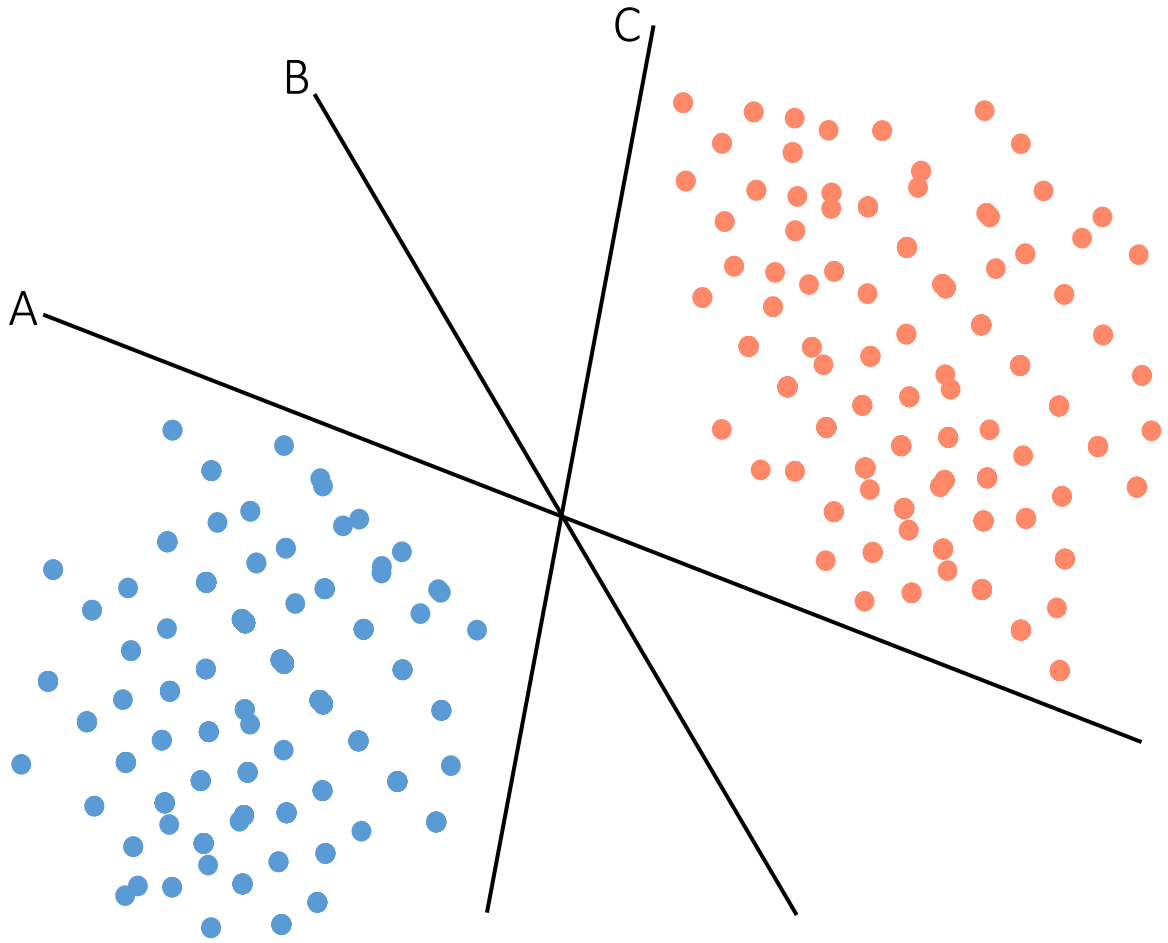
서포트 벡터 머신(SVM)

Support Vector Machine

Contents

1. 개요

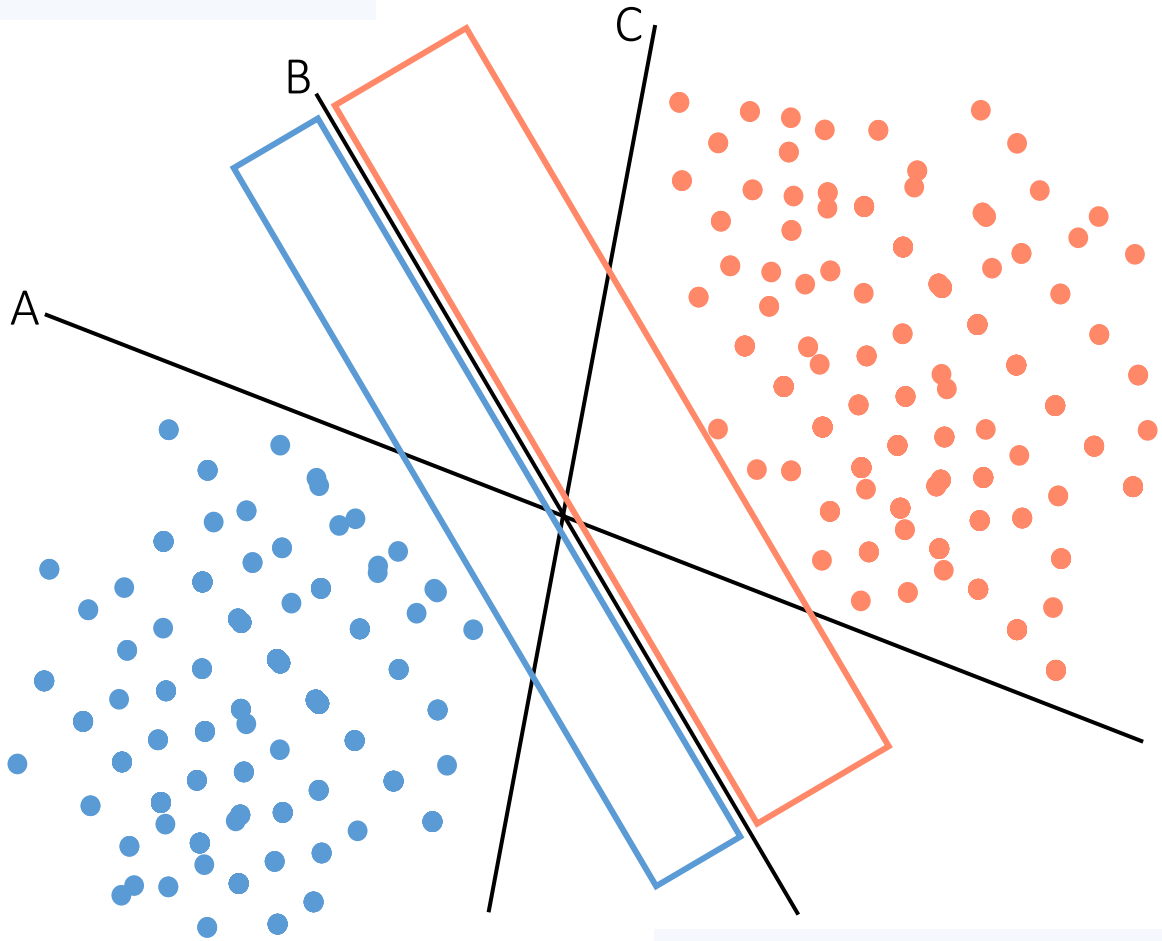
어떻게 분리할까?



