[머신러닝 기반 데이터 분석] 03. 지도학습 모델 적용하기

학습 목표

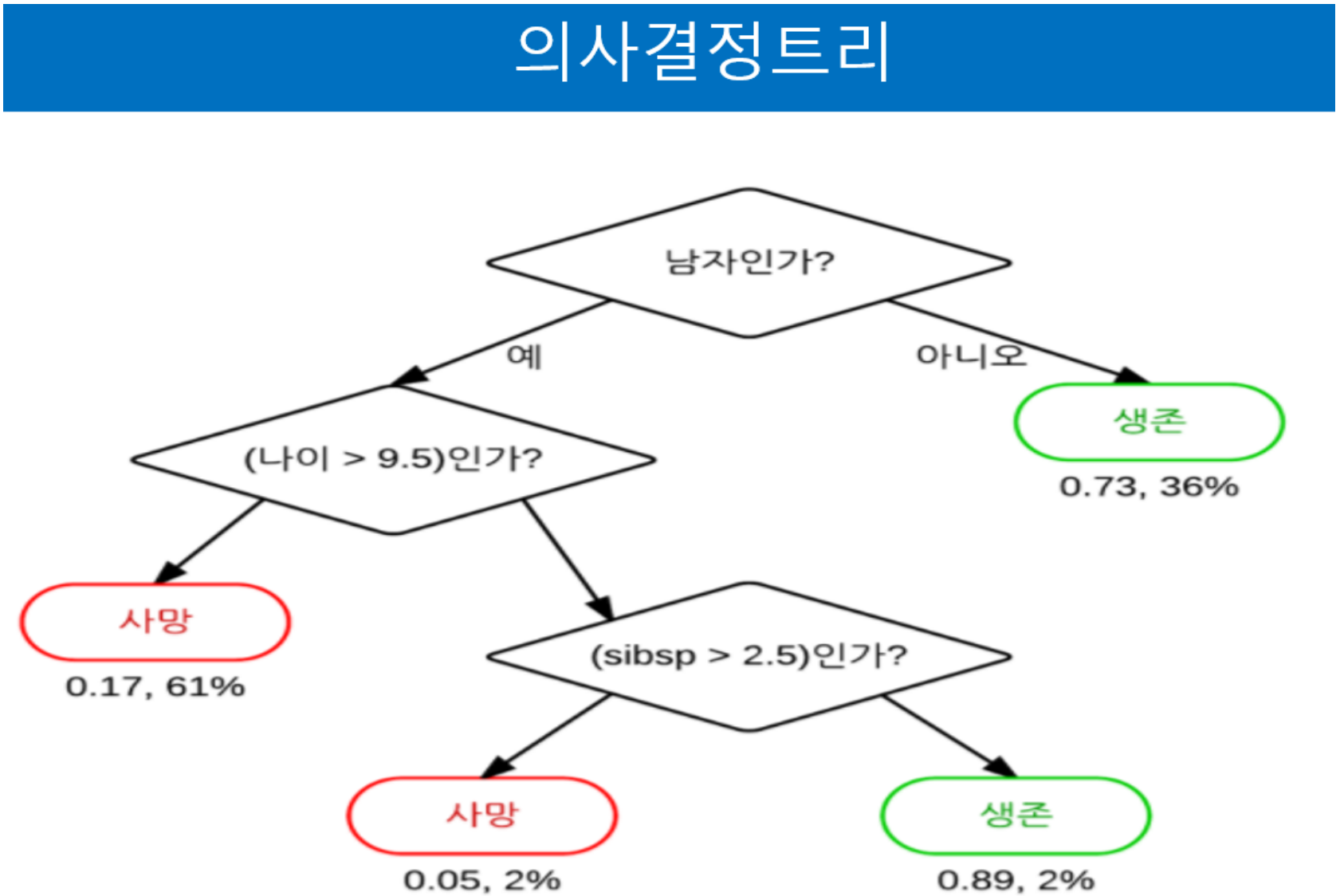
가. 의사결정트리 기법에 대해 이해해 본다.
나. 서포트 벡터 머신 개념에 대해 이해해 본다.

