한국경제학보 제26권 제2호
The Korean Journal of Economics Vol. 26, No. 2 (Autumn 2019)

머신 러닝을 이용한 경제분석

박 기 영* · 고 정 원**

요 약

본 논문은 경제학 전공자를 대상으로 인공 지능을 구현하는 핵심 기법인 머신 러닝의 개념과 주요 방법론, 경제학과 경제에 미치는 영향을 개괄적으로 소개하고자 한다. 먼저 머신 러닝의 주요 범주인 지도 학습, 비지도 학습, 강화 학습의 개념을 소개하고 기존 계량경제학 접근법과의 차이점을 설명한다. 그리고 학계와 산업계에서 널리 연구되고 활용되는 지도 학습 분야에서 분류 및 회귀를 위해 사용되는 주요 방법론을 예를 통해 설명한 뒤 머신 러닝 기법이 활용된 경제학 분야의 최신 연구들, 노동시장에 미치는 영향, 데이터의 가치를 둘러싼 논쟁에 대해 살펴본다.

주제분류 : B030104, B039900
핵심 주제어 : 인공 지능, 머신 러닝, 지도 학습, 빅데이터

Ⅰ. 서 론

많은 사람들이 인공 지능(AI, artificial intelligence)에 대해 처음 관심을 갖게 되고 인공 지능의 시대가 그리 멀지 않다고 느꼈던 시기는 2016년 3월 알파고 리(AlphaGo Lee)가 이세돌 9단에게 4승 1패로 승리했을 때였다.[1] 하지만 대중의 인식과 달리 인공 지능에 대한 연구는 이미 두 번의 인공 지능 겨울(AI winter)을 겪은 뒤 세번째 부흥기에 접어 든 상태이

* 교신저자, 연세대학교 경제학부, e-mail: kypark@yonsei.ac.kr
** 연세대학교 경제학부, email: kojw_ye@yonsei.ac.kr

1) 그 이후 알파고 리를 개선한 알파고 마스터(AlphaGo Master)는 중국의 커제 기사에게 3승 무패로 승리했다. 알파고 제로(AlphaGo Zero)는 72시간 독학 후 알파고 리에게 100승 무패, 40일 동안 독학 후 알파고 마스터에게 89승 11패로 승리했고 알파제로(AlphaZero)는 알파고 제로에 60승 40패로 승리했다.

며 이번 시기에는 이전의 전철을 밟지 않고 우리 사회와 경제를 전면적으로 바꿀 수 있을 것이라 기대되고 있다.[2] 예를 들어 자동차 자율 운행의 경우 기술적으로 거의 완성 단계라 할 수 있으며 오히려 관련 인프라, 규제 등의 미비로 지체되고 있다고 볼 수 있다. 또한 의료 진단, 신약 개발, 빅데이터 분석 등 다양한 분야에서 활발하게 연구 및 활용되고 있는 인공 지능 기술이 우리 사회를 획기적으로 변화시킬 것으로 보인다. 이런 예상은 경제학자들에게도 많은 질문을 던지고 있다: 이렇게 빠른 인공 지능 기술의 발전은 과연 우리가 살고 있는 사회의 경제를 어떻게 바꿀 것인가? 노동시장은 어떻게 변화할 것인가? 경제뿐만 아니라 경제학은 어떻게 바뀔 것인가? 또는 인공 지능을 경제학 분야에서 어떻게 활용할 것인가?

본 논문은 인공 지능 분야에 대한 사전 지식이 없는 경제학 전공자를 대상으로 인공 지능, 그리고 그 하위 개념이라 볼 수 있는 머신 러닝(ML, machine learning)의 개념, 주요 방법론 및 활용, 경제와 경제학에 미치는 영향을 개괄적으로 논의하고자 한다. 이를 통해 머신 러닝이 경제 및 경제학에 어떤 영향을 주는가에 대해 궁금해하는 독자들에게 유용한 정보를 제공하고 위에서 제기한 여러 질문들에 대한 단초를 제공하는 것을 목표로 한다. 또한 이를 통해 머신 러닝에 대한 연구자들의 관심을 환기하고 이를 이용한 경제학 연구가 활발해질 수 있는 계기가 되고자 한다.[3]

제Ⅱ장에서는 인공 지능과 머신 러닝의 개념을 살펴본 후 머신 러닝의 세 가지 범주인 지도 학습(supervised learning), 비지도 학습(unsupervised learning), 강화 학습(reinforcement learning)이 무엇인지 설명한다. 그리고 머신 러닝이 기존 계량경제학의 방법론과 어떤 차이가 있는지 설명

2) 인공 지능 연구에 대한 관심은 1950년 후반에 미국과 러시아의 군사안보 경쟁에서 촉발되었다. 러시아는 실시간 통역 시스템 개발을 추진했고 미국도 국방부와 CIA에서 비슷한 투자를 했지만 현실 적용이 어렵다는 것이 알려지며 미국 국방부 산하 국방고등연구계획국(DARPA, Defense Advanced Research Projects Agency)이 지원을 대폭 삭감하면서 70년대 초에 1차 인공 지능 겨울이 시작되었다. 이후 70년대 말, 80년대 초에 전문가의 지식 습득과 의사결정 방식을 모방하는 컴퓨터 시스템인 전문가 시스템(expert system)이 등장하면서 인공 지능이 다시 주목을 받았다. 그러나 예상보다 나쁜 연구 실적에 투자자들이 지원을 끊으면서 두번째 인공 지능 겨울이 시작되었다. 향후 인공 지능 발전에 대한 전문가들의 예상에 대해서는 Ford(2018)와 Lee(2018)를 참고하시오.

3) 머신 러닝 분야가 각광을 받으면서 경제학 분야에서도 활용되기 시작했는데 그런 맥락에서 널리 읽히는 서베이 논문으로는 Varian(2014), Mullainathan and Spiess(2017), Athey and Imbens(2019)를 들 수 있다.

한다. 제Ⅲ장에서는 학계와 산업계에서 널리 연구, 활용되고 있는 지도 학습의 주요 기법들을 소개한다. 구체적으로, 분류(classification)와 회귀(regression)를 위해 쓰이는 KNN(K-Nearest Neighbors), 의사결정 트리(decision tree), 서포트 벡터머신(SVM, support vector machine), 규제된 회귀(regularized regression) 기법들을 예를 통해 살펴본다. 제Ⅳ장에서는 이미 머신 러닝 기법이 활용되기 시작한 경제학 분야의 관련 연구들을 살펴본다. 제Ⅴ장에서는 인공 지능이 우리 사회에 미치는 영향을 이해하기 위해 인공 지능과 관련된 최근의 논의들을 소개한다. 먼저 인공 지능이 노동시장에 미치는 영향을 이해하기 위해 알고 있어야 할 배경 지식을 소개한다. 둘째, 엄청난 수익을 올리고 있는 구글, 페이스북 같은 거대 테크놀로지 기업들의 데이터가 사실 이들 회사의 무형 자본이 아니라 사용자들이 제공한 노동의 대가라는 주장들이 최근 학계와 정치계에서 나오고 있는 가운데 이에 대한 논리와 반론을 살펴본다. 마지막으로 머신 러닝이 경제학자들의 주된 관심사 중 하나인 인과관계 추론(causal inference)과 어떤 관계를 가지며 어떻게 기여할 수 있는지에 대한 최근 논의를 간단하게 소개한다. 결론에서는 향후 머신 러닝이 가져야 할 바람직한 성질들에 대해 논의한다.

Ⅱ. 머신 러닝의 개념 및 유형

머신 러닝의 개념과 종류, 기법 등을 살펴보기 전에 인공 지능의 개념을 먼저 살펴봄으로써 인공 지능과 머신 러닝의 개념을 명확히 하자. 인공 지능에 대한 정의는 매우 다양하지만 공통적인 정의는 “인간과 같은 지능을 실현하기 위한 컴퓨터 시스템 및 기술”을 의미한다. 인공 지능에는 두 가지 범주가 있는데 강한 인공 지능(strong AI)과 약한 인공 지능(weak AI)이다. 강한 인공 지능은 영화 ‘터미네이터’에 나온 기계 인간처럼 사람처럼 행동하고 다양한 업무를 수행하는 인공 지능을 가리키는데 현재까지 실현된 사례는 없다. 최근 일반 인공 지능(AGI, Artificial General Intelligence)에 대한 논의가 많이 되고 있는데 바로 강한 인공 지능에 대한 것이다. Ford(2018)에 따르면 앤드류 응(Andrew Ng) 등 이 분야 석학들은 일반

인공 지능의 개발은 아직 요원하다고 본다.[4] 반면 약한 인공 지능은 좁은 범위, 또는 단일한 업무를 처리하는 인공 지능을 의미하며 현 단계에서 상용화되어 있는 모든 인공 지능 기술은 이 범주에 속한다. 스팸 메일을 분류하는 작업, 사진을 분류하는 작업을 예로 들 수 있다. 약한 인공 지능에 대한 연구는 활발하게 진행되어 왔으며 컴퓨터 비전(computer vision), 자연어 처리(NLP: Natural Language Processing), 질병 진단 및 신약 개발 등 이미 여러 분야에서 폭넓게 활용되고 있다.

머신 러닝은 인공 지능을 구현하기 위한 기술이라 할 수 있다. 즉 인공 지능 개념의 부분 집합이라 볼 수 있으며 인공 지능을 구현하기 위한 방법론이라 볼 수 있다. Arthur Samuel(1959)은 머신 러닝을 "명시적인 프로그래밍 없이 컴퓨터가 자율적으로 학습하는 기능에 대한 연구 분야(field of study that gives computers the ability to learn without being explicitly programmed)"로 정의한다. Mitchell(1997)은 보다 현대적인 정의를 제안하는데 "어떤 컴퓨터 프로그램이 T(task)라는 작업을 수행하고 P(performance measure)라는 성능 측정 결과 성능이 E(experience)에 따라 향상된다면 이 프로그램은 E를 통해 학습한다고 할 수 있다."[5] 쉽게 설명하면 머신 러닝은 "명시적인 프로그래밍이나 지시 없이" 데이터 내부의 패턴을 자동적으로 인식하는 기법으로 이해할 수 있다.[6] 머신 러닝은 크게

---

4) 인공 지능 여부를 판별하는 테스트로는 영국의 수학자 앨런 튜링(Alan Turing)의 이름을 딴 "튜링 테스트"가 있다. 예를 들어, 컴퓨터를 벽 너머에 두고 인간과 대화를 하는데 벽 너머 컴퓨터의 반응을 인간의 반응이라 생각한다면 해당 컴퓨터는 인공 지능에 해당한다. 최근에는 일반 인공 지능을 판별하는 테스트로 애플의 공동창업자인 스티브 워즈니악(Steve Wozniak)이 제안한 "커피 테스트"가 많이 논의되고 있다. 이는 "모르는 집에 들어가서 커피를 끓일 수 있는가?" 여부로 판별한다. 비록 단순한 작업이라 할 수 있지만 위 작업을 수행하기 위해서는 커피와 커피가 있는 곳을 식별하고, 물을 끓여야 하는 등 다양한 작업을 수행해야만 완수할 수 있다.

5) "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."

6) 머신 러닝의 훈련을 위해서는 데이터가 많을수록 좋으며 최근 머신 러닝의 발전은 빅데이터의 구축과 활용, 그리고 컴퓨터 연산 능력의 발전과 불가분의 관계에 있다. Lee(2018)는 데이터를 '21세기 천연자원'이라 표현할 정도이며 Agarwal et al.(2018)의 지적처럼 프라이버시에 대한 규제 강도에 따라 데이터가 머신 러닝에 활용되는 정도가 달라질 수 있다. 예를 들면, 중국의 경우 상대적으로 프라이버시 보호보다는 데이터의 활용에 강조점을 두며, EU는 프라이버시 보호를 강하게 강조한다. 미국은 중국과 EU의 중간에 있다고 볼 수 있다. 빅데이터는 양(volume)뿐만 아니라 속도(velocity)에 대한 개념도 포함하고 있다. 예를 들어, 구글 트렌드(trends.

지도 학습, 비지도 학습, 강화 학습으로 분류하는데 아래에서 차례로 살펴보기로 한다.

1. 머신 러닝의 유형

(1) 지도 학습(Supervised Learning)

지도 학습이 비지도 학습과 구별되는 가장 큰 특징은 라벨(label)이 붙어있는 데이터(labeled data)를 이용한다는 것이다. 데이터의 형태는 (Y_i, X_i), $i = 1,2,...,N$인데 Y_i는 라벨(label), X_i는 특성(features)이다. 예를 들어 스팸 메일 여부가 0과 1로 표시된 변수가 Y_i, 스팸 메일의 전형적인 특징들(특정 문구 유무, 이메일 주소 등)이 X_i에 해당한다. 사진 속에 고양이가 있는지 여부를 가리는 작업의 경우 사진 속에 고양이가 있는지 여부가 0과 1로 표시된 변수가 Y_i, 고양이 얼굴의 특성에 대한 픽셀 정보 등이 X_i가 된다.

지도 학습의 목적은 X_i를 통해 Y_i를 예측하는 것인데 가장 간단한 경우의 지도 학습 예를 들어보자. 100만명에 대한 대출 심사 데이터 (Y_i, X_i)가 있다고 하자.