두 집단을 가장 잘 분리하는 직선은?

가장 좋은 분리 방법은?

선과 그룹 사이의 공간

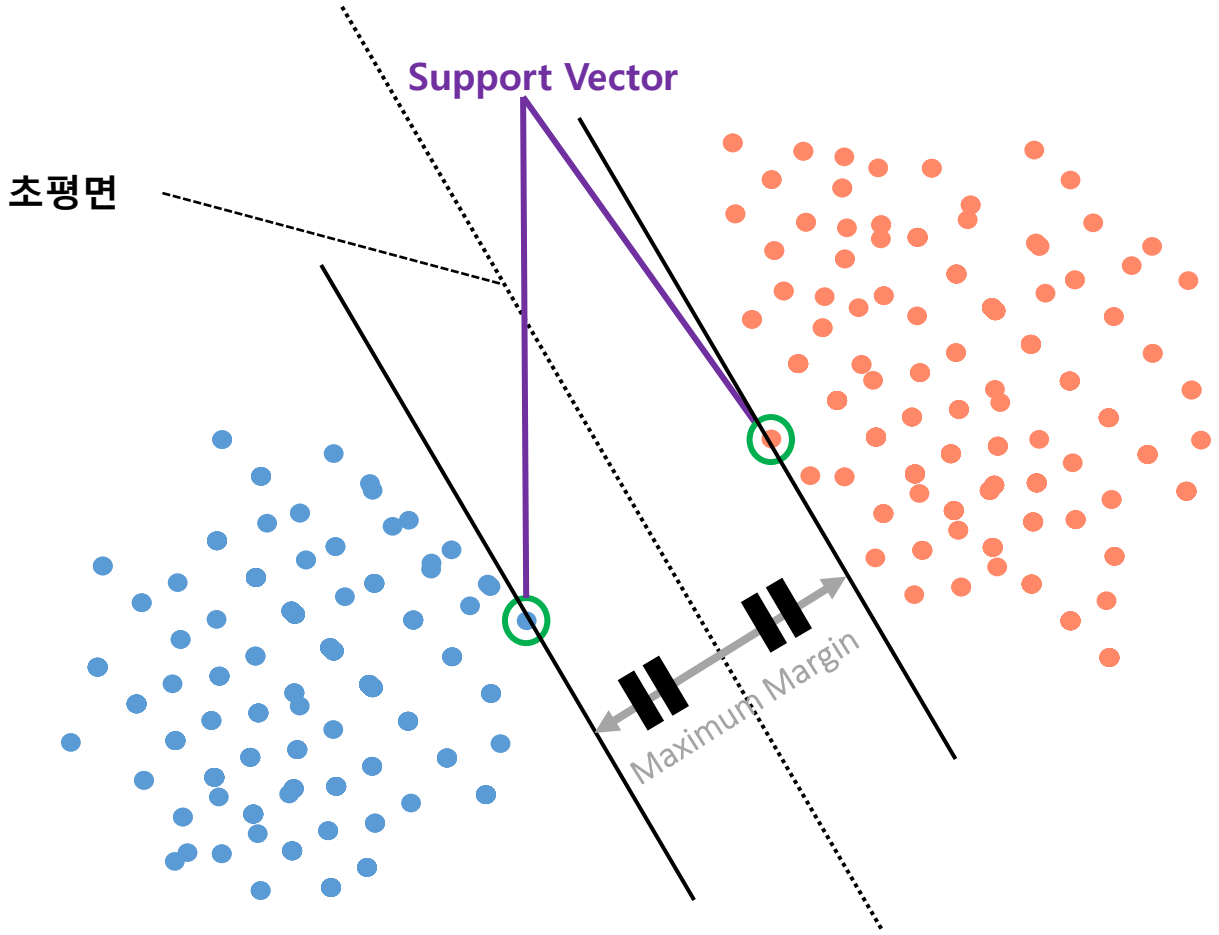


왜 B가 가장 좋은 선일까?

SVM이란?

