# 핵심 머신러닝 - 1

| | |
|---|---|
| 생성자 | ㊊ 재환 김 |
| 태그 | 머신러닝 |

## 1.1 머신 러닝이란

- **머신 러닝(Machine Learning, ML)**은 컴퓨터 과학의 한 분야로, 실제 현상에서 관측한 데이터를 토대로 문제를 해결하는 알고리즘을 구축하는 기법입니다.
    - 이 기법은 실제 데이터에서 수집한 정보를 기반으로 알고리즘을 만들어 문제를 해결하거나, 사람이 설계한 알고리즘을 활용하여 문제에 접근합니다.
- **머신 러닝의 주요 방법**:
    1. **데이터 기반 학습**: 실제 데이터에서 수집한 정보로 모델을 구축하는 방식입니다.
    2. **통계 모델 기반 학습**: 통계적 모델을 활용하여 문제를 수학적으로 해결하고 예측하는 방법입니다.

## 1.2 학습 유형

- 머신 러닝의 학습 유형은 크게 **지도 학습**, **준지도 학습**, **비지도 학습**, **강화 학습**으로 나뉩니다.

## 1.2.1 지도 학습 (Supervised Learning)

- **지도 학습**은 입력 데이터와 그에 대응하는 정답(label)이 함께 제공되는 학습 방법입니다.
    - 이 방법에서는 데이터셋이 (x, y) 쌍으로 구성되며, x는 특징 벡터(feature vector), y는 그에 대응하는 정답을 나타냅니다.
    - 예를 들어, x가 사람의 키와 체중을 나타내는 데이터라면, y는 그 사람의 분류(예: 성별이나 나이대)가 될 수 있습니다.
    - **목적**은 주어진 데이터셋을 이용해 입력 데이터를 기반으로 정답을 추정하는 모델을 학습하는 것입니다.
    - 이렇게 학습된 모델은 새로운 데이터가 입력되었을 때 해당 데이터에 맞는 정답을 예측하는 데 활용됩니다.

## 1.2.2 비지도 학습 (Unsupervised Learning)

- **비지도 학습**은 입력 데이터에 정답(label)이 없는 상태에서 학습하는 방법입니다.
    - 이 방법은 **특징 벡터**만을 사용하며, 입력 데이터를 분석하여 패턴이나 구조를 스스로 발견하는 데 초점을 맞춥니다.
    - 주로 **군집화(clustering)**, **차원 축소(dimensionality reduction)**, **이상치 탐지(outlier detection)** 등의 기법에 활용됩니다.
    - 예를 들어, 여러 사람의 데이터를 분석해 유사한 특성을 가진 그룹으로 나누는 군집화 알고리즘이 이에 해당합니다.

### 1.2.3 준지도 학습 (Semi-Supervised Learning)

- **준지도 학습**은 지도 학습과 비지도 학습의 중간에 위치한 방법입니다.
    - 이 방법에서는 일부 데이터에만 정답(label)이 주어지고, 나머지 데이터는 정답 없이 제공됩니다.
    - 정답이 없는 데이터를 포함하여 모델을 학습시킴으로써 모델의 성능을 향상시키는 것이 목적입니다.
    - 준지도 학습은 소량의 레이블이 있는 데이터와 다량의 레이블 없는 데이터를 혼합하여 사용합니다.

### 1.2.4 강화 학습 (Reinforcement Learning)

- **강화 학습**은 에이전트가 환경과 상호작용하며 얻는 보상을 기반으로 학습하는 방법입니다.
    - 에이전트는 주어진 환경에서 행동을 취하고, 그에 따른 보상을 받아 다음 행동을 결정합니다.
    - 반복적인 시행착오를 통해 보상을 최대화하는 방향으로 학습을 진행합니다.
    - **정책(policy)**은 환경에서 최적의 행동을 선택하기 위한 기준을 의미하며, 강화 학습의 목적은 이 최적의 정책을 학습하는 것입니다.
    - 강화 학습은 **평균 기대 보상(expected average reward)**을 최대화하는 방향으로 동작합니다.
- 강화 학습의 주요 특징:
    - **순차적 의사 결정(long-term sequential decision making)**: 현재의 결정이 미래의 결과에 미치는 영향을 고려합니다.
    - **자동화된 의사 결정**과 **게임 플레이** 등에 자주 활용됩니다.
- 적용 예시: 게임, 로봇 제어, 자율 주행, 물류 시스템 최적화 등