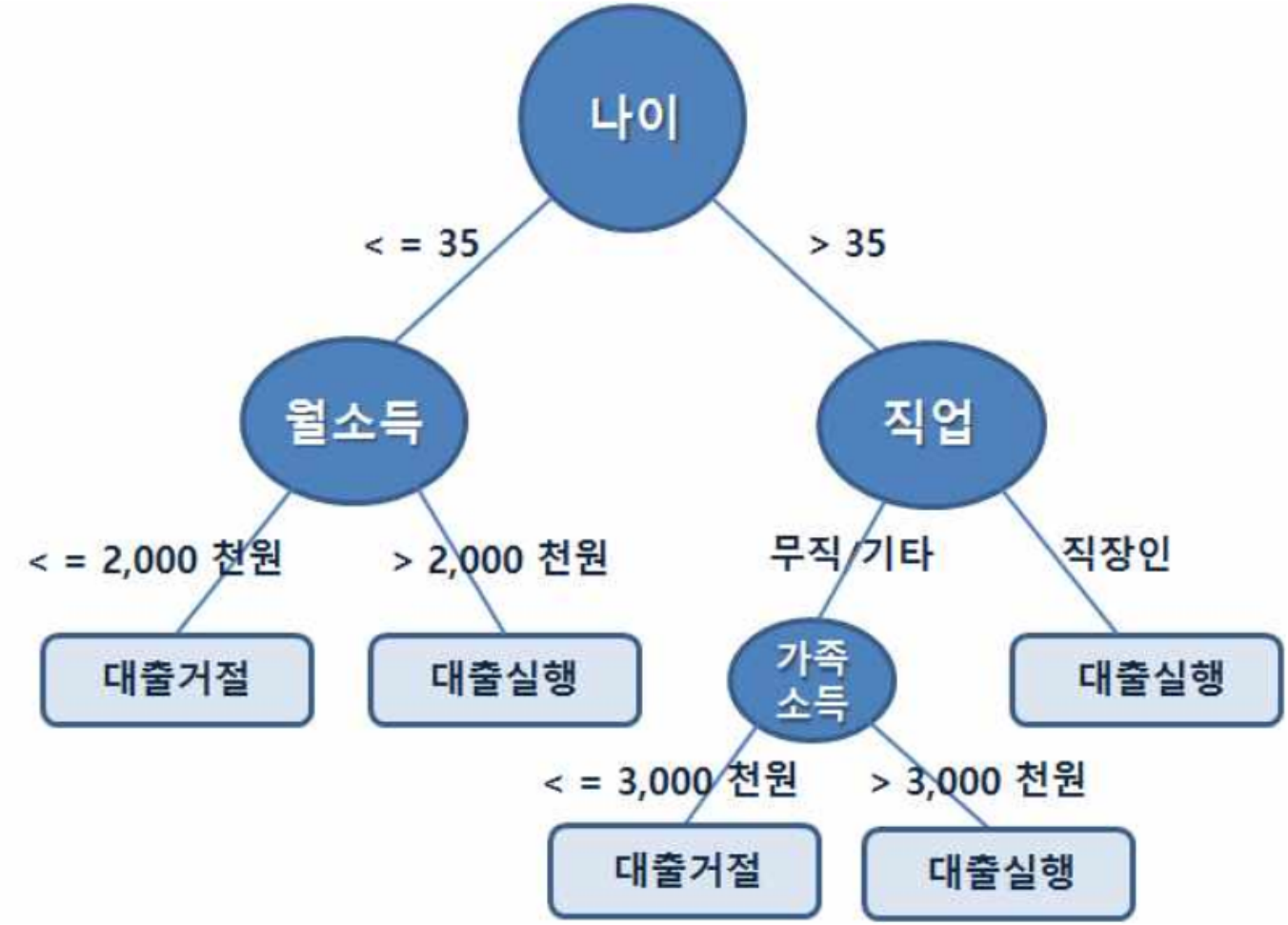
3-1 분류 목적의 머신러닝 기법 적용

가. 의사 결정 트리(decision) 기법

- 결정 트리 학습법은 지도 분류 학습에서 가장 유용하게 사용되고 있는 기법 중 하나이다.
- 예측 모델중의 하나이다. (분류)
- 통계, 데이터 마이닝, 기계 학습에 사용되는 예측 모델링 접근 방식 중 하나이다.
- 목표 변수(target variable)가 **이산형(discrete variable)**를 취할 수 있는 트리 모델을 **분류 나무(classification tree)**라고 한다.
- 목표 변수(target variable)가 **연속형(continuous values)** 값을 취할 수 있는 트리 모델을 **회귀 나무(regression tree)** 라고 한다.

