Ch 1. 서포트 벡터 머신(Support Vector Machine, SVM)

● 서포트 벡터 머신 모형의 개요

- 지도학습. 분류, 회귀 이상치 탐색에도 사용 가능한 다목적의 머신러닝 모델.
- 이진 분류 (binary classification) 모형.
  - x는 p 차원의 연속형 또는 범주형 변수
  - y는 이진변수로 -1 또는 1의 값만 가짐.
- 특징
  - 선형 및 비선형의 복잡한 분류 문제에 잘 들어맞는 강력한 학습 알고리즘임.
  - 작거나 중간 크기의 데이터셋에 적합함.

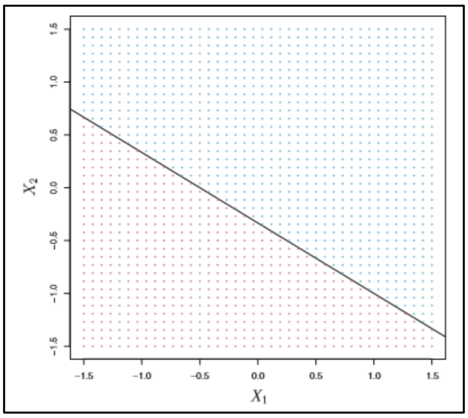
● 최대 마진 분류기(Maximal Margin Classifier)

- 초평면(Hyperplane)
  - p차원 공간에서, $p-1$ 차원의 편평(flat)한 공간을 말함.

    예) 2차원 공간에서의 초평면: $b + w_1x_1 + w_2x_2 = 0$

    $b + w_1x_1 + w_2x_2 = 0$을 만족하는 $x = (x_1, x_2)^T$는 초평면상에 놓이게 됨.

    $b + w_1x_1 + w_2x_2 > 0$ VS $b + w_1x_1 + w_2x_2 < 0$ 로 2차원 공간의 점들이 분류됨.



- **분리 초평면(Separating Hyperplane)을 사용한 분류**

  - 분리 초평면

    ▶ 훈련데이터 셋 $\left(\boldsymbol{x}^{(i)}, y^{(i)}\right), i = 1, \dots, m$이 주어졌을 때,

    $$\begin{cases} \boldsymbol{w}^T\boldsymbol{x}^{(i)} + b > 0\text{일 때,} & y^{(i)} = 1 \\ \boldsymbol{w}^T\boldsymbol{x}^{(i)} + b < 0\text{일 때,} & y^{(i)} = -1 \end{cases}$$

    을 만족하는 $\boldsymbol{w}^T\boldsymbol{x} + b = 0$로 정의됨.

    ▶ <u>모든 $i = 1, \dots, m$에 대하여 $y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) > 0$를 만족해야 함.</u>

    ▶ 분리초평면은 특성공간에 대한 **선형분리경계**(linear decision boundary)를 생성.

    ▶ 특성 공간이 p차원인 경우의 스칼라 표기법 :

    $\boldsymbol{x} = \left(x_1, x_2, \dots, x_p\right)$, $\boldsymbol{w} = \left(w_1, w_2, \dots, w_p\right)$이므로,

    $$\boldsymbol{w}^T\boldsymbol{x} + b = w_1x_1 + w_2x_2 + \cdots + w_px_p + b = 0$$

  - 분리 초평면을 활용한 예측

    ▶ 분리 초평면이 존재하여 $\widehat{\boldsymbol{w}}^T\boldsymbol{x} + \hat{b} = 0$으로 주어졌다고 할 때,

