인공지능 정리(1)

인공지능이란

인공지능은 사람처럼 학습하고 추론할 수 있는 지능을 가진 컴퓨터 시스템을 만드는 기술입니다. 인공지능의 역사는 약 80년 남짓 되었지만 인류는 훨씬 더 오래전부터 지능적인 시스템을 생각했습니다. 지능을 가진 로봇을 다룬 최초의 소설은 150년 전으로 거슬러 올라갑니다.

흔히 영화 속의 인공지능은 인공일반지능(Artificail General Intelligence) 혹은 강인공지능(Strong AI)이라고 부르는 인공지능입니다. 영화 <터미네이터>의 스카이넷 처럼 사람과 구분하기 어려운 지능을 가진 컴퓨터 시스템이 인공일반지능입니다. 반면 현실에서 우리가 마주하고 있는 인공지능은 약인공지능은(Week AI)입니다. 약인공지능은 아직까지는 특정 분야에서 사람의 일을 도와주는 보조 역할만 가능합니다. 예를 들면 음성비서, 자율 주행 자동차, 음악추천, 기계 번역 등입니다. 또 이세돌과 바둑 시합을 한 알파고가 좋은 예 입니다.

그럼 언제 인공일반지능에 도달할 수 있을까요? 아직은 아무도 그 시기를 정확히 알 수 없지만 가능성에 대해서는 대체로 긍정적입니다. 그렇지만 가까운 미래에 이런 기술이 도래하리라 낙관하지는 않았으면 좋겠습니다.

머신러닝이란

머신러닝은 규칙을 일일히 프로그래밍하지 않아도 자동으로 데이터에서 규칙을 학습하는 알고리즘을 연구하는 분야입니다. 인공지능의 하위 분야중에서 지능을 구현하기 위한 소프트웨어를 담당하는 핵심 분야입니다.

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.

사이킷런 라이브러리는 파이썬 API를 사용하는데 파이썬 언어는 배우기 쉽고 컴파일 하지 않아도 되기 때문에 사용하기 편리합니다. 머신러닝 분야에 관심이 높아지면서 파이썬과 함께 사이킷런 라이브러리가 큰 인기를 얻고 있습니다. 이제는 사이킷런 외에 대표적인 다른 머신러닝 라이브러리를 찾아보기 힘듭니다.

파이썬으로 print(“Hello World”)명령을 출력할 수 있다면 누구나 조금만 배워 머신러닝 프로그램을 만들 수 있습니다. 사이킷런 라이브러리에서 지공하는 클래스와 함수를 사용하여 필요한 작업을 수행할 수 있습니다.

머신러닝소개

머신러닝

머신러닝(기계학습)은 여러분에게 주어진 문제, 과제, 환경에 따라 컴퓨터 스스로 문제를 해결하려고 학습한 예측 결과를 활용하는 것입니다. 문제 해결 자체는 물론 문제 해결을 하는 구조전체를 뜻합니다.

(사진)

머신러닝의 유형

- 지도학습

- 비지도학습

- 강화학습

데이터의 중요성

머신러닝의 학습에는 어느 정도 정리한 데이터가 필요합니다. 데이터를 바탕으로 정답을 내는 교칙을 학습해서 예측과 추론 등을 실행합니다. 데이터가 없으면 머신러닝을 수행할 수 없습니다. 즉, 데이터 수집이 머신러닝의 첫 번째 작업입니다.

딥러닝이란

많은 머신러닝 알고리즘 중에 인공신경망(Artificial Neural Network)을 기반으로 한 방법들을 통칭하여 딥러닝(Deep Learning)이라고 부릅니다. 종종 사람들은 인공 신경망과 딥러닝을 크게 구분하지 않고 사용합니다.

국내에서는 2016년 이세돌과 알파고의 대국으로 인해 딥러닝에 대한 관심이 크게 높아졌습니다. 하지만 앞에서 소개했듯이 딥러닝은 2010년 초반부터 이렇게 새로운 혁명을 준비하고 있었습니다.

(바둑사진)

복잡한 알고리즘을 훈련할 수 있는 풍부한 데이터와 컴퓨터 성능의 향상, 그리고 혁신적인 알고리즘 개발입니다. 인공지능에 대한 과거의 시도와 달리 최근의 딥러닝 발전은 매우 긍정적이고 지속 가능해 보입니다.

