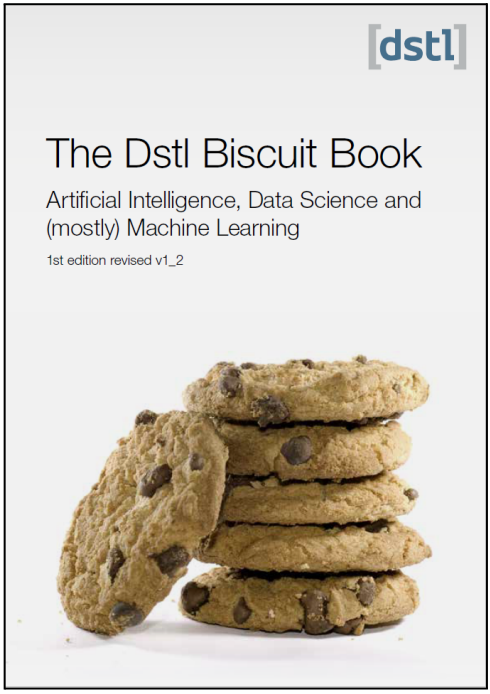
소개: 2019년 영국 국방과학기술연구소(Defense Science and Technology Laboratory, dstl)는 AI, 데이터 과학, 기계학습을 쉽게 설명하는 비스킷북을 발간하였습니다. 영국 정부는 인공지능 공공조달지침을 마련하면서 영국 공무원들에게 이 비스킷북을 일독할 것을 권장하고 있습니다.
정확하면서도 비교적 쉽게 인공지능의 개념을 설명한 이 책자를 번역하여 소개합니다.
번역: 정보인권연구소(초벌번역은 기계번역의 도움을 받았습니다.)

* 원문:
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/850129/The_Dstl_Biscuit_Book_WEB.pdf

Dstl 비스킷 북 (2019)

AI, 데이터 과학, (무엇보다) 기계학습이란

초판 개정 v1_2



머리말

최근 몇 년 동안 인공지능(AI), 데이터 과학, 기계 학습이 급속도로 부상하였고, 이는 여러 영역에서 기존의 컴퓨팅 접근 방식이 제공했던 기능을 넘어서는 혁신을 가져왔습니다. 이 기술들은 기계로 하여금 역대 최고치로 생성되고 있는 방대한 데이터에 대해 새로운 통찰력을 부여할 수 있게 하였고, 이전에는 인간만이 수행할 수 있었던 정교한 작업을 수행할 수 있게 하였습니다. 이러한 일이 가능해진 것은 데이터 생산 규모의 확대와 컴퓨팅 능력 향상이 결합되었기 때문입니다. 특히 기계 학습 방법의 개발을 통해 이런 발전이 가능해졌습니다. 따라서 이 가이드는 AI와 데이터 과학을 다루지만 대부분 기계 학습에 관한 것입니다.

이러한 혁명의 여파로 인해 용어 혼란과 과장 광고도 잇따릅니다. 이 짧은 가이드는 가장 보편적인 용어에 대한 설명을 제공하고 사실과 허구를 구분하는 데 도움을 주기 위한 것입니다.

국방과학기술연구소**(Dstl)**

Dstl은 영국의 국방, 안보, 번영을 위해 강력한 과학 기술을 제공합니다.

세계 최고 수준의 저희 기관 AI, 데이터 과학, 기계 학습 작업에는 업계, 학계, 동맹국 전문가들과 협력하는 업무가 포함됩니다. 저희는 기계가 인간과 상호 작용하는 방식을 살펴보는 초창기 연구부터 데이터 과학을 실제 문제 및 운영 요구 사항에 적용하는 것까지 모든 것을 검토하고 있습니다. 이것은 국방과 안보의 미래일 뿐만 아니라 우리가 살고 있는 세상의 미래에 대한 일입니다.

저희 역할에 따라 국방과 안보 상황에서 AI, 데이터 과학, 기계 학습의 관여와 사용에 대한 명시적 지침을 제공할 수 있습니다.

닥터 메르세데스 토레스 토레스[1], 글렌 하트[2], 토니 에머리[2]

1 노팅엄 대학교
2 DSTL

소개

이 가이드는 <비스킷 북>이라고 부릅니다. 여러분은 차와 비스킷과 함께 이 책자를 집어 들어 즐길 수 있습니다. 비스킷 북은 주제별로 쉽게 소화할 수 있는 덩어리들을 연속 배열하였고, 지나치게 기술적이지 않으면서 필수적인 정보를 제공하는 방식으로 구성되었습니다.

비스킷 북이 여러분에게 유익하고 쉽게 이해되길 바랍니다. 그렇다고 이 책자를 차에 담그지는 마세요!

정의

결론적으로 AI, 데이터 과학, 기계 학습이란 무엇입니까?

AI와 데이터 과학에 대해 보편적으로 승인된 정의는 없습니다. 반면 기계 학습은 일반적으로 더 잘 정의될 수 있습니다.

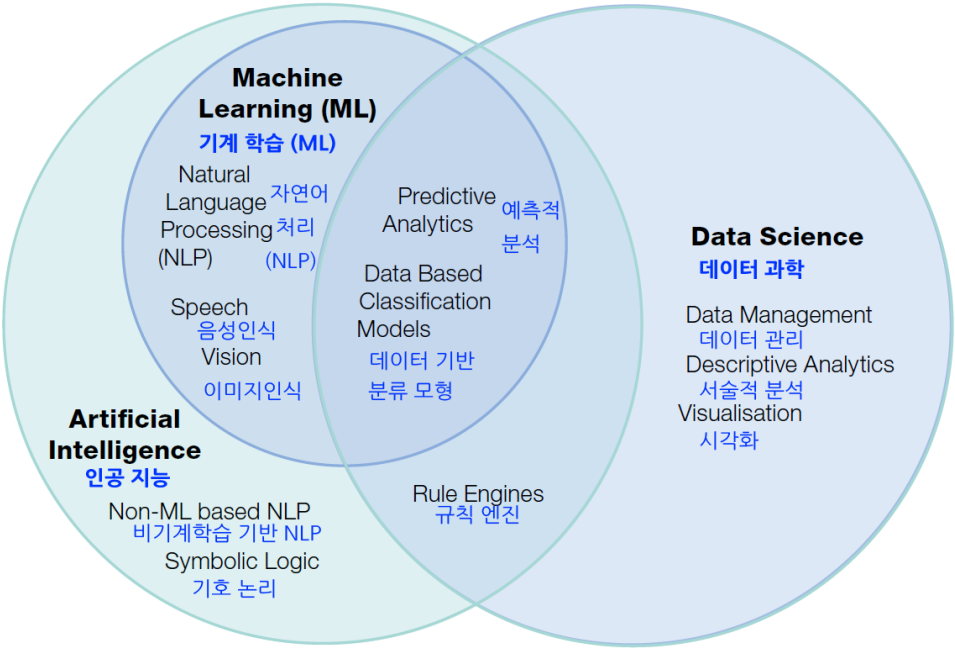
산업적, 상업적, 비전문적 환경에서는 세 가지 용어가 상호 교차적으로 사용되는 일이 드물지 않습니다. 많은 제품 설명서나 회사와 매체들에서 이러한 용어들이 매우 느슨한 방식으로 사용됩니다.

최근 한 연구에서 AI를 사용한다고 주장하는 2,800개 이상의 유럽 스타트업을 분석한 결과, 실제로는 40%만이 AI를 사용하는 것으로 나타났습니다. 이는 AI, 데이터 과학, 기계 학습을 둘러싼 다양한 정의가 있고, 무엇이 이들이고 아닌지에 대해 수많은 논쟁 가능성이 있다는 사실을 나타냅니다.

현재의 개념 혼란을 감안해 보았을 때 약간 명확해질 필요가 있습니다. 이에 상호 개념이 어떻게 다른지 이해하기 위해 다음과 같은 간단한 정의를 제시합니다.		
AI 컴퓨터 시스템이 일반적으로 인간 또는 생물학적 지능이 필요한 작업을 수행할 수 있도록 개발된 이론과 기법 (뒤에서 살펴 보다시피 이 지능은 매우 제한적임)	데이터 과학 통계학, 수학, 컴퓨터 과학 및 분야별 전문성을 융합하여 데이터에서 관련 통찰력 또는 지식을 추출하는 다학제 분야	기계 학습 명시적인 프로그래밍 없이도 컴퓨터 시스템이 자동으로 학습하고 개선할 수 있는 기능을 제공하려는 목표를 가진 분야

결국 AI는 작업의 지능적 수행에 대한 것이고, 데이터 과학은 데이터에서 통찰력을 구하는 것이고, 기계 학습은 자동화된 프로세스를 통해 두 가지 모두를 달성하는 수단입니다. 용어가 어떻게 혼동될 수 있는지 간략히 살펴봅시다. 여러분이 'AI'을 요청하면 '기계 학습'이 도착할 수 있습니다. 왜냐하면 기계 학습은 AI에 도달하기 위해 적용되는 방법이기 때문입니다.

