공정성: 편향의 유형

예상 시간: 5분

머신러닝 모델이라고 해서 본질적으로 객관적인 것은 아닙니다. 엔지니어는 학습 사례로 이루어진 데이터 세트를 입력하여 모델을 학습시키며 데이터의 사전준비와 선정에 사람이 관여하기 때문에 모델의 예측이 편향되기 쉽습니다.

모델을 만들 때 데이터에 나타날 수 있는 일반적인 사람들의 편향을 인식하여 영향을 최소화할 수 있도록 미리 조치하는 것이 중요합니다.

경고: 다음 편향 자료는 머신러닝 데이터 세트에서 흔히 발견되는 편향의 일부이며 *모든 편향을 다룬 것이 아닙니다*. 위키백과의 인지적 편향 카탈로그 (https://wikipedia.org/wiki/List_of_cognitive_biases)는 판단에 영향을 줄 수 있는 사람들의 100가지 편향 유형을 나열하고 있습니다. 데이터를 점검할 때 모델의 예측을 왜곡할 수 있는 잠재적인 편향 원인이 있는지 모두 살펴봐야 합니다.

보고 편향

보고 편향은 데이터 세트에 수집된 이벤트, 속성 및 결과의 빈도가 실제 빈도를 정확하게 반영하지 않을 때 나타납니다. 이 편향은 사람들이 '말할 필요도 없다고 느끼는' 일반적인 상황은 언급하지 않고 특별히 기억할 만하거나 특이한 상황만을 기록하려는 경향이 있기 때문에 발생합니다.

예: 인기 웹사이트에서 사용자가 제출하는 말뭉치를 바탕으로 도서 리뷰가 긍정적인지 또는 부정적인지 예측하도록 학습된 감정 분석 모델이 있습니다. 도서에 관해 별다른 의견이 없는 사람들은 리뷰를 제출할 가능성이 작기 때문에 학습 데이터 세트의 리뷰 대다수는 극단적인 의견(도서를 아주 좋아했거나 아주 싫어한 사람들의 리뷰)이 됩니다. 따라서 이 모델은 좀 더 미묘한 어휘를 사용한 도서 리뷰의 감정을 정확히 예측할 가능성이 작습니다.

자동화 편향

자동화 편향은 두 시스템의 오류율과 관계없이 자동화 시스템이 생성한 결과를 비자동화 시스템이 생성한 결과보다 선호하는 경향을 말합니다.

예: 톱니바퀴 제조업체에서 근무하는 소프트웨어 엔지니어가 톱니 결함을 파악하도록 학습한 새로운 '혁신적인' 모델을 배포하고 싶어 했지만 공장 감독자는 이 모델의 정밀도와 재현율이 사람 검사자보다 모두 15% 낮다고 지적했습니다.

표본 선택 편향

표본 선택 편향은 데이터 세트의 사례가 실제 분포를 반영하지 않는 방식으로 선정된 경우 발생합니다. 표본 선택 편향은 다음과 같은 여러 형태를 취할 수 있습니다.

- **포함 편향**: 선택된 데이터가 대표성을 갖지 않습니다.

예: 자사 제품을 구매한 소비자층을 대상으로 전화 설문조사를 하고 그 결과를 토대로 미래의 신제품 판매량을 예측하도록 학습된 모델이 있습니다. 경쟁업체 제품을 선택한 소비자는 설문조사 대상이 아니었기 때문에 결과적으로 이 소비자 그룹은 학습 데이터에 반영되지 않았습니다.

- **무응답 편향**(또는 **응답 참여 편향**): 데이터 수집 시 참여도의 격차로 인해 데이터가 대표성을 갖지 못합니다.

예: 제품을 구매한 소비자 샘플과 경쟁업체 제품을 구매한 소비자 샘플을 대상으로 한 전화 설문조사를 통해 미래의 신제품 판매량을 예측하도록 학습된 모델이 있습니다. 경쟁업체 제품을 구매한 소비자는 설문조사를 완료하지 않을 가능성이 80% 높았기 때문에 이들의 데이터가 샘플에서 실제보다 작게 표현되었습니다.

- **표본 추출 편향**: 데이터 수집 과정에서 적절한 무작위선택이 적용되지 않았습니다.

예: 제품을 구매한 소비자 샘플과 경쟁업체 제품을 구매한 소비자 샘플을 대상으로 한 전화 설문조사를 통해 미래의 신제품 판매량을 예측하도록 학습된 모델이 있습니다. 설문에서 소비자를 임의로 타겟팅하지 않고 선착순으로 이메일에 응답한 200명의 소비자를 선정했기 때문에 평균 구매자보다 제품에 관심이 많은 소비자가 다수 포함되었을 가능성이 높습니다.

