머신러닝과 경제학의 결합

17 여현규

1. 서론

머신러닝의 정의는 인간이 학습을 통해 정확도를 점진적으로 개선하는 방식을 모방하기 위한 데이터와 알고리즘의 사용에 초점을 맞추는 인공지능 및 컴퓨터 사이언스의 한 분야이다. 즉, 간단히 말하면, 몇 십년에 걸친 인간의 학습 과정을 시각화해 컴퓨터에 제공함으로써 지식의 본질을 이해하는 가이드라인을 제공하는 것이다. 1960년대부터 꾸준히 연구가 진행되고 있고, 2000년대 하드웨어의 성능이 연구를 뒷받침해줄 정도로 발전하면서, 두각을 나타내기 시작했다. 특히, 2006년 손글씨 인식 정확도를 획기적으로 높인 딥러닝 훈련 기법에 관한 논문이 발표된 후 비약적인 발전을 이뤘다. 머신러닝은 산업 전반에서 사용되고 있으며, 최근 경제학에서도 머신러닝 기법을 활용하고자 하는 연구가 활발하게 진행되고 있다. 특히, 위에서 언급한대로 개인 컴퓨터(PC)에서도 꽤 무거운 머신러닝 작업을 가능케하는 하드웨어와 더불어 용이해진 고차원(변수가 많은) 데이터 접근으로 실증분석에서 이를 활용한 다양한 연구들이 양산되고 있다.

2. 머신러닝이란?

인공지능과 머신러닝의 개념을 먼저 정립할 필요가 있다. 인공지능의 정의는 "인간과 같은 지능을 실현하기 위한 컴퓨터 시스템 및 기술"을 의미한다.

인공지능은 Strong AI와 Weak AI로 나눌 수 있다. 강한 인공지능은 인간의 지능을 가지고 생각할 수 있는 컴퓨터를 나타낸다. 즉, 의사결정이 가능한 기기를 의미한다. 약한 인공지능은 현재 모든 인공지능 분야를 포괄하는데, 특정 영역의 문제를 푸는 기술을 의미한다. 따라서, 기초 데이터와 알고리즘을 입력해야 한다.

머신러닝은 인공지능을 구현하기 위한 기술이다. 즉, 부분집합이다. 머신러닝의 큰 틀은 명시적인 프로그래밍이나 지시 없이 데이터 내부의 패턴을 자동으로 인식하는 기법이다. 머신러닝은 지도학습, 비지도 학습, 강화 학습의 3가지 유형으로 분류한다.

지도학습은 라벨이 붙어있는 데이터를 이용해서 모형을 학습시키고 스스로 모델을 평가한 뒤, 새로운 데이터의 라벨 값을 판단한다. 예를 들어, 고양이가 있는 사진을 1로 나타내고, 없는 사진을 0으로 나타낸다면, 많은 개수의 사진을 통해 고양이가 있고 없음을 구분하는 훈련을 시킨 후, 새로운 사진을 보고 고양이의 존재유무를 판단하는 일련의 과정을 의미한다.

비지도학습은 라벨이 없는 데이터를 그룹으로 나눌 때 사용한다. 그룹으로 나누는 과정을 clustering이라고 일컫는다. 위의 고양이 사진을 예로 들면, 사진에서 고양이의 유무를 판단하는 것이 아니라, 고양이와 다른 동물들을 구별하여 군집화한다. 또 다른 예로는 사진 앱에서 동일한 인물의 사진끼리 모아주는 사람 별로 사진들을 구별해 주는데 바로 비지도 학습을 사용한다.

마지막으로 강화학습은 주어진 환경 하에서 의사결정을 하고 그 결과에 따라 보상이나 패널티를 받는 방식으로 훈련하는 것을 일컫는다. 예를 들어, 생성자는 가짜를 만들고, 감별자는 가짜 여부를 가리는 방식으로 학습을 진행한다.

머신러닝 과정을 아주 간단하게 요약하자면, 훈련 데이터 셋과 머신러닝 알고리즘을 통한 모델 생성, 모델 검증, 실제 적용이라는 세가지 과정을 거친다. 모델 생성과 검증 과정에서는 다양한 방식들이 있고, 데이터 셋의 분포나 사용하고자 하는 분야에 따라 알고리즘들이 각기 다른 모델 모형과 예측력을 갖는다. 따라서, 적절한 훈련과정과 검증을 통해 적합한 모델을 찾는 것이 중요하다. 오류 함수를 통해 모델의 예측을 평가하고, 모델 최적화 프로세스를 통해 가중치 조정을 해 정확도를 올린다.

