

## 12강. Support Vector Machine

◆ 담당교수 : 김 동 하

### ■ 학습개요

분류, 회귀, 더 나아가 이상치 탐지 등에서 매우 뛰어난 성능을 가지고 있는 기계학습 방법론인 SVM에 대해서 학습한다. 이진 분류 문제에 대해 집중적으로 다룰 것이며, 선형 SVM, 비선형 SVM에 대해서 배운다. 특히, 비선형 SVM에서 사용되는 kernel trick이 무엇인지에 대해 공부한다.

### ■ 학습목표

1	Soft margin과 hard margin의 개념에 대해 학습한다.
2	선형 SVM의 개념과 최적화 문제에 대해 학습한다.
3	비선형 SVM에서 feature map과 kernel trick에 대해 학습한다.

### ■ 주요용어

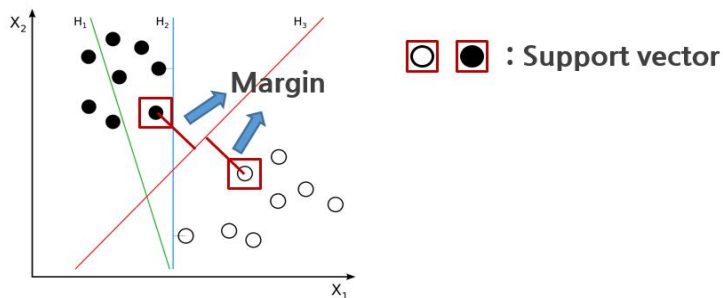
용어	해설
Margin	결정 경계(Decision boundary)와 가장 가까운 관측치 사이의 거리
Support vector	최종 학습된 결정 경계와 가장 가까운 관측값
Soft (Hard) margin SVM	자료가 하나의 초평면으로 완벽하게 구분이 될 경우 soft margin SVM을 사용하고, 그렇지 못할 경우 hard margin SVM을 사용한다.
Kernel trick	SVM 최적화 문제를 dual problem의 관점으로 해석할 때, feature map 자체보다는 feature map으로 만들어진 kernel 함수만 사용된다는 사실을 이용하여 무한차원의 feature map을 사용할 때도 손쉽게 SVM을 학습할 수 있는 특징을 의미한다.

### ■ 학습하기

## 01. SVM 개요

### Support Vector Machine

- Cortes and Vapnik (1995)
- 선형이나 비선형 분류를 할 수 있는 기계학습 방법론
  - > 특히 복잡한 분류 문제를 잘 해결.
- 회귀 문제, 이상치 탐색 문제로도 확장 가능.
  - > 본 강의에서는 이진 분류 문제만을 고려. 즉,  $y \in \{-1, 1\}$ 을 가정.
- 확률적 모형을 가정하지 않음.
  - > 확률 추정 없이 직접 분류 결과에 대해 예측.
- 두 클래스 사이에 가장 너비가 큰 분류 경계선을 찾기 때문에 Large margin classification이라고도 함.
- Margin을 최대로 하는 경계선은
- Support vector
  - > 분류 경계선과 가장 가까운 관측치

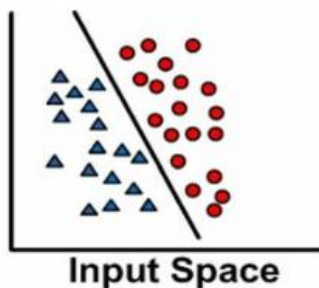


- SVM의 분류 경계선은 margin 계산에 민감하게 반응하므로 변수들 사이의 스케일을 맞춰 주는 작업이 필요.

### Linear SVM vs. Nonlinear SVM

- 선형 SVM
  - > 선형 분류 경계선으로 클래스를 분류.

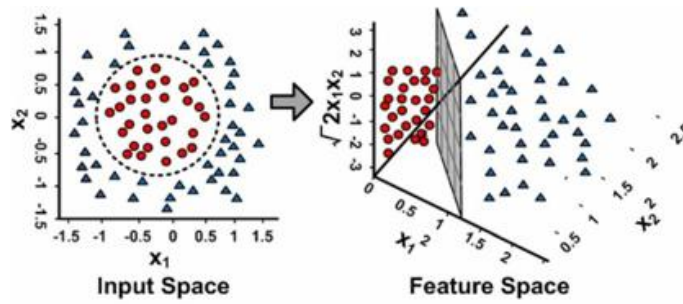
$$\text{sign}(w^T x + b)$$



- 비선형 SVM
  - > 비선형 분류 경계선으로 클래스를 분류.
  - > 비선형 함수를 통해 데이터를 변형하고, 변형된 공간에서 선형 경계선을 활용.