- 의사결정 트리는 설명변수(독립변수)의 특징이나 기준값에 따라 if-then의 형태로 분기된다. 이를 통해 **각 데이터가 주어졌을 때, 어떠한 카테고리로 분류**되는지 쉽게 알 수 있다.
- 아래 의사결정트리의 가장 상위의 루트 노드에서 분류하는데 있어, **성별(남,여)이 사용**되었다. 이는 **생존/사망을 판별**하는데 가장 유의한 변수는 **성별이 가장 유의한 변수**로 해석할 수 있다.
- 각 가지의 제일 마지막 사각형 노드가 리프(Leaf) 노드이다.





[그림 3-3] 의사결정 트리 예시

(NCS 교재 참조)

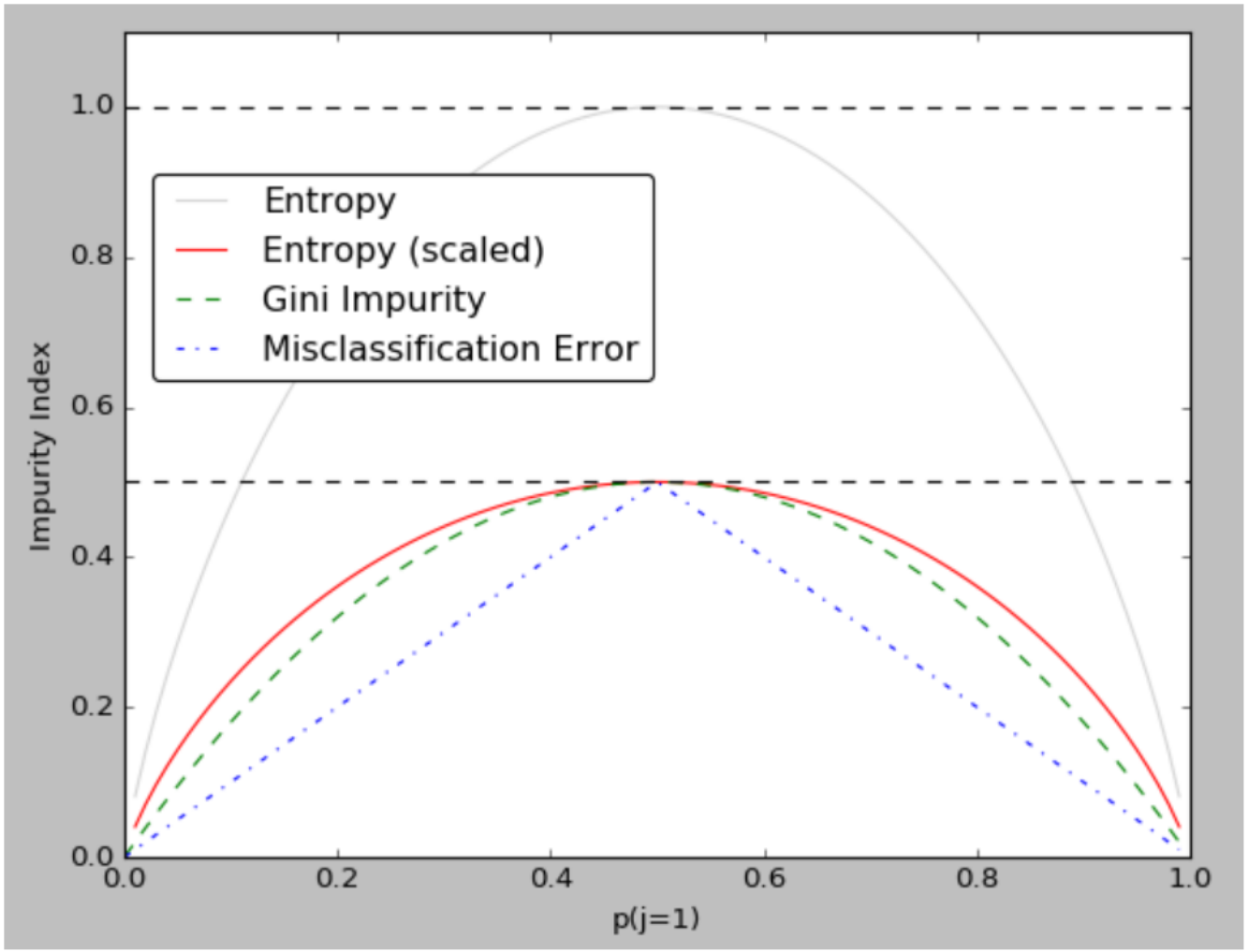
나. 의사 결정 트리(decision) 는 어떤 기준을 가지고 분류하는가?

