Machine Learning

Linear Model

START

스마트인재개발원
Smart Human Resources Development

학습목표

- 선형 분류모델을 이해하고 사용 할 수 있다.
- 다양한 분류평가 지표를 이해 할 수 있다.

Linear Model
(Classification)

학습목표

- 선형 분류모델을 이해하고 사용 할 수 있다.
- Logistic Regression 모델의 필요성을 이해할 수 있다.
- Logistic Regression 모델의 비용함수를 이해할 수 있다.

Linear Model - Classification

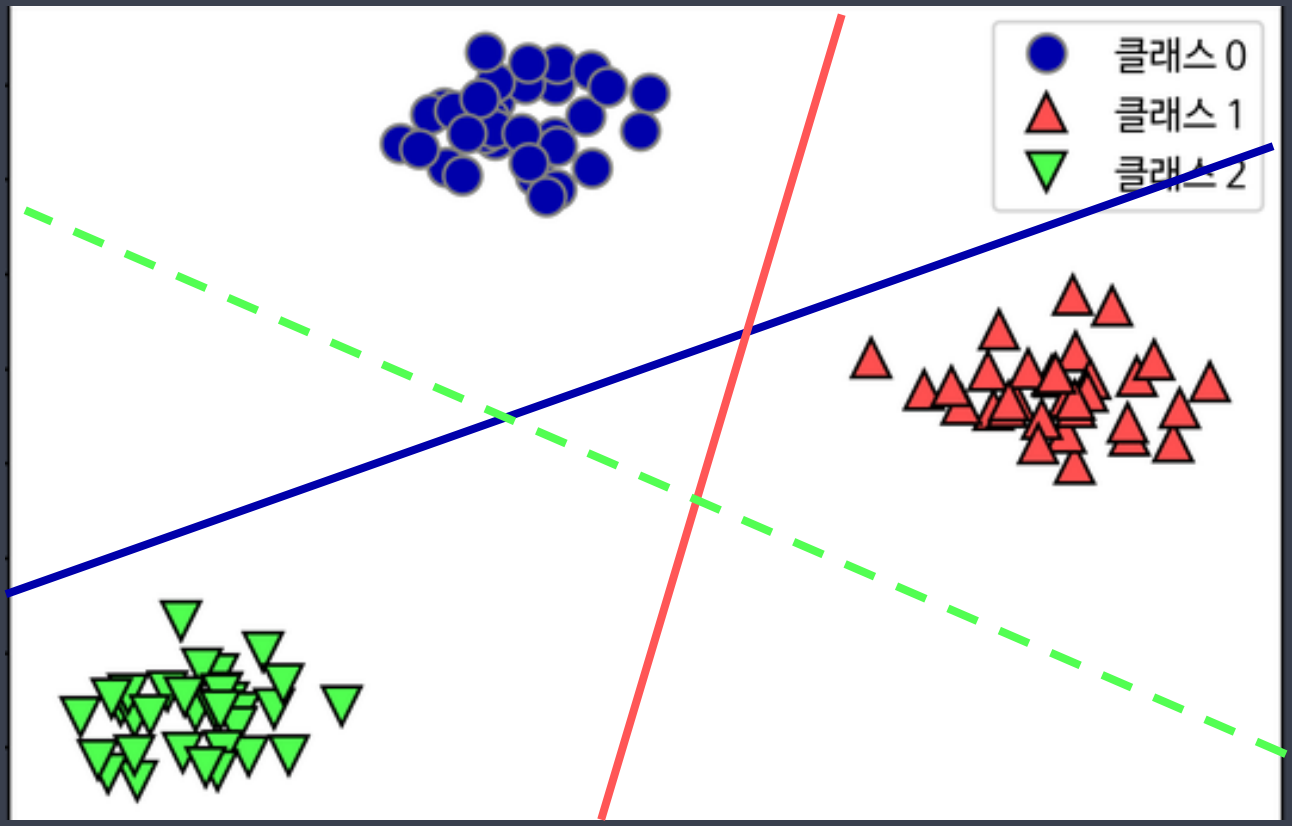
분류용 선형 모델

$$y = w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_px_p + b > 0$$

- 특성들의 가중치 합이 0보다 크면 class를 +1(양성클래스) 0보다 작다면 클래스를 -1(음성클래스)로 분류한다.
- 분류용 선형모델은 결정 경계가 입력의 선형함수
- 일대다 방법을 통해 다중 클래스 분류

