머신러닝
(MACHINE LEARNING)

LECTURE XI: 지도 학습 5
(Supervised Learning)

Dai-Gyoung Kim

Department of Applied Mathematics

Hanyang University ERICA

 ⓒ dgkim@hanyang.ac.kr

지도 학습 (Supervised Learning)

Contents

- 분류와 회귀
- 일반화, 과대적합, 과소적합
- 지도 학습 알고리즘
  - k-최근접 이웃
  - 선형모델
  - 나이브베이즈 모델
  - 결정트리
  - 결정트리의 앙상블
  - 커널 서포트 벡터 머신
  - 신경망, 딥러닝
- 분류 예측의 불확실성 추정

ⓒ dgkim@hanyang.ac.kr

❖ 분류용 선형 모델

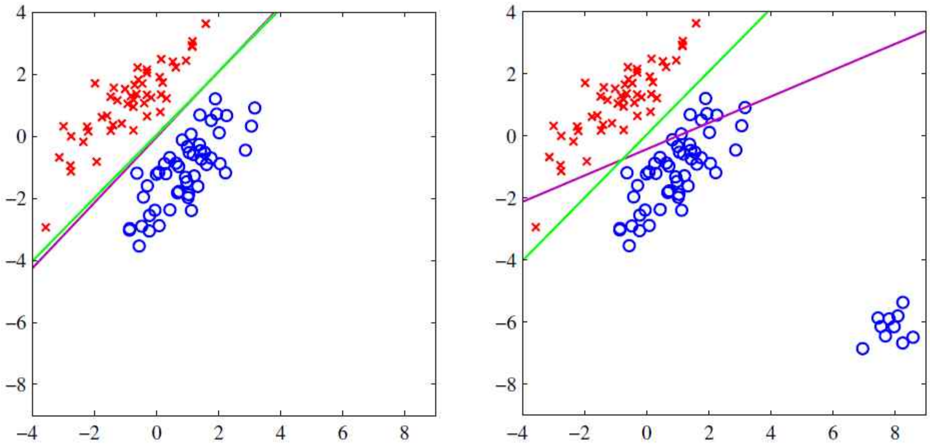
- 선형 모델은 분류에도 널리 사용됨.
- 선형 회귀 모델은 예측 값을 결정하는 직선, 평면, 초평면 등을 생성함 .
- 선형 분류 모델은 예측 값을 분류하는 **결정 경계**(직선, 평면, 초평면 등)를 생성함.

선형 이진 분류기

- 선형 이진 분류기의 예측 방정식은 다음과 같음.

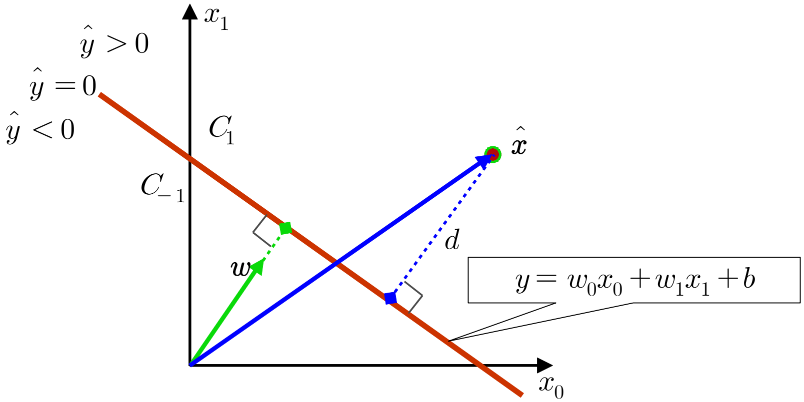
$$\hat{y} = w_0 x_0 + w_1 x_1 + \cdots + w_p x_p + b > 0$$

- 위 식에서 예측한 값 $\hat{y}$을 임계치 0과 비교함.
  - ✓ $\hat{y} < 0$이면 음성 클래스 -1로 분류.
  - ✓ $\hat{y} > 0$이면 양성 클래스 $+1$로 분류.
  - ✓ $\hat{y} = 0$이면 결정경계에 위치함.



