하반기 인과추론 with causalML 방향

상반기 인과추론에 대한 여러 이론들을 배우면서 같이 공부하신 여러분 고생 많으셨습니다. 그런데 저는 알고 넘어가는 지식은 그렇게 오래 머릿속에 남지 않더라구요. 제가 고등학생 때, 좋아하던 리처드 파인만이라는 사람은 이런 말을 남겼는데요 “What I cannot create, I do not understand”, 만들지 못하면 이해하지 못한다.

즉, 여러분이 알고 있는 인과추론 지식은 전체의 절반이라고 생각해요. 나머지 반은 구현할 수 있는 능력인데 그 도구를 causalML로 선택해서 활용하려고 합니다.

앞으로 우리가 다룰 **CausalML**은 특히 **CATE (Conditional Average Treatment Effect)** 추정에 강점을 가진 패키지입니다.

CausalML의 기본 개념 정리

전통적인 머신러닝은 주로 "무엇이 그렇다 **(What is)**" 에 답하려는 데 비해,
인과관계 머신러닝은 다음과 같은 질문에 초점을 맞춥니다:

> "만약 ~라면 어떻게 될까**? (What if)**"

즉, 단순히 예측이 아니라 개입**(intervention)**의 효과를 알고자 하는 것이죠.

CausalML은 무엇을 위한 도구인가**?**

- **CausalML**은 '인과 발견'보다는 '효과 추정'에 집중된 도구입니다.
- 이 패키지는 원인-결과 관계가 이미 가정된 상태(도메인 지식, 사전 실험 등)에서, 해당 개입의 효과가 모집단 내에서 어떻게 달라지는지를 정량적으로 추정하는 데 유용합니다.
- 실제 공식 문서에서도 다음과 같은 점이 명시되어 있습니다:
  “개입 W가 결과 Y에 미치는 인과적 영향을, 공변량 X를 조건으로 하여 추정하는 것이 목적이다.” [Link]

현재 **CausalML**의 한계

- 문서화된 기능 대부분이 추정 알고리즘 구현에 집중되어 있습니다.
- 즉, 인과 구조를 탐색하거나 가설을 세우는 데 직접적인 도움은 적습니다**.**

그래서 어떻게 공부할까**?**

머신러닝을 처음 배울 때처럼, 우리도 학습 로드맵을 명확히하고 한 단계씩 나아가는 방식으로 접근하려 합니다.
언제든 의견 주세요!

causalML 패키지 모듈 리스트

- `causalml.inference.meta`: S-leaner, T-leaner 등 메타 leaner를 위한 모듈
- `causalml.inference.tree`: 업리프트 트리, 인과 포레스트와 같은 트리 기반 방법론을 위한 모듈
- `causalml.inference.nn` (및 `.tf`, `.torch`): CEVAE, DragonNet과 같은 신경망 모델을 위한 모듈
- `causalml.inference.iv`: 도구 변수 방법론을 위한 모듈
- `causalml.metrics`: AUUC, Qini 곡선 등 평가지표를 위한 모듈
- `causalml.dataset`: 합성 데이터 생성을 위한 모듈
- `causalml.feature_selection`: 인과추론과 관련된 특징 선택 방법론을 위한 모듈
- `causalml.optimize`: 처치 최적화 알고리즘을 위한 모듈

공부 방향

1. Synthetic 데이터 소개 및 활용 방법
    a. 왜 사용해야 하는지 (실제 데이터 셋을 구하기 힘듦)
    b. 패키지는 어떻게 설계 되어있는지
    c. np.random과 비교
    d. Causalml 기본 문법에 대한 소개

2. META Learner 소개 및 구현
    a. S-leaner (`BaseSRegressor`)
    b. T-leaner (`BaseTRegressor`)
    c. X-leaner (`BaseXRegressor`)
    d. R-leaner (`BaseRRegressor`)
    e. DR-leaner (`BaseDRLearner`)
    f. Synthetic 데이터를 이용해 패키지와 결과 비교

3. 업리프트(처치 효과) 소개
    업리프트 트리는 단순히 반응 예측이 아니라, '개입(마케팅 등)이 실제로 행동 변화 유발하는 집단'을 찾아내는 데 특화된 인과추론 도구
    a. `UpliftRandomForestClassifier` 소개 및 간단한 예시
    b. Criteo Uplift Dataset 간단 소개 [원본 Link] 및 트리 모델 적용
    c. 탐색한 업 리프트 sample?의 결과를 DR-leaner와 결과 비교
    d. 다양한 메트릭
        i. 업리프트 곡선 (Uplift Curves)
        ii. AUUC (Area Under Uplift Curve)
        iii. Qini 곡선 (Qini Curves)
4. 신경망 기반 방법론 소개 및 작동 원리 설명
    a. model
        i. CEVAE (Causal Effect Variational Autoencoder):
        ii. DragonNet
    b. IHDP(Infant Health and Development Program) 데이터 셋 소개

    c. 메타 러너와 차이 확인
    d. 차이가 큰 raw data를 하나 골라 심층 분석해보기
5. 도구변수
    a. 도구 변수의 조건 [Link] 내용 확인하고, 검증 방법 (찾기) 소개
    b. KIPP 학교 프로그램의 효과 데이터 소개 및 분석
        i. 음.. 고수들의 계량 경제학 책을 찾아보면 될듯

    c. “2SLS” VS “DRIV” [Link] - 각 모델의 결과 및 차이 확인
    d. 도구변수 주의해야 할 점
    주요 장애물은 `causalml` 구현이 아니라 선택한 도구 변수의 유효성을 식별하고 방어하는 개념적 작업을 강조하면 좋을 듯

6. 실무 사례 공유
    a. 본인이 구할 수 있는 데이터를 마스킹 처리 혹은 출처를 표기하지 않는 방식 등… 무리가 가지 않는 선에서 분석

주의 사항

모든 결과물은 notebook 형태로 공유 하기 [깃헙 예시]

코드 나열하는 방식은 절대하면 안되고, 본인만의 설명 필수

공유하는 목적도 있지만, 본인의 성장에 도움이 될 수 있는 방식이라면 어떤 것도 허용 가능 (유튭 영상 등)

요즘 회사에 나오는 주제.

1. 광고 reward를 클릭하는 유저는 재화 소비에 소극적인가?