인공지능을 위한 머신러닝 알고리즘

5. 서포트 벡터 머신

Tacademy

CONTENTS

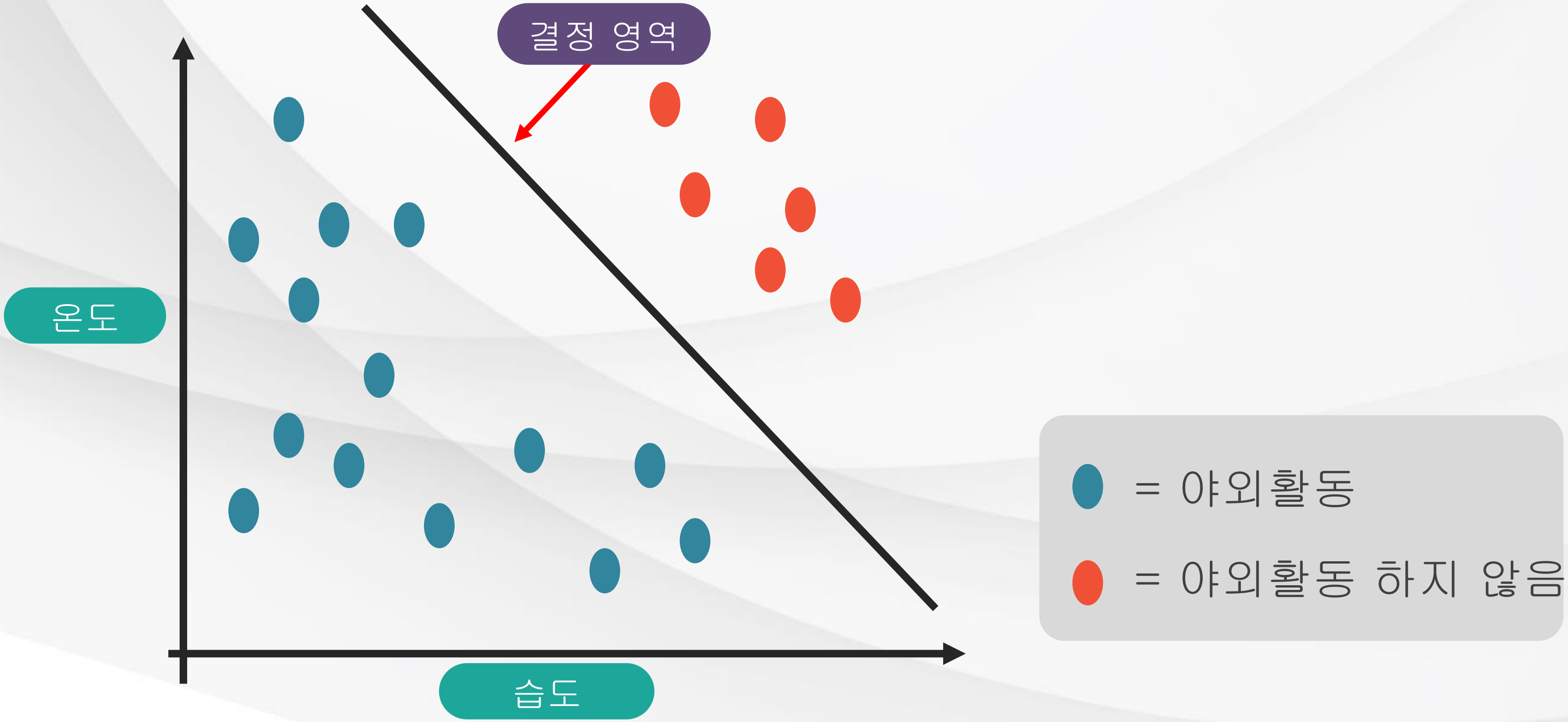
학습 목표

- 서포트 벡터 머신의 분류 원리에 대해서 이해할 수 있다.
- 서포트 벡터 머신의 학습이 최적화 문제로 변형되어 해결되는 과정을 이해할 수 있다.
- 소프트 마진 분류기와 커널을 사용한 비선형 분류기를 이해할 수 있다.