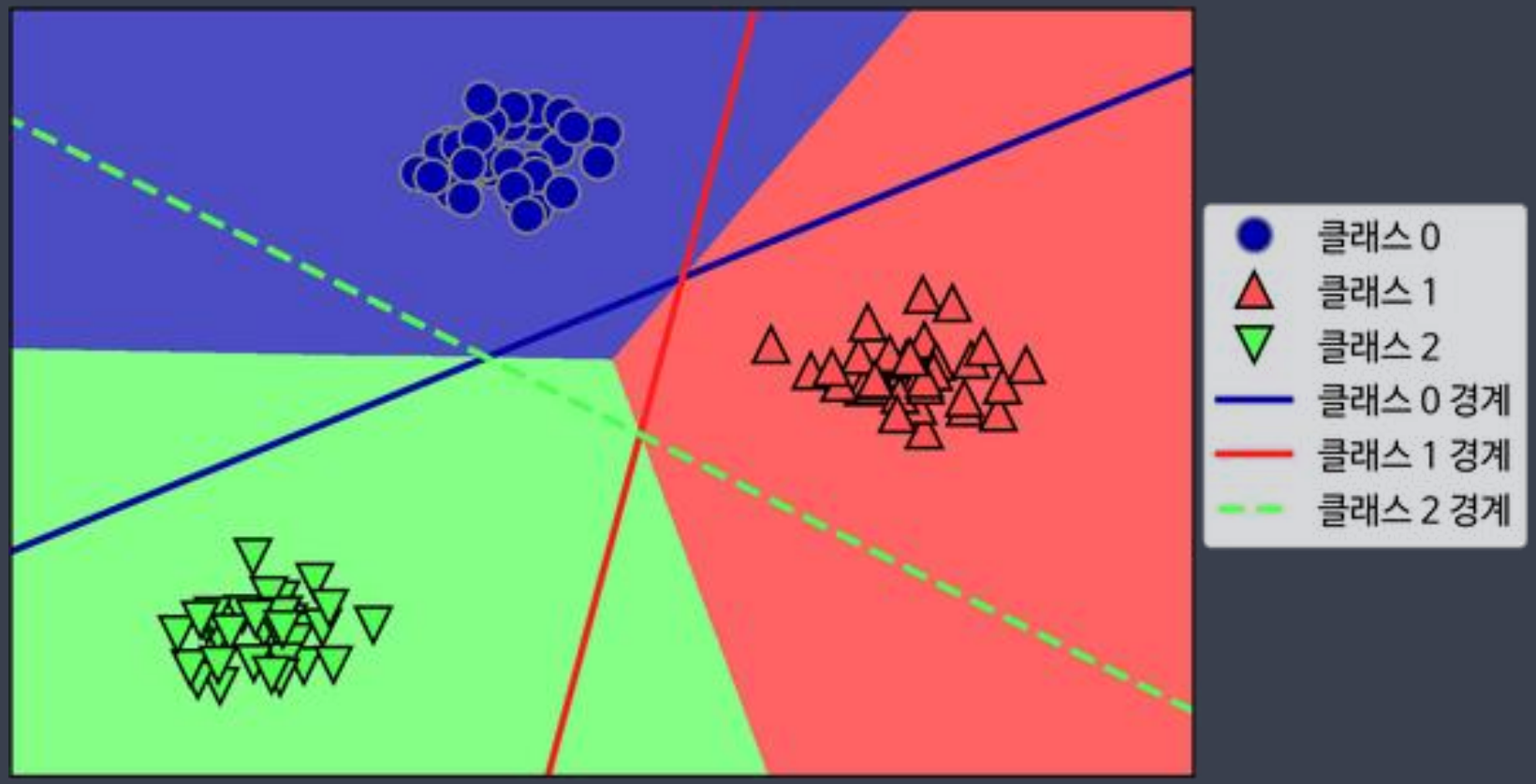
Linear Model - Classification

분류용 선형 모델



Linear Model - Classification

분류용 선형 모델

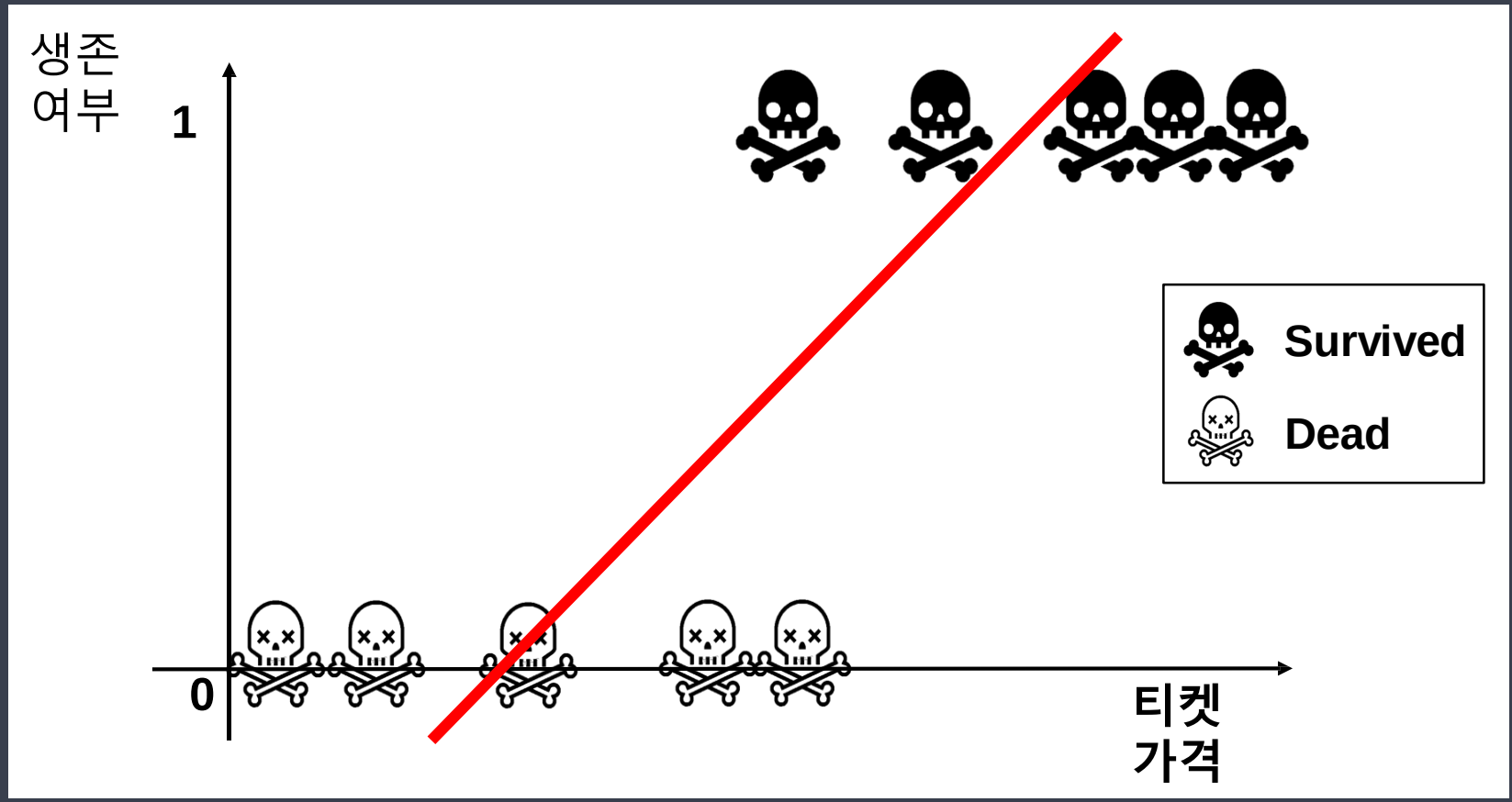


