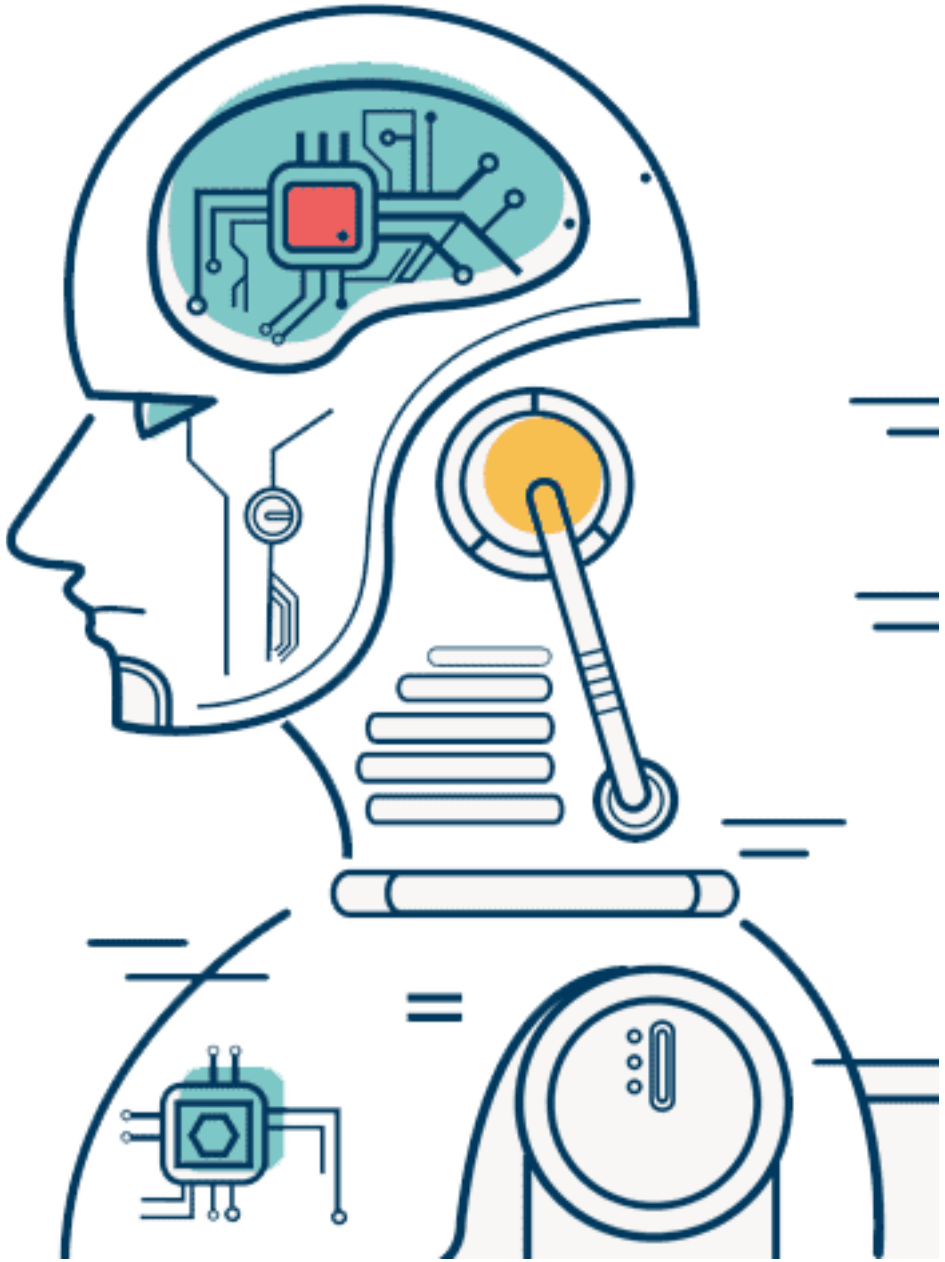
스마트인재개발원
Smart Human Resources Development

Machine Learning

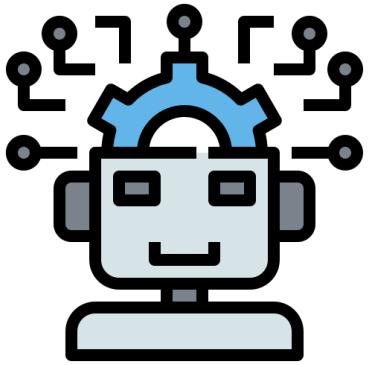


Chapter_6 지도학습
(Logistic Regression, SVM, 분류평가지표, GridSearch)

김은영

학습목표

- 선형 분류모델을 이해하고 사용할 수 있다.
- 다양한 분류평가 지표를 이해할 수 있다.
- GridSearch를 이용한 파라미터 튜닝을 할 수 있다.



Linear Model (Classification)

Linear Model – Classification

분류용 선형 모델

$$\hat{y} = w_1 x_1 + w_2 x_2 + w_3 x_3 + \cdots + w_p x_p + b > 0$$

- 특성들의 가중치 합 > 0 ➔ 클래스를 +1 (양성클래스)
- 특성들의 가중치 합 < 0 ➔ 클래스를 -1(음성클래스) 로 분류
- 분류용 선형모델에서 선형함수는 결정 경계 역할

분류용 선형 모델

- Logistic Regression

  → Regression 단어가 붙지만 분류용 모델

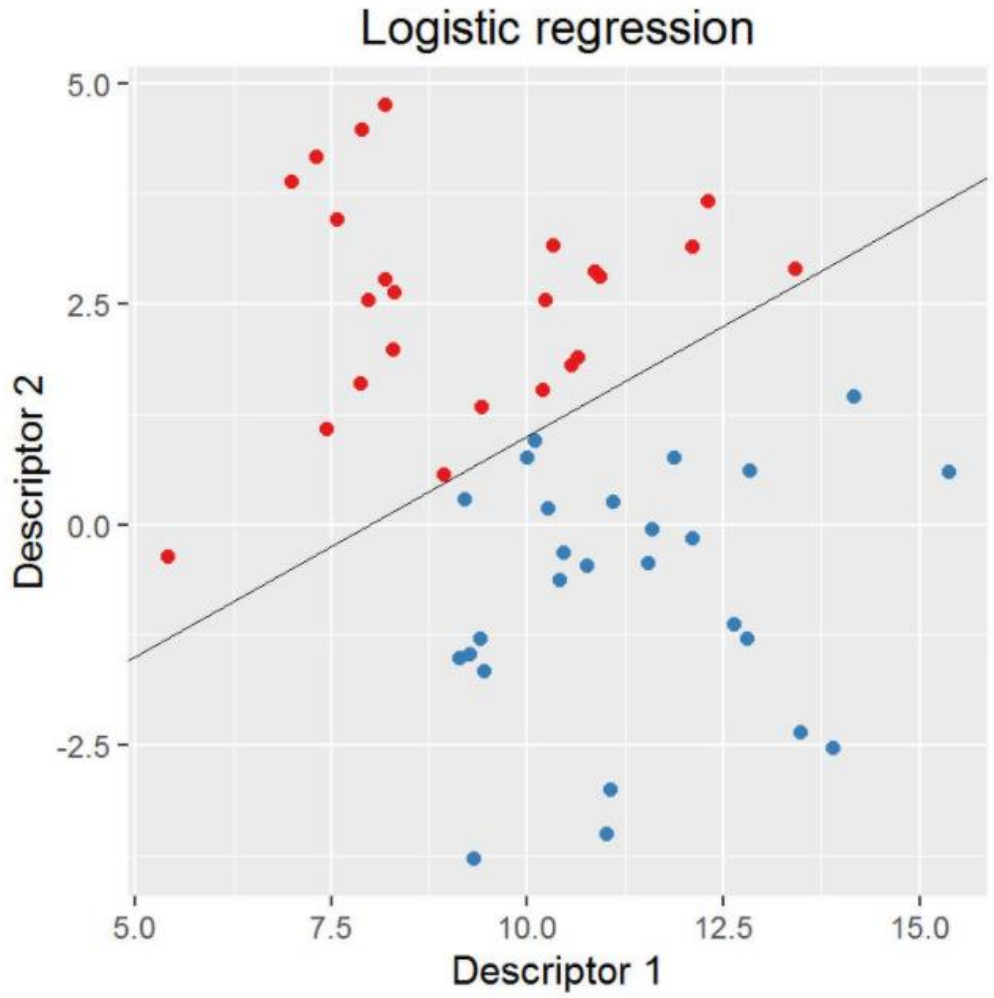
- Linear Support Vector Machines

Linear Model – Logistic Regression?

