

1. 독립된 주제로서의 금융 머신러닝

1.1 동기

- 수학적 이론과 실전의 조화로운 책의 필요성
- 금융이 사회에 도움이 해야 하는 바람 (사기꾼 etc)
- 투자와 ML의 응용은 굉장히 복잡함. '군중의 지혜'를 이기는 것은 자율 주행보다 훨씬 어려움. 특히 '백테스팅 과적합' 부분 확인

1.2 금융 머신러닝 프로젝트가 실패하는 주요 원인

실패의 공통점

- 시지프스 패러다임
백테스트 과적합으로 인한 FP 결론, 낮은 샤프 비율
- 메타 전략 패러다임
투자전략의 복잡도: 데이터 큐레이션, 처리, HPC 인프라구조, 소프트웨어 개발, 특성 분석, 실행 시뮬레이션, 백테스트 등의 작업 등이 필요함
저자가 아는 계량분석으로 성공한 모든 회사는 메타전략 패러다임을 적용함. 이 책은 개인의 대상이 아닌 팀을 대상으로 연구 매뉴얼을 제공함을 목적
금융 이외의 연구 성과에도 사용

1.3 책의 구조

1부: 데이터 구조화 방법

2부: 데이터 기반한 ML 기반. 우연으로 하지는 않음.

3부: 백테스트 & 결과가 잘못될 확률

4부: 특성 추출

5부: HPC 기법

1.3.1 생산체인으로 구조화

알파: 벤치마크 대비 초과 수익률.

베타: 벤치마크에 대한 초과 수익률의 민감도.

- 데이터 큐레이터: 자산군마다의 뉘앙스로 수집 정제 인덱싱 저장 수정 전달. 표, 계층화 구성, 정렬, 실시간 등.
- 특성분석가: 로데이터를 신호로 변환
- 전략가: 특성을 실질 투자 알고리즘으로 변환. 특성은 블랙박스에 의해 발견될 수 있지만, 전략은 화이트박스로 개발되어야 함.
- 백테스터: 다양한 시나리오 상황에서 투자 전략 수익률 측정. 과거 수익률은 여러 가능한 확률적 경로 중 하나. 여러 대안 시나리오가 평가되어야 함. → 제안된 전략의 약점과 강점에 부합되어야 함.
- 배포 팀: 자동화, 벡터화, 다중 연산, 분산 컴퓨팅, 고성능 컴퓨팅 등
- 포트폴리오 감독: 엠바고, 가상거래, 졸업, 재배포, 중단

1.3.2 전략 구성 요소에 따른 구조

- 데이터: GIGO, 다루기 힘든 데이터는 가치가 있을 수 있음
- 소프트웨어: 자신만의 클래스를 만드는 것은 가치가 있을 수 있음
- 하드웨어: HPC 전문가가 되는 것은 가치가 있을 수 있음
- 수학: 증명은 오래걸릴 수 있음. 실험 수학을 사용 & 실험 수학을 사용.
- 메타전략: 아마추어는 어떤 마법공식을 믿지만, 전문가는 대량 생산 전략을 개발.
- 과적합: 교차검증은 금융분야에 작동하지 않음. 과적합 가능성을 끊임없이 묻고, 과적합은 과학적 사기이며 시간 자원 기회의 낭비.

1.3.3 혼한 함정에 따른 구성

ML은 만병 통치약이 아니다. 통계적 요행을 패턴으로 혼동하거나, Low Signal to noise ratio와 결합하는 경우 False Discovery를 하게 될 수 있다.

범주	함정	해법	장
인신론적	시지프스 패러다임	메타 전략 패러다임	1
인신론적	백테스트	특성 중요도 분석	8
데이터처리	시간에 따른 표본 추출	거래량 시계	2
데이터처리	정수 미분	분수 미분	5
분류	고정된 시간 수평 레이블	삼중 배리어 기법	3
분류	방향성과 크기를 동시에 학습	메타 레이블링	3
분류	비 IID 표본 가중값	고유도 가중값: 순차적 부스트래핑	4
평가	교차 검정 누수	퍼지와 엠바고	7, 9
평가	워크 포워드 백테스트	조합적 제거 CV	11, 12
평가	백테스트 과적합	합성 데이터에 대한 백테스트: 축소 샤프 비율	10~16

1.6 Q&A

- ML이 금융에 어떻게 유용한가

많은 금융 연산은 정의된 규칙에 따라 결정을 내림. 컴퓨터는 주어진 규칙을 최대한 빨리 수행하는 것.

금융을 알고리즘화하는 흐름은 멈출 수 없음. 앞으로의 10년 후는 자동화 규칙을 따르는 것뿐만 아니라 스스로 판단하는 것까지 포함할 것으로 보임.

분야는 대출 승인, 채권 가격 산정, 회사 등급 분류, 인재 선발, 기업 이익 예측, 인플레이션 예상 등에 적용 가능.

- 금융 머신러닝과 계량경제학의 차이

- 계량 경제학: 근본 도구는 '다변량 선형 회귀'이다. 계량 경제학 때문에 경제학과 금융학이 지난 70년 동안 어떠한 의미 있는 진전도 없었다고 보는 시각도 존재한다.

- 금융머신러닝: 이론을 대체하는 것이 아니라 안내자 역할을 함. 비즈니스 관점의 성공을 위해서는 머신러닝이 필요할 수 있음.

- 머신러닝 알고리즘이 블랙박스라고 무시하는 의견

이 책은 ML알고리즘이 화이트 박스이고, 투명하고, 잘 정의 되어있고, 정확한 패턴 인식 함수임을 느끼게 한다.

- 특정 ML 알고리즘에 대한 설명은?

특정 ML을 사용하더라도 데이터 구조, 레이블, 가중값 계산, 정상성을 위한 변환, CV, 특성 추출, 특성 중요도, 과적합, 백테스트 등 공통된 일반적인 문제에 직면함. 이 책은 특정 모델이 아니라 위 부분에 대해 다룸.

- 책이 백테스트 과적합 문제에 집착하는 이유

백테스트 과적합은 모든 수학적 금융 문제에 있어 해결하지 못한 가장 중요한 과제 중 하나. 백테스트 과적합을 다루지 않는다면, 신뢰하기 어렵고 CAPM, APT(Arbitrage Pricing Theory), 자산배분기법, 위험관리 등 역시 false-discovery 가능성을 다루어야 함.

또한, 머신러닝에 있어 금융데이터는 짧고, snr 이 낮으며, 환경변수를 통제하며 실험할 수 없기 때문에 백테스트 과적합은 중요함.