## 6주차 - Handling Skewed Data

# Error Metrics for Skewed Classes

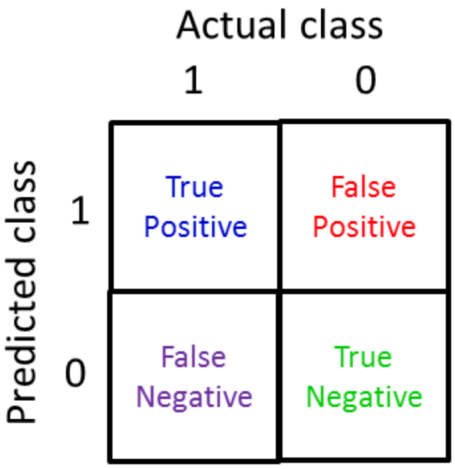
- Skewed Classes

분류문제에 있어서 주어진 test set의 y값 즉, class가 한쪽으로 몰려있는 상황.
예를들어, 스팸분류 문제 test set에 대부분은 스팸인 예제만 있는 상황.

- 문제점

accuracy로는 알고리즘의 성능을 정확히 판단할 수 없다.
왜냐하면 예를들어 위 상황처럼 test set에 대부분 스팸인 예제만 있다면 스팸으로 다 예측해 버리면 accuracy가 높게 나오기 때문이다.

- precision / recall

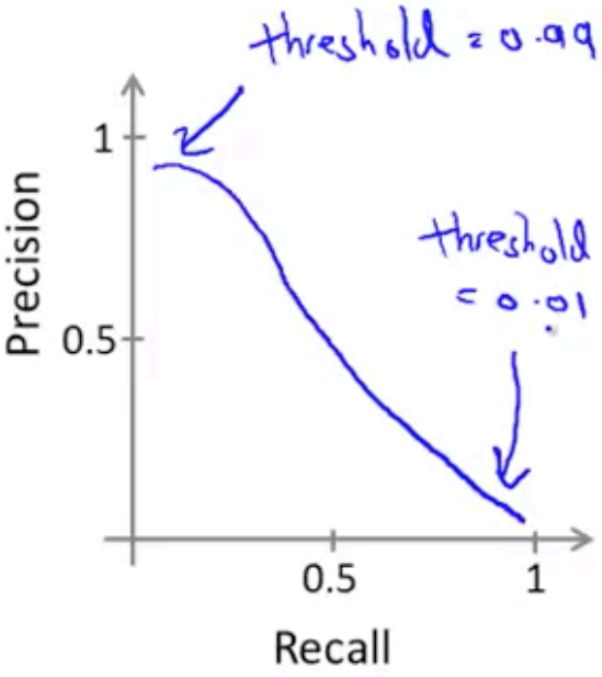



따라서 skewed classes가 있는 데이터로 평가하는 알고리즘에 대해서는 평가 기준을 좀 더 세분화 한다.
이진분류를 예를 들면 위의 그림처럼 4가지 경우를 나눌 수 있다.
precision = True Pos / (True Pos + False Pos)
정밀도는 참으로 예측한 데이터 중에서 실제로 참인 비율이고,
recall = True Pos / (Ture Pos + False Neg)
재현율은 실제 참인 데이터 중에서 참으로 예측한 비율이다.

좋은 성능의 알고리즘은 정밀도와 재현율 두가지 다 높아야 하며
스팸으로 다 예측해 버리는 것과 같은 엉터리 알고리즘은 상대적으로 재현율이 현저히 떨어지게 된다.

# Trading Off Precision and Recall

분류문제에서 문제 상황에 따라 precision이나 recall 중 더 중요한 기준이 있을 것이다.
예를 들어, 암을 예측할 때는 환자의 걱정을 덜어주기 위해서 정밀도를 더 중요한 기준을 생각할 수 있는데,
이런 경우에 threshold를 낮추거나 높임으로서 각 기준을 높이거나 줄일 수 있다.
threshould가 높아지면 y = 1 예측에 더욱 엄격해지므로 False Positive가 줄게 되고 precision이 높아지게 된다. 반면, True Negative가 증가하게 되므로 Recall은 줄어들게 되는 것이다.
반대로 threshould가 낮아지면 같은 이유로 recall은 증가하고 precision은 감소한다.




- trade off

서로 다른 알고리즘을 비교할 때 이 두가지 요소를 모두 고려해야하므로 비교가 어렵다.
따라서 한가지 기준을 두는 것이 빠른 결정에 도움이 되는데,
쉽게 생각할 수 있는 정밀도와 재현율의 산술평균은 극단적 값에 약하므로 이전에 예시를 들었던 것중 아무 이유 없이 모두 암으로 예측하는 것과 같은 알고리즘에 좋은 평을 주게 된다. 따라서 극단적인 값에 강한 조화평균을 사용한다.
F1 Score = 조화평균(Precision, Recall)
조화평균을 사용하면 둘 중 하나가 심각하게 0에 가까워지면 0의 값을 가지게 되므로 극단적인 값을 완화할 수 있다.

## How to compare precision/recall numbers?

| | Precision(P) | Recall (R) | Average | $F_1$ Score |
|---|---|---|---|---|
| Algorithm 1 | 0.5 | 0.4 | 0.45 | 0.444 |
| Algorithm 2 | 0.7 | 0.1 | 0.4 | 0.175 |
| Algorithm 3 | 0.02 | 1.0 | 0.51 | 0.0392 |

Predict y=1 all the time

Average: $\frac{P+R}{2}$

$F_1$ Score: $2\frac{PR}{P+R}$

<알고리즘을 고도화 시킬지 단순히 데이터를 추가할 지 고민할 때 방법>

1. 해당 문제의 전문가에게 주어진 feature만 보고 예측하게 해본다. (전문가는 아날로그 분류기라고 볼 수 있다.) -> 쉽게 예측이 가능하면 특성은 충분하니 데이터를 늘리면 된다.
2. learning curve를 그려서 error를 비교하고 충분한 데이터가 있는지 본다.(데이터가 충분하다면 train과 cv error가 비슷하게 낮을 것이다.