아래 다이어그램은 서로 관련된 세 가지 주제와 이에 적용할 수 있는 몇 가지 애플리케이션 및 기술을 보여줍니다.



예시

다음은 여러분이 매일 사용하고 있을 수 있고, AI, 데이터 과학, 기계 학습에 의존하는 일상 생활의 몇 가지 예시입니다.

검색 엔진:

구글, 빙, 여타의 검색 엔진은 정교한 기계 학습 방법을 사용하여 여러분의 검색 기준과 일치하는 웹 페이지를 찾고 순위를 매깁니다. 이 엔진은 기계 학습을 사용하여 여러분에게 관련된 결과를 제공할뿐 아니라, 데이터 과학과 기계 학습을 결합하여 여러분이 무언가를 검색할 때마다 이면의 알고리즘으로 여러분의 반응을 모니터링합니다. 어떤 페이지를 여러분이 열었는지, 얼마나 많이 열었는지, 각각 얼마나 오래 머물렀는지 등 말입니다. 검색 엔진들은 이러한 방식으로 검색 결과를 여러분에게 맞춤할 수 있습니다.

가상 개인 비서:

알렉사, 시리 또는 구글홈을 사용해 보셨나요? 이들 가상 개인 비서는 모두 데이터 과학을 적용하여 간단한 질문에 답하고, 뉴스나 날씨를 알려주고, 음악이나 팟캐스트를 재생하는 등 임무를 수행합니다. 이를 위해 이들은 여러분이 말하는 내용과 언제, 어디서, 어떻게 말하는지에 대한 정보를 수집합니다. 그후에 가상 개인 비서들은 이 정보를 사용하여 여러분의 선호에 맞춘 결과를 생성합니다. 이들은 또한 기계 학습을 사용하여 다음 업무도 수행합니다. 1) 여러분을 이해합니다(음성 처리 및 이해). 2) 여러분과 상호 작용한 결과에 기반하여 성능을 개선합니다. 3) 여러분과 다시 소통합니다(대화 관리).

교통 정보:

교통 또는 지도 앱이 어떻게 여러분에게 통근 중 차가 밀리는 지역을 알려줄 수 있는 건지 궁금한 적이 있었나요? 이들 앱이 사용자의 GPS 위치와 속도를 교통 관리용 중앙 서버에 계속 입력하기 때문입니다. 그후에 데이터 과학 방법론을 사용하여 현재의 교통 지도를 작성하고 교통 밀도를 추정합니다. GPS 정보를 사용할 수 없는 지역의 경우 과거 데이터를 사용해서 기계 학습으로 교통량이 많은 지역을 예측할 수 있습니다.

대출 승인:

은행과 기타 금융 기관은 대출을 신청하는 고객에 대하여 광범위한 정보를 수집합니다. 데이터 과학이 관련 데이터를 찾는 데 사용되는 한편으로, 기계 학습이 고객의 이력 및 고객과 유사한 프로필을 가진 사람들의 이력에 따라 고객의 대출 가능 여부를 분류하는 데 사용됩니다.

활동 추적:

핏빗과 같은 신체 활동 추적기는 사용자에 대한 방대한 정보를 수집합니다. 수집된 데이터에는 걸음 수, 오른 층계수, 칼로리 소모량, 수면 단계, 분당 심박수 등이 포함됩니다. 그후에 사용자가 허용하는 경우 데이터 과학을 사용하여 외부 파트너(예: 의료 전문가 및 보험 회사)와 공유할 수 있는 건강 통계를 생성하여, 여러분에게 더 맞춤한 서비스를 제공하기도 합니다.

챗봇(온라인 고객 지원):

점점 더 많은 웹사이트에서 고객 지원 서비스에 채팅을 사용하고 있지만, 그 채팅 상대는 사람이 아닌 챗봇인 경우가 많습니다. 이케아, 호텔스닷컴, E.ON과 같은 기업은 봇을 사용하여 연락이 필요한 고객을 필터링합니다. 챗봇은 기계 학습을 사용하여 텍스트에서 관련 정보를 식별하고 여러분의 질의에 부합하는 답변을 제공합니다. 봇으로 고객이 필요로 하는 정보를 제공할 수 없는 경우 인간 담당자에게 전달됩니다. 언어 학습을 위한 신규앱 Duolingo는 챗봇을 사용하여 사용자가 문자 메시지로 새로운 언어 기술을 연습하도록 돕습니다. 이 앱은 사용자별로 가장 잘 맞는 챗봇을 할당한다는 구상 속에, 사용자에 대한 정보를 수집할 때 데이터 과학을 사용하고 사용자의 성격과 학습 스타일을 분류할 때 기계 학습을 적용합니다.

추천 시스템:

아마존에서 여러분의 흥미를 끄는 제품에 대한 이메일을 받은 적이 있습니까? 또는 넷플릭스에서 "여러분을 위한 추천" 섹션을 본 적이 있나요? 이 두 가지는 추천 시스템의 예시들입니다. 추천 시스템은 사이트 내 여러분의 활동에서 데이터를 수집하고 사전 처리합니다(즉, 여러분이 검색한 내용, 여러분이 시청한 내용, 얼마나 오래 시청했는지, 무엇을 보관하거나 장바구니에 넣었는지, 영화의 어떤 부분을 되감거나 빨리 감았는지 등). 이를 통해 여러분의 행동을 사이트의 나머지 사용자와 비교하고 추천 항목을 생성합니다. 데이터 과학을 사용하면 행동에 따라 사용자를 그룹화하고 각 그룹 별로 권장 사항을 공유할 수 있습니다. 따라서 여러분과 비슷한 행동을 하는 사람들이 여러분이 보지 않은 영화를 본 적이 있다면 넷플릭스는 여러분에게 이 영화를 추천할 것입니다.

물론 직종 별로 다음과 같은 다양한 애플리케이션들이 사용됩니다.

- 분류: 예를 들어 이미지를 차량, 사람 등이 포함된 것으로 분류합니다.
- 인식: 이 분야 애플리케이션으로는 얼굴 인식이 일반적입니다.
- 필터링: 많은 양의 이미지, 비디오 또는 문서를 가져와서 그중 특정 이미지, 개체 또는 참조사항이 포함된 것들을 골라 냅니다.
- 이상 감지: 예를 들어 대량의 엔진 성능 데이터를 분석하여 고장일 수 있는 이상신호를 식별합니다.
- 예측: 예를 들어 음식이 상할 수 있는 시기를 예측하도록 할 수 있습니다.

그 외에도 애플리케이션의 범위가 너무 빨리 증가하고 있으므로 이 목록은 계속해서 추가될 수 있습니다.

인공지능(Artificial Intelligence, AI)

AI는 크게 좁은(Narrow, 전용) AI과 일반(General, 범용) AI의 두 가지 범주로 나눌 수 있습니다.

'일반' AI 또는 '강' 인공지능은 기계가 (꿈꿀수 있다면!) 꿈꿀수만 있는 존재인 반면, '약' 인공지능이라고도 불리는 '좁은' AI는 이미 존재합니다.

좁은 AI

좁은 AI는 하나의 주요 작업을 수행하는 데 집중하는 AI입니다. 현재 존재하는 모든 AI 시스템은, 설령 실제보다 더 똑똑해 보일지라도, 좁은 AI를 장착한 것입니다. 이에 대한 예시로는 아마존의 알렉사(아마존사의 인공지능 스피커 - 역주)를 들 수 있습니다. 이들은 각자 수행할 수 있는 제한적이고 사전 정의된 작업 범위를 가지고 있습니다. 이들이 지능이나 자기 인식 능력을 가지고 있다는 환상을 우리가 가진다 하더라도 이들은 기실 지능이나 자기 인식 능력을 가지고 있지 않습니다.