1. 좋은 선형 분류기 만들기

선형 분류의 예

❖ 예> 야외활동하기 좋은 날 분류하기

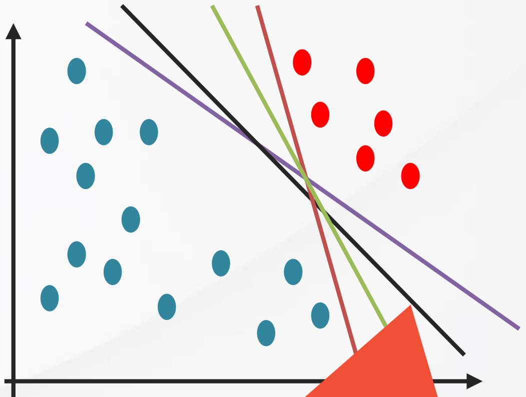


어떤 초평면을 선택해야 할까?

- 초평면이란?
  데이터 임베딩 공간에서 한 차원 낮은 부분 공간 (subspace)

  예> 3차원 공간의 초평면: 2차원 평면

- 가능한 **a, b, c**에 대해 많은 해답이 존재
- 여러 선형 분류기들이 있지만
  항상 최적의 초평면을 찾지는 않음 예> 퍼셉트론
- 최적의 초평면의 조건
  - 초평면과 결정영역 근처에 있는 '분류하기 애매한' 데이터의 거리가 최대가 되어야 함
  - 직관적인 이해 : 만약 결정 영역 근처에 데이터가 없다면,
    분류기는 분류 결정을 하기 위해 불확실하고
    애매한 결정을 내리는 경우가 조금 더 드물어질 것임



