

혼돈 매트릭스(confusion matrix)

혼돈 매트릭스(Confusion Matrix)는 오차 행렬이라고도 하는데, 기계 학습(machine learning)에 의해서 생성된 분류 분석 모델의 성능을 지표화할 수 있는 테이블을 말한다.

훈련(training)을 통하여 예측(prediction) 성능을 측정하기 위하여 예측된 값과 실제 값을 비교하기 위한 표를 말한다.

True/False, 그리고 Positive/Negative으로

항목	설명
True/False	분류를 제대로 한 경우 True이다. 모델이 정답을 잘 맞췄는가의 여부이다.
Positive/Negative	분류 결과가 Yes이면 Positive라고 표현한다. 분류 결과가 No이면 Negative라고 표현한다.

모델에 의해서 예측한 값과, 정답 레이블의 값으로 표시된다.

정답 레이블 분류 결과	Yes	No	
Yes	제대로 분류함(TP)	제대로 분류 못함(FP)	
No	제대로 분류 못함 (FN)	제대로 분류함(TN)	

예를 들어, FN은 실제 정답이 yes인데, no로 판단을 했다는 의미이다.
즉, 정답을 못 맞췄다(F)이고, 예측치가 부정(N)이라는 의미이다.

모델의 성능 평가

모델의 성능 평가에 사용 가능한 여러 가지 지표들은 다음과 같은 것들이 있다.

항목	설명
Accuracy(정확도)	정분류율이라고 한다. 전체 데이터에서 제대로 된 분류가 얼마나 있는가? (TN + TP)를 전체 관측치로 나눈 비율을 의미한다. $(TN + TP) / (TN + FP + FN + TP)$
에러율(Inaccuracy))	오분류율이라고 한다. (FP + FN)를 전체 관측치로 나눈 비율을 의미한다. $(FN + FP) / (TN + FP + FN + TP)$
정밀도(Precision)	정확률 또는 적합율이라고도 한다. 분류 결과가 참인 항목 중에서 정말로 참으로 분류된 경우를 의미한다. "값이 클수록 잘못 분류된 개수가 적다"는 의미이다. TP를 예측치 YES인 항목(FP + TP)으로 나눈 비율을 의미한다. $TP / (FP + TP)$

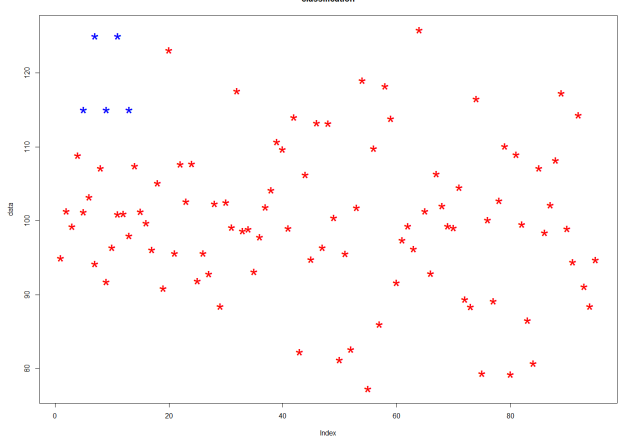
재현율(Recall)	민감도 혹은 True Positive Rate라고도 한다. 실제 정답이 참인 것 중에서 참으로 예측된 것들의 비율을 의미한다. $TP / (FN + TP)$
FR Rate	실제 정답은 거짓인데 참으로 잘 못 예측된 경우를 의미한다. $FP / (FP + TN)$

모델의 성능 평가에서 가장 많이 사용되는 항목은 Accuracy(정확도)이다.

에러율(Inaccuracy)은 모델의 오차 비율을 나타내는 척도로 수식은 (오분류율 = 1- 정확도)이다.