두 범주의 데이터를 초평면으로 margin을 최대화하도록 분류하는 방법

SVM의 기본 원리



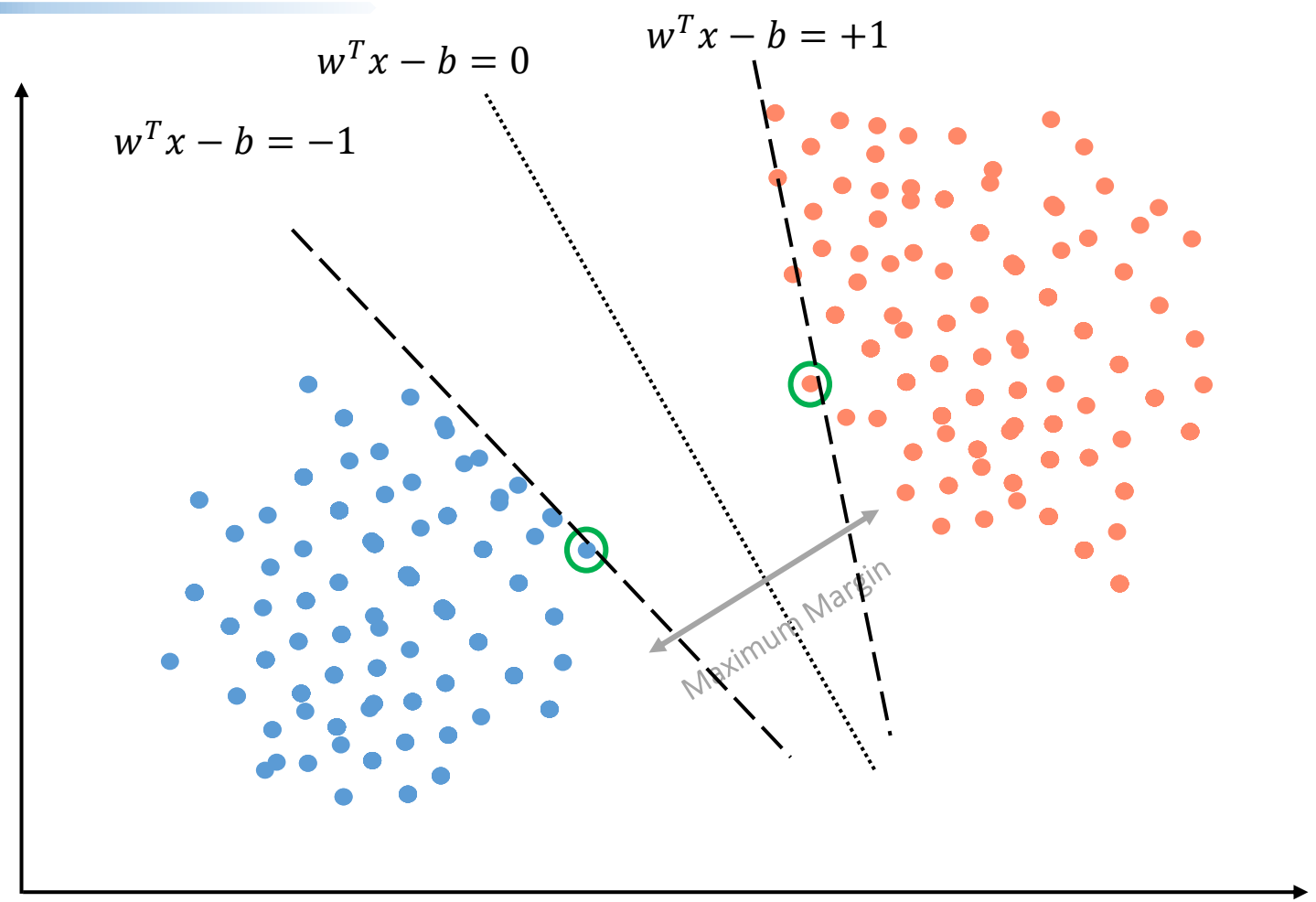
2. 선형 SVM

Linear Support Vector Machine

선형 SVM

선형 의사결정 경계를 학습하는 SVM

선형 SVM 기본원리



선형 SVM

선형 의사결정 경계를 학습하는 SVM

선형 SVM

- 초평면과 margin
  - ✓ 음의 초평면 : $w^T x - b = -1$
  - ✓ 양의 초평면 : $w^T x - b = +1$

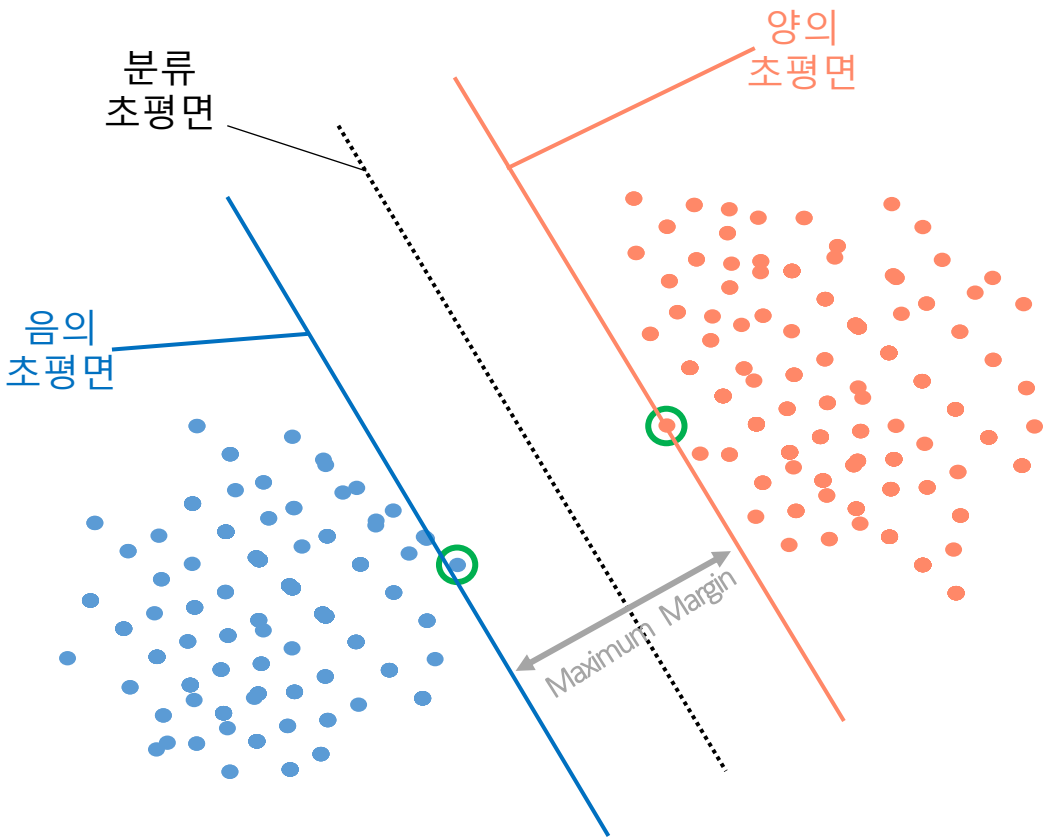
    $\Rightarrow$ 두 초평면 사이의 거리 : $2/\|w\|$
- 데이터의 분리
  - ✓ 음의 데이터점 x_i $(y_i = -1)$는 음의 초평면 아래쪽