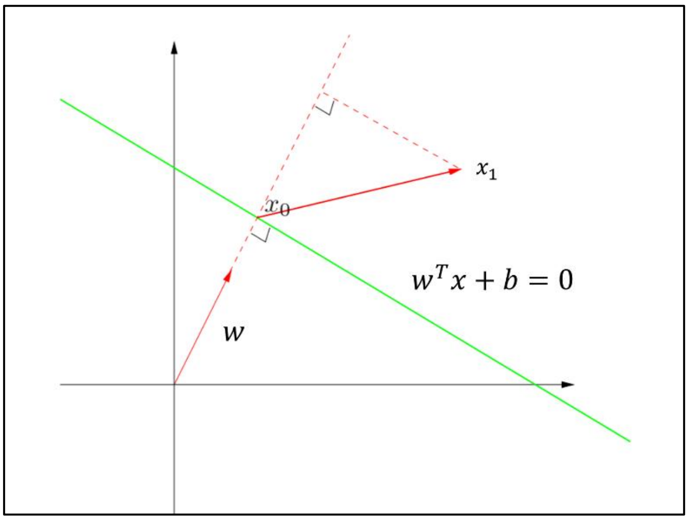
    ▶ 새로운 평가데이터 $\boldsymbol{x}^t$에 대한 $\widehat{\boldsymbol{w}}^T\boldsymbol{x}^t + \hat{b}$의 부호를 이용.

    $$\hat{y}^t = \begin{cases} 1 & , \quad \widehat{\boldsymbol{w}}^T\boldsymbol{x}^t + \hat{b} > 0 \\ -1 & , \quad \widehat{\boldsymbol{w}}^T\boldsymbol{x}^t + \hat{b} < 0 \end{cases}$$

    ▶ $\left|\widehat{\boldsymbol{w}}^T\boldsymbol{x}^t + \hat{b}\right|$의 크기로 $\boldsymbol{x}^t$에 대한 범주 예측값을 더욱 확신할 수 있음.

- 분리 초평면 훈련 알고리즘 (Rosenblatt's Perceptron Learning Algorithm)

▶ $\boldsymbol{x}$의 특성공간에서 임의의 한 점 $\boldsymbol{x_1}$과 분리초평면 $h(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b = 0$와의 수직 거리는 $\boldsymbol{w}^T\boldsymbol{x_1} + b$의 크기에 비례함.



▶ Rosenblatt의 분리초평면 훈련 알고리즘은 오분류된 관찰점들의 결정경계까지의 거리를 최소화하도록 함.

▶ 훈련 데이터셋에서 오분류된 관찰점의 인덱스 집합을 M이라고 할 때, 아래와 같이 정의되는 $\mathrm{D}(\boldsymbol{w}, b)$를 최소로 하는 문제임.

$$\mathrm{D}(\boldsymbol{w}, b) = -\sum_{i \in M} y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right)$$

▶ 경사하강법을 적용하여, 파라미터를 아래와 같이 업데이트 함.

▶ 그래디언트의 계산

$$\frac{\partial \mathrm{D}(\boldsymbol{w}, b)}{\partial \boldsymbol{w}} = -\sum_{i \in M} y^{(i)}\boldsymbol{x}^{(i)}$$

$$\frac{\partial \mathrm{D}(\boldsymbol{w}, b)}{\partial \boldsymbol{b}} = -\sum_{i \in M} y^{(i)}$$

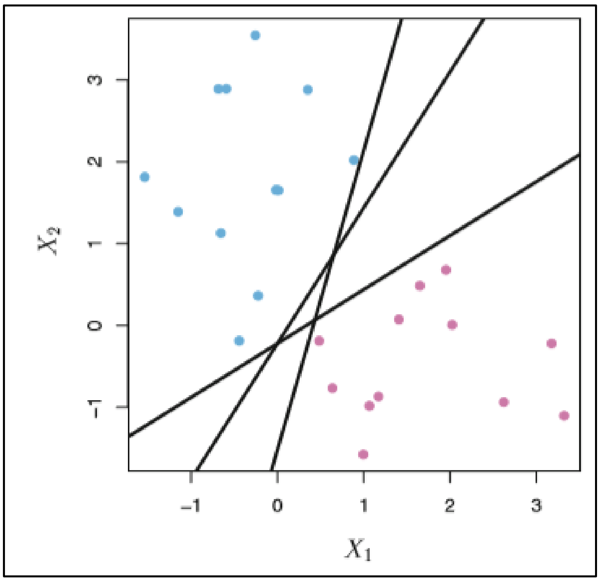
$$\begin{pmatrix} \boldsymbol{w} \\ b \end{pmatrix}^{next} = \begin{pmatrix} \boldsymbol{w} \\ b \end{pmatrix} - \alpha \begin{pmatrix} \frac{\partial \mathrm{D}(\boldsymbol{w}, b)}{\partial \boldsymbol{w}} \\ \frac{\partial \mathrm{D}(\boldsymbol{w}, b)}{\partial \boldsymbol{b}} \end{pmatrix}$$