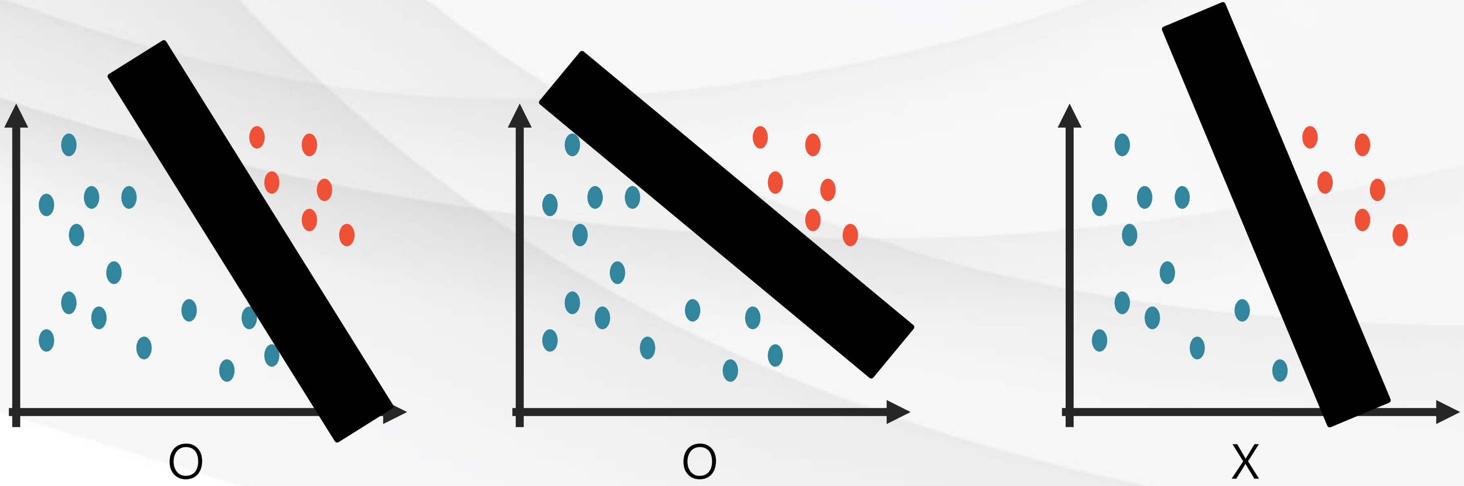
이 선은 결정영역을 나타냄
$ax + by - c = 0$

1. 좋은 선형 분류기 만들기

또 다른 직관

- 두꺼운 결정영역을 클래스 사이에 놓는다면, 선택의 가짓수는 보다 줄어듦

- 모델의 파라미터 개수를 줄일 수 있음



2. 서포트 벡터 머신

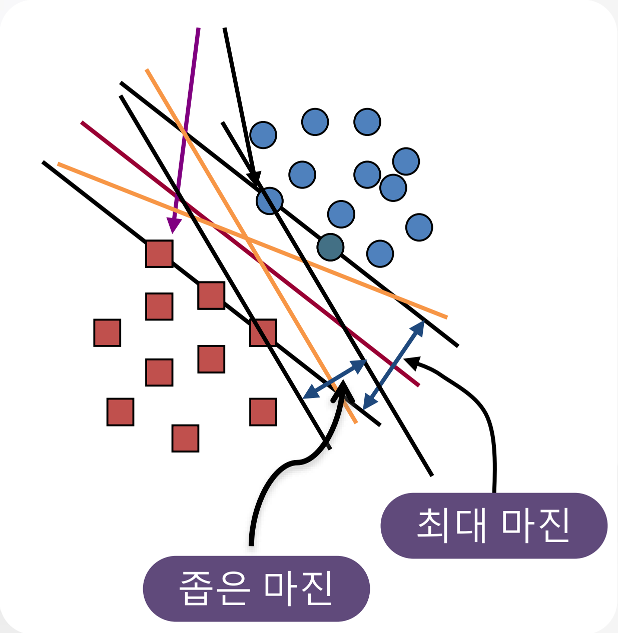
서포트 벡터 머신의 분류 원리

- 서포트 벡터 머신은 결정 영역의 초평면을 둘러싸고 있는 마진(**margin**)을 최대화시킴

예> 3차원 공간의 초평면 : 2차원 평면

- 클래스 결정 함수는 훈련 예제의 부분 집합(서포트 벡터)만으로 완전히 설명 가능
- 서포트 벡터 머신의 결정 함수를 구하는 것은 함수 최적화 문제
- 딥러닝이 나오기 전까지 가장 성공적인 분류기 역할을 했음

서포트 벡터 (**support vectors**)



서포트 벡터 머신 형식

w	정규화된 결정 초평면 벡터
x_i	데이터 포인트 i
y_i	데이터 포인트 i의 클래스 (+1 또는 -1)

서포트 벡터 머신 형식

- 분류기 형식

$$y_i = f(x_i) = \text{sign}(w^T x_i + b)$$

- x_i의 기능적 마진 (functional margin)

$$y_i(w^T x_i + b)$$

  - 주어진 데이터 포인트가 적절하게 분류되었는지 아닌지 가늠할 수 있는 테스트 함수의 역할
  - 값이 클수록 적절하게 분류되었는지 확인할 수 있음
  - w와 b의 크기를 키움으로써 마진의 크기도 증가시킬 수 있음

sign함수란?

- 수의 부호를 판별하는 함수

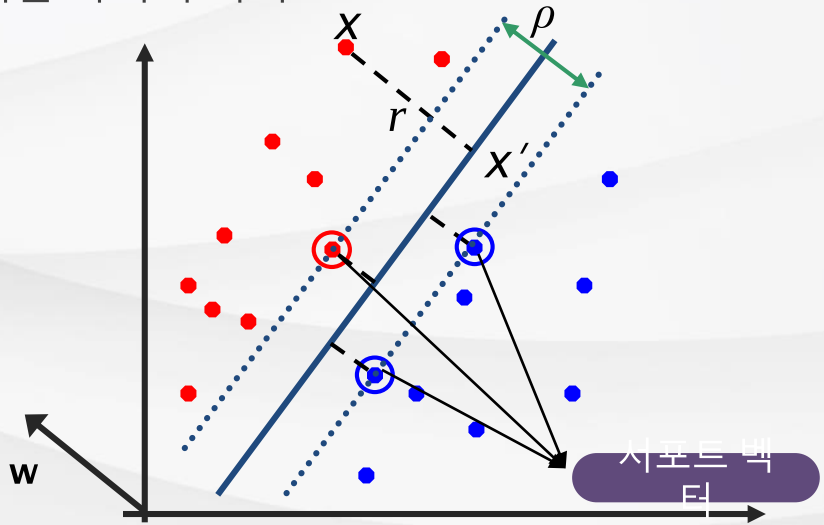
$$y_i = +1 \text{ when } w^T x_i + b \geq +1$$
$$y_i = -1 \text{ when } w^T x_i + b \leq -1$$

기하 마진 (geometric margin)

