공정성: 편향 평가

예상 시간: 5분

모델을 평가할 때 전체 테스트 또는 검증세트를 기준으로 계산된 지표가 모델의 공정성에 관해 항상 정확한 모습을 보여주는 것은 아닙니다.

환자의 의료 기록 1000개로 구성된 검증세트가 있다고 가정하고, 그를 바탕으로 종양의 유무를 예측하는 새로운 모델을 만들었다고 합시다. 500개 기록은 여성 환자의 기록이고 나머지 500개는 남성 환자의 기록입니다. 다음 혼동행렬 (https://developers.google.com/machine-learning/glossary#confusion_matrix)은 전체 1,000개 사례의 결과를 요약한 것입니다.

참양성(TP): 16	거짓양성(FP): 4
거짓음성(FN): 6	참음성(TN): 974

$$\text{정밀도} = \frac{TP}{TP + FP} = \frac{16}{16 + 4} = 0.800$$

$$\text{재현율} = \frac{TP}{TP + FN} = \frac{16}{16 + 6} = 0.727$$

정밀도가 80%이고 재현율이 72.7%이므로 유망한 결과로 보입니다. 하지만 각 환자 세트에 관한 결과를 따로 계산하면 어떻게 될까요? 결과를 두 개(여성 환자와 남성 환자)의 개별 혼동행렬로 나누어 봅시다.

여성 환자 결과

참양성(TP): 10	거짓양성(FP): 1
거짓음성(FN): 1	참음성(TN): 488

$$\text{정밀도} = \frac{TP}{TP + FP} = \frac{10}{10 + 1} = 0.909$$

$$\text{재현율} = \frac{TP}{TP + FN} = \frac{10}{10 + 1} = 0.909$$

남성 환자 결과

참양성(TP): 6	거짓양성(FP): 3
거짓음성(FN): 5	참음성(TN): 486

$$\text{정밀도} = \frac{TP}{TP + FP} = \frac{6}{6 + 3} = 0.667$$

$$\text{재현율} = \frac{TP}{TP + FN} = \frac{6}{6 + 5} = 0.545$$

여성 환자와 남성 환자의 측정항목을 따로 계산하면 각
그룹 모델의 성과가 크게 다른 것을 알 수 있습니다.

여성 환자:

- 실제로 종양이 있었던 여성 환자 11명에 관해 모델은 10명의 환자를 양성으로 정확하게 예측하여 90.9%의 재현율을 보였습니다. 다시 말해 **이 모델은 여성 환자 사례의 9.1%에 관해서는 종양 진단을 놓친 것입니다**.
- 마찬가지로 모델이 여성 환자의 종양에 양성을 반환할 때 11개 중 10개 사례에서 정확했습니다(정밀도: 90.9%). 다시 말해 **이 모델은 여성 환자 사례의 9.1%에 관해서는 종양을 잘못 예측한 것입니다**.

남성 환자:

- 하지만 실제로 종양이 있었던 남성 환자 11명에 관해 모델은 6명의 환자만을 양성으로 정확하게 예측했습니다(재현율: 54.5%). 따라서 **이 모델은 남성 환자 사례의 45.5%에 관해서는 종양 진단을 놓친 것입니다**.
- 모델이 남성 환자의 종양에 양성을 반환할 때 9개 중 6개 사례에서만 정확했습니다(정밀도: 66.7%). 다시 말해 **이 모델은 남성 환자 사례의 33.3%에 관해서는 종양을 잘못 예측한 것입니다**.

이제 모델 예측에 내재하는 편향을 이해하고 모델을 일반 인구를 대상으로 한 의료용으로 공개할 경우 각 하위 그룹에 발생할 위험도 이해하게 되었습니다.

추가 공정성 리소스

공정성은 머신러닝 분야에서 상대적으로 새로운 하위 분야입니다. 머신러닝 모델에서 편향을 파악하고 최소화하기 위한 새로운 도구 및 기술 개발을 목표로 하는 연구와 이니셔티브에 관해 자세히 알아보려면 Google 머신러닝 공정성 리소스 페이지
(https://developers.google.com/machine-learning/fairness-overview/)

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

다음
프로그래밍 실습 →
(https://developers.google.com/machine-learning/crash-course/fairness/programming-exercise)

3월 26, 2019에 마지막으로 업데이트되었습니다.