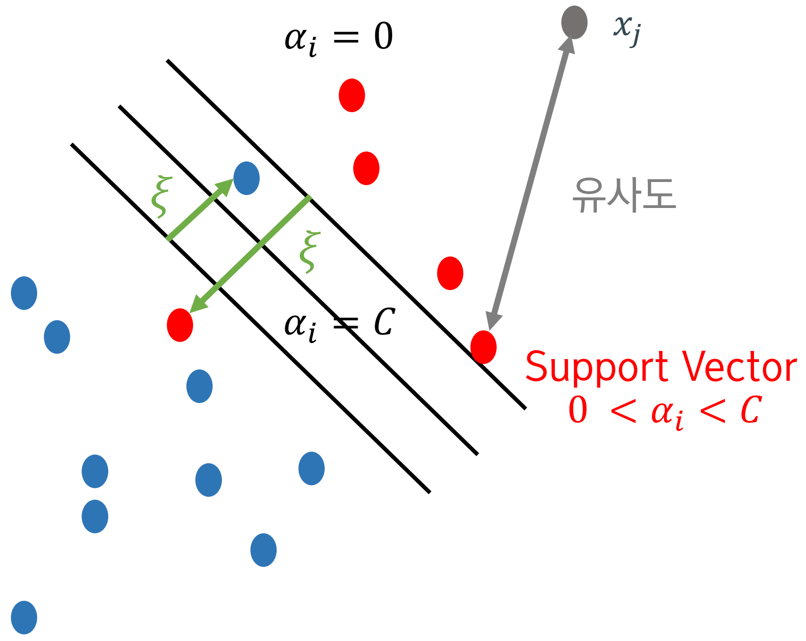
$$maximize \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j\, y_i y_j x_i^T x_j$$

✓ α_i : 관측치 i가 Support Vector인지에 대한 여부

$0 < \alpha_i < C$ 이면, Support Vector

$\alpha_i = 0$ 이면, 마진 밖에 위치

$\alpha_i = C$ 이면, 마진 내에 위치



✓ y_i : 기존 관측치 i의 y값

✓ $x_i^T x_j$: support vector인 관측치 i와 새로운 관측치 j의 내적 (유사도, Kenel 함수)

→ 이를 통해 새로운 관측치 j의 클래스를 분류해낼 수 있으며 비선형 SVM 해결 가능

Support Vector and Hyper Plane

- Dual

$$maximize \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j\, y_i y_j x_i^T x_j$$

$$w^* = \sum_{i=1}^{n} \alpha_i y_i$$

$$b^* = y_i - \sum_{i=1}^{n} \alpha_i y_i x_i^T x_j$$

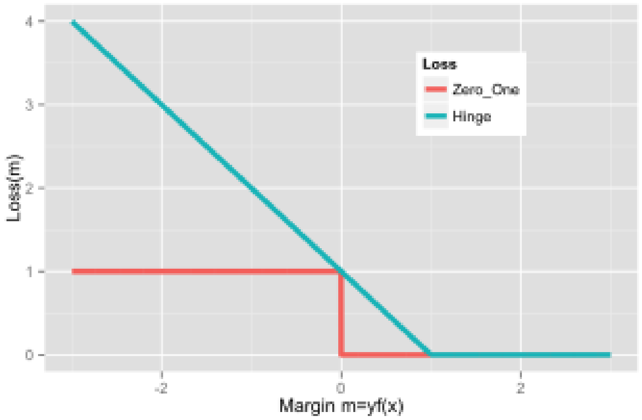
→ Dual은 관측치 관점에서 초평면 (Hyper Plane)의 기울기 w와 b의 최적해를 구할 수 있음.