스마트인재개발원
Smart Human Resources Development

선형 모델 방식을 기반으로

이진 분류를 수행하는 모델



이름은 회귀(Regression)이지만

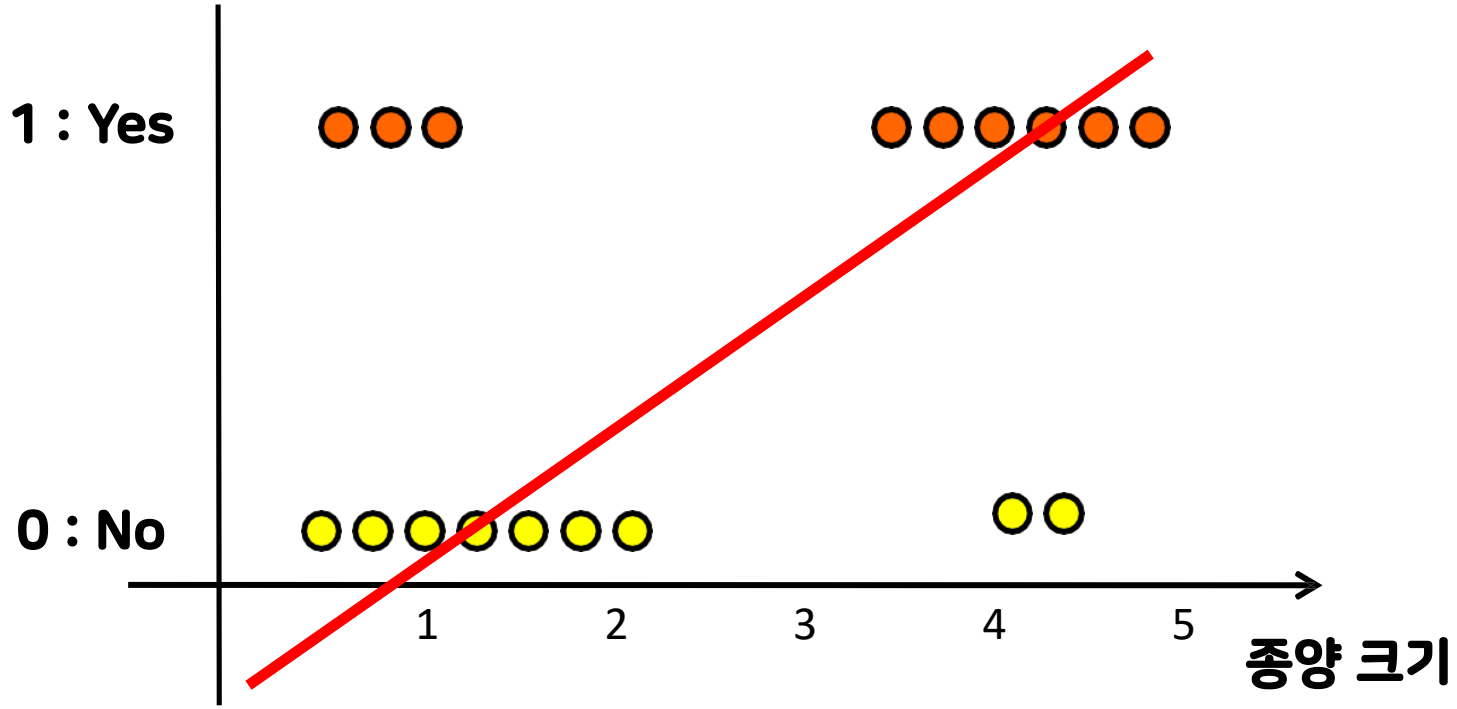
숫자 0과 1로 구분하는 분류 모델

선형회귀 직선을 사용하여 **두 집단을 분류** 할 수 있다는 점에서 착안

Linear Model – Logistic Regression

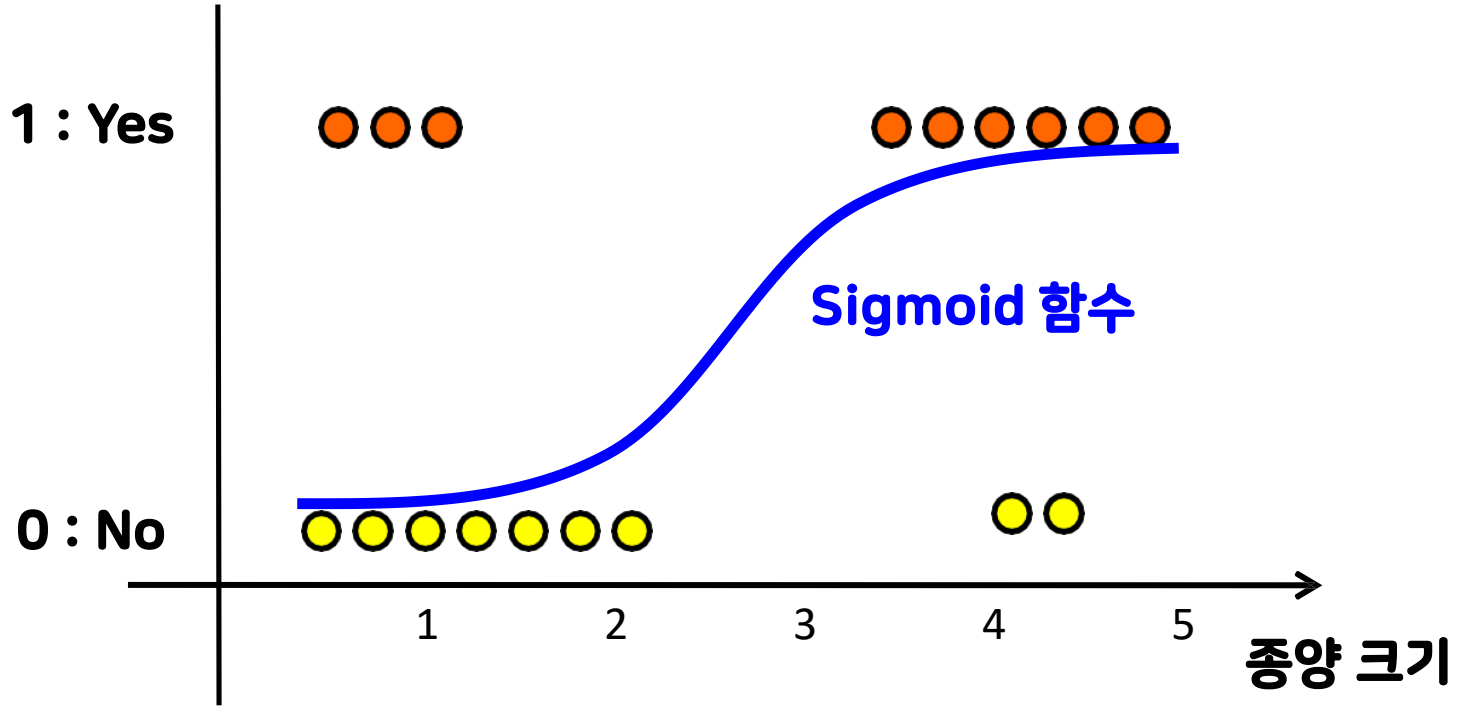
스마트인재개발원
Smart Human Resources Development

