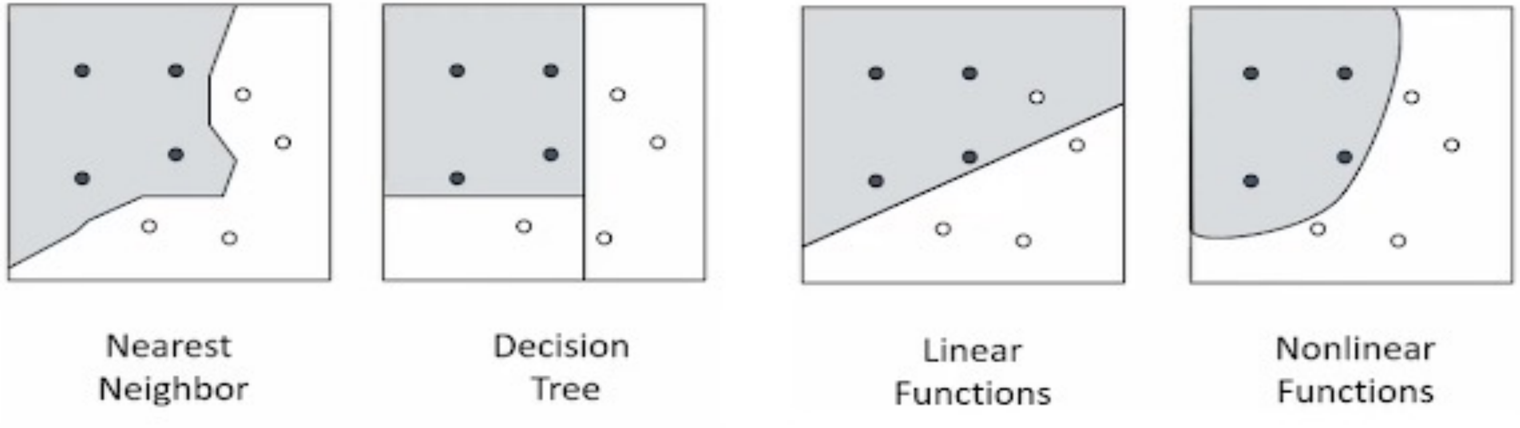
17기 정규세션

ToBig's 16기 김종우

SVM

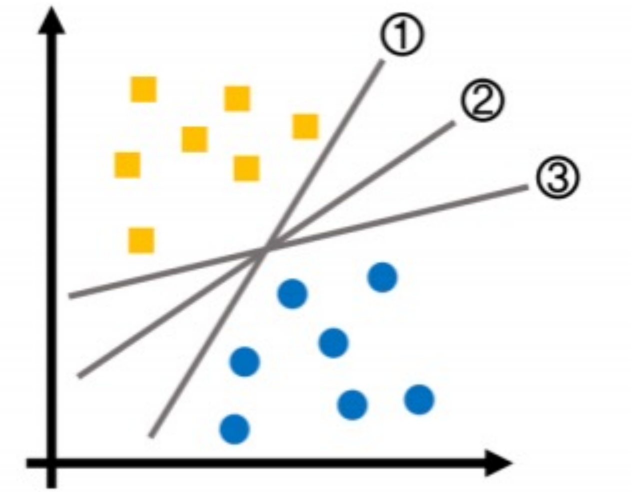
Contents

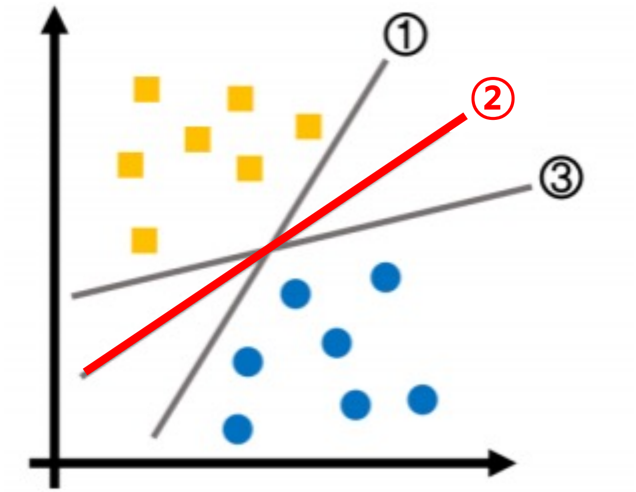
Discriminant Function in binary classification



$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$$

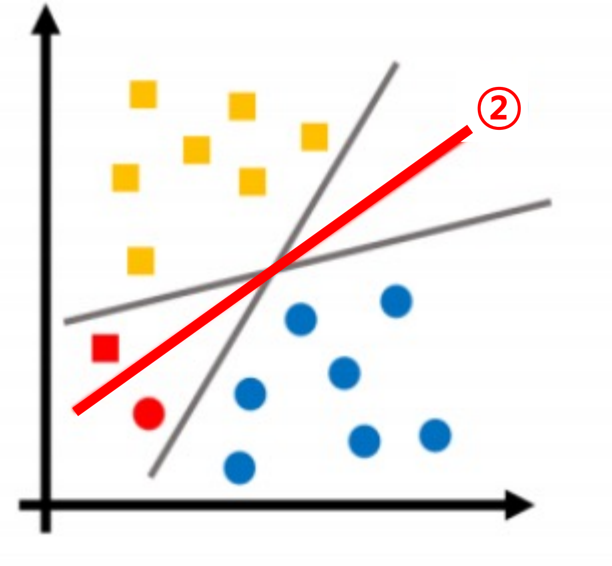
SVM (support vector machine)