- 한 데이터 포인트 $\hat{x}$ 에서 결정 경계 $y = w^T x + b$ 사이의 (부호)거리

$$d = \frac{w^T \hat{x} + b}{\| w \|} = \frac{\hat{y}}{\| w \|}$$

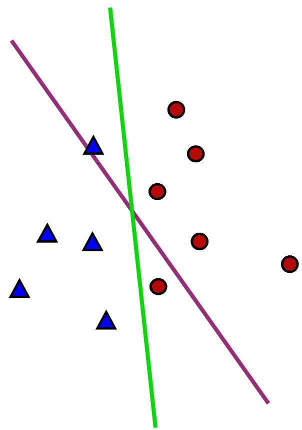


- 주어진 훈련 데이터셋 X와 타깃셋 $\{t_1, \ldots, t_m\}$에 대하여

$$X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}, \; t_i \in \{-1, 1\}$$

- 선형 분리의 필요충분조건

$$(\boldsymbol{w}^T \boldsymbol{x}_i + b) t_i > 0, \; i = 1, \ldots, m$$



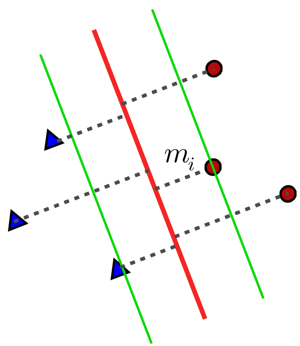
✓ w와 b의 스케일을 조정하여 얻은 조건

$$\left(\boldsymbol{w}^T \boldsymbol{x}_i + b\right) t_i > 1, \quad i = 1, \ldots, m$$

- 정규화된 마진(margin):

$$m_i = \frac{\left(\boldsymbol{w}^T \boldsymbol{x}_i + b\right) t_i}{\| \boldsymbol{w} \|}$$

- 효과적인 결정 경계를 구하기 위해 <u>최소 마진을 최대화</u>해야함.



✓ 최소 마진:

$$\min_{1 \le i \le m} \left\{ \frac{\left(\boldsymbol{w}^T \boldsymbol{x}_i + b\right) t_i}{\| \boldsymbol{w} \|} \right\} = \frac{1}{\| \boldsymbol{w} \|}$$

✓ 최소 마진의 최대화 문제:

$$\max_{w} \frac{1}{\| \boldsymbol{w} \|} \quad \text{subject to} \quad \left(\boldsymbol{w}^T \boldsymbol{x}_i + b\right) t_i > 1, \;\; i = 1, \dots, m$$

- 선형 SVM (하드 마진)분류기의 최적화 문제

$$\underset{w,\, b}{\text{Minimize}} \quad h(\boldsymbol{w}) = \frac{1}{2} \boldsymbol{w}^T \boldsymbol{w}$$

$$\text{subject to} \quad \left(\boldsymbol{w}^T \boldsymbol{x}_i + b\right) t_i > 1, \;\; i = 1, \dots, m$$

- 위 최적화 문제를 풀어 w와 b를 결정함.
- 이러한 해법을 **서포트 벡터 머신**(SVM, Support Vector Machine)이라고 함.

선형 분류 알고리즘

- 선형 분류 알고리즘의 대표적인 두 개는 다음과 같음.

1) 서포트 벡터 머신(SVM, support vector machine) 분류
   ✓ 사이킷런에서 "svm.LinearSVC"로 구현함.

2) 로지스틱 회귀(logistic regression)
   ✓ 사이킷런에서 "linear_model.LogisticRegression"으로 구현함.