이런 오픈소스 머신러닝 라이브러리의 영향력을 눈치챘던 것 일까요. 2015년 구글은 딥러닝 라이브러리인 텐서플로(TensorFlow)를 오픈소스로 공개했습니다. 텐서플로는 공개와 동시에 큰 인기를 얻었으며 아직까지 가장 널리 사용되는 딥러닝 라이브러리 입니다. 페이스북도 2018년 파이토치(PYTorch)딥러닝 라이브러리를 오픈소스로 발표했습니다. 이 라이브러리들의 공통점은 인공 신경망 알고리즘을 전문으로 다루고 있다는 것과 모두 사용하기 쉬운 파이썬 API를 제동한다는 점입니다.

(텐서플로 로고) (파이토치 로고)

코랩으로 파이선 환경 구축하기

머신러닝을 학습하려면 몇가지 준비물이 필요합니다. 보통은 특정 언어와 그 언어를 사용할 환경을 구축하는데, 이런 방식은 각자 가지고 있는 실습환경이 달라서 독자들이 실습에 어려움을 표하곤 합니다. 이 책은 사전 지식 없이도 누구나 쉽게 머신러닝의 기초를 학습할 수 있게 하겠다는 점을 목표로 하는 터라 ‘누구나 동일한 결과를 표현할 수 있게 쉽게 실습’할 수 있는 점을 고려했습니다.

이 책에서 사용한 실습 환경은 네트워크에 연결된 컴퓨터와 구글 계정입니다.

(구글 코랩 로고)

구글코랩

구글 코랩은 웹브라우저에서 무료로 파이썬 프로그램을 테스트하고 저장할 수 있는 서비스 입니다. 심지어 머신러닝 프로그램도 만들 수 있습니다. 좀 더 있어 보이는 말로 표현하자면 클라우드 기반의 주피터 노트북 개발 환경입니다. 머신러닝은 컴퓨터 사양이 중요한데, 구글 코랩을 사용하면 컴퓨터 성능과 상관없이 프로그램을 실습해 볼 수 있습니다.

구글 계정만 있다면 누구나 무료로 코랩을 사용할 수 있습니다.

구글 계정이 있다면 바로 코랩(http://colab.research .google.com)에 접속해보자. 다음과 같은 화면을 볼수 있습니다!

Graphical user interface, text, application

Description automatically generated

코랩은 웹 브라우저에서 텍스트와 프로그램 코드를 자유롭게 작성할 수 잇는 온라인 에디터라고 생각하면 쉽습니다. 이런 코랩 파일을 노트북(Notebook)혹은 코랩 노트북이라고 부릅니다.

텍스트 셀

셀(cell)은 코랩에서 실행할 수 있느 최소 단위입니다. 즉 셀 안에 있는 내용을 한 번에 실행하고 그 결과를 노트북에 나타냅니다. 하지만 텍스트 셀은 코드처럼 실행되는 것이 아니기 때문에 자유롭게 사용해도 괜찮습니다. 셀 하나에 아주 긴 글을 써도 되고 여러 셀에 나누어 작성해도 괜찮습니다.

Graphical user interface, text, application, email

Description automatically generated

코드 셀

코드 셀로 이동하면 코드와 결과가 함께 선택됩니다.

Graphical user interface

Description automatically generated with low confidence

지도학습

머신러닝의 유형

- 지도학습

- 비지도학습

- 강화학습

지도학습

지도학습(supervised learning)은 문제의 정답을 컴퓨터에 입력해 머신러닝 모델을 학습시키는 것입니다. ‘특징을 나타내는 데이터’와 ‘정답(목적)을 나타내는 데이터’를 미리 선택한 후 학습시킨다고 가정합니다.

예를들어 그림 1-3처럼 특징을 나타내는 데이터로 ‘키’와 ‘체중’을 선택했고, 정답을 나타내는 데이터로 ‘성별(남성/여성)’을 선택했다고 생각해 봅시다.

(그림1-3)

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

성별 같은 분류를 예측하는 문제를 분류(classification) 문제라고 합니다. 그림1-3의 예는 성별을 남자와 여자로 분류하므로 이진분류(binary classification)라고 합니다. 분류 문제는 이진 분류뿐만 아니라 10개의 속성으로 분류하는 예 등 다중 클래스분류(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 직접스팸 메일로 판단한 데이터를 종속 변수나 레이블, 이메일을 보내는 사람이나 스팸으로 판단하는 문장을 독립 변수나 특징으로 선택합니다. 레이블로 선택한 데이터가 많을수록 머신러닝을 수행했을 때 정확도가 더 높은 결과를 얻습니다.