일반 AI

일반 AI는 여러 기계들을 참조하여 많은 작업을 수행할 수 있고 인지적으로 자신이 무엇을 하고 있는지 알고 있으며, 스스로 배울 수 있고 조정할 수 있습니다.
일반 지능의 예시로는 영화 <2001: 스페이스 오디세이>의 할(HAL)이나 <스타워즈>의 알투디투(R2-D2), <스타 트렉>의 데이터를 들 수 있습니다. 이 모든 예시들이 SF물에서 유래했다는 사실로 보아 이 AI는 근미래에 사용할 수 있는 것이라기 보다 (최소한 누군가에게는) 장기적인 포부라는 점을 알 수 있습니다. SF물에서 일반 AI가 유행했기 때문에 우리는 종종 AI를 '인간과 같은' 지능체와 동일시합니다. 사실 SF물을 진실이라고 생각할 이유는 없으며, 기계 지능은 인간의 지능과 상당히 다르다는 징후가 매우 많습니다.

이 안내서 전체적으로 AI에 대해 계속되는 언급은 '좁은 AI'를 의미합니다.

AI 방법론

AI는 단일한 것이 아니라, 인간 또는 동물과 유사한 지능을 사용하여 행위를 한다는 일반 목표에 달성하기 위한 다양한 방법론의 모음입니다. 어떤 방법이 AI인지 아닌지조차 논쟁의 여지가 있습니다. 다음은 AI 제품 계열에 보통 포함되는 방법론을 요약한 것입니다.

기호 AI (Symbolic AI)

1990년대 후반까지 AI 분야를 지배했던 것은 현재 '기호 AI'라고 불리는 접근 방식이었습니다. 그 당시는 AI와 AI의 정의가 더 단순했던 시절이었습니다. 기호 AI는 일반적으로 1차 논리라고 하는 수학 분야를 적용하여 추론하는 것을 기반으로 합니다. 이 접근 방식은 조언과 지침을 제공할 때 규칙을 사용하는 전문가 시스템 등 특정 분야에서 성공적이었지만 여러 한계가 있었습니다. 큰 가능성을 보였지만 많은 실망도 안겨 주었습니다.

기호 AI는 간단한 문장을 사용하여 기본 지식을 제공합니다. 예를 들어 다음과 같습니다

***A*는 *B* 안에 있다.**
***B*는 *C* 안에 있다.**

로직에 기반하여 진술들에서 추론해낼 수 있는 프로그램인 Reasoner는 이로부터 다음을 추론할 수 있습니다.

***A*는 *C* 안에도 있다.**

기호 AI의 한 가지 문제는 이것이 엄격한 기준에 기반하여 구축되었기 때문에, 세계를 이해하는 복잡한 방식에 유연하게 대처할 수 있는 인간의 능력에 맞추는 것이 힘들다는 점입니다. 예를 들어, 새에 대해 생각해보라는 요청을 받으면 대부분의 사람들은 작고 깃털이 달린 무언가, 날고 있고 지지배배 소리를 내는 무언가를 생각할 것입니다. 이것이 모든 새에게 해당되는 것은 아니지만 사람들은 이것을 힘들어하지 않습니다. 그러나 이는 기호 AI에 있어 중대 문제입니다.

기호 AI는 결코 죽거나 사라진 것이 아니며 오늘날에도 여전히 많은 분야에서 사용되고 있습니다. 한 가지 흔한 용도는 온톨로지 개발입니다. 즉, 특정 주제 영역에 대한 형식화된 설명으로 기계가 해당 주제에 대한 데이터를 더 잘 이해할 수 있도록 하는 것입니다. 예를 들어, 육군 구조에 대한 온톨로지는 기계로 하여금 사단이 여단, 대대 등의 집단으로 구성되어 있음을 이해할 수 있도록 합니다.

일부에서 AI로 분류한 다른 기술로는 에이전트 기반 모델링 및 유전 알고리즘이 있습니다.

> 에이전트 기반 모델링은 "에이전트"를 생성하는 방법으로, 이는 개별적인 코드로서 일련의 규칙을 통해 다른 에이전트와 상호작용합니다. 이 접근 방식은 보통 복잡한 분야의 컴퓨터 시뮬레이션을 구축하는 데 사용됩니다.
>
> 유전 알고리즘은 자연 선택 과정을 시뮬레이션하는 진화적 접근 방식을 통해 문제를 해결하려는 접근 방식입니다.

현재의 최고점

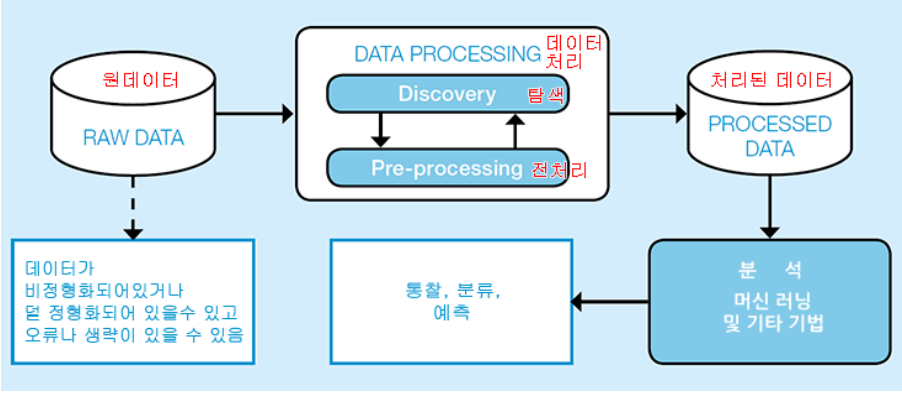
AI는 그간 다소간의 고점과 저점을 거쳐 왔고 현재 최고 수준을 향해 가고 있습니다. 이 최고점은 기계 학습 기법의 적용 덕분이며 이에 대해서는 이 비스킷 북의 뒤쪽에서 다룹니다.

데이터 과학

예전에 이 절의 제목은 '빅데이터' 또는 '데이터 마이닝'이었겠지만 시간과 이름은 계속 바뀌는 것입니다. 이들 용어들은 완전히 같은 수준이 아니지만 이제 데이터 과학이 효과적으로 이 모든 용어를 아우릅니다.

핵심을 말하자면, 데이터 과학은 데이터 준비 및 시각화 기술의 범주에 포함된 통계적 방법의 사용에 관한 것입니다. 데이터 과학 역시 정확히 무엇이고 아닌지에 대해 다소 논쟁의 여지가 있지만, AI 만큼은 정의되어 있습니다. 마찬가지로, 데이터 과학자는 한편으로 통계적 방법, 기계 학습, 시각화 기술의 전문가일 수 있고, 다른 한편으로는 더 나은 급여의 직업을 찾는 수학 또는 컴퓨터 과학 학위 소지자일 수도 있습니다(데이터 과학자로 직함을 변경하면 고용 가능성이 상당히 높아집니다).

데이터 과학을 이해할 수 있는 가장 좋은 방식은 데이터를 가져와 이로부터 어떤 통찰과 예측을 생성하는 프로세스라고 생각하는 것일 겁니다.
이 프로세스의 간단한 그림은 다음과 같습니다.



이 다이어그램이 단순하고 프로세스의 개념에 수많은 변형이 있다는 점을 짚긴 해야겠지만, 주요 요소는 분명히 드러나 있습니다. 기계 학습이 분석을 수행하는 유일한 방법이 아니며 기계 학습이 데이터 과학 외부에 존재할 수도 있다는 사실에 유의해야 합니다.

데이터

대부분의 데이터에 대해 적용될 수 있는 말은 '어렵다'는 것입니다.
여러분에게 데이터가 쉬운 상황이라면 운이 좋거나 착각한 것입니다.

데이터가 어려울 수 있는 흔한 이유 몇 가지는 다음과 같습니다.

- 상업적 사유, 법적 사유, 또는 기타 사유로 실제 확보하기 어렵고 때때로 덜 체계화되어 있음 - "이 데이터는 내 것이고, 당신은 쓸 수 없습니다."
- 문서화가 부족함
- 품질이 낮음(매우 흔한 상황)
- 필요한 정보가 부족함(데이터 양이 많은 경우일지라도)
- 윤리적으로 사용하기 어려움(예: 개인정보가 포함되어 있음)
- 다른 데이터와 결합하기 어려움

데이터 유형

데이터는 다양한 형태로 제공되며, 일부 사람들이 다른 사람들보다 더 운이 좋다면 이 이유 때문입니다. 주요 데이터는 정형 데이터, 텍스트 데이터, 디지털 신호 데이터, 이미지 데이터입니다.

정형 데이터**:**

대부분의 데이터베이스에 보관된 대부분의 데이터는 정형화되어 있습니다. 일반적으로 이러한 데이터는 표 또는 일련의 상호 연동된 표로 저장되고 표시됩니다. 데이터베이스에서 출력되는 일반적인 형식으로는 쉼표로 구분된 값(.csv), 엑셀 파일(.xlsx), XML 파일(.xml, 데이터에 계층이 있는 경우 특히 유용함) 등이 있습니다. 이러한 유형의 데이터가 가진 문제 중 하나는 특히 사람들에게 데이터를 입력하도록 하는 경우 품질이 매우 나빠질 수 있다는 점입니다. (데이터베이스에 데이터를 입력하는 사람들이 데이터베이스를 처음 설계했던 사람들보다 창의적이기 마련입니다. 이는 보통 데이터 품질에 좋지 않습니다.)