Linear Model - Classification

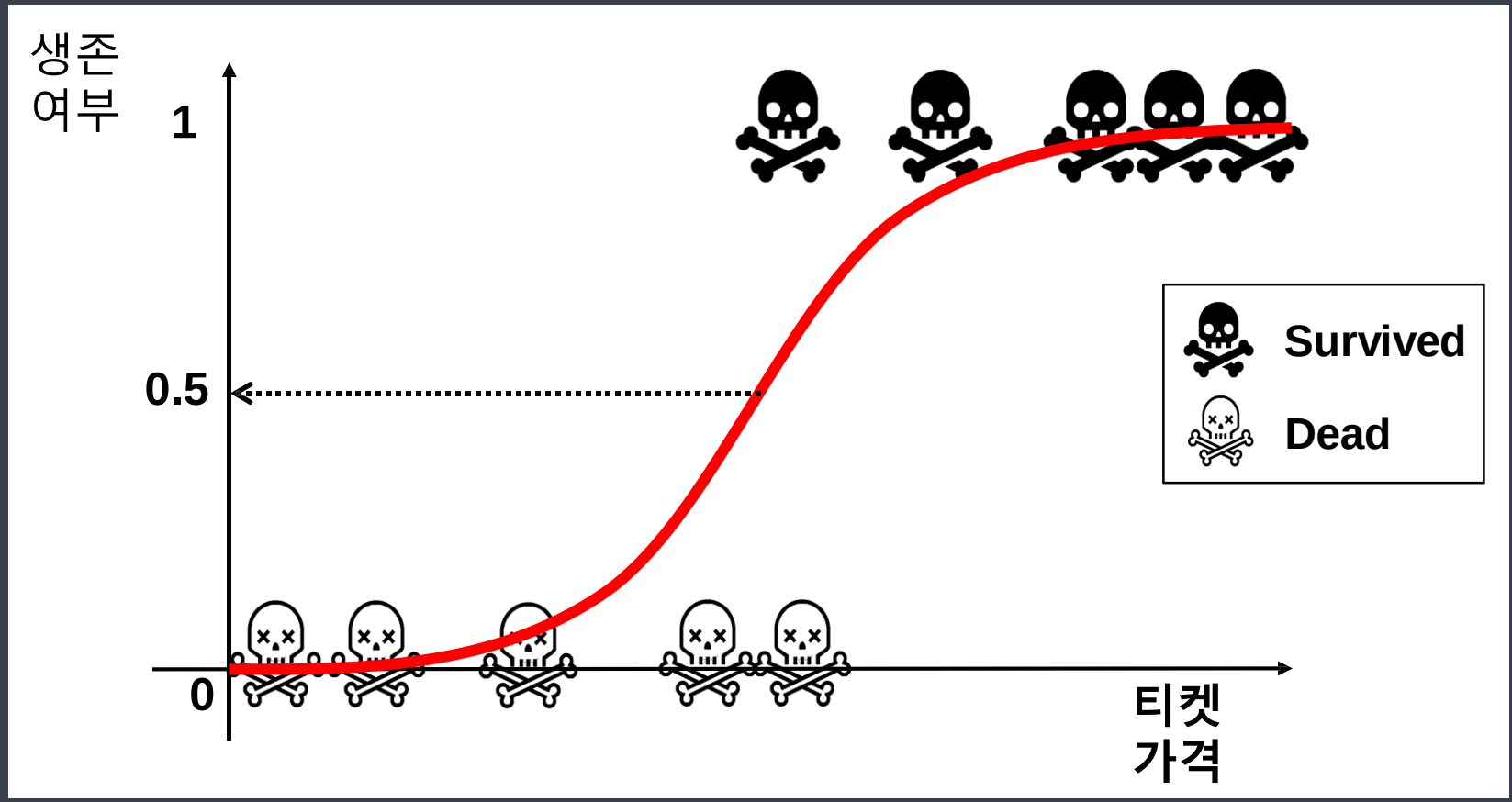
Logistic Regression

- 회귀공식을 사용해서 Regression이라는 이름이 붙음
- 결정 경계가 선형이기 때문에 선형 모델
- 시그모이드 함수의 최적선을 찾고 반환값을 확률로 간주