    $$w^T x_i - b \leq -1$$
  - ✓ 양의 데이터점 x_i $(y_i = +1)$는 양의 초평면 위쪽

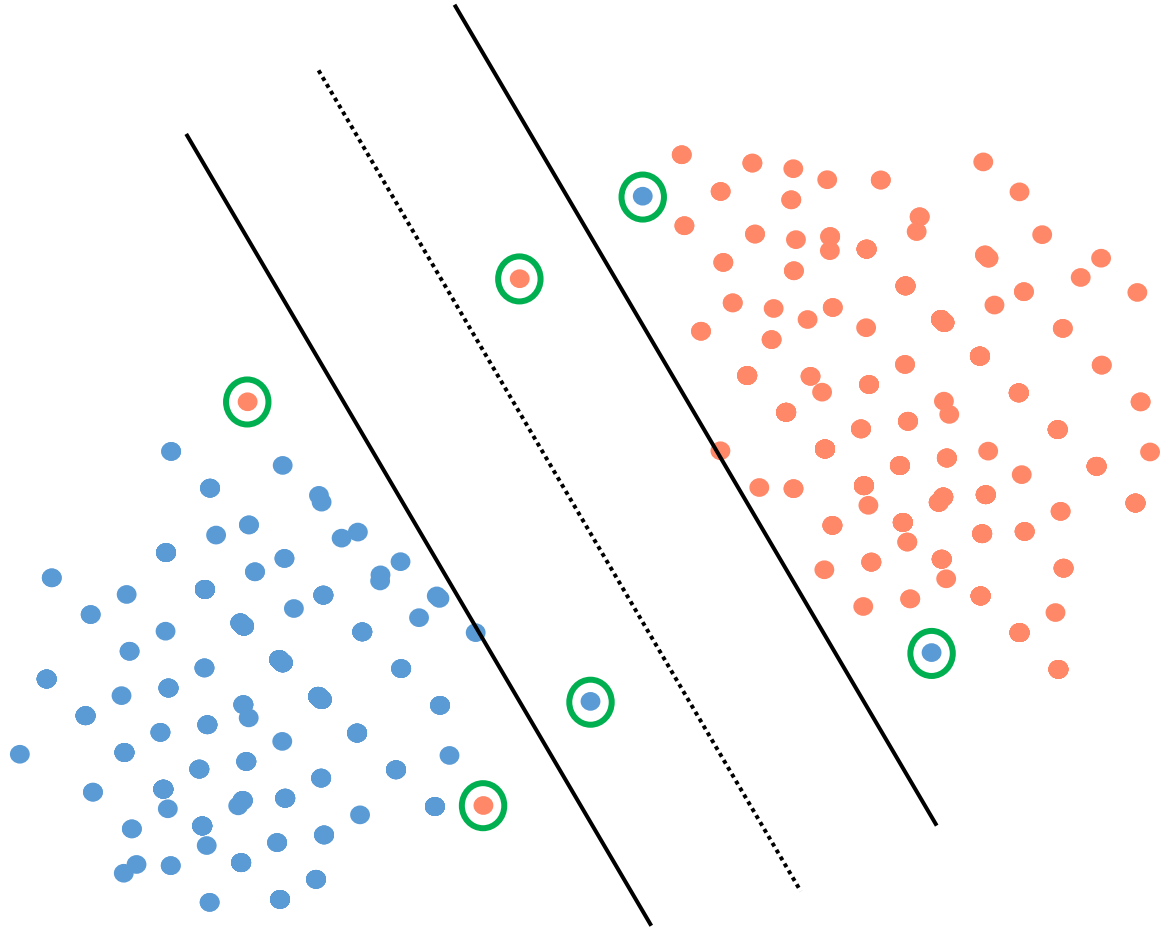
    $$w^T x_i - b \geq +1$$

    $$\Rightarrow y_i(w^T x_i - b) \geq 1$$
- 위 조건을 만족하면서 margin 거리 최대화 :

  $$\arg\min_{(w,b)} \|w\|$$



완벽한 분류가 어렵다면?



이런 데이터는 어떻게 학습할까?

선형 SVM

선형 의사결정 경계를 학습하는 SVM

Non Separable SVM

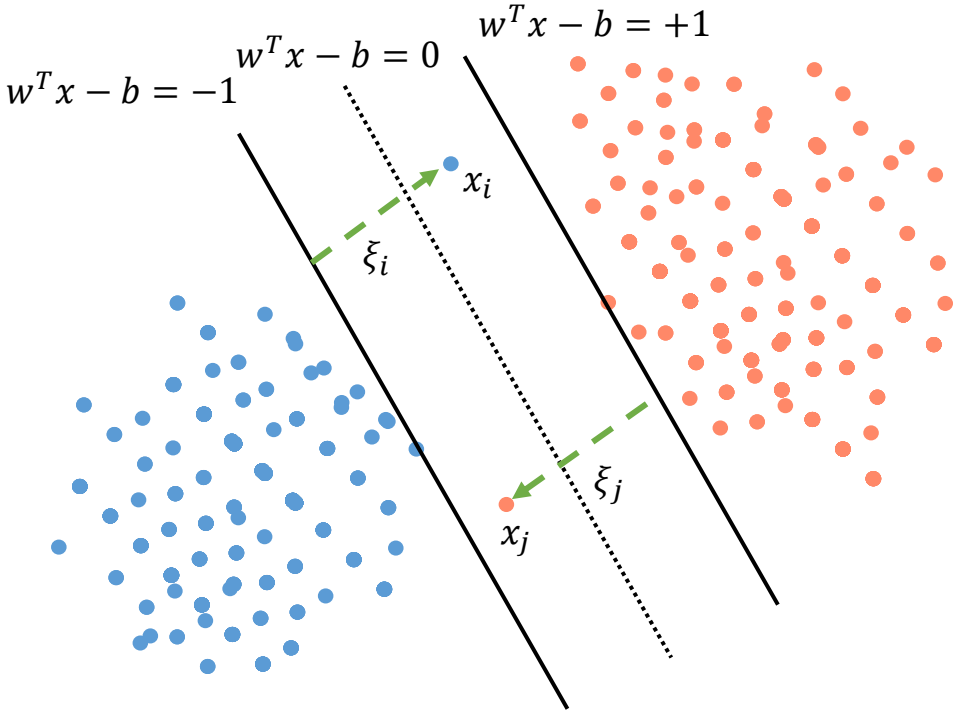
- 데이터가 선형 초평면에 의해 두 개의 범주로 분리되지 않는 경우, **기존 선형 SVM으로는 학습 불가**
- 오분류를 허용하는 것이 불가피하다면, 양/음의 초평면에서 **반대방향으로 침범한 정도(slack)에 비례해서 페널티를 주는 방법을** 이용
- 완전 분리만 고려하는 경우 Hard Margin
  불완전 분리를 허용하는 경우 Soft Margin
- 데이터 분리 조건의 완화:

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

- 페널티를 포함 :

$$\arg\min_{(w,b,\xi)} \left\{ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \right\}$$