종양의 크기에 따른 악성종양여부(Y/N)를 예측하는 선형 분류 모델



선형 회귀 직선을 사용하여 분류하면
종양의 크기가 10일 경우 y값은 1보다 커지게 됨

종양의 크기에 따른 악성종양여부(Y/N)를 예측하는 로지스틱 회귀 모델

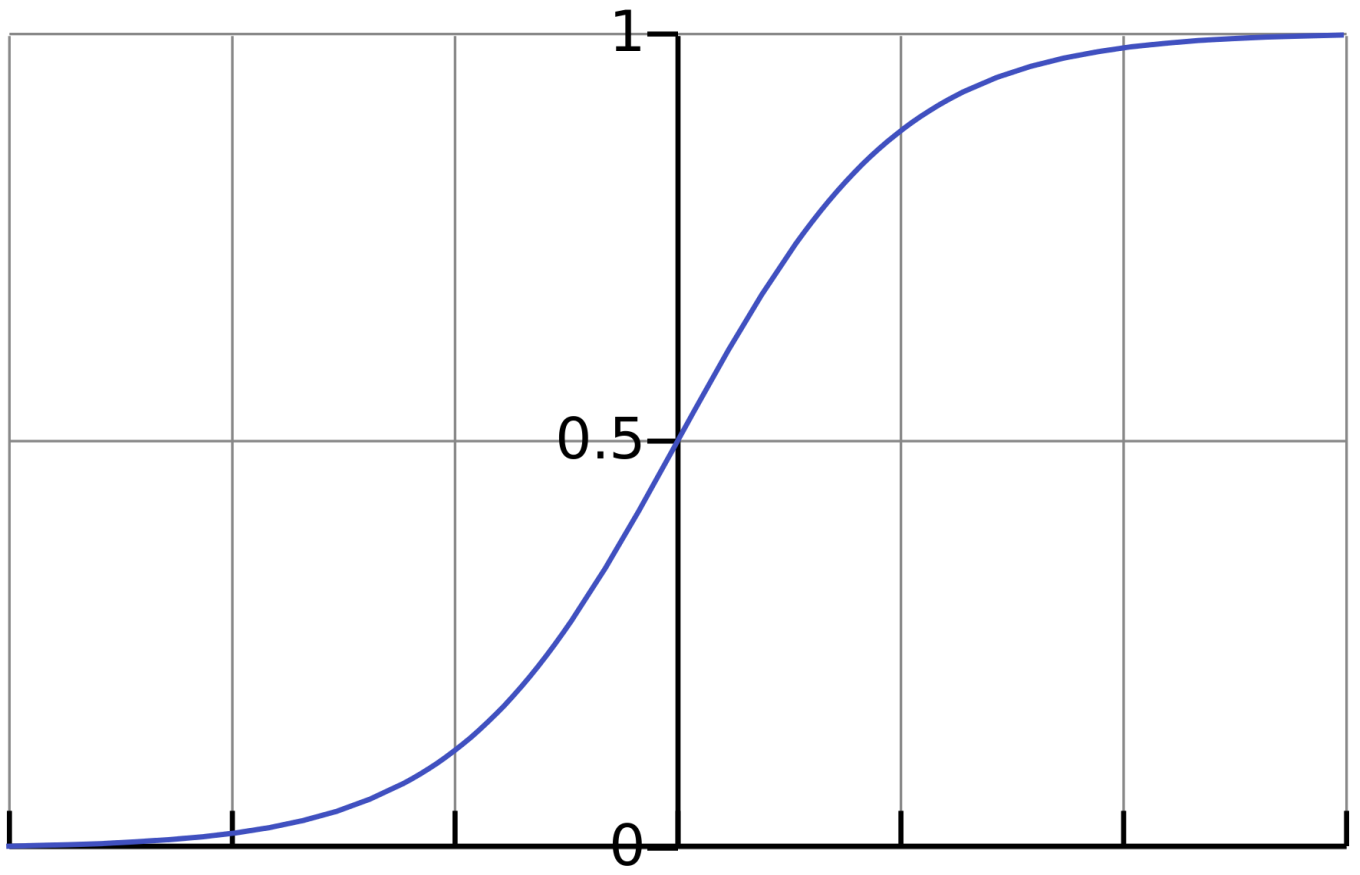


Sigmoid 함수를 사용하여 예측하면 0~1사이의 확률 정보로 표시 가능
0.5를 기준으로 낮으면 0, 높으면 1로 예측

Linear Model – Logistic Regression : sigmoid 함수

$$y = \frac{1}{1 + e^{-x}}$$

X	Y
∞	1
- ∞	0
0	0.5



교차 엔트로피 오차 함수 (Cross entropy error function)

$$E = -\sum_{n=1}^{N} t_n \log y_n \quad + (1-t_n)\log(1-y_n)$$

(N : 데이터 수, t (정답 1, 오답 0), y : 입력 데이터에 대한 출력)

주요 매개변수(Hyperparameter)

scikit-learn의 경우

LogisticRegression(C, max_iter)

- **규제 강도의 역수 : C**
  (값이 작을수록 규제가 강해짐)

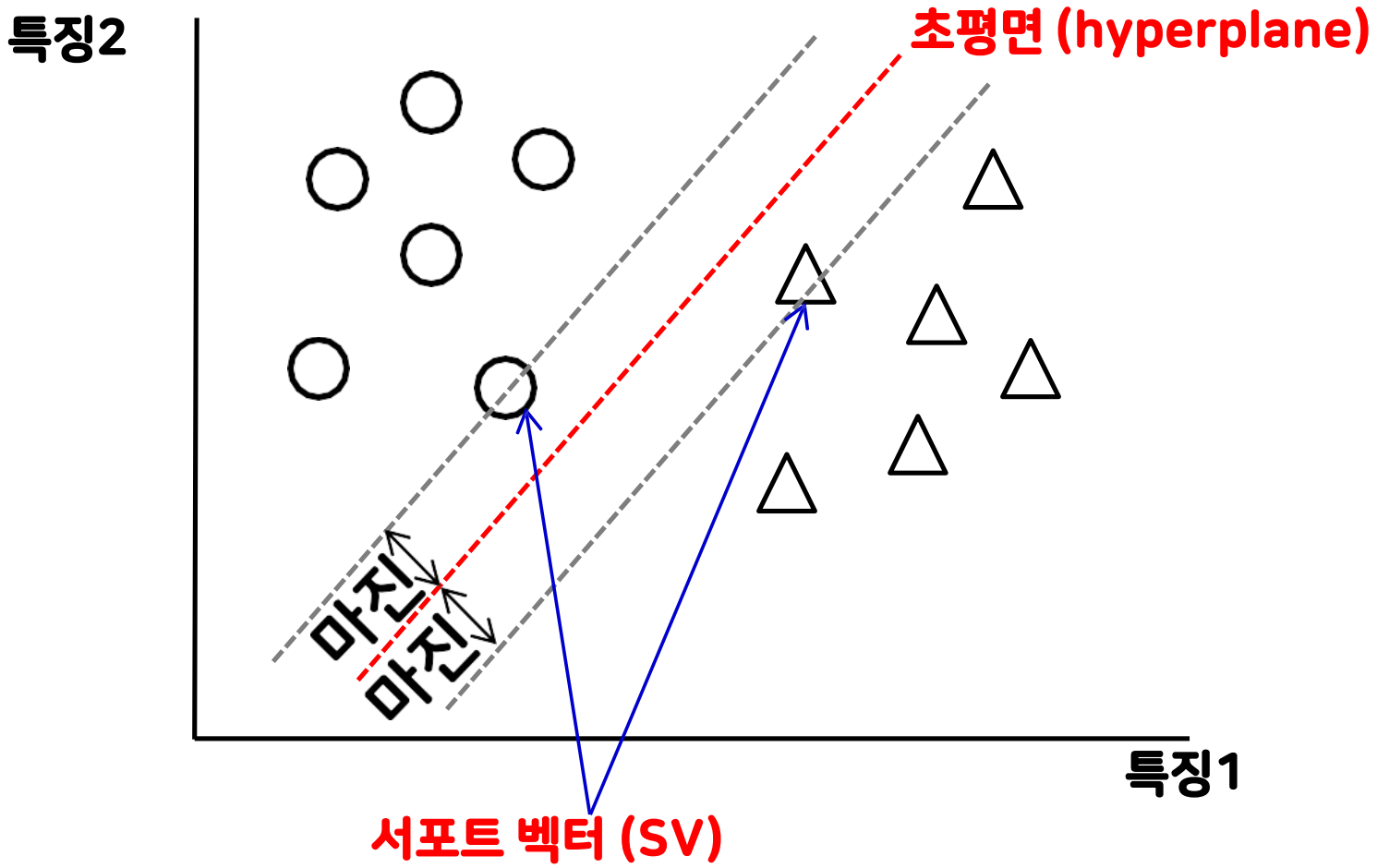
- **최대 반복횟수 : max_iter**
  (값을 크게 잡아 주어야 학습이 제대로 됨)

- **기본적으로 L2규제 사용, 중요한 특성이 몇 개 없다면 L1규제를 사용해도 무방**
  (주요 특성을 알고 싶을 때는 L1 규제를 사용하기도 함)

왜 선형 모델 방식을 분류에서 사용할까?

- 선형 모델은 간단한 함수식을 사용하므로 학습 및 예측 속도가 빠름
- 매우 큰 데이터 세트와 희소(sparse)한 데이터 세트에서도 잘 동작
- 특성이 많을수록 더 잘 동작
- 특성이 작은 데이터에서는 다른 모델이 더 좋은 경우가 많음.

Linear Model – Linear SVM(Support Vector Machines)



Linear Model – Linear SVM(Support Vector Machines)

스마트인재개발원
Smart Human Resources Development

주요 매개변수(Hyperparameter)

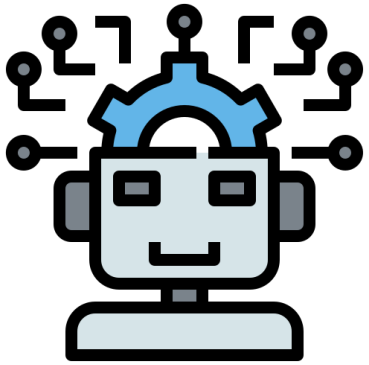
scikit-learn의 경우

LinearSVC (C)

- **규제 강도 : C**
  (값이 작을수록 규제가 강해짐)

- **기본적으로 L2규제를 사용, 하지만 중요한 특성이 몇 개 없다면 L1규제를 사용해도 무방**
  (주요 특성을 알고 싶을 때 L1규제를 사용하기도 함)

