EXPERT INSIGHT

Machine Learning for Algorithmic Trading

퀀트 투자를 위한 머신러닝·딥러닝 알고리듬 트레이딩 2/e

파이썬, Pandas, 텐서플로 2.0, Scikit-learn을 활용한 효과적인 트레이딩 전략

Second Edition

스테판 젠슨 지음 홍창수·이기홍 옮김

출판사 책소개

머신러닝과 딥러닝 기술을 이용해 알고리듬 트레이딩의 아이디어에서 실행까지 전반적인 프로세스를 서술하는 좋은 안내서다. 2 판에서는 전략 백테스팅, 오토인코더, 적대적 생성 신경망(GAN), 이미지 형식으로 변환된 시계열에 합성곱 신경망(CNN) 적용과 같은 최신 내용을 추가했다. 또한 금융 시장 분석으로 100 개 이상의 알파 팩터를 설명하고 있으며, 부록으로 기술적 분석에 관한 모든 것을 수록했다. 세계적으로 저명한 AI 콘퍼런스인 NeurIPS 에서 발표된 최신 금융 딥러닝 연구를 소개하고 구현한다.

저자소개

저자: 스테판 젠슨 Stefan Jansen

어플라이드 에이아이(Applied AI)의 창립자이자 CEO 다. 포춘지 선정 500 대 기업, 투자 기업 및 업계 전반에서 데이터와 AI 전략에 대해 조언하고, 데이터 과학 팀을 구성하고, 광범위한 비즈니스 문제에 대한 엔드투엔드 머신 학습 솔루션을 개발하고 있다.이전에는 국제 투자회사의 파트너이자 상무이사를 역임하며 예측 분석 및 투자 연구 업무를 구축했다. 15 개 시장에서 영업을 하는 글로벌 핀테크 기업의 고위 임원으로 신흥국 중앙은행에 자문, 세계은행과 상담하기도 했다.조지아 공대에서 컴퓨터 공학 석사 학위를, 하버드 및 자유 대학교 베를린에서 경제학 석사 학위를 취득했고 CFA 자격증을 보유하고 있다. 유럽, 아시아, 미주 지역에서 6 개 언어로 근무했으며 제너럴 어셈블리(General Assembly)와 데이터캠프(Datacamp)에서 데이터 과학을 가르쳤다.

역자: 홍창수

경북대학교에서 재무 전공으로 경영학 석사 학위를 받고, 한국외국어대학교에서「장외 개별주식옵션의 내재변동성에 관한 실증분석」으로 경영학 박사 학위를 받았다. 2000 년 외환선물㈜ 투자공학 팀 애널리스트로 첫 직장 생활을 시작한 이후로 리딩투자증권 파생상품운용 팀 옵션 딜러, 한국투자증권 리스크관리부 리스크퀀트로 근무했다. 한화투자증권에서 장외 파생상품겸영인가 취득에 참여한 이후 금융공학 팀, OTC 파생 팀에서 장외파생상품 담당 차장으로 근무했다. 이후 금융자산 평가회사인 NICE P&I㈜ 금융공학연구소 실장으로 근무했으며, 현재 정보사업본부에서 장외 파생상품 및 리스크 관련 금융공학 컨설팅 업무를 수행하고 있다. 지은 책으로는『장외 파생상품 실무입문』(서울경제경영, 2014)이 있다. SCI 급 저널인 PLOS ONE 을 비롯해 국내 유수의 재무금융 및 경영 학술지인 증권학회지, 재무관리연구, 금융공학연구, 경영연구 등에 재무 및 파생상품 관련 논문을 게재했다. 중앙대학교에서 금융공학 융합전공 강사로 '금융공학 실습' 과목을 강의했으며, 한국금융공학회 산학협력위원으로 활동하고 있다.

