CH 9. 서포트 벡터 머신

9-1. 최대 마진 분류기

9-1-1. 초평면(Hyperplane)은 무엇인가?

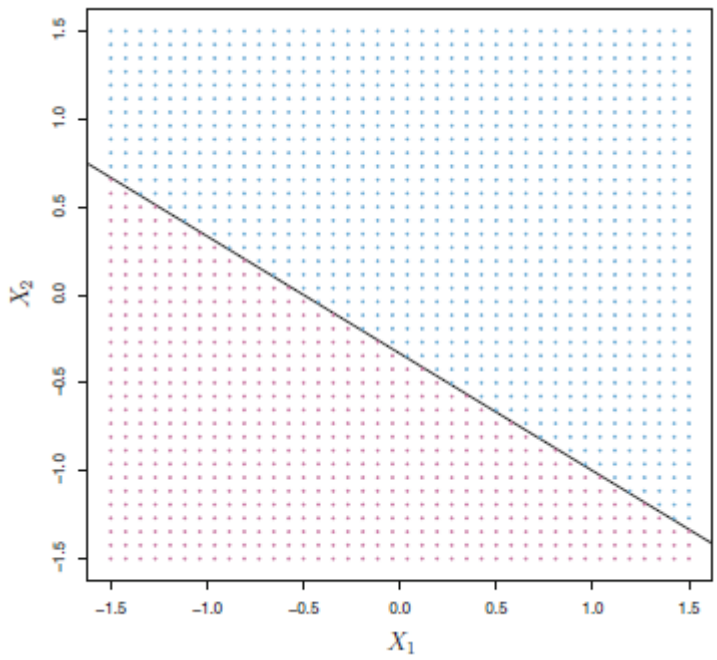
p 차원공간에서 **초평면**은 차원이 $p-1$인 평평한 아핀(affine) 부분공간이다.

ex) 2차원 : 평평한 1차원 부분공간, 선

p차원의 초평면

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = 0$$

초평면은 p차원의 공간을 **두 개의 부분으로 이등분**한다. 어떤 점(샘플)이 초평면의 어느 쪽에 있는지는 위 식의 부호를 계산함으로써 알 수 있다.



9-1-2. 분리 초평면(Seperating Hyperplane)을 사용한 분류

$$x_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{1p} \end{pmatrix}, \cdots, x_n = \begin{pmatrix} x_{n1} \\ \vdots \\ x_{np} \end{pmatrix} \; y_1, \cdots, y_n \in \{-1, 1\}$$

p차원의 공간에서 n개의 훈련 관측치로 구성되는 $n \times p$ 데이터 행렬 $\mathbf{X}$ 가 있다고 가정하자.

모든 훈련 관측치들을 클래스 라벨에 따라 **완벽하게 분리하는 초평면을 구성할 수 있다고 가정**하자.

이 때, 분리 초평면은 다음과 같은 식들을 만족한다.

$y_i = 1$이면, $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p > 0$

$y_i = -1$이면, $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p < 0$

이는 모든 $i = 1, \cdots, n$ 에 대하여 $y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) > 0$ 을 만족한다.

검정 관측치 x^*는 $f(x^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_p x_p^*$의 **부호를 기반으로 분류**되고

$f(x^*)$의 **크기는 검정 관측치가 초평면으로부터 얼마나 떨어져 있는지**를 의미한다.

$f(x^*)$가 0과 가까운 값이면 x^*가 초평면 근처에 놓여 있으므로 x^*의 클래스 할당(예측)에 대한 확신이 덜하다.