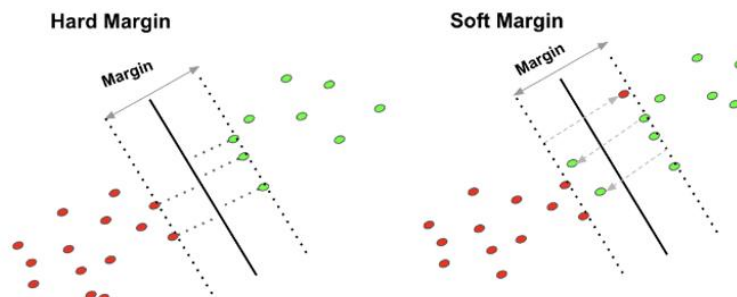
$$\text{sign}(w^T \phi(x) + b)$$

- >  $\phi(\cdot) : R^p \rightarrow R^m$ : feature function
- > 입력 변수를 선형 분리가 쉽도록 변형해주는 함수
- > 무한차원으로도 확장할 수 있음 (Kernel trick).



### Hard Margin vs. Soft Margin

- Hard Margin 방법
  - > 두 클래스가 하나의 선(혹은 평면)으로 완벽하게 나뉘지는 경우에 적용 가능.
- Soft Margin 방법
  - > 일부 샘플들이 분류 경계선의 분류 결과에 반하는 경우를 일정 수준 이하로 허용하는 방법.



## 02. Linear SVM

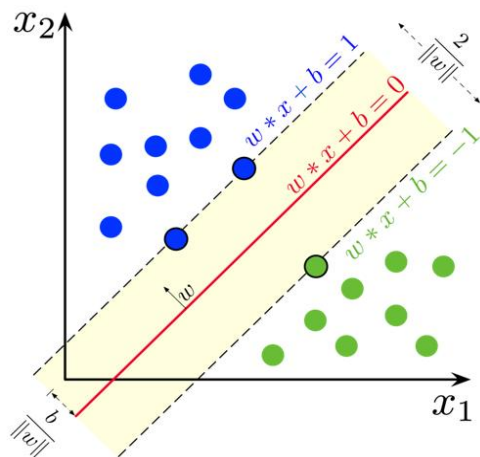
### Hard margin

- 학습 데이터:  $(x_1, y_1), \dots, (x_n, y_n)$
- 다음의 문제를 최소화하는 기울기  $w$ 와 절편항  $b$ 를 찾는다:

$$\min_{w,b} \|w\|_2^2$$

s.t.

$$y_i(w^T x_i + b) \geq 1 \text{ for } i = 1, \dots, n$$



### Soft margin

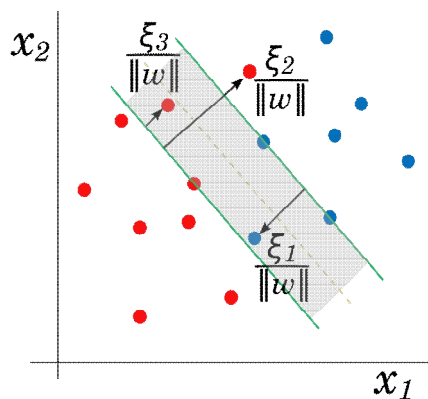
- $C$ : 분류 결과에 반하는 경우를 일정 수준으로 조정하는 조율 모수
- 다음의 문제를 최소화하는 기울기  $w$ 와 절편항  $b$ 를 찾는다:

$$\min_{w,b} \|w\|_2^2 + C \cdot \sum_{i=1}^n \xi_i$$

s.t.

$$y_i(w^T x_i + b) \geq 1 - \xi_i \text{ for } i = 1, \dots, n,$$

$$\xi_i \geq 0 \text{ for } i = 1, \dots, n$$



## 03. Nonlinear SVM

### Feature map

- 비선형 SVM은 입력값을 feature map  $\phi(\cdot): R^p \rightarrow R^m$ 을 이용해 변환한 후 선형 SVM을 사용하는 방법.
- Feature map
  - > 복잡한 함수일수록 변환된 feature space에서  $y$ 를 잘 나누는 선형 분류 경계선이 존재할 가능성이 높음.
- 무한 차원으로 변환하면 어떨까? (즉,  $m = \infty$ )

- 무한 차원의 feature space에서 선형 분류 경계선을 찾으려면 무한개의 모수가 필요함.
- 하지만, kernel trick 덕분에  $n$ 개의 모수만 필요함  
-> Kernel trick이란?