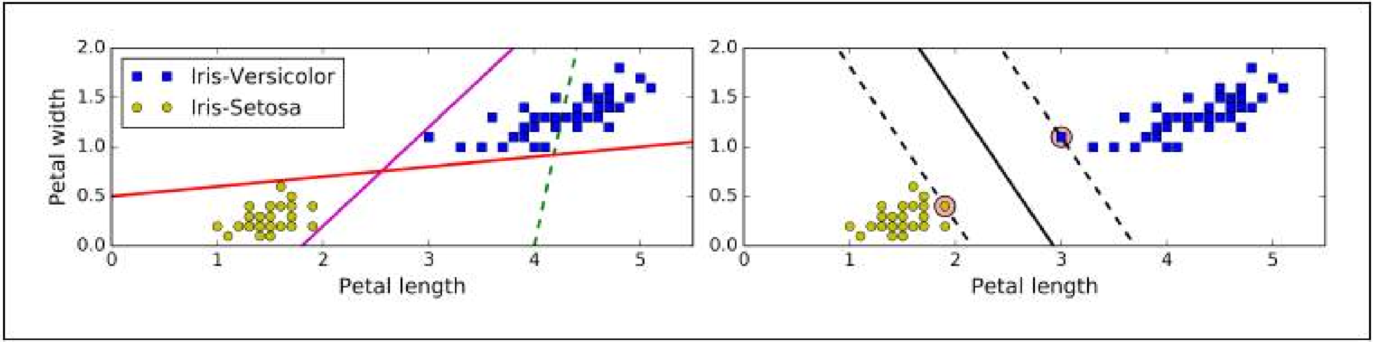
1) 서포트 벡터 머신 분류

✓ 하드 마진(hard margin) 분류

✓ 소프트 마진(soft margin) 분류

[하드 마진 분류]

- 아래 그림은 붓꽃 데이터셋의 일부를 나타낸 것이며, 두 클래스가 직선으로 잘 분류됨.

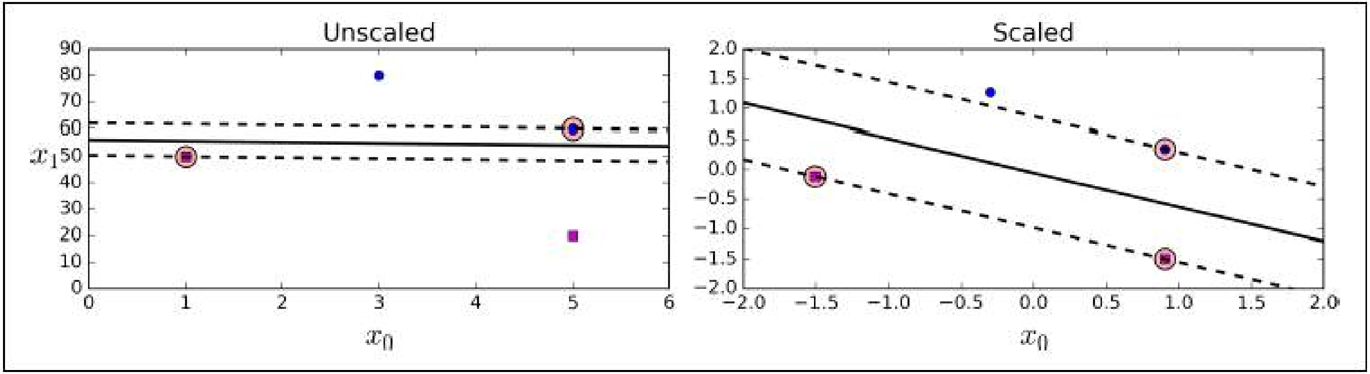


✓ 왼쪽 그래프의 세 개의 결정 경계 중 점선으로 나타난 결정 경계를 만든 모델은 클래스를 적절하게 분류하지 못하고 있음.

✓ 오른쪽 그래프의 실선은 SVM 분류기의 결정 경계이며 이 직선은 두 개의 클래스를 나누고 있을 뿐만 아니라 제일 가까운 훈련 샘플로부터 가능한 한 멀리 떨어져 있음. 이 경우 분류를 **라지 마진 분류**(large margin classification)이라고 함.

✓ 오른쪽 그래프의 두 점선 바깥쪽에 훈련 샘플을 더 추가해도 결정 경계에는 전혀 영향을 미치지 않음. 점선 경계에 위치한 샘플은 결정 경계에 영향을 미치므로 이 샘플을 서포트 벡터(support vector)라고 함.

- SVM은 특성의 스케일에 민감함.



✓ 위 그림의 왼쪽 그래프에서는 수직축의 스케일이 수평축의 스케일보다 훨씬 커서 결정 경계가 수평에 가까움. 특성의 스케일을 조정하면 결정 경계가 훨씬 좋아짐.