▶ 특성공간이 목표범주에 따라 선형 분리 가능한 경우(linearly separable), 이 알고리즘은 분리초평면으로 수렴하는 것이 이론적으로 증명됨.

- 한계점

  ▶ 선형분리가 불가능한 경우 해가 존재하지 않음.

  ▶ 선형분리 가능한 경우도 무수히 많은 해가 존재하며, 최적화 결과가 초기값에 의존한다는 문제가 있음.

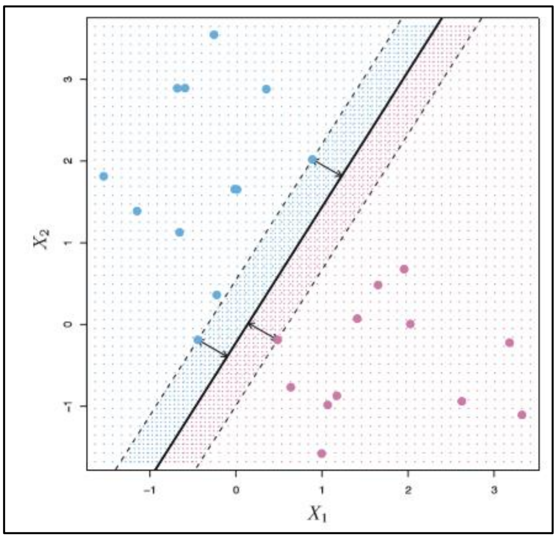


- **최대마진 분류기(Maximal Margin Classifier)**

  - **마진(Margin)** : 각 훈련 자료와 분리초평면과의 거리 중 최소값

  - 최대마진 분류기는 주어진 훈련자료에서 최대 마진을 가지는 분리초평면에 해당.

  - **서포트 벡터(support vector)** : 훈련 자료들 중 분리초평면에 대하여 최대 마진을 가지는 관찰점들.

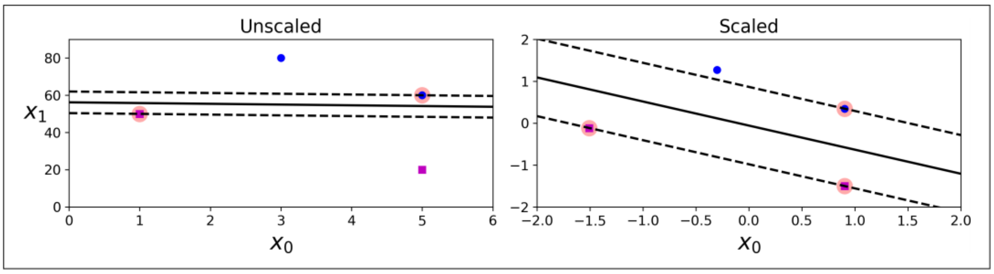


- 최대마진 분류기 특징

▶ 기존 분리초평면 알고리즘과는 달리, 유일한 해를 구해줄 뿐 아니라, 평가자료에 대한 분류 예측력도 더 좋은 편임.

▶ 최대마진 분류기는 서포트 벡터에 의해 전적으로 결정되며, 서포트 벡터가 아닌 나머지 관찰점들은 전혀 영향을 미치지 않음.