지도 학습의 다른 방법으로 회귀regression 문제가 있습니다. 회귀문제는 숫자 값의 크고 작음에 의미를 부여해 예측하는 것입니다.

예를들어 그림 1-4처럼 특징을 나타내는 데이터로 ‘성별’, ‘키’를 선택하고, 정답을 나타내는 데이터로 ‘신발 크기’를 선택했다고 생각해 봅시다.

(그림1-4)

분류문제는 ‘남자’를 뜻하는 값을 0, ‘여자를 뜻하는 값을 1로 정하는 방법 등으로 레이블을 숫자 값으로 표현할 수 있습니다. 이때 1과 0이라는 숫자의 크고 작음은 의미가 없습니다.

반면, 신발 크기인 ‘265mm’나 ‘240mm’는 숫자 값의 크고 작음에 의미가 있습니다. 이러한 수치를 예측하는 것이 회귀 문제입니다. 회귀 문제는 종속 변수가 연속적인 값이므로 ‘237mm’ 처럼 실제 만드는 신발 크기가 아닌 값도 예측할 수 있습니다.

지도학습 (분류)의 예

이번에는 지도 학습의 분류 문제를 구현하는 과정을 살펴보겠습니다. 먼저, 예제가 무엇인지 소개 한 후 구현 과정을 순서대로 살펴볼 것입니다.

예제 살펴보기

지금부터 소개할 예제는 미국 위스콘신 대학교에서 공개한 유방암 진단 데이터세트 입니다. 데이터 세트에는 특징이 30개 있으며, 종속 변수는 ‘양성’ 과 ‘악성’으로 구분합니다. 데이터는 총 569개가 있으며, 악성(M)으로 구분하는 데이터가 212개, 양성(B)으로 구분하는 데이터가 357개 입니다. 즉, 30개의 특징이 있는 569개 데이터로 악성인지 양성인지 예측합니다.

(1-6표)

사이킷런에는 유방암 진단 데이터세트가 포함되어 있습니다. 다음 샘플코드를 실행해 가져옵니다.

샘플코드

Text

Description automatically generated with low confidence

사이킷런에 포함된 유방암 진단 데이터세트를 사용할 수 있는 load\_breast\_cancer 함수를 불러 온 후 해당 함수를 실행한 결과를 변수 data에 저장합니다.

이어서 데이터세트에서 특징을 해당하는 data.data를 변수 X에 저장하고 종속 변수에 해당하는 data.target을 변수 y에 저장합니다.

x = data.data

y = data.target

X는 여러 개 특징을 나타내는 열 벡터 각각을 행렬로 만들어 다룹니다. 행렬은 수학에서 관례상 대문자로 표기하므로 변수 이름을 대문자로 정했습니다. y는 종속 변수를 나타내는 열 벡터입니다. 악성은 0으로 나타내고, 양성은 1로 나타냅니다.

x의 요소 수는 569 X 30개 이므로 569행 30열로 이루어진 행렬이라고 할 수 있습니다. 그리고 y는 열 벡터이고 569행 1열로 이루어진 행렬이라고 본다면 x와 y의 각 행은 대응합니다.

구현 방법

이제 유방암 진단 데이터세트를 이진 분류하는 학습 모델을 만들겠습니다. 분류에 사용할 알고리즘은 로지스틱 회귀(logistic regression)입니다. 알고리즘 이름에 ‘회귀’가 있지만 분류 문제입니다.

샘플코드

Text

Description automatically generated with medium confidence

로지스틱 회귀 모델을 사용하려고 사이킷런의 LogisticRegression 클래스를 불러왔습니다. 그리고 LogisticRegression 클래스의 인스턴스를 만들고 초기화한 모델을 변수 model에 저장합니다.

샘플코드



model 변수(LogisticRegression 클래스의 인스턴스)애 fit 메서드를 실행해 학습을 합니다. 전달인자로 특징 X와 종속 변수 y를 설정했습니다. fir 메서드는 학습한 후 model 변수에 학습 결과를 저장합니다. 출력 격과는 LogisticRegression 클래스로 학습할 때 설정한 주요 파라미터와 기본값을 보여줍니다. 각 파라미터가 무엇을 뜻하는지 궁금한 분은 사이킷런의 LogisticRegression 클래스 문서를 참고하기 바랍니다.

샘플코드



학습한 모델이 담긴 mode에 predict 메서드를 실행해 예측을 합니다. 학습할 때 사용한 특징 X로 예측한 후 결과를 y\_pred 변수에 저장합니다.