

머신러닝 입문 가이드

[illegible]

본 PDF 문서는 IDG Korea의 프리미엄 회원에게 제공하는 문서로, 저작권법의 보호를 받습니다.
IDG Korea의 허락 없이 PDF 문서를 온라인 사이트 등에 무단 게재, 전제하거나 유포할 수 없습니다.

머신러닝, 사이버 보안 꿈의 실현인가, 헛된 망상인가

Taylor Armerding | CSO

직원 가운데 누가 일을 게을리 하는지, 누가 곧 퇴사를 계획하고 있는지, 또는 누가 회사 자산 데이터를 훔칠 계획을 하고 있는지 알고 싶은가?

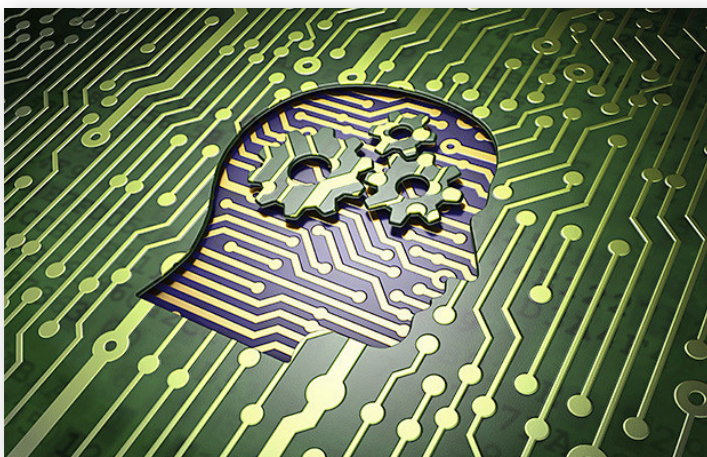
최고 보안 및 개인정보보호 책임자를 컨설팅하는 에어리얼 실버스톤에 따르면, 머신러닝(Machine Learning)이 기존 HR 부서보다 훨씬 더 빠르게 이런 사안을 파악할 수 있게 해준다. 실버스톤은 머신러닝이 “일주일 분량의 기본 데이터”를 가지고 이런 작업을 수행하는 모습을 본 적이 있다고 말했다.

머신러닝, 무너진 사이버 보안의 ‘희망의 불빛’

법률회사인 리틀러 멘델슨(Littler Mendelson) 데이터 분석 담당 글로벌 책임자 제브 J. 아이젠 역시 “사이버 보안에 머신러닝을 활용하는 것은 비교적 새로운 분야지만 머신러닝은 사이버 보안에 ‘혁명(revolutionize)’을 일으킬 잠재력이 있다”면서 머신러닝의 가치를 인정했다.

전 시만텍 CTO이자 KNRL 랩스의 관리자 아미트 미탈은 지난해 7월 포천이 후원한 패널 토론에서 “인공 지능(머신러닝은 인공 지능의 한 구성 요소)이 이 혼돈에서 얼마 안 되는 희망의 불빛 가운데 하나”라고 언급했다. 여기서 혼돈은 사이버 보안을 의미한다. 미탈은 사이버 보안이 “사실상 무너졌다”고 주장했다.

“조작이 불가능한 시스템이란 없다. 다만 ‘머신러닝이 없다면 이 문제가 얼마나 더 악화될 것인가?’라는 질문을 생각해봐야 한다.”



사이버 보안의 최신 망상

그러나 머신러닝의 혁신적 역량에 대해 회의적인 전문가들도 있다.

브로미움(Bromium) CTO 사이먼 크로스비는 최근 다크 리딩(Dark Reading)에 게시한 글에서 “머신러닝은 사이버 보안의 최신 망상”이라며, “보안에는 만병통치약이 존재하지 않는다. 머신러닝 도구들이 도움이 된다는 증거는 어디에도 없다”고 말했다.

가트너의 결론도 이런 회의론과 상당 부분 일치한다. 가트너는 2015 하이프 사이클(Hype Cycle)에서 ‘

“머신러닝이 만병통치약은 아니지만
머신러닝을 과소평가하는 것은
지나치게 근시안적인 생각이다.”

허황된 기대가 가장 극심한 수준인 5개 기술 가운데 하나로 머신러닝을 지목했다.

그러나 많은 전문가가 사이버 보안에 만병통치약이 존재하지 않는다는 데는 동의하지만 그것과 헛된 망상 사이에는 분명한 차이점

이 있다고 주장한다.

인터세트(Interset) CTO 스테판 조는 “머신러닝이 만병통치약은 아니다. 그러나 보안 기술에 엄청난 자금을 투자하고도 매주 막대한 손실이 일어나는 현 상황에서 기업이 머신러닝을 과소평가하는 것은 지나치게 근시안적인 생각”이라고 말했다.

과장은커녕 과소평가된 상황

하버드 대학 정량사회과학(Quantitative Social Science) 연구소 책임자 개리 킹 역시 머신러닝은 “결코 헛된 망상이 아니다”고 말했다.

킹은 “무슨 일이든 척척 해낸다는 의미는 아니다. 지금은 머신러닝이 효과를 제대로 발휘하지 못하는 분야가 있다. 머신러닝을 사용해 좋은 결과를 얻는 사람들도 있지만 별 성과를 얻지 못하는 경우도 많다”며, “숙련된 인력이 머신러닝을 이끌어야 한다”고 설명했다.

그러나 “이런 인력 역시 최대한 많은 도움을 받아야 하는데, 이렇게 되면 머신러닝이 큰 힘을 발휘할 수 있다”고 덧붙였다.

실버스톤은 “어떤 시스템이 해킹되지 않도록 막아줄 수 있는 것은 아무것도 없다는 의견에는 동의하지만 가트너의 하이프사이클 결론은 완전히 틀렸다”며, “머신러닝은 과장되기는커녕 심각하고도 중대하게 과소평가되고 있다”고 주장했다.

“데이터가 충분하고 해당 데이터가 특정 추세를 보이는 이유를 이해한다면 예측의 정확도를 90%보다 훨씬 더 높일 수 있다. 99% 이상도 충분히 가능하다”는 것이 실버스톤의 설명이다.

이는 즉, 머신에게 ‘다음 주에 공격을 받게 되는가?’정도가 아니라 ‘다음 주 화요일 오후 3시에 중국에서 실행되는 공격을 받게 되는가?’, 나아가 ‘다음 해킹이 일어날 대상과 시점은 언제이며, 그 해킹의 근원지와 실행자는 누구인가?’라는 질문까지 던질 수 있다는 의미다.

실버스톤은 “지금도 매우 높은 정확도로 가능한 이야기”라면서, “또한 이것보다 훨씬 더 복잡한 알고리즘도 가능할 뿐만 아니라 실제로 사용되고 있다고 생각한다. 우리가 지금 할 수 있는 일은 나 자신도 2년 전에는 믿지 못했던 일들”이라고 말했다.

머신러닝, 효과났다는 변칙성 찾기에서도 효과 입증

브로미움의 크로스비도 머신러닝이 ‘강력한 도구’라는 데에는 동의하며 구글 검색이나 아마존, 넷플릭스 등이 운용하는 추천 엔진 등의 경우 그 효과가 크다는 점도 인정했다. 그러나 크로스비는 “구글이 유행성 독감을 파악하고자 시도했지만 그 결과가 ‘터무니없이 부정확’했다”고 지적했다.

크로스비는 머신러닝이 요소 간의 유사성을 찾는 데는 아주 뛰어나지만 “변칙성을 찾는 데는 그다지 효과적이지 않다. 사실 변칙적 행동에 대한 모든 논의는 정상적 행동에 대한 설명이 가능하다는 것을 전제로 한다”며, “이것이 매우 어려운 부분”이라고 말했다.