Day	Outlook	Temp.	Humidity	Wind	Play tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

텍스트 데이터**:**

텍스트에는 많은 정보가 포함되어 있으며(이것이 텍스트의 최종 목적이니까요), 컴퓨터는 자연어 처리(NLP)를 통해 이러한 정보에 점점 더 많이 접근할 수 있게 되었습니다.

이미지 데이터**:**

여기에는 이미지와 비디오가 포함됩니다. 어떤 의미에서 이는 고도로 정형화되어 있습니다. 그림은 픽셀 격자로 표현되기 때문입니다. 그러나 이미지 그 자체는 매우 비정형적이어서 기계가 해석하기 어렵습니다.

디지털 신호 데이터**:**

디지털 신호 데이터는 시간과 진폭이 불연속적인 값을 갖는 것입니다. 이 데이터는 물리적 신호를 샘플링하고 정량화함으로써 얻어집니다. 음성 데이터는 디지털 신호 처리의 특수한 사례로서, 이중 녹음된 소리에는 한 명 이상의 사람이 말하는 소리가 녹음되어 있습니다.

데이터 탐색

분석 및 시각화

일부 데이터를 실제 구할 수 있을 만큼 운이 좋았더라도, 가지고 있는 데이터가 포장에 표시된 내용과 정확히 일치하지 않을 수 있으므로 데이터를 파악해야 합니다.

이 탐색의 여정을 당신이 어떻게 수행할지는 분명 데이터의 특징에 크게 좌우됩니다. 여기에는 통계적 측정, 품질 평가는 물론, 상자그림(boxplot), 히스토그램, 산점도와 같은 시각화 기법이 포함될 수 있습니다. 일부 데이터의 경우 데이터 수집 경로나 이 경로로 인해 데이터가 편향될 수 있는지 등 데이터의 미묘한 상황을 파악하는 것이 중요할 수 있습니다. 예를 들어 이미지 데이터는 이미지화에 사용되는 센서의 특징에 따라 편향이 생길 수 있으며, 근적외선 센서는 초목을 강조 표시하는 데에는 매우 뛰어나고 이를 쉽게 감지하지만 다른 물체에 대해서는 덜 두드러지게 감지할 수 있습니다. 이러한 미묘한 차이들은 데이터가 실제로 나타내는 것에도 발생할 수 있습니다. 데이터가 차량을 기록하고 있다고 해도, 수집기의 차량에 대한 정의는 여러분이 생각하는 차량의 정의와 다를 수 있습니다. 이 전체 과정은 어느 정도 실망스럽기도 하지만 또 한편으로는 여러분이 데이터를 개선하기 위해 해야 할 일이 무엇인지 알려주기도 합니다.

데이터 랭글링

데이터가 낙관적인 기대치와 차이가 나는 것을 발견했다고 해서 모든 것을 잃은 것은 아닙니다. 문제를 해결하기 위해 할 수 있는 일이 있습니다. 설령 데이터의 품질이 양호하더라도 여러분이 보유하고 있는 분석 도구에 적합하도록 데이터를 처리하고 다른 데이터와 결합해야 할 수도 있습니다. 이러한 프로세스를 합성(conflation)이라고 합니다. 이들 전처리 단계는 데이터 랭글링(Data Wrangling)이라고 하구요.

이 프로세스를 지원하는 도구와 기술들이 여럿 있지만 각 데이터셋에 고유한 방식으로 접근해야 합니다. 이런 업무는 가내 수공업이라고 묘사되곤 하지만 장인이라는 표현이 더 적합할 수 있습니다. 다른 것은 몰라도 비용 문제를 감안해야 합니다.

랭글링이 데이터 과학 프로젝트 일정의 80%를 차지할 수 있다고들 말하지만 그렇게까지 드는 것은 아닙니다. 정확한 비율이 어떻든 대부분의 프로젝트에서 많이 소요되는 것은 사실입니다. 어떤 사람들은 이 프로세스가 비용이 많이 들고 시간이 많이 걸리며 실제 분석을 수행하는 재미있는 부분에 끼어들기 때문에 이 절차를 생략하고 싶은 유혹을 느낄 수 있습니다.

저희의 조언은 이 단계를 진지하게 받아들이라는 것입니다. 이 단계는 좋은 결과를 얻을 수 있는 유일한 방법입니다. 성공하고 싶지 않은 프로젝트에 대해서만 이 단계를 질러가세요. 또 이 단계는 탐색 단계와 독립적으로 취급되어서도 안 됩니다. 대부분의 경우 밀접하게 연관됩니다. 탐색은 문제에 주목하고, 문제를 해결한 후 또다른 문제를 발견하는 식으로 진행됩니다.

데이터 랭글링 방법

데이터 정제(Data cleaning):

데이터 정제는 중복되거나 누락된 데이터, 이상값 또는 잡음이 있는 데이터(임의의 값을 포함하는 데이터 – 라디오에서 들리는 잡음을 생각해 보십시오)를 처리합니다. 동일한 정보가 두 번 이상 발생하는 중복 데이터는 제거되어야 합니다. 더 큰 문제는 누락된 데이터입니다. 일부 항목이나 측정과 관련된 데이터가 너무 많이 누락되었을 경우 삭제하는 것이 가장 좋습니다. 누락된 데이터가 많지 않은 경우 누락된 값을 추론하거나 평균 또는 중앙값을 사용할 수 있습니다. 이는 완벽한 솔루션이 아니며 어떤 식으로든 결과에 영향을 미칠 수 있으므로 주의가 필요합니다.

이상값은 여러분의 데이터셋에서 다른 모든 항목의 저장값(instance)과 멀리 떨어져 있는 저장값입니다. 일반적으로 측정에 오류가 있을 때 발생합니다. 그러나 이상값들은 여러분의 데이터셋이 매우 비대칭적이라는(어떤 식으로든 편향되어 있다는) 사실을 나타낼 수도 있습니다. 이상값이 오류인 경우 저장값을 삭제할 수 있지만, 비대칭 데이터에 속하는 사례인 경우 이를 삭제하면 데이터의 일부를 무시하는 결과를 낳습니다. 신호 데이터와 같은 일부 데이터에는 잡음이 포함되어 잡음 감소 기술을 종종 적용해야 할 수 있습니다. 다른 종류로는 체계적으로 제거할 수 있는 불일치가 있을 수 있습니다. 예를 들어 어떤 데이터셋은 회사 IBM을 "IBM", "I.B.M." 또는 "International Business Machines"로 [서로 다르게] 지칭하였을 수 있습니다.

데이터 변환**(Data transformation)** 및 데이터 축소**(Data reduction)**: 통계 데이터에 적용할 수 있는 방법이며 모든 데이터가 비교 가능한 형식이 되도록 데이터를 변환하거나(이를 정규화 normalising 라고 함) 중복되거나 불필요한 정보를 제거하는 데 사용됩니다(축소).

합성**(Conflation):** 프로세스가 충분히 복잡하지 않다면 두 개 이상의 서로 다른 데이터셋의 구성 요소를 병합해야 하는 경우가 더러 있습니다. 이러한 프로세스를 합성이라고 합니다. 주의해야 할 부분은, 두 데이터셋에 있는 요소에 두 개의 서로 다른 식별자가 적용되었을 수 있으므로 속성 데이터를 사용하여 두 항목을 연결시켜야 한다는 점입니다. 이러한 프로세스는 잘못된 상관관계 및 누락된 상관관계로 오류를 유발할 수 있으므로 주의해야 합니다.

분석론

데이터를 획득, 탐색, 시각화, 랭글링한 후 마침내 분석을 수행할 때가 왔습니다. 이는 현재 기계 학습을 위한 시간이 왔다고 말하는 것과 거의 같은 의미가 되었지만, 기계 학습이 분석을 수행할 수 있는 유일한 방법인 것은 아닙니다.

다음 절에서 기계 학습을 살펴보기 전에, 기계 학습이 도입되기 전에는 분석론이 수행되었다는 점을 염두에 둘 필요가 있습니다. 고전적인 통계 분석은 여전히 건재하고 잘 작동하며, 지리 정보와 같은 특수 데이터는 지리 정보 시스템(GIS)을 사용하여 여전히 건전하게 분석됩니다. 이러한 방법들은 특히 데이터 양이 제한적일 때 기계 학습보다 이점을 가지고 있으므로 여전히 고려할 가치가 있습니다.

