3. 머신러닝 소개

1) 데이터의 중요성

머신러닝은 데이터의 규칙을 바탕으로 정답을 냅니다. 데이터가 없으면 머신러닝을 수행할 수 없습니다. 따라서 머신러닝의 학습에는 어느정도 정리된 데이터가 필요합니다.

최근에는 정보화의 영향으로 엄청난 양의 데이터가 쏟아져 빅 데이터(Big Data) 기술이 크게 발전했습니다. 더불어 쏟아지는 빅 데이터를 이용한 딥러닝 기술도 크게 발전하고 있지요.



2)머신러닝(Machine Learning)은 무엇일까요?

머신러닝은 데이터를 사용하여 인공지능을 만들 수 있습니다. 마치 요리사가 밀가루 반죽을 오븐에 넣어 빵을 만들듯이 데이터라는 재료를 머신러닝을 사용하여 하나의 인공지능으로 만들 수 있습니다.

물론 인공지능을 만드는 방법에는 머신러닝 뿐만 아니라 여러 방법이 있습니다. 하지만 오늘날 인공지능을 만들 때 가장 많이 사용하는 방법이 바로 머신러닝을 사용하는 방법입니다.

인공지능은 기계가 스스로 생각할 수 있게끔 만들어서 사람이 맞닥뜨리는 여러 문제를 해결하도록 돕는 역할을 합니다. 이러한 문제 역시 굳이 인공지능이 아닌 여러 프로그램을 사용하여 해결할 수 있습니다. 전통적인 프로그래밍 방식으로 말이죠.

우리가 사용하는 자동문을 생각해 봅시다. 사람 혹은 어떤 물체가 다가가면 자동으로 문이 열립니다. 그 이유는 문을 언제 여는지. 어떤 속도로 열고 언제 닫는지에 대한 명령을 프로그래머가 자동문 프로그램에 작성해서, 이를 자동문 시스템에 넣어두었기 때문입니다.

하지만 머신러닝의 방법은 조금 다릅니다. 머신러닝을 직역하면 ‘기계가 공부하는 것’입니다. 즉, 사람이 직접 프로그램을 만들어서 기계에 넣어주는 것이 아니라, 문을 언제 어떻게 열 것인지 등 기계가 스스로 문제를 해결하는 방법을 학습하는 것입니다.

그러면 머신러닝은 무엇을 가지고 학습할까요?

바로 데이터입니다. 수많은 데이터, 즉 빅데이터 속에서 스스로 규칙을 찾고, 그 규칙을 학습하는 것입니다. 다시 말해, 머신러닝은 데이터를 사용해서 기계가 스스로 학습하는 방식이라고 볼 수 있습니다.

전통적인 프로그래밍에서는 어떤 결과를 내기 위해서 프로그램을 만들어야 했습니다. 그런 다음 그 프로그램에 어떤 값을 주면, 정해진 프로그램에 따라 그 결과값이 나옵니다. 예를 들어 2를 곱하는 프로그램이 있습니다. 이 프로그램은 특정한 숫자인 10을 넣으면, 20이라는 결과값을 출력합니다. 이처럼 전통적인 프로그래밍에서는 데이터를 넣으면 그 데이터를 어떠한 식으로 처리하는지 프로그래머가 프로그램을 작성했습니다.

하지만 머신러닝의 경우 데이터로 인해 나타나는 결과를 사용하여 학습하기 때문에 스스로 프로그램을 만들 수 있습니다. 전통적인 방식에서 프로그램을 사람이 만들었다면, 머신러닝에서는 그 프로그램을 기계가 스스로 만들 수 있습니다. 특정한 데이터와 함께 그 데이터로 인해 나타나는 결과를 같이 넣어주면 기계가 그 관계를 찾아냅니다. 기계가 스스로 관계를 찾아내므로 예전처럼 사람이 직접 프로그래밍할 필요가 없어졌습니다. 이와 같이 기계가 스스로 프로그램을 만드는 것, 이것을 가리켜 ‘기계가 학습한다’고 보는 것이지요. 그리고 그 학습 재료가 바로 데이터이므로 오늘날 데이터의 중요성 또한 상당히 강조되고 있습니다. 그래서 데이터를 다른 말로 ‘인공지능 시대의 석유’라고도 하지요.

머신러닝은 통계학과 깊은 관련이 있습니다. 통계학에서 유래된 머신러닝 알고리즘이 많으며 통계학과 컴퓨터 과학 분야가 상호 작용하면서 발전하고 있습니다. 대표적인 오픈소스 통계 소프트웨어인 R에는 다양한 머신러닝 알고리즘이 구현되어 있습니다.

하지만 최근 머신러닝의 발전은 통계나 수학 이론보다 경험을 바탕으로 발전하는 경우도 많습니다. 컴퓨터 과학 분야가 이런 발전을 주도하고 있습니다. 컴퓨터 과학 분야의 대표적인 머신러닝 라이브러리는 사이킷런(scikit-learn)입니다.



3) 머신러닝의 유형

머신러닝에는 몇가지 유형이 있습니다. 입력 데이터가 무엇인지에 따라 크게 다음의 세가지로 나눕니다.

* 지도 학습(Supervised Learning) : 문제의 정답을 컴퓨터에 입력해 머신러닝 모델을 학습시키는 것입니다. ‘특징을 나타내는 데이터’와 ‘정답(목적)을 나타내는 데이터’를 미리 선택한 후 학습시킨다고 가정합니다.
* 비지도 학습(Unsupervised Learning) : 지도 학습은 특징과 종속 변수(정답)를 조합해 학습하는 방법입니다. 하지만 비지도 학습은 정답을 나타내는 종속 변수가 없는 방법입니다.
* 강화 학습(Reinforcement Learning) : 어떤 환경 안에서 정의한 에이전트가 최고의 보상을 얻는 쪽으로 행동하도록 학습하는 방법입니다.

위와 같이 간단하게 머신러닝의 유형을 알아보았습니다. 뒷부분에 더 자세히 다뤄보도록 하겠습니다.