C : 데이터를 잘못 분류하는 것을 어느정도 허용할지 결정하는 hyperparameter



선형 SVM

선형 의사결정 경계를 학습하는 SVM

선형 SVM의 적용 및 해석

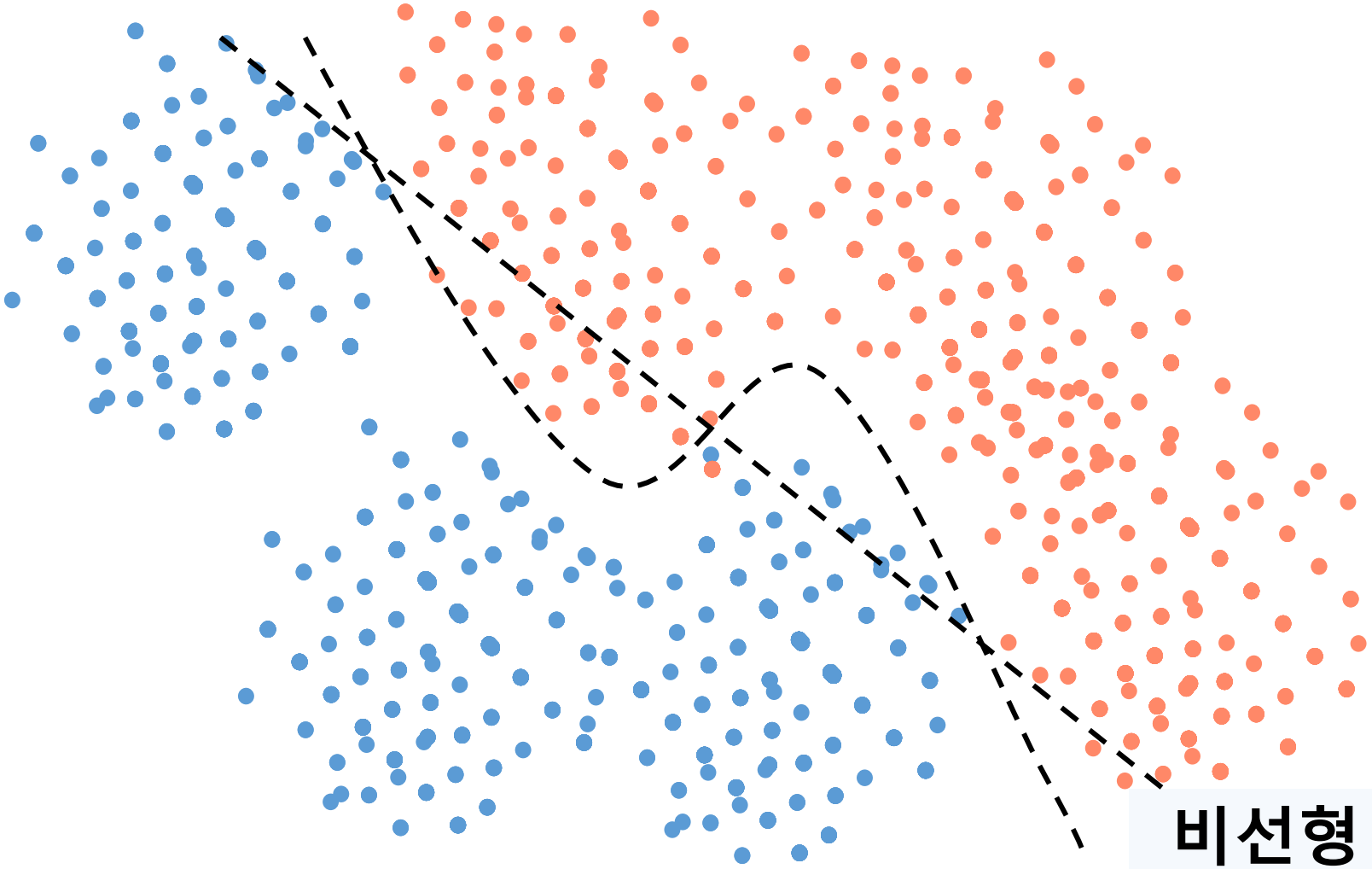
- soft margin 버전이 기본적으로 적용되며, 추천됨
- 튜닝 파라메터 : soft margin 가중치 C
- 학습 파라메터 : weight w, 상수항 b, 학습데이터 점 x_i 별 서포트 벡터 여부, slack ξ_i
  - ✓ ξ_i가 0이거나 양수이면서 margin 영역에 있는 데이터 점 x_i 는 서포트 벡터
  - ✓ ξ_i가 양수이면 는 영역 침범 벡터
- 의사결정 함수 : $D(x) = w^T x - b$ (새 데이터 x 는 $D(x)$ 의 부호로 그룹 판별)
  - ✓ $|w|$가 큰 변수 : 그룹 분류에 영향력이 큰 변수
  - ✓ ξ_i가 0 인 서포트 벡터 데이터 : 그룹의 경계로 학습된, 기준 역할을 하는 데이터 점
  - ✓ 영역 침범 데이터 : 분류가 어려운 데이터 점

3. 비선형 SVM

Non-Linear Support Vector Machine

분리하기 힘든경우?

오분류를 줄이는 방법?



비선형 함수로 분류

비선형 SVM

비선형 의사결정 경계를 학습하는 SVM

비선형 SVM 기본원리

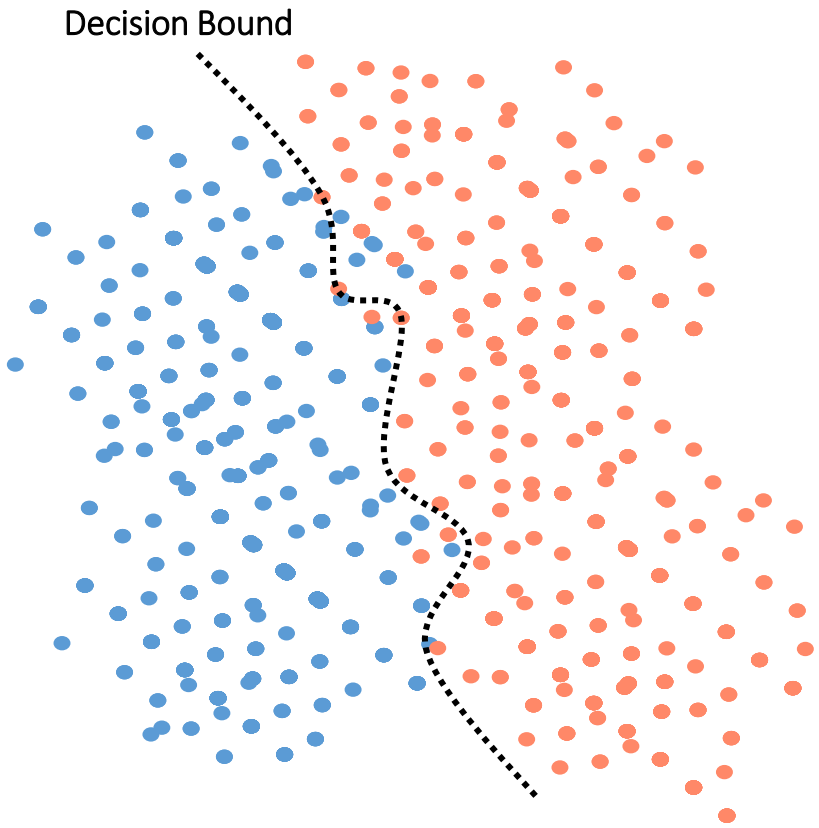
- 선형 SVM으로 데이터를 분류하기 어려운 경우 비선형 함수를 통한 분류를 고려

- 선형 SVM의 초평면 대신, 변수 변환 함수 $\Phi(\cdot)$를 설명변수에 먼저 적용해서 차원(관점)을 늘린 뒤 **일반화된 초평면**을 구성