## 1.3 지도 학습의 원리

- **지도 학습**은 입력 데이터와 그에 대응하는 정답이 주어진 상태에서 학습하는 방식입니다.
    - 이 방법은 **입력(feature vector)**과 **정답(label)**을 포함한 데이터셋으로 모델을 훈련시킵니다.
    - 입력 데이터는 텍스트 메시지, 이미지, 센서 데이터 등 다양한 형태를 취할 수 있습니다.
- **라벨링(Labeling)**은 각 입력 데이터에 정답을 할당하는 과정입니다.
    - 예를 들어, 이메일 스팸 필터링에서는 "스팸"과 "정상"이라는 두 가지 범주로 라벨을 지정합니다.
    - 각 이메일은 0(정상) 또는 1(스팸)으로 분류되며, 이를 통해 모델이 정답을 학습합니다.
- 지도 학습으로 훈련된 모델은 새로운 데이터를 분류하거나 예측하는 데 활용됩니다.
    - 이때 중요한 점은 모델이 학습 데이터셋에 맞춰 **예측 성능을 최적화**해야 한다는 것입니다.
    - 이 과정에서 사용되는 대표적인 알고리즘으로 **SVM(Support Vector Machine)**이 있습니다.
    - **SVM**은 **고차원 공간에서 데이터를 분리**하여 새로운 데이터를 효과적으로 예측합니다.

## SVM에서의 결정 경계와 수식

- **결정 경계(decision boundary)**는 주어진 데이터에서 각 클래스 간의 경계를 정의하는 선 혹은 면입니다.
    - 이 경계는 두 클래스의 데이터를 가장 효과적으로 분리하며, SVM에서는 이 경계 간의 거리를 최대화하는 방향으로 학습이 진행됩니다.

## wx - b = 0 수식

- **wx - b = 0**은 **결정 경계**를 나타내는 수식입니다.
    - 여기서 w는 가중치 벡터, x는 입력 벡터, b는 편향(bias)입니다.
    - 이 수식을 통해 특정 벡터 x가 어느 클래스에 속하는지 판단할 수 있습니다.

## 결정 경계에서의 예측

- 주어진 입력 벡터 x에 대해 예측된 레이블 y는 다음과 같이 정의됩니다:
  - y = sign(wx - b)는 결정 경계를 기준으로 데이터 포인트가 어느 쪽에 속하는지를 나타냅니다.
    - $wx - b \geq 0$인 경우, 데이터는 한 클래스(예: +1)에 속합니다.
    - $wx - b < 0$인 경우, 데이터는 다른 클래스(예: -1)에 속합니다.
  - **sign 함수**는 이 값이 양수인지 음수 인지를 판별하여 +1 또는 -1 값을 반환합니다.

## 초평면과 거리의 최적화

- **초평면(hyperplane)**은 두 클래스 간의 가장 가까운 경계를 정의하는 평면입니다.
  - SVM의 목표는 이 초평면을 두 클래스 사이에서 최대한 멀리 위치시키는 것입니다.
  - **초평면의 최적화**는 두 클래스의 데이터 포인트와 초평면 사이의 거리를 최대화하는 방식으로 이루어집니다.

## 최적화 문제

- **최적화 문제**는 다음과 같이 수식으로 표현됩니다:

  $\min_w \frac{1}{2}||w||^2$
  - 이 식은 초평면을 구성하는 가중치 벡터 w의 크기를 최소화하여, 두 클래스 간의 마진을 최대화하는 것을 의미합니다.
  - 이때 ||w||는 벡터 w의 크기(노름)를 나타내며, 벡터의 길이를 최소화하는 방향으로 최적화를 수행합니다.