- SVM의 목적함수

$$Minimize\ \frac{||w||^2}{2} + C\sum_{i=1}^{n}\max(0, 1\ - y(x_i^T w + b))$$

미분 가능

Hinge Loss (미분 불가능)



✓ SVM의 목적함수를 또다른 관점에서 보면 $\frac{||w||^2}{2}$ 와 Hinge Loss의 결합 형태

✓ 목적함수를 최소화 시키는 w와 b를 구해야 하지만 Hinge Loss는 미분 불가능

어떻게 최적화 하는 것일까 ?

04 | Hinge Loss and Subgradient

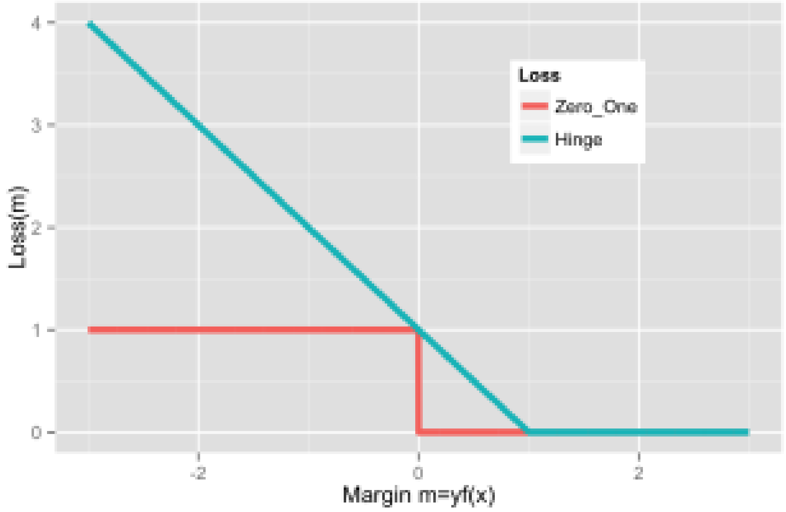
What is Hinge Loss ?

- Hinge Loss

  ✓ SVM에서 사용되는 손실함수

  ✓ 마진 (Margin) 을 최대화하려는 목표

$$L\left(y,\left(x_i^T w+b\right)\right)=\max(0, 1-y\left(x_i^T w+b\right))$$



$y_i\left(x_i^T w+b\right)$

$$L\left(y,\left(x_i^T w+b\right)\right)=\begin{cases} if\;\; y_i\left(x_i^T w+b\right) \geq 1, & ;\; 0 \\ if\;\; y_i\left(x_i^T w+b\right) < 1, & ;\; 1-y\left(x_i^T w+b\right) \end{cases}$$

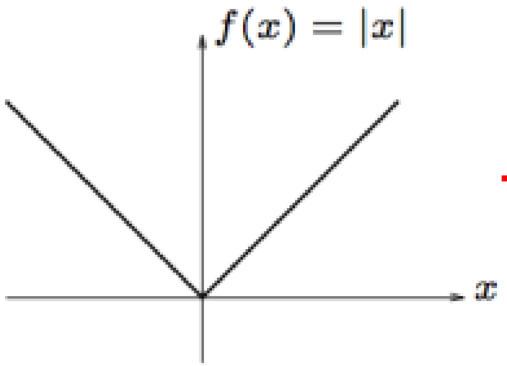
$$L'\left(y,\left(x_i^T w+b\right)\right)=\begin{cases} if\;\; y_i\left(x_i^T w+b\right) > 1, & ;\; 0 \\ if\;\; y_i\left(x_i^T w+b\right) < 1, & ;\; -y \\ if\;\; y_i\left(x_i^T w+b\right) = 1, & ;\; ? \end{cases}$$

$\longrightarrow$ Subgradient Descent

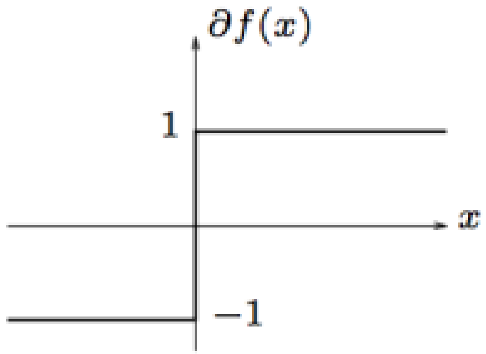
04 Hinge Loss and Subgradient

What is Subgradient ?