크로스비는 “이는 악의적인 사람에게 ‘정상 속에 숨을 수 있는’ 기회를 풍부하게 제공하며, 나아가 악의적 행동을 정상으로 인식하도록 시스템을 교육시킬 기회까지 제공한다”고 지적했다.

그러나 리틀러 멘델슨의 아이젠은 어려움은 있지만 그렇다고 머신러닝에 부가적인 가치가 없다는 의미는 아니라고 말했다. 아이젠은 “조작이 불가능한 시스템이란 없다. 다만 ‘머신러닝이 없다면 이 문제가 얼마나 더 악화될 것인가?’라는 질문을 생각해봐야 한다”고 말했다.

스테판 조의 입장은 더욱 단호하다. 조는 “머신러닝은 무엇이 정상인지 정의한 다음 비정상(변칙성)을 정의할 수 있음을 이미 입증했다”고 말했다.

조는 머신러닝이 사람을 대신하지는 않겠지만 사람이 하는 일(패턴을 인식한 다음, 이 패턴에 맞지 않는 변칙성을 인식하는 것)을 자동화한다는 데 동의하면서 “머신러닝은 데이터 셋을 받아 패턴을 찾고 무엇이 정상이고 무엇이 ‘이상(weird)’인지를 정의하는 것”이라고 말했다.

또한 조는 공격자가 자신의 행동을 정상으로 인식하도록 시스템을 속이기 위해 사용하는 기법인 ‘모델 감염(model poisoning)’에 대해 “데이터 소스별로 여러 모델을 사용하는 방법으로 대처할 수 있다”고 반박했다.

즉, 상대방은 위험한 행동을 탐지하기 위해 사용되는 모든 모델에 대한 완전한 지식을 갖추어야 하며, 또한 모든 모델과 데이터 소스를 동시에 오염시켜야 함을 의미한다는 것이 조의 설명이다.

조와 실버스톤은 머신러닝이 조직 내에서 퇴사할 가능성이 높거나 악의를 품고 데이터를 훔칠 직원을 예측할 수 있음을 입증했다고 말했다. 조는 “이 방법을 사용해 악의적인 사람들을 계속 포착하고 있다”고 밝혔다.

실버스톤은 “일주일 분량의 기본 데이터가 있으면 어느 직원이 일을 게을리 하는지, 퇴사할 가능성이 높거나 악의적으로 행동할 가능성이 높은 직원이 누구인지 찾아낼 수 있다. 또한 그날 특정 시간에 어느 정도의 대역폭이 필요한지, 어느 포트가 필요한지, 사람들이 어느 사이트를 방문할 것인지도 예측이 가능하다. 내가 직접 경험한 것”이라고 말했다.

맥락을 학습한다

조는 머신러닝의 강점 가운데 하나는 맥락(context)을 학습할 수 있다는 점이라고 말했다. 조는 “‘행위자가 언제, 어디서 특정 행동을 수행할 권한을 갖고 있는가?’를 예로 들 수 있다. 이는 하나의 단순한 예측을 통해 60GB 방화벽 대신 6GB의 방화벽만 사용할 수 있게 됨을 의미한다. 게다가 이는 수많은 가능성 가운데 하나일 뿐”이라고 설명했다.

조는 “머신러닝이 정상과 비정상의 차이를 학습하고 집어낼 수 없다고 주장하는 사람은 ‘과거의 머신러닝에 대해 이야기하는 것’이다. 유사점과 차이점을 파악하는 데 있어 머신러닝보다 더 좋은 것은 없다”면서, “데이터 비정상 찾아내기에서 머신과 대결을 원하는 사람은 얼마든지 환영한다”고 덧붙였다.

수십년 전 머신러닝이 현재에 각광받는 이유

제브 J. 아이젠은 머신러닝이 수십 년 전부터 시작되었다는 점을 강조했다. 실제 최소한으로 잡아도 유명한 영국 과학자 앨런 튜링까지 거슬러 올라간다. 튜링은 2차대전 당시 팀장으로 나치의 ‘에니그마(Enigma)’ 코드를 크랙한 머신을 고안했으며, 최근 영화 <디 이미테이션 게임(The Imitation Game)>을 통해 조명되기도 했다. 1950년 논문에서 튜링은 “머신이 생각할 수 있는가?”라는 질문을 던졌다.

아이젠은 새삼 현 시점에서 머신러닝이 ‘새롭게’ 화두가 된 이유에 대해 “데이터 스토리지가 발달했고 더 높은 품질의 데이터를 가지고 있으며 이 데이터들을 더 신속하게 처리할 수 있게 되었기 때문”이라고 설명했다.

조는 “머신러닝이 과장됐다는 인식은 사이버 보안에서 머신러닝의 사용이 비교적 새로운 분야라는 데 기인한다”고 말하며, “다른 영역에서 거둔 것과 같은 효과를 입증하면 사이버 보안에 혁명을 일으키게 될 것이다”고 주장했다.


사이버 보안에서의 머신러닝, 빠르게 도입될 것

물론 머신러닝을 지지하는 사람들도 머신러닝이 공공 및 민간 분야 전반에 걸쳐 보편적으로 사용될 수 있을 정도로 성숙했다고 주장하지는 않는다.

실버스톤은 “대기업 조직이 머신러닝으로 혁신하는 경우는 많지 않다”면서, “그러나 연구 기관, 대학, 그리고 금융 분야에서는 현실화되고 있다”고 말했다.

조는 “보안 종사자들은 데이터를 공유하는 데 익숙하지 않다. 즉, ‘방금 침투당했어. 여기 내 방화벽 트래픽 로그를 보여줄게. 네 것도 보여줘’ 등의 대화는 오가지 않는다는 의미다. 이런 이유로 사이버 보안에서 머신러닝의 도입이 더 어렵다. 또한 많은 기업이 빅데이터 기업이 되었음을 이제 막 인식하고 있는 단계다”고 말했다.

조는 “그러나 다른 분야에서 머신러닝을 성공적으로 활용하는 방법에 대한 학습이 이미 대부분 이루어진 만큼 사이버 보안 분야에도 머신러닝이 빠르게 도입될 것”이라고 말했다.

실버스톤은 “가능성이 대단히 크다”면서 “머신러닝 개념을 디지털 통화에 적용한다면 누군가 돈을 마음껏 훔칠 수 있다 하더라도 그 돈을 사용할 수는 없게 된다는 것을 의미한다”고 말했다. 

머신러닝 실전 입문

SriSatish Ambati | Infoworld

구글, 페이스북과 같은 기업이 머신러닝(machine learning)을 사용해 자동차를 운전하고 음성을 인식하고 이미지를 분류한다는 이야기를 들어봤을 것이다. 그런데 머신러닝이 당면한 자사의 비즈니스와는 무슨 관계가 있을까? 우선 기업들이 현재 머신러닝을 어떻게 사용하고 있는지 살펴보자.

- 한 지불 결제 처리 업체는 실시간으로 10억 개 이상의 거래 중에서 숨겨진 사기를 탐지함으로써 손실 금액을 월 100만 달러 가량 줄이고 있다.
- 어느 자동차 보험업체는 상세한 지역 관련 데이터를 사용해 심각한 기상 악화가 비즈니스에 미치는 영향을 모델링하고 보험 청구로 인한 손실 금액을 예측한다.
- 한 제조업체는 차량 텔레매틱스에서 생성하는 데이터를 사용해 운영 지표 상의 패턴을 발견하고 선제적인 정비를 유도한다.

이런 성공 사례를 관통하는 테마는 두 가지다. 첫째, 각 애플리케이션은 빅데이터에 의존한다. 데이터는 그 양이 방대하고 포맷이 다양하며 속도가 빠르다. 둘째, 각각의 경우 머신러닝은 새로운 시각을 발견하고 가치를 창출한다.