손 글씨 숫자 데이터 분류 실습

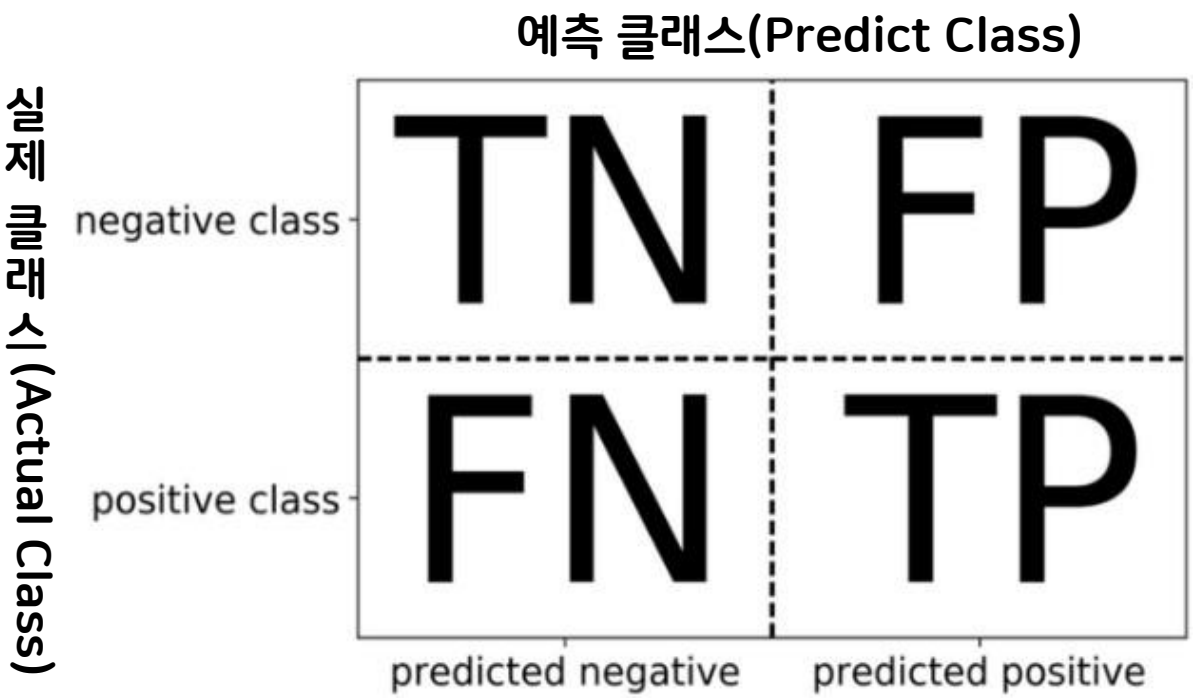


분류 평가 지표

분류 평가 지표 – 이진 분류

스마트인재개발원
Smart Human Resources Development

Confusion_matrix



True Negative(TN)
- 실제 False인 정답을 False라고 예측(정답)

False Positive(FP)
- 실제 False인 정답을 True라고 예측(오답)

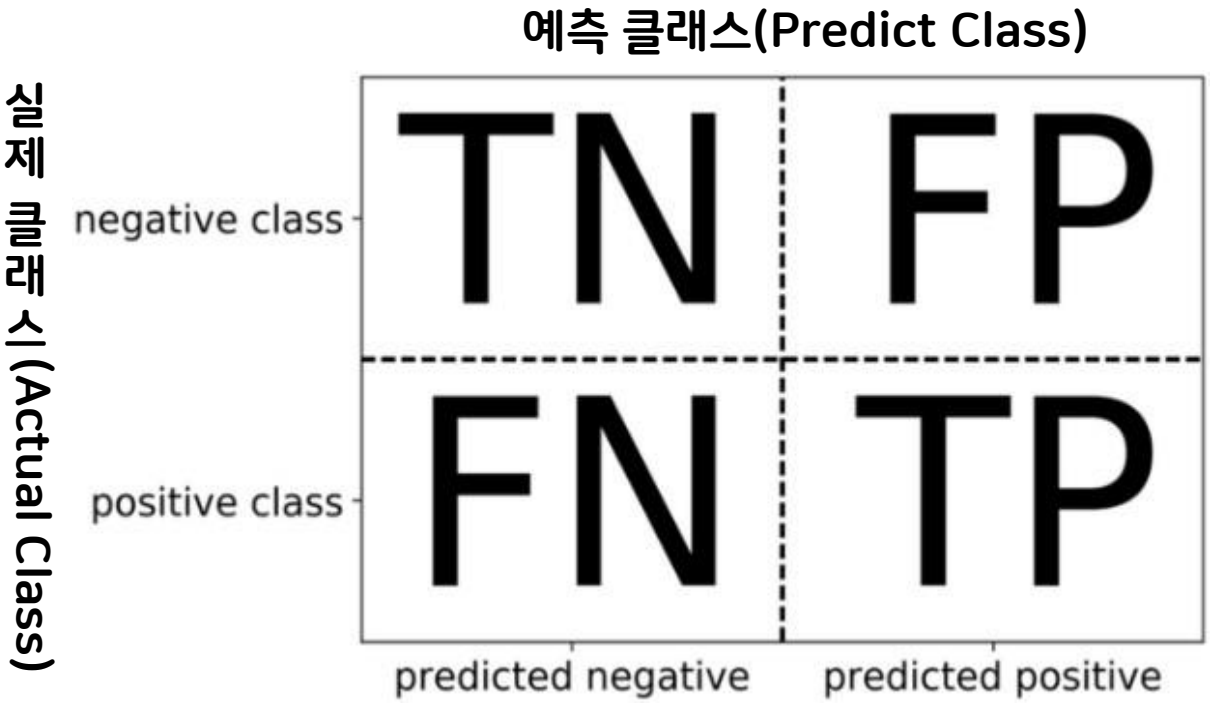
False Negative(FN)
- 실제 True인 정답을 False라고 예측(오답)

True Positive(TP)
- 실제 True인 정답을 True라고 예측(정답)

분류 평가 지표 – 이진분류 : 정확도(Accuracy)

스마트인재개발원
Smart Human Resources Development

Confusion_matrix



정확도(Accuracy)

정확히 예측한 수를

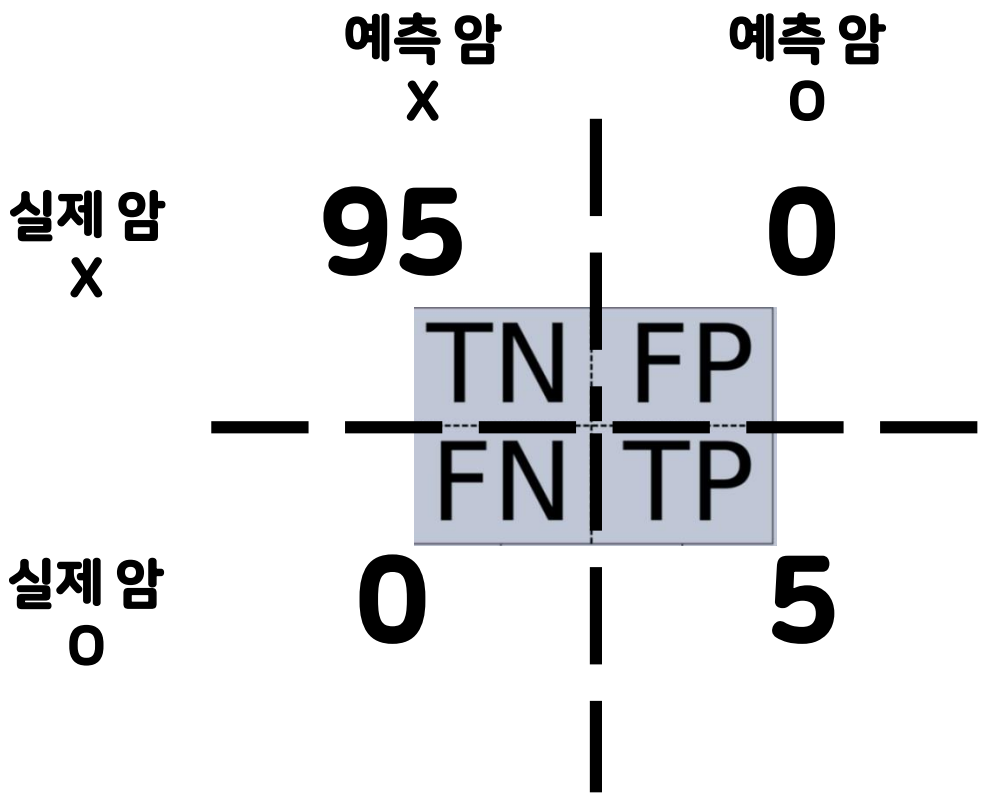
전체 샘플 수로 나눈 것

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

분류 평가 지표 – 이진분류 : 정확도(Accuracy)

스마트인재개발원
Smart Human Resources Development

Confusion_matrix



100명 중 암 환자는 5명

(95명 : 암 X, 5명 : 암O 예측)