- 데이터 포인트에서 결정 영역까지의 거리 : $r = y\frac{\mathbf{w}^T\mathbf{x} + b}{\|w\|}$
- 결정 영역까지 가장 가까운 데이터들을 서포트 벡터(support vectors)라고 함
- 결정 영역의 마진 ρ 는 클래스들의 서포트 벡터들 사이의 거리

r 을 계산하기 위한 과정

점선 **x'−x** 은 결정 영역에 직교하고 **w**에 평행함
유닛 벡터는 **w/|w|** 이고, 점선은 **rw/|w|**임
x' = x – yrw/|w|
x'는 $\mathbf{w^Tx'+b = 0}$을 만족함
그러므로 $\mathbf{w^T(x - yrw/|w|) + b = 0}$
$\mathbf{|w| = sqrt(w^Tw)}$이고, $\mathbf{w^Tx - yr|w| + b = 0}$
r에 대해서 정리하면, $\mathbf{r = y(w^Tx + b)/|w|}$



선형 서포트 벡터 머신

- 모든 데이터의 기능적 마진이 항상 1 이상이라고 가정한다면, 다음 두 개의 조건이 훈련 데이터 집합 $\{(x_i, y_i)\}$에 대해 만족함

$$w^T x_i + b \geq 1 \quad \text{if } y_i = 1$$

$$w^T x_i + b \leq -1 \quad \text{if } y_i = -1$$

- 서포트 벡터는 부등호가 등호로 바뀜
- 각 데이터 포인트로부터 결정영역까지의 거리
  - 기능적 마진을 파라미터 **w**에 따라 크기를 바꿔줌
  - 데이터 포인트가 적절하게 분류가 잘 되었는지, |**w**|로 스케일링 된 척도로 알려줌

$$r = y\frac{w^T x_i + b}{\|w\|}$$

- 마진 ρ : $\frac{2}{\|w\|}$

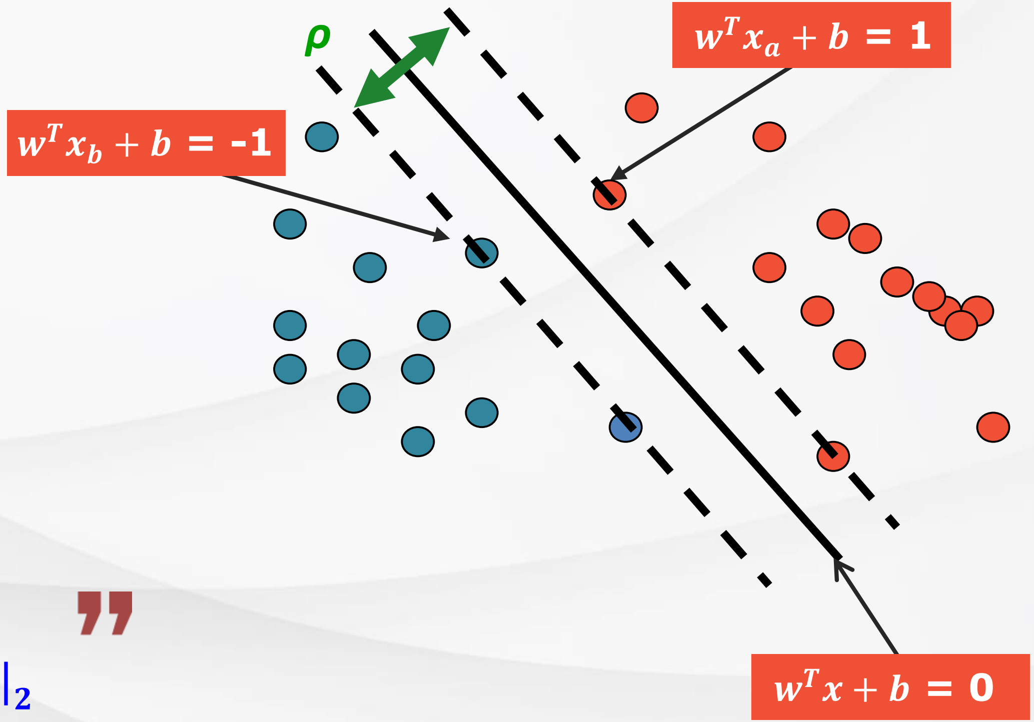
선형 서포트 벡터 머신 정리

- 초평면

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

- 추가 조건

$$min_{i=1\ldots n} |w^T x_i + b| = 1$$



즉,

$$w^T(x_a - x_b) = 2$$

$$\rho = \|x_a - x_b\|_2 = 2/\|w\|_2$$

최적화 문제를 사용한 파라미터 계산 (1/3)

