



머신러닝의 네 가지 분류

전에 배운 머신러닝 문제, 이진 분류, 다중 분류, 스칼라 회귀. 모두 **지도 학습**의 예이다.

지도 학습의 목표는 훈련 데이터의 입력과 타겟 사이에 있는 관계를 학습하는 것이다.

지도 학습은 빙산의 일각이다. 전체 머신 러닝은 복잡한 하위 분류를 가진 방대한 분야이다. 일반적으로 머신 러닝 알고리즘은 다음 절에서 소개하는 4개의 범주 안에 속한다.

지도 학습

지도 학습이 가장 흔한 경우이다. 샘플 데이터(사람이 레이블링한)가 주어지면 알고 있는 타겟(꼬리표)에 입력 데이터를 매핑하는 방법을 학습한다. 앞서 본 예제도 지도학습의 고전적인 예이다.

문자 판독, 음성 인식, 이미지 분류, 언어 번역 같은 딥러닝의 거의 모든 애플리케이션이 일반적으로 이 범주에 속한다. 지도학습은 대부분 분류와 회귀로 구성되지만 다음과 같은 경우도 있다.

- **시퀀스 생성** : 사진이 주어지면 이를 설명하는 캡션을 생성한다. 시퀀스 생성은 이따금 일련의 분류 문제로 재구성할 수 있다.
- **구문 트리 예측** : 문장이 주어지면 분해된 구문 트리를 예측한다.
- **물체 감지** : 사진이 주어지면 사진 안의 특정 물체 주위에 경계 상자를 그린다. 이는 분류 문제로 표현되거나, 경계 상자의 좌표를 벡터 회귀로 예측하는 회귀와 분류가 결합된 문제로 표현할 수 있다.
- **이미지 분할** : 사진이 주어졌을 때 픽셀 단위로 특정 물체에 마스킹한다.

비지도 학습

어떤 타겟도 사용하지 않고 입력 데이터에 대한 흥미로운 변환을 찾는다. 데이터 시각화, 데이터 압축, 데이터의 노이즈 제거 또는 데이터에 있는 상관관계를 더 잘 이해하기 위해 사용한다.

비지도 학습은 데이터 분석에서 빼놓을 수 없는 요소이며, 종종 지도 학습 문제를 풀기 전에 데이터셋을 잘 이해하기 위해 필수적으로 거치는 단계이다. **차원 축소**와 **군집**이 비지도 학습에서 잘 알려진 범주이다.

자기 지도 학습

자기 지도 학습은 지도 학습의 특별한 경우이며, 별도의 범주로 할 만큼 충분히 다르다. 자기 지도 학습은 지도 학습이지만, 사람이 만든 레이블을 사용하지 않는다. 즉 **학습 과정에 사람이 개입하지 않는 지도 학습**이다.

레이블이 여전히 필요하지만 보통 경험적인 알고리즘을 사용해서 입력 데이터로부터 생성한다.

예를 들어 오토인코더가 잘 알려진 자기 지도 학습의 예이다. 여기에서 생성된 타겟은 수정하지 않은 원본 입력이다. 지난 프레임이 주어졌을 때 다음 프레임을 예측하는 것이다.

지도 학습, 자기 지도 학습, 비지도 학습의 구분은 모호하다. 명확한 경계가 없고, 연속적이다.

강화 학습

강화학습에서 에이전트는 환경에 대한 정보를 받아 보상을 최대화하는 행동을 선택하도록 학습된다.

예를 들어 강화학습으로 훈련된 신경망은 비디오 게임 화면을 입력으로 받고 게임 점수를 최대화하기 위한 게임 내의 행동을 출력할 수 있다.

분류와 회귀에서 사용하는 언어

- **샘플 또는 입력** : 모델 주입될 하나의 데이터 포인트
- **예측 또는 출력** : 모델로부터 나오는 값
- **타겟** : 정답. 외부 데이터 소스에 근거하여 모델이 완벽하게 예측해야 하는 값
- **예측 오차 또는 손실 값** : 모델의 예측과 타겟 사이의 거리를 측정한 값
- **클래스** : 분류 문제에서 선택할 수 있는 가능한 레이블의 집합. 예를 들어 고양이와 강아지 사진을 분류할 때 클래스는 '고양이'와 '강아지' 2개이다.

- **레이블** : 분류 문제에서 클래스 할당의 구체적인 사례. 예를 들어 사진 #1234에 '강아지' 클래스가 들어 있다고 표시한다면 '강아지'는 사진 #1234의 레이블이 된다.
- **참 값 또는 꼬리표** : 데이터셋에 대한 모든 타깃. 일반적으로 사람에 의해 수집된다.
- **이진 분류** : 각 입력 샘플이 2개의 배타적인 범주로 구분되는 분류 작업
- **다중 분류** : 각 입력 샘플이 2개 이상의 범주로 구분되는 분류 작업. 예를 들어 손글씨 숫자 분류를 말한다.
- **다중 레이블 분류** : 각 입력 샘플이 여러 개의 레이블에 할당될 수 있는 분류 작업. 예를 들어 하나의 이미지에 고양이와 강아지가 모두 들어 있을 때는 '고양이' 레이블과 '강아지' 레이블을 모두 할당해야 한다. 보통 이미지마다 레이블의 개수는 다르다.
- **스칼라 회귀** : 타깃이 연속적인 스칼라 값인 작업. 주택 가격 예측이 좋은 예이다. 각기 다른 타깃 가격이 연속적인 공간을 형성한다.
- **벡터 회귀** : 타깃이 연속적인 값의 집합인 작업. 예를 들어 연속적인 값으로 이루어진 벡터이다. 여러 개의 값에 대한 회귀를 한다면 벡터 회귀이다.
- **미니 배치 또는 배치** : 모델에 의해 동시에 처리되는 소량의 샘플 묶음. 샘플 개수는 GPU의 메모리 할당이 용이하도록 2의 거듭제곱으로 하는 경우가 많다. 훈련할 때 미니 배치마다 한 번씩 모델의 가중치에 적용할 경사 하강법 업데이트 값을 계산한다.