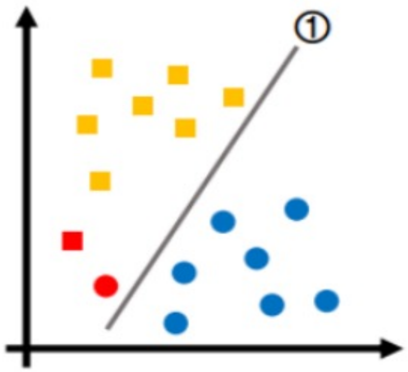


Unit 01 | SVM

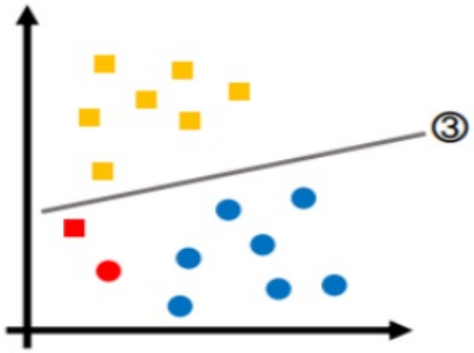
SVM (support vector machine)



새로운 점이 추가 되었을 경우



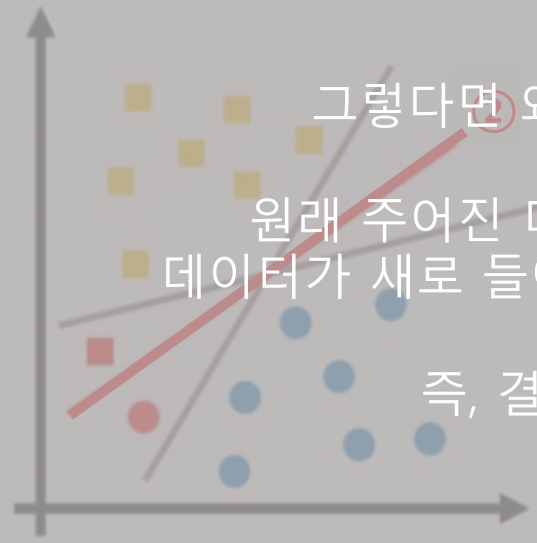
동그라미를 분류 하지 못함



네모를 분류 하지 못함

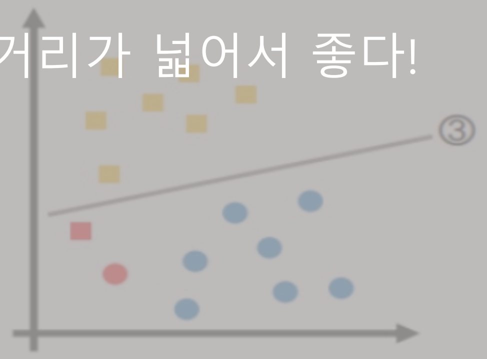
Unit 01 | SVM

SVM (support vector machine)



새로운 점이 추가 되었을 경우

동그라미를 분류 하지 못함



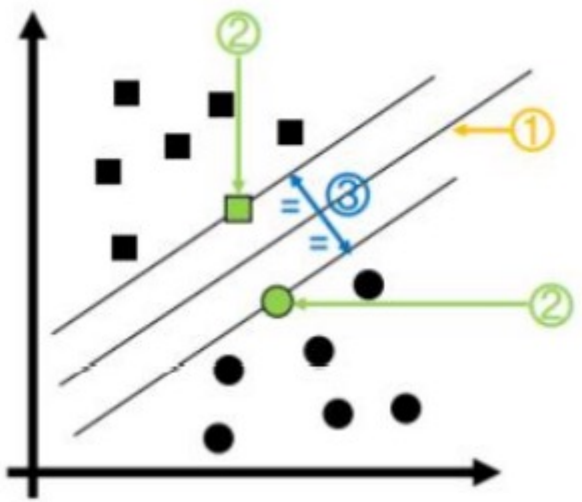
네모를 분류 하지 못함

그렇다면 왜 2번 직선이 데이터를 가장 잘 나눈 선일까?

원래 주어진 데이터들과 경계선이 충분히 많이 떨어져 있어서
데이터가 새로 들어왔을 때 더 잘 분류할 가능성이 높다 (일반화 성능)

즉, 결정경계와 데이터의 거리가 넓어서 좋다!

SVM (support vector machine)



Hyper plane(초평면) : 데이터를 나누는 기준이 되는 경계

Support vector : Hyper plane과 가장 가까운 데이터

Margin : 결정 경계 사이의 거리

SVM (support vector machine)

Sample drawn i.i.d from set $X \in \mathbb{R}^d$ according to some distribution D

$$S = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\} \in X \times \boxed{\{-1, +1\}}$$

수학적 계산을 위해서
class를 -1과 1로 표시

Find hypothesis $h : X \to \{-1, +1\}$ in H(classifier) with small generalization error $\boxed{R_D(h)}$