역자: 이기홍

카네기멜론 대학교에서 석사 학위를 받았고, 피츠버그 대학교 Finance Ph.D, CFA,FRM 이며, 금융, 투자, 경제분석 전문가다. 삼성생명, HSBC, 새마을금고 중앙회, 한국투자공사 등과 같은 국내 유수의 금융 기관, 금융 공기업에서 자산운용 포트폴리오 매니저로 근무했으며, 현재 딥러닝과 강화학습을 금융에 접목시켜 이를 전파하고 저변을 확대하는 것을 보람으로 삼고 있다. 저서로는『엑셀 VBA 로 쉽게 배우는 금융공학 프로그래밍』(한빛미디어, 2009)이 있으며, 번역서로는『포트폴리오 성공 운용』(미래에셋투자교육연구소,

2010),『딥러닝 부트캠프 with 케라스』(길벗, 2017),『프로그래머를 위한 기초 해석학』(길벗, 2018)과『실용 최적화 알고리즘』(에이콘, 2020),『초과 수익을 찾아서 2/e』(에이콘, 2020),『자산운용을 위한 금융 머신러닝』(에이콘, 2020),『실전 알고리즘 트레이딩 배우기』(에이콘, 2021),『존 헐의 비즈니스 금융 머신러닝 2/e』(에이콘, 2021) 등이 있다. 누구나 자유롭게 머신러닝과 딥러닝을 자신의 연구나 업무에 적용해 활용하는 그날이 오기를 바라며 매진하고 있다.

작가의 말

이 글을 읽고 있다면, 여러분은 아마 투자 산업을 포함한 많은 산업에서 머신러닝(ML)이 전략적인 역량이 됐다는 것을 알고 있을 것이다. ML 의 부상과 밀접한 관련이 있는 디지털 데이터의 폭발은 특히 투자에 강력한 영향을 미치고 있으며, 이미 정교한 모델을 사용해 정보를 처리해 온 오랜 역사를 갖고 있다. 이러한 추세는 계량 투자를 새로운 방식으로 접근할 수 있게 만들었으며, 재량적 거래 전략과 알고리즘 거래 전략 모두에 대한 데이터 과학의 수요를 증가시키고 있다.자산클래스 간 거래 범위는 주식과 국채에서 상품과 부동산에 이르기까지 광범위하다. 이는 매우 광범위한 새로운 대체 데이터 소스가 시장 위와 그 밖의 데이터, 과거에 대부분의 분석 노력의 중심에 있었던 기초 데이터와 관련될 수 있음을 시사한다.
머신러닝이나 데이터 과학을 성공적으로 적용하려면 개인 또는 팀 차원의 통계 지식, 컴퓨터 기술과 도메인 전문 지식이 통합돼야 한다. 다시 말해 올바른 질문을 하고, 답을 제공할 수 있는 데이터를 식별하고 이해하며, 결과를 얻기 위한 광범위한 도구를 배포하고, 올바른 결정을 내리는 방식으로 이를 해석하는 것이 필수다. 따라서 이 책은 머신러닝의 투자 및 트레이딩 영역에 대한 통합적 관점을 제공한다.

목차

3 부. 자연어 처리

출판사 서평

◈ 이 책에서 다루는 내용 :
◆ 투자와 트레이딩 문제를 푸는 머신러닝 기법 구현
◆ 시장, 기본, 대체 데이터를 활용한 알파 팩터 연구
◆ 지도학습, 비지도학습, 강화학습 모델 설계
◆ 판다스, 넘파이와 Scikit-learn 을 이용한 포트폴리오 위험과 성과 최적화
◆ 머신러닝 모델을 Quantopian 플랫폼 위에서 실전 트레이딩 전략으로 통합
◆ 신뢰성 있는 시계열 백테스트 방법으로 전략 평가
◆ 케라스, 파이토치와 텐서플로를 사용한 딥신경망 설계와 평가
◆ 강화학습을 이용한 OpenAI Gym 에서의 트레이딩 전략 개발