최근 머신러닝 인기가 폭발하는 이유

머신러닝의 기술적인 토대가 만들어진 시기는 50여 년 전이지만 얼마 전까지만 해도 학계를 제외하면 머신러닝의 역량에 대한 인지도는 미미했다. 머신러닝에는 막대한 컴퓨팅 파워가 필요한데, 얼리어댑터들이 비용 효율적이게 해줄 인프라가 없었기 때문이다.

최근 머신러닝에 대한 관심과 활동이 폭증한 이유 중에는 여러 가지 융합이 이뤄지는 추세도 있다.

- 무어의 법칙(Moore's Law)으로 컴퓨팅 비용이 급격히 낮아져 지금은 최소한의 비용으로 강력한 컴퓨팅 성능을 폭넓게 이용할 수 있다.
- 새롭고 혁신적인 알고리즘이 더욱 빠른 결과를 제공한다.
- 데이터 과학자들이 머신러닝을 효과적으로 적용하기 위한 이론과 실무 지식을 축적했다.

무엇보다 빅데이터가 대대적으로 도입되면서 일반적인 통계로는 해결이 불가능한 분석 문제가 발생했다. 필요가 곧 발명을 낳는다는 말이 있듯 이전 분석 방법이 더 이상 현재의 비즈니스 환경에서 통하지 않게 된 것이다.



머신러닝 기법의 종류

머신러닝의 알고리즘은 수백 가지다. 최근 한 논문은 분류 한 가지 항목에 대해 150개 이상의 알고리즘을 벤치마킹했다. 여기서는 데이터 과학자들이 현재 가치 창출을 위해 사용하는 핵심적 기법을 설명한다.

데이터 과학자들은 지도 러닝(supervised learning)과 자율 러닝(unsupervised learning)을 위한 기법을 구분한다. 지도 러닝 기술에는 결과에 대한 사전 지식이 필요하다. 예를 들어 마케팅 캠페인의 과거 데이터를 다루는 경우 잠재 고객이 응답했는지 여부에 따라 각 임프레션을

분류하거나 이런 고객이 소비한 금액을 확인할 수 있다. 이 때 지도 러닝 기법은 예측과 분류를 위한 강력한 도구가 된다.

그러나 현실에서는 이벤트의 '최종적' 결과를 알지 못하는 경우가 많다. 예를 들어, 사기(fraud)의 경우 이벤트가 끝나고 오랜 시간이 지나기 전까지는 거래의 사기 여부를 여부를 알 수 없다. 이 경우 사기 거래 예측을 시도하기보다는 머신러닝을 사용해 일반적이지 않은 거래를 식별해 추가 조사하도록 표시하는 방법을 사용할 수 있다. 자율 러닝은 구체적인 결과에 대한 사전 지식이 없지만 데이터를 통해 유의미한 지식을 얻고자 하는 경우 사용된다. 가장 광범위하게 사용되는 지도 러닝 기법은 다음과 같다.

- **일반화 선형 모델(Generalized linear models, GLM)**: 선형 회귀(linear regression)의 발전된 형태로, 다양한 가능성 분산과 연결 함수를 지원해 분석가가 더 효과적으로 데이터를 모델링할 수 있도록 한다. 그리드 탐색(grid search)으로 강화된 GLM은 전통적인 통계와 가장 발전된 머신러닝의 조합이다.
- **의사결정 트리(Decision trees)**: 모 집단을 대상 변수에 대해 동질적인 더 작은 조각으로 점진적으로 분할하는 규칙 집합을 학습하는 자율 학습 방법.
- **랜덤 포레스트(Random forests)**: 널리 사용되는 총체적 학습 방법으로, 다수의 의사결정 트리를 학습한 다음, 트리 전반에 걸친 평균을 구해 예측을 산출한다. 이 평균 프로세스는 일반화 가능한 솔루션을 제공하며 데이터의 불규칙 잡음(random noise)을 걸러내는 효과가 있다.
- **그래디언트 부스팅 머신(Gradient boosting machine, GBM)**: 의사결정 트리의 시퀀스 교육을 통해 예측 모델을 생성하는 방법으로, 연속되는 트리가 이전 트리의 예측 오류를 수정해 나간다.
- **딥 러닝(Deep learning)**: 데이터의 고수준 패턴을 복합적인 다계층 네트워크로 모델링하는 방법. 문제를 모델링하는 가장 일반적인 방법이며 머신러닝의 가장 어려운 문제를 해결할 잠재력을 지녔다.

자율 러닝의 주요 기술은 다음과 같다.

- **클러스터링(Clustering)**: 개체를 다수의 메트릭스에서 상호 유사한 세그먼트 또는 클러스터로 그룹화하는 기법. 고객 세분화가 클러스터링의 실제 예다. 클러스터링 알고리즘은 무척 다양하는데, 가장 널리 사용되는 것이 k-평균(k-means)이다.
- **비정상 탐지(Anomaly detection)**: 예상치 못한 이벤트 또는 결과를 식별하는 프로세스. 보안, 사기 등의 분야에서는 모든 거래를 철저하게 조사하기란 불가능하므로 가장 비일반적인 거래에 체계적으로 플래그를 지정해야 한다. 앞서 지도 러닝에서 언급한 기법인 딥 러닝도 비정상 탐지에 사용

할 수 있다.

- **차원 축소(Dimension reduction)**: 고려 대상 변수의 수를 줄이는 프로세스. 조직이 더 많은 데이터를 캡처할수록 예측에 사용 가능한 예측 변수(또는 특성)의 수도 급격히 증가한다. 특정 문제에 대해 가치있는 정보를 제공하는 데이터를 식별하는 것만 해도 상당한 작업이다. 주성분 요소 분석(Principal components analysis, PCA)은 일련의 원시 특성을 평가해 이를 상호 독립적인 인덱스로 축소한다.

일부 머신러닝 기법이 다른 기법에 비해 지속적으로 좋은 결과를 내는 경우도 있지만, 특정 문제에 대해 어느 기법이 가장 효과적인지 사전에 판단할 수 있는 경우는 극히 드물다. 따라서 대부분의 데이터 과학자는 많은 기법을 시도한 후 최적의 모델을 선택한다. 즉, 데이터 과학자가 더 적은 시간에 더 많은 방법을 시도하기 위해서는 높은 성능이 필수적이다.

머신러닝의 실제 사용 사례

여러 산업과 업종을 불문하고 기업들은 머신러닝을 사용해 사람 손을 거치는 것보다 효율적으로 작업을 수행함으로써 수익을 늘리거나 비용을 줄이려고 한다. 머신러닝의 폭넓은 활용과 다용도성을 보여주는 다음과 같은 7가지 사례가 있다.