## 모델 일반화 및 최적화

- 일반화(generalization)는 학습된 모델이 새로운 데이터에 대해 얼마나 잘 예측할 수 있는지를 의미합니다.
  - SVM은 데이터를 효과적으로 분류하면서도, 과적합(overfitting)을 방지하고 새로운 데이터에도 잘 적용되도록 초평면을 설정합니다.
  - **마진**을 최대화하여 모델의 성능을 높이는 것이 SVM의 핵심 목표입니다.

## SVM과 결정 경계

- SVM(Support Vector Machine)은 주어진 데이터 포인트를 두 개의 클래스로 나누기 위해 **결정 경계**를 설정하는 기법입니다.

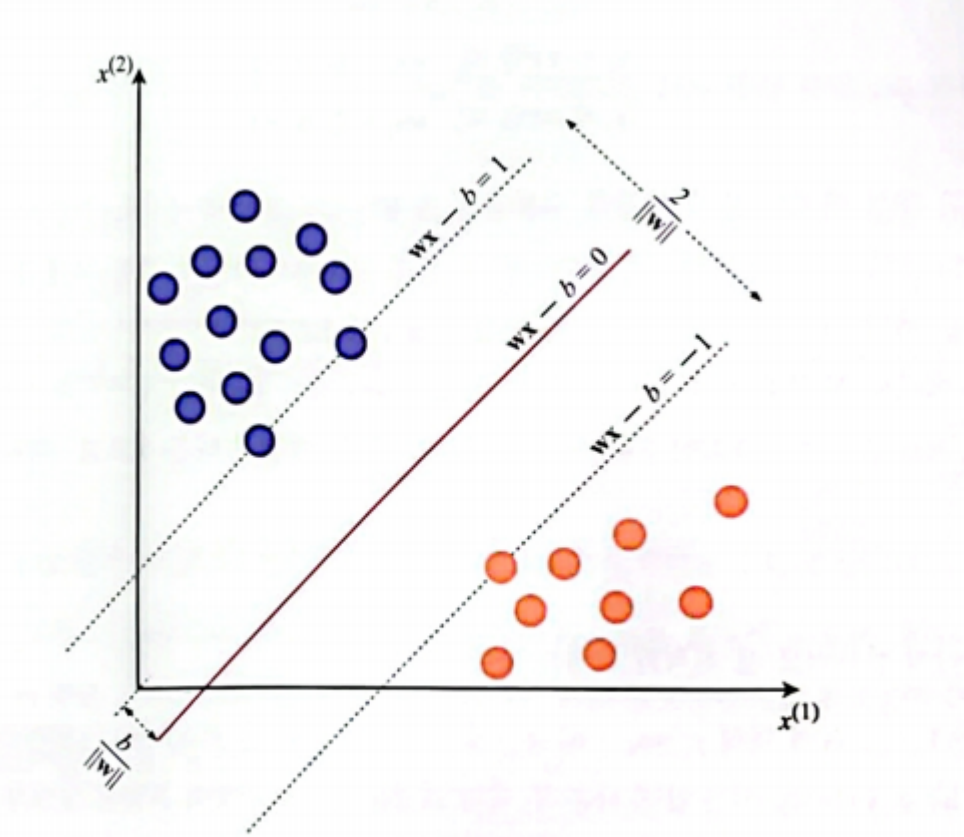


그림 1.1 2차원 특징 벡터에 대한 SVM 모델의 예

- 그림 1.1은 2차원 데이터를 시각화한 예시입니다. 여기서 두 클래스는 파란색과 주황색으로 구분되어 있습니다.
- 결정 경계$(wx - b = 0)$는 이 두 클래스를 분리하는 선으로 그려지며, 이 선을 기준으로 데이터를 분류하게 됩니다.

