공정성: 편향 식별하기

예상 시간: 10분

모델에서 데이터를 가장 잘 표현할 (https://developers.google.com/machine-learning/crash-course/representation/) 방법을 찾기 위해 데이터를 살펴볼 때 공정성 문제를 염두에 두고 편향의 원인이 될 수 있는 요소를 사전에 점검하는 것이 중요합니다.

어디에 편향이 숨어 있을까요? 데이터 세트에서 주의 깊게 확인해야 할 세 가지 위험 요소를 확인해 보세요.

특성 값 누락

데이터 세트의 다수의 예에서 값이 누락된 특성이 하나 이상 있는 경우 데이터 세트의 주요 특성 중 일부가 제대로 표현되지 않았음을 나타내는 지표일 수 있습니다.

예를 들어 아래 표는 pandas `DataFrame`에 보관되어 있고 `DataFrame.describe` (https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.describe.html)를 통해 생성된 캘리포니아 주택 데이터 세트 (https://developers.google.com/machine-learning/crash-course/california-housing-data-description)에 있는 특성의 하위 집합에 관한 주요 통계 요약을 보여줍니다. 모든 특성의 `count`가 17000이라는 것은 누락된 값이 없음을 나타냅니다.

	longitude	latitude	total_rooms	population	households	median_income	median_house_va
count	17000.0	17000.0	17000.0	17000.0	17000.0	17000.0	17000.0
mean	-119.6	35.6	2643.7	1429.6	501.2	3.9	207.3
std	2.0	2.1	2179.9	1147.9	384.5	1.9	116.0
min	-124.3	32.5	2.0	3.0	1.0	0.5	15.0
25%	-121.8	33.9	1462.0	790.0	282.0	2.6	119.4
50%	-118.5	34.2	2127.0	1167.0	409.0	3.5	180.4

	longitude	latitude	total_rooms	population	households	median_income	median_house_va
75%	-118.0	37.7	3151.2	1721.0	605.2	4.8	265.0
max	-114.3	42.0	37937.0	35682.0	6082.0	15.0	500.0

3가지 특성(`population`, `households`, `median_income`)의 count가 `3000`이라고 가정해 보겠습니다. 다시 말하면 각 특성에 14,000개의 누락된 값이 있는 것입니다.

	longitude	latitude	total_rooms	population	households	median_income	median_house_va
count	17000.0	17000.0	17000.0	3000.0	3000.0	3000.0	17000.0
mean	-119.6	35.6	2643.7	1429.6	501.2	3.9	207.3
std	2.0	2.1	2179.9	1147.9	384.5	1.9	116.0
min	-124.3	32.5	2.0	3.0	1.0	0.5	15.0
25%	-121.8	33.9	1462.0	790.0	282.0	2.6	119.4
50%	-118.5	34.2	2127.0	1167.0	409.0	3.5	180.4
75%	-118.0	37.7	3151.2	1721.0	605.2	4.8	265.0
max	-114.3	42.0	37937.0	35682.0	6082.0	15.0	500.0

14,000개의 값이 누락되어 가구의 평균 소득과 주택 가격의 중앙값을 정확히 연관시키기 훨씬 어려워졌습니다. 이 데이터의 모델을 학습하기 전에 누락된 값의 원인을 신중하게 조사하여 소득 및 인구 데이터 누락의 원인이 될 수 있는 잠재적인 편향이 없는지 확인하는 것이 좋습니다.

예기치 않은 특성 값

또한 데이터를 살펴볼 때 특이하거나 비정상적인 특성 값을 포함하는 예가 있는지 확인해 보아야 합니다. 이와 같이 예기치 않은 특성 값이 있다는 것은 데이터 수집 중에 문제가 발생했거나 편향을 일으킬 수 있는 기타 부정확성이 있음을 나타낼 수 있습니다.

예를 들어 캘리포니아 주택 데이터 세트에서 발췌한 다음 예를 살펴보겠습니다.

	longitude	latitude	total_rooms	population	households	median_income	median_house_value
1	-121.7	38.0	7105.0	3523.0	1088.0	5.0	0.2
2	-122.4	37.8	2479.0	1816.0	496.0	3.1	0.3
3	-122.0	37.0	2813.0	1337.0	477.0	3.7	0.3
4	-103.5	43.8	2212.0	803.0	144.0	5.3	0.2
5	-117.1	32.8	2963.0	1162.0	556.0	3.6	0.2
6	-118.0	33.7	3396.0	1542.0	472.0	7.4	0.4

예기치 않은 특성 값을 정확히 찾아낼 수 있나요?

정답을 보려면 드롭다운 화살표를 클릭하세요.

	longitude	latitude	total_rooms	population	households	median_income	median_house
1	-121.7	38.0	7105.0	3523.0	1088.0	5.0	0.2
2	-122.4	37.8	2479.0	1816.0	496.0	3.1	0.3
3	-122.0	37.0	2813.0	1337.0	477.0	3.7	0.3
4	-103.5	43.8	2212.0	803.0	144.0	5.3	0.2
5	-117.1	32.8	2963.0	1162.0	556.0	3.6	0.2
6	-118.0	33.7	3396.0	1542.0	472.0	7.4	0.4

4번 예의 경도 및 위도 좌표(각각 -103.5 및 43.8)는 미국 캘리포니아주에 속하지 않습니다. 이 좌표는 사실 사우스다코타주의 러시모어 국립 기념공원 (https://wikipedia.org/wiki/Mount_Rushmore)의 대략적인 좌표로 데이터 세트에 넣은 가짜 예입니다.