◈ 이 책의 대상 독자 ◈
금융시장에 대한 이해와 트레이딩 전략에 관심이 있는 애널리스트, 데이터 과학자, ML 엔지니어라면 이 책이 도움이 될 것이다. 또한 ML 을 활용해 더 나은 의사결정을 내리고자 하는 투자 전문가로서 가치도 찾을 수 있다.소프트웨어와 ML 에 대한 배경지식이 있는 경우 이 영역에 대한 소개자료를 생략할 수 있다. 마찬가지로 전문 지식이 투자에 관한 것이라면 이 책이 제공하는 다양한 배경의 금융 맥락 일부 또는 전부를 잘 알고 있을 것이다.이 책은 독자가 매우 역동적인 이 영역을 계속 배우고 싶어 한다고 가정한다. 이를 위해 깃허브(GitHub) 저장소 내의 각 장에 대해 README 파일에 연결된 수많은 학술 참고 자료와 추가 자료를 책의 끝에 제공한다.파이썬(Python) 3 과 넘파이, 판다스, SciPy 와 같은 과학 컴퓨팅 라이브러리를 사용하는 것이 편할 것이며, 계속해서 다른 많은 자료를 얻을 수 있을 것을 기대한다. ML 과 사이킷런에 대한 경험이 도움이 되겠지만 기본 워크플로를 간략하게 다루고 격차를 메꾸거나 더 깊이 파고들려면 다양한 리소스를 참조하길 바란다. 마찬가지로 금융과 투자에 기본 지식이 있다면 용어를 더 쉽게 사용할 수 있을 것이다.

◈ 이 책의 구성 ◈
ML 이 트레이딩 전략의 설계와 실행에 가치를 더할 수 있는 방법을 종합적으로 소개한다. 데이터 소싱과 전략 개발 프로세스의 다양한 측면과 ML 과제에 대한 다양한 솔루션을 다루는 4 개의 부로 구성돼 있다.
1 부에서는 머신러닝을 활용하는 트레이딩 전략 전반에 걸쳐 관련된 기본적인 측면을 다룬다.
1 장, '트레이딩용 머신러닝: 아이디어에서 주문 집행까지'에서는 ML 이 거래에서 중요한 이유와 방법을 요약하고 투자 프로세스를 설명하며, ML 이 가치를 높일 수 있는 방법을 간략히 설명한다.
2 장, '시장 데이터와 기본 데이터: 소스와 기법'에서는 거래소 제공 틱 데이터와 보고된 재무 정보를 포함해 시장 데이터를 소싱하고 활용하는 방법을 다룬다.
3 장, '금융을 위한 대체 데이터: 범주와 사용 사례'에서는 폭발적으로 증가하는 수의 출처와 공급자를 평가하기 위한 범주와 기준을 설명한다.

4 장, '금융 특성 공학: 알파 팩터 리서치'에서는 예측 시그널을 포착하는 데이터 변환을 생성하고 평가 프로세스와 팩터 성과를 측정하는 방법을 설명한다.
5 장, '포트폴리오 최적화와 성과 평가'에서는 전략의 실행에 따른 포트폴리오를 관리, 최적화, 평가하는 방법을 소개한다.