- 서포트 벡터 머신의 파라미터를 찾기 위해서 최적화 문제로 변형시킬 수 있음

Find w and *b* such that

1 / ||w|| is maximized; and for all $\{(x_i, y_i)\}$

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

- 보다 나은 형식으로 변형 (min ||w|| = max 1/ ||w||)

Find w and *b* such that

$\Phi(\mathbf{w}) = \frac{1}{2} w^T w$ **is minimized;**

and for all $\{(x_i, y_i)\}$: $y_i\,(w^T x_i + b) \geq 1$

최적화 문제를 사용한 파라미터 계산 (2/3)

Find w and b such that
$\Phi(w) = ½\ w^T w$ is minimized ;
and for all $\{(x_i, y_i)\}$: $y_i\ (w^T x_i + b) \geq 1$

- 선형 조건에 부합하도록 이차함수를 최적화 시키는 문제
- 이차함수의 최적화 문제는 수학적 프로그래밍 문제에서 잘 알려진 분야로, 해결할 수 있는 많은 알고리즘이 존재함
- Lagrange multiplier α_i 을 사용하여 다음의 primal과 dual problem으로 변형 가능

Maximize
$L(w,b) = 1/2 w^T w\ - \Sigma \alpha_i \{ y_i (w^T x_i - b) - 1 \}$
(1) $\alpha_i \geq 0$ for all α_i

Find $\alpha_1 \ldots \alpha_N$ such that
$Q(\alpha) = \Sigma \alpha_i\ - ½ \Sigma\Sigma \alpha_i \alpha_j y_i y_j x_i^T x_j$ is maximized and
(1) $\Sigma \alpha_i y_i = 0$
(2) $\alpha_i \geq 0$ for all α_i

Tacademy

최적화 문제를 사용한 파라미터 계산 (3/3)

솔루션은 다음과 같은 형식을 가짐

$\mathbf{W} = \sum \alpha_i y_i X_i \quad b = y_k - w^T X_k$, k는 $\alpha_k \neq 0$ 을 만족

- 0이 아닌 α_i 는 해당하는 x_i가 서포트 벡터임을 의미

그러므로 분류함수는 다음과 같은 형식임

$f(\mathrm{X}) = \sum \alpha_i y_i X_i^{\mathrm{T}} \mathrm{X} + b$

- 분류는 새로운 테스트 데이터 $\mathbf{x}$와 서포트 벡터 x_i의 내적에 의해 계산됨

“ **하지만, 모델의 훈련 과정 때는**
모든 훈련 데이터 쌍 (x_i, x_j)에 대해 내적 $x_i^T x_j$을 계산 ”

3. 비선형 분류기 만들기

소프트 마진 분류 (soft margin classification)

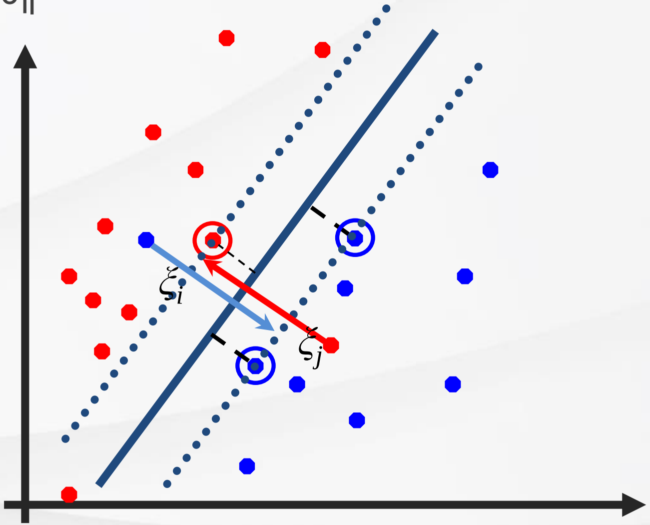
- 만약 훈련 데이터가 선형으로 분리되지 않을 경우,
  슬랙 변수 ξ_i 가 잘못 분류되거나 노이즈가 포함된 데이터에
  추가됨