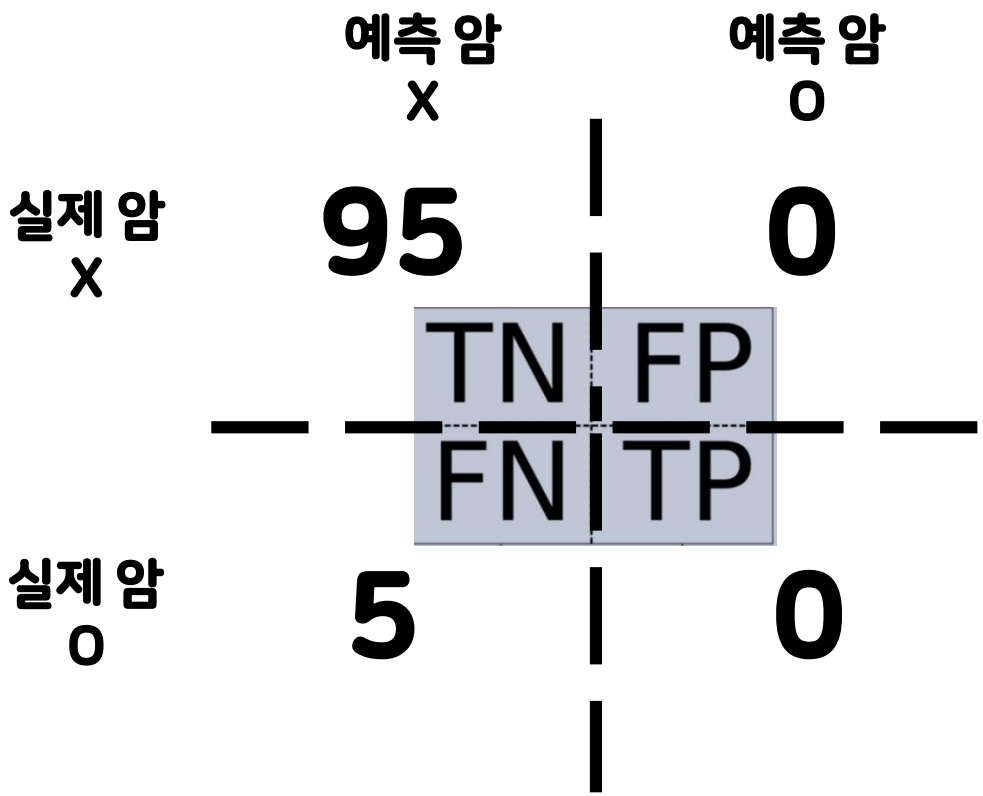
$$\frac{100}{100}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

분류 평가 지표 – 이진분류 : 정확도(Accuracy)

스마트인재개발원
Smart Human Resources Development

Confusion_matrix



100명 중 암 환자는 5명

(100명 : 암 X 예측)

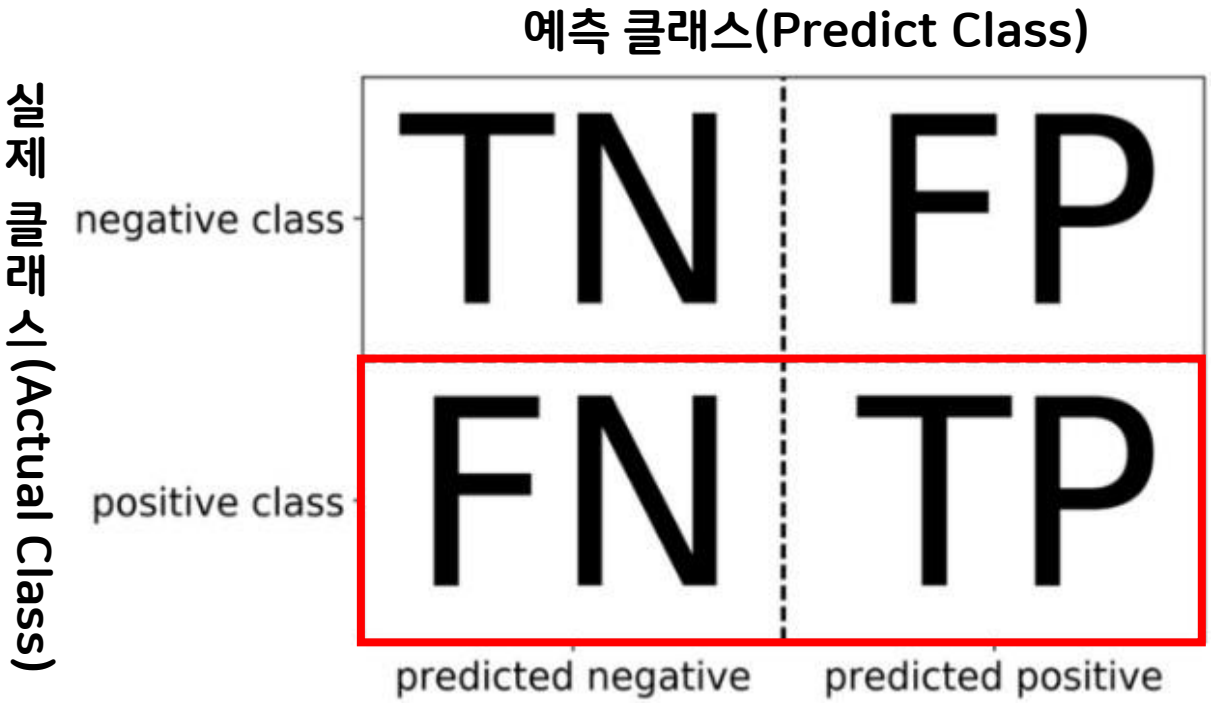
$$\frac{95}{100}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

분류 평가 지표 – 이진분류 : 재현율(Recall)

스마트인재개발원
Smart Human Resources Development

Confusion_matrix



재현율(Recall)

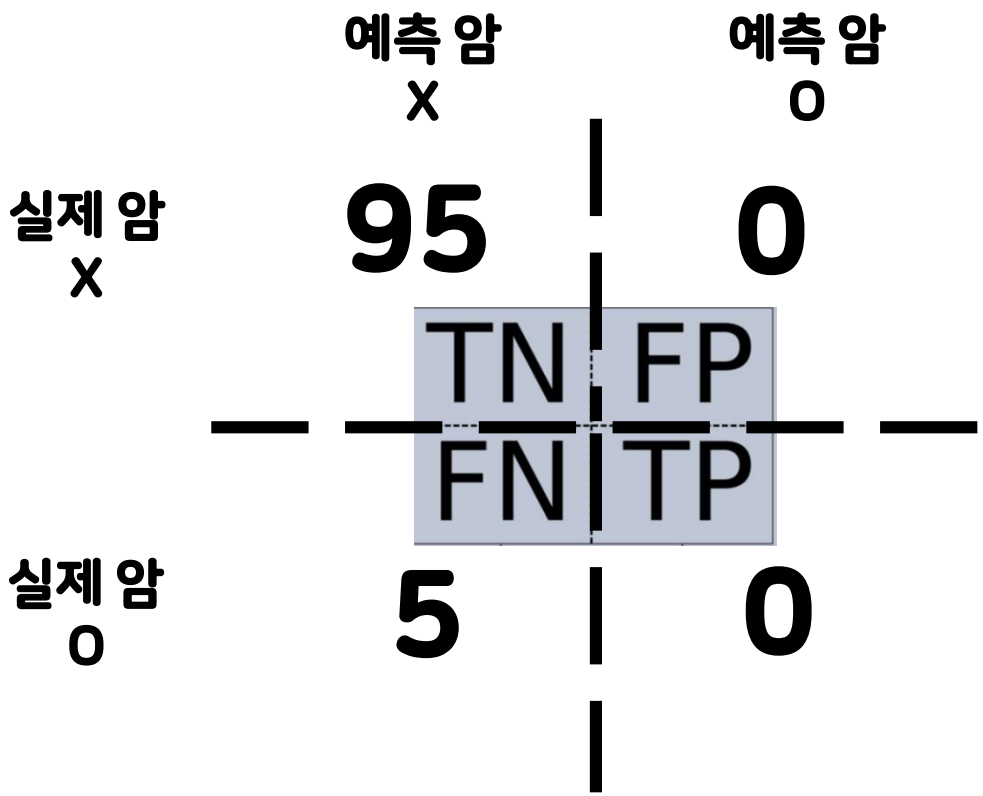
전체 양성 샘플 중에서 얼마나 많은 샘플이 양성 클래스로 분류되는가

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

분류 평가 지표 – 이진분류 : 재현율(Recall)

스마트인재개발원
Smart Human Resources Development

Confusion_matrix



100명 중 암 환자는 5명

(100명 : 암 X 예측)