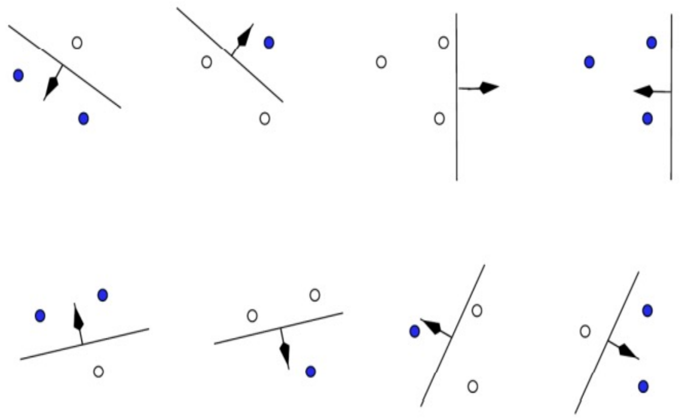
Error R을 최소화하는 H
classifier를 만드는 것이 목표

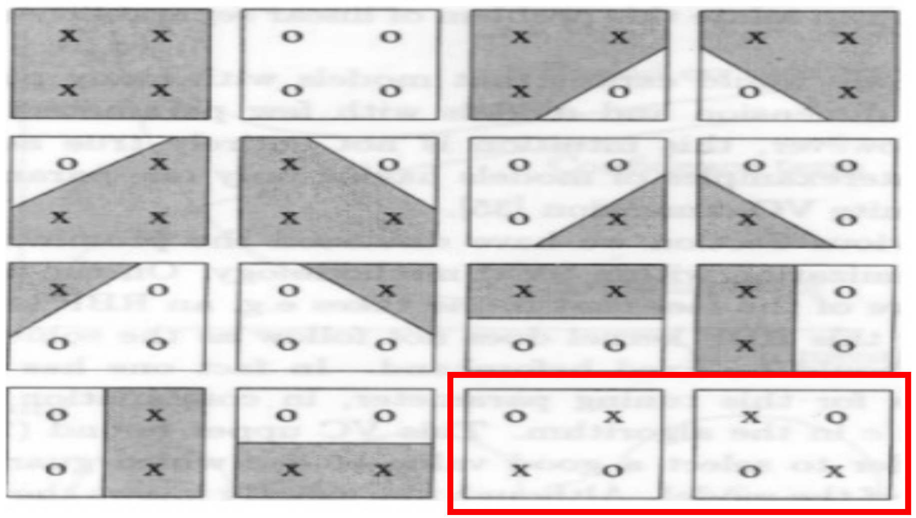
$$H = \{x \to sign(\mathbf{w} \cdot \mathbf{x} + b) : \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}\}$$

직선 그래프를 기준으로 위
쪽은 +1, 아래쪽은 -1을 반환

SVM (support vector machine)
- VC dimension

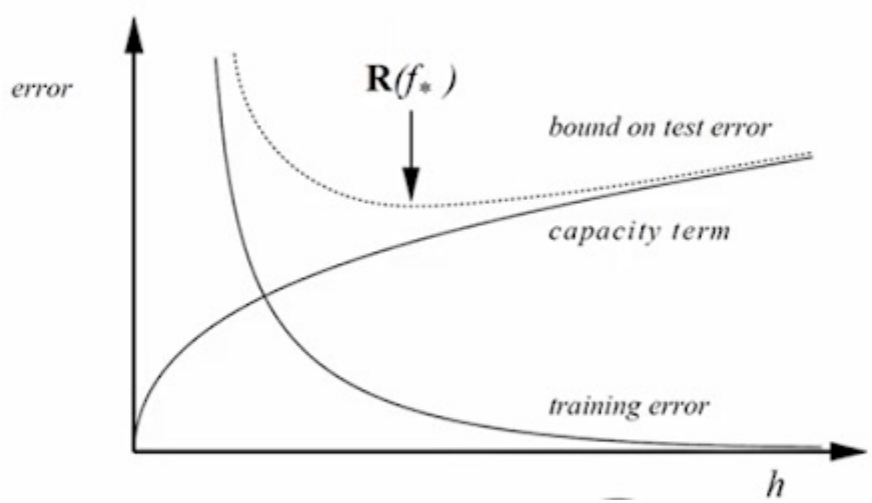
Shatter : 개별 인스턴스를 가능한 모든 조합의 이진 레이블로 구분해 낼 수 있다는 것
VC dimension : n차원이 최대 shatter 할 수 있는 점의 수 (input dimension +1)





SVM (support vector machine)

- Structural Risk Minimization



모델의 복잡도(h)가 증가할수록

- Capacity(Flexibility)는 지수적으로 증가하고
- Training Error는 지수적으로 감소

Trade-off를 고려해서 모델의 위험을 최소화 하는 (Test error bound가 가장 작은) 모델을 선택

Unit 01 | SVM

SVM (support vector machine)
- Structural Risk Minimization

$$R[f] \le R_{emp}[f] + \sqrt{\frac{h(ln\frac{2n}{h}+1) - ln(\frac{\delta}{4})}{n}}$$

R_{emp} : xi에 대해서 f(xi)가 정답 yi와 같다면 0을 반환하고, 틀리다면 1을 반환

h : H라는 분류기에 의해서 최대로 분류 될 수 있는 point의 수(VC dimension)

n : 학습데이터 의 수

$$R = R_{emp} + (Capacity\ Term)$$

< Remp[f]이 동일하다는 가정>

- n이 증가한다면 logn/n의 꼴 임으로 Capacity term은 감소
- h가 증가한다면 h/logh의 꼴 임으로 Capacity term은 증가

SVM (support vector machine)
- Structural Risk Minimization

margin을 최대화 하는 것이 Test error bound 을 최소화 하는 것인지 그 이유를 알아보자

$$h \leq min\left(\left\lceil \frac{R^2}{\Delta^2} \right\rceil, D\right) + 1$$

D = input의 차원수(dimensionality)
R = 전체 data를 감싸는 가장 작은 초구(Hypersphere)의 반지름
Δ = margin의 크기

- input data가 주어졌을 때 D와 R은 고정. 그렇다면 margin을 충분히 크게 하여, h≤ $\left\lceil \frac{R^2}{\Delta^2} \right\rceil$ +1≤ Dimension + 1
- margin을 최대화 하는것이 CapacityTerm을 최소화 하여 전체적인 loss를 줄인다.

Maximize the margin ⇒ Minimizing the VC dimension ⇒ Minimizing the Expected Risk(test error)

Support Vector Machine : Hard margin vs. Soft margin

SVM은 **margin**을 벗어나는 예외를 허용하는 **Soft margin** 방법과, 허용하지 않는 **Hard margin** 방법으로 구분

	Hard margin	Soft margin
Linearly separable	Basic form (Case 1)	Introduce **penalty terms** (Case 2)
Linearly non-separable	Utilize **Kernel trick**	Introduce **penalty terms** Utilize **Kernel trick** (Case 3)

