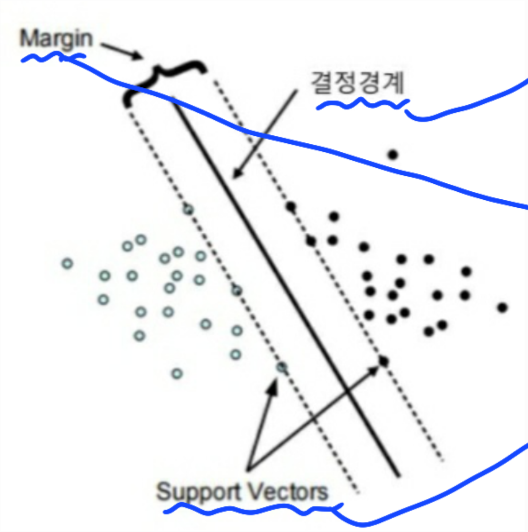
SVM(Support Vector Machine)

: Large Margin Classification이라고도 함

: SVM 분류기는 클래스 사이에 가장 폭이 넓은 마진을 찾는 것

: 특성의 스케일에 민감



데이터의 분류 기준이 되는 경계

결정 경계에서 서포트 벡터까지의 거리

학습 데이터중 결정 경계와 가장 가까이 있는 데이터들의 집합

하드 마진 SVM (Hard Margin SVM)

- 오분류를 허용하지 않는 SVM → 마진오류 엄격
- 노이즈로 인해 최적의 결정 경계를 잘 못 구하거나, 못 찾을 수 있음

소프트 마진 SVM (Soft Margin SVM)

- 오분류를 허용하는 SVM
- 하드 마진 SVM은 적용하기가 어려우므로 어느 정도의 오류를 허용하는 소프트 마진 SVM을 주로 이용

마진오류 결정파라미터 : C → C↑ → 마진오류⇓ ⇒ Hard margin classification

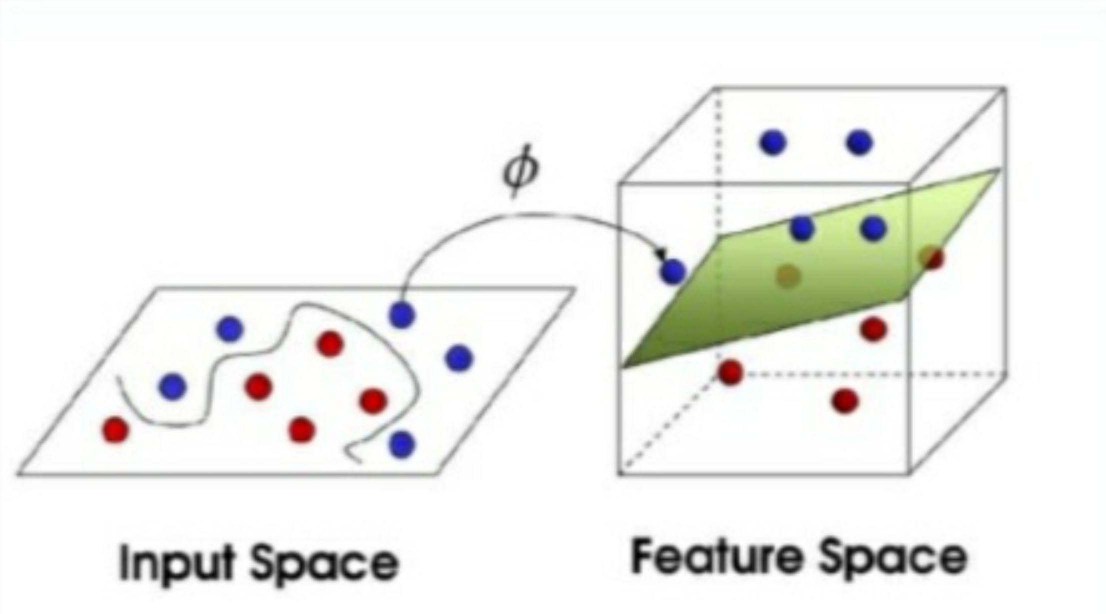
C↓ → 마진오류⇑ ⇒ Soft Margin classification

If svm overfit ⇒ C⇓

비선형 SVM

Kernel Trick을 이용하여 매핑

-> 고차원 공간으로 매핑하는 경우에 증가하는 연산량의 문제를 해결



SVM kernel => poly로 설정

coef0 => 다항식 커널에 있는 상수항 — coef0 ⇑
=> 고차항의 영향 ⇓

유사도 특성

: 각 샘플이 특정 랜드마크와 얼마나 닮았는지 측정하는 것

- 가우시안 RBF (유사도 함수)

$\phi_\gamma(x, \ell) = \exp(-\gamma \|x - \ell\|^2)$ => 0~1 사이 (0은 랜드마크와 아주 멀리 떨어진 경우)
종모양 (1은 랜드마크와 같은 위치인 경우)