2 부에서는 근본적인 지도학습과 비지도학습 알고리듬이 엔드투엔드 워크플로의 맥락에서 트레이딩 전략에 어떻게 정보를 제공할 수 있는지 설명한다.
6 장, '머신러닝 프로세스'에서는 ML 모델의 예측 성과를 체계적으로 수립, 훈련, 튜닝, 평가하는 방법을 개략적으로 설명함으로써 단계를 설정한다.
7 장, '선형 모델: 리스크 팩터에서 수익률 예측까지'에서는 선형 회귀와 로지스틱 회귀를 추론과 예측에 사용하는 방법과 규제화를 사용해 과적합 위험을 관리하는 방법을 설명한다.
8 장, 'ML4T 작업 흐름: 모델에서 전략 백테스팅까지'에서는 지금까지 별도로 다룬 ML4T 워크플로의 다양한 구성 요소를 통합한다.
9 장, '시계열 모델'에서는 벡터 자기 회귀 모델뿐 아니라 변동성 예측에 대한 ARCH/GARCH 모델을 비롯한 일변량 및 다변량 시계열 진단과 모델을 다룬다.
10 장, '베이지안 머신러닝: 동적 샤프 비율과 페어 트레이딩'에서는 확률적 모델과 마르코프 체인 몬테카를로(MCMC) 샘플링 및 변분 베이즈 방법이 대략적인 추론을 어떻게 용이하게 하는지 제시한다.
11 장, '랜덤 포레스트: 일본 주식 롱/숏 전략'에서는 통찰력과 예측을 위한 비선형 트리 기반 모델을 구축, 훈련, 조정하는 방법을 설명한다.
12 장, '거래 전략 강화'에서는 그래디언트 부스트(Gradient Boost)를 소개하고 고성능 훈련과 예측을 위해 XGBoost, LightBGM, CatBoost 라이브러리를 사용하는 방법을 보여준다.
13 장, '비지도학습을 활용한 데이터 기반 리스크 팩터와 자산 배분'에서는 알고리듬 거래를 위해 차원 축소와 군집화를 사용하는 방법을 설명한다.

3 부는 텍스트 데이터에 초점을 맞추고 대체 데이터의 핵심 출처에서 고품질 시그널을 추출하기 위한 최첨단 비지도학습 기법을 도입한다.
14 장, '트레이딩을 위한 텍스트 데이터: 감성 분석'에서는 텍스트 데이터를 수치형으로 변환하는 방법을 보여주고 감성 분석을 위한 2 부의 분류 알고리듬을 대규모 데이터 세트에 적용한다.
15 장, '토픽 모델링'에서는 비지도학습을 사용해 많은 문서를 요약한 주제를 추출하고 분류 모델의 특성으로 텍스트 데이터를 탐색하거나 주제를 사용하는 좀 더 효과적인 방법을 제공한다.
16 장, '어닝 콜과 SEC 공시 보고서를 위한 단어 임베딩'에서는 신경망을 사용해 기존 텍스트 특성보다 의미론적 콘텍스트를 훨씬 더 잘 포착하고 텍스트 데이터에서 거래 시그널을 추출하는 매우 유망한 방법을 나타내는 단어 벡터 형태의 최신 언어 특성을 학습한다.

4 부에서는 딥러닝과 강화학습을 소개한다.
17 장, '딥러닝'에서는 4 부에서 사용할 가장 인기 있는 딥러닝 프레임워크인 텐서플로 2 와 파이토치(PyTorch)를 소개한다.
18 장, '금융 시계열과 인공위성 이미지를 위한 CNN'에서는 대규모의 비정형 데이터를 활용하는 분류 작업을 위해 매우 강력한 합성곱 신경망(CNN)을 다룬다.
19 장, '순환 신경망'에서는 순환 신경망(RNN)이 어떻게 시퀀스 대 시퀀스 모델링에 유용한지 보여준다. 이는 예측할 일변량 시계열과 다변량 시계열을 포함한다.

20 장, '조건부 위험 요인과 자산 가격 결정을 위한 오토인코더'에서는 고차원 데이터의 비선형 압축을 위한 오토인코더를 다룬다.
21 장, '합성 시계열 데이터를 위한 적대적 생성 네트워크'에서는 딥러닝에서 흥미로운 발전 중 하나를 소개한다.
22 장, '심층 강화학습: 트레이딩 에이전트의 구축'에서는 강화학습이 환경에 대응해 의사결정을 최적화하는 방법을 배우는 에이전트의 설계와 훈련을 어떻게 허용하는지 설명한다.
23 장, '결론과 다음 단계'에서는 학습한 교훈을 요약하고, 계속해서 학습하고 트레이딩 전략을 수립하고자 취할 수 있는 몇 가지 단계를 개략적으로 설명한다.
부록, '알파 팩터 라이브러리'에서는 거의 200 개의 인기 있는 금융 특성을 나열하고 근거를 설명하며, 이러한 특성을 계산하는 방법을 보여준다.