Unit 02 | Hard margin & Linearly separable

Hard margin & Linearly separable

$$x_1 = x_0 + pw$$

$$w^T x_1 + b = w^T(x_0 + pw) + b = 1$$

$$w^T x + b = 1$$

$$w^T x_0 + b + p \cdot w^T w = 1$$

$$w^T x_0 + b = 0$$

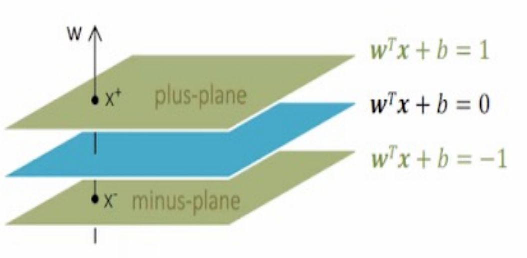
$$p \cdot w^T w = 1$$

$$p = -\frac{1}{w^T w} = -\frac{1}{||w||^2}$$

$$margin = |p| = \frac{1}{||w||^2}$$

$$maximize \quad (margin) = \frac{2}{||w||^2}$$

우리의 목표는
margin을 최대화



x_1; plus plane point, x_0 ; hyperplane point

$$if \;\; y_i = +1, \;\; w^T x_i + b \geq 1$$

$$if \;\; y_i = -1, \;\; w^T x_i + b \leq -1$$

Objective Function

$$\boxed{minimize \quad (\frac{1}{margin}) = \frac{1}{2}||w||^2}$$

Constraint

$$\boxed{y_i(w^T x_i + b) \geq 1}$$

Hard margin & Linearly separable

$$x_1 = x_0 + pw$$

$$w^T x_1 + b = w^T(x_0 + pw) + b = 1$$

$w^T x + b = 1$

$$w^T x_0 + b + p \cdot w^T w = 1$$

$w^T x_0 + b = 0$

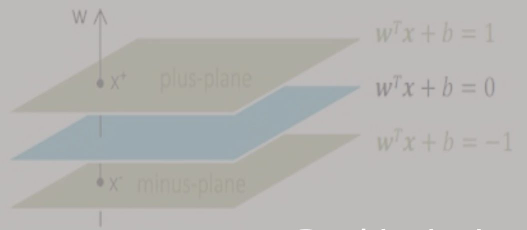
$$p \cdot w^T w = 1$$

$$p = -\frac{1}{w^T w} = -\frac{1}{||w||^2}$$

$$margin = |p| = \frac{1}{||w||^2}$$

$$maximize \quad (margin) = \frac{2}{||w||^2}$$

우리의 목표는
margin을 최대화



x_1; plus plane point, x_0 ; hyperplane point

$$if \ \ y_i = +1, \ \ w^T x_i + b \geq 1$$

$$if \ \ y_i = -1, \ \ w^T x_i + b \leq -1$$

Objective Function

$$\boxed{minimize \quad (\frac{1}{margin}) = \frac{1}{2}||w||^2}$$

Constraint

$\boxed{x_1 ; \text{plus plane point}, x_0 ; \text{hyperplane point}}$

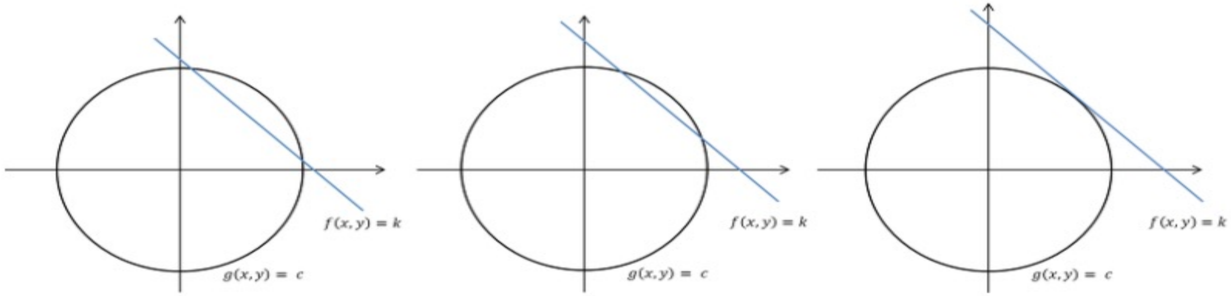
Constraint를 고려해서 최적의 해를 찾는 Objective Function을 찾아야 해

Lagrangian Multiplier > Dual problem > check KKT condition

Hard margin & Linearly separable

- **Lagrange Multiplier Method :** 제약 조건이 있는 최적화 문제를 풀기 위해 고안한 방법

"제약 조건 g를 만족하는 f의 최솟값 또는 최댓값은 f와 g가 접하는 지점에 존재할 수도 있다."



g(x,y)=c와 f(x,y)가 접할 때 f(x,y)는 최대가 된다

두 함수 f와 g가 접한다는 것??? 두 함수의 gradient가 서로 상수배인 관계에 있다는 것!!

$$\nabla f(x,\ y) = \lambda \nabla g(x,\ y)$$

$$L\ (x,\ y,\ \lambda\ )= f\ (x,\ y) - \lambda g\ (x,\ y)$$

Hard margin & Linearly separable
- Dual problem

optimization problem

<primal problem>

일반적인 부등식(제약), 등식(목적식)으로 이루어진 최적화 문제

$$p^\star = \min_{\mathbf{x}} f(\mathbf{x})$$

$$\text{subject to:} \quad g_i(\mathbf{x}) \le 0, \quad i = 1, \dots, m$$

$$h_j(\mathbf{x}) = 0, \quad j = 1, \dots, k$$

<dual problem>

라그랑주 승수를 통해 만들어진 제약식이 없는 최적화 문제

$$d(\mu, \lambda) = \min_{\mathbf{x}} L(\mathbf{x}, \mu, \lambda)$$

$$= \min_{\mathbf{x}} \left(f(\mathbf{x}) + \sum_{i=1}^{m} \mu_i g_i(\mathbf{x}) + \sum_{j=1}^{k} \lambda_j h_j(\mathbf{x}) \right)$$

Hard margin & Linearly separable

$$L(\mathbf{x},\mu,\lambda) = f(\mathbf{x}) + \mu^T\mathbf{g}(\mathbf{x}) + \lambda^T\mathbf{h}(\mathbf{x})$$
$$= f(\mathbf{x}) + \sum_{i=1}^{m} \boxed{\mu_i g_i(\mathbf{x})} + \sum_{j=1}^{k} \boxed{\lambda_j h_j(\mathbf{x})} \quad \mu \geq 0$$

$\mu \geq 0$ 이고, $g_i(\mathbf{x}) \leq 0$ 이기 때문에
$\mu_i g_i(\mathbf{x})$는 항상 0보다 작거나 같다.