가우시안 RBF kernel

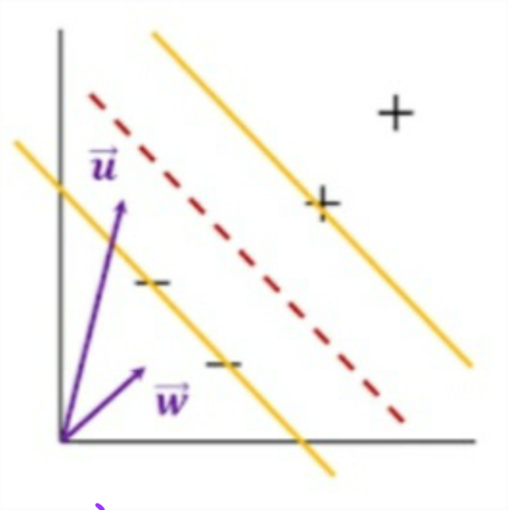
kernel = rbf, gamma

- γ ⇑ → 종모양(RBF) 좁아짐 => 샘플의 영향 범위 작아짐
- γ ⇓ → 종모양(RBF) 넓어짐 => 결정경계 부드러워짐

⇓

규제역할!! => SVM Overfit => γ ⇓
∴ underfit => γ ⇑

SVM 수식



$\vec{u}$가 경계면을 기준으로 어디 (+, -)에 속하는가?

⇒ $\vec{w}\cdot\vec{u} \ge C$ (상수) ⇒ + else −

$$\boxed{\begin{array}{l}\vec{w}\cdot x_+ + b \ge 1 \\ \vec{w}\cdot x_- + b \le -1\end{array}}$$

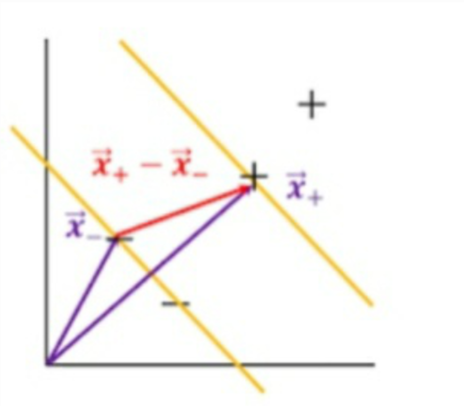
(x_+: + 샘플, x_-: − 샘플)

$$\Rightarrow \boxed{y_i \begin{cases} 1 & \text{for } + \\ -1 & \text{for } - \end{cases}}$$

$$\Rightarrow \underbrace{y_i(\vec{w}\cdot x_i + b) - 1 \ge 0}_{\text{s.t (제약식)}}$$

$\vec{w}$: 경계면에 직교하는 벡터

$\vec{u}$: 임의의 벡터



목적: Margin 폭을 최대로

↳ $(\vec{x}_+ - \vec{x}_-)\cdot \dfrac{\vec{w}}{\|\vec{w}\|}$

$$= \underbrace{\begin{array}{l}\vec{w}\cdot x_+ + b \ge 1 \\ -\ \vec{w}\cdot x_- + b \le -1\end{array}}_{\vec{w}(x_+ - x_-) \ge 2}$$

→ WIDTH = $\dfrac{2}{\|\vec{w}\|}$

라그랑지 승수법 이용

$$L(w, b, \alpha) = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{N} \alpha_i [y_i(\vec{w}\cdot\vec{x}_i + b) - 1]$$

$\vec{w}$에 대해 편미분 : $\nabla_{\vec{w}} L = \vec{w} - \sum \alpha_i y_i \vec{x}_i = 0$

b에 대해 편미분 : $\nabla_b L = -\sum \alpha_i y_i = 0$

→ $\vec{w} = \sum_i \alpha_i y_i \vec{x}_i$

α에 대한 maximization 문제로 정리

1. α를 구하면 $\vec{w}$ 구할 수 있음
2. α ≠ 0 , $\vec{x}$가 경계선을 정하는 샘플 즉, Support Vector
3. SVM으로 구한 해는 라그랑지 최적화 이론에 의해 최적의 해 증명 (local Minima에 안빠짐)

SVM 강의

$w^T x + b = 0$

- w → normal vector hyperplane
- b → bias

SVM에서 좋다는 기준

⇒ Margin을 최대화 = test 에러 최소화

$w^T x + b = 0$ → 이걸 찾는게 목적

Margin : 각 클래스에서 가장 가까운 관측치 사이의 거리
: w로 표현 가능

$$w^T x^+ + b = 1$$
$$w^T (x^- + \lambda w) + b = 1 \quad (x^+ = x^- + \lambda w)$$
$$\underbrace{w^T x^- + b}_{-1} + \lambda w^T w = 1$$
$$-1 + \lambda w^T w = 1 \quad \therefore \lambda = \frac{2}{w^T w}$$

