분류 모델의 성능측정

– 확률적 해석 –

정확도 : 예측이 들어맞을 확률

확률적 경사 하강법을 이용한 분류 모델

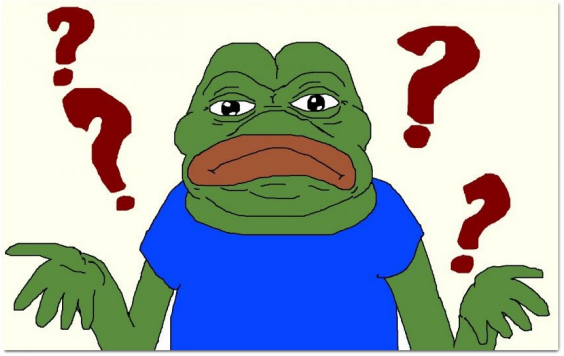
```
[19] from sklearn.model_selection import cross_val_score
     cross_val_score(sgd_clf, X_train, y_train_5, cv=3, scoring="accuracy")
```

```
array([0.95035, 0.96035, 0.9604 ])
```

모든 입력에 '5가 아님 ' 으로 예측한 모델

```
never_5_clf = Never5Classifier()
cross_val_score(never_5_clf, X_train, y_train_5, cv=3, scoring="accuracy")
```

```
array([0.91125, 0.90855, 0.90915])
```



데이터의 **90%**가 **5**가 아니기 때문.

오늘의 데이터

양성(positive) : ‘긍정’으로 예측한 것

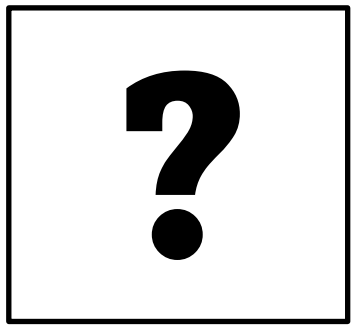
음성(negative) : ‘부정’으로 예측한 것

물음 : 이것은 강아지 입니까?

양성 : 그렇습니다. 이것은 강아지입니다.

음성 : 그렇지 않습니다. 이것은 강아지가 아닙니다.

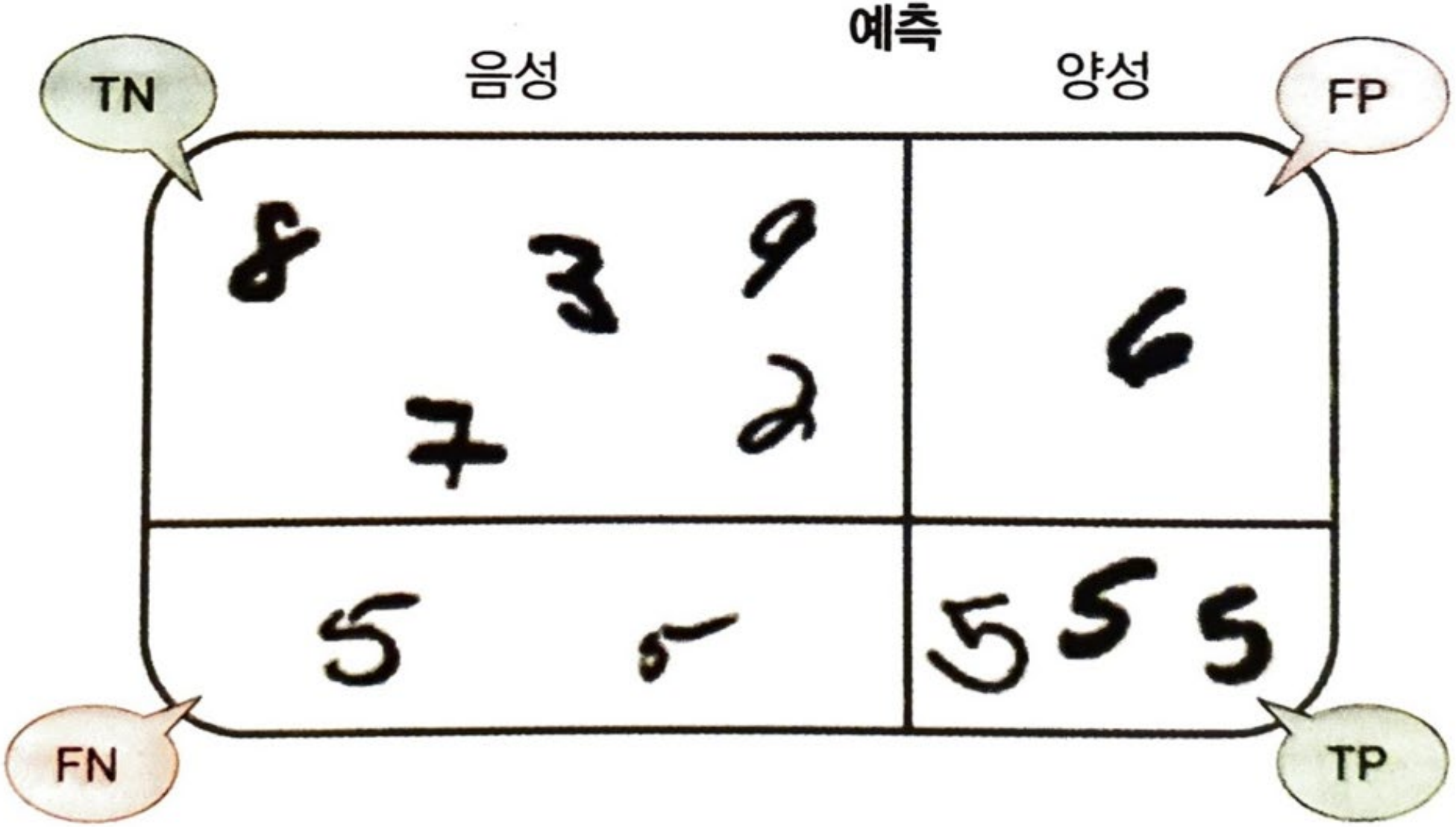
양성 / 음성의 의미는 물음에 의존한다.