- 각 노드마다 질문을 던지고 그 응답에 따라 가지를 쳐서 데이터를 분리한다.
- 데이터가 얼마나 잘 분리되어 있는지 평가하기 위해 사용되는 기준은 불순도(impurity)기준으로 사용한다.
- 불순도는 노드에 여러 분류가 섞여 있을 수록 높고, 노드에 하나의 분류만 존재할 때 가장 낮아진다.

-> 노드 분리 후 각 노드의 불순도가 낮아질수록 트리 분류가 잘 된 것으로 판단할 수 있다.

$$Gini\ Coff. = 1 - \sum_{i=1}^{K} p_i^2 \quad , \qquad Entropy\ Coff. = -\sum_{i=1}^{K} p_i \log_2 p_i$$

대표적인 사용되는 불순도
지니 불순도(Gini Impurity) :
엔트로피 지수(Entropy) :



(참조 : https://www.bogotobogo.com/python/scikit-learn/scikt_machine_learning_Decision_Tree_Learning_Informatioin_Gain_IG_Impurity_Entropy_Gini_Classification_Error.php)

다. 의사결정트리 기법

- 입력 변수(input variables)로 부터 목표 변수(target variable)를 예측하는 모델을 생성하는 것이다.
- 결정트리의 '학습'은 학습에 사용되는 자료 집합을 적절한 분할 기준 또는

분할 테스트에 따라 부분 집합으로 나누는 과정이다.

$$(\mathbf{x}, Y) = (x_1, x_2, x_3, \ldots, x_k, Y)$$

종속 변수 Y는 분류를 통해 학습하고자 하는 **목표 변수(target variable)와 벡터 x는 x1,x2,x3등의 입력변수(input variable)**로 구성된다.

라. 의사결정트리 기법 사용하기 위해 어떤 패키지가 있을까?

R에서 의사결정트리 기법을 구현한 여러가지 패키지가 존재한다.

대표적인 패키지 rpart (rpart 함수)
party 패키지 (ctree 함수)
C50 패키지 (C5.0 함수)

[rpart 패키지의 rpart 함수]

rpart는 대표적인 의사결정 트리 기법(CART : Classification and Regresson Trees)를 구현한 패키지.

[party 패키지의 ctree 함수]

[C50 패키지의 C5.0 함수]

C5.0 알고리즘을 구현한 함수

마. 의사결정트리 장단점은 무엇일까?

<표 3-5> 의사결정트리 기법의 장단점

장 점	단 점
• 분류문제 및 수치예측 모두 활용 가능하다. • 결측치가 있는 데이터 효과적으로 처리 가능. • 중요한 변수만 선별할 수 있고, 이를 통해 다른 추가 분석 위한 통찰력을 얻을 수 있다. • 선형성, 정규성, 등분산성 등의 가정이 필요 없는 비모수적 모형이다. • 수학적 지식이 없는 사람도 모형의 결과 이해가 쉽고, 어떤 입력변수가 목표변수를 설명하는데 영향력이 높은지를 알 수 있다. • 분류결과에 대한 Rule 기반의 해석 가능하여 분류결과 이유를 설명해야 할 경우 유용하다,	• 연속형 입력변수를 비연속적인 값으로 취급하므로, 분리의 경계점 근방에서 예측오류 가능성 있음. • 선형 또는 주효과 모형과 같은 해석이 불가능하므로 모형식을 수립해야 하는 경우 적용이 어렵다. • 훈련데이터에 대한 약간의 변경이 발생 시 트리 분류 결정 논리에 큰 변화를 가져온다. (특정 데이터 변화에 분석결과 변화가 민감함) • 모델이 쉽게 과적합화 되거나 과소적합 될 수 있다. • 트리가 너무 커질 경우 패턴 이해하기가 쉽지 않다.