한쪽으로 쏠림 현상이 많은 데이터

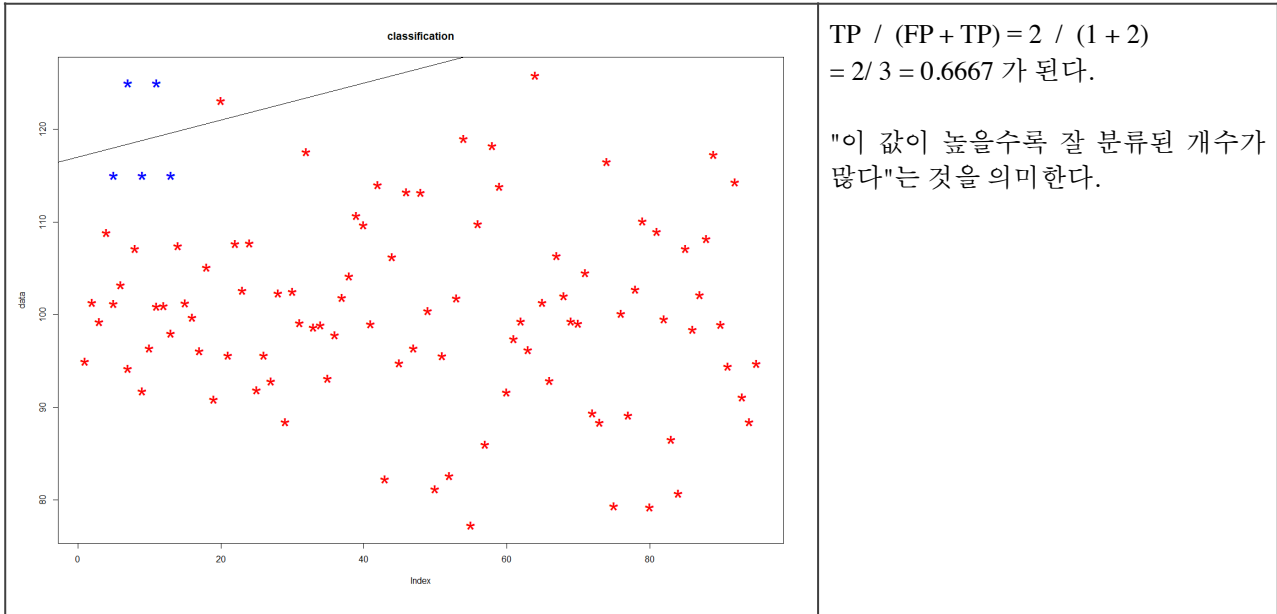
다음과 같은 데이터 셋이 있다고 가정하자.

그림	내용
	<p>다음과 같은 가정을 해보자. 가정 : "그림의 모든 데이터는 negative이다." 이렇게 말하는 경우 정확도는 95%가 된다. 하지만 전체를 Negative라고 분류하는 모델이 과연 좋은 모델이라고 말할 수 있나?를 의구심을 가진다. => 다른 지표가 필요하다.(정밀도, 재현율)</p>

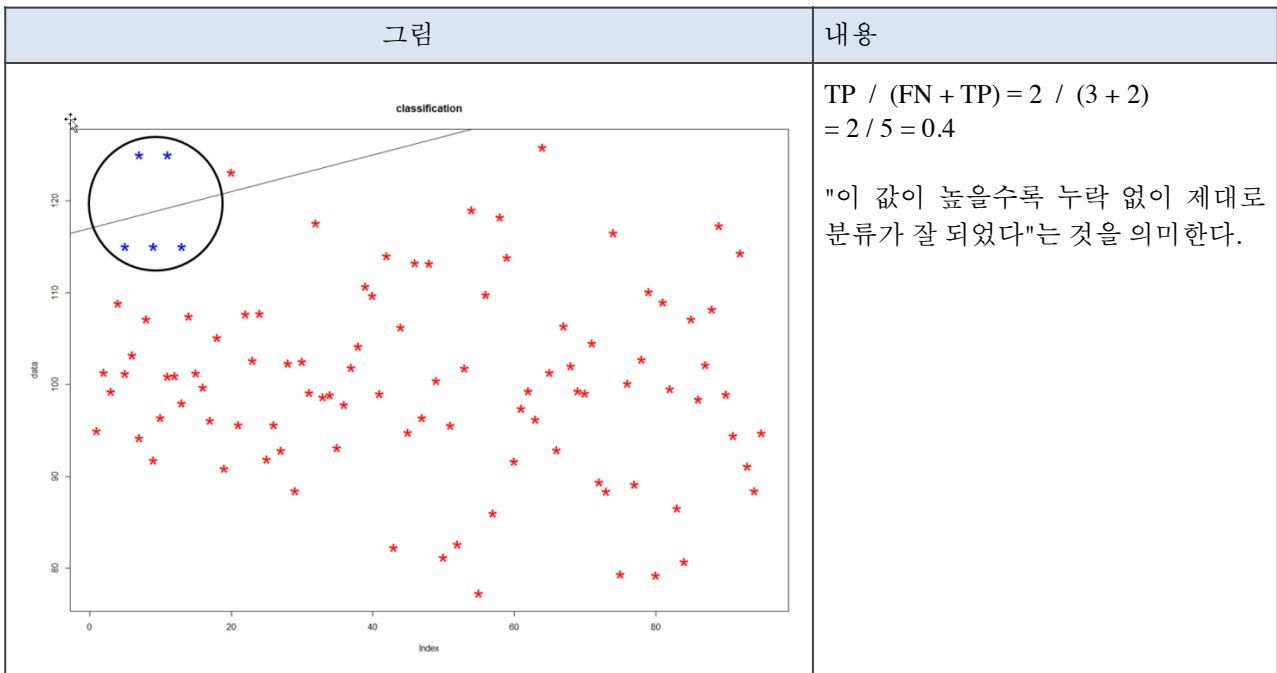
정밀도(정확률)은 모델이 Yes로 판단한 것 중에서 실제로 Yes인 비율을 말한다.

즉, 정밀도(Precision)는 TP와 FP에만 주목하는 비율이다.