Linear Model - Classification



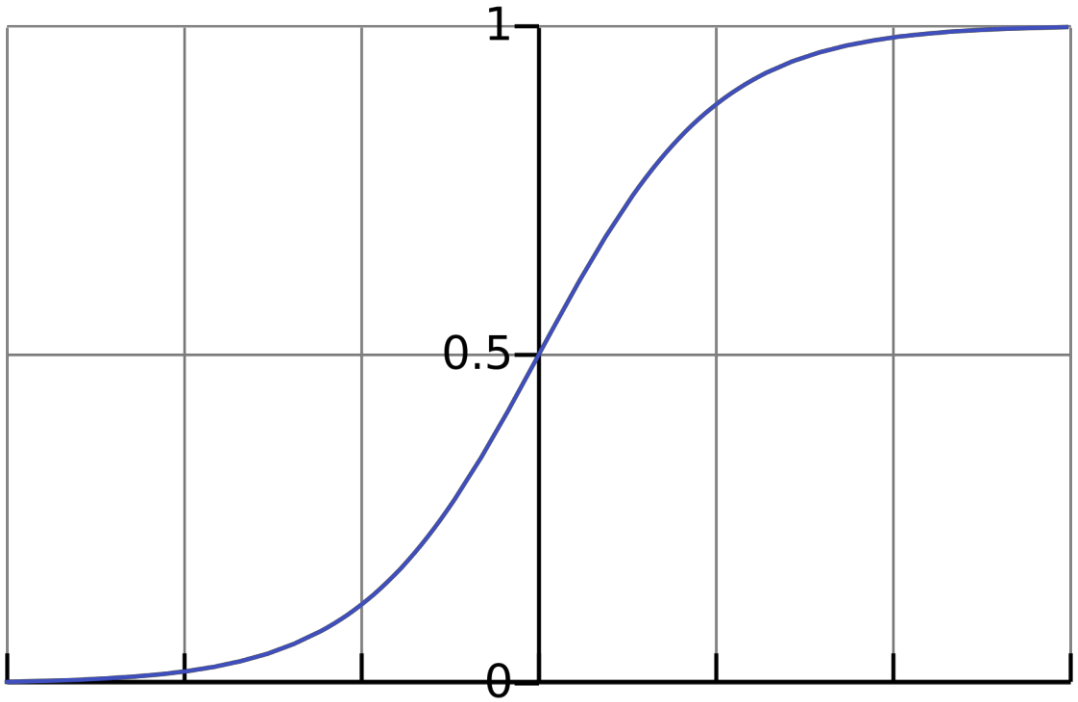
Linear Model - Classification



Linear Model - Classification

Logistic Regression

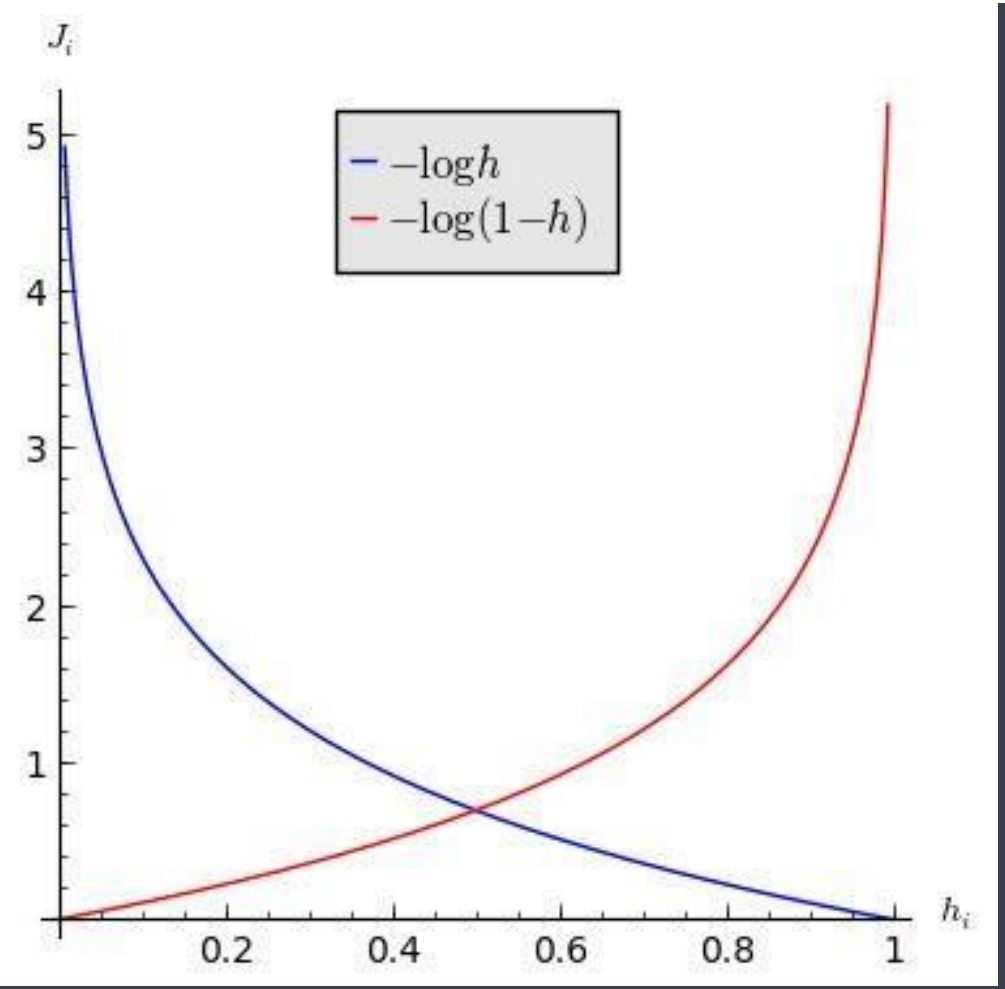
- 선형 함수의 결과 값을 Sigmoid Function(Logistic Function)을 이용해 0과 1로변환한다.



$$sigmoid(x) = \frac{1}{1 + e^{-x}}$$

Linear Model - Classification

$$\text{Cost}(h_\theta(x), y) = -\log(h_\theta(x)) \quad \text{if y} = 1$$
$$\text{Cost}(h_\theta(x), y) = -\log(1 - h_\theta(x)) \quad \text{if y} = 0$$



Linear Model

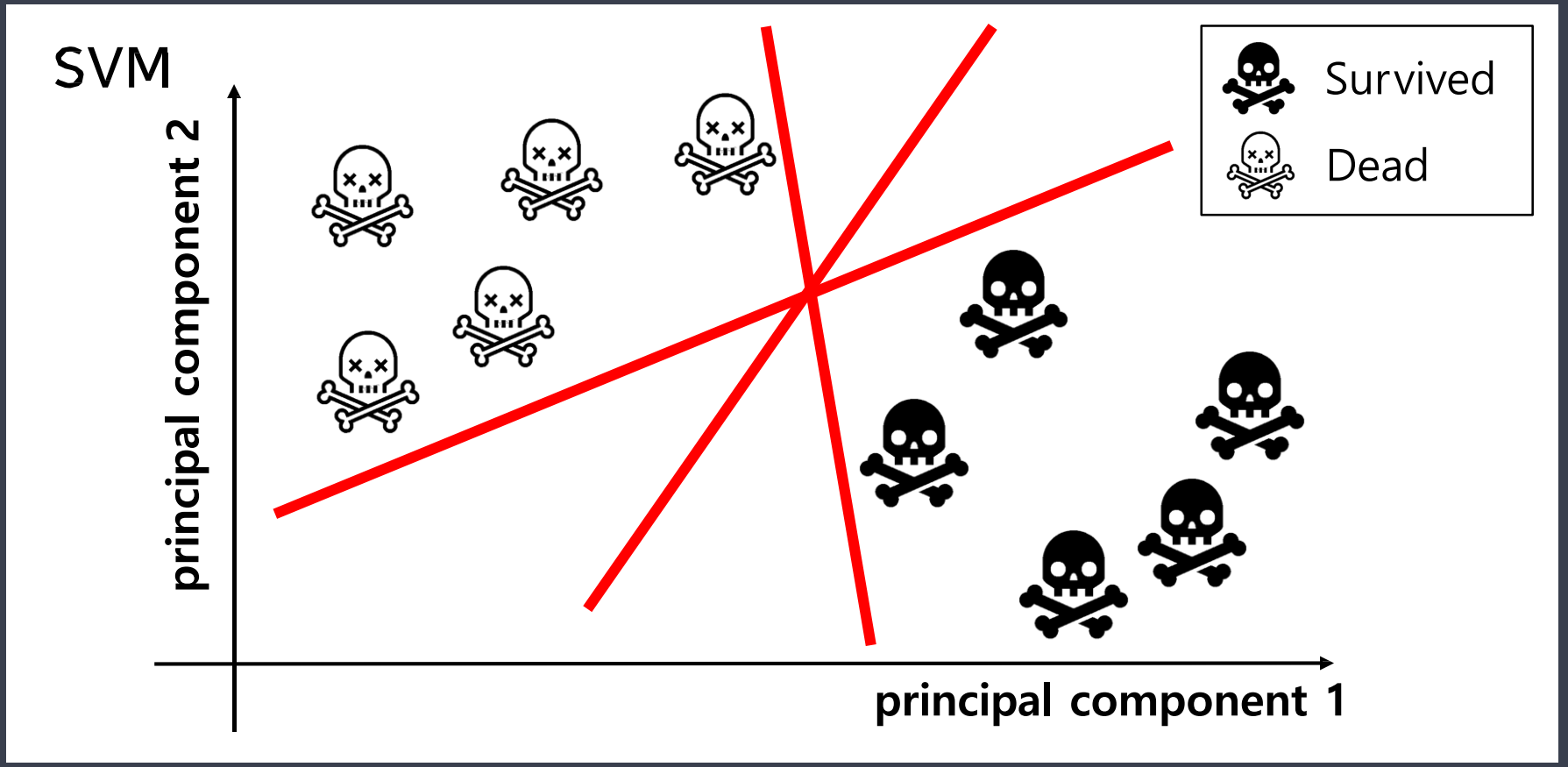
주요 매개변수(Hyperparameter)

- 선형 분류 모델 : C
  (값이 클수록 규제가 약해진다.)

- 기본적으로 L2규제를 사용, 하지만 중요한 특성이 몇 개 없다면 L1규제를 사용해도 무방
  (주요 특성을 알고 싶을 때 L1 규제를 사용하기도 한다.)