$h_j(\mathbf{x}) = 0$ 이기 때문에 항상 0이다.

$$f(\mathbf{x}) \geq L(\mathbf{x},\mu,\lambda) \geq \min_{\mathbf{x}} L(\mathbf{x},\mu,\lambda) = d(\mu,\lambda)$$

μ에 의해 *F(x)* 의 최대 하한선 (lower bound)이 결정, max $d(\mu,\lambda)$ 를 찾는 것이 중요(*F(x)* 의 최솟값을 최대화, page 20에 이유 설명)
*F(x)*와 max $d(\mu,\lambda)$ 의 gap이 0 이 되도록 하여 strong duality(zero duality)를 형성하게 함

Hard margin & Linearly separable

- KKT condition

- $\partial_x(f(x) + \sum u_i g_i(x) + \sum \lambda_j h_j(x)) = 0$ 라그랑주 상수를 제외한 상수로 편미분
- $u_i g_i(x) = 0,$ for all i
- $g_i(x) \le 0, h_j(x) = 0$, for all i,j
- $u_i \ge 0$ for all i

Hard margin & Linearly separable

- KKT condition & duality example

Objective Function $\quad minimize \ (\frac{1}{margin}) = \frac{1}{2}||w||^2$

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

<Lagrange problem>

$$\min L_p(w, b, \alpha_i) = \frac{1}{2}\|w\|_2^2 - \boxed{\sum_{i=1}^{n} \alpha_i (y_i(w^T x_i + b) - 1)}$$

Lagrange Multiplier

KKT condition

- $\partial_x (f(x) + \sum u_i \, g_i(x) + \sum \lambda_j \, h_j(x)) = 0$

$$\frac{\partial L(w, b, \alpha_i)}{\partial w} = 0 \quad \rightarrow \quad w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$\frac{\partial L(w, b, \alpha_i)}{\partial b} = 0 \quad \rightarrow \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

Hard margin & Linearly separable

- KKT condition & duality example

$$\min L_p(w,b,\alpha_i) = \boxed{\frac{1}{2}\|w\|_2^2} \boxed{-\sum_{i=1}^{n}\alpha_i(y_i(w^T x_i + b) - 1)} = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j \quad \max L_D(\alpha_i) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

a_i의 최고차항 계수는 (-)임으로 minimize이 maximize로 변경
(page 17의 최대 하한선 설정이유, 뒤에 reference 참조)

$$\begin{aligned}
\frac{1}{2}\|w\|_2^2 &= \frac{1}{2}w^T w \\
&= \frac{1}{2}w^T \sum_{j=1}^{n}\alpha_j y_j x_j \\
&= \frac{1}{2}\sum_{j=1}^{n}\alpha_j y_j (w^T x_j) \\
&= \frac{1}{2}\sum_{j=1}^{n}\alpha_j y_j (\sum_{i=1}^{n}\alpha_i y_i x_i^T x_j) \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j
\end{aligned}$$

$$\begin{aligned}
-\sum_{i=1}^{n}\alpha_i(y_i(w^T x_i + b) - 1) &= -\sum_{i=1}^{n}\alpha_i y_i(w^T x_i + b) + \sum_{i=1}^{n}\alpha_i \\
&= -\sum_{i=1}^{n}\alpha_i y_i w^T x_i - b\sum_{i=1}^{n}\alpha_i y_i + \sum_{i=1}^{n}\alpha_i \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^{n}\alpha_i
\end{aligned}$$

Unit 02 | Hard margin & Linearly separable

Hard margin & Linearly separable
- KKT condition & duality example

$$\min L_p(w,b,\alpha) = \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{n}\alpha_i(y_i(w^Tx_i+b)-1) = \ldots \qquad \max L_D(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

α_i의 최고차항 계수는 (-)임으로 minimize이 maximize로 변경
(page 17의 최대 하한선 설정이유, 뒤에 reference 참조)

$$\begin{aligned}\frac{1}{2}\|w\|_2^2 &= \frac{1}{2}w^Tw \\ &= \frac{1}{2}w^T\sum_{j=1}^{n}\alpha_j y_j x_j \\ &= \frac{1}{2}\sum_{j=1}^{n}\alpha_j y_j(w^Tx_j) \\ &= \frac{1}{2}\sum_{j=1}^{n}\alpha_j y_j(\sum_{i=1}^{n}\alpha_i y_i x_i^T x_j) \\ &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j\end{aligned}$$

$$\begin{aligned}-\sum_{i=1}^{n}\alpha_i(y_i(w^Tx_i+b)-1) &= -\sum_{i=1}^{n}\alpha_i y_i(w^Tx_i+b)+\sum_{i=1}^{n}\alpha_i \\ &= -\sum_{i=1}^{n}\alpha_i y_i w^T x_i - b\sum_{i=1}^{n}\alpha_i y_i + \sum_{i=1}^{n}\alpha_i \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^{n}\alpha_i\end{aligned}$$

1. Margin을 최대화 시키는 것이 오류를 줄이는데 도움이 됨
2. 식 변형을 통해 Margin을 최소화 하고자 함
3. Primal problem은 너무 복잡해 편안한 Dual problem으로 만듦
4. KKT condition,strong duality에 의해 objective function의 최소값을 최대화 시킴

Hard margin & Linearly separable

$$H = \{x \rightarrow sign(\mathbf{w} \cdot \mathbf{x} + b) \ : \ \mathbf{w} \in \mathbb{R}^d, \ b \in \mathbb{R}\}$$