※ 알아야 할 것

Vector norm $= \|W\|_p = \left(\sum_i |w_i|^p\right)^{\frac{1}{p}}$

ex) L_2 Norm

$$\|W\|_2 = \left(\sum |w_i|^2\right)^{\frac{1}{2}} = \sqrt{w_1^2 + w_2^2 + \cdots w_n^2} = \sqrt{w^T w}$$

$$\text{Margin} = \text{distance}(x^+, x^-) = \|x^+ - x^-\|_2 = \|x^- + \lambda w - x^-\|_2 = \|\lambda w\|_2$$
$$= \lambda\sqrt{w^T w} = \frac{2}{w^T w} \cdot \sqrt{w^T w} = \frac{2}{\sqrt{w^T w}} = \frac{2}{\|w\|_2}$$

max Margin $= \max \frac{2}{\|W\|_2} \iff \min \frac{1}{2}\|w_2\|$ (역수) → 계산 편의상 제곱 ⇒ $\min \frac{1}{2}\|w\|_2^2$

$\therefore$ minimize $\frac{1}{2}\|w\|_2^2$ ← 목적식

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

↳ 마진을 1보다 크게하자

- 결정변수 w, b
- 목적함수 : margin의 역수
- 제약식 : train data를 완벽하게 분리하는 조건
- 목적식은 2차식 이고 제약식은 선형이다 ⇒ 2차계획법 (QP) ⇒ convex optimization (전역최적해 존재)

라그랑지 Primal

$$\boxed{\begin{aligned} &\min L(w,b,a) = \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{n} a_i\left(y_i(w^T x_i + b) - 1\right) \\ &\text{subject to } a_i \geq 0,\ i = 1,2,3 \cdots n \end{aligned}}$$

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

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

① $\frac{1}{2}\|w\|_2^2 = \frac{1}{2}w^T w = \frac{1}{2} w^T \sum_j^n a_j y_j x_j = \frac{1}{2}\sum_j^n a_j y_j (w^T x_j) = \frac{1}{2}\sum_j^n a_j y_j \left(\sum_i^n a_i y_i x_i^T x_j\right) = \frac{1}{2}\sum_i^n\sum_j^n a_i a_j y_i y_j x_i^T x_j$

② $-\sum_i^n a_i\left(y_i(w^T x_i + b) - 1\right) = -\sum_i^n a_i y_i (w^T x_i + b) + \sum_i^n a_i = -\sum_{i=1}^n a_i y_i w^T x_i - b\underbrace{\sum_i^n a_i y_i}_{0} + \sum_i^n a_i = -\sum_i^n\sum_j^n a_i a_j y_i y_j x_i^T x_j + \sum_i^n a_i$

→ $\underset{a}{\text{maximize}} \sum_{i=1}^n a_i - \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n a_i a_j y_i y_j x_i^T x_j$

subject to $\sum_{i=1}^n a_i y_i = 0$, $a_i \geq 0$, $i = 1,2,3,4 \cdots n$

$$\underset{a}{\text{maximize}} \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j y_i y_j x_i^T x_j$$

2차

내적

subject to $\sum_{i=1}^{n} \alpha_i y_i = 0$, $\alpha_i \geq 0$, $i = 1, 2, 3, 4 \cdots n$

↳ linear

전역 최적해 존재

=> 내적으로 표현 되어
non-linear 확장시 좋은 성질

KKT (쿤 커터) 상태

① Stationarity (미분시 0이 되는 점 존재)

$\frac{\partial L(w,b,a)}{\partial w} = 0 \Rightarrow w = \sum_{i}^{n} a_i x_i y_i$ $\quad \frac{\partial L(w,b,a)}{\partial b} = \sum_{i}^{n} a_i y_i = 0$

② Primal feasibility $y_i(w^T x_i + b) \geq 1$

③ Dual feasibility $a_i \geq 0$, $i = 1, 2 \ldots n$

④ Complementary feasibility (②×③)

$a_i(y_i(w^T x_i + b) - 1) = 0$

$a_i(y_i(w^T x_i + b) - 1) = 0$

① $\alpha_i > 0$ and $y_i(w^T x_i + b) - 1 = 0$ → $y_i(w^T x_i + b) = 1$ ⇒ $a_i > 0$ plus-plane

② $\alpha_i = 0$ and $y_i(w^T x_i + b) - 1 \neq 0$ $\quad a_i = 0$ → Hyper plane 를 구축하는데 영향 X

↳ 이상치에 강건한 이유

$$\boxed{\begin{aligned} w^* &= \sum_{i \in sv} a_i^* y_i x_i \\ (\alpha_i^* &\geq 0) \end{aligned}}$$

$w^{*T} + b^* = y_{sv}$

$w^{*T} + b^* = \sum_{i}^{n} a_i^* y_i x_i^T x_{sv} + b^* = y_{sv}$

$$\boxed{= b^* = y_{sv} - \sum^{n} a_i^* y_i x_i^T x_{sv}}$$

$w^{*T} x_{new} + b^* < 0 \Rightarrow \hat{y}_{new} = -1$

$w^{*T} x_{new} + b^* > 0 \Rightarrow \hat{y}_{new} = 1$

$\hat{y}_{new} = \text{sign}(w^{*T} x_{new} + b^*)$

$= \text{sign}(\sum a_i^* y_i x_i^T x_{new} + b^*)$