(1) 데이터를 입수한 뒤 누락된 값(missing value)를 제거하고 필요에 따라 특성 변수들을 0과 1, 또는 -1과 1사이의 변수로 스케일링하는 등의 전처리(preprocessing)를 한다.
(2) 무작위 추출을 해서 80만명으로 구성된 훈련 세트(training set)와 20만명으로 구성된 검증 세트(test set)을 구성한다.
(3) 훈련 세트를 로짓, SVM, 의사결정 트리 기법 등 다양한 모형으로 추정한다. 머신 러닝에서는 이 과정을 모형을 훈련시킨다고 표현한다.
(4) 추정된 모형의 성능을 검증 세트를 이용해서 평가한다.[7]

---

google.com)를 이용해서 실시간 경기예측을 하는 나우캐스팅(nowcasting)이 있다. MIT의 Billion Prices Project(http://www.thebillionpricesproject.com/)는 원래 아르헨티나의 물가를 더 정확하게 추정하기 위해 시작되었으나 현재는 온라인 상점의 가격 정보를 이용하여 미국의 실시간 물가상승률(real-time inflation)을 제공하고 있다.

(5) 가장 우수한 추정 기법을 정하고 새로운 대출 신청자의 를 이용해서 대출 허가 여부를 결정한다.[8]

요약하면 지도 학습은 라벨이 붙어 있는 데이터를 이용해서 모형을 학습시키고 검증한 뒤, 새로운 데이터를 보여주고 해당 데이터의 라벨을 예측하는 것이다.

라벨이 붙은 데이터를 얻는 방법으로는 은행의 기존 대출 심사 자료와 같이 라벨이 이미 있는 경우, 그리고 사람이 직접 입력하는 경우가 있다. 후자의 경우 최근 들어 온라인 상에서 노동력을 제공받는 크라우드 소싱(crowd-sourcing)이 널리 활용되고 있다. 2005년에 시작된 아마존 미케니컬 터크(Amazon Mechanical Turk)는 일종의 온라인 노동시장으로 요청자(requesters)가 온라인 상에 필요한 업무(예를 들어, 설문조사, 데이터 검증 등)를 올려 놓으면 약속된 대가를 받고 사람들이 노동력을 제공하게 만들어 놓은 사이트이다. 이미지넷(ImageNet) 데이터베이스는 2006년 스탠퍼드 대학 페이페이 리(Fei-Fei Li) 교수의 주도로 시작되었는데 현재 무엇에 대한 사진인지 라벨이 붙은 1,400백만개의 이미지가 관리되고 있다. 라벨은 참여자들이 자발적으로 손수 붙인 것이며 머신 러닝 훈련 등 학문적 목적으로 사용되고 있다.[9]

지도 학습의 알고리듬으로는 로짓, 나이브 베이즈(naïve Bayes), KNN(K-Nearest Neighbors), 의사결정 트리(decision tree), SVM(Support Vector Machine), 규제된 회귀(regularized regression), 랜덤 포레스트(random forest), 신경망(neural net), 딥러닝(deep learning) 등이 있다. 이들 알고리듬은 오픈소스로 제작된 패키지가 제공되어 비교적 손쉽게 이용할 수 있다. 예를 들어, 파이썬(python)에서는 사이킷런(scikit-learn), 텐서플로(TensorFlow), 케라스(Keras) 등이 있고 R에서는 CARET(Classification and Regression Training)을 많이 사용한다.

---

7) 평가 방식은 아래 3장에서 설명한다.

8) 독자의 이해를 돕기 위해 가장 단순한 형태의 작업 단계를 설명한 것이며 실제로는 교차검증(cross-validation), 초매개변수 조정(hyperparameter tuning) 등의 추가 작업이 필요하며 앙상블 기법(ensemble method)를 사용할 수도 있다. 이는 아래에서 설명한다.

9) http://www.image-net.org/

(2) 비지도 학습(Unsupervised Learning)

비지도 학습은 지도 학습과 달리 라벨이 없는 데이터, 즉 Y_i없이 X_i만 있는 데이터를 군집화(clustering)하거나 차원 축소(dimensionality reduction) 할 때 사용한다. 구글 포토 등 사진 앱에서 누구의 사진이라고 이름을 붙이지 않아도 앱에서 자동으로 사람 별로 사진들을 분류해 주는데 바로 비지도 학습의 전형적인 예이다. 이외에도 의료 영상을 판독한다든가 차원 축소를 통해 DNA 정보를 축약하는 목적 등에 사용한다.

인공 지능 관련 세계적 석학들은 비지도 학습이 지도 학습에 비해 현재까지 학문적, 기업적 측면의 활용은 적으나 일반 인공 지능(AGI)의 개발과 관련해서 매우 중요하다고 본다.[10] 이는 지도 학습과 비교해 볼 때 비지도 학습의 원리가 인간의 인지 방식과 더 유사하기 때문이다. 예를 들면, 아기가 고양이를 인식하는 방식은 수많은 데이터(실제 고양이나 고양이 사진)를 통해 훈련한 뒤 고양이를 구별하는 것이 아니라 고양이를 몇 번 보고 약간의 시행착오를 겪은 뒤 고양이를 다른 동물과 구별하게 되는데 이는 비지도 학습의 원리와 더 가깝다. 그리고 이런 원리와 관련해서 비지도 학습이란 용어 자체가 올바르지 않다는 주장도 있다. 페이스북 AI 수석 엔지니어인 얀 르쿤(Yann LeCun)은 비지도 학습이란 용어의 의미가 모호하다며 자기 지도 학습(self-supervised learning), 또는 예측 학습(predictive learning)으로 대체하자고 주장한다.

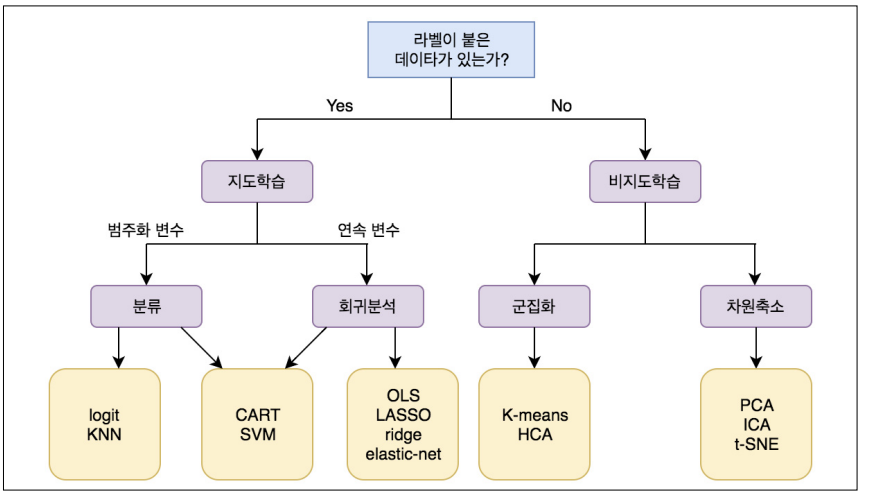
군집화에 많이 쓰이는 알고리듬으로는 k-means, HCA(Hierarchical Cluster Analysis), 기대값 극대화(expectation maximization) 등이 있고 시각화, 차원 축소와 관련해서는 주성분 분석(PCA, Principal Component Analysis), ICA (Independent Component Analysis), kernel PCA, LLE (Locally-Linear Embedding), t-SNE(t-distributed Stochastic Neighbor Embedding) 등이 많이 사용된다.

아래 〈그림 1〉은 지도 학습과 비지도 학습의 구분, 그리고 관련 알고리듬을 요약해서 보여준다. 먼저 라벨이 붙은 데이터 여부에 따라 지도 학습과 비지도 학습으로 분류된다. 지도 학습의 경우 다시 분류와 회귀로 구분되고, 비지도 학습의 경우 군집과 차원 축소로 구분된다. 그 아래에는 각 영역에서 대표적인 알고리듬을 표시했으며 CART와 SVM은 분류와 회귀분석

10) Ford(2018)에 나온 석학들과의 인터뷰에서 공통적으로 언급된 사안이다.

모두에 사용될 수 있다.

〈그림 1〉 지도 학습과 비지도 학습의 구분(Supervised learning and unsupervised learning)



(3) 강화 학습(Reinforcement Learning)

주어진 환경 하에서 의사결정을 하고 그 결과에 따라 보상 또는 벌칙을 받는 방식으로 훈련을 하는 것을 강화 학습이라 하며 원리상 마코프 프로세스(Markov process)에 기반한 동태적 최적화(dynamic optimization)의 개념과 일치한다. 알파고가 대표적인 강화 학습 모형의 예이며 이미지와 텍스트의 생성, 게임 등에서 활용되고 있다. 특히 강화 학습에서 최근 자주 쓰이는 생성적 적대 신경망(GAN: Generative Adversarial Network)은 생성자(generator)와 감별자(discriminator)가 서로 경쟁하게 만드는 방식으로 학습한다.[11] 예를 들어, 생성자는 가짜 뉴스 텍스트를 만들고 감별자는 가짜 뉴스 여부를 가리는 방식으로 모형을 훈련시킨다. 최근에는 이 방법을 이용해서 정치인의 가짜 연설 동영상을 만들 수도 있게 되어 가짜 뉴스인지 식별하기 어려운 '딥페이크(Deepfake)'에 대한 우려도 나오고 있다.

11) Goodfellow et al.(2014)에서 처음 제안되었다.

2. 기존 계량경제학 방법론과의 차이

머신 러닝의 방법론, 특히 지도 학습의 방법론은 기본적으로 계량경제학의 방법론과 근본적 차이는 없지만, 목적, 진행 방식, 용어에서 차이를 보인다. Varian(2014)은 기존 통계학과 계량경제학의 분석 목적은 예측, 요약, 추정, 가설 검정으로 이루어져 있으나 머신 러닝은 예측(prediction)의 문제를 다룬다고 본다. 통상적으로 계량경제학의 기본적인 작업은 표본 전체를 이용해서 평균자승오차나 우도 함수를 최적함으로써 모수를 추정하고 가설을 검증하는 방식으로 이루어진다. 반면 머신 러닝 작업에서는 통상적으로 표본의 70-80% 정도를 추출해서 훈련 세트를 구성하고 이 훈련 세트를 대상으로 추정(모형을 훈련)한다. 그리고 얻어진 결과를 검증 세트라 부르는 나머지 표본을 이용해서 예측하고 평가한다. 예측의 문제에서 가장 중요한 것은 좋은 표본 외 예측(out-of-sample prediction)을 하는 것인데 상대적으로 복잡한 모형일수록 훈련 세트 내 예측에 비해 검증 세트 내 예측의 성능이 떨어지는 문제가 발생한다. 이를 과대 적합(overfitting)이라 부른다. 이 문제를 부분적으로 해결하기 위해 모형의 복잡도(complexity)에 대해 패널티를 주는데 이를 규제(regularization)라 부른다. 또 다른 용어의 차이로 계량경제학에서 흔히 사용하는 설명변수(explanatory variables) 또는 독립변수(independent variables)란 용어 대신에 머신 러닝 분야에서는 특성(features)이라 부른다. 그리고 피설명변수, 종속 변수는 흔히 타겟 변수(target variable)라 한다.

표본을 사용하는 방식도 상이하다. 위에서 언급한 바와 같이 계량경제학에서는 일반적으로 표본 전체를 추정에 사용하지만 머신 러닝은 훈련 세트와 검증 세트로 나누어서 각각 훈련(추정)과 표본 외 예측을 위해 사용한다. 이때 검증 세트 추출의 자의성을 줄임으로써 모형의 성능을 제고하는 방법 중 하나가 k겹 교차 검증(k-fold cross-validation)이다. 예를 들어 k=5인 경우 표본을 다섯 개의 서브 샘플(1,2,3,4,5)로 나눈 후 첫번째 단계에서는 서브 샘플 1,2,3,4를 훈련 세트, 5를 검증 세트로 사용하고 두번째 단계에서는 2,3,4,5를 훈련 세트, 1을 검증 세트로 사용하는 방식이다. 이렇게 5번의 추정을 종합하여 모형을 평가한다. 이런 방식으로 추정할 경우 검증 세트의 자의성을 줄일 수 있는데 Varian(2014)은 기존의 계량경

제학 기법에서도 교차 검증을 적극 사용할 것을 권장하고 있다.

또 한 가지 차이점은 초매개변수 조정(hyperparameter tuning)의 필요성이다. 좋은 표본 외 예측 능력을 갖추기 위해서 초매개변수 조정이 필요한데 초매개변수에 대해서는 아래에서 예를 들어 설명을 한다.

Ⅲ. 방법론

Ⅲ장에서는 위에서 소개한 머신 러닝의 세 가지 범주 중에서 상업적으로나, 학술적으로 가장 널리 쓰이고 있는 지도 학습의 대표적인 기법들을 예를 통해 직관적으로 설명한다. 그리고 모형을 평가하는 방법에 대해서도 설명한다.[12)]

1. 분류 및 회귀

지도 학습의 기본적인 목적은 분류와 회귀인데 쉽게 이해하자면 데이터의 라벨, y_i가 이산형 변수(discrete variable)인 경우 분류, 연속형 변수(continuous variable)인 경우 회귀로 볼 수 있다.

(1) 로짓

로짓은 경제학에서 사용되는 대표적인 분류 및 회귀 기법 중 하나로 로짓 함수를 이용해 관측치가 각 카테고리에 해당할 확률을 구한다. 이때 특성들의 값을 0과 1 사이로 정규화하는 스케일링(scaling)같은 전처리 과정이 불필요하며, 모델링 및 해석이 쉬워 널리 쓰인다. 하지만 많은 변수를 한번에 다루지 못하고 과대 적합의 문제가 발생하기 쉽다는 단점이 있다. 또한 비선형 문제를 다루지 못한다.