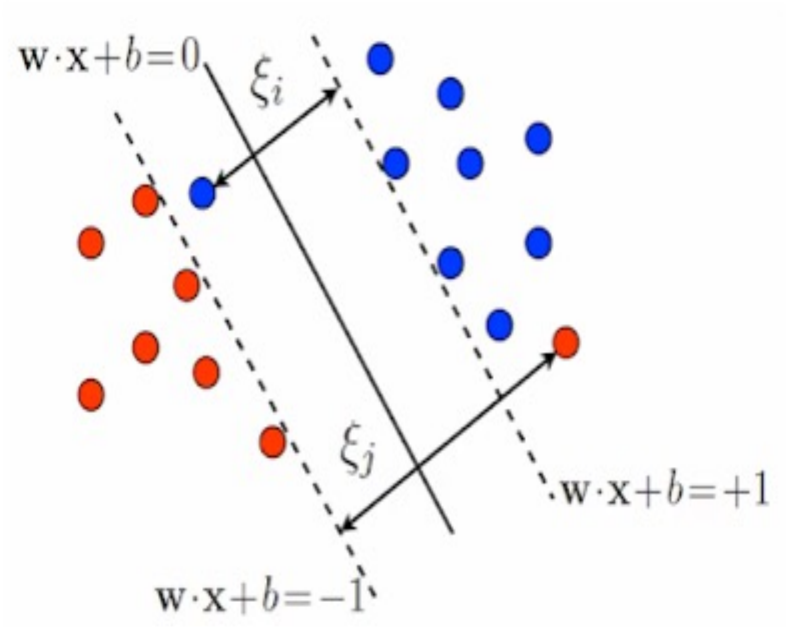
$$max \ \ L_P(\alpha_i) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i x_j$$

$$\mathbf{w} = \textstyle\sum_{i=1}^{N} \alpha_i y_i x_i$$

$$f(\mathbf{x}_{new}) = sign\left(\sum_{i=1}^{N} \alpha_i y_i x_i^T \mathbf{x}_{new} + b \right)$$

Soft margin & Linearly separable

- Soft margin : margin 안쪽에도 존재할 수 있도록 penalty를 주어 예외를 허용해주는 방법



Objective Function $$\min \quad \frac{1}{2}||\mathbf{w}||^2 + C\sum^{N} \xi_i$$

Constraint $$s.t. \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \ \forall i$$

Soft margin & Linearly separable

- KKT condition & duality example

<Lagrange problem>

Objective Function $\min \quad \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i$

Constraint $s.t. \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \ \forall i$

Lagrange Multiplier

$$\min \quad L_P(\mathbf{w}, b, \alpha_i) = \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i \boxed{- \sum_{i=1}^{N}\alpha_i(y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i) - \sum_{i=1}^{N}\mu_i\xi_i}$$

$$s.t. \quad \alpha_i \geq 0, \quad \mu_i \geq 0$$

KKT condition

- $\partial_x(f(x) + \sum u_i g_i(x) + \sum \lambda_j h_j(x)) = 0$

$$\frac{\partial L_P}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{N}\alpha_i y_i \mathbf{x}_i = 0 \qquad \mathbf{w} = \sum_{i=1}^{N}\alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial L_P}{\partial b} = \sum_{i=1}^{N}\alpha_i y_i = 0 \qquad \sum_{i=1}^{N}\alpha_i y_i = 0$$

$$\frac{\partial L_P}{\partial \xi_i} = C - \alpha_i - \mu_i = 0 \qquad C - \alpha_i - \mu_i = 0$$

Soft margin & Linearly separable

$$L_P = \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i(y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i) - \sum_{i=1}^{N}\mu_i\xi_i$$

$$L_D = \frac{1}{2}||\sum_{i=1}^{N}\alpha_i y_i \mathbf{x}_i||^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i(y_i((\sum_{i=1}^{N}\alpha_i y_i \mathbf{x}_i)\mathbf{x}_i + b) - 1 + \xi_i) - \sum_{i=1}^{N}\mu_i\xi_i$$

$$= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i x_j + \sum_{i=1}^{N}\alpha_i + \sum_{i=1}^{N}(C - \alpha_i - \mu_i)\xi_i \qquad s.t. \qquad C - \alpha_i - \mu_i = 0$$

$$\frac{\partial L_P}{\partial \xi_i} = C - \alpha_i - \mu_i = 0$$

$\alpha_i \geq 0$, $\mu_i \geq 0$ 이므로 $0 \leq \alpha_i \leq C$

$$max \quad L_D(\alpha_i) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j$$

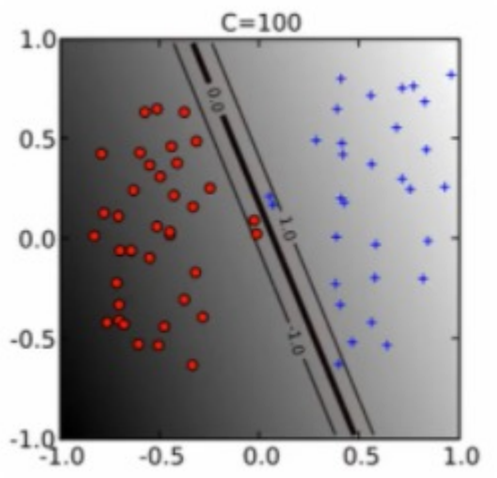
$$s.t. \qquad \sum_{i=0}^{N}\alpha_i y_i = 0 \quad and \quad \boxed{0 \leq \alpha_i \leq C}$$