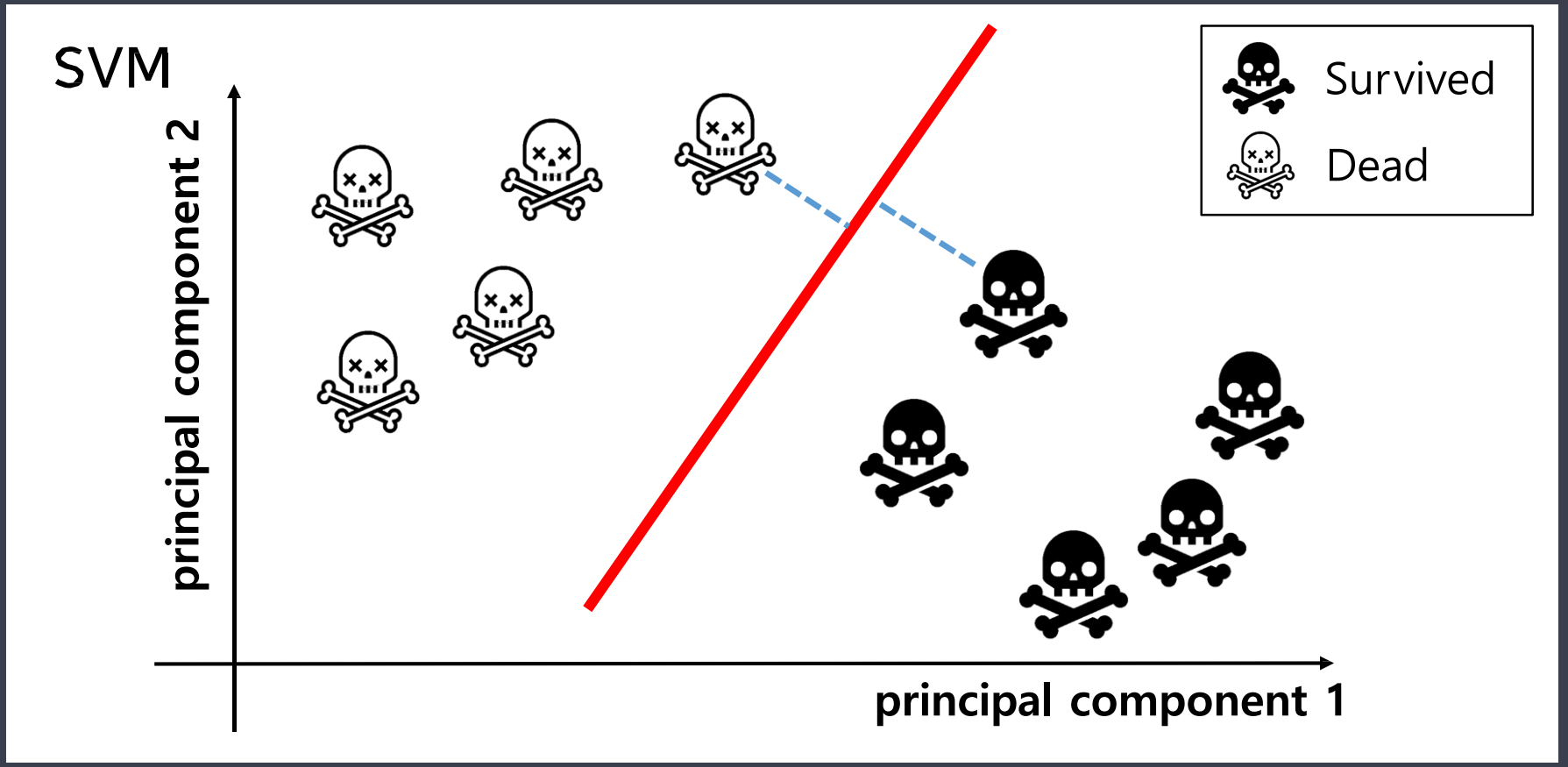
Linear Model - Classification

Support Vector Machine (SVM)



Linear Model - Classification

Support Vector Machine (SVM)



Linear Model

장단점 및 주요 매개변수(Hyperparameter)

- 회귀 선형 모델 : alpha
  (값이 클수록 규제가 강해진다.)
- 선형 분류 모델 : C
  (값이 클수록 규제가 약해진다.)
- 기본적으로 L2규제를 사용, 하지만 중요한 특성이 몇 개 없다면 L1규제를 사용해도 무방
  (주요 특성을 알고 싶을 때 L1 규제를 사용하기도 한다.)

Linear Model

장단점 및 주요 매개변수(Hyperparameter)

- 선형 모델은 학습 속도가빠르고 예측도 빠르다.
- 매우 큰 데이터 세트와 희소 (sparse)한 데이터 세트에서도 잘 동작한다.
- 특성이 많을 수록 더욱 잘 동작한다.
- 저차원(특성이 적은)데이터에서는 다른 모델이 더 좋은 경우가 많다.