◈ 옮긴이의 말 ◈

금융시장의 흐름을 이해하려면 시장 데이터, 공시, 뉴스, SNS 등 수많은 요인을 분석해야 합니다. 데이터를 효율적으로 관리하기 위해 금융시장에서도 점차 머신러닝과 딥러닝 기술이 사용되고 있으며, 트레이딩 분야에도 활용도가 높아진 상황입니다. 이 책은 머신러닝과 딥러닝 기술을 이용해 알고리즘 트레이딩의 아이디어에서 실행까지 전반적인 프로세스에 초점을 맞춰 서술하고 있는 이 분야의 바이블이 될 수 있는 좋은 안내서입니다. 2 판의 새로운 내용은 저자가 말했듯 더 많은 알파 팩터 사례를 소개하고 있으며, 최신 금융 딥러닝 논문에서 발표된 중요한 내용을 꼼꼼히 구현했다는 점이 특징입니다. 실제로 각 장의 내용을 상당 부분 고쳐가며 개정판 수준으로 설명해 초기에 생각했던 번역 기간보다 8 개월 정도가 더 소요됐습니다. 1 판보다 번역 품질을 높이려고 했고, 의역으로도 이해가 쉽지 않은 부분은 각주를 많이 달아 독자의 이해를 높이려고 노력했습니다. - 홍창수

최근에 일부 계량 금융학자들이 서클을 이루면서 AI 와 금융을 융합하려고 노력하고 있다. 이 책의 저자 스테판 젠슨도 이러한 노력에 상당한 공헌을 하고 있으며, 새로운 AI 의 흐름을 적극 반영하고자 1 판을 개정한 2 판을 출간했다. 이 책은 머신러닝의 전통 이론에서 최신 이론뿐 아니라 최신 딥러닝을 금융에 응용하는 부분도 다루고 있어 계량 금융과 머신러닝의 금융 응용에 관심 있는 사람들에게 특히 도움이 될 것이다. 1 판에서의 실험적인 내용들이 2 판에서는 많이 정립돼 독자들이 더욱 체계적으로 접근할 수 있을 것이다. 특히 2020 년, 일각에서 상당한 공헌을 하던 퀀토피안(Quantopian)이 문을 닫으면서 계량 금융 커뮤니티에 커다란 충격을 줬는데, 많은 학자와 실무 경험자가 모여 혁신적이면서도 실전적인 연구와 작품들을 만들어내고 있던 상황이라 충격은 더욱 컸다. 특히 이들이 만든 라이브러리인 집라인(zipline)과 파이폴리오(pyfolio)는 실제 선진 대형 자산운용사에서 사용하는 수준을 과시하고 있었으므로 이들이 사장되는가 하는 아쉬움이 많았다. 심지어 이 책의 초판도 이들을 광범위하게 사용하고 있어 개정판에서는 어떻게 전개할지 궁금하기도 했다. 놀랍게도 저자는 자체적으로 집라인을 개선한 집라인 리로디드(zipline reloaded)를 만들어내 모든 내용을 살렸을 뿐 아니라 최신 파이썬 버전에서도 잘 작동하게 만들어 다시금 계량 금융 커뮤니티에 집라인 열풍을 일으키는 큰 공헌을 했다(참고로 요즘 핫한 인터넷 증권사인 알파카(Alpaca)와 인터랙티브 브로커(InteractiveBroker)와 같은 증권사가 거래 API 에 최근 집라인을 사용할 수 있도록 하고 있다). 저자의 이런 노력에 찬사를 보내며, 지속적인 개발이 이뤄지기를 바란다.
마지막으로 저자도 언급하듯이 방대한 내용인 만큼 모든 내용을 자세히 다루기는 힘들어 핵심 요점 위주로 서술돼 있으며 오히려 저자의 깃허브에 있는 노트북에 많은 내용을 실었다는 것을 다시 한번 상기시키고 싶다. 이 책을 통해 이론과 실전이 통합된 최첨단의 고급 계량 금융의 여정을 즐기기를 바란다. - 이기홍