(NCS 교재 참조)

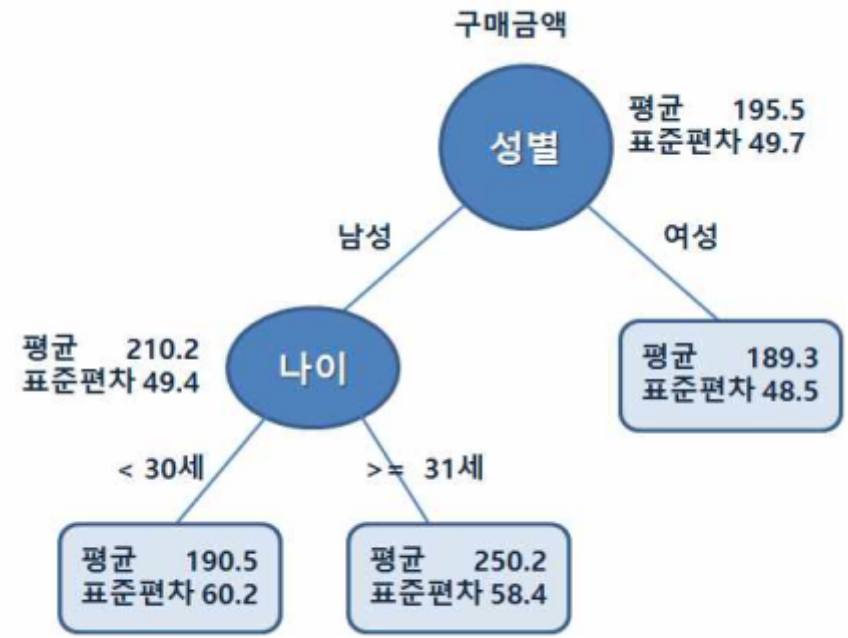
바. 의사결정트리 - 목적변수가(Target)이 수치형에도 사용이 가능한가?

의사결정트리는 분류 목적의 머신러닝 기법이다. 단, 수치예측의 목적으로도 사용할 수 있다.
분류 목적의 의사결정트리를 분류나무(Classification Tree)라 부르고,
수치예측의 의사결정트리를 회귀나무(Regression Tree)라 구별해 부른다.

분류 목적은 노드 분리를 위해
카이제곱 통계량, 지니계수, 엔트로피 지수 등의 불순도 측정으로 트리를 구성.

수치예측의 목적의 의사결정트리 -
목표변수의 평균과 표준편차 혹은 평균과 절대 편차 같은 통계치를 이용하여

마디가 분리된다.



[그림 3-9] 수치예측 목적의 의사결정 트리 예시

구매 금액을 목표변수로 하고 이에 영향을 미치는 설명변수들을 마디로 분리한다고 할 때,
첫번째 분리변수로 성별이 사용되었다.
남성의 구매금액의 평균이 210만원이고, 여성의 구매금액이 189만원이다.
성별에 따라 구매금액의 높고 낮음이 잘 예측될 수 있다.

의사결정트리의 분리기준
수치 예측 목적의 의사결정트리의 경우 F통계량의 p값이나 분산(혹은 표준편차)의 감소량 등을 통해 가지를 분리
(자세한 내용은 NCS 교재 p63 참조)

3-2 분류 목적의 머신러닝 기법 적용 - SVM(Support Vector Machine)