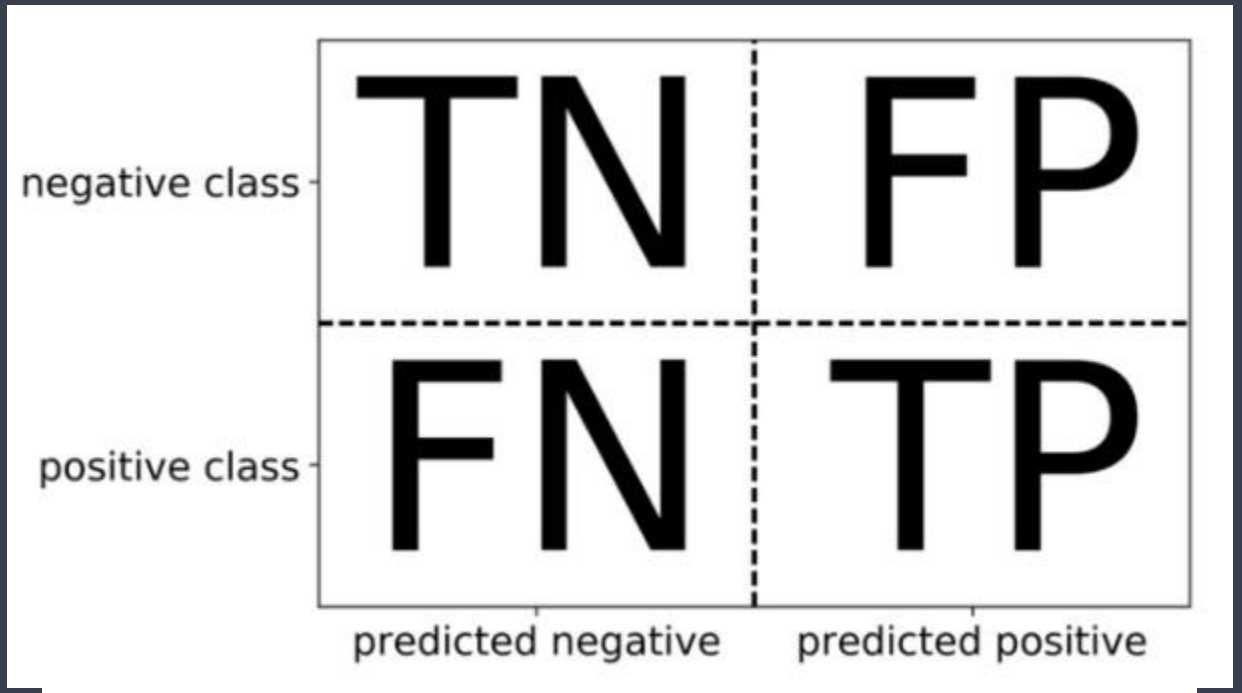
Linear Model

손글씨 데이터 분류 실습

분류 평가 지표

분류 평가 지표

Confusion_matrix



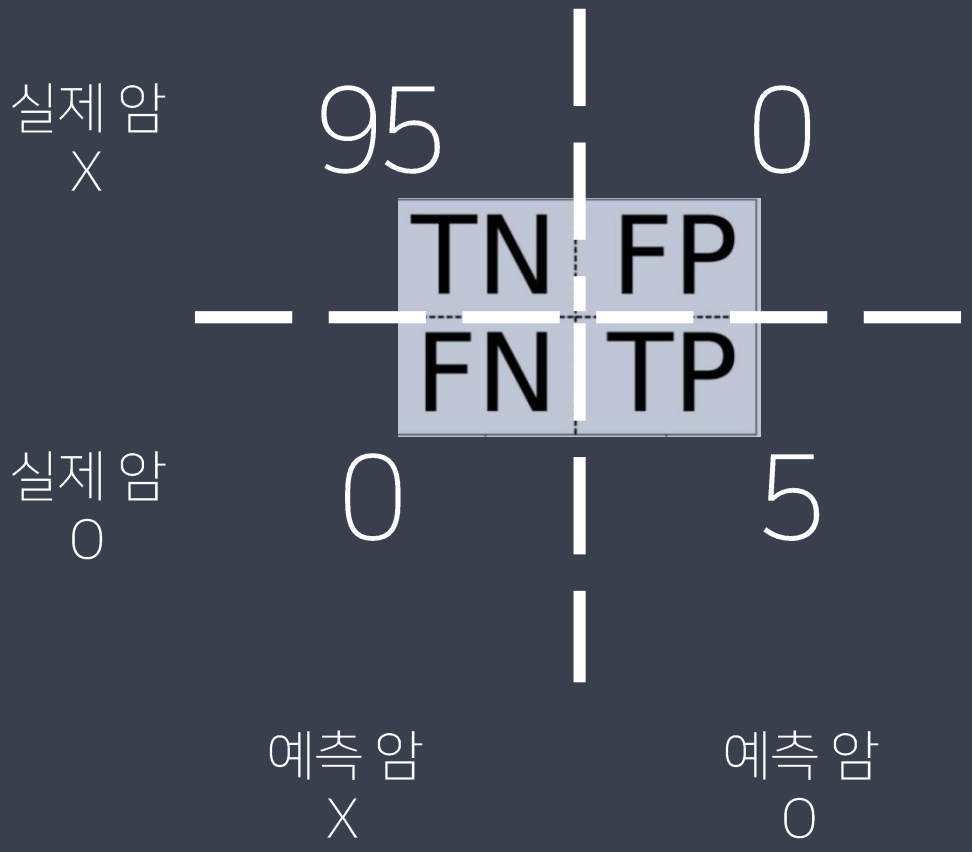
$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

정확도 (Accuracy)