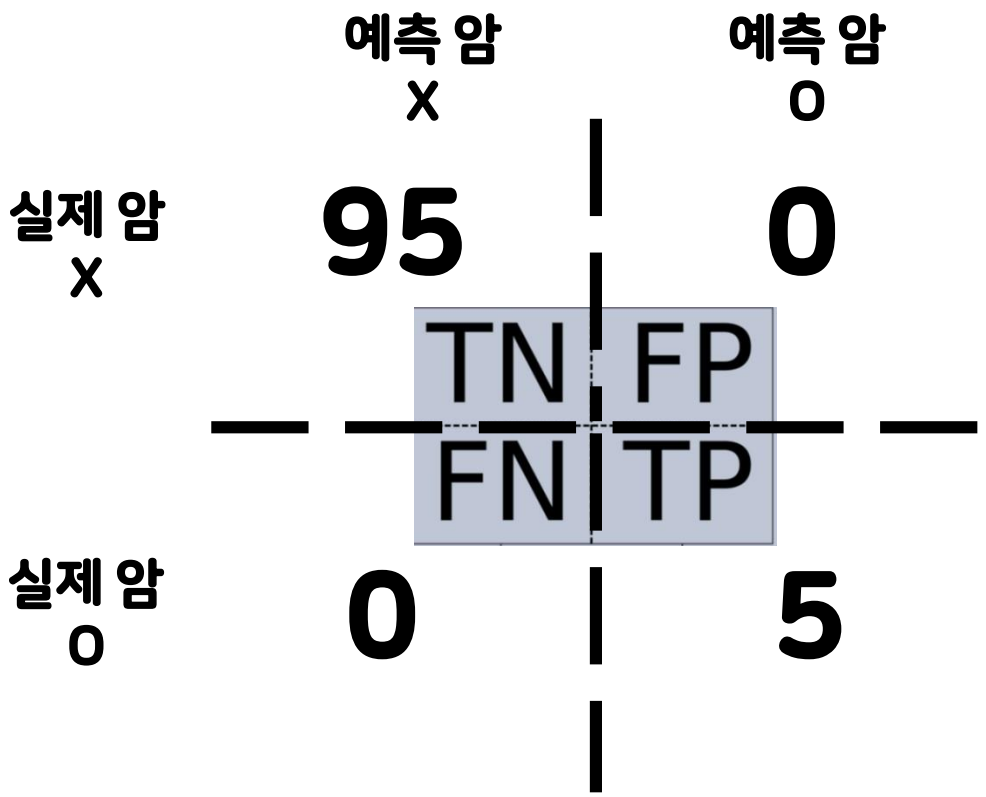
$$\frac{0}{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

분류 평가 지표 – 이진분류 : 재현율(Recall)

스마트인재개발원
Smart Human Resources Development

Confusion_matrix



100명 중 암 환자는 5명

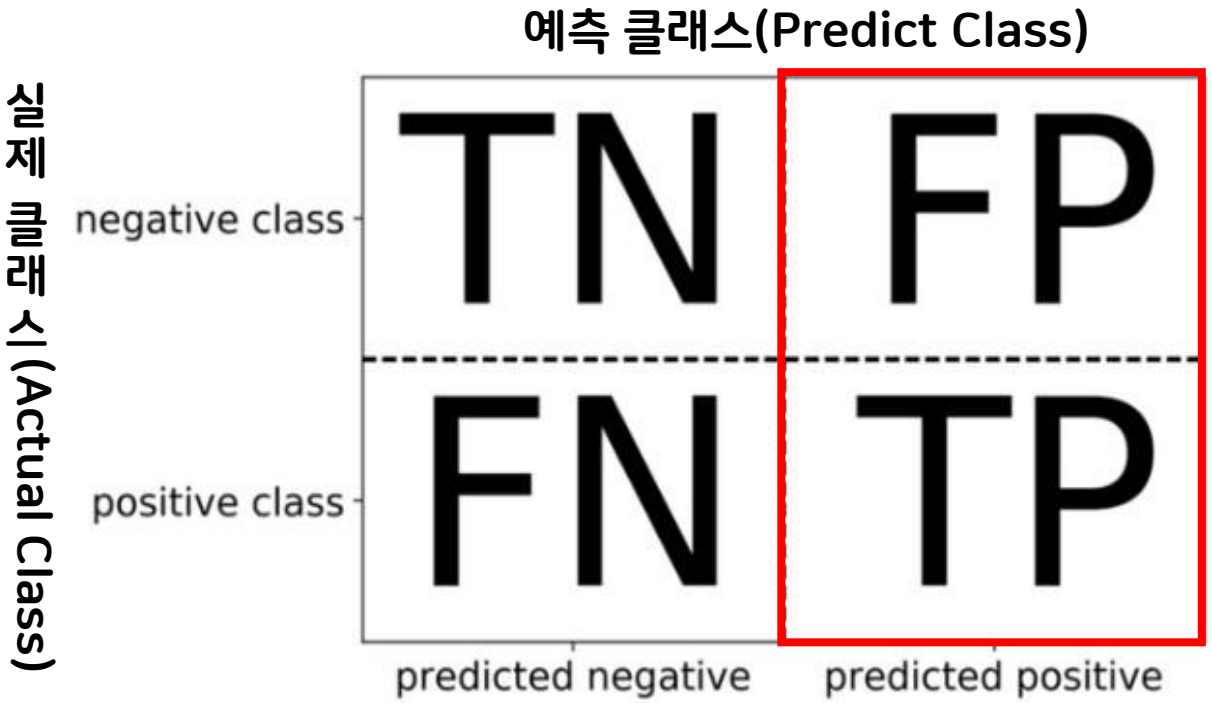
(95명 : 암 X, 5명 : 암 O 예측)

$$\frac{5}{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

분류 평가 지표 – 이진분류 : 정밀도(Precision)

Confusion_matrix



정밀도(Precision)

양성으로 예측된 것 중 얼마나 많은

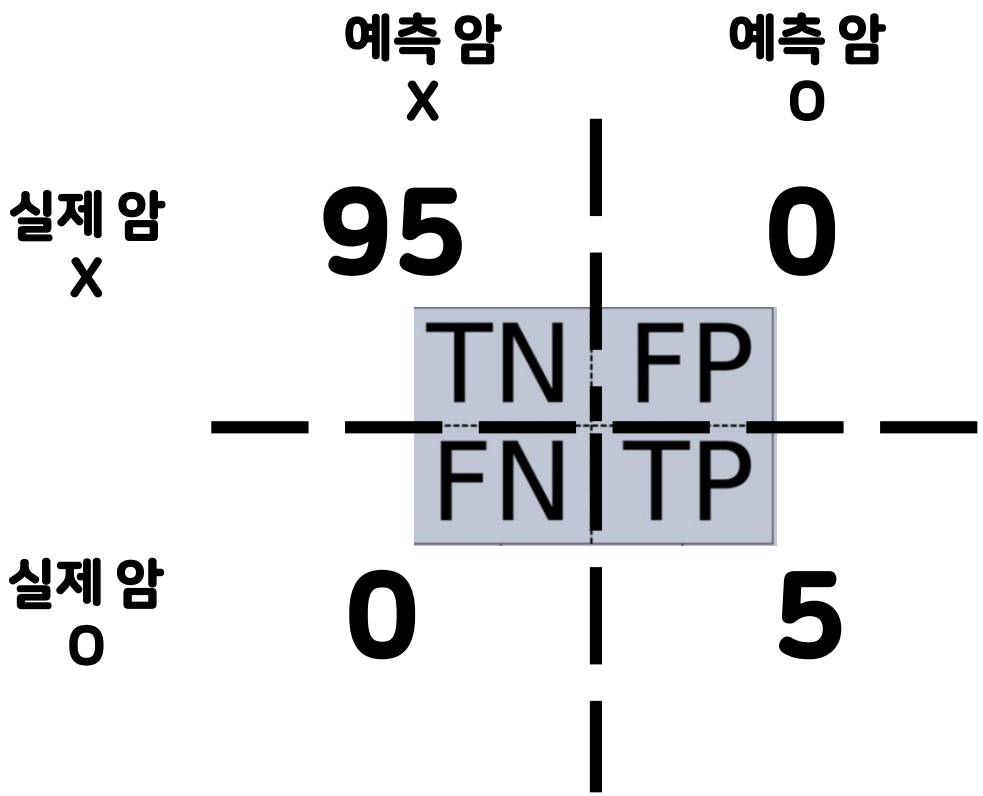
샘플이 진짜 양성인지 측정하는 것

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

분류 평가 지표 – 이진분류 : 정밀도(Precision)

스마트인재개발원
Smart Human Resources Development

Confusion_matrix



100명 중 암 환자는 5명

(95명 : 암 X, 5명 : 암 O 예측)