$$D(x) = w^T \Phi(x) - b$$

- 늘어난 차원에서는 초평면이지만 **원본 차원에서는 곡면 형태인** 의사결정경계를 구성 가능

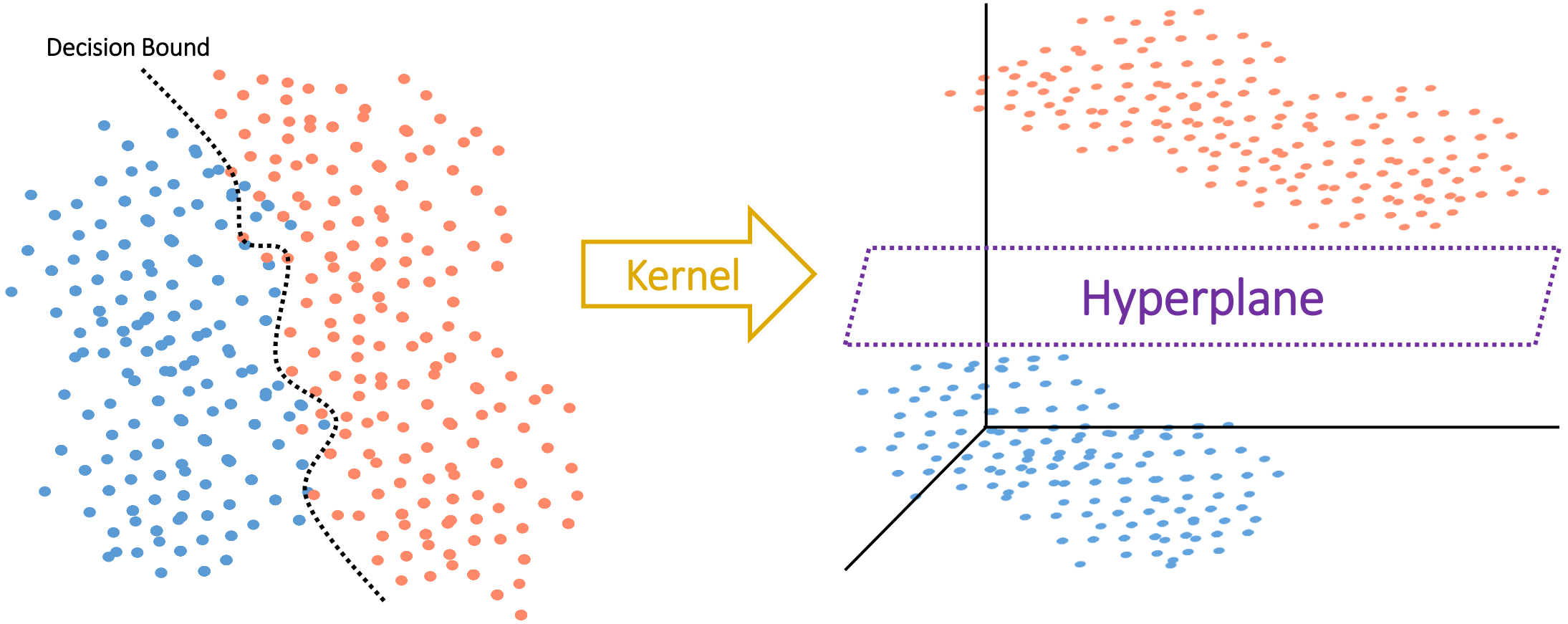
  ✓ 예시 : $\Phi(x_1, x_2) = (x_1, x_2, x_3 = x_1^2, x_4 = x_1 x_2)$
  $\mathrm{D}(x) = 2x_1 + 3x_4 - 1 = 2x_1 - 3x_1 x_2 - 1$



비선형 SVM

비선형 의사결정 경계를 학습하는 SVM

비선형 SVM 기본원리



비선형 SVM

비선형 의사결정 경계를 학습하는 SVM

Kernel SVM

- 변환 함수 Φ를 직접 계산하는 대신,
  $K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$로 연관되는
  커널함수 K를 이용하는 커널 트릭을 통해서도
  비선형 SVM을 구현 가능함

- 의사결정 함수 :

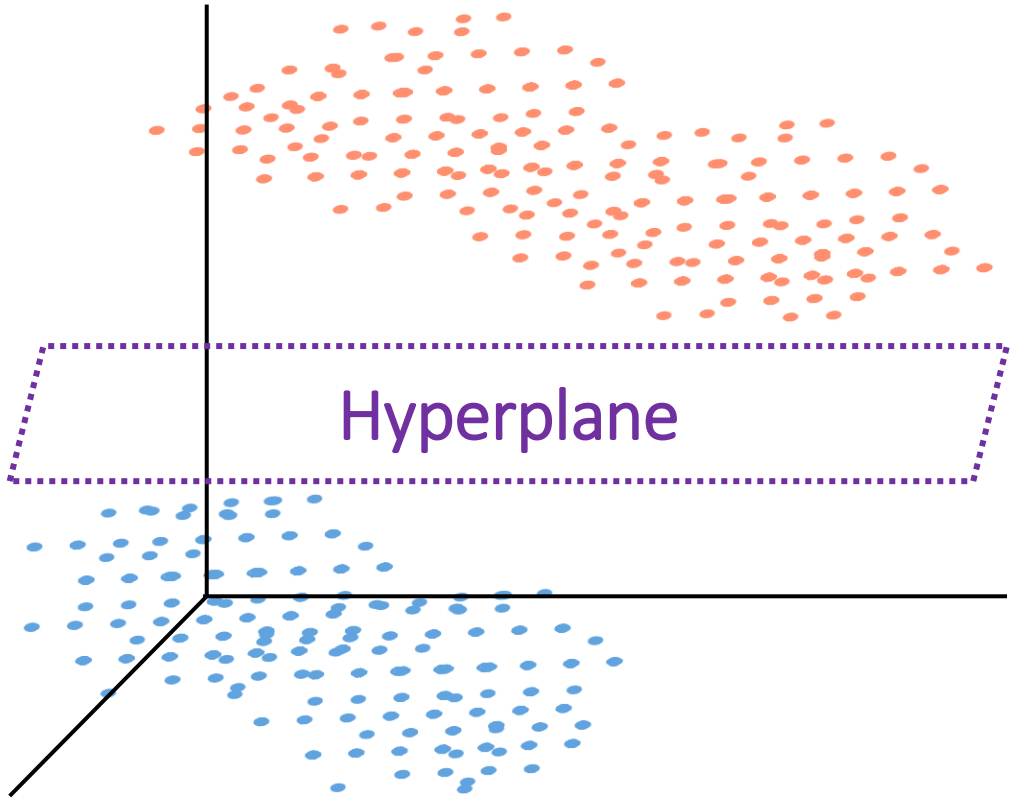
  $$D(x) = \sum_j \alpha_j K(x, x_j) - b = \left\{\sum_j \alpha_j \Phi(x_j)\right\}^T \Phi(x) - b$$

  ✓ 변수별 가중치 w 대신,
  학습데이터 점 x_j 별 가중치 α_j로 표현
  (서포트 벡터만 값을 가짐)