- **사기 방지**: 1억 5,000만 개의 디지털 월릿을 통해 연간 2,000억 달러 이상의 결제를 처리하는 페이팔(PayPal)은 온라인 결제업계의 선두 주자다. 이 정도 규모에서는 사기 비율이 낮다 해도 그 비용은 상당 규모에 이른다. 창업 초기에는 월별 사기 피해 금액이 1,000만 달러에 이르렀다. 페이팔은 이 문제를 해결하기 위해 최고의 연구원들로 팀을 꾸렸고 이 팀은 최신 머신러닝 기법을 사용해 사기성 결제를 실시간으로 식별하는 모델을 구축했다.
- **타겟팅 디지털 디스플레이**: 광고 기술 기업 디스틸러리(Dstillery)는 머신러닝을 사용해 버라이즌(Verizon), 윌리엄스-소노마(Williams-Sonoma)와 같은 기업이 실시간 입찰 플랫폼에서 타겟팅 디지털 디스플레이 광고를 진행하도록 한다. 디스틸러리는 개인의 브라우징 내역, 방문, 클릭 및 구매에 대해 수집된 데이터를 사용해 한 번에 수백 개의 광고 캠페인을 처리하며 초당 수천 건의 예측을 실행한다. 덕분에 디스틸러리는 투자대비 최적의 결과를 얻기 위한 타겟 광고에서 인간 마케터보다 훨씬 더 좋은 성과를 내고 있다.
- **콘텐츠 추천**: 컴캐스트(Comcast)는 X1 인터랙티브 TV 서비스 고객을 위해 각 고객의 이전 시청 습관을 기반으로 한 실시간으로 개인 맞춤형 콘텐츠 추천을 제공한다. 컴캐스트가 운영하는 머신러닝은 수십억 개의 내역 기록을 사용해 각 고객별로 고유한 취향 프로필을 작성한 다음, 공통적인 취향을 가진 고객을 클러스터로 묶는다. 그런 후 각 고객 클러스터를 대상으로 가장 인기있는 콘텐츠를 실시간으로 추적, 표시해 고객이 현재 인기있는 콘텐츠를 볼 수 있도록 한다. 더 정확한 추천으로 사용률을 높이고 고객 만족도도 높아진다.
- **자동차 품질 개선**: 재규어 랜드 로버(Jaguar Land Rover)의 신형 차량에는 60개의 온보드 컴퓨터가 탑재되며 이 컴퓨터는 2만 개 이상의 메트릭스를 기준으로 매일 1.5GB의 데이터를 생성한다. 재규어 랜드 로버 엔지니어들은 머신러닝을 사용해 이 데이터에서 고객이 차량을 실제로 어떻게 다루는지를 파악해낸다. 이렇게 얻은 정확한 사용 데이터를 통해 설계자는 부품 고장과 잠재적 안전 위험을 예측할 수 있다. 이는 예상되는 조건에 따라 적절히 차량을 엔지니어링하는 데 도움이 된다.

- **유망 잠재 고객에 집중:** 마케터들은 최적의 판매와 마케팅 기회, 그리고 최적의 제품을 판단하기 위한 도구로 '구매 성향(propensity to buy)' 모델을 사용한다. 라우터부터 케이블 TV 박스에 이르기까지 방대한 제품을 보유한 시스코(Cisco)의 마케팅 분석팀은 몇 시간 만에 6만 개의 모델을 교육시키고 1억 6,000만 명의 잠재 고객을 확보했다. 이 팀은 의사결정 트리부터 그래디언트 부스팅 머신까지 다양한 기법을 테스트함으로써 모델의 정확도를 대폭 개선했다. 이는 판매량 증가, 무익한 판매 전화 감소, 영업 담당자들의 만족도 향상으로 이어진다.
- **미디어 최적화:** NBC 유니버설(NBC Universal)은 국제 케이블 TV 배포를 위해 수백 테라바이트 용량의 미디어 파일을 저장한다. 또한 전세계 고객을 대상으로 한 배포를 지원하기 위한 효율적인 온라인 리소스 관리가 필요하다.
NBC 유니버설은 머신러닝을 사용해 척도의 조합을 기반으로 각 항목의 미래 수요를 예측한다. 이런 예측을 기반으로 수요가 낮을 것으로 예상되는 미디어를 저렴한 오프라인 스토리지로 옮긴다. 머신러닝을 통한 예측은 파일 수명과 같은 하나의 척도를 기반한 임의의 규칙에 비해 훨씬 더 효과적이다. 결과적으로 NBC 유니버설은 고객 만족도를 그대로 유지하면서 전체 스토리지 비용을 줄이고 있다.
- **의료 보건 서비스 개선:** 병원에게 있어 환자의 재입실은 심각한 문제다. 환자의 건강과 복지도 문제지만 의료 보험 공단과 민간 보험사가 재입실 비율이 높은 병원에 불이익을 주기 때문이다. 따라서 향후 건강한 상태를 유지할 가능성이 충분히 높은 환자만 퇴원시키는 역량이 병원의 재무에 큰 영향을 미치게 된다. 캐롤리나 헬스케어 시스템(Carolinas Healthcare System, CHS)은 머신러닝을 사용해서 환자의 위험 점수를 계산하고 병원 사례 관리자는 이를 바탕으로 퇴원 결정을 내린다. 이 시스템은 각 사례의 위험과 복잡성에 따라 환자에 우선 순위를 부여함으로써 간호사와 사례 관리자의 능력을 높여준다. 그 결과 CHS는 재입실 비율을 21%에서 14%로 낮췄다.

머신러닝 소프트웨어 요구 사항

머신러닝을 위한 소프트웨어는 다양하고 머신러닝 역량을 개발하고자 하는 조직이 선택할 수 있는 옵션도 많다. 머신러닝을 평가할 때는 다음과 같은 요구 사항을 고려해야 한다. 각 항목을 차례로 살펴보자.


- 속도(Speed)
 - 가치 창출 시간(Time to value)
 - 모델 정확도(Model accuracy)
 - 손쉬운 통합(Easy integration)
 - 유연한 구축(Flexible deployment)
 - 사용 편의성(Usability)
 - 시각화(Visualization)
- **속도:** 시간은 돈이다. 빠른 소프트웨어는 몸값이 비싼 데이터 과학자의 생산성을 더 높여준다. 실용 데이터 과학은 많은 경우 반복적이고 실험적이다. 프로젝트에 수백 가지 테스트가 필요할 수 있으므로 작은 속도 차이도 큰 폭의 효율성 개선으로 이어질 수 있다. 현재의 방대한 데이터 볼륨을 감안할 때 고성능 머신러닝 소프트웨어는 분산 플랫폼에서 실행되어 다수의 서버에 걸쳐 워크로드를

배분할 수 있어야 한다.

- **가치 창출 시간:** 런타임 성능(Runtime performance)은 가치를 얻기까지 소요되는 총 시간에서 일부분에 불과하다. 비즈니스에서 핵심적인 메트릭스는 데이터 수집부터 구축에 이르기까지 프로젝트를 완료하는 데 필요한 시간이다. 즉, 머신러닝 소프트웨어는 보편적으로 사용되는 하둡 및 클라우드 포맷과 통합되어야 하고 예측 모델을 조직의 어느 곳이나 배포할 수 있는 코드로 내보내야 한다.
- **모델 정확도:** 특히 위험성이 높은 경우 정확도가 중요하다. 사기 탐지와 같은 분야에서는 정확도를 조금만 개선해도 연간 수백만 달러를 절약할 수 있다. 머신러닝 소프트웨어는 데이터 과학자가 샘플이 아닌 모든 데이터를 사용할 수 있도록 해야 한다.
- **손쉬운 통합:** 머신러닝 소프트웨어는 실무 환경의 복잡한 빅데이터 소프트웨어 사양과 공존해야 한다. 일반 하드웨어에서 실행되며 특수한 HPC 머신 또는 GPU 칩과 같은 특이한 하드웨어를 필요로 하지 않는 머신러닝 소프트웨어를 찾는 것이 좋다.
- **유연한 구축:** 머신러닝 소프트웨어는 하둡 또는 프리스탠딩 클러스터에서의 코로케이션을 포함한 다양한 구축 옵션을 지원해야 한다. 아키텍처에 클라우드가 포함되어 있다면 아마존 웹 서비스, 마이크로소프트 애저, 구글 클라우드 플랫폼과 같은 다양한 클라우드 플랫폼에서 실행되는 소프트웨어를 찾아야 한다.
- **사용 편의성:** 데이터 과학자는 R, 파이썬(Python), 스칼라(Scala)와 같은 분석 언어를 포함한 다양한 소프트웨어 도구를 사용해 작업을 수행한다. 머신러닝 플랫폼은 데이터 과학자가 이미 사용하고 있는 도구와 손쉽게 통합되어야 한다. 또한 잘 설계된 머신러닝 알고리즘은 다음과 같이 시간을 절약해주는 기능을 포함한다.
 - 누락된 데이터를 처리하는 기능
 - 범주형 데이터를 변환하는 기능
 - 복잡성 관리를 위한 규칙화 기법
 - 자동 테스트 및 학습을 위한 격자 탐색
 - 자동 교차 검증(과도 학습 방지를 위함)
- **시각화:** 성공적인 예측 모델링에는 데이터 과학자와 비즈니스 사용자 간의 협업이 중요하다. 머신러닝 소프트웨어는 비즈니스 사용자에게 예측 모델의 품질과 특징을 시각적으로 평가하기 위한 도구를 제공해야 한다.