로짓은 분류 카테고리 수에 따라 이진 분류와 다중 분류로 나뉜다. 이진 분류의 경우, 로짓은 추정된 확률이 0.5가 되는 지점을 기준으로 결정 경계(decision boundary)를 구하여 관측치를 분류한다. 예를 들어, 로짓을 이

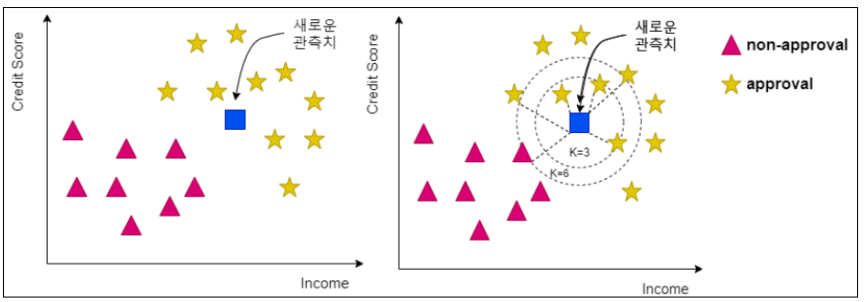
12) 머신 러닝 분야에서 널리 알려진 용어나 기법을 설명할 때는 Efron and Hastie(2016)과 Géron(2017)을 주로 참고하였다.

용해 새로 관측된 메일이 스팸 메일일 확률을 구한다고 하자. 이때, 결정 경계 이상의 값을 가지면 스팸 메일로, 그 이하의 값을 가지면 일반 메일로 분류한다. 다중 분류는 와인의 종류를 분류하는 것과 같이 관측치가 두 가지 이상의 카테고리로 분류될 수 있을 때 쓰인다.

(2) KNN

KNN(K-Nearest Neighborhood)은 비모수적(non-parametric) 분류 기법으로, 기존 데이터를 단순 기억하여 학습하는 사례 기반 학습이다. 이 기법은 비선형 데이터에도 쓰일 수 있으며 회귀 기법과도 함께 쓰일 수 있다.

〈그림 2〉 KNN의 분류과정(Classification using KNN)



〈그림 2〉에 보이는 ☆와 △는 기존 관측치이며 ☆는 신용카드 발급을 승인 받은 사람들, △는 승인을 받지 못한 사람들의 소득과 신용 점수를 보여준다. 이때 새로운 관측치 □가 어떤 그룹에 속할 지 예측하기 위해서는 먼저 기존 관측치와의 거리를 구해야 한다. 신용카드 발급 승인을 받은 사람의 소득과 신용 점수가 상대적으로 높다고 할 때, □의 소득과 신용 점수를 이와 비교하여 유사도(similarity)를 측정하는 것이다. 그 후, 예측할 관측치와 가장 가까운 K개의 이웃을 찾고, 이 이웃들이 속한 그룹을 대상으로 다수결 투표를 한다. K=6인 경우, □와 가장 가까운 이웃들은 승인을 받지 못한 사람 1명과 승인을 받은 사람 5명이다. 따라서 다수결 투표에 따라 □은 신용카드 승인을 받은 그룹으로 예측 분류된다.

이 기법에서 이웃 간 유사도를 측정할 때 기준이 되는 특성들의 수가 적을수록 분류의 정확도가 높아진다. 고려할 특성의 수가 많아지면 과대 적합

의 문제가 발생하기 때문에 학습에 필요한 데이터도 그만큼 커져야 한다. 새로운 관측치를 예측할 때 학습한 모든 데이터를 사용하는 이 기법의 특성상, 기존 데이터가 커질 경우 예측에 많은 시간과 컴퓨터 자원이 소요된다. 한편 학습 자체는 단순하기 때문에 훈련 시간은 다른 기법보다 짧다.

(3) 의사결정 트리(Decision Tree)

Breiman et al.(1984)의 저서 제목을 따서 CART(Classification and Regression Trees)라고도 불리는 의사결정 트리는 비모수적(nonparametric)인 분류 및 회귀 기법으로 결과를 시각적으로 이해할 수 있을 뿐만 아니라 해석하기 쉽기 때문에 널리 쓰이고 있다. 또한 스케일링(scaling)같은 전처리 과정이 상대적으로 불필요하며 비선형 관계가 결과에 큰 영향을 미치지 않는 장점이 있다. 반면 아주 복잡한 트리를 생성해서 훈련 세트는 매우 잘 설명할 수 있지만 검증 세트는 잘 설명하지 못 하는 과대 적합(overfitting) 문제가 발생할 수 있으며 데이터의 작은 변화에 결과가 크게 변할 수 있다.[13] 각각의 단계에서 국소 최적해(local optimum)를 구해 전역 최적해(global optimum)를 근사화하는 탐욕 알고리듬(greedy algorithm)이므로 전역적으로 최적인 의사결정 트리를 보장하지는 못 한다.

의사결정 트리는 CART(Classification and Regression Trees) 손실 함수를 최소화해서 구해진다. 아래 식 (1)은 CART 손실 함수를 보여준다:

$$J(k,t_k) = \frac{m_{left}}{m} + \frac{m_{right}}{m} G_{right} \tag{1}$$

m은 표본 수, m_{left}는 왼쪽 노드의 표본수를 표시한다. $G_i (i = left, right)$는 노드 i의 불순도(impurity)를 표시하는데 통상 지니(Gini), 엔트로피(entrophy), 분류오류(misallocation) 지표를 많이 사용한다. 지니의 경우 아래 식 (2)와 같이

13) 이를 분산(variance) 문제라 부르는데 아래에서 설명하는 앙상블 기법 중 배깅(bagging), 부스팅(boosting) 기법을 이용해서 부분적으로 해결할 수 있다.

$$G_i = 1 - \sum_{k=1}^{n} p_{i,k}^2 \tag{2}$$

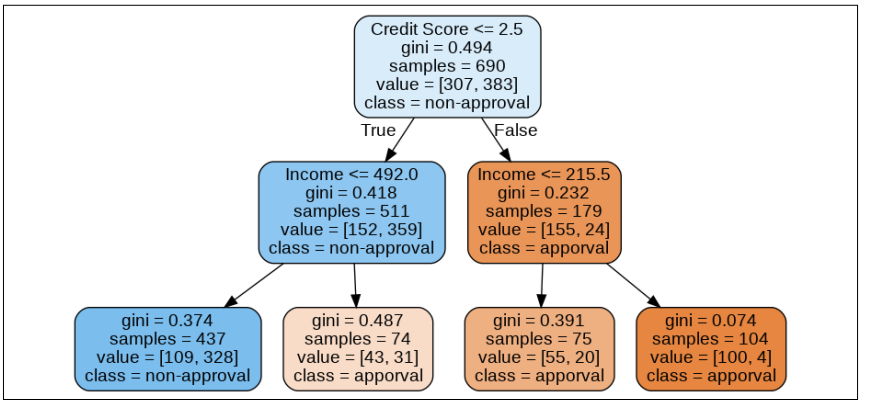
이며 $p_{i,k}$는 노드 i에서 분류집단 k에 속하는 확률이다. 아래 식 (3)과 (4)는 각각 엔트로피와 분류오류를 나타낸다.

$$G_i = -\sum_{k=1}^{n} p_{i,k} \log(p_{i,k}) \tag{3}$$

$$G_i = 1 - \max(p_{i,k}) \tag{4}$$

아래에서는 UCI 머신 러닝 데이터 저장소(UCI Machine Learning Data Repository)의 신용카드 발급심사 데이터를 이용해서 의사결정 트리를 직관적으로 이해하고자 한다.[14] 사생활 보호를 위해 익명화 된 데이터이며 신용카드 발급 심사에 필요한 특성 변수(성별, 나이, 채무, 결혼 여부, 은행 고객 여부, 교육 수준, 인종, 근무 연수, 파산 여부, 고용 여부, 신용점수, 운전면허 유무, 시민권 여부, 우편번호, 소득)와 발급 승인 여부를 담고 있다. 여기에서는 설명의 편의를 위해 발급 승인 여부와 소득, 신용점수 두 가지 특성만을 이용했는데 아래 〈그림 3〉이 결과를 보여준다.

〈그림 3〉 의사결정 트리를 이용한 대출 심사의 예(Credit card approval using decision tree)



14) UCI 데이터 저장소(https://archive.ics.uci.edu/ml/index.php)는 캘리포니아 어바인 대학교에서 관리하는 머신 러닝 훈련 및 검증에 사용하는 데이터 저장소이며 현재 481 종류의 데이터를 제공하고 있다.

맨 위의 출발점을 뿌리 노드(root node)라 하고 뿌리 노드의 깊이(depth)는 0이다. 뿌리 노드 아래에 분화된 2개의 노드는 깊이가 1이 된다. 추가적인 노드로 분화되는 노드를 결정 노드(decision node)라 하며 더 이상 분화되지 않는 노드를 잎(leaf)이라 한다. 우리 예에서 깊이 1의 노드는 모두 결정 노드이며 깊이 2의 노드는 모두 잎에 해당한다.

뿌리 노드를 보면 표본 수는 690명이며 데이터에 비승인이 307명, 승인이 383명이 있다는 것을 보여준다. 690명을 신용점수 2.5를 기준으로 분류하면 깊이 1에서 2.5 이하 511명, 2.5 초과 179명으로 분류된 것을 볼 수 있다. 신용 점수가 2.5이하이기 때문에 비승인으로 분류된 511명 중 실제로 승인된 사람은 359명이다. 추가로 소득 492를 기준으로 승인과 비승인으로 구분할 경우 437명이 비승인으로 분류되는데 이중 실제 비승인된 사람은 109명이다. 불순도의 척도로 지니계수를 보면 0.374로 높게 나온 것을 알 수 있다.[15)]

위의 예에서 짐작할 수 있겠지만 분류를 위해 사용하는 특성 변수가 많은 경우 깊이와 노드의 수는 크게 증가할 수 있다. 극단적으로 표본 수만큼의 노드를 만들 수 있으며 이 경우 훈련 세트는 완전하게 설명하지만 검증 세트는 거의 설명하지 못 하는 과대 적합 문제가 발생한다. 이런 경우를 방지하기 위해 초매개변수를 조정해서 모형을 적절하게 규제(regularization)을 할 필요가 있다. 의사결정 트리 모형의 경우 다수의 초매개변수가 있다. 결정 트리의 최대 깊이(maximum depth), 분할되기 위해 노드가 가져야 할 최소 표본수, 잎 노드가 가져야 할 최소 표본수, 잎 노드의 최대 수, 각 노드에서 분할에 사용할 특성의 최대 수 등이 있다. '최소' 관련 초매개변수를 증가시키거나 '최대' 관련 초매개변수를 감소시키면 규제(regularization)가 증가한다. 예를 들어, 결정 트리의 최대 깊이를 증가시키면 규제가 감소해서 주어진 훈련 세트는 더 잘 설명할 수 있으나, 일반화가 부족해진다.

의사결정 트리는 분류뿐만 아니라 회귀분석에도 쓰일 수 있다. 분류의 경우 각 노드에서 어떤 집단에 속하는지 예측하는 것이 목적이라면 회귀분석의 경우 각 노드에서의 값을 예측하는 것이라 볼 수 있다. 이때, 아래 식(5)와 같이 손실 함수에 불순도 대신 MSE를 사용한다.

---

15) $0.374 = 1 - \left(\frac{109}{437}\right)^2 - \left(\frac{328}{437}\right)^2$로 구해진다.

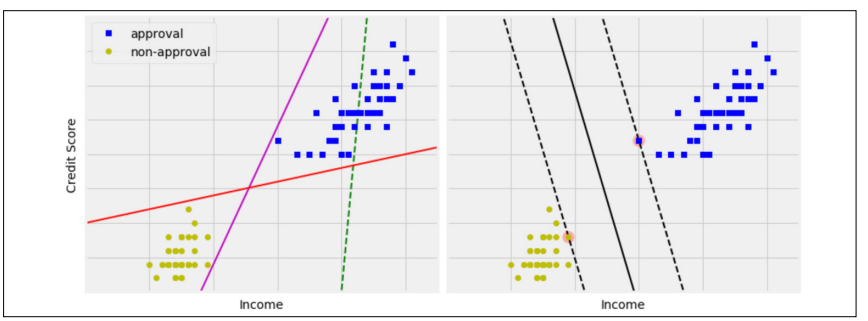
$$J(k,t_k) = \frac{m_{left}}{m} MSE_{left} + \frac{m_{right}}{m} MSE_{right} \quad (5)$$

(4) SVM(Support Vector Machine)

SVM(Support Vector Machine)은 현재 가장 널리 쓰이는 분류 기법 중 하나로 과대 적합 문제가 상대적으로 적으며 소규모, 중규모의 데이터에 적합한 기법이다(Vapnik, 1998; Scholkopf and Smola, 2001). 분류의 경우 선형 SVM과 비선형 SVM으로 구분할 수 있는데 먼저 간단한 예로 선형 SVM을 설명한다.

아래 〈그림 4〉는 신용 점수와 소득을 이용해서 신용카드 발급 심사를 할 경우 SVM을 적용하는 예시를 보여준다. 사각형의 파란 점들은 발급을 승인 받은 사람들의 소득과 신용 점수이며 원형의 노란 점들은 승인을 받지 못 한 사람들의 소득과 신용 점수를 보여준다. 좌측 그림의 보라색과 적색 실선은 대출 승인을 받은 집단과 그렇지 않은 집단을 잘 구분하는 반면 초록색 점선은 두 집단을 구분하는데 실패하고 있다. 그렇다면 어떤 직선이 두 집단을 가장 잘 구분한다고 할 수 있을까? 직관적으로 생각해 보면 두 집단의 '거리'를 가장 크게 하는 방식으로 구분하는 선이 가장 강건한(robust)한 구분이라 할 수 있다. 예를 들어, 좌측 보라색 실선은 현재 상태에서 두 집단을 잘 구분하고 있지만 파란 점들 중에서 가장 좌측에 있는 점보다 조금 더 낮은 소득이 가진 관찰치가 나타날 경우 적절한 구분이 되지 못 할 수 있다. 이에 반해 우측 그림의 검은색 실선의 경우 두 집단에서 조금씩 소득이 높거나 낮은 관찰치가 주어지더라도 현재의 검은색 실선으로 구분을 잘 할 가능성이 여전히 높다.