▶ 특성 스케일에 민감하므로, 특성변수를 스케일을 조정하면 더 좋은 결정경계를 얻을 수 있음.



▪ 최대 마진 분류기의 최적화 문제

- 최대마진 분류기의 목적함수와 제약식

$\text{Maximize}_{\boldsymbol{w},b}\ \ M$
조건 : $\boldsymbol{w}^T\boldsymbol{w} = 1$이고, $i = 1,\ldots,m$에서 $y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq M$

$\boldsymbol{w}^T\boldsymbol{w} = 1$ 제약 하에서, $y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right)$는 $\boldsymbol{x}^{(i)}$와 분류기와의 거리가 됨 → M이 마진.

이는 다음과 같이 표현가능함.

$\text{Minimize}_{\boldsymbol{w},b}\ \ \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$
조건 : $i = 1,\ldots,m$일 때, $y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1$

이는 선형 제약하의 이차함수 최적화인 Quadratic Programming 문제에 해당하며, 일반화된 라그랑주 함수는 아래와 같이 표현됨.

▶ $L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - \sum_{i=1}^{m}\alpha^{(i)}\left(y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) - 1\right)$를 최소로 하는 $\boldsymbol{w}, b, \boldsymbol{\alpha}$ 찾기

조건 : $i = 1,\ldots,m$일 때, $\alpha^{(i)} \geq 0$

해가 있다면 아래의 **Karush-Kuhn-Tucker (KKT) 조건**을 만족하는 정류점$(\hat{\boldsymbol{w}}, \hat{b}, \hat{\boldsymbol{\alpha}})$ 중에 있게 됨.

① $\frac{\partial}{\partial w} L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \boldsymbol{w} - \sum_{i=1}^{m} \alpha^{(i)} y^{(i)} \boldsymbol{x}^{(i)} = 0$

② $\frac{\partial}{\partial b} L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = -\sum_{i=1}^{m} \alpha^{(i)} y^{(i)} = 0$

③ 모든 $i = 1, \dots, m$ 에 대해, $y^{(i)}\left(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b\right) - 1 \geq 0$

④ 모든 $i = 1, \dots, m$ 에 대해, $\alpha^{(i)} \geq 0$

⑤ 모든 $i = 1, \dots, m$ 에 대해, $\alpha^{(i)}\left(y^{(i)}\left(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b\right) - 1\right) = 0$: $\alpha^{(i)} = 0$이거나, 그렇지 않고 $\alpha^{(i)} > 0$인 경우면 i번째 제약이 등식 조건 $y^{(i)}\left(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b\right) = 1$ 을 만족해야 함. → 상보적 여유성 조건(complementary slackness condition)

▶ $\alpha^{(i)} > 0$인 경우는 i번째 관찰점이 분류경계면에 놓인 서포트 벡터임을 의미

KKT조건은 정류점이 부등식의 제약이 있는 최적화 문제의 해가 되기 위한 필요 조건이자, 일정 조건하에서는 충분 조건이기도 함. 최대마진 분류기는 이 조건을 만족하므로, KKT 조건을 만족하는 어떤 정류점도 최적화의 해임이 보장됨.

- 최대 마진 분류기의 쌍대 문제

  - 제약이 있는 최적화 형태인 원 문제(primal problem)가 주어지면 이와 깊게 연관된 쌍대 문제(dual problem)로 표현 가능함.

  - 일반적으로 쌍대 문제의 해는 원 문제 해의 하한값이지만, 일정 조건 하에서는 원 문제와 동일한 해를 제공함. 최대 마진 분류기의 원 문제는 이 조건을 만족시키므로 원 문제와 쌍대 문제는 동일한 해를 제공하게 됨.

  - 최대 마진 분류기의 쌍대 형식

    ▶ $L(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha^{(i)}\alpha^{(j)} y^{(i)} y^{(j)} \boldsymbol{x}^{(i)^T} \cdot \boldsymbol{x}^{(j)} + \sum_{j=1}^{m} \alpha^{(i)}$ 를 최대로 하는 $\boldsymbol{\alpha}$ 찾기.