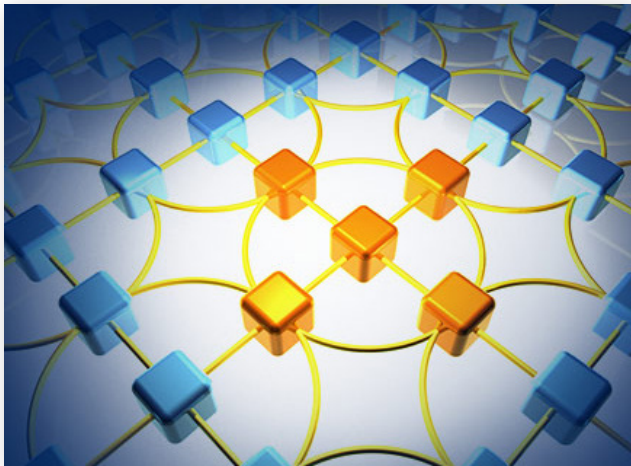
미래에는 검색처럼 머신러닝을 이용할 것이다

머신러닝이 지금의 검색처럼 보편화되고 사용하기 쉬워지고 강력해질 것으로 믿는다. 구글, 야후를 비롯한 여러 기업들은 무수히 많은 페이지에서 관련된 결과를 손쉽게 찾을 수 있도록 함으로써 일반적인 사용자에게 웹의 강력한 힘을 제공했다. 마찬가지로 머신러닝은 가치있는 통찰력을 손쉽게 얻을 수 있도록 함으로써 모든 형태의 기업에게 현대 데이터 셋이 지닌 강력한 힘을 제공하게 될 것이다.

물론 아직 거기까지 도달하진 못했다. 이 목표에 이르기 위해서는 머신러닝 개발자, 그리고 일반적인 방법으로는 충족할 수 없는 데이터 볼륨과 분석 요구 사항을 가진 비즈니스 사용자 모두의 투자가 더 많이 필요하다. 

딥 러닝에 대한 실전 문제 해결 가이드

SriSatish Ambati | InfoWorld



데이터의 고수준 패턴을 복잡한 다계층 네트워크로 모델링하는 딥 러닝(Deep learning)은 현재 빠르게 성장 중인 영역이다. 딥 러닝은 가장 보편적인 문제 해결 방법으로, 머신러닝과 인공지능 분야의 극히 난해한 문제를 해결할 잠재력을 지녔다.

마이크로소프트, 구글과 같은 기업들은 딥 러닝을 사용해 음성 인식, 이미지 인식, 3D 사물 인식, 자연어 처리와 같은 분야에서 어려운 문제를 해결하고 있다.

이미 실전에 쓰이는 딥 러닝

그러나 딥 러닝에서 유용한 모델을 구축하기 위해서는 상당한 컴퓨팅 성능이 필요하다. 얼마 전까지만 해도 컴퓨팅 비용과 높은 진입 장벽으로 인해 딥 러닝의 실용적인 활용은 제한될 수밖에 없었다. 게다가 연구자들에게도 실무적 문제에 딥 러닝을 적용하기 위한 이론과 경험이 부족했다. 제한된 시간과 자원 내에서 다른 방법이 딥 러닝보다 더 효과적인 경우가 많았다.

그러나 무어의 법칙에 따른 발전으로 컴퓨팅 비용은 급격히 낮아졌으며 혁신적인 알고리즘으로 훨씬 더 빠르고 효율적으로 모델을 학습시킬 수 있게 됐다. 경험과 지식이 누적되면서 데이터 과학자들도 딥 러닝에서 가치를 끌어내기 위한 이론과 실무적 지침을 갖췄다.

미디어에서는 미래에 가능해질 음성 및 이미지 인식 분야의 활용 사례에 초점을 맞추는 경향이 있지만, 현재 데이터 과학자들은 딥 러닝을 사용해 비즈니스 각 분야에서 실무적인 문제를 해결하고 있다. 예를 들면 다음과 같다.

- 결제 시스템 업체는 딥 러닝을 사용해 의심스러운 거래를 실시간으로 파악한다.
- 대규모 데이터센터와 컴퓨터 네트워크를 운영하는 기업은 딥 러닝을 사용해 로그 파일을 마이닝하고 위협을 탐지한다.
- 자동차 제조업체와 운송업체는 딥 러닝을 사용해 센서 데이터를 마이닝해서 부품 및 차량 고장을 예측한다.
- 딥 러닝은 대규모의 복잡한 공급망을 운영하는 기업은 딥 러닝을 사용해 생산 지연과 병목을 예측한다.

딥 러닝 소프트웨어가 증가하고 이런 소프트웨어를 효과적으로 사용하기 위한 기술이

발전함에 따라 앞으로 몇년 동안 상업용 애플리케이션이 빠르게 늘어날 것으로 전망된다.

딥 러닝의 장점

다른 머신러닝과 비교할 때 딥 러닝이 갖는 4가지 핵심 이점은 다음과 같다.

- 특징 간의 복잡한 상호작용을 탐지하는 능력
- 최소한으로 처리된 원시 데이터에서 저수준의 특징을 학습하는 능력
- 높은 기수(high-cardinality) 클래스 멤버십을 다루는 능력
- 미분류(unlabeled) 데이터를 다루는 능력

이런 4가지 강점을 통해 딥 러닝은 다른 방법으로는 불가능한 유용한 결과를 도출할 수 있고, 다른 방법보다 더 정확한 모델을 구축할 수 있으며 유용한 모델을 구축하는 데 필요한 시간을 단축할 수 있다.

보이지 않는 변수 간 상호작용 탐지

딥 러닝은 표면적으로는 보이지 않을 수 있는 변수 간의 상호작용(interactions)을 탐지한다. 상호작용이란 상호 조합되어 움직이는 두 개 이상 변수의 효과다. 예를 들어 한 약품이 젊은 여성에게 부작용을 일으키지만 노령의 여성에게는 부작용을 일으키지 않는다고 가정해 보자. 이 경우 성별과 연령을 조합한 효과를 수용하는 예측 모델이 성별만 기반으로 하는 모델에 비해 훨씬 더 효과적일 것이다.

기존의 예측 모델링 방법도 이런 효과를 측정할 수 있지만 많은 가설 테스트를 수작업으로 수행해야만 한다. 딥 러닝은 이런 상호작용을 자동으로 탐지하며 분석가의 전문 지식이나 기존 가설에 의존하지 않는다.

특히 딥 뉴런 네트워크(deep neural networks) 사용 시 비선형적 상호작용을 자동으로 생성하고 충분한 뉴런으로 임의 함수를 근사치로 계산할 수 있다.

일반적인 예측 분석 방법을 사용할 경우 분석의 성공 여부는 특징 가공(feature engineering)을 사용해 데이터를 준비하는 데이터 과학자의 역량에 크게 좌우되는데, 이 준비 단계에는 상당한 분야별 지식과 기술이 필요하다.