〈그림 4〉 SVM을 이용한 대출 심사의 예(Credit card approval using SVM)



위의 예에서 볼 수 있는 것처럼 두 집단 사이의 '거리'를 극대화하는 방식으로 분류하는 기법을 SVM이라 하며 우측 그림에 나온 점선 두 개를 서포트 벡터(support vector)라 부른다. 달리 표현하면, 구분하기 가장 어려운 점들을 모아 놓은 집합이 서포트 벡터(support vector)가 된다.[16] 수학적으로 이차 계획법(quadratic programming)을 사용해서 구할 수 있으며 위의 예처럼 2차원이 아니라 n-차원 공간의 경우 직선이 아니라 분리 초평면(separating hyperplane)을 찾는 문제와 다름없다. 분리 초평면으로 집단을 예외없이 구분할 수 있는 경우를 하드마진(hard margin)이라 하며, 서포트 벡터들 사이에 관찰치를 허용하는 경우를 소프트마진(soft margin)이라 한다. 서포트 벡터 사이의 '폭(width)'을 얼마로 정하는지에 따라 분류하지 못 하는 관찰치의 개수가 영향을 받으므로 바로 이 거리가 SVM의 초매개변수가 된다.

선형 SVM은 대체적으로 잘 작동한다고 알려져 있으나, 직선이나 초평면으로 구분되지 않는 데이터들이 있다. 이 경우 비선형 SVM을 사용한다. 〈그림 5〉의 좌측 그림에서는 파란 점들과 노란 점들을 직선으로 분류할 수 없다. 이때 원래 변수의 제곱인 x_1^2을 새로운 특성으로 추가할 경우 우측 그림과 같은 모양이 되며 이 경우에 붉은 점선과 같이 직선으로 두 집단을 분

16) 수학적으로 표현하면 우측 그림의 실선은 $w \cdot x_i + b = 0$로 표시되며 대출 승인을 $y_i = 1$, 비승인을 $y_i = -1$로 표현하면 서포트 벡터는 아래와 같이 표시할 수 있(비승인을 $y_i = 0$로 표시해도 수학적 논리는 그대로 유지된다):

$$w \cdot x_i + b \geq 1 \text{ for } y_i = +1$$
$$w \cdot x_i + b \leq -1 \text{ for } y_i = -1$$

위 두 식을 결합하면 아래의 식으로 표현할 수 있다:

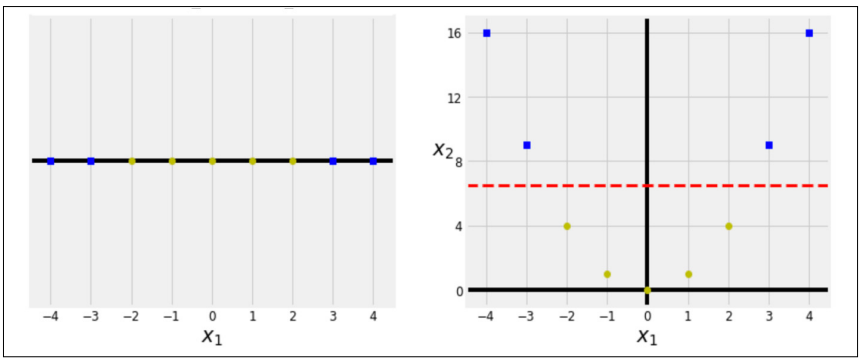
$$y_i(w \cdot x_i + b) - 1 \geq 0 \text{ for } y_i = +1, -1$$

원점에서 아래쪽 점선($w \cdot x_i + b = -1$)까지 거리는 $(-1-b)/|w|$이고 원점에서 위쪽 점선 ($w \cdot x_i + b = 1$)까지 거리는 $(1-b)/|w|$이므로 두 직선 사이 거리는 $2/|w|$가 된다. SVM은 두 직선 사이의 거리를 극대화하는 문제이므로 $2/|w|$의 역수를 극소화하는 문제와 같다. 따라서 SVM의 최적화 문제는 아래와 같이 표현할 수 있으며 경제학에서 자주 접하는 제약 조건하 최적화 문제가 된다:

$$\min \cdot \frac{\|w\|}{2} \; such \; that \; y_i(w \cdot x_i + b) - 1 \geq 0$$

류할 수 있다. 이렇게 데이터를 저차원에서 고차원으로 맵핑(mapping)해 주는 함수를 커널(kernel)이라 부르며 위의 예처럼 x^n을 취하는 방식을 다항 커널을 이용한다고 한다. 이외에도 유사도 특성(similarity feature)을 이용하는 방법이 있다. RBF(Radical Basis Function) 커널과 같이 특정 관찰치('랜드마크')와의 유사도를 0과 1사이의 값으로 표현한 뒤 랜드마크와의 거리를 이용해서 분류하는 방법도 있다.

〈그림 5〉 다항 커널을 이용한 비선형 SVM의 예(Nonlinear SVM using polynomial kernel)



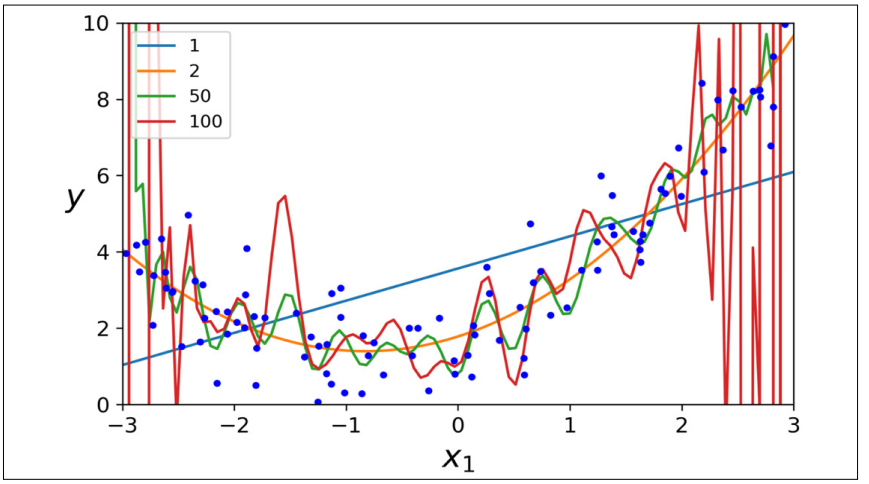
SVM은 분류 뿐만 아니라 회귀분석에도 사용될 수 있다. 이 경우 개념을 정반대로 적용하면 된다. 분류할 때 사용하는 SVM의 개념은 서포트 벡터 사이에 관찰치가 가급적 적게 들어가게 하는 것이나 회귀분석시 목적 함수는 관찰치들을 최대한 많이 포함시키게끔 서포트 벡터를 정하는 것이다. 그리고 두 서포트 벡터 사이의 직선 또는 초평면이 회귀선이 된다.

(5) 규제된 회귀분석(Regularized Regression)

머신 러닝에서 회귀분석을 할 경우 과대 적합을 피하기 위해 많은 경우 규제된 회귀분석(regularized regression)을 사용한다. 〈그림 6〉은 과대 적합의 예를 보여준다. 2차 함수에서 생성된 데이터를 1, 2, 50, 100차 함수로 회귀분석한 뒤 예측값을 보여준다. 1차 함수로 회귀분석한 경우 데이터를 제대로 설명하지 못 하는 과소적합(underfitting)이 나타나는 것을 볼 수 있고, 50차, 100차 함수의 경우 개별 관찰치를 과도하게 고려하려다 보니 비선형성 정도가 매우 큰 모양이 나타난다. 이 경우 기존의 데이터는

잘 설명할지 몰라도 새로운 데이터는 설명하지 못 하는 과대 적합의 문제가 발생한다.

〈그림 6〉 과대 적합의 예(An example of overfitting)



아래 식 (6)은 규제된 회귀분석의 손실함수를 보여준다:

$$\min_{\beta} \cdot \left[\sum_{i=1}^{N} (y_i - \beta^T x_i)^2 + \lambda \left\{ \alpha \sum_{k=1}^{K} |\beta_k| + (1-\alpha) \sum_{k=1}^{K} \beta_k^2 \right\} \right] \qquad (6)$$

$\lambda = 0$인 경우, 즉 식 앞 부분의 $\sum (y_i - \beta^T x_i)^2$만 고려한다면 통상적인 OLS의 경우 손실 함수가 된다. $\alpha = 0$일 경우, 즉 추정치 제곱값의 합을 최소화시키는 것도 함께 고려할 경우 릿지 회귀(ridge regression) 또는 티코노프(Tikhonov) 규제라 한다(Hoerl and Kennard, 1970). 제곱항과 관련되어 있으므로 L2 규제라고도 한다. λ의 값이 크게 주어질수록 데이터를 설명하는 능력에 비해 계수값이 큰 추정치를 가급적 작게 만들게 되며, λ의 값이 아주 크게 되면 데이터를 직선으로 설명하게 된다. 릿지 회귀는 특성의 크기에 민감하므로 스케일링을 반드시 해 주어야 한다.

$\alpha = 1$인 경우를 라쏘(LASSO: Least Absolute Shrinkage and Selection Operator) 회귀라고 하며 L1 규제에 해당한다.[17] 이 경우 절

17) 함수의 특성상 라쏘의 손실함수는 0에서 미분가능하지 않다.

대값의 합을 최소화하는 것에도 신경을 쓰기 때문에 덜 중요한 특성에 대한 추정치를 가급적 0으로 만들려 한다(Tibshirani, 1996). 이런 특성이 있기 때문에 라쏘 회귀는 변수 선택(variable selection)에도 사용할 수 있다.[18] $\alpha \in (0,1)$인 경우는 엘라스틱넷(elastic net)이라 하며 릿지 회귀와 라쏘 회귀를 절충한 것이다(Zou and Hastie, 2005). 규제는 훈련 세트에서만 적용하며 검증 세트에서 평가하거나 실제 예측을 할 때는 규제 없는 모형을 적용한다.[19]

Géron(2017)은 규제가 약간 있는 것이 대부분의 경우 좋으므로 OLS는 피하고 릿지 회귀를 기본으로 하되 실제로 사용되는 특성이 소수라면 릿지 회귀나 엘라스틱넷을 사용하는 것을 추천한다.

(6) 앙상블(Ensemble) 기법

1907년 영국의 한 박람회에서 수소의 무게를 알아 맞히는 대회가 열렸다. 800여 명이 참석해서 제각각 예측을 내 놓았는데 예측의 평균값과 중위수는 실제 무게와 1% 오차 범위 이내였다. Surowiecki(2005)는 저서에서 이런 사례를 소개하며 군중의 지혜(wisdom of crowds)라 불렀는데 앙상블 기법이 바로 이런 아이디어에 기반하고 있다. 앙상블 기법은 동일한 데이터에 다수의 모형들을 적용한 뒤 그 결과들에 기반해서 예측하는 기법이다. 정확도가 51%인 분류기(classifier)가 다수 있다고 하자. 더 좋은 분류기를 만드는 쉽고도, 좋은 방법은 각 분류기의 예측을 모아서 가장 많이 선택된 클래스를 선택, 즉 다수결로 선택하는 것이다. 이와 같이, 51%의 분류기처럼 아무렇게나 분류를 했을 때보다 조금 더 나은 정도의 분류기를 약한 학습기(weak learner)라 부르며 이들을 이용하여 강한 학습기(strong learner)를 만드는 기법을 앙상블이라고 부른다.[20] 각 분류기에서 예측된 클래스를 이용해서 다수결 투표를 할 경우 직접 투표(hard

18) 축약형 모형(reduced form)의 경우에도 어떤 변수를 설명 변수로 사용할 지에 대해 이론적 근거가 있을 수 있으나, 이론적 지침이 전혀 없거나 아무런 정보가 없는 경우가 있기 때문에 변수 선택 과정이 필요한 경우가 있다. 예를 들면, 의학 분야에서 특정 질병을 예측하는데 사용되는 DNA 정보는 수십 만개의 특성으로 이루어진 경우도 있기 때문에 이들로부터 예측 가능성이 높은 특성들을 추려낼 필요가 있다.

19) 절편항에는 규제를 적용하지 않는다.

20) 수학적으로 51%의 정확도를 가진 분류기 1,000개를 이용해서 75%의 정확도를 달성할 수 있다.

voting), 얻어진 확률들을 평균 내어 사용하는 것을 간접 투표(soft voting)라 하는데 후자의 성능이 더 좋은 것으로 알려져 있다. 분류 작업의 경우 직접 투표의 예를 들자면 의사결정 트리, SVM, 로짓 모형 등을 훈련시킨 뒤 새로운 관찰치에 대한 예측을 다수결로 정하는 것이다. 회귀분석의 경우 단일 모형들에서 얻어진 값들의 평균치를 사용한다.