✓ 모든 샘플이 점선 경계 바깥쪽에 올바르게 분류되어 있을 때 <u>하드 마진 분류</u>(hard margin classification)이라고 함.

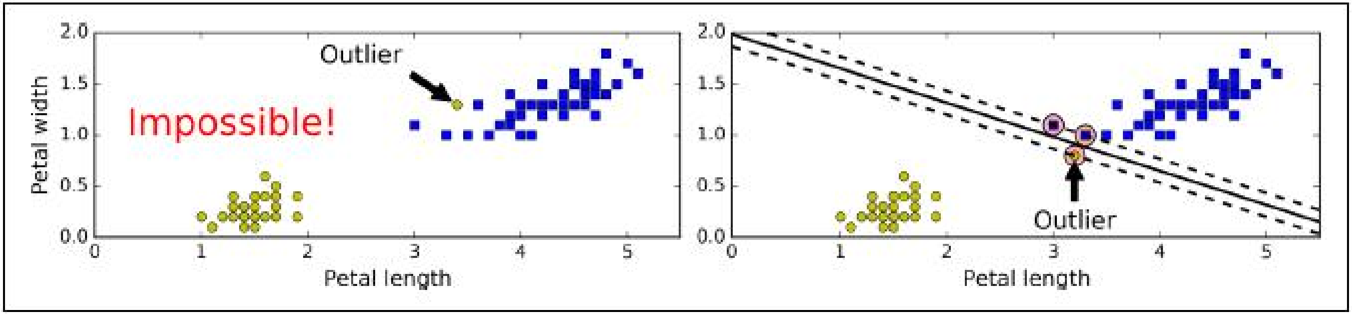
- <u>하드 마진</u> 선형 SVM 분류기의 최적화 문제

$$\underset{w,\,b}{\text{Minimize}} \quad h(w) = \frac{1}{2} w^T w$$

$$\text{subject to} \quad \left(w^T x_i + b\right) t_i > 1, \ \ i = 1, \ldots, m$$

[소프트 마진 분류]

- 하드 마진 분류는 두 가지 문제점이 있음.
  - ✓ 데이터가 선형적으로 구분될 수 있어야 잘 작동함.
  - ✓ 이상치에 민감함.



- 위와 같은 문제를 다루려면 좀 더 유연한 모델이 필요함.
  - ✓ 서포트 벡터들을 지나는 경계선들이 가능한 넓게 유지하는 것과 **마진 오류**(margin violation, 샘플이 결정 경계 또는 반대쪽에 있는 경우) 사이에 적절한 균형을 이루어야 함. 이를 고려한 모델을 **소프트 마진 분류**(soft margin classification)라고 함.

- 소프트 마진 분류기의 최적화 문제를 구성하기 위해서는 각 샘플에 대해 **마진 슬랙 변수**(margin slack variable) $\xi_i \geq 0$을 도입하여 i번째 샘플이 얼마나 마진을 위반할지를 정함.
  - ✓ 이 문제는 두 개의 상충된 목표를 가지고 있음.
    - ① 마진 오류를 최소화하기 위해 가능한 한 슬랙 변수의 값을 작게 만들어야 함.
    - ② 마진을 크게 하기 위해 $\| w \|$를 가능한 한 작게 만들어야 함.
  - ✓ 위의 두 목표 사이의 트레이드오프(tradeoff)가 있으며, 이를 정의하는 것이 다음과 같은 하이퍼파라미터 C임.

- **소프트 마진** 선형 SVM 분류기의 최적화 문제

$$\underset{w,\,b,\,\zeta}{\text{Minimize}} \qquad h(w,\zeta) = \frac{1}{2}w^T w + C\sum_{i=1}^{m}\xi_i$$

$$\text{subject to} \qquad \left(w^T \boldsymbol{x}_i + b\right)t_i > 1 - \xi_i,$$

$$\xi_i \geq 0, \qquad\qquad i = 1, \ldots, m$$

▷ $C > 0$

▷ $\xi = (\xi_1, \ldots, \xi_m)$

✓ 하이퍼파라미터 C의 역할은 허용오차 정도를 결정함. 즉, 마진을 위반하는 정도를 결정함.