### Dual problem


- 선형 SVM 문제를 다음과 같은 문제로 변환할 수 있음. (Dual Problem)

$$\begin{aligned} \min_{c_1, \dots, c_n} & -\sum_{i=1}^n c_i + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j c_i c_j x_i^T x_j \\ \text{s.t.} & \\ & \sum_{i=1}^n c_i y_i = 0, \\ & 0 \leq c_i \leq \frac{1}{2nC} \text{ for } i = 1, \dots, n. \end{aligned}$$

### Kernel trick

- 마찬가지로 비선형 SVM을 dual problem으로 변환하면 다음과 같음.

$$\begin{aligned} \min_{c_1, \dots, c_n} & -\sum_{i=1}^n c_i + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n y_i y_j c_i c_j \boxed{K(x_i, x_j)} \\ \text{s.t.} & \\ & \sum_{i=1}^n c_i y_i = 0, \\ & 0 \leq c_i \leq \frac{1}{2nC} \text{ for } i = 1, \dots, n \end{aligned}$$



$\phi(x_i)^T \phi(x_j)$   
**Kernel!**

- 따라서, 최적의 분류 경계선을 찾기 위해 알아야 하는 것은 feature map 그 자체가 아닌 kernel 값들임.
- 즉, 무한차원의 feature map이더라도 kernel만 정의되면 문제를 쉽게 풀 수가 있음  
-> Kernel trick
- 위의 문제의 해를  $\hat{c}_1, \dots, \hat{c}_n$ 이라 할 때, 분류 예측값은 다음과 같이 계산할 수 있다:

$$\begin{aligned} \hat{f}(x) &= \text{sign}(w^T \phi(x) + b) \\ &= \text{sign}\left(\left[\sum_{i=1}^n \hat{c}_i y_i K(x_i, x)\right] + b\right) \end{aligned}$$

$$\rightarrow w = \sum_{i=1}^n \hat{c}_i y_i x_i$$

$$\rightarrow b = -w^T \phi(x_k) + y_k = -\left[\sum_{j=1}^n \hat{c}_j y_j K(x_j, x_k)\right] + y_k$$

->  $(x_k, y_k)$ : support vector

### 다양한 kernel 함수들

- 앞서 언급했듯이, 비선형 SVM은 feature map을 이용한 kernel이 정의되면 최적의 분류 경계선을 구할 수 있음.
- 널리 사용되는 kernel 함수는 다음과 같다:
  - > Polynomial kernel :  $K(a,b) = (\gamma \cdot a^T b + r)^d$
  - > Radial Basis kernel :  $K(a,b) = \exp(-\gamma \cdot \|a - b\|_2^2)$
  - > Sigmoid kernel :  $K(a,b) = \tanh(\gamma \cdot a^T b + r)$

#### ■ 연습문제

(객관식)1. 다음 SVM에 대해 올바른 설명을 한 것을 고르시오.

- ① 분류 문제만을 해결할 수 있고, 회귀 문제에는 적용이 불가능하다.
- ② 확률적 모형을 가정한다.
- ③ Kernel trick을 통해 비선형 SVM도 쉽게 적합할 수 있다.
- ④ Margin을 최소화하는 방향으로 모형을 학습한다.

정답 : ③

해설 : SVM은 회귀 문제에도 확장이 가능하고, 확률적 모형을 가정하지 않는다. 또한, margin을 최대화하는 방향으로 학습한다.

(주관식)2. 최종 학습된 결정 경계선과 가장 가까운 관측치를 무엇이라 하는가?

정답) Support vector

해설) 최종 학습된 결정 경계선과 가장 가까운 관측치를 support vector라 한다.

(O/X)3. Kernel만 잘 정의가 된다면 무한차원의 feature map을 이용한 비선형 SVM도 구축할 수 있다.

정답 : O

해설 : Kernel trick에 의해서 kernel만 정의가 된다면 입력 자료를 무한차원으로 변형하는 feature map을 이용한 SVM도 손쉽게 적합할 수 있다.

#### ■ 정리하기

1. SVM은 선형 혹은 비선형 함수를 이용하여 분류를 할 수 있는 기계학습 방법론 중 하나이며, 일반적으로 매우 뛰어난 성능을 가지고 있다.
2. SVM은 margin을 최대화하는 방향으로 결정 경계가 학습되며, 학습 데이터를 완벽히 나눌 수 있는지 여부에 따라 각각 soft margin SVM, hard margin SVM을 사용할 수 있다.
3. 무한차원의 feature map을 사용하여 비선형 SVM을 적합할 때에도 kernel trick을 이용해 손쉽게 모델을 학습할 수 있다.

■ 참고자료 (참고도서, 참고논문, 참고사이트 등)

박창이, 김용대, 김진석, 송종우, 최호식. 『R을 이용한 데이터마이닝』. 서울:교우사, 2018.