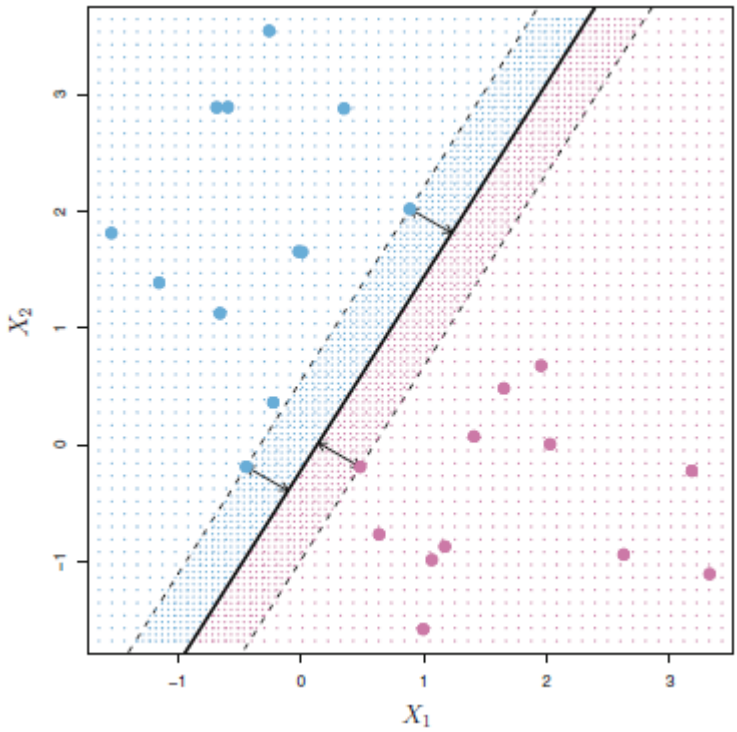
9-1-3. 최대 마진 분류기

초평면을 사용하여 데이터가 완벽하게 분류될 수 있으면 무한개의 초평면이 존재할 것이다. (약간의 이동과 회전)

그 중 어떤 초평면을 선택할 것이냐가 중요하다.

-> 훈련 관측치들로부터 가장 멀리 떨어진 분리 초평면인 **최대 마진 초평면을 선택**

- 마진 : 관측치들 중에서 초평면까지의 가장 짧은 거리
- 최대 마진 초평면 : 마진이 가장 큰 분리 초평면
- 평판 (slab) : 최대 마진 초평면으로부터 양 쪽으로 마진만큼 떨어진 초평면 사이의 공간
- 서포트 벡터 (support vector) : 평판의 경계에 놓여진 벡터
    - 서포트 벡터의 위치에 따라 최대 마진 및 최대 마진 초평면이 결정된다. 즉 최대 마진 초평면은 서포트 벡터에 직접적으로 의존적이다.
    - 하지만 마진 밖의 다른 벡터들에게는 전혀 영향을 받지 않는다.
- 단점 : p값이 클 때 (차원이 클 때) 과적합에 이를 수 있다.



9-1-4. 최대 마진 분류기의 구성

최대 마진 초평면은 다음 최적화 문제의 해(솔루션)이다.

$$\underset{\beta_0,\beta_1,\cdots,\beta_p}{maximize} \quad M$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 = 1$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M \qquad \forall \quad i = 1, \cdots, n$$

M이 양수이면 각 관측치가 초평면의 올바른 쪽에 있게 되도록 보장한다. (각 관측치가 일부 완충공간(cushion)을 갖고 초평면의 올바른 쪽에 있도록 한다.)