✓ $C \gg 1$ ⇨ $\sum \xi_i \approx 0$: 하드 마진 분류와 같아짐.

✓ $C \approx 0$ ⇨ ξ_i의 범위가 넓어짐: 마진 위반을 많이 허용함

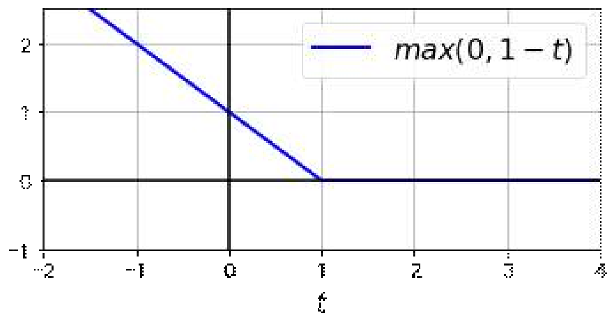
- SVM 분류기의 최적화 문제를 풀기 위해서는 다음의 선형 SVM 비용 함수를 최소화하는 경사 하강법을 적용함.

$$J(w,b) = \frac{1}{2} w^T w + C \sum_{i=1}^{m} \max\left(0, 1 - \left(w^T x_i + b\right) t_i\right)$$

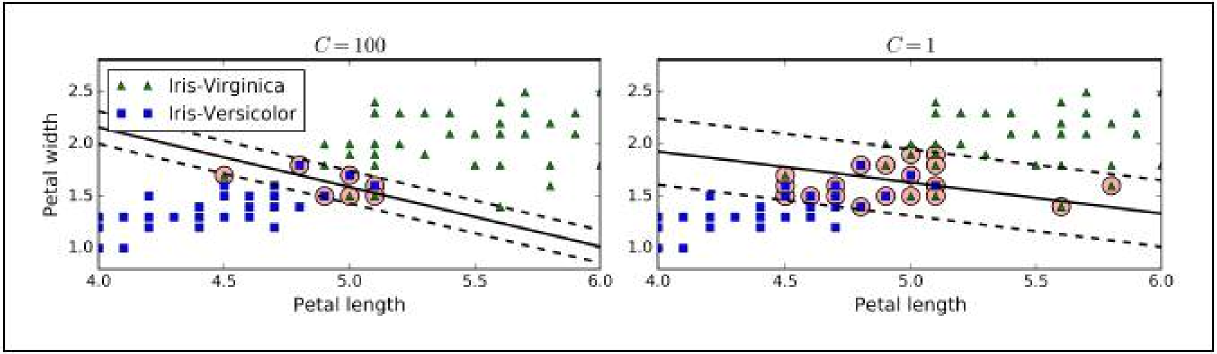
✓ 이 비용함수의 첫 번째 항은 모델이 작은 가중치 벡터를 가지도록 제약을 가해 최소 마진을 극대화 하는 것이고, 두 번째 항은 모든 마진 오류를 계산하는 것임.

✓ 어떤 샘플이 서포트 벡터 경계선에서 올바른 방향으로 벗어나 있다면 마진 오류는 0이고, 그렇지 않다면 마진 오류는 올바른 방향의 서포트 벡터 경계선까지의 거리에 비례함. 이 항을 최소화하려면 마진 오류를 가능한 한 줄이고 크기도 작게 만들어야 함.

- **힌지 손실 함수**(hinge loss function)



- 사이킷런의 SVM 모델에서는 하이퍼파라미터 C를 사용해 소프트 마진 분류의 균형을 조절할 수 있음.
  - ✓ C 값을 줄이면 서포트 벡터를 지나는 직선들의 폭이 넓어지지만 마진 오류도 커짐.
  - ✓ 다음 그림은 선형적으로 구분되지 않는 데이터셋에 두 개의 소프트 마진 SVM 분류기로 만든 결정 경계와 마진을 보여줌.



  - ✓ 두 번째 분류기(작은 C 값)가 더 잘 일반화가 되는 경향이 있음. 대부분의 마진 오류는 결정 경계를 기준으로 올바른 클래스로 분류되기 때문에 이 훈련 세트에서 예측 에러는 마진 오류보다 작음.