3. 머신러닝이 주목받는 이유

최근 경제학에서 머신러닝이 주목받고 있는 가장 큰 이유는 바로 예측에 있다. 기존 계량경제학은 인과관계 추론에 강점이 있다면, 머신러닝은 예측에 비교우위가 있다. 기존 계량경제학은 독립변수들과 종속변수를 연결하는 상수 값에 관심이 있다면, 머신러닝은 독립변수를 통한 종속 값 추론에 관심을 갖는다. 상관관계를 파악하고자 함을 의미한다. 머신러닝은 개념에 대한 별다른 가정 없이 특정 독립변수를 기반으로 종속변수 예측치의 정확도를 획기적으로 높일 수 있는 알고리즘을 제공한다. 다시 말해, 머신러닝은 예측 정확성을 향상시키는데 주된 목적이 있다.

조금 더 자세히 기존의 계량경제학의 방법론과 비교하자면, 다음과 같다. 계량경제학의 기본적인 작업은 표본 전체 활용해 최적화 과정을 거쳐 모수를 추정하고 가설을 검증하는 방식으로 이루어진다. 반면 머신러닝은 표본의 80%가량을 훈련에 사용하고 나머지 20%를 검증에 사용한다. 즉, 예측력을 평가하는 과정에서 사용하는 것이다.

머신러닝을 경제학에서 활용할 수 있는 방식엔 크게 네 가지가 있다.

첫째, 기존에 없던 통계를 이용한 연구가 가능하다. 다시 말해, 머신러닝은 빅데이터를 처리하는데 특화돼 있다. 빅데이터라 함은 많은 수의 표본만을 의미하는 것이 아니라 고차원 데이터를 의미한다. 즉, 변수 값이 많아 데이터 자체가 비대한 것을 나타낸다. 대표적인 예로, 데이터 수집이 어려운 국가의 도시나 작은 단위의 지역, 다양한 이유로 지속적인 설문조사가 어려운 개발도상국에서 위성사진을 이용해 경제성장 정도나 빈곤율을 파악할 수 있다. 또한, 인공위성의 특성을 이용하여 넓은 지역의 종경지에 목표 수확량을 예측할 수 있다. 이외에도 로드 뷰를 이용하여 지역 소득 수준을 높은 정확도로 예측해낸 연구도 존재한다. 또한, 기존의 데이터를 개선하는 데에

도 도움을 줄 수 있다.

둘째, 예측 정책 문제에 활용될 수 있다. 정책 시행에 따른 효과를 예측하고 누가 적절한 정책 대상자가 되어야 하는가를 파악하는데 도움이 된다. 위의 빈곤함을 파악하는 데에서 더 나아가 어떠한 집단이 지원의 대상이 되어야 하는 지를 파악할 수 있다. 이는 정책의 이질적 효과를 살펴보는데 머신러닝이 탁월하기 때문이다. 개인 수준에서 정책효과를 예측할 수 있기 때문이다.

셋째, 자연어 처리가 가능하다. 자연어란 말그대로 사람들이 일상적으로 쓰는 언어를 의미한다. 전 세계 데이터의 80%가 자연어와 같은 비정형 데이터로 이루어진 것을 고려하면, 텍스트 마이닝을 통해 자연어를 수치화 한다는 것이 얼마나 큰 자산이 될 수 있는지 짐작할 수 있다. 특히 경제학에서는 특정 기간내에 특정 단어의 빈도를 측정하여 예측을 하는 방식이 주로 사용된다. 우리나라의 예를 살펴보면, 뉴스 기사를 이용해 금융 시장의 감성 지표를 구축한 뒤, 이 지표가 국채 금리나 환율 등을 예측하는 데에 사용될 수 있는 지 검증하는 연구가 수행됐다.

넷째로 이론의 검증 과정에서 사용할 수 있다. 이론을 검증하기 위해서 이론의 대상이 모형 또는 예측과 얼마나 일치하는 지를 확인해야한다. 머신러닝은 이 과정에서 벤치마크를 제공한다. 기존 이론이 제시하는 예측을 벤치마크와 비교하여 어느 정도의 설명력을 지니는지 측정할 수 있다.

4. 머신러닝의 한계

위의 장점들에도 불구하고, 머신러닝은 현재 몇 가지 문제들이 한계점으로 지적되고 있다.

첫째, 블랙박스 모델이다. 다시 말해, 어떻게 데이터를 분석했는지 구조적으로 명확하게 설명할 수 없는 경우들이 많다. 인과관계를 나타낼 수 없다는 것이다. 이를 해석 가능성이라고도 하는데, 해석 가능성이 보장되어야 해당 모델을 새로운 데이터 또는 더 큰 규모의 데이터에 적용했을 때 생길 수 있는 문제들에 대해 쉽게 파악할 수 있다. 예측력이 높지만, 높은 이유를 논리적으로 설명할 수 없기 때문에 거부감을 일으킬 수 있다.

둘째, 차별 또는 공정성에 대해 자유롭지 않다. 머신러닝은 데이터셋을 활용한 학습을 기반으로 추론을 진행한다. 따라서, 편향성을 지닌 데이터를 표본으로 한다면, 잘못된 예측을 내놓을 가능성이 높다. 실제로, 아마존의 경우 인공 지능에 기반한 사내 채용 프로그램을 활용했는데, 이는 여성 차별적인 것으로 드러나 사용을 중단하기도 했다.