전체 중에 정확히 맞춘 비율

분류 평가 지표

Confusion_matrix



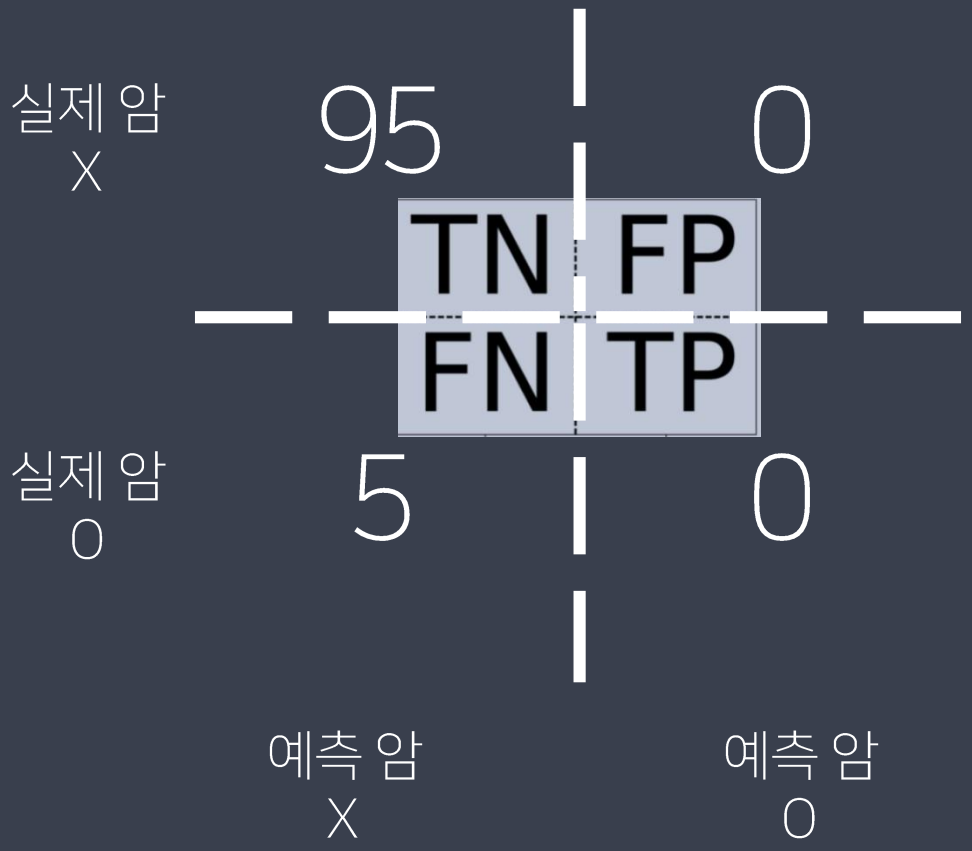
100명 중 암 환자는 5명

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

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

분류 평가 지표

Confusion_matrix



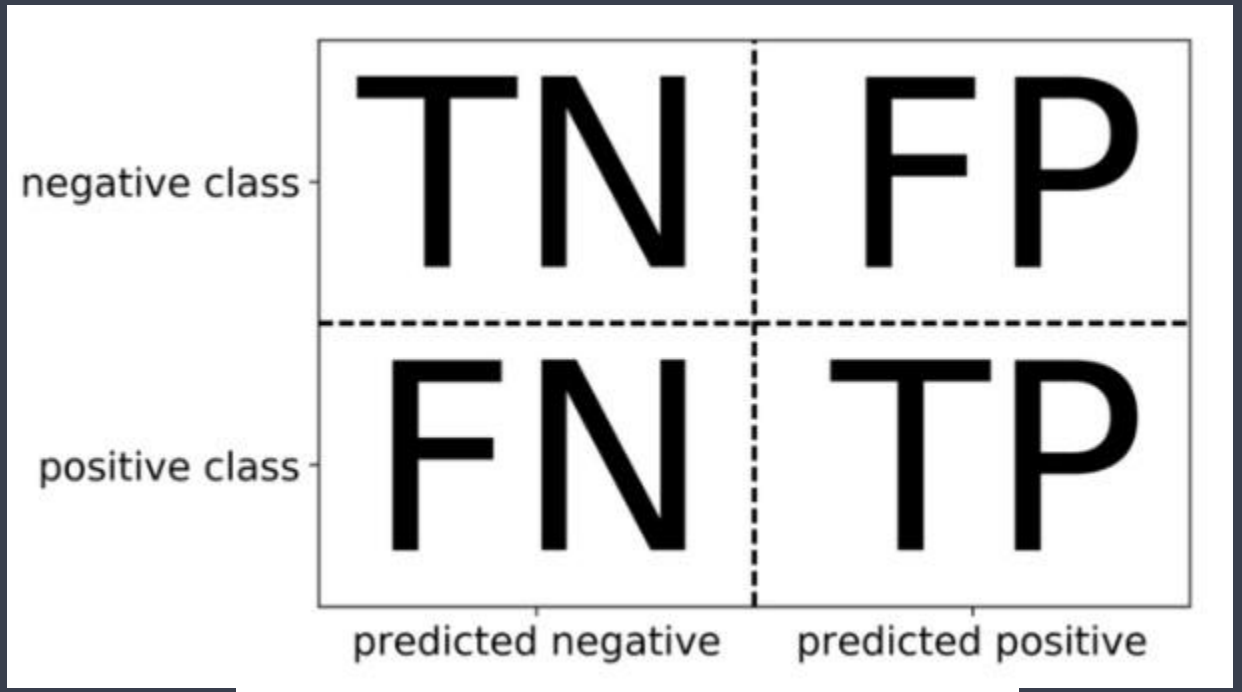
100명 중 암 환자는 5명

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

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

분류 평가 지표

Confusion_matrix



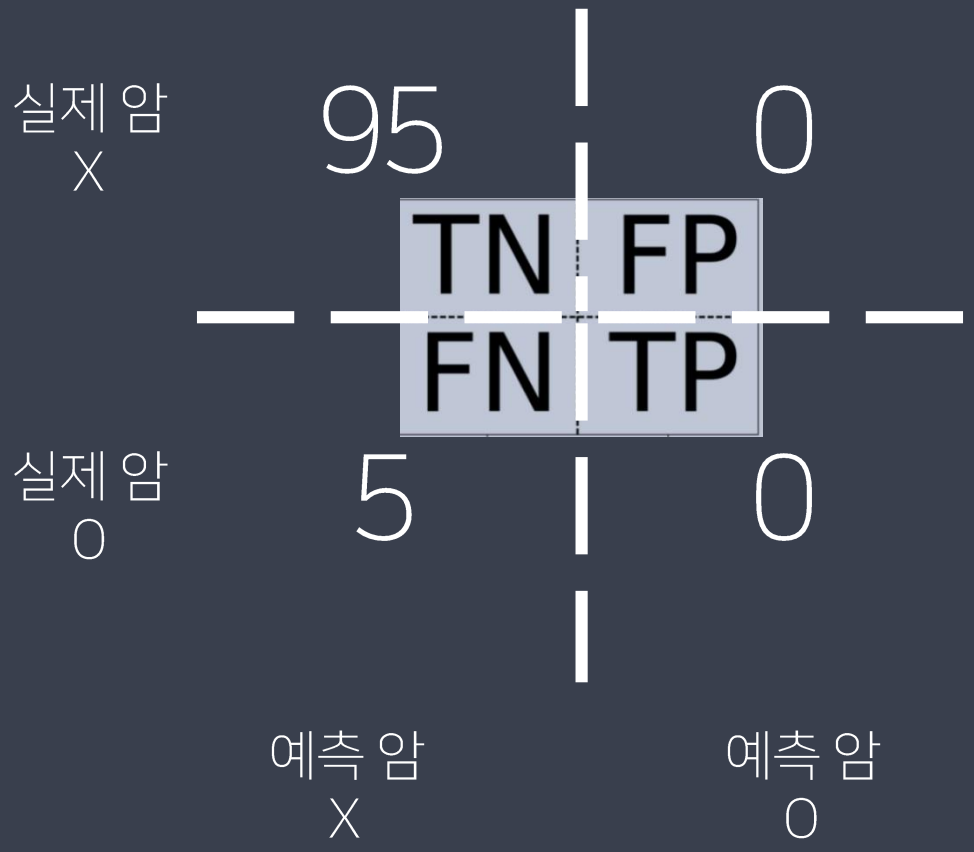
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

재현율 (Recall)