- Subgradient



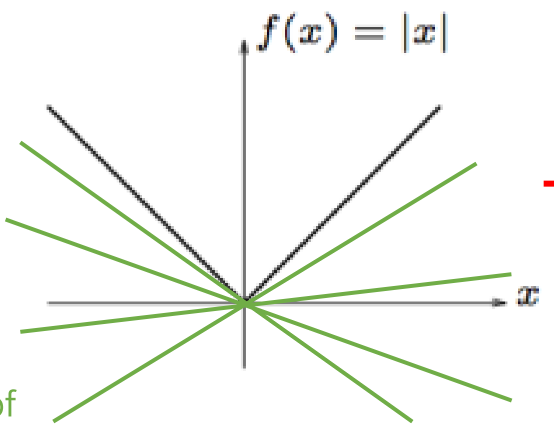
미분



$$f'(x) = \begin{cases} if\ x > 0, & ;\ 1 \\ if\ x < 0, & ;\ -1 \\ if\ x = 0, & ;\ ? \end{cases}$$

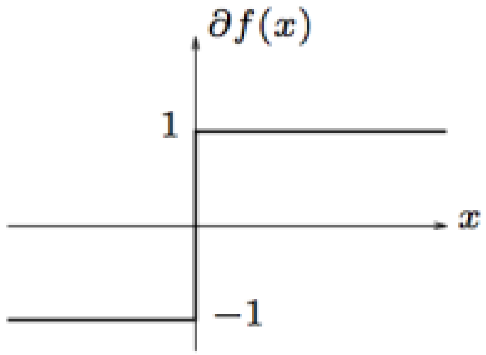
✓ $x = 0$ 인 지점은 미분이 되지 않아 closed form이 도출되지 않음.

- Subgradient



Subgradient of $f(x) = (g)$

미분



$$f'(x) = \begin{cases} if\ x > 0, & ;\ 1 \\ if\ x < 0, & ;\ -1 \\ if\ x = 0, & ;\ ? \end{cases}$$

✓ $x = 0$ 인 지점은 미분이 되지 않아 closed form이 도출되지 않음.

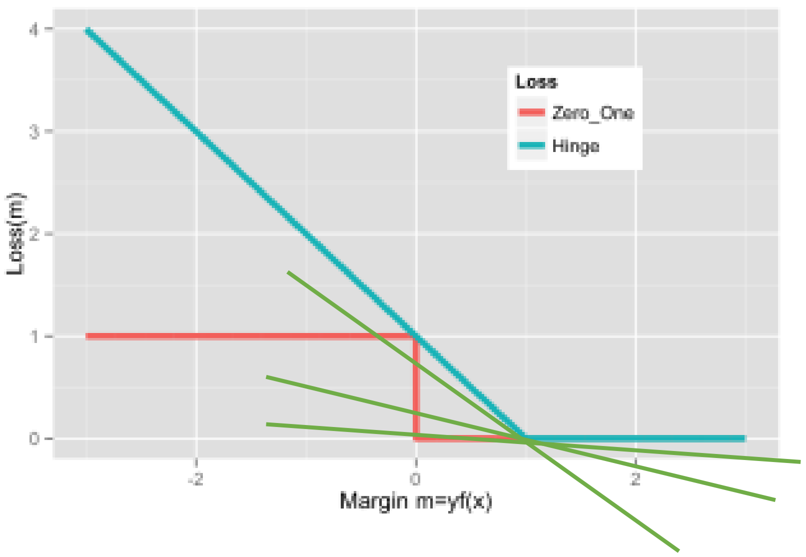
✓ 이를 해결하고자 위와 같은 직선을 만들었을 때 주 함수인 $f(x)$보다 아래에 위치하게 됨.