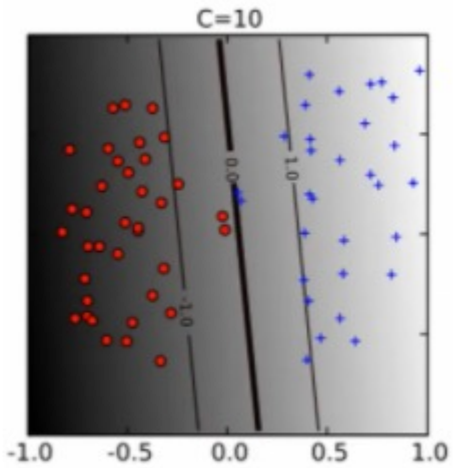
Hard margin과 다른 점

Unit 03 | Soft margin & Linearly separable

Soft margin & Linearly separable

- C parameter



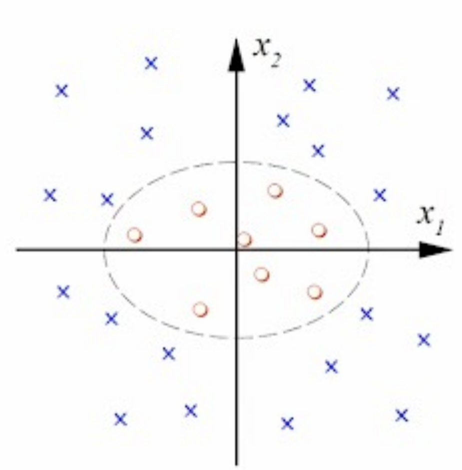


C가 크다면, ξi에 대한 penalty를 크게 하여 ξi를 최대한 축소.
- 예외를 최대한 허용하지 않겠다는 의미.
- margin의 크기가 작아짐.

C가 작다면, ξi에 대한 penalty를 적게 하여 ξi를 어느 정도 크게 함.
- 예외를 어느 정도 허용하겠다는 의미.
- margin의 크기가 커짐.

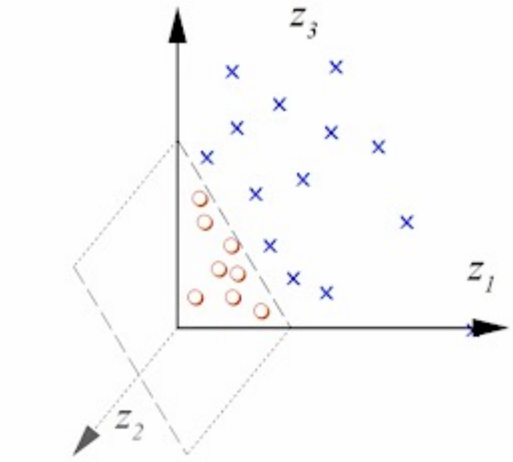
Soft margin & Linearly non-separable

- Mapping function : 비선형 문제 해결



<Mapping function>

$$\Phi : \mathbb{R}^2 \rightarrow \mathbb{R}^3$$
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{2}\, x_1 x_2, x_2^2)$$



Soft margin & Linearly non-separable

- KKT condition & duality example

Objective Function $min \quad \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i$

Constraint $s.t. \quad y_i(\mathbf{w}^T \boxed{\Phi(}\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \ \forall i$

Soft margin과 다른 점

<Lagrange problem>

Lagrange Multiplier

$$min \quad L_P(\mathbf{w}, b, \alpha_i) = \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i \boxed{- \sum_{i=1}^{N}\alpha_i(y_i(\mathbf{w}^T\Phi(\mathbf{x}_i) + b) - 1 + \xi_i) - \sum_{i=1}^{N}\mu_i\xi_i}$$

<Dual problem>

$$max \quad L_D(\alpha_i) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \boxed{\Phi(\mathbf{x}_i)^T\Phi(\mathbf{x}_j)}$$

Soft margin과 다른 점

$$s.t. \quad \sum_{i=0}^{N}\alpha_i y_i = 0 \quad and \quad 0 \leq \alpha_i \leq C$$

Soft margin & Linearly non-separable

- Kernel function

식이 매우 복잡하여, 서로 내적을 할 때 계산 시간이 많이 걸린다.

$$max \quad L_D(\alpha_i) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boxed{\Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j)} \rightarrow \mathbf{K}(\mathbf{x}_i^T, \mathbf{x}_j)$$

$$x = (x_1, x_2), \quad z = \Phi(x) = (1, x_1, x_2, x_1^2, x_2^2, x_1 x_2)$$

$$K(x, x') = z^T z' = 1 + x_1 x_1' + x_2 x_2' + x_1^2 x_1'^2 + x_2^2 x_2'^2 + x_1 x_1' x_2 x_2'$$

- Mercer`s condition

Scalar

$$\Phi(x) = Ax \qquad K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j) = \boxed{x_i^T A^T A x_j}$$

$$
\begin{aligned}
linear &: \quad K(x_1, x_2) = x_1^T x_2 \\
polynomial &: \quad K(x_1, x_2) = (x_1^T x_2 + c)^d, \quad c > 0 \\
sigmoid &: \quad K(x_1, x_2) = \tanh\{a(x_1^T x_2) + b\}, \quad a, b \geq 0 \\
gaussian &: \quad K(x_1, x_2) = exp\left\{-\frac{\|x_1 - x_2\|_2^2}{2\sigma^2}\right\}, \quad \sigma \neq 0
\end{aligned}
$$

For any symmetric function K: X × X → R which is square integrable (L_2-space) in X × X and which satisfies

$$\int_{X \times X} f(\mathbf{x}) K(\mathbf{x}, \mathbf{x}') f(\mathbf{x}') dx dx' \geq 0 \text{ for all } f \in L_2(X)$$

there exist functions $\phi_i : X \rightarrow R$ and numbers $\lambda_i \geq 0$ such that

$$K(\mathbf{x}, \mathbf{x}') = \sum_i \lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{x}') \text{ for all } \mathbf{x}, \mathbf{x}' \in X$$

Soft margin & Linearly non-separable

- Kernel function

<Polynomial Kernel>

$$K(x,x') = (1 + x^T x')^Q = (1 + x_1 x_1' + x_2 x_2' + \cdots + x_d x_d')^Q \rightarrow K(x,x') = (a x^T x' + b)^Q$$