- 잘못 분류된 데이터 포인트를 본래 속하는 클래스로
  비용을 들여 이동시켜줌

$$y_i(w^T x_i + b) \geq 1$$
$$\min\|w\|$$

➡

$$y_i(w^T x_i + b) \geq 1 - \xi_i$$
$$\min\|w\| + C\|\xi\|$$



- 모델의 학습 방법은 여전히 결정 영역을
  각 클래스로부터 가장 멀리 위치하는 것임
  (large margin)

소프트 마진 분류의 파라미터 계산

- 예전 문제 형식

> **Find w and b such that**
> **$\Phi(\mathrm{w}) = \frac{1}{2}\, \mathrm{w}^T\mathrm{w}$ is minimized and for all $\{(x_i, y_i)\}$**
> **$y_i(w^T x_i + b) \geq 1$**

- 새로운 형식은 슬랙 변수를 포함하고 있음

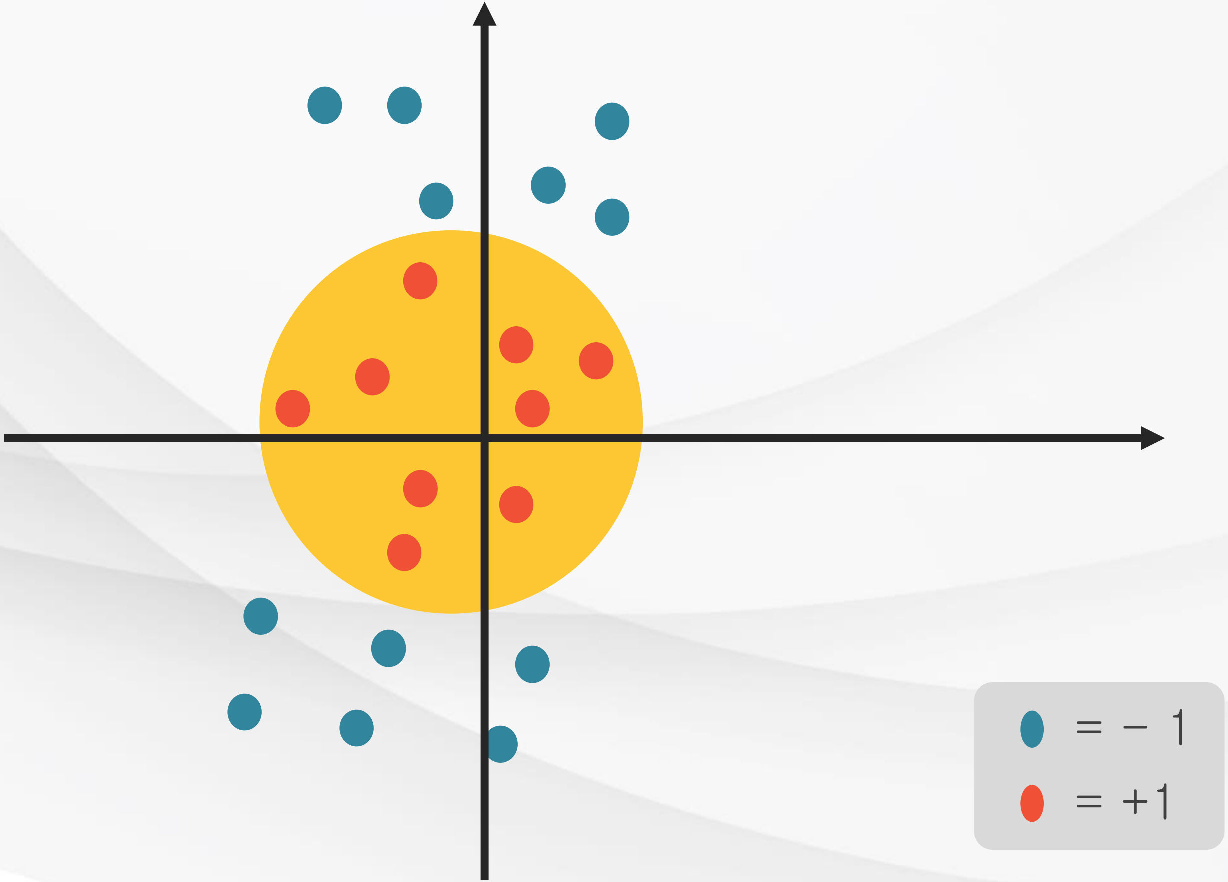
> **Find w and b such that**
> **$\Phi(\mathrm{w}) = \frac{1}{2}\, \mathrm{w}^T\mathrm{w} + C\Sigma\xi_i$ is minimized and for all $\{(x_i, y_i)\}$**
> **$y_i(w^T x_i + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$ for all i**