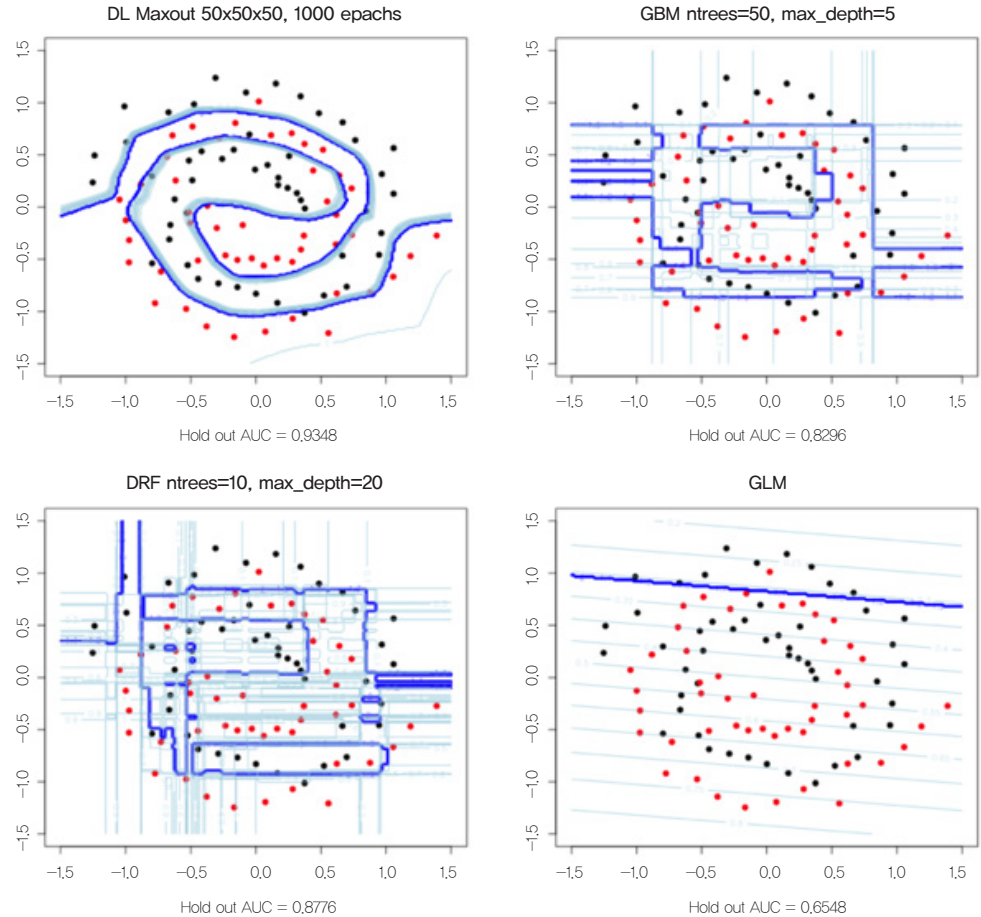
특징 가공에는 시간도 소요된다. 딥 러닝은 올바른 데이터 분포에 대한 전제 없이 최소한으로 변형된 원시 데이터를 처리해 가장 예측성 높은 특징을 자동으로 학습한다.

(그림)에서 딥 러닝의 장점을 볼 수 있다. 4개의 도표는 복잡한 패턴을 각각의 기법에서 어떻게 모델링하는 지를 보여준다. 오른쪽 아래 도표를 보면 일반화 선형 모델(Generalized Linear Model)은 데이터를 관통하는 직선을 그린다.

왼쪽 아래 랜덤 포레스트(Random Forests), 그리고 오른쪽 위의 그라디언트 부스팅 머신(Gradient Boosted Machines)과 같은 트리 기반 방법은 일반 선형 모델보다 효과적이다.

이 두 가지 방법은 하나의 직선을 그리는 대신 데이터를 관통하는 다수의 직선을 그려 모델의 '조화(fit)'를 현저하게 개선한다. 왼쪽 위의 딥 러닝은 데이터에 복잡한 곡선을 그

그림 | 딥 러닝 모델과 타 모델과의 비교



려 가장 정확한 모델을 제공한다.

딥 러닝, 미분류 데이터를 통한 학습 가능

딥 러닝은 데이터 과학자들이 ‘높은 기수 클래스 멤버십(high-cardinality class memberships)’이라고 칭하는 매우 많은 수의 개별 값을 가진 데이터 유형과 잘 맞는다.

이와 같은 문제의 실질적 예를 들면 ▲소리가 많은 단어 가운데 하나를 나타낼 수 있는 음성 인식 ▲특정 이미지가 큰 이미지 클래스에 속할 수 있는 이미지 인식 ▲제안할 최적의 항목이 많은 항목 가운데 하나일 수 있는 추천 엔진 등이 있다.

딥 러닝의 또 다른 강점은 미분류 데이터(unlabeled data)를 통해 학습할 수 있다는 점이다. 미분류 데이터에는 당면 문제와 관련된 확고한 ‘의미’가 없다. 일반적으로 태그되지 않은 이미지, 비디오, 뉴스 기사, 트윗, 컴퓨터 로그 등을 예로 들 수 있다. 사실 오늘날 정보 경제에서 생성되는 데이터의 대부분은 미분류 데이터다. 딥 러닝은 이런 데이터에서 기본적인 패턴을 감지해 비슷한 항목을 그룹으로 묶거나 조사를 위한 예외적 값을 식별할 수 있다.

딥 러닝의 결점

그러나 딥 러닝에는 단점도 있다. 딥 러닝으로 생성된 모델은 다른 머신러닝 방법에 비해 해석하기가 무척 어려울 수 있다. 이런 모델에는 많은 계층과 수천 개의 노드가 포함될 수 있는데, 이런 각 요소를 개별적으로 해석하기는 불가능하다. 데이터 과학자들은 이 모델이 얼마나 정확히 예측하는지를 기준으로 딥 러닝 모델을 평가하며 아키텍처 자체는 ‘블랙 박스(black box)’로 취급한다.

일부 비판적인 이들은 딥 러닝의 이런 측면에 대해 반기를 든다.

그러나 중요한 것은 분석의 목표다. 예를 들어 분석의 주 목표가 변동을 설명하거나 결과와 행위를 연결하는 것이라면 딥 러닝은 적합하지 않을 수 있다. 그러나 예측 변수(predictor variables)는 그 중요성을 기준으로 순위화하는 것이 가능하며 많은 경우 데이터 과학자들이 찾는 데이터는 이것에 다름 아니다. 부분적인 종속 플롯(partial dependency plots)은 데이터 과학자에게 딥 러닝 모델을 시각화할 수 있는 대안을 제공한다.

훈련 데이터 과잉 학습


또한 딥 러닝 역시 다른 머신러닝 방법과 마찬가지로 훈련 데이터를 과잉 학습한다. 이는 이 모델이 사용되는 생산 환경으로 일반화될 수도, 되지 않을 수도 있는 훈련 데이터의 특징을 알고리즘이 ‘기억’함을 의미한다. 이 문제는 딥 러닝에 국한되지 않으며 독립적인 검증을 통해 피할 수 있는 방법이 있다.

딥 러닝 모델은 복잡하므로 이를 구축하기 위해서는 상당한 컴퓨팅 성능이 필요하다. 컴퓨팅 비용이 급격히 낮아지긴 했지만 그렇다고 공짜는 아니다. 소량의 데이터를 사용한 비교적 단순한 문제라면 더 간편한 방법 대신 굳이 비용과 시간을 들여 딥 러닝을 사용할 이유가 없다.

실제 구축시 문제점, ‘복잡성’과 ‘비용’

복잡성 역시 구축 시 잠재적인 문제점이다. 넷플릭스(Netflix)는 100만 달러의 자사 상금을 차지한 딥 러닝 모델을 지나치게 높은 엔지니어링 비용을 이유로 구축하지 않았다. 테스트 데이터에서 우수한 성능을 발휘한 예측 모델이라도 실제 구현할 수 없다면 쓸모가 없다.

딥 러닝은 새로운 기술이 아니다. 딥 러닝 기술의 시초는 1950년대까지 거슬러 올라간다. 컴퓨팅 비용이 하락하고 데이터 볼륨이 커지고 기술이 향상되면서 딥 러닝에 대한 관심이 폭증한 것이다.