✓ 이때, $f(y) \geq f(x) + g^T(y - x)$ 식 성립

$$g = \begin{cases} if\ x > 0, & ;1 \\ if\ x < 0, & ;\ -1 \\ if\ x = 0, & ;\ [-1,1] \end{cases}$$

04 | Hinge Loss and Subgradient

Hinge Loss and Subgradient

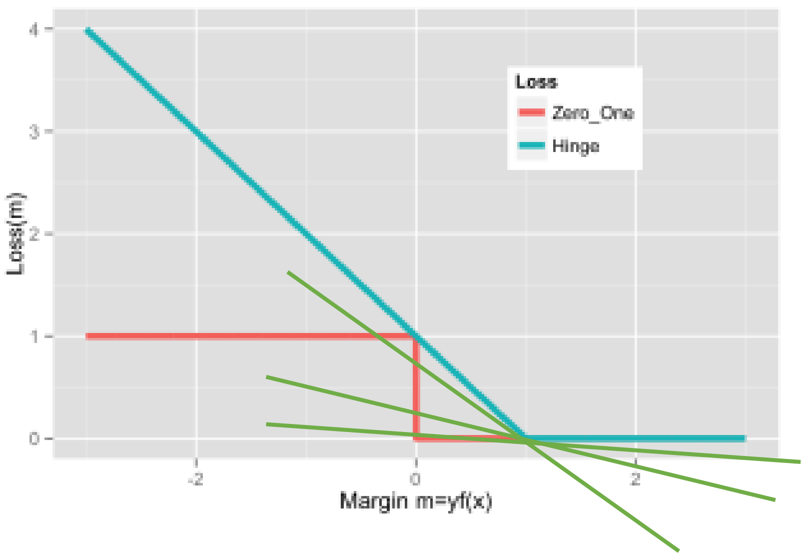


Subgradient of $f(x) = g$

$$g = \begin{cases} if\ \ y_i\left(x_i^T w + b\right) > 1 & ;\ 0 \\ if\ \ y_i\left(x_i^T w + b\right) < 1 & ;\ -y \\ if\ \ y_i\left(x_i^T w + b\right) = 1, & ;\ [-y, 0] \end{cases}$$

04 | Hinge Loss and Subgradient

Solve Hinge Loss with Subgradient Descent



Subgradient of $f(x) = g$

$$g = if \;\; y_i(x_i^T w + b) > 1 \qquad ;\; 0$$

$$\frac{\partial L}{\partial w} = if \;\; y_i(x_i^T w + b) > 1 \qquad ;\; -y_i x_i$$

$$\frac{\partial L}{\partial b} = if \;\; y_i(x_i^T w + b) > 1 \qquad ;\; -y_i$$

Subgradient를 사용하여 gradient를 계산한 후,

w와 b 업데이트

$$w = w \; - t(C\frac{\partial L}{\partial w} + 2\mathrm{w})$$

$$b = b \; - t(C\frac{\partial L}{\partial b})$$

04 | Hinge Loss and Subgradient

Solve Hinge Loss with Subgradient Descent

- t (Step Size)

  = Subgradient Descent 알고리즘에서 각 업데이트 단계 시 현재 위치에서 이동할 거리를 결정하는 하이퍼파라미터

1. Fixed Step Size

2. Diminishing Step Size

   = 초기에는 큰 t를 사용함으로써 초기 수렴속도를 높이고 학습이 진행됨에 따라 점차 t를 줄여나감으로써
   최적점에 도달할 수 있게 함.

✓ 문제점 : Dual과 달리 Subgradient Descent로는 초평면의 w과 b를 찾는 방법론은 최적점이 근사치로 나옴. (non-convex이기 때문)

Q&A

감사합니다.