기계 학습

기계 학습의 목표는 명시적으로 프로그래밍하지 않고도 입력 데이터 및 분류, 또는 여러분이 수행하려는 작업 간의 관계를 자동으로 학습할 수 있는 시스템(즉, 알고리즘)을 만드는 것입니다.

대중적인 기계 학습 문제에는 개체 분류, 추적, 이미지 분할, 오디오 인식 또는 토지 피복 분류(Land Cover Classification) 등이 있으며, 자동 자막 생성, 텍스트 생성, 이미지 색상 지정 및 자연재해예측 등 기계 학습이 탐구되기 시작한 새롭고 흥미로운 문제들이 많이 있습니다. 여러분은 알렉사와 대화하거나 지문을 사용하여 스마트폰 잠금을 해제할 때 기계 학습을 만나게 됩니다.

기계 학습은 인간이 프로그래밍한 솔루션을 필요로 하는 고전적인 기호 AI와 다릅니다.

차이점을 이해하기 위해,
미로를 찾아가는 두 개의 로봇 쥐를 떠올려 보십시오.
첫 번째 로봇 쥐는 기호 AI를 사용하여 프로그래밍하고
두 번째 로봇 쥐에는 기계 학습 기능이 있다고 합시다.

첫 번째 로봇 쥐는 보상을 받고 미로를 찾는 방법을 알려주는 많은 규칙을 가지고 있습니다.
누군가[프로그래머]가 이 규칙을 만들어야 했습니다. 규칙은 꽤 잘 작동하겠지만, 새로운 유형의 장애물 등 프로그래머가 미처 생각하지 못한 어떤 상황을 만나면 멈출 수 있습니다.

두번째 로봇 쥐는 자체적으로 학습하며, 미로를 꽤 많이 시행하는 동안 계속 헤맬 것입니다. 그러나 미로를 충분히 시행하면 이 로봇 쥐가 자체적으로 규칙을 알아낼 것입니다. 비록 쥐의 로봇 두뇌에 규칙이 표시되는 방식이 인간인 우리가 규칙을 떠올리는 방식과 같지는 않겠지만 말입니다.

현재 이 분야에는 네 가지 주요 접근 방식이 있습니다.

- 비지도 학습
- 준지도 학습
- 지도 학습
- 강화 학습

비지도 학습

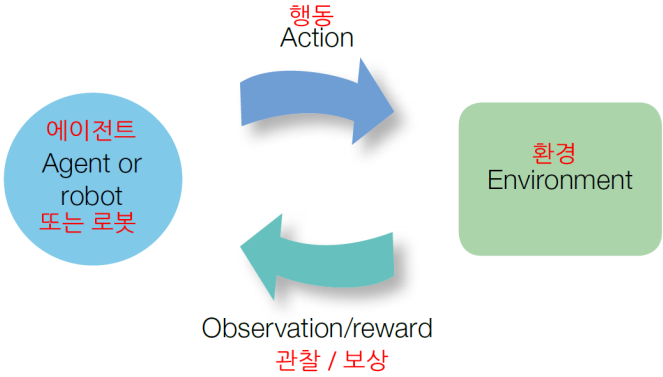
비지도 학습은 누군가가 라벨을 지정하거나 태그를 지정한 데이터에 의해 감독(또는 지도)되지 않는 학습 솔루션(모델이라고 합니다)으로 정의됩니다. 즉 기계에게 무엇을 학습해야 하는지 알려주지 않습니다. 나중에 라벨링에 대해 자세히 설명하겠습니다. 가장 대중적인 비지도 학습 제품 계열은 클러스터링 알고리즘입니다. 클러스터링은 어떤 식으로든 여러분의 데이터에서 서로 유사한 그룹을 찾는 프로세스입니다. 클러스터링 기법마다 유사한 속성의 여러 집합으로 정의된 서로 다른 그룹을 가지고 있습니다. 예를 들어 구글의 비지도 이미지 분류기는 고양이와 개를 구별하는 방법을 학습하였습니다. 비지도 학습을 통해 우리는 데이터에 숨겨져 있는 관계를 발견할 수 있습니다. 예를 들어, 어느 날의 특정 활동에서 다음 날 특정 범죄가 발생할 수 있는 징후를 나타내는 (명백하지 않을 수 있지만) 지역을 제시할 수 있습니다.

지도 학습

지도 학습은 라벨링 예제 또는 데이터 출력에 의해 학습이 감독되는 접근 방식군을 말합니다. 기계에는 우리가 관심이 있을 예시들 또는 학습 데이터(라벨링된 데이터)가 제공되며, 또한 우리가 관심이 없을 예시들도 제공됩니다. 그러면 모델은 학습 데이터의 입력을 일정한 출력으로 매핑하는 함수를 학습할 수 있습니다. 이미지들 중에서 고양이를 식별하도록 기계를 학습시킨다고 가정해 보겠습니다. 먼저 합리적으로 가능한 한 많은 이미지에서 고양이 이미지에 라벨링을 합니다. 이러한 이미지는 고양이가 포함되어 있지 않다고 기계에게 알려준 다른 많은 이미지들과 함께 기계를 학습시키는 데 사용됩니다. 지도 기법은 학습 과정에서 우리의 관심사가 무엇인지 또는 아닌지 기계에 명시적으로 알려주기 때문에 가장 정확한 결과를 얻습니다. 그러나 이 또한 여러분의 데이터셋에 라벨이 지정된 데이터가 수집되어야 하므로, 데이터 라벨링에 전문가 또는 특수 장비가 필요한 경우 문제가 될 수 있습니다. 이에 비해 비지도 학습을 위해 라벨이 지정되지 않은 데이터를 구하는 일은 상대적으로 쉽습니다.

강화 학습

또 다른 대안은 문제를 재구성하는 것입니다. 강화 학습에서는 입력 데이터를 라벨에 연결시키는 기능을 갖추기 위해 모델을 학습시키는 대신 더 작은 결정을 내리는 에이전트를 학습시킵니다. 이 에이전트는 각 결정에 대해 보상을 받게 됩니다. 그러면 에이전트가 누적 보상을 최대화하려는 목표를 갖게 됩니다. 강화 학습은 게임에 성공적으로 사용되고 있으며, 이때 게임은 "환경"과 그 "보상"이 되는 높은 점수(또는 [이용자가] 얼마나 잘하고 있는지를 알려주는 다른 방법)를 제공합니다.



분류 및 회귀

수행하려는 작업에 따라 여러분은 분류, 예측 또는 회귀 문제를 처리하게 됩니다. 분류의 목표는 각각의 클래스(자동차나 고양이 등 여러분이 관심이 있는 사물의 유형)를 식별하는 것입니다. 대중적인 분류 문제로는 객체 인식, 토지 이용 분류, 이미지 분할 등이 있습니다. 반면에 여러분이 온도, 연령, 위험 수준 등 연속적인 값을 예측하려는 경우에는 회귀 방법을 사용하게 됩니다.

회귀

회귀 또는 (보다 정확하게는) 회귀 분석은 변수 간의 관계를 추정하기 위한 일련의 통계 프로세스입니다. 회귀 분석은 다른 변수는 고정된 상태로 유지되고 어떤 변수가 변경이 될 때, 그 변수의 값이 어떻게 변하는지 이해하는 데 도움을 줍니다. 간단한 예를 들자면 변수가 X와 Y 두 개뿐일 때 X를 변경함에 따라 Y가 어떻게 변하는지 이해하고 싶을 수 있습니다. 여러분은 여러 오븐 온도(X)에 따라 저녁 식사가 얼마나 빨리 되는지(Y)를 계산할 수 있습니다. 변수가 3개 이상인 경우, 다른 변수를 고정된 값으로 유지한 채 하나의 변수만 변경하였을 때 또다른 변수가 어떻게 변하는지 확인할 수 있습니다.

데이터 및 기계 학습

학습 프로세스에는 일반적으로 학습, 검증, 테스트 단계를 둡니다. 각 단계는 서로 다른 목적을 가지고 있으며, 데이터의 서로 다른 하위 집합을 사용합니다.

1. 학습 단계:

학습 단계에서는 학습셋이라고 하는 데이터셋의 하위 집합을 사용하여 모델을 학습시킵니다. 학습셋이란 여러분의 기계 학습 알고리즘을 학습시키기 위해 제공하는 예제셋이라고 생각하십시오. 지도 학습에서는 기계가 라벨을 통해 학습할 수 있도록 데이터에 라벨을 지정해야 합니다. 이 라벨링에는 상당한 시간이 소요될 수 있으며, 라벨이 제대로 지정되어 있지 않으면 결과도 좋지 않기 때문에 라벨을 올바르게 지정하는 것이 중요합니다.