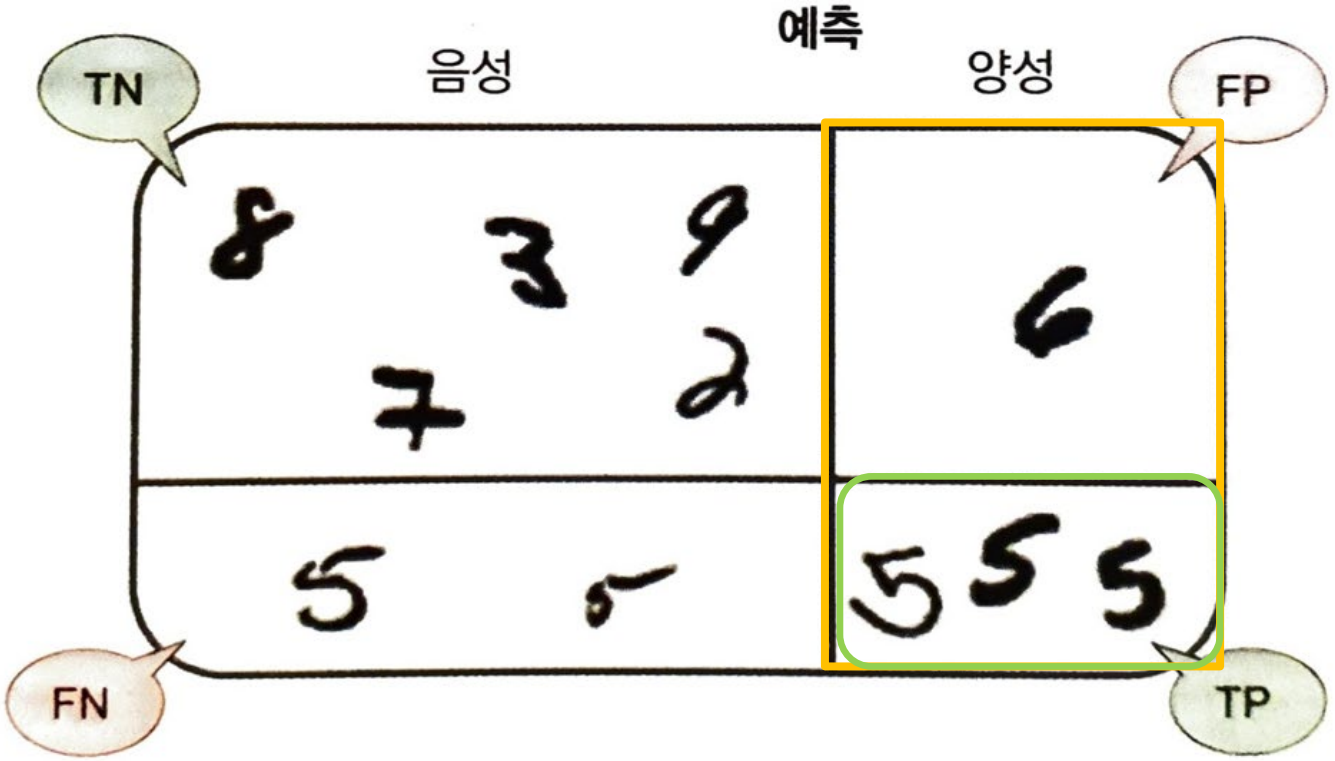
물음 :  는 5입니까?