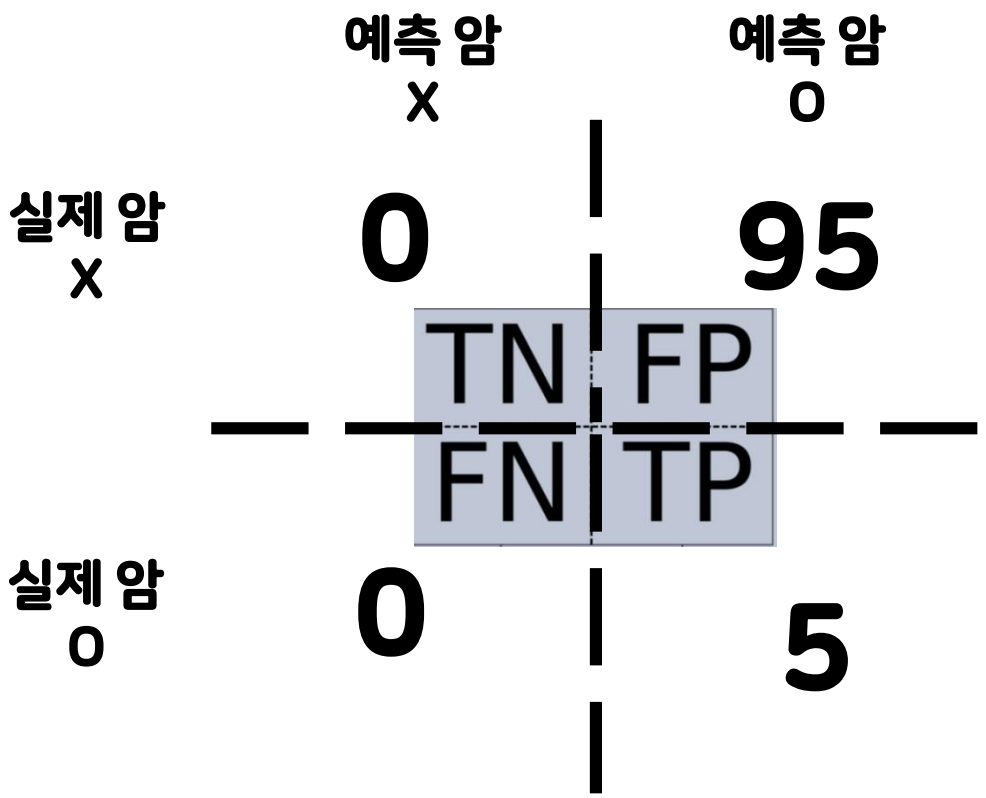
$$\frac{5}{5}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

분류 평가 지표 – 이진분류 : 정밀도(Precision)

스마트인재개발원
Smart Human Resources Development

Confusion_matrix



100명 중 암 환자는 5명

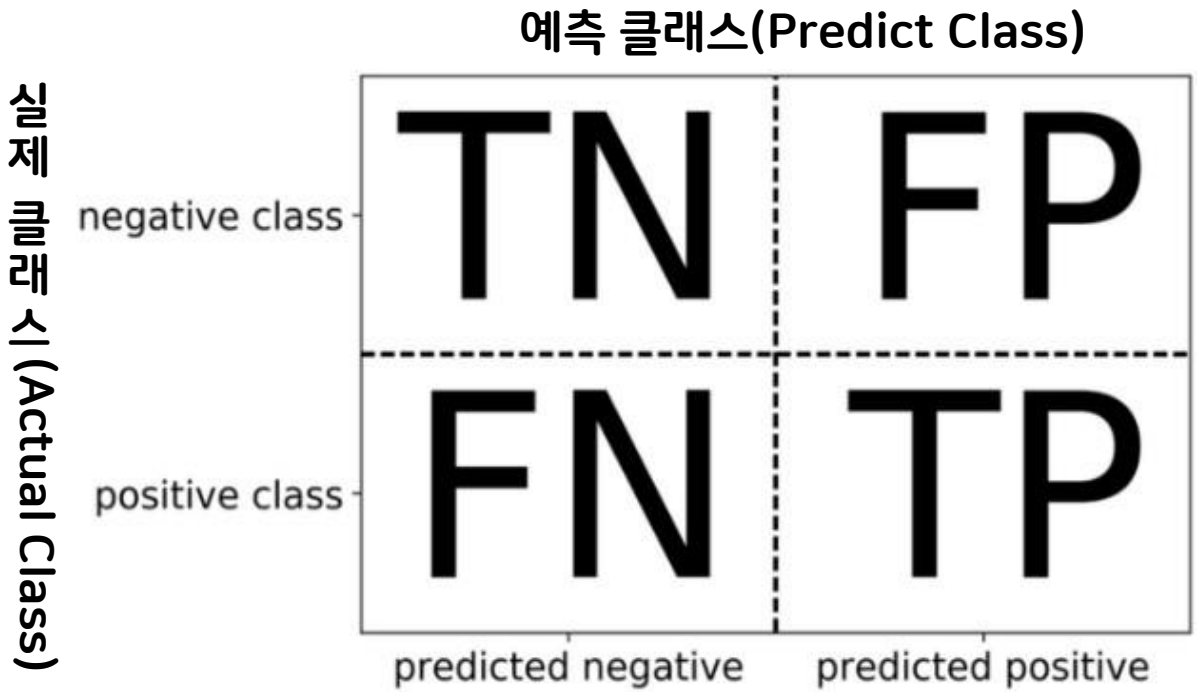
(100명 : 암 O 예측)

$$\frac{5}{100}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

분류 평가 지표 - 다중분류

Confusion_matrix



F1-score

정밀도와 재현율의 조화 평균

$$F = 2\frac{precision \cdot recall}{precision + recall}$$

분류 평가 지표

스마트인재개발원
Smart Human Resources Development

- **낮은 재현율보다 높은 정밀도를 선호하는 경우**

  - 어린아이에게 안전한 동영상(양성)을 걸러내는 분류기를 훈련시킬 경우 좋은 동영상이 많이 제외되더라도(낮은 재현율) 안전한 것들만 노출시키는(높은 정밀도) 분류기가 더 좋음

- **낮은 정밀도보다 높은 재현율을 선호하는 경우**

  - 감시카메라로 좀도둑(양성)을 잡아내는 분류기를 훈련시킬 경우 경비원이 잘못된 호출을 종종 받지만(낮은 정밀도) 거의 모든 좀도둑을 잡는(높은 재현율) 분류기가 더 좋음

분류 평가 지표

주요 매개변수(Hyperparameter)

scikit-learn의 경우

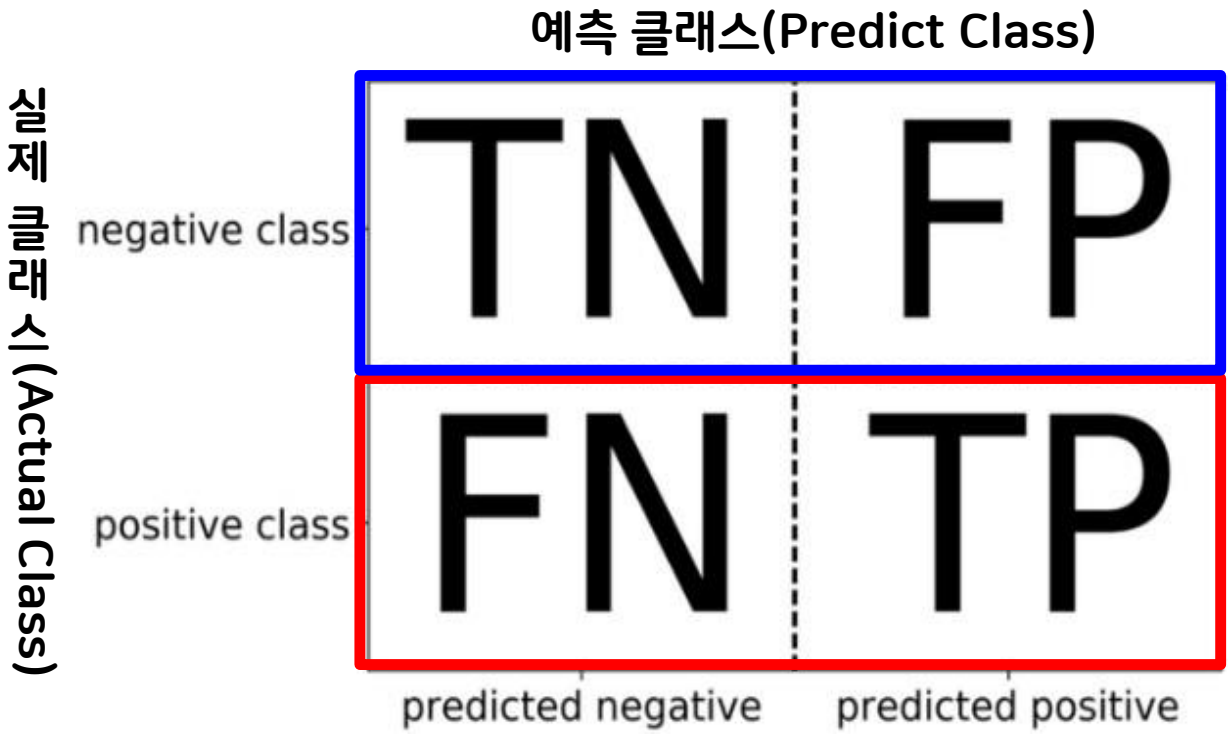
```
classification_report(실제값, 예측값)
```

분류 평가 지표 시각화 – ROC(Receiver Operating Characteristic) curve

- 여러 임계값(0 ~ 1)에서 분류기의 특성을 분석하는데 널리 사용되는 도구
- 클래스의 분포가 다르고 겹치는 부분이 존재한 경우에 정확도(Accuracy)의 단점을 보완하기 위한 것
- 진짜 양성 비율(TPR) 에 대한 거짓 양성 비율(FPR)을 나타냄