데이터 격차

특정 그룹이나 특성이 실제보다 과소 또는 과대 표현되는 모든 종류의 데이터 격차로 인해 모델에 편향이 생길 수 있습니다.

검증 프로그래밍 실습
(https://developers.google.com/machine-learning/crash-course/fairness/machine-learning/crash-course/validation/programming-exercise)
을 완료했다면 학습 세트와 검증세트로 나누기 전에 캘리포니아 주택 데이터 세트를 무작위로 섞지 않아서 확연한 데이터 격차가 생겼던 것을 기억하실 겁니다. 그림 1은 전체 데이터 세트에서 추출한 데이터의 하위 집합을 캘리포니아 북서부 지역만 나타내도록 시각화한 것입니다.

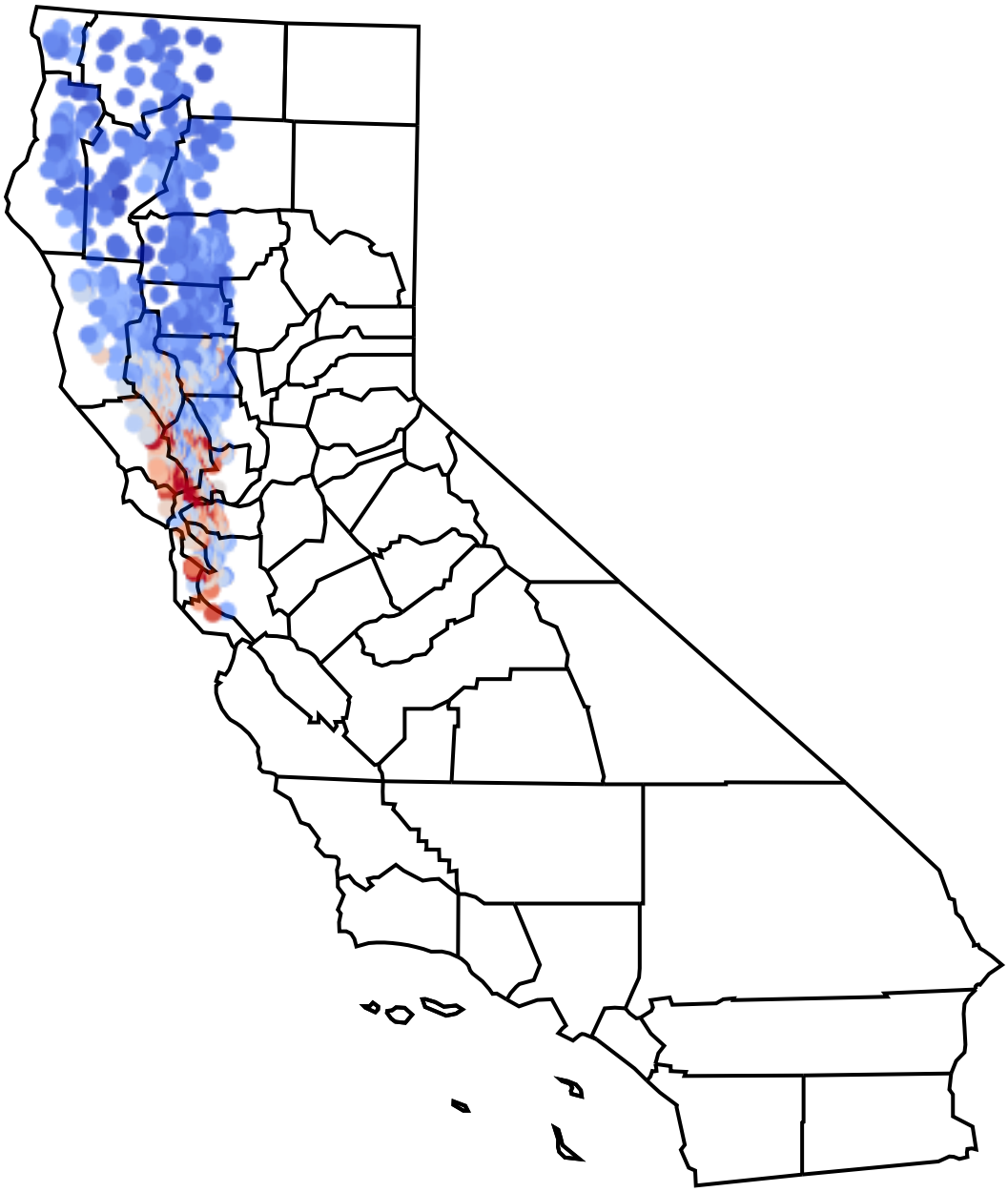

그림 1. 캘리포니아주 지도 위에 캘리포니아 주택 데이터 세트의 데이터를 오버레이한 그림. 각각의 점은 주택 단지를 나타내며, 파란색은 주택 가격 중앙값이 낮은 곳을, 빨간색은 주택 가격 중앙값이 높은 곳임을

나타냅니다.

이와 같이 전체를 잘 대표하지 못하는 샘플을 캘리포니아주 전체의 주택 가격 예측을 위한 모델을 학습하는 데 사용했다면 캘리포니아주 남부의 주택 데이터가 없는 것이 문제가 될 수 있습니다. 모델에 인코딩된 지리적 편향은 데이터에 표현되지 않은 커뮤니티의 주택 구매자에게 부정적인 영향을 미칠 수 있습니다.

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

이전
← 편향 유형
(https://developers.google.com/machine-learning/crash-course/fairness/types-of-bias)

다음
편향 평가하기 →
(https://developers.google.com/machine-learning/crash-course/fairness/evaluating-for-bias)

3월 26, 2019에 마지막으로 업데이트되었습니다.