P (positive, 긍정)　　N (negative, 부정)

T (True, 사실)　　F (False, 사실이 아님)

T P : 5일 것으로 예측했고 사실임 => 5
T N : 5가 아닐 것으로 예측했고 사실임 => 5가 아님
F P : 5일 것으로 예측했고 사실이 아님 => 5가 아님
F N : 5가 아닐 것으로 예측했고 사실이 아님 => 5


예측
음성
양성
TN
FP
FN
TP



정밀도 : $\frac{TP}{TP+FP} = \frac{TP}{P} = \frac{TP/n(\Omega)}{P/n(\Omega)} = \frac{P(T,P)}{P(P)} = \frac{P(T \mid P) * P(P)}{P(P)} = P(T \mid P)$

정밀도는 긍정이라 예측했을 때 그것이 사실일 확률이다.

재현율 = $\frac{TP}{TP+FN} = \frac{TP}{n("5")} = \frac{n(P,"5")}{n("5")}$ = **P**(P | ”5”)

재현율은 입력값이 5일 때 긍정으로 예측할 확률 (알아볼 확률)

정밀도가 중요한 경우




정밀도 : 2/2 = 100 %

재현율 : 2/3 = 67 %

재현율이 중요한 경우

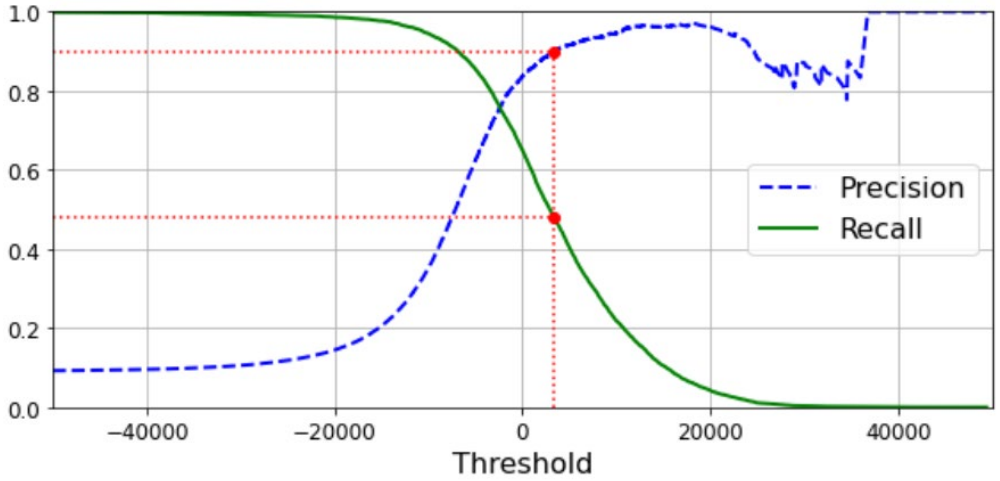


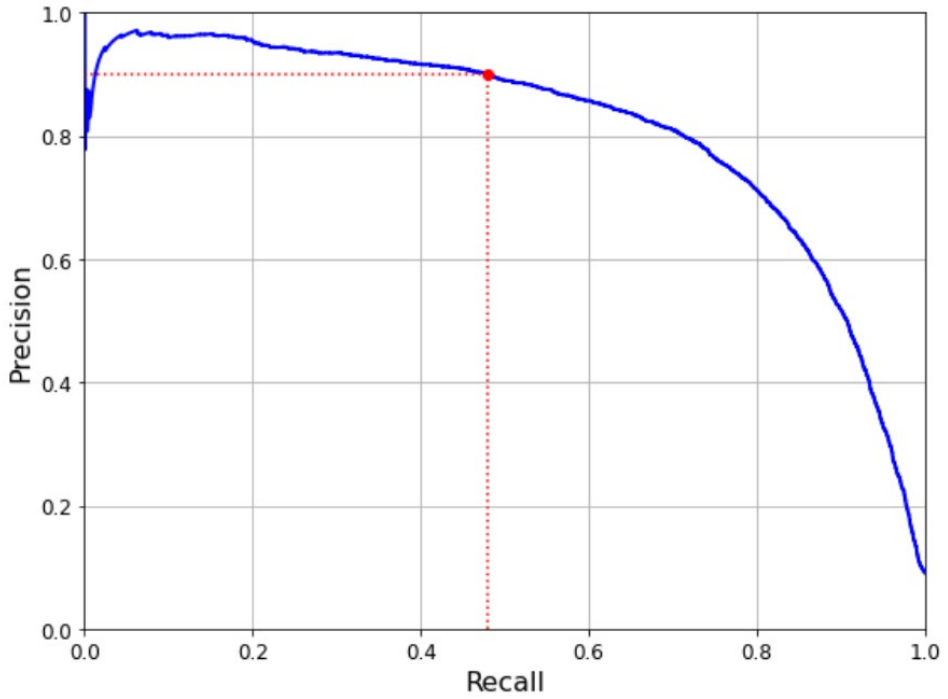

a b c

정밀도 : 4/5 = 80 %

재현율 : 4/4 = 100 %

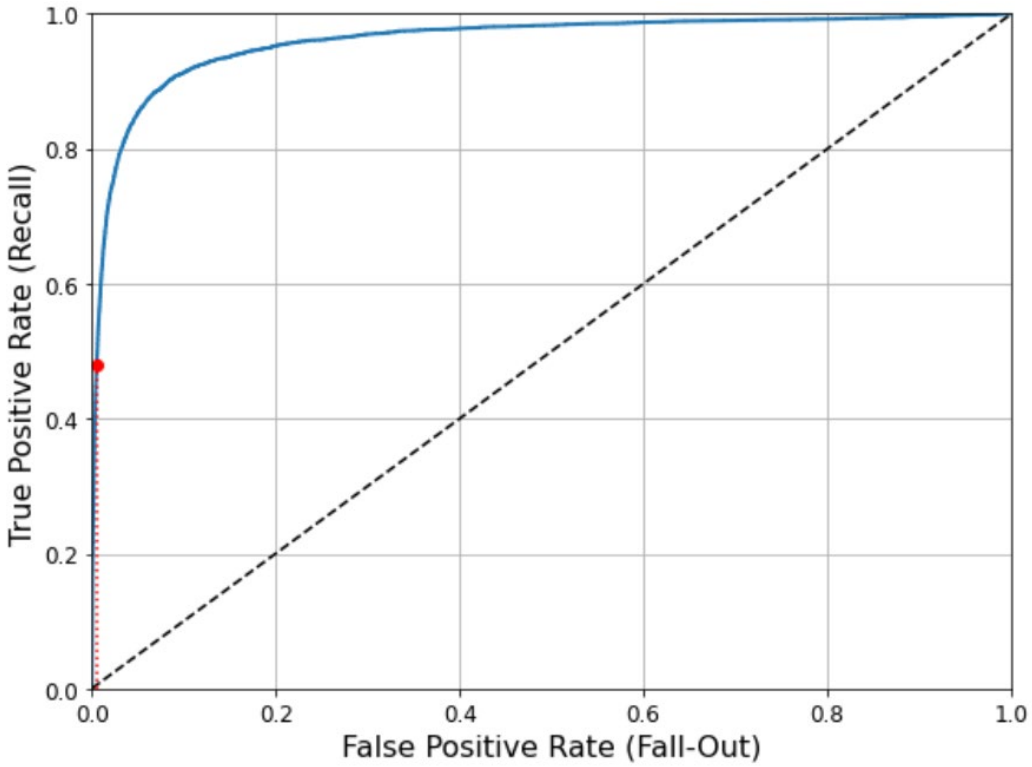
정밀도/재현율 트레이드오프



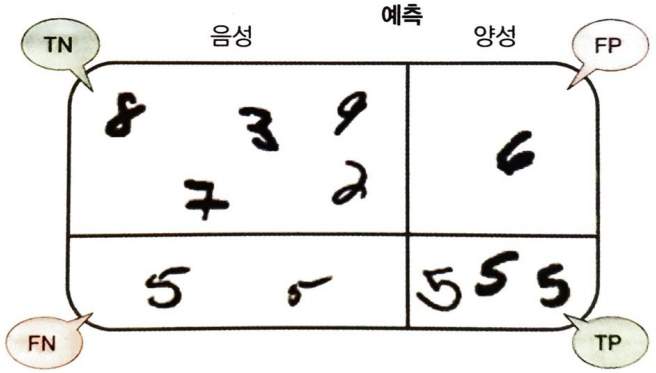


y = 정밀도 = f(재현율)

ROC 곡선 : y = 재현율 = f(FPR)



FPR : $\frac{FP}{FP+TN}$: 5가 아닌 수를 5라고 예측할 확률