- 솔루션은 다음과 같은 형식을 가짐 (일반 서포트 벡터 머신과 유사)

$$W = \sum \alpha_i y_i X_i$$

$$b = y_k(1 - \xi_k) - w^T X_k \quad \text{where } k = \operatorname{argmax}\, \alpha_{k'}$$

결정 영역이 선형이 아닌 경우



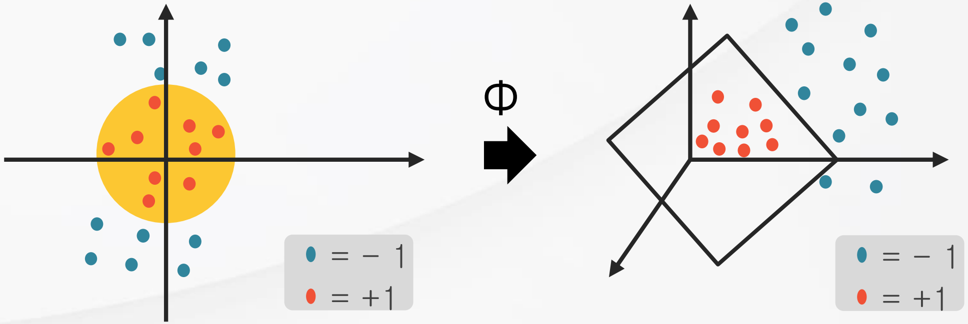
Tacademy

결정 영역이 선형이 아닌 경우

- 데이터 포인트를 선형으로 분류하기 위해 차원을 더 생성

$$(x_1, x_2) \Rightarrow (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

  - Φ : x → Φ(x)
  - f(x) = w^TΦ(x) + b를 학습
  - Φ(x)는 피처 맵이라고 부름



- 파라미터 결정

x대신 Φ(x)을 사용하여 다음 식을 계산

$$\text{maximize} \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} y_i y_j \alpha_i \alpha_j \langle \phi(x_i) \cdot \phi(x_j) \rangle$$

  - 커널 K($\boldsymbol{x_i}$, $\boldsymbol{x_j}$) = $\langle \phi(\boldsymbol{x_i}) \cdot \phi(\boldsymbol{x_j}) \rangle$ 이며 이 경우, $\Phi(\boldsymbol{x_i})^T \Phi(\boldsymbol{x_j}) = (\boldsymbol{x_i}^T \boldsymbol{x_j})^2$ 임

분류기로써 서포트 벡터 머신의 성능

- 서포트 벡터 머신은 실세계 데이터에 좋은 성능을 보여줌
  - 프로그래머는 커널 함수를 설계 해야 함
  - 나머지 파라미터는 자동으로 계산
- 데이터 집합의 크기가 클 수록 시간 소모가 큼
  - 초평면의 최대 마진을 구하기 위해서
    훈련 데이터 개수의 제곱에 해당하는 계산량 필요
  - 모든 서포트 벡터를 저장 해야 함

**"만약, 문제에 어떠한 알고리즘을 사용할지 모르겠다면,
서포트 벡터 머신은 좋은 출발선이 될 수 있음"**

학습정리

학습정리

지금까지 [서포트 벡터 머신]에 대해서 살펴보았습니다.

좋은 선형 분류기 만들기

서포트 벡터 x_i가 결정 영역의 초평면 w를 결정

$$y = f(\mathbf{x}) = \mathrm{sign}(\boldsymbol{w}^T\boldsymbol{x} + b) = \mathrm{sign}(\sum \alpha_i y_i X_i^T \boldsymbol{X}) + b$$

서포트 벡터 머신

이차함수의 최적화 문제를 통해 서포트 벡터와,
Lagrangian multiplier α_i를 계산할 수 있음

비선형 분류기 만들기

슬랙 변수 ξ_i 를 잘못 분류되거나 노이즈가 포함된 데이터에 추가하여
본래 속하는 클래스로 비용을 들여 이동시켜줌

$$y_i(\boldsymbol{w}^T\boldsymbol{x} + b) \geq 1 - \xi_i$$