    조건 : $i = 1, \dots, m$일 때, $\alpha^{(i)} \geq 0$ 이고, $\sum_{i=1}^{m} \alpha^{(i)} y^{(i)} = 0$.

- 쌍대 형식은 $\alpha^{(i)}$에 관한 선형 제약 하에서 이차함수에 관한 최적화 문제로, 원 문제보다 단순한 이차 최적화 문제(Quadratic Programming)임. 이 쌍대 형식의 해인 $\alpha^{(i)}$를 찾고, 이를 원 문제의 KKT 조건에 다시 대입하여, 원 문제의 목적함수를 최소로 하는 $\boldsymbol{w}, b$를 아래와 같이 구할 수 있음.

  $\widehat{\boldsymbol{w}} = \sum_{i=1}^{m} \hat{\alpha}^{(i)} y^{(i)} \boldsymbol{x}^{(i)}$

  $\hat{b} = \frac{1}{n_s} \sum_{i=1}^{m} \left(y^{(i)} - \widehat{\boldsymbol{w}}^T \cdot \boldsymbol{x}^{(i)} \right)$, ($n_s : \alpha^{(i)} > 0$인 관찰점(서포트 벡터)의 개수)

- 쌍대문제로 표현하는 경우 원 문제에서 적용이 안되는 **커널 트릭**이 가능해짐.

- 예제 1

  두 개의 훈련 자료 $x_1 = \binom{2}{3}, y_1 = 1$과 $x_2 = \binom{4}{1}, y_2 = -1$가 2차원 공간에 놓여 있을 때, 최대마진 분류기를 도출하여라.

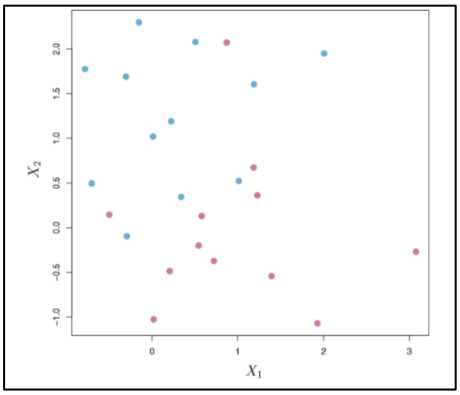
- 예제 2

세 개의 훈련 자료 $x_1 = \begin{pmatrix} 2 \\ 3 \end{pmatrix}, y_1 = 1$과 $x_2 = \begin{pmatrix} 4 \\ 1 \end{pmatrix}, y_2 = -1$과 $x_3 = \begin{pmatrix} 5 \\ 1 \end{pmatrix}, y_2 = -1$가 2차원 공간에 놓여 있을 때, 최대마진 분류기를 도출하여라.

- 최대 마진 분류기를 활용한 예측
  - 최대마진 분류기가 $\hat{\boldsymbol{w}}^T\boldsymbol{x}^t + \hat{b} = 0$으로 주어졌다고 할 때, 새로운 평가데이터 $\boldsymbol{x}^t$는 아래와 같이 분류함.

$$\hat{y}^t = \begin{cases} 1 & ,\hat{\boldsymbol{w}}^T\boldsymbol{x}^t + \hat{b} > 0 \\ -1 & ,\hat{\boldsymbol{w}}^T\boldsymbol{x}^t + \hat{b} < 0 \end{cases}$$

- 최대마진 분류기의 한계
  - 최대 마진 분류기는 훈련 데이터가 선형 분리 가능한 경우에만 해가 존재함. 따라서 아래와 같이 선형 분리 불가능한 경우에는 적용할 수 없음.



  - 선형 분리 불가능한 경우 특성공간을 확장하여 비선형의 분류기를 만들거나, 마진 내에 약간의 훈련 데이터가 허용되는 좀 더 유연한 분류 알고리즘이 대안이 될 수 있음.