5. 지도 학습과 비지도 학습

1) 지도 학습

지도 학습이라고 하니 뭔가 어색하죠? 바로 단어의 뜻이 와 닿지 않아서 그럴 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 포도니? 오렌지니?” 하고 물어도 정확히 대답할 수 없습니다.

인공지능이 포도를 학습할 때 포도 사진을 보여 주면서 “이것은 포도야.”라고 말해주고 오렌지를 학습할 때도 똑같이 “이것은 오렌지야.”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 포도 사진과 오렌지 사진은 서로 다른 수많은 종류의 사진이지만, 포도와 오렌지라는 ‘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 포도 사진을 보며 포도만의 특성을 찾아내고, 여러 장의 오렌지 사진을 보며 오렌지만의 특성을 찾아냅니다. 이렇게 수많은 데이터를 사용하여 학습한 인공지능에게 새로운 사진을 보여주며 “이것은 포도야? 오렌지야?”라고 물으면 인공지능은 학습한 내용을 바탕으로 답을 말할 수 있습니다.

지도 학습 사례 살펴보기

Diagram

Description automatically generated

(지도학습으로 포도, 오렌지 구분)

위 그림처럼 지도학습은 미리 오렌지 사진을 보여주며 ‘이건 오렌지야!’라고 정답을 알려주고, 포도 사진을 보여주며 ‘이건 포도야!’라고 정답을 알려주게 되면 머신러닝은 각 포도와 오렌지 사진을 분석을 통해 보라색이고 삼각형 모양이면 포도라고 인식하고 주황색이고 원형 모양이면 ‘이건 오렌지야’라고 인식할 수 있습니다.

하지만 포도나 오렌지 사진을 각각 한 장씩 넣고 인식을 하려고 하면 인식률이 현저하게 떨어지게 됩니다. 그 이유는 만약 포도의 사진이 사람이 들고 있는 포도 사진을 데이터를 주게 되면 컴퓨터는 포도가 사람인지 포도인지 인식을 하지 못해 이상한 알고리즘을 설정하기 때문이죠. 그래서 인공지능에는 빅데이터가 필요합니다. 인공지능에 대해 얘기할 때 데이터는 인공지능의 연료 역할을 하기 때문에 데이터가 많으면 많을수록 좋습니다.

그럼 데이터만 많으면 될까요? 라는 의문을 가질 수 있습니다. 정답은 아닙니다. 정확한 포도사진이 아닌 키위 사진을 넣고 포도라고 컴퓨터에 넣게 되면 컴퓨터는 또 다른 알고리즘을 설정하기 때문에 정확한 데이터가 많이 필요합니다. 그래서 요즘에는 데이터분석전문가라는 직업이 생기기도 하였습니다.

2) 비지도 학습

비지도 학습은 지도 학습이 아닌 학습을 의미합니다.

지도 학습은 정답이 있는 데이터를 사용하여 인공지능을 학습시켰지만, 비지도 학습은 정답이 있는 데이터가 아닌, 정답이 없는 데이터를 사용합니다. 그렇다면 비지도 학습에서의 인공지능은 어떻게 학습하는 것일까요?

인공지능이 포도와 오렌지를 학습한다고 생각해 봅시다. 여기서 달라지는 것은 바로 무엇이 포도 사진이며, 무엇이 오렌지 사진인지 정답을 알려주지 않는다는 것입니다.

인공지능이 무엇이 포도인지, 오렌지인지 알 수 없기 때문에 포도와 오렌지를 정확하게 구분하라고 할 수는 없습니다. 그 대신 인공지능에게 수많은 사진을 보여주고 그것을 2개로 나누어 보라고 하면 인공지능은 여러 사진을 비교하며 스스로 형태를 나눕니다. 대상의 특징을 살펴보고 스스로 그 특징에 따라 구분해 나가는 것입니다. 이때 인공지능은 높은 확률로 포도는 포도대로, 오렌지는 오렌지대로 구분해 나갑니다. 하지만 정답이 없기 때문에 무엇이 포도인지, 무엇이 오렌지인지는 결코 알지 못하죠.

만약 포도와 오렌지 사진으로 2개가 아니라 3~4개로 구분하라고 한다면 어떻게 될까요? 아마 포도와 오렌지의 품종별로 구분될 수도 있겠죠? 이처럼 정답이 없는 데이터에서 그 데이터의 특징을 찾아서 스스로 구분해 나갑니다. 이렇게 구분해 나가면 새로운 데이터가 들어왔을 때 그 데이터가 어떤 그룹에 속하는지 스스로 판단할 수 있습니다. 정답이 없는 데이터를 사용해서 스스로 판단할 수 있는 지능을 가지게 되는 것이죠.

.A picture containing diagram

Description automatically generated

(데이터를 보고 스스로 구분하는 비지도 학습)

이러한 비지도 학습은 지도 학습을 보완해 주는 중요한 역할을 합니다. 지도 학습이 가능한 상황은 모범 답안, 즉 데이터의 정답(레이블)이 있을 때입니다. 만약 미지의 상황, 즉 모범 답안이 없는 상황에서는 지도 학습을 사용할 수 없습니다. 그리고 현실의 여러 문제를 분류하는 문제에서는 답이 없는 경우가 더 많기 때문에 이러한 상황에서 비지도 학습은 아주 유용하게 사용됩니다.

예를 들어 사진에서 사람의 얼굴을 판별하여 사람별로 사진을 정리해 주는 인공지능을 살펴봅시다. 이 인공지능은 처음에는 각 사람의 얼굴 특징을 바탕으로 사람들을 구별해 나갑니다. 왜냐하면 누가 누구인지 정답이 없기 때문입니다. 이럴 때는 비지도 학습 방식을 사용합니다. 누가 누구인지 지도해 주는 사람이 없어도 얼굴 특징만을 바탕으로 구별하는 것이죠. 즉, 지도 학습이 정답이 있는 데이터로 학습한다면, 비지도 학습은 정답이 없는 데이터로 학습하는 방식입니다.

비지도 학습 사례 살펴보기

Diagram

Description automatically generated

(비지도학습으로 포도, 오렌지 구분)

지도 학습은 데이터와 정답을 미리 컴퓨터에 알려주면서 인식을 하게 했습니다. 하지만 비지도 학습은 데이터만 컴퓨터에게 넣어주고 분류를 할 수 있습니다.

위에 그림처럼 포도와 오렌지 사진을 컴퓨터에게 주게 되면 컴퓨터는 각 사진의 특징을 뽑아 분류할 수 있게 됩니다. 이것을 군집화(Clustering)라고 합니다.

A picture containing tableware, dishware

Description automatically generated

(비지도 학습 군집화)

포도와 오렌지의 사진을 넣게 되면 컴퓨터는 군집분류를 하여 그림처럼 만들게 됩니다.

파란색 원 근처에 있으면 포도일 사진일 확률이 높아지고 노란색 근처에 있으면 오렌지일 확률이 더 커지게 됩니다.

비지도 학습도 지도 학습처럼 정확한 데이터가 많이 필요로 합니다.