2. 검증 단계:

학습이 얼마나 잘 이루어졌는지 측정하기 위해서는, 여러분의 모델을 결과에 최적화된 추가 데이터에 노출시켜 볼 수 있습니다. 학습의 한 가지 문제점은 데이터가 "과적합"(데이터의 특정 측면에 너무 민감해짐)되어 부적절한 결과를 낳을 수 있다는 것입니다. 예를 들어 탱크가 포함된 이미지를 식별하려는 경우 학습 데이터는 모델을 전반적으로 녹색 사물에 민감한 상태로 만들 수 있습니다. 검증 데이터셋은 이를 교정하는 데 도움이 될 수 있습니다. 또한 충분한 데이터가 사용되어 더 이상 학습이 필요하지 않은 시기를 나타내는 데에도 도움이 될 수 있습니다.

3. 테스트 단계:

테스트 단계에서는 테스트셋, 즉 [모델이] 보지 못한 샘플셋을 제공하고 모델의 예측을 여러분이 예상하는 실제 출력과 비교하는 방식으로 모델을 테스트합니다. 상호 유사할수록 모델이 "데이터에 더 잘 맞추어졌다"는 의미입니다.

> 기억해야 할 한 가지 중요한 점은 데이터셋의 저장값을 두 단계 이상에서 사용하지 않는 것입니다. 그 경우 테스트 단계가 무효화될 수 있으며, 이는 마치 실제 시험에서 모의 고사와 똑같은 문제를 받는 경우와 비슷합니다. 여러분이 이해해야 할 또 다른 점은 이런 과정이 많은 데이터를 요구할 수 있다는 것입니다. 일반적으로는 데이터가 많을수록 좋지만, 반환값(return)이 급격히 감소하기 시작하는 순간이 올수 있습니다. 데이터가 많지 않은 경우 "로우샷 학습"이라고 하는 몇몇 기술을 사용할 수 있지만, 이 경우는 사실 기계 학습이 최선의 접근법이 아닐 수 있습니다.

기계 학습 방법

여러분이 기계 학습에 사용되는 수많은 방법론에 대한 아이디어를 얻으려면 다음과 같은 일부 목록을 참고하세요.

자동 인코더(AE)
볼츠만 머신(BM)
컨볼루션 신경망(CNN)
의사 결정 트리(DT)
디컨볼루션 네트워크(DN)
심층 신념 네트워크(DBN)
심층 컨볼루션 네트워크(DCN)
심층 컨볼루션 역전사 그래픽 네트워크(DCIGN)
딥 피드 포워드(DFF)
딥 Q-네트워크
딥 레지듀얼 네트워크(DRN)
노이즈 제거 자동 인코더(DAE)
에코 상태 네트워크(ESN)
익스트림 러닝 머신(ELM)
피드 포워드(FF)
게이트 순환 유닛(GRU)
생성적 적대 신경망(GAN)
홉필드 네트워크(HN)
코호넨 네트워크(KN)

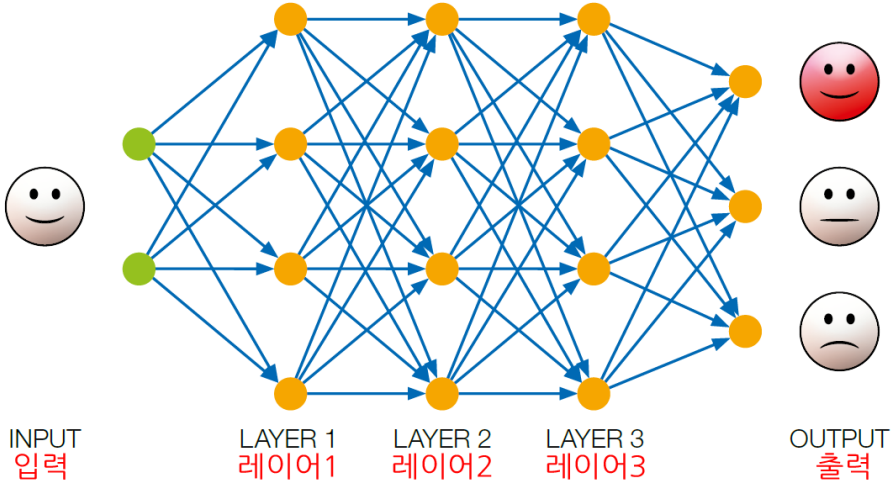
액체 상태 기계 (LSM)
장단기 기억망(LSTM)
마르코프 체인(MC)
신경 튜링 기계(NTM)
퍼셉트론
방사형 기본 네트워크(RBF)
랜덤 포레스트(RF)
순환 신경망(RNN)
제한적 볼츠만 기계(RBM)
스파스 자동 인코더(SAE)
지원 벡터 기계(SVM)
가변 자동 인코더(VAE)

… 등등

당연하게도 이 목록은 계속 늘어나고 있습니다. 이 항목들은 단지 방법론일 뿐입니다. 상업용이든 오픈 소스이든 이 방법론들을 사용하는 도구의 수는 훨씬 더 많습니다.

인공신경망

위에 열거된 많은 방법론은 인공 신경망(Artificial Neural Networks, ANN)으로 알려진 방법론 클래스에 속하는데, 이 방법론은 우리 자신의 두뇌가 연결되는 원리에 기반을 두고 있습니다. 인공 신경망은 오늘날 기계 학습 방법의 가장 일반적인 형태이며 가장 발전된 형태입니다.



인공 신경망은 구조적으로 여러 레이어의 노드(뉴런)들로 구성되며 각각의 노드들은 다음 레이어의 노드들에 연결됩니다. 첫 번째 레이어는 입력 레이어, 마지막 레이어는 출력 레이어이고, 중간 레이어는 숨겨진 레이어로 불립니다. 대략적으로 신경망이 작동하는 방식은 뉴런 간 연결 각각에 가중치가 부여되고 이 가중치는 원하는 출력과 입력이 일치할 때까지 조정되는 것입니다. 각 출력 뉴런은 출력이 입력과 얼마나 일치하는지에 대한 확률을 나타냅니다. 위 다이어그램은 ANN에 서로 다른 행복한 얼굴, 슬픈 얼굴, 중립적인 얼굴을 연달아 제시하면서 학습시켰습니다. 행복한 얼굴을 제시하면 출력에서도 행복한 얼굴을 나타내는 상황에 도달할 때까지 각 얼굴은 어떤 방식으로든 가중치를 조정합니다.

딥 러닝

딥 러닝 방법론은 인공 신경망 계열의 방법론으로서, 개발된 이후 지속적으로 기계 학습 작업 대부분에서 최고 수준의 성과를 내고 있습니다. 딥 러닝 네트워크는 비지도 학습 및 강화 학습에도 사용될 수 있지만 일반적으로는 지도 학습 문제에 더 자주 사용됩니다.

딥 러닝은 여러 레이어(multiple layers)를 연속적으로 사용하는 방식으로 작동하는데 각 레이어셋은 그 자체로 신경망 역할을 효과적으로 수행합니다. 예를 들어 이미지 처리 문제의 경우 첫 번째 레이어셋은 원시 픽셀 데이터를 입력으로 받아 다양한 모양의 선과 같은 기본 모양을 출력할 수 있습니다. 다음 레이어에서는 이를 가져와 원, 직사각형 등 특정 모양을 출력할 수 있으며, 이러한 과정은 출력물이 다양한 유형의 차량 모양을 형성할 때까지 계속될 수 있습니다.

입력 데이터 유형과 여러분이 해결하려는 문제 유형에 따라, 여러분에게 특정 유형의 딥 네트워크가 다른 유형보다 더 유용할 수 있습니다.

표 형식 데이터셋:

표 형식 데이터셋에 사용할 수 있는 딥 네트워크는 많지 않습니다. 왜냐하면 자주 쓰이는 네트워크 대부분은 특징을 자동으로 추출하기 위해 공간 정보(즉, 어떤 속성이 서로 인접해 있는지)를 사용하기 때문입니다. 표 형식 데이터를 만나면 초기 데이터셋으로는 기본 신경망(basic neural networks)이 최선일 것입니다.

오디오/비디오/시간대 데이터셋:

시간 정보(즉, 시간에 따라 변하는 데이터)가 있는 경우 순환 신경망(Recurrent Neural Networks, RNN)과 장단기 기억망(LSTM)이 가장 좋은 두 후보입니다.