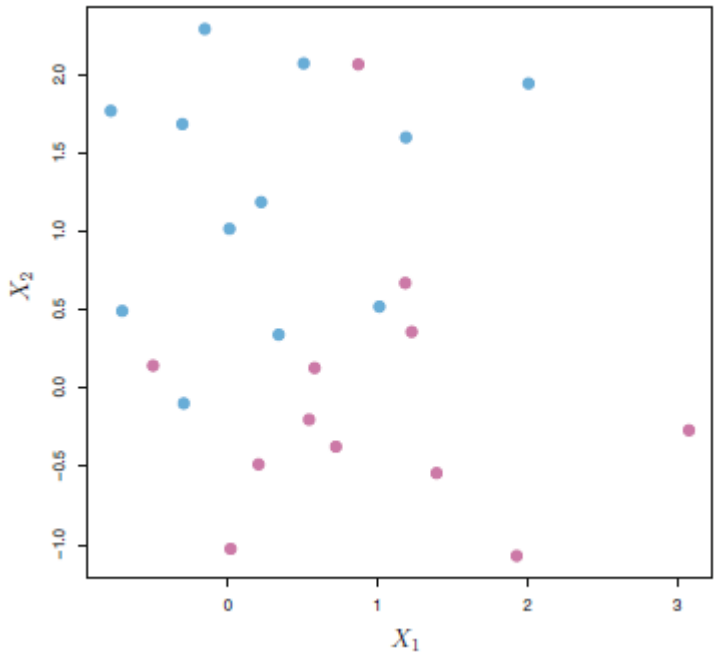
9-1-5. 분류 불가능한 경우

만일 앞서서 계속 가정했던

> 모든 훈련 관측치들을 클래스 라벨에 따라 **완벽하게 분리하는 초평면을 구성할 수 있다고 가정**

조건이 만족되지 않는다면 분리 초평면은 존재할 수 없고 최대 마진 분류기 또한 없다.

이 경우 위의 최적화 문제인 $M > 0$의 해 또한 없다.



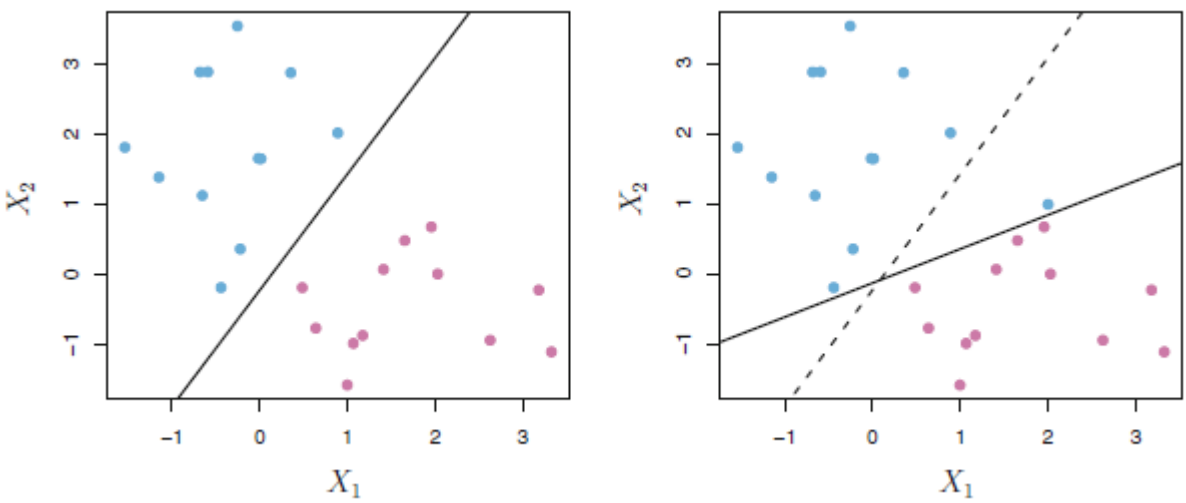
모든 클래스를 ***정확하게 (exactly)*** 분류할 수 없으므로

소프트 마진(soft margin)을 사용하여 클래스들을 ***거의 (almost)*** 분류하는 초평면을 사용한다.

서포트 벡터 분류기 (Support Vector Classifier)이다.

