- 2025 건강·돌봄 AI 센터 10월 세미나 -
인공지능 편향성 문제와 이를 해결하는 방법론들

◈ Title : 인공지능 편향성 문제와 이를 해결하는 방법론들
◈ Speaker : 김용대 교수(서울대학교 통계학과)
◈ Date : 2025. 10. 14

1 서론: AI 시스템에 내재된 비공정성 문제와 그 심각성

인공지능(AI) 기술은 신용 평가, 채용 심사, 의료 진단 등 사회적으로 중대한 의사결정 과정에 깊숙이 관여하고 있습니다. 그러나AI 모델은 학습 과정에서 과거 데이터에 내재된 사회적 편향(Bias)을 그대로 학습하여, 특정 그룹에 대해 차별적이거나 불공정한 결과를 초래하는 심각한 문제를 안고 있습니다. 이러한 비공정성은AI의 신뢰도를 저해하고 사회적 불평등을 심화시킬 수 있습니다.

주요 차별 사례:

- **재범 확률 예측(컴파스 사례):** 재범 확률 예측 시스템이 백인 피의자보다 흑인 피의자에게 더 높은 위험 점수를 부여하는 편향이 발견되었습니다.
- **언어 모델 편향(LLM/Word2Vec):** 대규모 언어 모델이나 단어 임베딩(Word2Vec)이 성별에 따른 직업 선호도(예: '의사'는 남성, '간호사'는 여성)와 같은 고정관념을 학습하여 편향된 결과를 생성합니다.
- **광고 노출 편향:** 흑인 이름으로 검색했을 때 범죄 관련 광고가 더 많이 노출되는 사례가 보고되었습니다.
- **채용 및 금융 서비스:** 이력서 심사, 신용카드 발급 승인, 숙소 가격 추천 등에서 인종이나 성별에 따른 차별적 결과가 나타났습니다.

2 공정한 AI (Fair AI)의 목표 및 공정성 정의

Fair AI는 데이터 편향으로 인해 발생하는AI의 불공정성을 완화하고, 모델의 예측 성능을 최대한

유지하면서도 **공정성 기준을 만족**시키는AI 모델 학습을 목표로 합니다.

민감 속성(Sensitive Attributes): 공정성 논의에서 다루는 핵심 요소로, 성별, 인종, 나이 등 사회적으로 공평하게 대우받아야 하는 속성을 의미합니다.

공정성의 다양한 정의: 공정성을 수학적으로 정의하는 방식은 다양하며, 크게 세 가지 범주로 분류될 수 있습니다.

분류	주요 개념	설명
그룹 공정성(Group Fairness)	통계적 평등	서로 다른 민감 그룹 간의 예측 결과 통계치(예: 참 긍정률, 오 긍정률)가 유사해야 함.
개인 공정성 (Individual Fairness)	유사한 개인은 유사하게 취급	유사한 특성을 가진 개인들에게는 유사한 예측 결과가 나와야 함.
인과적 공정성(Causal Fairness)	개입의 효과	민감 속성을 제거했을 때 예측 결과가 변하지 않아야 함.

3 공정한 알고리즘 학습 방법론(3가지 범주)

공정한AI 알고리즘을 학습시키기 위한 방법론은 데이터 처리 시점과 방식에 따라 세 가지 주요 범주로 나뉩니다.

Pre-processing (사전 처리):

모델 학습 이전에 원본 데이터를 변환하거나 수정하여, 데이터셋 내에서 민감 속성이 다른 속성들과 통계적으로 독립되도록 만드는 방식입니다.

In-processing (동시 처리):

모델 학습 과정 자체에 개입하는 방법입니다. 손실 함수(Loss Function)를 최소화하는 과정에서 **공정성 제약 조건(Fairness Constraint)**을 추가하여 최적화를 진행합니다.

Post-processing (사후 처리):

이미 학습이 완료된 편향된 모델의 예측 결과를 최종적으로 조정하거나 변환하여 공정성을 확보하는 방식입니다.

4 핵심 연구: 베이지안 기반Fair Clustering 알고리즘

연구팀은 특히 클러스터링(군집화) 과정에서 발생하는 편향 문제를 해결하기 위해 **베이지안 방법**을 활용한 **Fair Clustering** 알고리즘을 제안했습니다.

기존 클러스터링의 문제점: 일반적인 클러스터링 기법은 데이터의 자연스러운 분포를 따르기 때문에, 인종이나 성별과 같은 민감 속성에 따라 클러스터 내의 그룹 비율이 불균형하게 나뉘는 편향이 발생할 수 있습니다.

제안된Fair Clustering의 작동 원리: 제안된 방법은 베이지안MCMC(Markov Chain Monte Carlo) 기법을 사용하여 **매칭 함수(Permutation)**를 학습합니다. 이 학습을 통해 각 클러스터 내에 존재하는 민감 그룹(예: 남성, 여성)의 비율을 **완벽하게 동일하게** 맞추도록 제약합니다.

클러스터 개수(K)에 대한 시사점: 이 접근법의 중요한 발견은 공정성 제약 조건을 만족시키려는 시도가 **클러스터 개수(K) 자체에 영향**을 미친다는 점입니다. 따라서K를 사전에 고정하기보다는, 공정성 제약 조건을 만족시키면서K를 학습 과정에서 유연하게 결정하는 것이 더 효과적일 수 있음을 시사합니다.

현재 연구팀은 이 클러스터링 결과를 바탕으로 새로운 데이터 포인트에 대한 공정한 할당(Assignment) 방법을 연구 중에 있습니다