이미지 데이터셋:

딥 러닝의 발전은 대부분 이미지 데이터셋에서 이루어졌습니다. 분류, 분할 또는 추적하려는 이미지가 있는 경우 CNN(컨볼루션 신경망)이 좋은 딥 러닝 후보 중 하나입니다.

복잡성에 대한 이해

상황이 얼마나 복잡해지고 있는지 알아보기 위해 두 가지 대중적인 방법에 대해서 다른 경우보다 더 자세히 설명해 보겠습니다. 요점은 방법론 그 자체를 이해하는 것이 아닙니다. 이는 부수적입니다. 핵심은 세부 사항이 실제로 얼마나 복잡하고 전문적인지 이해하는 데 있습니다.

세부 사항에 앞서 긴 이야기를 요약하면 다음과 같습니다.

컨볼루션 신경망(Convolutional Neural Networks, CNN)은 강력한(on steroids) 인공 신경망입니다.

장단기 기억망(Long Short-term Memory Networks)은 타임머신을 장착한 강력 인공 신경망입니다.

컨볼루션 신경망(CNN)

시각적 데이터셋과 자연어 처리에 주로 적용되는 CNN은 가장 널리 사용되는 딥 네트워크 중 하나입니다. CNN은 입력 레이어, 여러 숨겨진 레이어, 출력 레이어로 구성됩니다.

입력 레이어는 일반적으로 학습셋에 있는 이미지(들)입니다. 숨겨진 레이어는 대부분 컨볼루션 레이어, 정류 선형 단위(ReLu) 레이어, 풀링 레이어, 완전 연결 레이어입니다. (헉!)

CNN의 핵심 블록인 컨볼루션 레이어는 이미지 어딘가에서 특정한 특징이 인식될 때 활성화되는 필터를 학습합니다. 풀링 레이어는 입력 이미지를 겹치지 않는 창으로 분할하고 각 영역의 최대값을 출력하는 하위 샘플링 레이어입니다. ReLu 레이어는 활성화 맵에서 음수 값을 0으로 설정하여 제거하는 데 사용됩니다. 완전 연결 레이어는 분류가 실제로 수행되는 곳입니다. 이들은 어떤 활성화가 어떤 클래스와 관련이 있는지 학습합니다.

마지막으로, 여러분의 예측에 대응하는 출력 레이어와 손실 레이어가 있습니다. 손실 레이어는 예측과 실제 출력 간 편차에 대해 학습에서 어떻게 불이익을 줄지 지정합니다. 여러분의 문제에 따라 여러분은 여러 종류의 손실 레이어를 선택할 수 있습니다. 예를 들어 소프트맥스 손실은 N개의 상호 배타적인 클래스 중 하나의 클래스에 대한 예측에 사용되는 반면, 유클리드 손실은 연속 값을 예측(회귀)하는 데 사용됩니다.

CNN은 딥 Q-네트워크(강화 학습에 사용), Fast RCNN, Fully CNN 등 다른 많은 방법론의 기초입니다.

복잡하다고 말씀드렸었죠!

장단기 기억망(LSTM)

LSTM은 순환 신경망(RNN)의 변형입니다. LSTM은 단일 데이터 포인트(예: 이미지 또는 표 형식 데이터) 뿐만 아니라 순서형 데이터(예: 비디오 또는 음성)에 적합한 심층 네트워크 유형입니다. 1997년에 도입되어서 기계 학습 분야에서 비교적 오래된 기술이지만, 인간 행동 인식 및 음성 인식과 관련된 대중적인 기계 학습 문제에서 여전히 최고 수준의 성과를 내고 있습니다. 이들은 또한 영화나 TV 용 자동 음악 작곡이나 자막 생성 등 일부 특이한 문제용으로 사용되어 왔습니다.

LSTM은 입력 게이트, 출력 게이트, 셀 및 망각 게이트로 구성됩니다.

셀은 여러 시간 간격으로 과거 값을 기억할 수 있고, 망각 게이트는 셀 안팎의 정보 흐름(즉, 기억해야 할 사항과 적용해야 할 시기)을 제어합니다.

로우샷 학습

네트워크는 먼저 데이터의 특징을 학습해야 하기 때문에 딥 러닝 네트워크는 통상 방대한 양의 데이터를 필요로 해 왔습니다. 현재 기술로는 이미지 인식 및 오디오 분류를 위해 라벨링된 수백만 개의 데이터 포인트를 사용해야 하는데, 이것이 항상 달성가능한 것은 아닙니다.

이러한 데이터를 입수해야 하는 불가능한 요구 사항에 대응하기 위하여 지난 몇 년 동안 로우 또는 퓨샷 학습이라고 하는 딥 러닝의 새로운 분야가 주목을 받고 있습니다. 로우샷 학습은 각 클래스가 매우 적은 샘플(일반적으로 20개, 10개, 5개, 심지어 1개)을 가진 데이터셋에서 자동으로 특징을 학습하는 것을 목표로 합니다. 수백만 개의 샘플로 학습된 네트워크 만큼 결과를 향상시킬 수는 없어도 로우샷 학습법은 꽤 경쟁력이 있습니다.

성공 측정

여러분은 데이터를 구했고 기계 학습 모델에 데이터를 넣었으며 이제 몇 가지 결과를 얻었습니다. 문제는 이겁니다. 그 결과가 좋은가요? 여러분이 해결하려는 문제에 따라 그 성공 수준을 측정하는 데 사용할 수 있는 몇 가지 측정법(metrics)이 있습니다. 가장 일반적으로 알려진 것은 평균 제곱근 오차, 혼동 행렬, 재현율, 정밀도입니다.

평균 제곱근 오차(Root Mean Square Error, RMSE):

이 오류 지표는 모델에서 예측한 값과 데이터셋에서 관찰된 값 간의 차이를 측정합니다. 예측 오류의 표준 편차로 정의됩니다(문헌에서 종종 '잔차 residuals'라고 부릅니다). 그 특성상 RMSE는 예측 문제에 특히 적합합니다.

혼동 행렬**(Confusion matrix)**: 분류 문제를 다루는 경우 모델의 성능을 측정하는 한 가지 방법은 혼동 행렬을 사용하는 것입니다. 이를 통해 클래스 간의 "혼동(confusion)", 즉 클래스가 자꾸 다른 클래스로 잘못 표시되는 것을 볼 수 있습니다. (이 설명이 여러분에게 '혼동'을 주지 않았기를 바랍니다!) 여러분의 문제가 이진법적이라면(즉, 단 두 개의 클래스를 가진 경우) 혼동 행렬은 이와 같을 것입니다. 혼동 행렬은 다음 정보를 가지고 있습니다.

- 참양성(True Positive, TP) : 관찰이 양성이며 예측도 양성입니다.
- 위음성(False Negative, FN) : 관찰은 양성이지만 예측은 음성입니다.
- 참음성(True Negative, TN) : 관찰이 음성이며 예측도 음성입니다.
- 위양성(False Positive, FP) : 관찰은 음성이지만 예측은 양성입니다.

이진법 문제에 대한 혼동 행렬

		실제 클래스	
		클래스 1	클래스 2
다음과 같이 예측됨	클래스 1 (예) 남성	TP	FP
	클래스 2 (예) 여성	FN	TN

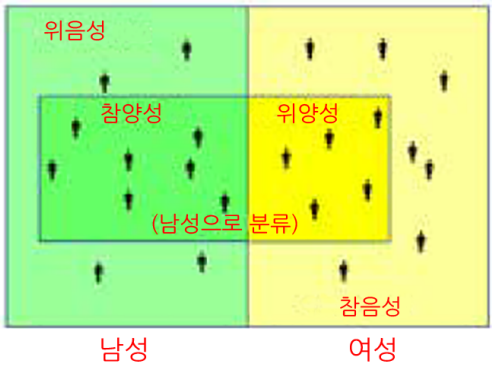
위 행렬은 두 개의 클래스만 보여줍니다. 클래스가 많을수록 혼동 행렬이 더 복잡해집니다. 혼란하지 않도록 더 복잡한 예시는 보여드리지 않겠습니다.

정확도**(Accuracy)**: 올바르게 분류된 저장값의 비율, 즉 여러분의 분류기가 성공한 횟수입니다. 다음과 같이 계산할 수 있습니다.

$$\text{정확도} = \mathbf{(TP+TN)/(TP+TN+FP+FN)}$$

정밀도**(Precision):** 관련 선택 항목의 총수를 나타냅니다. 따라서 이는 참양성을 양성으로 선택된 모든 값(참양성+위양성)으로 나눈 값입니다.