9-2. 서포트 벡터 분류기

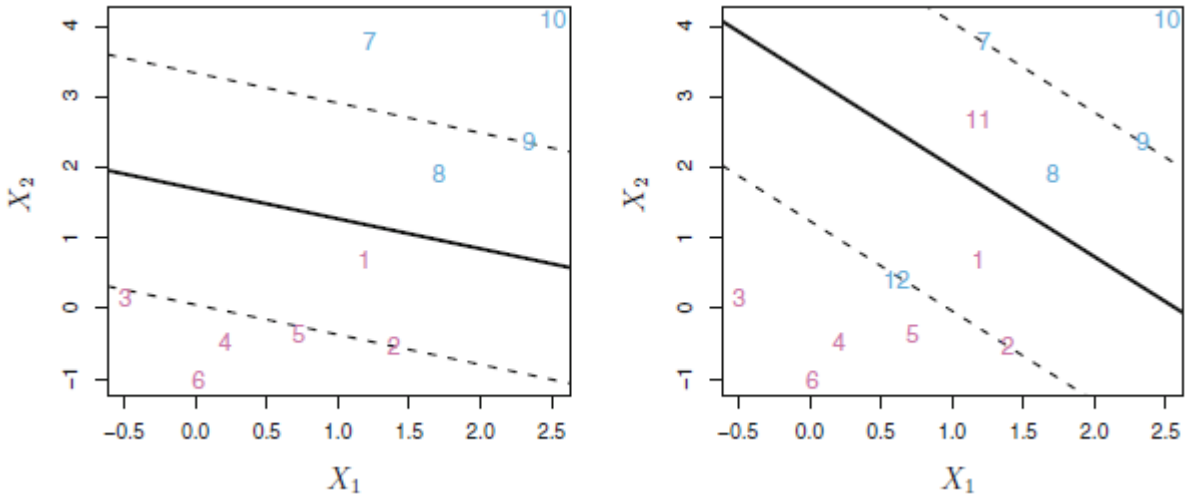
9-2-1. 서포트 벡터 분류기의 개요



- 단점 : 개별 관측치 (특히 분리 초평면으로부터 가까운 벡터들)에 대해서 **민감**하다.
  - 과적합의 위험도가 높다.

1. 개별 관측치에 대해 *robust*하다.
2. *대부분*의 훈련 관측치들을 더 잘 분류한다.

   라는 목적을 갖고 초평면에 기반한 새로운 분류기를 생성한다.



- **가능한 한** 모든 관측치가 마진의 올바른 쪽에 위치하도록 **가장 큰 마진**을 찾는다.
- 하지만 **일부** 관측치들은 마진의 옳지 않은 쪽에 있거나, 심지어 초평면의 올지 않은 쪽에 있을 수 있도록 **허용**.

9-2-2. 서포트 벡터 분류기의 세부 사항

서포트 벡터 분류기는 다음 최적화 문제의 해(솔루션)이다.

$$\underset{\beta_0,\beta_1,\cdots,\beta_p}{maximize} \quad M$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 = 1$$