분류 평가 지표 시각화 – ROC(Receiver Operating Characteristic) curve

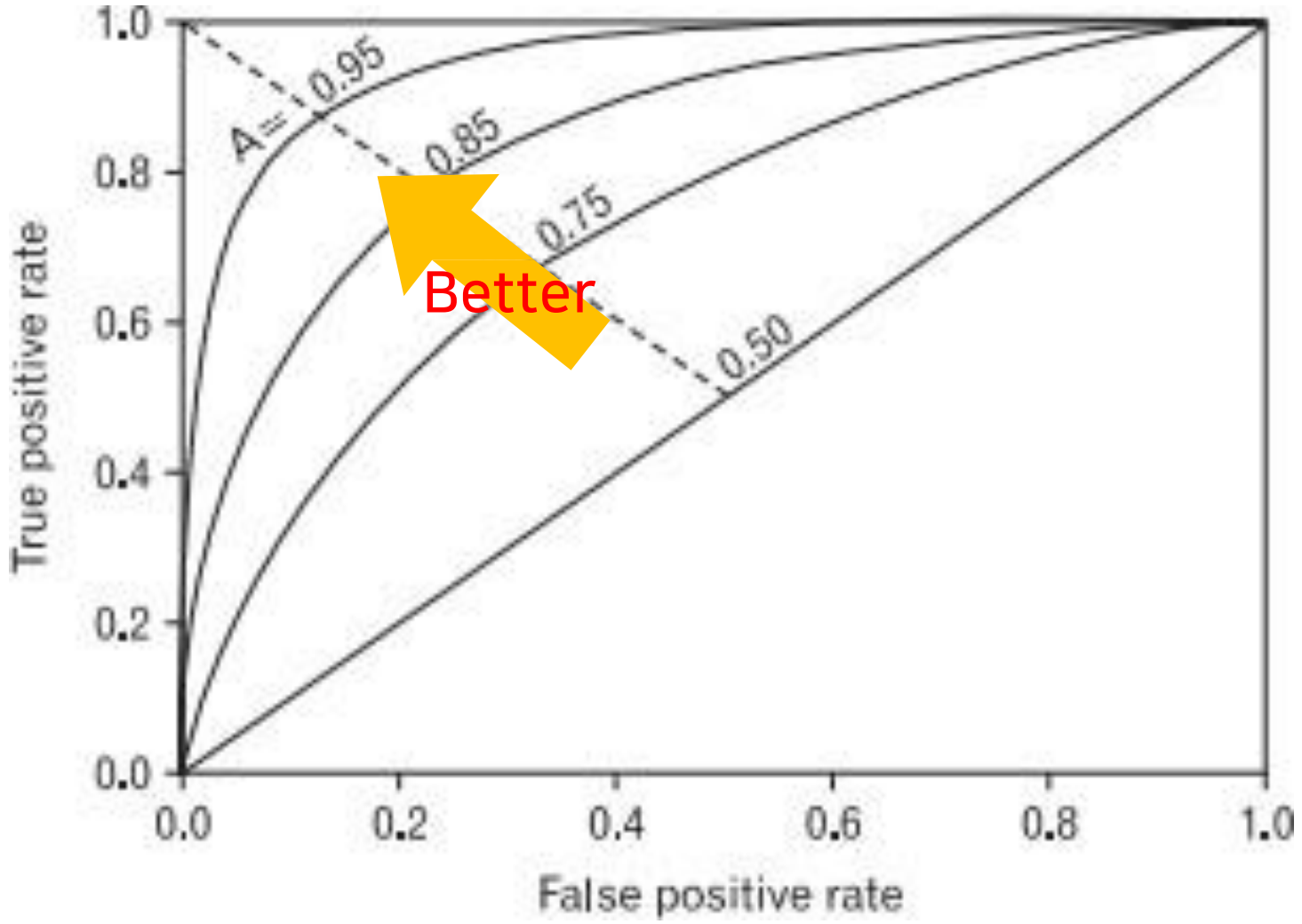
- 가짜 양성비율(FPR) : 전체 음성 샘플 중에서 거짓 양성으로 잘못 분류한 비율
- 진짜 양성비율(TPR) : 재현율



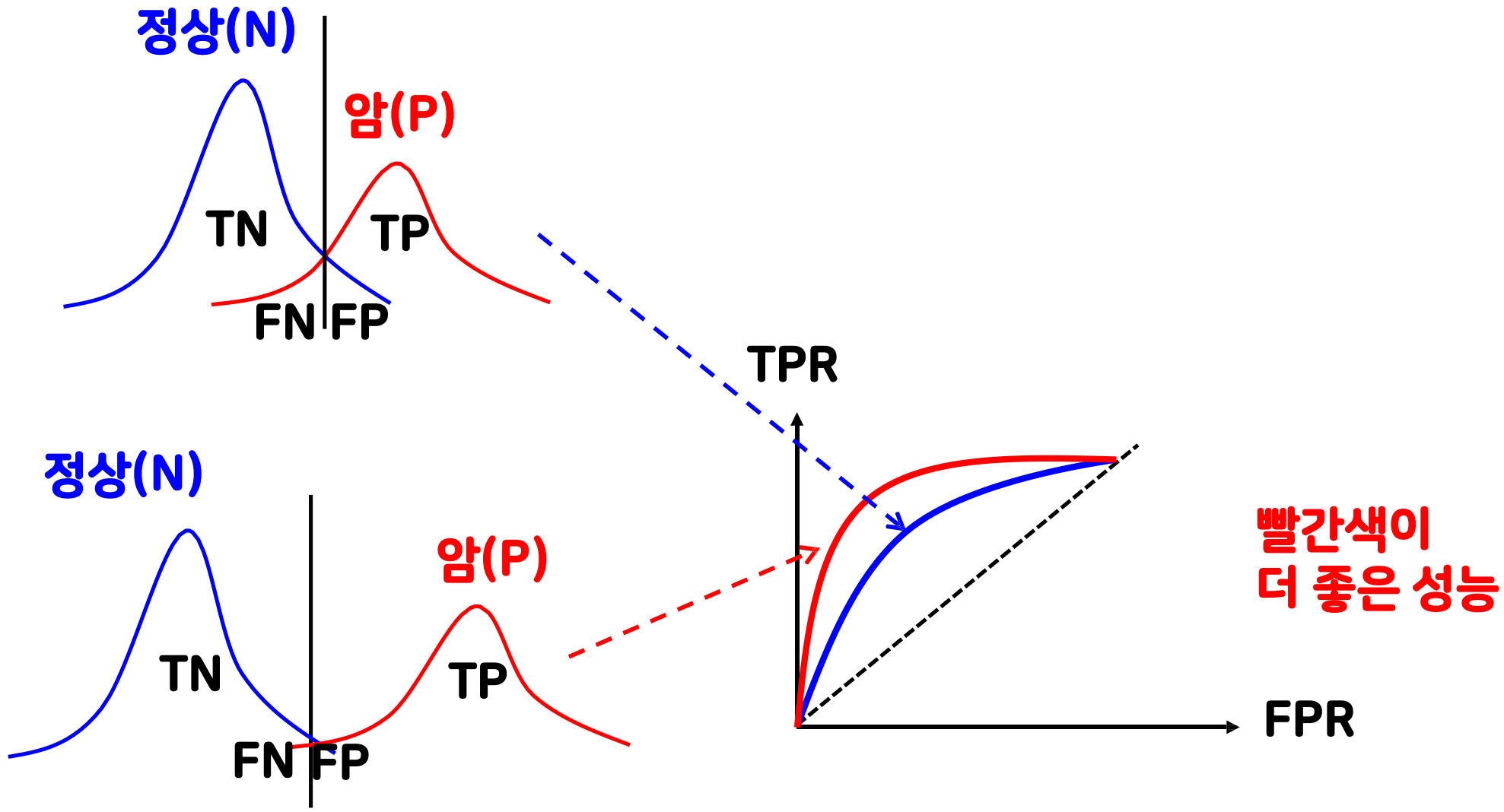
$$FPR = \frac{FP}{FP+TN}$$

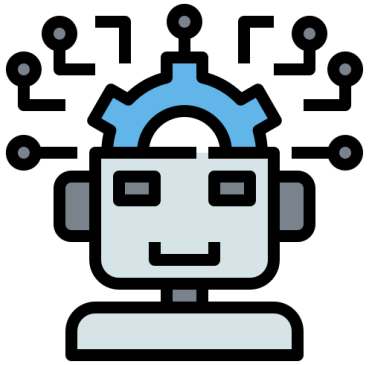
$$TPR = \frac{TP}{TP+FN} = recall$$

분류 평가 지표 시각화 – ROC(Receiver Operating Characteristic) curve



분류 평가 지표 시각화 – ROC(Receiver Operating Characteristic) curve





GridSearch

GridSearch?

- 매개변수를 선택하는 것은 머신러닝에서 중요한 일
- 관심 있는 매개변수들을 대상으로 가능한 모든 조합을 시도하는 것

주요 매개변수(Hyperparameter)

scikit-learn의 경우

GridSearchCV(모델, 모델의 파라미터목록, cv)

- cv : 교차검증시 나눌 데이터분할 수