5. 지도 학습과 비지도 학습

1) 지도 학습

지도 학습이라고 하니 뭔가 어색하죠? 바로 단어의 뜻이 와 닿지 않아서 일 것입니다. 그렇다면 단어의 뜻부터 알아보겠습니다.

‘지도’는 어떤 목적이나 방향으로 남을 가르쳐 이끈다는 의미입니다. 즉, 지도 학습은 인공지능을 누군가가 직접 가르치고 이끄는 학습 방법입니다. 그렇다면 누가 어떻게 인공지능을 가르치는 것일까요?

인공지능은 처음부터 강아지와 고양이를 구별할 수 있을까요? 그렇지 않습니다. 처음 상태의 인공지능은 마치 어린아이와 같아서 아무것도 알지 못합니다. 그렇기 때문에 이러한 인공지능에게 “이것이 사과니? 배니?” 하고 물어도 정확히 대답할 수 없습니다.

인공지능이 사과를 학습할 때 사과 사진을 보여 주면서 “이것은 사과야.”라고 말해주고 배를 학습할 때도 똑같이 “이것은 배야.”라고 말해줍니다. 이렇게 인공지능이 어떤 사진으로 학습할 때 학습하는 하나하나에 대해 감독자의 입장에서 학습시키는 것을 바로 지도 학습이라고 합니다.

그렇다면 이런 감독은 사람이 하는 것일까요? 아닙니다. 지도 학습은 머신러닝의 학습 방법 중 하나이기 때문에 인공지능이 데이터를 보고 스스로 학습합니다. 여기서 감독자는 데이터 그 자체입니다. 지도 학습은 데이터 중에서도 정답이 있는 데이터를 이용하여 학습합니다.

이 점이 다른 학습 방법과의 가장 큰 차이점입니다.

그렇다면 데이터인데 정답이 있는 데이터란 무엇을 의미하는 것일까요? 데이터의 모습을 살펴보면 그 안에 정답이 들어있습니다. 바로 레이블이라는 이름으로 말이죠.

이때 사과 사진과 배 사진은 서로 다른 수많은 종류의 사진이지만, 사과와 배라는 ‘이름’은 일정합니다. 이 이름을 전문 용어로 ‘레이블’이라고 합니다. 그리고 이 레이블이 바로 데이터의 정답을 의미합니다.

이제 이러한 형태의 데이터를 인공지능에 입력하여 학습시키면 인공지능은 여러 장의 사과 사진을 보며 사과만의 특성을 찾아내고, 여러 장의 배 사진을 보며 배만의 특성을 찾아냅니다. 이렇게 수많은 데이터를 사용하여 학습한 인공지능에게 새로운 사진을 보여주며 “이것은 사과야? 배야?”라고 물으면 인공지능은 학습한 내용을 바탕으로 답을 말할 수 있습니다.

지도 학습 사례 살펴보기

그림 처럼 특징을 나타내는 데이터로 ‘키와 ‘체중을 선택했고, 정답을 나타내는 데이터로 ‘성별(남성/여성)’을 선택했다고 생각해 봅시다.

(그림)

이렇게 선택한 데이터 조합을 컴퓨터로 학습시켜 키와 체중으로 성별을 예측하는 모델을 만듭니다. 그리고 새로운 ‘키’와 ‘체중’데이터를 모델에 넣어 ‘성별’을 예측해 봅니다.

성별 같은 분류를 예측하는 문제를 분류(Classification)문제라고 합니다. 그림의 예는 성별의 남자와 여자로 분류하므로 이진 분류(Binary Classification)라고 합니다. 분류 문제는 이진 분류뿐만 아니라 10개의 속성으로 분류하는 예 등 다중 클래스 분류(Multi-Class Classification)도 있습니다. 이처럼 정답을 나타내는 데이터에 해당하는 변수가 연속적인 값이 아닌 일정 범위 안에서 선택한 값(Discrete Values)이면 분류 문제로 해결합니다.

참고로 , 특징을 나타내는 데이터는 특징(Feature)혹은 독립 변수(Independent Variable)로 나타냅니다. 정답을 나타내는 데이터를 레이블(Label) 혹은 종속 변수(Dependent Variable)로 나타냅니다.

분류 문제의 대표적인 예로는 스팸 메일 필터 구현이 있습니다. 사용자가