영역별 전문 지식이나 많은 시간을 소모하는 특징 가공, 도는 광범위한 데이터 준비 없이 막대한 규모의 데이터 집합에서 숨은 관계를 풀어낼 수 있는 딥 러닝은 갈수록 복잡해지는 비즈니스 문제 해결을 위한 매력적인 방법으로 부상했다. 

개발자를 위한 성공적인 머신러닝 핵심 요소 10가지

Alexander Gray | ITWorld

머신러닝이 데이터 깊숙이 묻혀 있는 패턴을 발견해준다는 점에서 애플리케이션의 성능을 높이고, 사용자의 수요에 더욱 민감하게 만들 수 있다는 잠재력이 있다. 제대로 고안한 알고리즘을 통해 인간의 사고와 분석적 한계를 뛰어넘어 엄청난 양의 이질적인 데이터로부터 가치를 뽑아낼 수 있다.

머신러닝은 개발자가 비즈니스에 필수적인 분석을 그 어떤 애플리케이션에도 적용하고, 고객 경험을 높이거나 제품 추천을 제공하고, 더욱 개인화된 콘텐츠를 제공하는 것까지 달성하게 해준다.

도구의 적절한 활용이 중요

아마존과 마이크로소프트와 같은 클라우드 제공업체들은 개발자가 머신러닝을 손쉽게 통합할 수 있는 클라우드 기반의 솔루션을 제공함으로써 최근 화제를 불러일으켰다. 무엇인가 굉장해 보이지만, 개발자들의 주의가 필요해 보인다.

클라우드 기반의 머신러닝 도구는 개발자가 머신러닝을 활용해 참신한 기능을 구현할 수 있도록 한다. 하지만 이런 도구를 적절하게 활용하지 않으면 형편없는 결과로 사용자에게 좌절감을 줄 수 있다. 마이크로소프트의 나이 감지 머신러닝 도구를 시험해 본 사람이라면 알겠지만, 사용 편의성이 뛰어난 만큼 중대한 정확도 문제가 대두한 바 있으며, 신뢰하거나 중요한 의사를 결정할 때 참조할 수 없는 경우도 많다.

성공을 위한 일부 핵심 요소

머신러닝을 자신의 애플리케이션에 도입하려는 개발자는 성공을 위한 일부 핵심 요소를 고려해야 한다.

1. 알고리즘의 데이터가 많으면 더욱 정확해진다. 따라서 가능하다면 부차 표본 추출은 피한다.

머신러닝 이론에는 예측 오차에 대한 매우 직관적인 특성이 있다. 쉽게 말해서 머신러닝 모델과 (이론상 최고의 오류를 달성하는) 최적 예측변수 사이의 예측 오차의 공백은 다음과 같은 세 부분으로 분류할 수 있다.

- 모델을 위한 적절한 기능적인 형태가 없기 때문에 발생하는 오차
- 모델을 위한 최적의 파라미터가 없기 때문에 발생하는 오차
- 모델에 충분한 데이터를 제공하지 않기 때문에 발생하는 오차

훈련 데이터가 제한된 경우 문제를 위해 필요한 모델 복잡성을 뒷받침하지 못할 수 있다. 통계의 기능적 법칙을 통해 우리는 가능하다면 부차 표본이 아닌, 우리가 가진 모든 데이터를 이용해야 한다.



2. 주어진 문제에 가장 적절한 머신러닝 학습법을 선택하는 것이 핵심이며, 이는 성공과 실패를 결정하기도 한다.

예를 들어, 정확도가 높은 GBT(Gradient Boosting Tree)는 업계 실무자들이 널리 활용하고 있는 인기 감독 학습 알고리즘이다. 하지만 그 높은 인기에도 불구하고 모든 문제를 위한 알고리즘으로써 맹목적으로 사용해서는 안 된다. 대신에 항상 가장 정확한 결과를 위해 데이터의 특성에 가장 적합한 알고리즘을 항상 사용해야 한다.

이 개념을 입증하기 위해 GBT와 선형 SVM(Support Vector Machine) 알고리즘 사이의 정확성을 인기있는 텍스트 범주와 데이터세트 rcv1에서 비교하는 실험을 해봐도 된다.

실제 테스트해 본 결과, 선형 SVM이 이 문제에 대한 오류율 측면에서 GBT보다 우월하다는 사실을 발견했다. 이는 텍스트 영역에서 데이터가 종종 고차원적이기 때문이다. 선형 분류자는 N개의 예시를 N-1 차원으로 완벽하게 분리할 수 있어, 단순한 모델은 이런 데이터에서 제대로 기능하게 된다. 게다가 모델이 간단할수록 한정된 수의 훈련 예제로 파라미터를 학습할 때 문제가 덜 발생하여 과적응을 방지하고 정확한 모델을 제공할 수 있다.

한편, GBT는 매우 선형적이며, 더욱 강력한 성능을 자랑하지만, 학습이 더 어렵고 이런 설정에서 과적응의 경향이 더욱 크다. 때로는 정확도가 떨어질 수도 있다.

3. 뛰어난 모델을 얻기 위해서는 방법과 그 방법에 관한 파라미터를 반드시 잘 선택해야 한다.

데이터가 엔지니어가 아닌 사람들에게는 간단하지 않을 수 있다. 현대의 머신러닝 알고리즘은 변경할 수 있는 부분이 많다. 예를 들어, 인기 있는 GBT 알고리즘 단독으로도 트리(Tree) 크기 제어 방법, 학습률, 행이나 열의 샘플 채취 방법론, 손실 함수, 조직화 옵션 등을 포함해 최대 12개의 파라미터를 설정할 수 있다.

일반적으로 프로젝트에서는 각 파라미터에 대한 최적값을 찾아 주어진 데이터 셋에 대해 가장 높은 정확도를 얻어야 하는데, 그리 쉬운 일이 아니다. 직관과 경험이 도움되긴 하지만, 데이터 엔지니어는 최선의 결과를 위해 다수의 모델을 훈련하면서 교차 검증 점수를 파악하고, 다음에 시도할 파라미터를 결정하는 일을 고민해야 할 것이다.

4. 머신러닝 모델이 데이터와 마찬가지로 수도 있다. 부적절한 데이터 수집과 정리로 일반화가 가능한 예측 가능한 머신러닝 모델을 구축하는 능력이 저하될 수 있다.

주제와 관련된 전문가와 데이터를 신중하게 검토해 데이터와 그 이면의 생성 프로세스에 대한 통찰력을 얻는 것이 좋다. 종종 이 과정으로 기록, 기능, 값, 샘플 채취 등과 관련된 데이터 품질 문제를 식별할 수 있다.

5. 데이터의 특징을 이해하고 새로운 기능을 만들어내면서 기존의 것들을 없애 향상시키면 예측 가능성을 높일 수 있다.

머신러닝의 기본적인 역할 가운데 하나는 머신러닝 알고리즘을 효과적으로 활용할 수 있는 풍부한 기능 공간에서 미가공 데이터를 표현하는 것이다. 예를 들어, 수학적 변화를 통해 기존의 기능을 토대로 새로운 기능을 개발하는 '기능 변화'는 이를 인기있는 방법이다. 그 결과 기능 공간(즉, 데이터를 특징짓기 위해 사용하는 기능의 집합)은 (여러 기능들 사이의 비선형성과 상호작용 등) 데이터의 여러 복잡한 특성을 잘 잡아내며, 이는 다음 학습 프로세스에 중요하다.

6. 기업 가치에 부합하는 적절한 목적/손실 함수의 선택은 애플리케이션의 궁극적인 성공에 중요하다.

거의 모든 머신러닝 알고리즘이 최적화 문제로 표현되고 있다. 기업의 특성에 기초해 최적화의 목적 함수를 적절히 설정하거나 조정하는 것이 머신러닝의 성공을 위한 핵심이다.