그룹 귀인 편향

그룹 귀인 편향은 개인의 특성을 개인이 속한 그룹 전체의 특성으로 일반화하려는 경향을 말합니다. 이 편향의 두 가지 주요 양상은 다음과 같습니다.

- **내집단 편향**: *자신이 소속된 그룹* 또는 본인도 공유하는 특성을 가진 그룹의 구성원을 선호하는 경향입니다.

예: 소프트웨어 개발자 모집을 위한 이력서 심사 모델을 학습시키는 두 명의 엔지니어는 자신들과 같은 컴퓨터 공학 아카데미에 다녔던 지원자가 더 직무에 적합하다고 믿습니다.

- **외부 집단 동질화 편향**: *자신이 속하지 않은 그룹*의 개별 구성원에 관해 고정 관념을 갖거나 그들이 모두 동일한 특징을 가진다고 판단하는 경향입니다.

예: 소프트웨어 개발자 모집을 위한 이력서 심사 모델을 학습시키는 두 명의 엔지니어는 자신들과 같은 컴퓨터 공학 아카데미에 다니지 않은 지원자가 직무에 필요한 전문지식이 부족하다고 믿습니다.

내재적 편향

내재적 편향은 일반적으로 적용할 필요가 없는 자신의 정신적 모델과 개인적 경험을 바탕으로 가정할 때 발생합니다.

예: 동작 인식 모델을 학습시키는 중인 엔지니어가 '아니요'라는 단어를 나타는데 고개 가로 젓기 (https://wikipedia.org/wiki/Head_shake)를 기능으로 사용하려고 합니다. 하지만 일부 지역에서 고개를 가로 젓는 것은 반대로 '예'를 의미하기도 합니다.

내재적 편향의 일반적인 형태는 **확증 편향**으로 모델을 만드는 사람이 자기도 모르게 이미 가지고 있는 믿음이나 가설을 긍정하는 방향으로 데이터를 처리하는 것을 말합니다. 경우에 따라 모델을 만드는 사람이 자신의 원래 가설과 일치할 때까지 반복해서 모델을 학습시키기도 하는데 이를 **실험자 편향**이라고 합니다.

예: 엔지니어가 다양한 특징(키, 체중, 종, 환경)을 바탕으로 개의 공격성을 예측하는 모델을 만들고 있습니다. 이 엔지니어는 어릴 때 활동성이 강한 토이 푸들로 인해 불쾌한 일이 있었기 때문에 이후 항상 토이 푸들을 공격적인 종이라고 생각하고 있었습니다. 학습된 모델이 대부분의 토이 푸들이 상대적으로 유순하다고 예측했을 때 엔지니어는 크기가 작은 푸들이 더 공격적이라는 결과가 나올 때까지 모델을 여러 번 다시 학습시켰습니다.

주요 용어

- 자동화 편향 (https://developers.google.com/machine-learning/glossary#automation_bias)
- 포함 편향 (https://developers.google.com/machine-learning/glossary#selection_bias)
- 내집단 편향 (https://developers.google.com/machine-learning/glossary#in-group_bias)
- 내재적 편향 (https://developers.google.com/machine-learning/glossary#implicit_bias)
- 외부 집단 동질화 편향 (https://developers.google.com/machine-learning/glossary#out-group_homogeneity_bias)
- 확증 편향 (https://developers.google.com/machine-learning/glossary#confirmation_bias)
- 실험자 편향 (https://developers.google.com/machine-learning/glossary#confirmation_bias)
- 그룹 귀인 편향 (https://developers.google.com/machine-learning/glossary#group_attribution_bias)
- 무응답 편향 (https://developers.google.com/machine-learning/glossary#selection_bias)
- 보고 편향 (https://developers.google.com/machine-learning/glossary#reporting_bias)

- 표본 추출 편향 (https://developers.google.com/machine-learning/glossary#selection_bias)
- 표본 선택 편향 (https://developers.google.com/machine-learning/glossary#selection_bias)

고객센터 (HTTPS://SUPPORT.GOOGLE.COM/MACHINELEARNINGEDUCATION)

이전
← 동영상 강의 (https://developers.google.com/machine-learning/crash-course/fairness/video-lecture)

다음
편향 식별하기 → (https://developers.google.com/machine-learning/crash-course/fairness/identifying-bias)

3월 26, 2019에 마지막으로 업데이트되었습니다.