  ✓ $K(x_i, x_j)$는 x_i, x_j의 유사도 처럼 해석

- 커널 K의 선택에 따라 어떤 변환 Φ를 사용하는
  것과 같은 의미인지가 달라짐

  ✓ 예 : $K(x_i, x_j) = x_i^T x_j$ 이면 변환 안함/선형

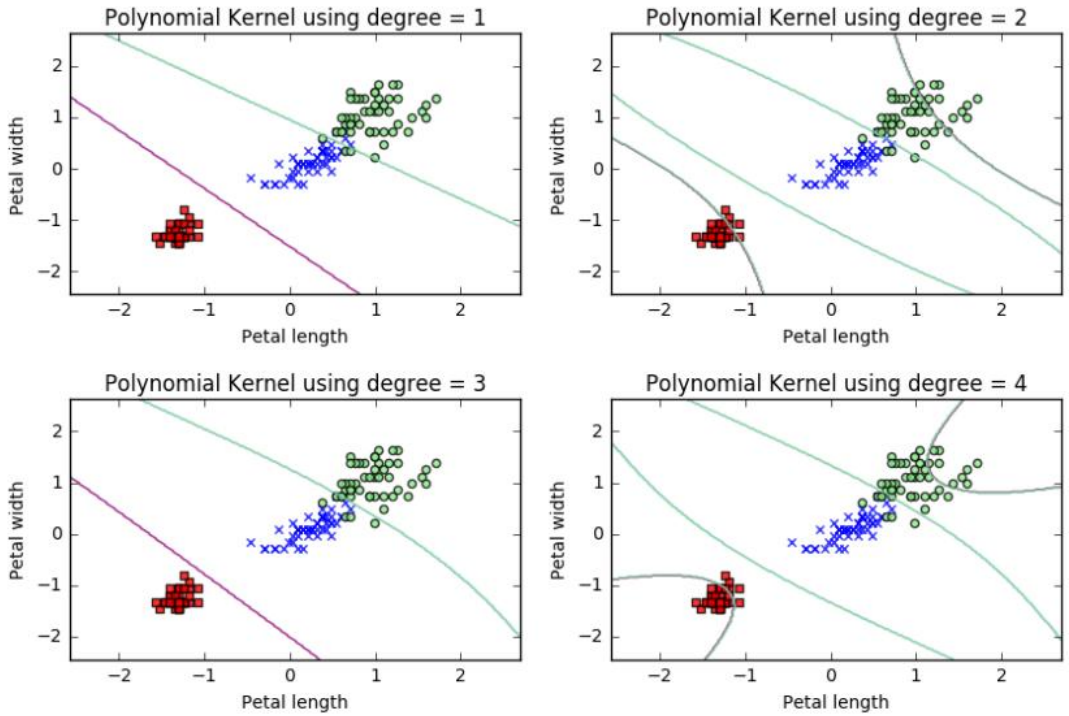


비선형 SVM

비선형 의사결정 경계를 학습하는 SVM

다항 커널

- $K(x_i, x_j) = (< x_i, x_j > + c)^d$
- 대응되는 변수변환 Φ : 모든 변수의 d 차까지의 다항식
- 차수 d 를 높일수록 일그러진 형태의 의사결정 경계를 학습

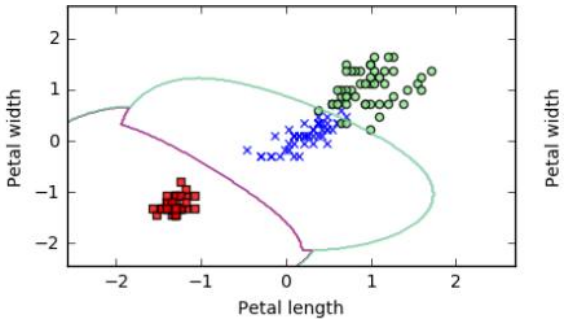


비선형 SVM

비선형 의사결정 경계를 학습하는 SVM

RBF(Radial Basis Function) 커널

- $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|)^2$
- 대응되는 변수변환 Φ : 모든 변수의 무한차원 다항식
- 특정 서포트 벡터들의 근방 단위로 의사결정 영역이 형성됨
  - ✓ 다항 커널보다도 유연



비선형 SVM

비선형 의사결정 경계를 학습하는 SVM

Kernel SVM의 적용 및 해석

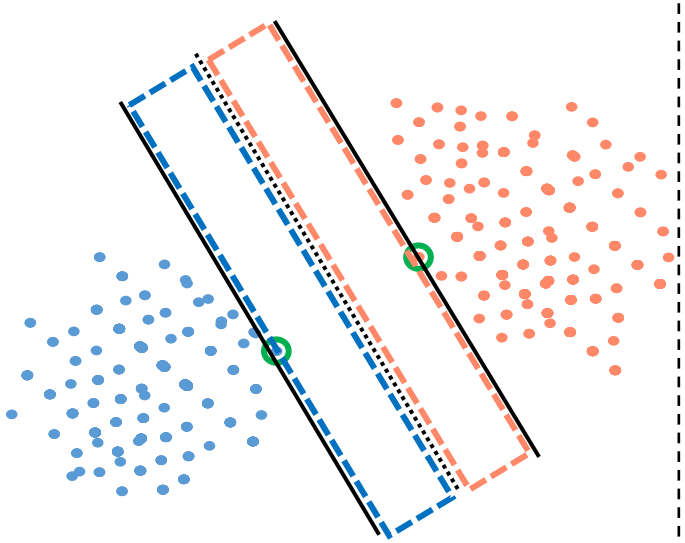
- 좋은 변수 변환/의사결정 경계를 모르면 선형 SVM 이후 kernel을 바꿔가면서 적용
- 튜닝 파라메터 : soft margin 가중치 C, kernel의 종류 K 및 그 파라메터 (예 : 다항 커널의 지수 d, 상수 c)
  - ✓ 원하는 의사결정 경계의 형태에 따라 kernel 및 그 파라메터 선정
- 학습 파라메터 : 데이터별 weight α_j, 상수항 b, slack ξ_i
  - ✓ kernel을 이용하면 변수별 weight는 얻을 수 없음
  - ✓ α_j가 양수이면 서포트 벡터
- 의사결정 함수 : $D(x) = \sum_j \alpha_j K(x, x_j) - b$ (새 데이터 x 는 $D(x)$ 의 부호로 그룹 판별)
  - ✓ $|\alpha_j|$가 큰 학습 데이터 x_j : 그룹 분류시 유사도를 확인해야 하는 랜드마크 역할을 하는 점

SVM 요약

선형 SVM

- 데이터 :
$$y_i(w^T x_i - b) \geq 1$$

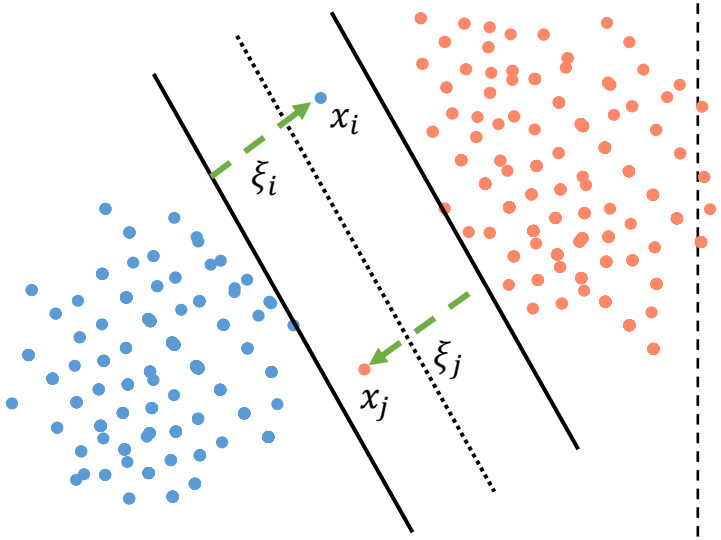
- 최대 Margin :
$$\arg\min_{(w,b)} \|w\|$$