<Gaussian(RBF) Kernel>

$$\begin{aligned} K(x,x') &= exp(-(x-x')^2) \\ &= exp(-x^2)exp(-x'^2)exp(2xx') \\ &= exp(-x^2)exp(-x'^2)\sum_{k=0}^{\infty}\frac{2^k x^k x'^2}{k!} \end{aligned}$$

<Taylor expansion of exponential function>

$$\boxed{e^x = \sum_{n=0}^{\infty}\frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \quad for\ all\ x}$$

Soft margin & Linearly non-separable

- Kernel function

<Kernel Function의 장점>

<Polynomial Kernel>

$$K(x,x') = (1 + x^T x')^Q = (1 + x_1 x'_1 + x_2 x'_2 + \cdots + x_d x'_d)^Q \rightarrow K(x,x') = (a x^T x' + b)^Q$$

- Efficiency : 일반적으로 내적을 계산하는 것 보다 Kernel function을 사용해서 한번에 결과를 얻는 것이 효율적.
- Flexibility : Mercer's condition을 만족하는 모든 함수를 Kernel function으로 사용할 수 있어 유연성이 높다.

<Gaussian(RBF) Kernel>

$$\begin{aligned} K(x,x') &= exp(-(x-x')^2) \\ &= exp(-x^2)exp(-x'^2)exp(2xx') \\ &= exp(-x^2)exp(-x'^2)\sum_{k=0}^{\infty}\frac{2^k x^k x'^2}{k!} \end{aligned}$$

<Taylor expansion of exponential function>

$$\boxed{e^x = \sum_{n=0}^{\infty}\frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \quad for\ all\ x}$$

Unit 05 | Summary

Summary

1. 분류를 잘하는 이진 선형 분류기를 선택해야 해
2. Margin이 넓을 수록 VC dimension이 줄어 test error가 줄어들어
3. 제약식이 있어 쉽게 해를 찾지 못해
4. 라그랑주 승수를 사용해서 dual problem으로 식을 변형 해줘
5. KKT condition과 strong duality 성질을 이용해서 적절한 식으로 변형 시켜줘 (minimize -> maximize)
6. 예외 허용과 비선형적 condition으로 인해 3가지 case로 분리

Unit 05 | Summary

Summary

	Hard Margin & Linear	Soft Margin & Linear	Soft Margin & non-Linear
Summary	Margin을 최대화 하여 test error를 줄이는 방법	Margin을 최대화 하나,margin 안쪽에도 data 존재할 수 있도록 penalty를 주어 예외를 허용해주는 방법	Margin을 최대화 하나, non-linear data에 대해서 linear data으로 변형하는 방법
Objective function	$minimize \quad (\frac{1}{margin}) = \frac{1}{2}\|\|w\|\|^2$ $y_i(w^T x_i + b) \geq 1$	$min \quad \frac{1}{2}\|\|\mathbf{w}\|\|^2 + C\sum_{i=1}^{N}\xi_i$ $s.t. \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \ \forall i$	$min \quad \frac{1}{2}\|\|\mathbf{w}\|\|^2 + C\sum_{i=1}^{N}\xi_i$ $s.t. \quad y_i(\mathbf{w}^T\Phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \ \forall i$
Dual problem	$\max L_D(\alpha_i) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j$	$\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j$	$\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \Phi(\mathbf{x}_i)^T\Phi(\mathbf{x}_j)$

Unit 05 | Summary

Assignment

1. Multiclass SVM을 직접 구현하시는 것입니다. 기본적으로 사이킷런에 있는 SVM은 멀티클래스 SVM을 지원하지만 과제에서는 절대 쓰면 안됩니다! Iris 데이터는 총 세 개의 클래스가 있으므로 이 클래스를 one- hot인코딩한 뒤, 각각 binary SVM을 트레이닝하고 이 결과를 조합하여 multiclass SVM을 구현하시면 됩니다

2. 기본적으로 one vs one, one vs rest 방법이 있으며 둘 중 자유롭게 구현해주세요. 만약 투표결과가 동점으로 나온 경우(예를 들어, 각각의 SVM 결과가 A vs B 의 경우 A로 판별, B vs A 의 경우 B로 판별, C vs A 의 경우 C로 판별한 경우 투표를 통해 Class를 결정할 수 없음)

1) decision_function을 활용하시거나
2) 가장 개수가 많은 클래스를 사용하시거나
3) 랜덤으로 하나를 뽑거나 하는 방법 등을 이용해 동점자인 경우를 판별 해주시면 됩니다.

공식문서를 통해 사이킷런이 어떤 방법으로 구현했는지 참고하셔도 됩니다

3. 과제코드에는 iris 데이터를 로드하고 스케일링 부분까지 구현되어 있습니다.

4. Iris의 클래스는 3개입니다. Iris 데이터셋 뿐만 아니라 다른 데이터셋에도 적용 가능한, 클래스의 수와 무관한 Multiclass SVM을 만들어주세요.

Unit 05 | Summary

Assignment

• One vs One(OVO)
➔ 클래스가 N개 있을 때, 모든 클래스에 대해 1:1로 binary 분류를 하고, 제일 많이 승리한 클래스에 대해 투표로 결정!
➔ N개의 클래스에 대해 서로 다른 Classifier를 가지고 있어야 하기 때문에 n(n-1)/2 개의 Classifier가 필요함

• One vs Rest(OVR)
➔ 클래스가 N개 있으면 모든 클래스에 대해 1:N-1 로 binary 분류하여 이 클래스가 맞는지 아닌지를 투표로 결정!
➔ N개의 Classifier가 필요함

https://ratsgo.github.io/convex%20optimization/2018/01/25/duality/
https://ratsgo.github.io/machine%20learning/2017/05/23/SVM/3
https://yngie-c.github.io/machine%20learning/2021/03/07/svm/#fn:1
https://yupsung.blogspot.com/2021/01/022-kernel-based-learning-support.html

Q & A

들어주셔서 감사합니다.