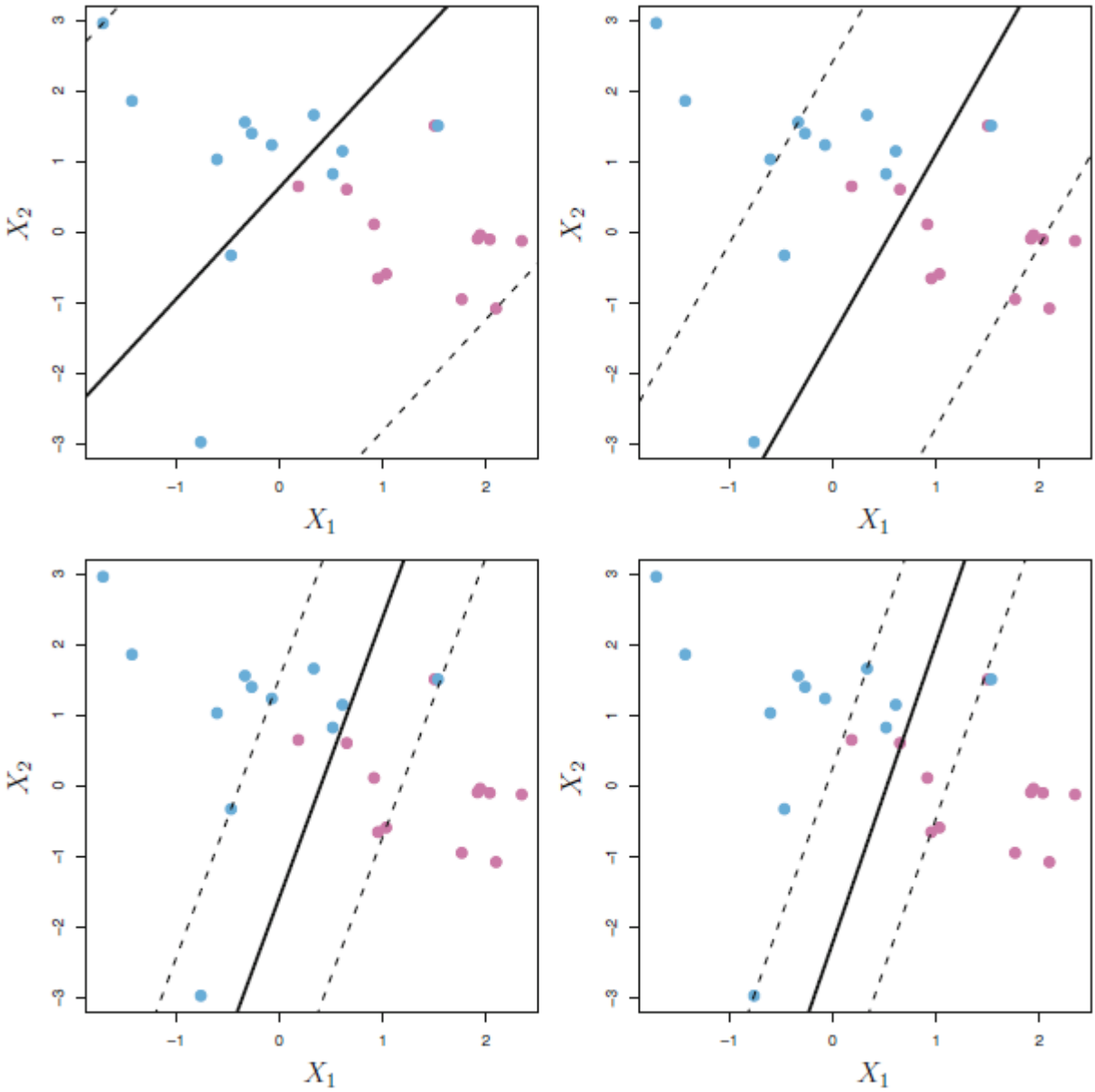
$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M(1 - \epsilon_i)$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C$$

- C는 음수가 아닌 하이퍼파라미터
- M은 마진의 폭이고 가능한 한 이 값을 크게 하는 것이 목표
- **슬랙변수 (slack variable)** : $\epsilon_i \qquad \forall i = 1, \cdots, n$
  - 개별 관측치들이 마진 또는 초평면의 **옳지 않은 쪽에 있도록 허용**해주는 변수
  - i번째 관측치가 **초평면과 마진에 관해** 어디에 위치하는가를 알려줌
  - $\begin{cases} \text{마진의 올바른 쪽에 위치} & if \quad \epsilon_i = 0 \\ \text{마진의 옳지 않은 쪽에 위치} & if \quad \epsilon_i > 0 \\ \text{초평면의 옳지 않은 쪽에 위치} & if \quad \epsilon_i > 1 \end{cases}$
- **하이퍼파라미터** C