Non Separable SVM

- 데이터 :
$$y_i(w^T x_i - b) \geq 1 - \xi_i$$

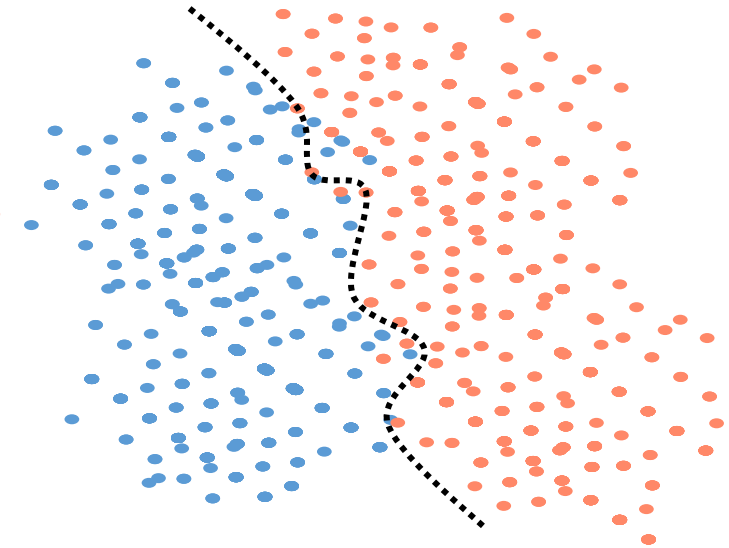
- 최대 Margin :
$$\arg\min_{(w,b,\xi)} \left\{ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \right\}$$



Kernel SVM

- 데이터 :
$$y_i(\{\sum_j \alpha_j \Phi(x_j)\}^T \Phi(x_i) - b) =$$
$$y_i(\sum_j \alpha_j K(x_j, x_i) - b) \geq 1 - \xi_i$$

- 최대 Margin :
w대신 α_j에 대해 최소화



SVM 요약

SVM 적용시 고려사항

- 명목형 변수를 입력에 사용할 경우 해석이 잘 안됨 (선형, Kernel SVM 둘 다)
- Kernel SVM의 경우 데이터의 건수가 많으면(대략 1만 단위 이상) 유사도 계산 단계의 메모리 소비량이 많음
  - ✓ 필요에 따라 부분 표본 추출, 차원 축소 기법 적용후 학습
- 튜닝 파라미터가 있으므로 비교하여 결정 (EDA로 미리 좋은 설정 파악하기 어려움)
- 선형 SVM에 한해 의사결정 경계면 및 변수별 가중치를 설명하기 용이
- 결과가 그룹 경계 근처에 있거나 분류가 어려운 위치에 있는 데이터 점(Support vector)에만 크게 의존
  - ✓ 이 영역 밖의 이상점이나 데이터의 그룹 비율에 둔감(강건)
  - ✓ 이 영역의 데이터 수가 적으면 결과가 불안정함
- 의사결정 함수 $D(x)$를 확률로 볼 수 없음 (비교 대상 : 일반화 선형 모형의 로지스틱 회귀 모형)
  - ✓ $|D(x)|$가 클수록 그 그룹에 속할 가능성이 높다 정도의 해석은 가능

4. SVR

SVM for Regression

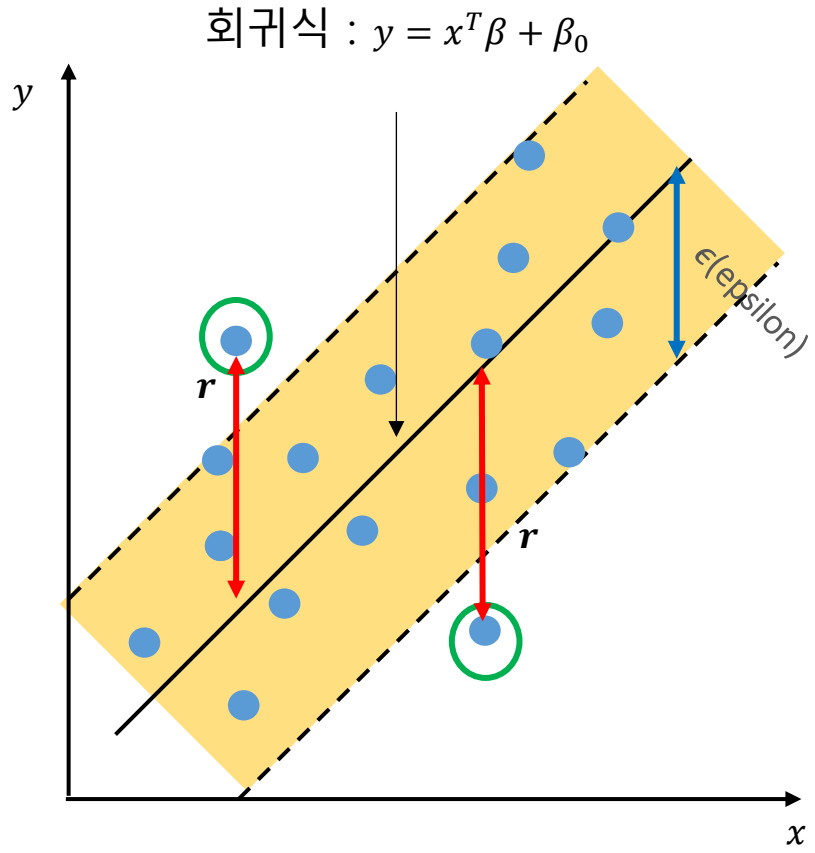
SVR(SVM for Regression)

SVM의 아이디어를 예측에 적용한 모형

Linear SVR(SVM for Regression)

- 두 그룹으로 구분하는 분류(Classification)에서는 두 점선간 거리인 마진(Margin)을 최대로 갖는 값을 찾았다면 (각 그룹이 최대로 떨어질 수 있는 최대 거리)
- 회귀(Regression)에서는 정해진 마진(회귀식 인접 공간) 내부에 데이터가 모두 들어가도록 학습하는 것이 목표
- Soft 마진 : 마진(margin) 밖으로 넘어가는 에러(Error)가 최소가 되도록 학습
- 마진의 폭은 ϵ(epsilon)을 직접 설정해서 조절 (hyperparameter)
  - ✓ 데이터 점별 조건 : $|y_i - x_i{}^T\beta - \beta_0| \le \epsilon + \xi_i$

$$\min\{\frac{1}{2}\|\beta\|^2 + \lambda(\sum_{i=1}^{n} \max(0, |y_i - x^T\beta - \beta_0| - \epsilon))\}$$



SVR 적용시 고려 사항

SVM의 아이디어를 예측에 적용한 모형

SVR의 특징

- 해석상의 서포트 벡터/계산과 관련된 대부분의 특징을 SVM과 공유함
- 결과(기울기 및 예측치)가 회귀선에서 ϵ보다 먼 위치에 있는 데이터 점(Support vector)에만 크게 의존
  - ✓ 이상치를 무시하고 회귀선에 가까운 데이터에만 크게 의존하는 강건회귀모형(robust regression)과는 다른 특성
- 선형 SVR의 경우 회귀식을 설명하기 용이함