널리 쓰이는 기법으로는 배깅(bagging, bootstrap aggregating의 약자), 부스팅(boosting), 스태킹(stacking)이 있다. 배깅은 부트스트래핑(bootstrapping)을 이용하여 훈련 세트의 양을 늘리고 이를 훈련시키는 기법을 말하는데 새로운 데이터에 대해 모형의 결과가 크게 바뀌는 분산(variance)의 문제를 줄일 수 있다. 예를 들어, 표본에서 부트스트랩핑을 통해 10개의 소표본을 만들고 여기에 각각 모형을 훈련시킨 뒤 결과를 종합해서 판단한다. 즉 배깅은 표본의 수를 늘린 뒤 동일한 모형을 병렬적(parallel)으로 적용하는 방식이다. 비복원 추출을 할 경우에는 패이스팅(pasting)이라 부르고, 일반적으로 배깅의 성능이 더 좋으나 더 오랜 시간과 컴퓨터 연산 능력이 필요하다. 랜덤 포레스트는 배깅 기법을 이용한 의사결정 트리의 앙상블 기법이라 이해할 수 있다.

부스팅은 배깅과 달리 약한 학습기를 순차적으로(sequentially) 적용시키는 방법인데 이전 단계 학습기의 오류를 다음 단계의 학습기가 교정함으로써 강한 학습기를 만드는 기법이다.[21] 대표적으로 애더부스팅(AdaBoosting, adaptive boosting의 약자)과 경사 부스팅(gradient boosting)이 사용되는데 두 방법의 가장 큰 차이는 전자의 경우 예측을 하지 못 한 데이터에 더 많은 가중치를 두면서 학습을 시킨다는 점에 있다. 경사 부스팅의 변종으로 엑스지부스트(XGBoost, Stochastic Gradient Boosting)가 있는데 경사 부스팅과 달리 매 단계마다 훈련 세트 전체를 쓰는 것이 아니라 훈련 세트에서 무작위 추출한 표본을 사용한다.

대개의 경우 동일한(homogenous) 모형을 병렬적으로 사용하거나 순차적으로 사용하는 부스팅과 달리 스태킹은 상이한 모형들을 사용하고 이 모형들의 결과를 종합하는 메타 모형이 존재한다. 예를 들어, 분류 문제를 해결하기 위해 약한 학습기로 KNN, 의사결정 트리, SVM 등의 모형들을 사용하고 메타 모형으로는 뉴럴 넷(neural net)을 사용하는 방식을 들 수 있다.

---

21) 자세한 논의는 Schapire and Freund(2012)를 참고하시오.

2. 평가(evaluation)

머신 러닝 모형의 학습 과정이 끝나면 모형에 대한 평가가 이루어져야 한다.[22] 다양한 평가 기법이 제안되고 사용되고 있으며 지도 학습 회귀의 경우 널리 알려진 MSE(mean squared error), MAE(mean absolute error)를 사용한다.

분류의 경우 여러 지표들이 사용된다. 먼저 실제 값과 모형의 예측에 따라 다음과 같은 2*2의 오차 행렬(confusion matrix)을 고려해보자.

〈표 1〉 오차 행렬(Confusion matrix)

		모형의 예측	
		긍정(Positive)	부정(Negative)
실제 값	긍정(Positive)	TP	FN
	부정(Negative)	FP	TN

오차 행렬은 TP(True Positive), TN(True Negative), FP(False Positive), FN(False Negative)의 구성 요소를 가지는데, 예를 들어 설명하자면 TP는 실제 값도 스팸 메일인데 모형의 예측도 스팸 메일인 경우이다. 반면 FN은 실제로 스팸 메일 임에도 불구하고 모형은 스팸 메일로 분류하지 않은 경우를 말한다.[23]

오차 행렬의 구성 요소를 이용해서 분류의 경우 널리 쓰이는 평가 지표들을 아래 식 (7)-(10)과 같이 정의할 수 있다. 각각의 식은 평가 지표인 정확도(accuracy), 정밀도(Precision), 재현율(Recall), F1 score를 정의한다.

$$\text{정확도(accuracy)} = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

$$\text{정밀도(precision)} = \frac{TP}{TP+FP} \tag{8}$$

22) 아래 설명은 김수현・이영준・신진영・박기영(2019)를 참고하였다.
23) FP는 가설이 참임에도 불구하고 기각하는 1종 오류(type I error)에 해당하고 FN은 2종 오류에 해당한다.

$$\text{재현율(recall)} = \frac{TP}{TP+FN} \tag{9}$$

$$\text{F1 score} = 2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{10}$$

정확도는 전체 개수($TP+TN+FP+FN$) 중에서 실제 값과 모형의 예측이 일치한 갯수($TP+TN$)의 비중을 보여주는 지표이다. 스팸 메일의 예를 들자면, 전체 메일 중에 스팸 메일을 정확하게 스팸 메일로 구분한 것과 스팸 메일이 아닌 것을 아니라고 정확하게 구분한 메일의 비중이다. 정밀도는 모형이 긍정(예를 들어, 모형이 스팸 메일이라고 분류)이라고 분류한 것 중에서 실제로 스팸이라고 예측한 것의 비중이다. 재현율은 실제 값이 참인 것들(예를 들어, 실제로 스팸 메일들) 중에서 모형이 긍정이라고 맞춘 것의 비율이다. F1 score는 정밀도와 재현율 지표의 조화 평균으로 계산한다.

이들 지표의 값들은 모두 높을수록 좋지만, 분석하는 데이터의 특성에 따라 조심스럽게 선택되어 사용해야 한다. 예를 들어 1개의 스팸 메일과 999개의 정상 메일로 구성된 표본이 있다고 하자. 모든 메일을 정상 메일로 분류하는, 판별력이 전혀 없는 모형의 경우에도 정확도가 99.9% (=999/1,000)나 된다. 즉 데이터가 참, 거짓 중 한 쪽으로 치우친 경우 정확도는 좋은 평가 지표가 될 수 없다. 정밀도는 FP가 중요할 때 유용한 지표가 될 수 있다. 스팸 메일의 예를 들자면 FP는 정상 메일인데도 스팸 메일로 분류된 경우이다. 만약에 스팸 메일을 받더라도 중요한 업무 메일을 놓치지 않는 알고리듬을 만들고자 한다면 정확도를 중시해서 평가하는 것이 바람직하다. 반대로 재현율은 FN이 중요할 때 사용한다. 어떤 사람이 특정 질환에 걸려서 치료를 받아야 함에도 환자로 분류하지 않는 상황을 중시할 경우를 들 수 있다. 정밀도와 재현율의 조화 평균으로 계산되는 FI score는 FP와 FN을 모두 고려하는 장점이 있으며, 정확도의 약점인 한쪽으로 치우친 데이터의 경우에도 유용하게 사용될 수 있다.

오차 행렬의 구성 요소를 이용해서 모형의 성능을 시각적으로 보이는 방법으로는 ROC(Receiver Operating Characteristic) 곡선이 널리 쓰인다. ROC 곡선의 세로축과 가로축에는 각각 TP 비율(TPR, True Positive Rate), FP 비율(FPR, False Positive Rate)을 사용한다.[24] 아래 식 (11)과 (12)는 TP 비율과 FP 비율을 나타낸다:

$$\text{TPR} = \frac{TP}{TP + FN} \tag{11}$$

$$\text{FPR} = \frac{FP}{FP + TN} \tag{12}$$

분류의 '정확도'는 ROC 곡선의 아래 부분 면적으로 측정하는데 이를 AUC(Area under Curve)라 하면 0과 1 사이의 값을 가지며 분류를 오류 없이 하는 경우 AUC는 1의 값을 가진다.[25]

Ⅳ. 머신 러닝의 경제학 분야 활용

이 장에서는 머신 러닝 기법을 활용한 경제학 분야의 최신 연구들을 살펴본다. 상호 배타적이지는 않지만, 이해를 돕기 위해 관련 문헌을 새로운 데이터를 이용한 연구, 공공정책 관련 예측 정책 문제(prediction policy problem), 텍스트 마이닝 등으로 분류하여 소개한다.

1. 기존에 없던 통계를 이용한 연구 - 새로운 데이터

경제학 분야에 머신 러닝 기법이 쓰이면서 새로운 종류의 데이터가 새로운 방식으로 쓰이기 시작했다. 대표적으로 원격 탐사 이미지의 한 종류인 위성 사진이 있다. 물론 과거에도 위성 사진이 경제학 연구에 사용된 바가 있지만, 머신 러닝을 통해 다양한 정보를 이용하는 것은 최근의 일이며 위성 사진을 통해 전통적인 데이터보다 더욱 광범위한 정보를 주기적으로 쉽게 확보할 수 있다(Donald and Storeygard, 2016). 예를 들어, Henderson et al.(2012)는 데이터 수집이 어려운 국가의 도시나 작은 단위의 지역의 경제성장 정도를 측정하기 위해 야간 위성 사진에 나타난 인공조명의 밝기를 활용하였다. 또한 식민 지배, 내전, 재정 부족 등의 이유로 비용이 많이 드는 설문조사를 지속하는 것이 어려운 개발도상국에서는 위성

24) TP 비율은 위에서 정의한 재현율과 일치한다.
25) FP 비율은 전 구간에서 0의 값을 가지며 TP 비율은 1의 값을 가지는 경우에 AUC는 1의 값을 가진다.

사진을 이용해 빈곤율과 관련된 데이터를 상대적으로 쉽게 얻기도 한다(Blumenstock, 2016). 구체적으로 Jean et al.(2016)은 위성 사진에 담긴 야간 조명의 세기 혹은 사회 간접 자본의 확충 정도를 각 나라의 소득 수준에 대한 경제 지표로 사용하여 빈곤율을 예측하는 알고리즘을 생성하였다. 이때 사회 인프라의 구축 정도를 관찰하기 위해서는 사진을 통해 자세한 구조물까지 식별 가능해야 하는데, 이는 최근 위성 사진의 높아진 해상도로 인해 가능했다. 뿐만 아니라, 위성 사진은 연구자로 하여금 일정한 주기로 넓은 지역을 관찰할 수 있게 한다는 이점이 있다. 하나의 예로, 위성 사진은 넓은 지역에 걸쳐 있는 각 농경지의 특색을 담고 있어, 필드 실험이나 모델 시뮬레이션과 달리 목표 수확량과 실제 수확량의 차이에 대한 분석을 전체 지역에 대해 일반화할 수 있도록 만든다(Lobell, 2013).

이외에도 구글 스트리트 뷰를 이용하여 뉴욕의 소득 수준을 학습하고 이를 기반으로 보스턴의 소득 수준을 높은 정확도로 예측해낸 연구 사례도 존재한다(Glaeser et al., 2018). 또한 설문이 어렵거나 인터넷, 소셜미디어 기반의 데이터 수집이 어려운 개발도상국에서 빈곤율을 측정하고 예측하기 위해 개인 휴대폰 사용률에 관한 메타 데이터를 사용하기도 한다(Blumenstock et al., 2015).

머신 러닝은 기존의 데이터를 개선하는 데에도 도움을 줄 수 있다. Feigenbaum(2015a, 2015b)은 머신 러닝의 분류 기법과 텍스트 인식 기법을 사용하여 과거의 인구 센서스 데이터와 현재의 데이터를 각 개인별로 매칭하여 경제학자들로 하여금 장기(long-term) 데이터 확보를 가능하도록 했다. 이러한 방법으로 미국 대공황 이전 세대인 아버지와 이후 세대인 아들을 연결시켜 당시의 경제 불황이 세대 간 계층 이동에 미친 영향을 연구했다. Bernheim et al.(2013)은 기존 계량 기법의 사용이 어려운 상황에서 정책 개입의 효과를 예측해야 할 때, 머신 러닝을 이용할 수 있다고 한다. 추정하고자 하는 정책과 비슷한 개입이 과거에 존재하지 않았거나, 개입에 내생성이 존재한다면 기존의 추정 방식으로는 정책 효과를 정확하게 예측할 수 없는 경우, 머신 러닝을 이용해서 개인의 행동에 대한 설문과 그들의 실제 행동에 관한 데이터를 학습하여 해당 정책이 개인의 선택에 미칠 영향을 예측할 수 있다.

2. 예측 정책 문제(prediction policy problem)

Kleinberg et al.(2015)는 경제학 연구에서는 주로 인과관계 추론(causal inference)에 관심을 많이 가지지만 단순한 예측 능력의 향상만으로 다수의 공공정책을 개선할 여지가 많다고 주장하며 이를 예측 정책 문제(prediction policy problem)라 부른다. 이들은 예측 정책 문제를 비가 오게 하려고 돈을 들여 기우제를 하려는 사람과 비가 올 지 몰라 출근길에 우산을 가져갈까 고민하는 사람을 비교하여 설명한다. 전자의 경우 기우제와 실제 강우와의 인과관계가 매우 중요하지만, 후자의 경우 단순히 강우 여부만을 예측하는 것이 중요하다. 머신 러닝은 예측에 강점을 가지므로 후자의 문제 같은 경우에 머신 러닝을 활용해서 공공정책을 개선할 수 있다고 강조한다. 일례로 와 Kang et al.(2013)과 Glaeser et al.(2016)은 옐프(yelp) 온라인 리뷰를 이용한 음식점의 위생 상태 예측이 정부가 어떤 음식점에 위생 점검원을 파견해야 할 지 결정하는 데에 도움을 줄 수 있음을 보였다.[26]

같은 맥락에서, 예측 모델은 정책의 대상을 가려내는 데에 사용될 수 있다. 예를 들어, Abelson et al.(2014)은 위성 사진을 이용하여 무조건부 현금지급(unconditional cash transfer)의 대상이 되는 극빈곤층 마을을 분류했다. 이들은 지붕의 소재를 빈곤의 지표로 삼아 학습한 후 현금지원 대상을 효율적으로 가려냈다. 이와 유사하게, McBride and Nichols (2016)는 경제 발전 분야에서 빈곤층 수혜자를 가려내기 위해 흔히 사용되는 PMT(Proxy Means Test)가 머신 러닝 기법으로 개선되었음을 설명한다.