그림	내용
----	----



재현율은 관측치가 Yes인 것 중에서 모델이 Yes로 판단한 비율이다.



정밀도(Precision)와 재현율(Recall)은 한쪽이 커지면, 다른 쪽은 상대적으로 낮아지므로 일반적으로 판단하기가 조금 힘들어 다음과 같은 지표를 제시한다.

다음의 예시를 살펴 보자.

예제 : 어느 모델이 좋은 모델인가

다음과 같은 2가지 형태의 모델이 있다고 가정하자.

어느 모델이 좋은 모델인가요?

항목	정밀도	재현율	평균
----	-----	-----	----

모델 A	0.6	0.39	$0.495 = (0.6 + 0.39) / 2$
모델 B	0.02	1.0	$0.51 = (0.02 + 1.0) / 2$

2개 모델의 산술 평균으로 보면 모델 B가 좋은 모델이라고 볼 수 있다.

그렇지만 모두 positive로 분류하는 모델에서 정밀도가 0.02이면 좋은 모델이라고 볼 수 없다.

F값(F measure)

F1 점수(score)라고도 한다.

두 분류기를 비교할 때 정밀도와 재현율을 하나의 숫자로 만들어 사용하면 편리할 때가 많다.

F1 점수는 정밀도와 재현율의 조화 평균이다.

F1 점수 공식

$$F1 = 2 / ((1/\text{정밀도}) + (1/\text{재현율})) = 2 * \text{정밀도} * \text{재현율} / (\text{정밀도} + \text{재현율})$$

참조 사이트

<http://freshrimpsushi.tistory.com/571>

<https://www.waytoliah.com/1222>

<https://datascienceschool.net/view-notebook/731e0d2ef52c41c686ba53dcdf346f32/>

ROC 곡선

ROC(Receiver Operating Characteristic) Curve는 **이진 분류의 진단 능력을 보여주는 곡선**이다.

ROC Curve는 x축, y축 모두 [0,1]의 범위의 값을 가지고, (0,0) 에서 (1,1)을 잇는 곡선이다.

FPR(False Positive Rate)은 특이도라고 하는 데, 0(False)인 케이스에 대해 1(True)로 잘못 예측한 비율을 말한다.

예를 들어서 당뇨병 환자가 아님에도 불구하고 당뇨병 환자라고 진단을 받는 경우이다.

TPR(True Positive Rate)은 민감도라고 하는 데, 1(True)인 케이스에 대해 1로 잘 예측한 비율을 말한다.

예를 들어서 당뇨병 환자를 진찰해서 당뇨병이라고 진단을 하는 경우이다.

ROC Curve는 x축에 FP Rate(FPR)를, y축에 TP Rate(TPR)를 표시한다.

ROC 곡선에서 왼쪽 상단의 계단 모양의 빈 공간이 분류 정확도에서 오분류(missing)를 의미한다.

ROC 곡선 그리는 절차

```
pr <- prediction( 예측_값, 정답_label )  
prf <- performance(pr, measure = "tpr", x.measure = "fpr")  
plot(prf, main='ROC Curve 그래프')
```

ROC 커브의 면적은 항상 0.5 이상, 1이하의 범위를 갖는다.

면적이 1에 가까울수록(좌측 상단의 꼭지점에 가까워질 수록) 좋은 성능이라고 보면 된다.

TPR과 FPR은 서로 비례하는 관계에 있다.

당뇨병 환자를 진단할 때, 의사는 아주 조금의 징후만 보여도 당뇨병 환자라고 얘기할 것이다.

이 경우 TPR은 1에 가까워짐과 동시에 FPR의 값도 1에 가까워진다.

반대로 초보 의사라서 당뇨병 환자를 잘 식별해내지 못한다고 하면, 모든 환자에 대해 당뇨병 환자가 아니라고 말할 것이다.

이 경우 TPR은 매우 낮아져 0에 가까워짐과 동시에 FPR 또한 0에 가까워질 것이다.

그런데 좋은 성능에 대한 지표인 TPR을 높이려다 보면, 나쁜 성능에 대한 지표인 FPR도 같이 높아져 버린다.

이러한 TPR과 FPR 측정을 시각화 한 것이 바로 ROC 커브이다.

이 ROC커브는 두 가지 장점이 있다.

먼저 그 커브의 면적을 재어 다양한 기준에서의 TRP과 FPR을 복합적으로 평가할 수 있다는 점이다.

또 한가지는 실제로 당뇨병 환자를 판단할 때, 어디를 기준으로 잡을 지 결정하는 데 도움이 될 수 있다.

예를 들어 보자.

질병에 걸릴 확률은 매우 낮지만 치사율이 극히 높은 병에 대해서는 일단 환자라고 의심을 할 수록 좋다.

반대로 질병에 걸릴 확률은 높지만 위험성이 매우 낮은 병은 FPR이 좀 낮은 기준을 선택하는 것이 괜찮을 것이다.