예를 들어, SVM(Support Vector Machine)은 모든 유형의 오류의 가중치가 동등하다고 가정함으로써 바이너리 분류 문제에 대한 일반화의 오류를 최소화한다.

이는 고장 감지 등 특정 유형의 오류 비용이 다른 것보다 더욱 중요할 수 있는, 비용에 민감한 문제에 적합하지 않다. 이때, 가중치를 고려하기 위해 특정 유형의 오류에 더 많은 패널티를 더함으로써 SVM 손실 함수를 조정하는 것이 좋다.

7. 적절한 훈련 및 테스트 데이터를 취급함으로써 모델을 제품에 배치할 때 테스트 데이터를 유입되는 데이터처럼 보이도록 한다.

이 점이 시간에 의존하는 데이터일 경우 얼마나 중요한지 알 수 있다. 이때, 훈련, 조율, 시험 모델을 위해 표준 교차 검증 접근방식을 사용하면 잘못되거나 정확하지 않은 결과로 귀결될 수 있다. 그 이유는 배치 단계에서 유입되는 데이터의 특성을 적절히 모방하지 않기 때문이다. 이를 바로잡기 위해서는 배치 시 모델이 사용되는 방식을 반드시 모방해야 한다. 훈련한 모델을 시간의 측면에서 더욱 새로운 데이터에 대해 검증하는 시간 기준 교차 검증을 이용해야 한다.

8. 배치 전 모델의 일반화의 오류를 이해한다.

일반화의 오류는 모델이 보이지 않는 데이터를 얼마나 잘 처리하는지를 측정한다. 모델이 훈련 데이터를 잘 처리한다고 해서 반드시 보이지 않는 데이터에 잘 일반화되는 것은 아니다. 모델의 일반화의 오류를 예측하기 위해 실제 배치 용법을 모방한 신중하게 설계한

모델 평가 프로세스가 필요하다.

일반화의 오류는 인지하지도 못한 채 교차 검증의 규칙을 위반하기 쉬우며, 교차 검증을 올바르게 수행하는 방식이 명확하지 않아 연산을 위한 지름길을 이용하려 시도할 때 자주 발생한다. 배치 성능에 대한 과학적인 예측을 얻기 위해 모델을 배치하기 전에 적절한하고 성실한 교차 검증에 주목하는 것이 중요하다.

9. 텍스트, 시계열, 공간, 그래프 데이터, 이미지 등의 비구조화 및 준구조화 데이터를 처리하는 방법을 파악한다.

대부분의 머신러닝 알고리즘은 각각 객체의 특성을 기술하는 일련의 기능으로 객체를 표현하는 기능 공간에서 데이터를 다룬다. 실제로 이런 형식으로 해당 세트에 도입되는 대신 데이터는 종종 미가공 형태로 유입되며, 머신러닝 알고리즘의 소비를 위해 반드시 바람직한 형태로 만들어야 한다. 예를 들어, 이로부터 다양한 특징을 추출하기 위해 다양한 컴퓨터 비전 기법을 사용하는 방법이나 텍스트를 특징짓기 위해 자연어 처리 기법을 적용하는 방법을 알아야 한다.

10. 기업 문제를 머신러닝 알고리즘으로 변화하는 문제를 학습한다.

사기 감지, 제품 추천, 표적 광고 등 기업에서 중요하게 여기는 일부 문제를 실제로 해결한 '표준' 머신러닝 공식이 있다. 이런 잘 알려진 문제뿐만 아니라, 덜 알려졌지만 예측 정확성이 더 높은 더욱 강력한 공식이 존재한다. 블로그와 포럼에서 일반적으로 논의하는 일련의 소규모 예시 외의 기업 문제라면 적절한 머신러닝 공식이 덜 명확하다.



자동화야말로 머신러닝의 핵심

개발자에게 있어서 이런 성공을 위한 10가지 핵심 요소를 학습하기가 그리 쉽지 않아 보일 수 있지만 낙담할 필요는 없다. 사실 개발자들은 데이터 엔지니어가 아니다. 개발자가 머신러닝이 제공하는 모든 도구를 활용할 수 있다고 생각하는 것 자체가 무리일 수 있다.

하지만 그렇다고 해서 개발자가 자신의 애플리케이션의 성능을 높이기 위해 일정 수준의 데이터 엔지니어링을 배우지 않아도 된다는 것은 아니다. 적절한 기업 솔루션과 향상된 자동화가 있으면 개발자는 높은 정확성을 보유한 머신러닝 모범

사례를 이용해 모델 구축부터 배치까지 모든 것을 할 수 있다.

자동화는 애플리케이션 내 머신러닝 확산의 핵심이다. 개발자와 밀접히 협력할 수 있는 소수의 데이터 엔지니어를 확보할 수 있다 하더라도 충분한 인력을 확보할 수는 없다. 스카이트리(Skytree)의 오토모델(AutoModel)의 사례가 모델 정확성 최대화를 위한 최적의 파라미터와 알고리즘을 자동으로 결정하는 데 도움이 될 수 있다. 사용이 간편한 인터페이스를 통해 개발자는 훈련, 조율, 시험 모델의 과정을 거치면서 통계적 실수를 방지할 수 있다.

머신러닝 프로세스 내의 자동화는 여러 측면에서 데이터 엔지니어나 개발자를 위해 인

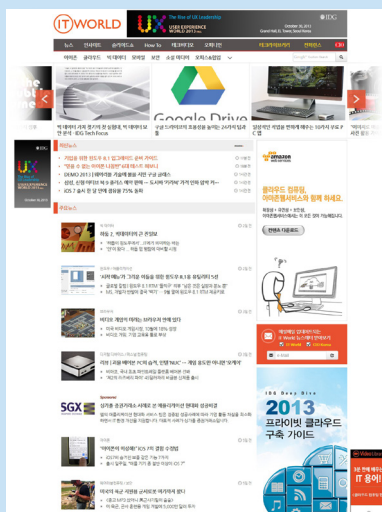
공지능의 원리를 통합하고, 알고리즘이 생각하고 학습하는 모델 구축 작업의 부담을 크게 덜어줄 수 있다.

즉, 데이터 엔지니어를 머신러닝과 분리할 수 있다는 생각이 실수이며, 특히 업무에 필수적인 모델일 경우에는 더욱 그렇다. 기초 기술의 정확함, 정교함, 확장성 등에 대한 생각 없이 적용할 수 있는 간편한 머신러닝 기능의 가능성을 인지하자.

이를 통해 높은 예측 정확성과 머신러닝이 제공해야 하는 이로 인한 높은 비즈니스적 가치를 얻을 수 있다. 게다가 애플리케이션에서 형편없는 모델을 제공하면 실제로 역효과를 낳고 사용자들 사이에서 제품이나 서비스에 대한 불신이 신속하게 쌓일 수 있다. **ITWORLD**

ITWORLD

테크놀로지 및 비즈니스 의사 결정을 위한 최적의 미디어 파트너



기업 IT 책임자를 위한 글로벌 IT 트렌드와 깊이 있는 정보

ITWorld의 주 독자층인 기업 IT 책임자들이 원하는 정보는 보다 효과적으로 IT 환경을 구축하고 IT 서비스를 제공하여 기업의 비즈니스 경쟁력을 높일 수 있는 실질적인 정보입니다.

ITWorld는 단편적인 뉴스를 전달하는 데 그치지 않고 업계 전문가들의 분석과 실제 사용자들의 평가를 기반으로 한 깊이 있는 정보를 전달하는 데 주력하고 있습니다. 이를 위해 다양한 설문조사와 사례 분석을 진행하고 있으며, 실무에 활용할 수 있고 자료로서의 가치가 있는 내용과 형식을 지향하고 있습니다.

특히 IDG의 글로벌 네트워크를 통해 확보된 방대한 정보와 전세계 IT 리더들의 경험 및 의견을 통해 글로벌 IT의 표준 패러다임을 제시하고자 합니다.