재현율 = f($\frac{FP}{FP+TN}$) (y = 재현율)

= f($\frac{FP}{9(TP+FN)}$) 각 레이블의 비율이 일정하다고 가정

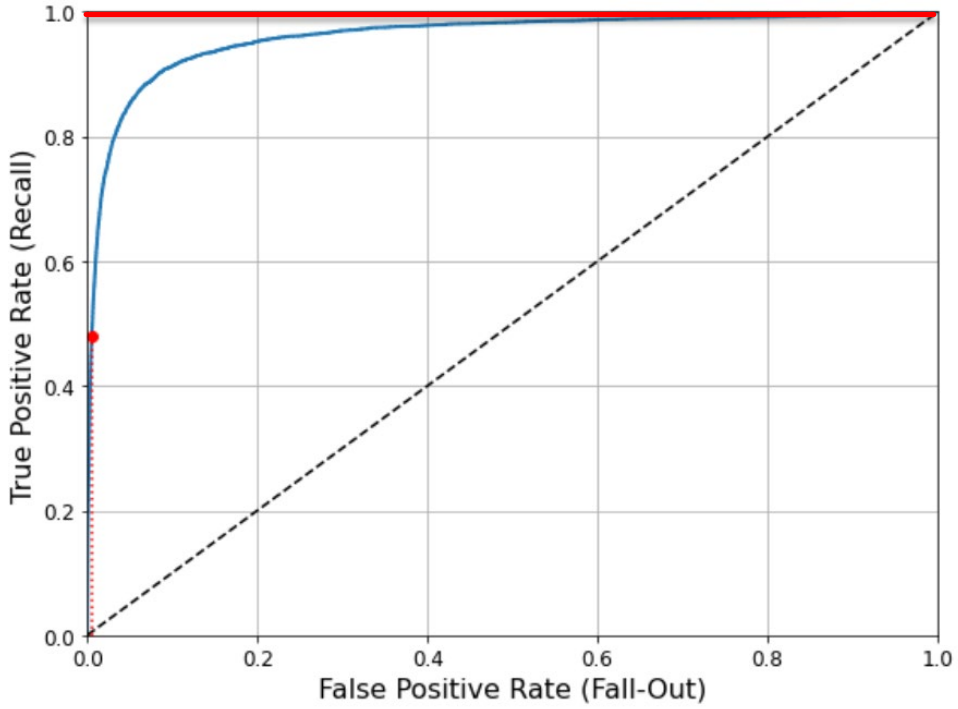
= f($\frac{1}{9}\frac{P-TP}{TP+FN}$)

= f($\frac{1}{9}$(10**P**(P)- 재현율))

y = f(FPR) = 9(x) +2y - 10**P**(P)

y = 10(**P**(P)-0.9(FPR))

곡선 아래 면적이 1인 경우



그래프 전체의 면적이 1이고
ROC 곡선은 순증가 함수이므로
FPR = 0 에서 재현율 = 1 이어야 함.

y = 10(**P**(P)-0.9(FPR)) 에서
1 = 10**P**(P) 이므로
P(P) = 0.1

100개의 표적이 한 색당 10개씩, 10개의 색으로 이루어져 있다 하자.
이때 정확히 10번 사격하여 빨간색 표적만을 전부 관통하는 경우.