- 마진과 초평면에 대해 **허용**될 ***위반의 수***와 그 ***정도***를 결정
- n개의 관측치에 의해 마진이 위반될 수 있는 양에 대한 *예산(budget)*
- $if \quad C = 0$: 마진을 위반할 예산이 없다. $\rightarrow \epsilon_1 = \cdots = \epsilon_n = 0 \rightarrow$ 단순 최대 마진 초평면의 최적화
- $if \quad C > 0$: C개 이하의 관측치들이 초평면의 옳지 않은 쪽에 있을 수 있다. (최대 C개)

  $\because \quad \epsilon_i > 1$은 초평면의 옳지 않은 쪽에 위치 & $\sum_{i=1}^{n} \epsilon_i \leq C$
- $C \uparrow \rightarrow$ 마진 위반 허용 정도 $\uparrow \rightarrow$ 마진 폭 $\uparrow \rightarrow$ 덜 엄격하게 적합 $\rightarrow$ 분산 $\downarrow$ 편향 $\uparrow$
- $C \downarrow \rightarrow$ 마진 위반 허용 정도 $\downarrow \rightarrow$ 마진 폭 $\downarrow \rightarrow$ 더 엄격하게 적합 $\rightarrow$ 분산 $\uparrow$ 편향 $\downarrow$



서포트 벡터 분류기는 마진 상에 놓이거나 마진을 위반하는 관측치들 (서포트 벡터)로부터만 영향을 받는다.

즉, 엄격하게 마진의 올바른 쪽에 놓인 관측치 (높은 확신을 갖은 관측치)들은 분류기에 영향을 주지 않는다.

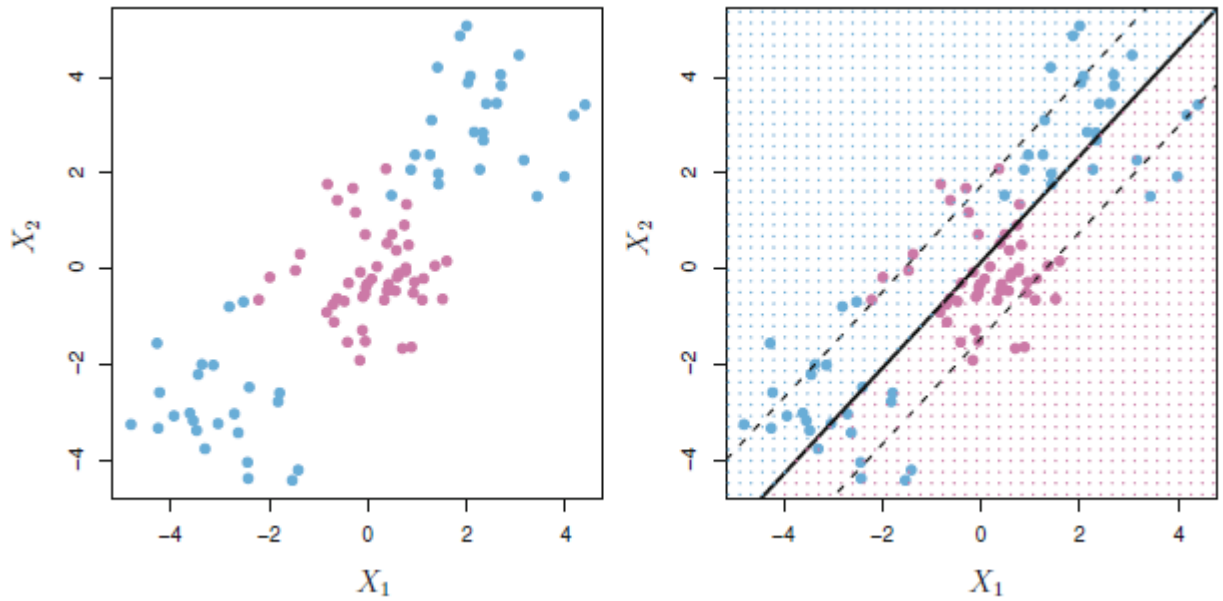
$\because$ 높은 확신을 갖은 관측치들의 ϵ_i 값은 0이고 C에 영향을 주지 않기 때문이다.