## 마진과 최적화

- **마진(margin)**이란 두 클래스 사이의 가장 가까운 데이터 포인트와 결정 경계 사이의 거리를 말합니다.
  - SVM의 목표는 이 마진을 최대화하는 것입니다.
  - **마진을 최대화하는 방식**은 두 클래스 간의 경계가 가장 멀리 떨어진 데이터를 기준으로 결정 경계를 설정하는 것입니다.
  - 즉, 마진이 최대가 되도록 $wx - b = 1$과 $wx - b = -1$을 설정하여, 두 클래스 간의 구분을 명확히 합니다.

## 수학적 표현

- **wx - b = 1**과 **wx - b = -1**은 각각 두 클래스 간의 가장 가까운 데이터 포인트를 기준으로 설정된 선입니다.
    - 이때 마진은 두 선 사이의 거리로 계산되며, 마진의 크기는 $\frac{2}{||w||}$로 나타냅니다.
    - 따라서 SVM은 ||w||를 최소화하여 마진을 최대화하는 방향으로 학습을 진행합니다.

## SVM의 일반화

- SVM 알고리즘은 결정 경계를 기준으로 새로운 데이터 포인트를 분류합니다.
    - 즉, 주어진 데이터로 학습한 후, 새로운 데이터가 입력되면 결정 경계를 기준으로 해당 데이터의 클래스를 예측합니다.

## 1.4 훈련 데이터로 만든 모델의 효율성

- SVM 모델이 훈련 데이터와 새로운 데이터 모두에 대해 뛰어난 성능을 보이는 이유는 다음과 같습니다:
    - SVM은 결정 경계를 최대한 일반화(generalization)하여 새로운 데이터에도 잘 적용될 수 있도록 설계되었습니다.
    - **결정 경계의 최적화**를 통해 두 클래스의 분류 성능을 극대화함으로써, 새로운 데이터에 대해서도 높은 예측 정확도를 유지합니다.

## 예측 성능과 예제의 선택

- 학습 과정에서 사용된 각 예제는 서로 독립적으로 선택됩니다.
    - 새로운 예제가 학습된 데이터와 근접할수록 높은 예측 정확도를 가질 가능성이 큽니다.
    - 반면, 새로운 예제가 기존 학습 데이터와 멀리 떨어질수록 예측 오류 가능성이 증가합니다.
- 음성 오류(negative error)와 양성 오류(positive error)의 가능성도 고려해야 합니다.
    - 새로운 예제가 잘못 분류될 확률이 존재하며, 특히 학습된 모델의 결정 경계에 가까운 예제일수록 이러한 오류 발생 가능성이 높아집니다.

## 학습에서의 일반화와 성능

- 직관적으로, 학습된 모델은 훈련 예제와 유사한 새로운 데이터를 잘 예측할 수 있습니다.
    - 훈련에 사용된 예제와 가까운 새로운 데이터는 높은 정확도로 예측되지만, 멀리 떨어진 데이터는 오류 발생 확률이 높아집니다.

- SVM의 목표는 마진을 최대화하여 새로운 예제에 대한 예측 확률을 극대화하는 것입니다.
    - 결정 경계(decision boundary)가 두 클래스의 데이터를 최대한 멀리 분리하므로, 새로운 데이터의 클래스를 더 쉽게 예측할 수 있습니다.

## 학습 가능성(learnability)과 PAC 이론

- 학습 가능성(learnability)은 모델이 새로운 데이터를 얼마나 잘 예측할 수 있는지를 설명하는 개념입니다.
    - PAC 이론(Probably Approximately Correct learning)은 모델이 특정 조건에서 좋은 성능을 낼 수 있는지에 대한 수학적 방법론을 제공합니다.
    - 이 이론은 주어진 조건에서 학습 알고리즘이 얼마나 정확한 분류기를 제공할 수 있는지 분석합니다.