가. 용어 이해하기

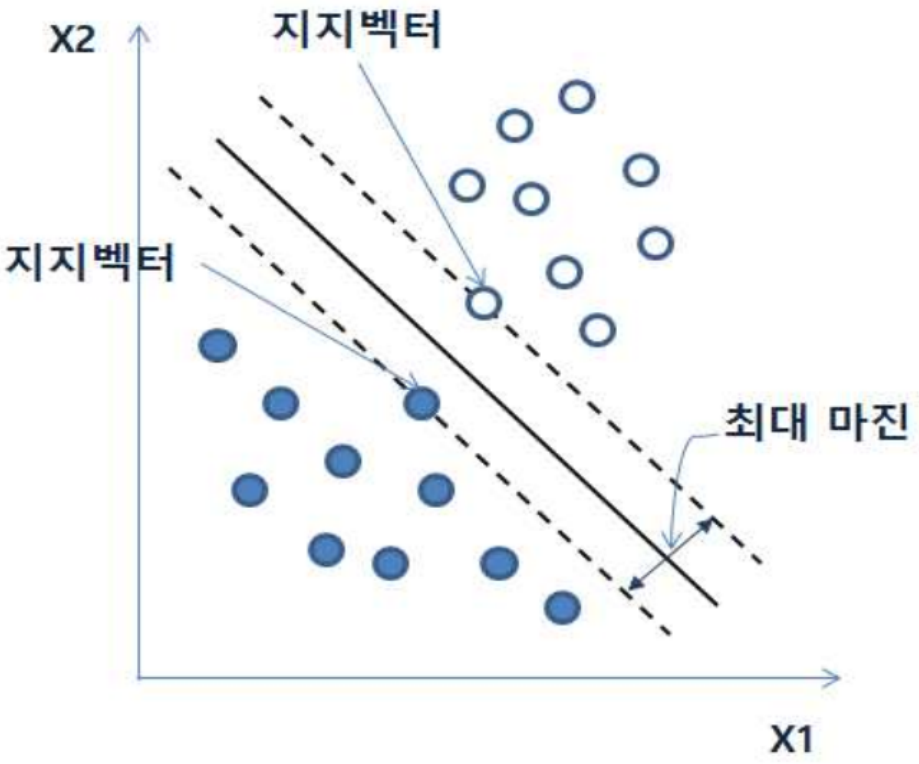
서포트 벡터 머신은?

- 두 범주 간의 데이터를 최대로 나눌 **최대 마진 초평면**을 찾아서 각 데이터를 분류한다.
- 최대 마진 초평면을 찾는 이유는 현재의 훈련 데이터가 아닌

  평가용 데이터나 미래의 데이터에서 경계선 주변의 점들이 **약간 변경되더라도 분류 경계선을 넘어가는 가능성을 최소**로 하기 위한 것이다.

서포트 벡터(Support Vector)

경계선과 가장 가까운 각 분류에 속한 점들을 서포트(혹은 지지 벡터)라고 한다.
하나 이상의 서포트 벡터를 가지고 있어야 한다.



[그림 3-7] 서포트 벡터 머신 개념

나. 모든 데이터를 위와 같이 선형적으로 분류할 수 없을때,

이런 경우에는 **커널 트릭(Kernel Trick)이라는 방법**을 써서 주어진 데이터를 적절한 고차원으로 사상한 뒤, 변환된 차원에서 데이터를 잘 분류할 수 있는 초평면을 찾는다.

주로 사용되는 커널 함수는 **다항 커널, 가우시안 커널, 시그모이드 커널** 등이 있다.

<표 3-7>서포트 벡터 머신 기법의 장단점

장 점	단 점
• 범주분류나 수치예측 문제에 모두 활용 가능하다. • 노이즈 데이터에 영향을 크게 받지 않고, 과적합화가 잘 되지 않는다. • 일반적으로 분류 문제에서 다른 알고리즘보다 분류 성능이 높은 것으로 알려져 있으며, 특히 분류 경계가 복잡한 비선형 문제일 경우 타 기법대비 성능이 좋은 것으로 알려져 있다.	• 최적 분류를 위해 커널함수 및 매개변수 등에 대한 반복적인 조합 테스트가 필요하다. • 입력 데이터의 양이나 변수가 많은 경우 훈련에 오랜 시간이 소요된다. • 배경이 되는 이론 및 알고리즘 구현 시 타 기법에 비해 상대적으로 난해한 면이 있다. • 결과 해석이나 이유 설명 등이 쉽지 않다.

다. 서포트 벡터 머신 기법의 활용 분야

- 서포트 벡터 머신은 분류와 수치예측 문제에 모두 활용 가능하다.
- 분류 성능이 좋으면서도 과적합화가 잘 되지 않고, 일반화 능력이 높아서 정교한 분류 성능이 필요한 분야

  (유전자 데이터 분류, 언어식별, 보안 결함, 이상치 거래 탐색 등)