$\therefore$ C가 서포트 벡터 분류기의 bias-variancec trade-off를 제어하는 하이퍼마라미터이다.

& 초평면으로부터 멀리 떨어진 관측치들에 대해 상당히 robust하다. (민감도가 낮다) : 로지스틱 회귀와 관련

9-3. 서포트 벡터 머신

9-3-1. 비선형 결정경계를 가진 분류



위와 같이 클래스 경계가 비선형일 때 기존의 서포트 벡터 분류기의 성능은 매우 나쁘다. 이 경우 설명변수들의 2차, 3차 등의 다항식 함수를 사용하여 변수 공간을 확장함으로써 비선형적으로 문제를 해결할 수 있다.

설명변수들의 2차항을 추가했다고 가정하자. 그러면 다음의 최적화 문제에 대한 해가 비선형 분류기이다.

$$\underset{\beta_0,\beta_{11},\beta_{12},\cdots,\beta_{p1},\beta_{p2},\epsilon_1,\cdots,\epsilon_n}{maximize} \quad M$$

subject to $y_i\left(\beta_0+\sum_{j=1}^{p}\beta_{j1}x_{ij}+\sum_{j=1}^{p}\beta_{j2}x_{ij}^2\right)\geq M(1-\epsilon_i)$

$y_i(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip})\geq M(1-\epsilon_i)$

$\epsilon_i\geq 0, \sum_{i=1}^{n}\epsilon_i\leq C, \sum_{j=1}^{p}\sum_{k=1}^{2}\beta_{jk}^2=1$

9-3-2. 서포트 벡터 머신

서포트 벡터 머신(SVM) : 서포트 벡터 분류기의 확장으로, *커널(kernels)*을 사용하여 특정한 방식으로 변수공간을 확장한 결과

서포트 벡터 분류기 문제에 대한 해는 관측치들의 ***내적(inner products)***만으로 해결된다.

두 관측치 $x_i, x_{i'}$의 내적은 다음과 같다.

$$\langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j}$$

선형 서포트 벡터 분류기는 다음과 같다.

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle$$

여기서 함수 $f(x)$를 평가하기 위해선 새로운 점 x와 각 훈련 포인트 x_i사이의 내적을 계산해야 한다.

이 때 α_i는 서포트 벡터에 대해서만 $\neq 0$, 서포트 벡터가 아닌 관측치들(확신을 갖은 관측치)에 대해선 $\alpha_i = 0$

(그래야만 $f(x)$가 높은 확신을 갖는 관측치들에 대해 robust함)

서포트 포인트들의 인덱스 모임을 S라 하면

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i \langle x, x_i \rangle$$

위의 계산보다 더 적은 항의 계산만을 필요로 한다.