Athey(2017)는 지도학습을 이용한 머신 러닝을 정책 결정에 사용할 때 발생할 수 있는 한계에 대해 지적한다. 머신 러닝에 기반한 예측은 데이터 학습을 통한 단순 예측이기 때문에 공공정책이 해결하고자 하는 문제의 원

26) 인과관계의 규명 없이 예측 그 자체가 정책 결정에 필요조건이 될 수 있는 추가적인 예로 Kleinberg et al.(2015)은 골관절염 수술 관련 메디케어(Medicare) 수혜자를 결정하는 기준을 예로 든다. 이 수술은 금전적인 비용뿐 아니라 통증과 회복 기간 동안의 불편함 때문에 비금전적인 비용도 매우 크다. 골관절염 수술을 받은 환자의 삶의 질은 즉각적으로 개선되기 보다는 회복에 따라 점진적으로 개선되기 때문에 수술의 혜택을 누릴 수 있을 정도로 긴 수명이 남은 사람을 수혜자로 선택해야 한다. 따라서 환자의 수명 예측이 정책 효율성을 결정짓는 주요 요인이 된다.

인을 파악하지 못한다고 한다. 예를 들어, 앞선 예시에서 다룬 위생 정책의 목표는 음식점들의 양호한 위생 상태 유지이다. 따라서 비위생적인 음식점들이 발생하는 원인을 찾아서 제거하는 것이 정책결정의 내용이어야 하지만, 온라인 리뷰를 통한 예측 내용은 이러한 결정과 관련이 없다. 또한, 음식점 주인이 불시 검문을 받을 확률이 낮다는 생각이 들면 위생 유지를 위한 노력을 줄이려 하고 이를 악용하여 이득을 취하려는 유인이 생기는데 이를 예측으로 설명하기 어렵다.

3. 자연어 처리(NLP, Natural Language Processing)

자연어 처리 기법은 언어학, 컴퓨터 공학, 인공 지능 분야의 접점에 있는 분야로 대량의 자연어를 처리하고 분석하는 기법을 지칭하며 텍스트 마이닝(text mining)이라고도 불린다.[27] 전 세계 데이터의 80%가 자연어(텍스트)와 같은 비정형 데이터(unstructured data)로 이루어진 것을 고려하면 이들 데이터의 정보를 수치화하는 자연어 처리 기법은 학문적 연구 뿐만 아니라 기업 의사 결정에서도 매우 중요한 역할을 할 수 있다.[28]

경제학 분야에서는 이 기법을 이용해서 표본 기간, 빈도(frequency) 등을 확장해서 기존의 데이터를 개선하거나 또는 새로운 데이터를 구축해서 활용하는 방식 등으로 활발하게 활용되고 있다. 빈도 관련한 대표적인 예로는 구글 트렌드(trends.google.com)의 검색어 빈도를 이용해서 실시간 경기 예측을 하는 나우캐스팅(nowcasting)을 들 수 있다. 기존에는 GDP, 민간 소비 및 투자 등 거시 변수를 이용해서 고작해야 분기별, 월별 예측을 할 수 있었는데, 검색어 정보를 이용해서 실시간 예측이 가능하게 되었다. 이 분야의 가장 유명한 연구로는 Baker et al.(2016)의 경제정책 불확실성 지표(EPU: Economic Policy Uncertainty)를 들 수 있다. 이들은 매스 미디어에 출현하는 경제 정책 불확실성과 관련된 용어들의 상대적 빈도를 지표화했는데 이 지표를 이용해서 생산, 투자, 고용을 예측할 수 있음을

27) 경제학 분야의 자연어 처리 기법에 대한 서베이 논문으로는 Gentzkow et al.(2019)이 있다. Kim et al.(2020)도 해당 분야의 서베이 논문이며 관련 문헌, 방법론에 대해 더 자세한 설명을 제공한다.

28) https://www.forbes.com/sites/forbestechcouncil/2019/01/29/the-80-blind-spot-are-you-ignoring-unstructured-organizational-data/#41c9f931211c

보였다. Kelly et al.(2018)은 특허 문서의 유사성을 이용해서 1840-2010년 기간의 기업별, 산업별 '기술 혁신ge' 지표를 구축해서 생산성 예측에 유용함을 보였다. Gentzkow and Shapiro(2010)는 언론사의 논조가 사주의 정치적 성향을 반영하는 것이 아니라 이윤극대화의 결과임을 보였다. 위에서 예시한 연구들과 마찬가지로 이 연구도 자연어 처리 기법 없이는 실증적으로 보이기 힘든 연구이다.

중앙은행의 통화정책, 건전성 감독과 관련된 연구들이 다수 있다.[29] Lucca and Trebbi(2011)는 미 연준 FOMC 의결문에 사용된 단어들의 극성(polarity)를 분석하여 지표를 만든 뒤 의결문이 향후 통화정책 방향에 대해 많은 정보를 가지고 있음을 보였고, Picault and Renault(2017)는 ECB에 대해 유사한 결과를 보였다. Acosta(2015), Acosta and Meade (2015)는 유사도 분석을 통해 FOMC 위원들 발언이 2002년 금리 결정 찬성과 반대 공개 이후 점차 유사해지는 것을 보였다.

우리나라에서도 자연어 처리 기법을 경제학 분야에 활용한 연구들이 나오고 있다. Kim and Pyo(2019)은 뉴스 기사를 이용해 금융 시장의 감성 지표(sentiment index)를 구축한 뒤 이 지표가 국채 금리, 환율 등을 예측하는지 검증하였다. Lee et al.(2019a)는 약 20만 건의 보도를 이용해서 중앙은행 의사록의 감성을 1과 -1 사이의 값을 가지는 지표로 만들었는데 이 지표가 테일러 준칙 하에서 기존의 거시경제 변수들보다 중앙은행의 기준 금리 결정을 더 잘 예측한다는 것을 보였다. Lee et al.(2019b)는 새로운 통화정책 충격 지표를 제안하였다. 기존에는 통화정책 충격을 식별하기 위해서 VAR을 사용한다든가, 기준 금리의 선물 금리를 사용해 왔는데, 전자의 경우 모형 설정(specification)에 따라 결과가 달라질 가능성이, 후자의 경우 해당 금융 상품이 미국에만 존재한다는 약점이 있다. 이런 약점을 보완하고자 이들은 금통위 회의 하루 전 날과 하루 후에 보도된 관련 뉴스의 감성을 측정하고 이 둘의 차이를 통화정책 충격이라 정의하였다.[30]

---

29) 글로벌 금융위기 기간 동안 중앙은행의 언설(words)에 기대는 선제적 지침(forward guidance)의 중요성이 커졌는데 자연어 처리 기법이 적절한 분석 수단이 될 수 있다. 중앙은행에 국한한 자연어 처리 연구에 대한 서베이 논문으로는 Bholat et al.(2015)를 들 수 있다.

30) 시장 참여자들이 대체적으로 금통위가 금리를 인하할 것이라 예상하는 경우를 고려해 보자. 이 상황에서 금통위가 금리를 동결한다면 비록 기준 금리 변화는 0이지만 금통위의 결정은 hawkish한 것으로 받아 들여질 수 있고 이들 지표는 이런 경우를

VAR로 식별된 충격이 주로 단기 금리와 상관 관계가 높은 것과 달리, 이들의 지표는 주로 장기 금리와 연관성이 높은 것으로 나타났다.

4. 이론의 검증

지도 학습은 이론을 검증하는 데에도 적용될 수 있다. 이론을 검증하기 위해서는 대상이 모형 또는 이론의 예측과 얼마만큼 일치하는지 확인해야 한다. 머신 러닝은 이러한 기준에 비추어 이론의 예측을 비교, 평가할 수 있는 벤치마크를 제시한다. Kleinberg et al.(2017)은 가장 정확한 예측 변수(predictor)를 이용한 결과와 이론의 예측력을 비교한다. 마찬가지로 Naecker and Peysakhovich(2015)는 행동 경제 모형에서 경제 주체가 위험 혹은 불확실성 아래에 있을 때 어떤 의사결정을 내리는지 예측한다. 기존의 이론이 제시하는 예측을 벤치마크와 비교하여 평가한 결과 이론이 벤치마크보다 설명력이 떨어짐을 보였다. Gu et al.(2018)은 머신 러닝 기법을 다중요인 모형(multi-factor model)에 응용하였는데 기존 학계에서 제안된 수많은 요인들 중에서 모멘텀, 유동성, 변동성 변수가 주식들의 횡단면 수익률을 설명하는데 가장 우수한 요인들이라는 것을 보였다.

Ⅴ. 머신 러닝, 경제, 경제학

1. 인공 지능과 노동시장

인공 지능이 노동시장에 미치는 영향은 너무나 광범위하므로 중요한 두 가지 질문, 첫째 과연 인공 지능이 일자리를 줄일 것인가 그리고 둘째, 저숙련 노동자가 상대적으로 먼저 대체될 것인가에 대해 고려할 점들을 간단하게 언급하기로 한다.

Agrawal et al.(2018)에 따르면 인공 지능은 기본적으로 예측(prediction)에 대한 기술이기 때문에 인공 지능이 경제에 미치는 영향은 예측을 더 저렴하게(cheaper)하게 만드는 것에 나온다. 이런 관점에서 보

포착할 수 있는 장점이 있다.

면 일자리에 대한 영향은 개선된 예측을 통해 (1) 일자리 자체를 대체('replacing'), (2) 업무의 일부분만을 더 효율적으로 만들어 주고 일자리 자체를 대체하지는 않음('enabling'), (3) 없어지는 일자리로 새로운 일자리가 창출되는 등 여러 경우가 생길 수 있으며 이들 일자리의 상대적 비중, 분포 등에 따라 노동시장에 대한 영향은 상이할 수 있다.[31] (1)의 예로는 패스트푸드점에서 키오스크를 통해 주문을 자동화하면서 고용을 줄이는 것을 들 수 있고, (2)의 예로는 비행기 조종사를 들 수 있다. Agrawal et al.(2018)에 나온 예처럼 조종사의 경우 실제 조종하는 시간은 전체 비행시간의 7%에 불과하며 인공 지능이 발달할수록 이 비중은 점차 줄어들 것이나 이 수치가 0%에 도달하기 전까지는 조종사의 일자리는 그대로 유지될 것이다. (3)의 예로는 트럭 운전수를 들 수 있다. 자율주행이 가장 쉽게 활용될 수 있는 분야는 미국의 경우 주간 고속도로를 통해 화물을 운송하는 트럭인데 이 경우 운전사 일자리는 사라질 수 있으나 트럭을 지키는 보안요원 일자리가 창출된다.[32] 이런 맥락에서 PWC(2018)은 영국의 경우 향후 20년간 인공 지능으로 인해 700만개의 일자리가 사라질 것으로 예상되지만, 생산비용 감소와 지출 증가로 인해 720만개의 일자리가 생길 것으로 전망한다. 오히려 일자리가 20만개 증가하는 것이다.

인공 지능이 발전하면서 저숙련의 일자리가 먼저 대체될 것이라는 우려가 있다. 그러나 꼭 그렇지만은 않다. Agrawal et al.(2018)은 중국의 농부와 미국의 회계사의 경우를 예로 든다. 미국의 회계사가 하는 일이 중국의 농부가 하는 일보다 더 많은 교육과 훈련이 필요한 전문적인 일이지만, 머신 러닝을 이용한 업무 자동화의 관점에서 보면 전자의 경우가 훨씬 용이하다. 게다가 임금 수준을 고려하면 전자를 자동화할 유인이 훨씬 더 크므로 중국의 농부보다는 미국의 회계사가 인공 지능으로 대체될 가능성이 더 크다고 볼 수 있다.

---

31) Acemoglu and Restrepo(2019)는 1990-2017년 동안 산업용 로봇으로 인해 미국 노동시장에서 고용과 임금이 감소한 것을 보였다.

32) 왜냐하면 자율주행의 특성상 사람 형태가 나타나면 정지를 하거나 우회를 하는 방식으로 프로그램이 될 텐데 이를 이용한 절도의 가능성이 상존하기 때문이다.

2. 노동으로서의 데이터

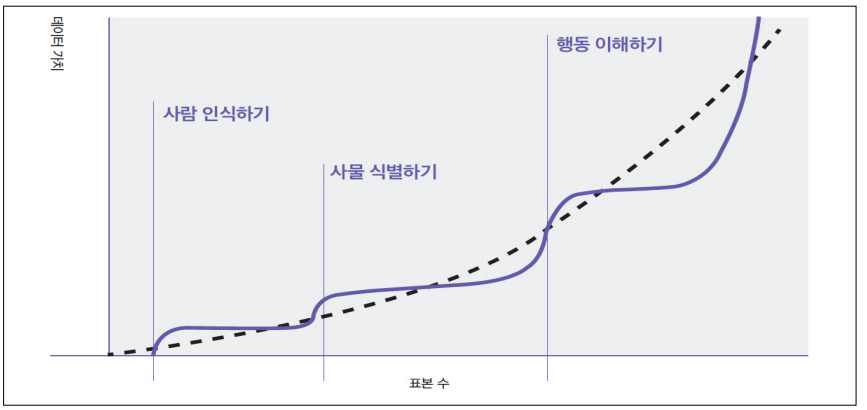
일반적인 생산함수는 $Y = F(K,L)$의 형태를 가지고 있는데 데이터는 자본(K)과 노동(L) 중 어디에 속할까? FAANG이라고도 부르는 페이스북, 아마존, 애플, 넷플릭스, 구글과 같은 거대 테크 기업들의 관점에서 보면 데이터는 기업이 소유한 무형 자본에 속하겠지만, 최근 들어 이들 기업들이 이용하는 데이터는 사용자들이 제공한 것이므로 사용자들의 노동이라 간주하고 이에 대한 보상이 주어져야 한다고 주장들이 나오고 있다. 2020년 대선에 출마를 선언한 민주당 후보 앤드류 양(Andrew Yang)은 데이터로부터 나오는 혜택 또는 이윤을 기본 소득의 형태로 사회에 환원하자는 주장을 하며 캘리포니아 주지사 개빈 뉴섬(Gavin Newsom)은 데이터 배당금(data dividend)을 주장하고 있다.[33] 데이터에 대한 보상이 주어질 경우 동기 부여가 되어 더 양질의 데이터가 제공될 가능성이 커지며 인공 지능의 성능이 더 향상될 수 있다.