셋째, 조작가능성이 있다. 위 문단에서 지적된 문제점의 연장선으로 생각할 수 있다. 공공정책에 활용할 때, 데이터나 알고리즘을 조작한다면, 의도한 평가를 내놓도록 수정할 수 있다.

넷째, 현재 경제학에서 활용되는 머신러닝 알고리즘들은 모두 컴퓨터 공학분야에서 개발된 기성품들이다. 즉, 경제학적 특성을 고려하지 않았다. 기존의 방식들은 분석하고자 하는 경제학 모델에 적절한 머신러닝 모델을 적용시켜 예측을 진행했다. 보다 나은 분석을 위하여 경제학 특유의 방법론을 고려한 모델이 개발된다면 빅데이터와 결합되어 경제학계의 방향 자체를 바꿀 여지

가 있다.

5. 머신러닝의 활용 방안

머신러닝의 경제학적 활용에서 가장 큰 약점으로 작용하는 한계점은 인과관계를 설명하는데 한계가 있다는 점이다. 예를 들어, 경찰관과 범죄율의 관계를 생각해보면, 경제학적 관점은 경찰관의 증원이 어느 정도의 범죄율 하락을 가져올 것인가에 대한 인과성이다. 하지만, 통계적으로 번죄율이 높은 지역일수록 많은 경찰관들이 배당되는 경향이 있음으로, 두 변수는 양의 관계를 갖는다. 따라서, 이 통계를 이용하여 머신러닝 학습을 진행하면 인과관계에 대한 해답을 내놓지 못한다. 머신러닝을 통해 해결할 수 있는 궁금증은 범죄율에 따른 경찰관 배치 인원 추론에 불과하다. 따라서, 머신러닝 기법을 인과관계를 밝히기 위해 직접적으로 사용하는 것이 아니라, 기존의 계량 경제 방법론들을 머신 러닝을 통해 개선하는 방식을 택할 수 있다. 가령 변수 간 관계를 분석하는 여러 머신러닝 알고리즘을 통해 이제까지 임의적이었던 계량 모형 설정에 정확성을 가미할 수 있다. 고차원 데이터에서 변수 간 관계를 일일이 추론하는 것은 불가능 했지만, 머신러닝의 도움을 받으면 해결할 수 있다.

실생활의 빅데이터를 통해 경제 현상 분석 및 예측을 통한 정책, 전략 수립, 실증 분석이 가능하다. 넓은 범용성을 통해 분석하고자 하는 분야와 방향에 따라 데이터만 수집된다면 어떠한 상황에서도 예측할 수 있다. 다시 말해, 어떤 데이터를 활용하느냐, 어디에 접목을 시키느냐에 따라 머신러닝 예측 모델의 그 활용도가 무궁무진하다. 사소한 개인 사용자의 검색 기록을 통한 광고 노출부터 회사 단위의 알고리즘 트레이딩, 신용카드 데이터 등 민간거래 빅데이터를 활용한 지역별 소비수준 파악 및 지역 생산, 고용효과 비교, 국가 정책 차원에서 둘째 출산 결정 요인 분석 등이 가능하다.

6. 결론

머신러닝은 높은 예측력이라는 큰 장점을 토대로 다양한 분야에서 각광받고 있다. 이에, 경제학에서도 그 중요성이 커지고 있다. 따라서, 기본 개념 및 활용성, 한계에 대해 살펴보았다. 머신러닝은 다양한 경제현상 분석 및 예측, 기존의 계량 모델 검증 등에 폭 넓게 활용할 수 있다. 특히, 목적에 따른 다양한 모델 생성이 가능하다는 점에서 경제학의 대부분 분야에서 활용될 수 있다. 하지만, 경험에 따른 학습 및 추론이 머신러닝 알고리즘의 주를 이루기 때문에 인과관계를 설명할 수 없다는 점에서 한계가 있다.

기존 분석방법으로는 해결할 수 없는 복잡한 문제 해결 툴, 기존 실증분석 보완, 빅데이터를 통한 가격 예측, 정책 평가 등, 머신러닝은 기존 경제학의 분석 방법을 근본적으로 바꿀만한 잠재력이 있다고 평가되고 있다.

참고문헌

https://blogs.nvidia.co.kr/2016/08/03/difference_ai_learning_machinelearning/ (인공 지능과 머신 러닝, 딥 러닝의 차이점을 알아보자)

https://www.ibm.com/kr-ko/cloud/learn/machine-learning (머신 러닝)

머신러닝을 이용한 경제분석 (박기영, 고정원)

"The Evolving U.S. Occupational Structure: A Textual Analysis" (Enghin Atalay)

빅데이터를 활용한 경제 데이터 분석 사례 및 방법론 연구 (국회예산정책처)

머신러닝을 활용한 정책설계: 출산 결정요인을 중심으로 (정재현)