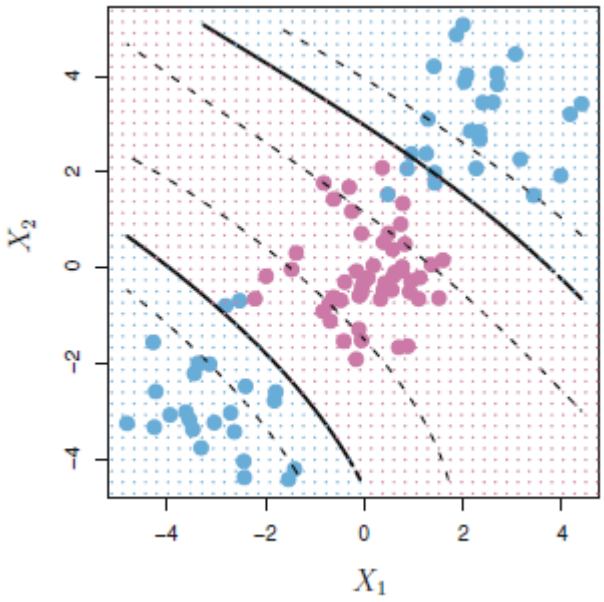
앞으로 모든 내적 표현을 다음과 같은 일반화된 형태로 바꾼다.1

$K(x_i, x_{i'})$

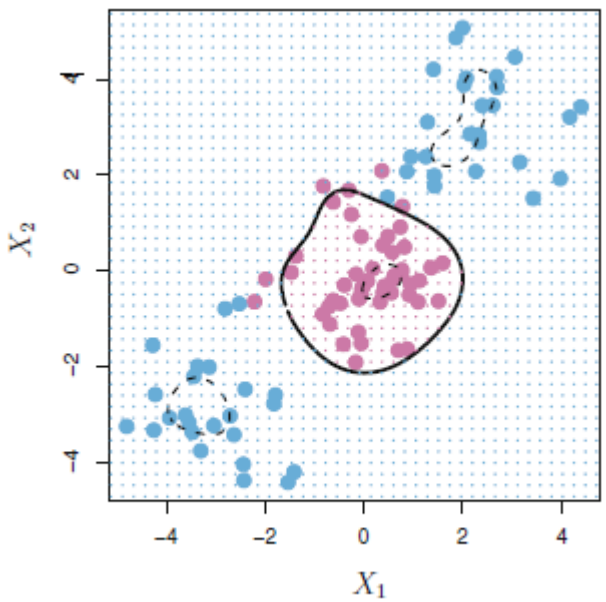
- 여기서 K는 커널(kernel)이라고 언급될 어떤 함수이다.
- *커널은 두 관측치들의 유사성(similarity)을 수량화하는 함수이다.*

- **선형 커널**
  - $K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j}$
  - 피어슨 (Pearson)(표준) 상관을 사용하여 관칙치 쌍의 유사성을 수량화한다.

- **다항식 커널**
  - $K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j}\right)^d$
  - 차수가 d인 다항식 커널 (d는 양의 정수)
  - 표준 선형 커널 대신 $d > 1$인 다항식 커널은 사용하면 더 유연한 결정경계가 형성된다.



- 방사 커널 (radial kernel)
  - $K(x_i, x_{i'}) = \exp\left(-\gamma \sum_{j=1}^{p}(x_{ij} - x_{i'j})^2\right)$
  - γ는 양의 상수
  - 검정 관측치 $x^* = (x_1^*, \cdots, x_p^*)^T$ 가 훈련 관측치 x_i로부터 유클리드 거리 (Euclidean distance)로 멀리 떨어져 있으면, $\sum_{j=1}^{p}(x_{ij} - x_{i'j})^2$ 값은 큰 값이 될 것이다. $\therefore K(x_i, x_{i'}) = \exp\left(-\gamma \sum_{j=1}^{p}(x_{ij} - x_{i'j})^2\right)$ 값은 아주 작은 값이 된다. 즉, x_i가 사실상 $f(x^*)$에 아무런 역할을 하지 않을 것임을 의미한다.
  - x^*로부터 멀리 떨어진 관측치는 예측된 클래스 라벨에 아무런 영향을 주지 않는다.
  - 주변 관측치들만이 클래스 라벨에 영향을 준다 : 방사 커널은 국소적인 (local) 방식으로 동작한다.
  - 높은 확신을 갖는 관측치들에 대해서는 모델이 robust하다.



이와 같이 커널을 사용하는 이유 (장점)

- 원래 변수들의 함수를 이용하여 실제로 변수공간을 확장하지 않고
- $\binom{n}{2}$개의 서로 다른 모든 쌍 i, i'에 대해 $K(x_i, x_{i'})$ 만 계산하면 된다.