반면 반론도 있다. 구글의 수석 경제학자인 홀 베리언(Hal Varian)은 기름이 정제 과정을 거쳐야 하듯이 데이터도 전처리와 분석 과정을 거쳐야만 가치가 있다고 주장한다. 또한 고전적 통계학에서는 표본 수가 늘어날수록 데이터의 가치는 급격하게 감소하기 때문에 데이터 제공에 대한 보상을 한다고 해도 매우 미미하고 이메일과 소셜 미디어를 무료로 사용하는 것으로 충분히 보상된다고도 볼 수 있다. 즉 데이터의 가치는 표본 수에 대해 오목한(concave)한 함수 형태를 가지며 표본 수가 아주 커질 경우 데이터의 한계 가치는 0으로 수렴한다. Ibarra et al.(2018), Posner and Weyl(2019)에 따르면 이러한 논리는 고전적 통계학에서 성립하는 논리이며 머신 러닝에서는 해결하고자 하는 문제의 복잡도(complexity)에 따라 데이터의 가치가 비선형적으로 변화한다고 주장한다. 아래 〈그림 7〉은 머신 러닝 영역에서 전형적으로 나타나는 표본 수와 데이터의 가치 사이 관계를 보여 준다. 사진에 특정 사람이 있는지 여부를 판단하는 것, 사진 속 특정 사물을 식별하는 것, 사진 속 사람이 어떤 행동을 하고 있는지 판단하는 것

33) https://www.cnbc.com/2019/09/12/andrew-yang-says-he-will-give-1000-a-month-ubi-to-10-more-families.html
, https://www.nytimes.com/2019/03/25/us/newsom-hertzberg-data-dividend.html

순서로 문제의 복잡도가 늘어나는데, 데이터의 가치는 특정 복잡도 문제를 해결하기 직전부터 매우 가파르게 증가한다. 사람을 인식하는 작업에 필요한 데이터가 확보되면 한동안 데이터의 한계 가치는 0이나 다름없지만 더 복잡한 단계의 작업을 하기 위해서는 추가적인 데이터 유무가 매우 중요해지며 따라서 데이터의 가치도 점증적으로 증가하는 구간이 생긴다.

〈그림 7〉 머신 러닝 영역에서 전형적으로 나타나는 표본 수와 데이터 가치의 관계(A typical relationship between sample size and data value in machine learning)



출처: Posner and Weyl(2019, 304쪽).

이런 맥락에서 보면 데이터의 가치를 어떻게 측정하고, 그 가치를 어떤 방식으로 분배할 것인가에 대한 연구가 필요하다. Acemoglu et al. (2019)는 이 방면의 선도적 연구로 온라인 데이터의 가치와 시장 구조에 대해 분석하고 있다.

3. 인과관계 추론

예측과 인과관계 추론(causal inference)은 관련은 있으나 별개의 개념이며 머신 러닝은 기본적으로 예측에 초점을 맞추고 있다. 그렇기 때문에 머신 러닝의 방법론을 경제학 분야에 직접적으로 원용하여 인과 관계를 보이는 것은 쉽지 않다.[34] 예를 들어, Levitt(1997)에 나온 경찰관 수와 범

34) Athey(2019)는 바로 이런 이유 때문에 인과관계보다는 예측에 강점이 있는 머신 러닝 기법들이 경제학 분야에 예상보다 빨리 활용되지 않는다고 본다. 컴퓨터 과학

죄율의 관계를 생각해 보자. 경제학자들의 관심은 "경찰관을 어떤 지역에 10명을 증원했다면 범죄율은 평균적으로 얼마 감소할 것이라 예상할 수 있는가?"라는 인과관계 추론에 기반한 질문이다. 반면 통계를 살펴보면 범죄율이 높은 지역일수록 경찰관이 많이 배당되는 경향이 있으므로 두 변수의 관계는 (+)의 관계를 가지기 때문에 이 통계를 이용해서 모형을 학습을 시키면 위와 같은 인과관계에 기반한 답을 제시하지 못 한다. 머신 러닝 기법으로 대답할 수 있는 적절한 질문은 "어떤 지역의 인구 대비 범죄율이 아주 높다면 해당 지역의 경찰관 수는 어떨까"이다.[35]

Ascarza(2018)은 소비자 이탈(churn)을 막기 위한 기업들의 정책(예를 들어, 무료 이용 기간을 연장)이 머신 러닝 기법에 의존한 경우 비효율적일 수 있음을 보였다. 지도 학습을 이용해서 이탈의 가능성이 높은 소비자 집단과, 경제학적 방식으로 일종의 처치 효과(treatment)로 볼 수 있는 기업 정책에 따라 이탈을 하지 않은 집단을 비교해 보니 50% 정도만 겹친다는 것을 보였다. 즉 나머지 50%의 경우 기업 입장에서 어떤 정책을 쓰든 어차피 이탈할 집단인 것이다. Ascarza(2018)의 연구는 지도 학습 결과에 따라서 이탈 가능성이 높은 군에 자원을 배분하는 것은 비효율적이며 인과관계 파악의 중요성을 보였다.

최근 들어서 컴퓨터 과학 분야의 기성("off-the-shelf") 기법 사용을 탈피하고 인과 관계 파악을 위한 경제학자들의 시도가 나타나고 있다. 기본적으로 머신 러닝 기법을 직접적으로 이용해서 인과 관계를 밝히기 보다는 기존 방법론을 머신 러닝을 통해 개선하는 방식을 택하고 있다. Mullainathan and Spiess(2017)는 기존 계량경제학에서는 $\hat{\beta}$이라 표현할 수 있는 모수추정(parameter estimation)을 중요시하는 반면 머신 러

분야의 노벨상이라 할 수 있는 튜링상을 2011년 수상한 Pearl(2018)은 경제학자들처럼 인과 관계를 강조하며 머신 러닝을 통해 반사실적(counterfactual) 결과를 구현할 수 있음을 보이며 관련 연구를 진행하고 있다. 이런 방식의 연구를 통해 어떤 환자에게 약 A가 아니라 약 B를 처방했다면 환자의 상태는 어떠했을까 식의 반사실적 질문에 대답할 수 있다.

35) Athey(2017)에 나온 호텔방 점유율과 방값 사이의 관계도 같은 맥락에서 볼 수 있다. 호텔은 이윤극대화를 위해 이용하고자 하는 고객이 많을수록 방값을 인상시키기 때문에 두 변수는 데이터에서 (+)의 관계를 가진다. 경제학자들이 알고자 하는 것은 "가격이 한 단위 변하면 호텔 수요가 얼마나 감소하는가?"에 대한 대답이나 머신 러닝을 통해 적절하게 답할 수 있는 질문은 "호텔 가격이 매우 높은 시기에 점유율을 어떻게 될까?"이며 답은 "매우 높을 가능성이 높다"이다.

닝에서는 $\hat{y}$으로 표현할 수 있는 예측(prediction)을 중요시하기 때문에 $\hat{y}$이 중요한 분야에서 머신 러닝이 기여할 여지가 있다고 강조한다. 도구변수(IV, Instrumental Variable) 추정을 예로 들 수 있다. 선형회귀의 경우 도구변수를 z라 표시한다면 1단계 추정은 $x = z\gamma + u$로 이루어지고 2단계 추정 $y = x\beta + \varepsilon$에서 x대신 1단계 추정에서 구한 $\hat{x}$을 이용하게 된다. 이 경우 중요한 것은 추정량 $\hat{\gamma}$ 자체가 아니라 오차항 ε과 무관한 x의 예측치 $\hat{x}$이다. 이렇게 예측에 강점을 보이는 머신 러닝 기법을 이용해서 도구변수 추정의 편의(bias)를 줄일 수 있다. 이외에도 이질적 처치 효과(heterogenous treatment effect), 패널 데이터, synthetic control, 이중차분법(difference-in-differences method) 등의 맥락에서 인과 관계 추정을 개선하려는 노력들이 시도되고 있다.[36]

Ⅵ. 결 론

Mullainathan and Spiess(2017)은 경제학자들이 무작위 대조 실험(randomized control trials)을 활발하게 이용하면서 경제학 실증 분석의 질문 자체가 변화한 것과 같은 유사한 가능성이 머신 러닝에서도 보인다고 주장한다. 이들에 따르면 머신 러닝이 새로운 데이터, 새로운 방법론을 만드는 것뿐만 아니라 새로운 질문에 답하는 데도 유용하게 쓰일 것이라 강조한다. 이런 낙관적 전망에도 불구하고 해결해야 할 여러 문제들이 산재해 있다. 먼저 부족한 해석 가능성(explainability, interpretability)이다. 머신 러닝의 기본적 목표는 예측에 있으므로 데이터에 기반해서 인공 지능이 예측을 했지만 어떻게 데이터를 분석했는지 명확하게 설명할 수 없는 경우들이 많다. 특히 딥러닝의 경우 특성과 목표 사이 관계가 블랙박스인 경우가 대다수이다. 이와 관련된 문제로 확장성(scalability)을 들 수 있다. 확장성은 새로운 데이터 또는 더 큰 규모의 데이터에서도 모형의 성능이 그대로 유지되는지 여부를 의미하는데 해석 가능성이 전제되어 있다면 확장성

36) 관련 분야의 최신 연구로는 Athey and Wager(2018), Athey, Tibshirani, and Wager(2019), Athey, Bayati, Imbens, and Qu(2019)를 참고하시오.

문제가 발생할 경우 왜 문제가 발생했는지 상대적으로 용이하게 파악할 수 있다.[37]

차별 또는 공정성(fairness)에 대한 고려도 필요하다. 민간 회사가 만들고 미국 20여개 주에서 사용되었던 COMPAS(Correctional Offender Management Profiling for Alternative Sanctions)는 체포 직후 피의자에게 137개의 질문을 해서 재범율을 예측하는 인공 지능 알고리듬인데 흑인을 차별한 것으로 밝혀져 큰 사회적 반향을 일으켰다.[38] 아마존의 경우 인공 지능에 기반한 사내 채용 프로그램을 활용해 왔는데 여성 차별적인 것으로 드러나 2018년 10월부터 사용을 중단하였다. 이런 사례는 "쓰레기가 들어가면 쓰레기가 나온다(garbage in, garbage out)"이라는 머신 러닝 분야의 금과옥조를 상기시킨다.

또한 공공정책에 활용할 때 조작 가능성(manipulation)을 염두에 두어야 한다. 앞에서 살펴 본 Kang et al.(2013), Glaeser et al.(2016)에 나온 것처럼 온라인 평가를 이용해 효과적으로 검침원을 배치할 수 있지만 해당 알고리듬이 밝혀질 경우 이에 맞춰서 온라인 평가를 조작한다든가, 또는 그런 평가가 나오도록 행동할 수 있다. 따라서 예측에 기반한 알고리듬 모형을 대규모로 현실 사회에서 사용할 경우 대상자들의 반응도 염두에 두고 정책 설계를 할 필요가 있다.

위에서 언급한 머신 러닝 분야의 문제 뿐만 아니라 경제학계의 문제도 있다. Athey and Imbens(2019)는 머신 러닝 고유의 문제를 해결해야 할 뿐만 아니라 경제학계에 어울리는 고유의 머신 러닝 방법론을 개발할 필요성이 있다고 강조한다. 현재까지 많지 않은 연구 중 다수는 컴퓨터 공학 분야에서 개발한 기성품(off-the-shelf)의 방법론을 차용했는데, 이들은 경제학 분야에 특화된 방법론도 개발할 필요가 있다고 강조한다. 위에서 언급한 문제들이 점차 해결되고, 경제학 분야 특유의 방법론이 함께 개발될 경우 머신 러닝은 빅데이터와 결합되어 경제학계의 데이터, 방법론, 질문 자체를 바꿀 여지가 매우 크다고 하겠다.

투고 일자: 2019. 10. 6. 심사 및 수정 일자: 2019. 10. 30. 게재 확정 일자: 2019. 10. 31.

37) 이와 유사한 개념으로 강건성(robustness), 안정성(stability)이 있다.

38) 프로퍼블리카의 관련 기사는 다음을 참고하시오: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

◈ 참고문헌 ◈

김수현 · 이영준 · 신진영 · 박기영 (2019), "경제분석을 위한 텍스트 마이닝," 『한국경제의 분석』, 제26권 1호, 게재예정.

Kim, Soohyon, Youngjoon Lee, Jhinyoung Shin, and Ki Young Park (2020), "Text Mining for Economic Analysis," *Panel for Korean Economic Analysis*, 26(1), forthcoming.