재현율**(Recall):** 한 클래스에서 올바르게 분류된 요소의 수(참양성)를 해당 클래스에 속하는 것으로 지정된 요소의 총수(참양성, 위음성)로 나눈 값입니다. 위양성(다른 클래스 요소를 해당 클래스로 잘못 분류한 것)은 불이익을 받기 때문에 '정밀도'보다 더 제한적입니다.



따라서 기계가 남성을 남성이라고 분류하려는 위 예시(그림)에서 정밀도는 다음과 같습니다.

$$\frac{\text{남성 7명 (TP)}}{\text{남성 7명(TP) + 여성 5명(FP)}} = 0.583$$

재현율은 다음과 같습니다.

$$\frac{\text{남성 7명 (TP)}}{\text{남성 7명(TP) + 남성 4명(FN)}} = 0.636$$

두 경우 모두 값이 높을수록 좋습니다. 따라서 위 결과는 그다지 좋다고 볼 수 없으므로 여러분은 이 솔루션을 구입하지 않는 것이 좋겠습니다.

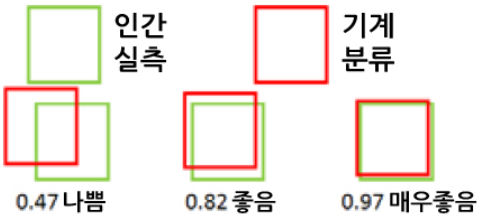
F1: 정밀도와 재현율의 관계를 동시에 확인하는 데 유용합니다. 다음과 같이 정의됩니다.

F1 = 2 x (정밀도 **x** 재현율**)/(**정밀도 **+** 재현율**)**

다시 말하지만, 값이 높을수록 좋습니다. 1은 완벽합니다!
(믿기 어려울 수는 있겠습니다.)

자카드 지수(Jaccard Index) 또는 유사성 계수(Similarity Coefficient)

자카드 지수는 서로 다른 샘플셋 간의 유사성을 측정하는 데 사용되는 통계입니다. 이는 이미지에서 인간 실측과 기계 분류 모두에 공통인 이미지 부분을 둘 중 하나에 해당하는 이미지의 해당 부분으로 나눈 값으로 정의됩니다. 0과 1 사이의 값이 클수록 좋습니다.



자카드 지수를 보완하는 자카드 거리(Jaccard Distance)는 두 영역 간의 비 유사성을 측정하므로 점수가 높을수록 차이가 커집니다.

제한 사항

AI, 데이터 과학, 기계 학습은 훌륭하지만 완벽하지 않습니다. 한계가 있다는 사실을 분명히 하지 않았을 수 있으며 과대 광고로 인해 성능을 오해하고 가능성을 과대 평가하기 쉽습니다. 성공적으로 사용된 경우라 하더라도 알려질 때 보였던 부분보다 더 제한적인 의미로 적용되는 경우가 많습니다. 따라서 이들 분야가 인도주의적 재난, 범죄 예방, 질병 진단, 기타 여러 가지 대응을 어떻게 지원했는지에 대한 보고서를 읽을 때, 해당 성과가 종종 특정 상황에 매우 특화되어 있음을 이해할 필요가 있습니다.

우리 모두 개인적으로는 일이란 것이 항상 우리가 원하는 대로 작동하지 않는다는 사실을 알고 있습니다. 시리, 알렉사, 구글이 종종 여러분을 오해하지 않던가요? 아마존과 넷플릭스가 여러분의 취향과 전혀 맞지 않는 내용을 추천하지 않던가요? SatNav(내비게이션)를 따라간 후 예기치 않은 교통 체증을 만난 적은 없던가요? 챗봇이 여러분의 질문에 제대로 답하지 않아 결국 은행 고객센터에 전화해야 했던 적은 없었나요?

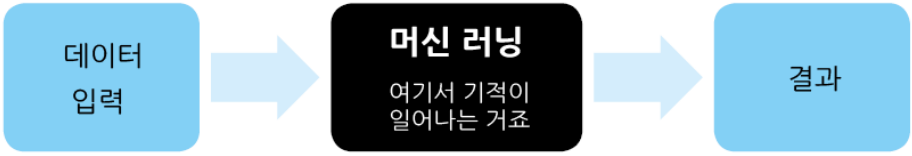
따라서 강점을 이해하는 것만큼 한계를 인정하는 것이 중요합니다.

[AI의] 가장 진보된 기법에도 한 가지 분명한 한계가 있기 때문입니다. 즉, 모두 데이터에 의존적이라는 사실입니다. 대부분의 기계 학습 모델과 데이터 과학 방법론은 주석이 달린 대규모 데이터셋으로 학습합니다. 이러한 주석들은 대개 사람들이 생성합니다. 데이터의 품질, 완전성, 정확성 및 주석은 결과의 품질에 직접적인 영향을 미칩니다. 이런 식으로 생각해보세요. 여러분이 어떤 일을 어떻게 해야 할지 모를 때에는, 어떻게 해야 하는지에 대한 예시를 찾아볼 것입니다. 기계 학습 모델에도 동일한 것이 필요합니다. 여러분이 분류하거나 예측하려는 클래스에 대한 충분한 예시가 없으면, 기계가 성공적으로 학습하는 것이 불가능합니다.

> 불균형, 편향 또는 불완전한 데이터로 인한 피해 사례는 우려될 만큼 많습니다. 2014년 한 회사는 채용 절차를 최적화하고 편향을 제거하기 위해 이력서를 자동으로 필터링하는 기계학습 기반 채용 시스템을 만들었습니다. 그러나 이 시스템은 10년 간의 해당 회사 [채용] 데이터로 학습했는데, 이 기간 동안의 채용 관행은 편향적이었습니다. 따라서 시스템이 산출한 결과는 [편향된] 해당 관행을 강화하는 것이었습니다. 조사 결과 이 시스템은 "여성", "여성 체스 클럽 회장"과 같은 단어를 포함한 이력서에 불이익을 주고 두 개 여자 대학 졸업생을 거부한 것으로 나타났습니다.

기계 학습의 큰 문제

직설적으로 말하자면, 기계 학습의 가장 큰 문제는 그것이 무엇을 하는지 아무도 모른다는 점입니다!



기계 학습 단계는 딱 블랙박스 같습니다. 우리는 인공 신경망의 일반 원칙을 이해하지만 그것이 적용될 때 실제로 어떤 일이 벌어지는지 자세히 이해하지 못합니다. 모든 의도와 목적을 불문하고 이는 마법입니다!

좋은 결과가 나오는데 왜 이것이 문제가 되는지 물을 수 있습니다. 문제는 무슨 일이 일어나고 있는지 모르기 때문에 문제가 발생했을 때 바로잡을 수 없고 주어진 답변을 이해하기 어렵다는 데 있습니다. 이 문제는 우리가 어디에 기계 학습을 적용할 수 있는지를 볼 때 심각한 의미를 갖습니다. 우리가 하려는 일이 새끼 고양이 사진을 찾기 위해 인터넷을 뒤지는 것 뿐이라면 별다른 문제가 되지 않습니다. 그러나 항공기를 제어하기 위해 기계 학습을 적용하려면 실제로 무슨 일이 일어나고 있는지 이해해야 합니다. 이 분야에서는 기호 AI와 기존 프로그래밍이 엄청난 이점을 갖고 있습니다. 해당 상황에서 우리는 이들이 무엇을 하는지 알고 있기 때문입니다. 따라서 이런 접근 방식들은 투명합니다. 투명성과 기계 학습 분야에서 현재 여러 노력들이 이루어지고 있습니다만, 통상 "설명 가능한 AI"로 알려져 있는 이 솔루션이 현실화되려면 아직 멀었습니다. 기껏해야 기계가 강아지 사진을 강아지로 분류하는 데 사용된 픽셀을 판단할 수 있을 뿐입니다. 하지만 기계의 픽셀 선택은 우리에게 약간 이상할 수 있고 그 작동 방식에 대한 정보가 그리 많지 않습니다.

마무리

여러분의 차가 아직 식지 않았기를, 여러분이 비스킷을 맛있게 드셨기를 바랍니다. 무엇보다 여러분에게 이 시간이 AI에 대해 유용하고 재미있는 학습 시간이 되었기를 바랍니다! AI는 "새로운 전기"로 불리면서 점점 더 많은 산업에 동력을 공급하고 일상 생활의 여러 측면에 영향을 미칠 것으로 예상되는 기술입니다. AI에 대한 이해, 특정 응용 분야에 대한 전문 지식, 그리고 확고한 윤리 의식을 결합하는 것이야말로, 이 강력한 기술의 가치를 창출하고 모두를 위해 더 나은 세상을 만드는 방식으로 이 기술을 사용하는 방법을 찾아가는 길일 것입니다.