Abelson, Brian, Joy Sun, and Kush R. Varshney (2014), "Targeting Direct Cash Transfers to the Extremely Poor," In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '14*, 1563-72. New York, New York, USA: ACM Press. https://doi.org/10.1145/2623330.2623335

Acemoglu, Daron, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar (2019), "Too Much Data: Prices and Inefficiencies in Data Markets," Cambridge, MA: National Bureau of Economic Research, September. https://doi.org/10.3386/w26296

Acemoglu, Daron, and Pascual Restrepo (2019), "Robots and Jobs: Evidence from US Labor Markets," *Journal of Political Economy*, no. Not Available. https://doi.org/10.1086/705716

Acosta, Miguel (2015), "FOMC Responses to Calls for Transparency," *FEDS Working Paper* No. 2015-060, Board of Governors of the Federal Reserve System (U.S.), (July 10). http://dx.doi.org/10.17016/FEDS.2015.060

Acosta, Miguel, and Ellen E. Meade (2015), "Hanging on Every Word : Semantic Analysis of the FOMC's Postmeeting Statement," *FEDS Notes*, 2015-09-30, Board of Governors of the Federal Reserve System (U.S.).

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb (2018), *Prediction Machines: The Simple Economics of Artificial Intelligence*, Harvard Business Review Press.

Arrieta-Ibarra, Imanol, Leonard Goff, Diego Jiménez-Hernández, Jaron Lanier, and E. Glen Weyl (2018), "Should We Treat Data as Labor? Moving beyond 'Free'," *AEA Papers and Proceedings*, 108, 38-42.

Ascarza, Eva (2018), "Retention Futility: Targeting High-Risk Customers Might Be Ineffective," *Journal of Marketing Research*, 55(1), February, 80-98. https://doi.org/10.1509/jmr.16.0163

Athey, Susan (2017), "Beyond Prediction: Using Big Data for Policy Problems," *Science*, 355(6324), February 3, 483-485. https://doi.org/10.1126/science.aal4321

Athey, Susan (2019), "The Impact of Machine Learning on Economics," *In The Economics of Artificial Intelligence: An Agenda*, 1 edition., 507-547, National Bureau of Economic Research Conference Report, University of Chicago Press.

Athey, Susan, and Guido W. Imbens (2019), "Machine Learning Methods That Economists Should Know About," *Annual Review of Economics*, 11(1), August 2, 685-725. https://doi.org/10.1146/annurev-economics-080217-053433

Athey, Susan, and Stefan Wager (2018), "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association*, 113(523), July 3, 1228-1242. https://doi.org/10.1080/01621459.2017.1319839

Athey, Susan, Julie Tibshirani, and Stefan Wager (2019), "Generalized Random Forests," *The Annals of Statistics*, 47(2), 1148-1178. https://doi.org/10.1214/18-AOS1709

Athey, Susan, Mohsen Bayati, Guido Imbens, and Zhaonan Qu (2019), "Ensemble Methods for Causal Effects in Panel Data Settings," *AEA Papers and Proceedings*, 109, 65-70.

Baker, Scott R., Nicholas Bloom, and Steven J. Davis (2016), "Measuring Economic Policy Uncertainty," *The Quarterly Journal of Economics*, 131(4), November, 1593-1636. https://doi.org/10.1093/qje/qjw024

Bernheim, B. Douglas, Daniel Bjorkegren, Jeffrey Naecker, and Antonio Rangel (2013), "Non-Choice Evaluations Predict Behavioral Responses to Changes in Economic Conditions," Cambridge, MA: National Bureau of Economic Research, August. https://doi.org/10.3386/w19269

Bholat, David, Stephen Hansen, Pedro Santos, and Cheryl Schonhardt-Bailey (2015), "Text Mining for Central Banks."

Availabe at SSRN: https://ssrn.com/abstract=2624811

Blumenstock, J. E. (2016), "Fighting Poverty with Data," *Science*, 353(6301), August 19, 753-754. https://doi.org/10.1126/science.aah5217

Blumenstock, J., G. Cadamuro, and R. On (2015), "Predicting Poverty and Wealth from Mobile Phone Metadata," *Science*, 350(6264), November 27, 1073-1076. https://doi.org/10.1126/science.aac4420

Breiman, Leo, Jerome Friedman, Charles J. Stone, and R. A Olshem (1984), *Classification and Regression Trees*, 1 edition. Wadsworth Statistics/Probability, Chapman and Hall/CRC.

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan (2016), "Productivity and Selection of Human Capital with Machine Learning," *American Economic Review*, 106(5), May, 124-127. https://doi.org/10.1257/aer.p20161029

Donaldson, Dave, and Adam Storeygard (2016), "The View from Above: Applications of Satellite Data in Economics," *Journal of Economic Perspectives*, 30(4), November, 171-198. https://doi.org/10.1257/jep.30.4.171

Efron, Bradley, and Trevor Hastie (2016), *Computer Age Statistical Inference; Algorithms, Evidence, and Data Science*, 1 edition. Institute of Mathematical Statistics Monographs, (Book 5), Cambridge University Press.

Eric A. Posner, and E. Glen Weyl (2018), *Radical Markets: Uprooting Capitalism and Democracy for a Just Society*, Princeton University Press.

Feigenbaum, James J (2015b), "Intergenerational Mobility during the Great Depression." http://scholar.harvard.edu/jfeigenbaum/publications/jmp

Feigenbaum, James J. (2015a), "Automated Census Record Linking." http://scholar.harvard.edu/jfeigenbaum/publications/automated-census-record-linking

Ford, Martin (2018), *Architects of Intelligence: The Truth about AI from the People Building It*, Packt Publishing.

Gentzkow, Matthew, and Jesse M. Shapiro (2010), "What Drives

Media Slant? Evidence From U.S. Daily Newspapers," *Econometrica*, 78(1), 35-71. https://doi.org/10.3982/ECTA7195

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (2019), "Text as Data," *Journal of Economic Literature*, 57(3), September 1, 535-574. https://doi.org/10.1257/jel.20181020

Géron, Aurélien (2017), *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 1 edition, O'Reilly Media.

Glaeser, Edward L., Andrew Hillis, Scott Duke Kominers, and Michael Luca (2016), "Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy," *American Economic Review*, 106(5), May, 114-118. https://doi.org/10.1257/aer.p20161027.

Glaeser, Edward L., Scott Duke Kominers, Michael Luca, and Nikhil Naik (2018), "Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life: Big Data and Big Cities," *Economic Inquiry*, 56(1), January, 114-137. https://doi.org/10.1111/ecin.12364

Goodfellow, Ian, Mehdi Mirza, Jean Pouget-Abadie, Bing Xu, David Warde-Farley, Sherjil Ozair, and Aaron Courville (2014), "Generative Adversarial Nets," *Neural Information Processing Systems*, December 9.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu (2018, 2019), "Empirical Asset Pricing via Machine Learning," *NBER Working Paper* No. 25398. https://www.nber.org/papers/w25398.pdf

Henderson, J. Vernon, Adam Storeygard, and David N Weil (2012), "Measuring Economic Growth from Outer Space," *American Economic Review*, 102(2), April, 994-1028. https://doi.org/10.1257/aer.102.2.994

Hoerl, E., and R. W Kennard (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55-67. https://doi.org/10.1080/00401706.1970.10488634

Jean, N., M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon (2016), "Combining Satellite Imagery and Machine Learning to Predict Poverty," *Science*, 353(6301), August 19, 790-794. https://doi.org/10.1126/science.aaf7894

Kang, Jun Seok, Polina Kuznetsova, Michael Luca, and Yejin Choi (2013), "Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews," In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1443-1448. Stroudsburg, PA: Association for Computational Linguistics.

Kelly, Bryan, Dimitris Papanikolaou, Amit Seru, and Matt Taddy (2018), "Measuring Technological Innovation over the Long Run," *NBER Working Paper* No. 25266, November. https://www.nber.org/papers/w25266.pdf

Kim, Jungho, and Dong-Jin Pyo (2019), "News Media Sentiment and Asset Prices in Korea: Text-Mining Approach," *Asia-Pacific Journal of Accounting & Economics*, July 17, 1-23. https://doi.org/10.1080/16081625.2019.1642115

Kim, Soohyon, Young Joon Lee, Ki Young Park and Jhinyoung Shin (2020), "Text Mining for Macroeconomic Analysis (in Korean)," *Journal of Korean Economic Analysis*, April.

Kleinberg, Jon, Annie Liang, and Sendhil Mullainathan (2017), "The Theory Is Predictive, but Is It Complete? An Application to Human Perception of Randomness," *ArXiv:1706.06974 [Cs, Stat]*, June 21. http://arxiv.org/abs/1706.06974

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer (2015), "Prediction Policy Problems," American Economic Review, 105(5), May, 491-495. https://doi.org/10.1257/aer.p20151023

Kreif, Noémi, and Karla DiazOrdaz (2019), "Machine Learning in Policy Evaluation: New Tools for Causal Inference," In *Oxford Research Encyclopedia of Economics and Finance*, by Noémi Kreif and Karla DiazOrdaz, Oxford University Press. https://doi.org/10.1093/acrefore/9780190625979.013.256.

Lee, Kai-Fu (2018), *AI Superpowers: China, Silicon Valley, and the New World Order*, 1 edition. Houghton Mifflin Harcourt.

Lee, Young Joon, Soohyon Kim, and Ki Young Park (2019), "Deciphering Monetary Policy Board Minutes Through Text Mining Approach: The Case of Korea," January 7, Bank of Korea WP 2019-1. https://ssrn.com/abstract=3312561

Lee, Young Joon, Soohyon Kim, and Ki Young Park (2019), "Measuring Monetary Policy Surprises Using Text Mining: The Case of Korea," April 9, Bank of Korea WP 2019-11. https://ssrn.com/abstract=3347429

Levitt, Steven D. (1997), "Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime," The American Economic Review, 87(3), June, 270-290.

Lobell, David B. (2013), "The Use of Satellite Data for Crop Yield Gap Analysis," Field Crops Research, 143, March, 56-64. https://doi.org/10.1016/j.fcr.2012.08.008

Lucca, David O., and Francesco Trebbi (2011), "Measuring Central Bank Communication: An Automated Approach with Application to FOMC Statements," Technical Report 15367,2009 version, NBER Working Paper. https://faculty.arts.ubc.ca/ftrebbi/research/lt.pdf

McBride, Linden, and Austin Nichols (2016), "Retooling Poverty Targeting Using Out-of-Sample Validation and Machine Learning," *The World Bank Economic Review*, October 28, lhw056. https://doi.org/10.1093/wber/lhw056

Mitchell, Thomas M. (1997), *Machine Learning*, New York, NY, USA: McGraw-Hill, Inc.

Moritz, Benjamin, and Tom Zimmermann (2016), "Tree-Based Conditional Portfolio Sorts: The Relation between Past and Future Stock Returns," *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2740751

Mullainathan, Sendhil, and Jann Spiess (2017), "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 31(2), May, 87-106. https://doi.org/10.1257/jep.31.2.87

Naecker, Jeffrey, and Alexander Peysakhovich (2017), "Using Methods from Machine Learning to Evaluate Behavioral Models of Choice under Risk and Ambiguity," *Journal of Economic Behavior & Organization*, 133, January, 373-384. https://doi.org/10.1016/j.jebo.2016.08.017

Pearl, Judea, and Dana Mackenzie (2018), *The Book of Why: The New Science of Cause and Effect*, 1 edition. Basic Books.

Picault, Matthieu, and Thomas Renault (2017), "Words Are Not All Created Equal: A New Measure of ECB Communication," *Journal of International Money and Finance*, 79, December, 136-156. https://doi.org/10.1016/j.jimonfin.2017.09.005

Posner, Eric C., and E. Glen Weyl (2019), 〈〈래디컬 마켓: 공정한 사회를 위한 근본적 개혁 (Radical Markets: Uprooting Capitalism and Democracy for a Just Society)〉〉, 박기영 옮김, 부키.

Samuel, A. L. (1959), "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal of Research and Development*, 3(3), July, 210-229.

Schapire, Robert E., and Yoav Freund (2014), *Boosting: Foundations and Algorithms*, Adaptive Computation and Machine Learning Series, The MIT Press.

Schlkopf, Bernhard, and Alexander J. Smola (2001), *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, 1 edition. Adaptive Computation and Machine Learning Series, The MIT Press.

Surowiecki, James (2005), *The Wisdom of Crowds*, First Paperback Edition edition, Time Warner Books Uk.

Tibshirani, Robert (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), January, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Vapnik, Vladimir N. (1998), *Statistical Learning Theory*, Wiley-Interscience.

Varian, Hal R. (2014), "Big Data: New Tricks for Econometrics," Journal of Economic Perspectives, 28(2), May, 3-28. https://doi.org/10.1257/jep.28.2.3

Zou, Hui, and Trevor Hastie (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), April, 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

A Short Guide to Machine Learning for Economists

Ki Young Park* · Jeong Won Ko**

Abstract

With the development of various AI (Artificial Intelligence) techniques and increased availability of big data, ML (Machine Learning) is expected to become the essential technology that would affect many aspects of our economy and society. With this in mind, the purpose of this paper is to provide an overview of ML techniques with emphasis on its application to economics. Contrasting the key differences in ML techniques and econometric methodologies, we first explain the key techniques used in supervised learning, which are widely used in industry and academia. Then we provide a survey of recent economic research that uses ML techniques and introduce debates on its impact on labor market and the value of data. We conclude with discussing the current limitations of ML technique in terms of economic research, while we believe that ML will fruitfully complement the current methodologies of economics.

KRF Classification : B030104, B039900
Key Words : AI (Artificial Intelligence), ML (Machine Learning), supervised learning, big data

* Corresponding Author, Associate Professor, School of Economics, Yonsei University, e-mail: kypark@yonsei.ac.kr
** Graduate Student, School of Economics, Yonsei University, e-mail: kojw_ye@yonsei.ac.kr