실제 양성 중에 예측 양성 비율

분류 평가 지표

Confusion_matrix



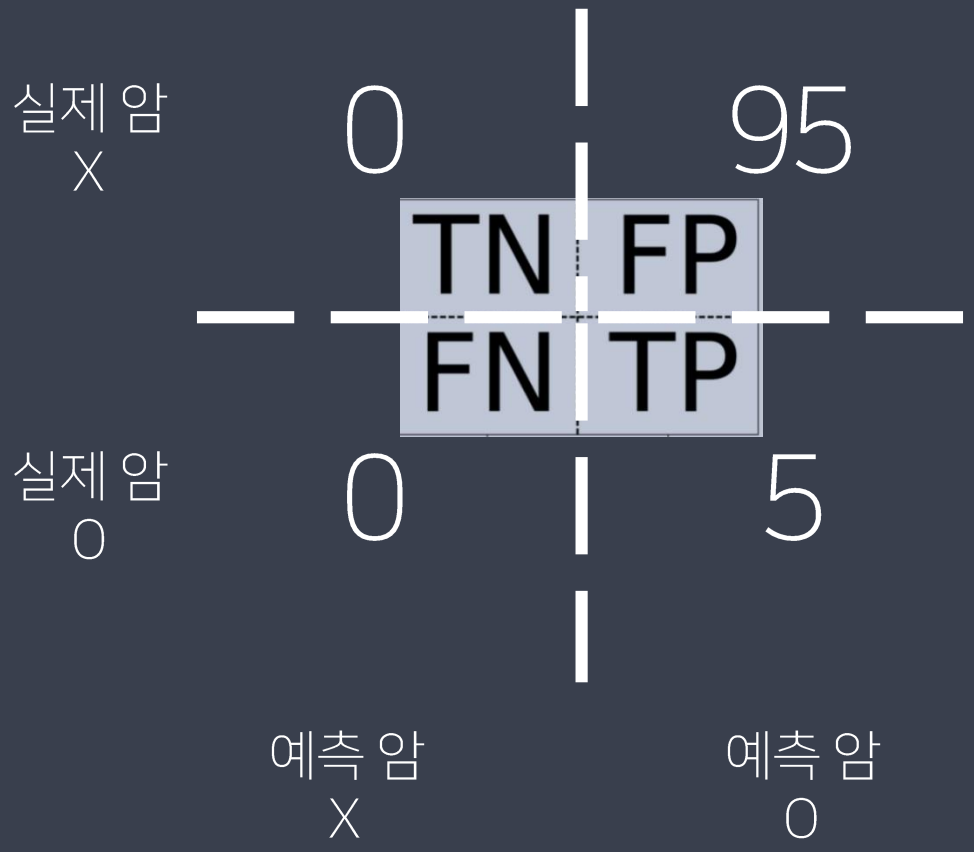
100명 중 암 환자는 5명

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

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

분류 평가 지표

Confusion_matrix



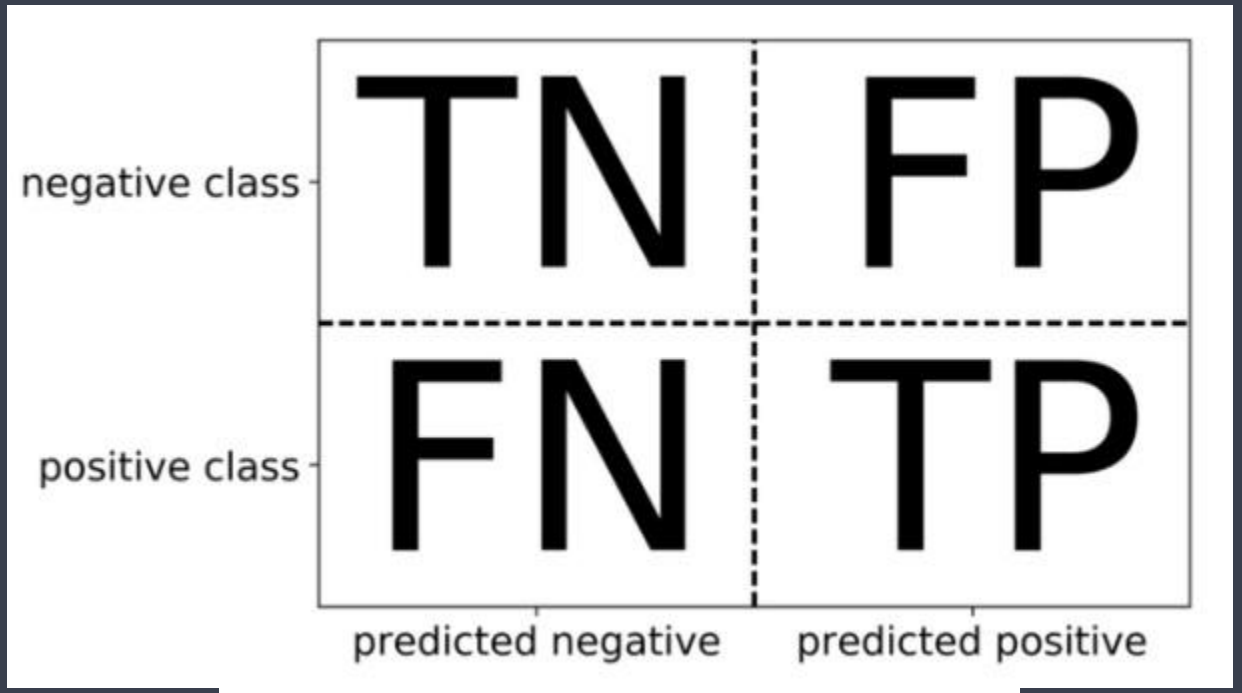
100명 중 암 환자는 5명

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

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

분류 평가 지표

Confusion_matrix



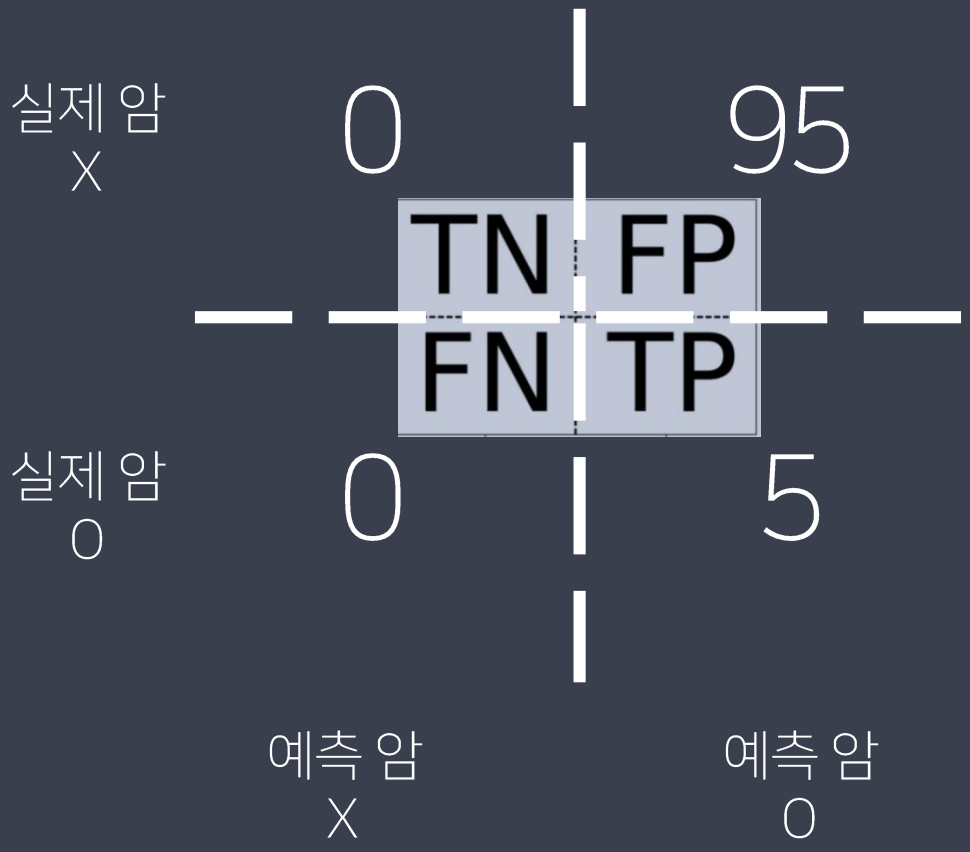
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

정밀도 (Precision)

예측 양성 중에 실제 양성 비율

분류 평가 지표

Confusion_matrix



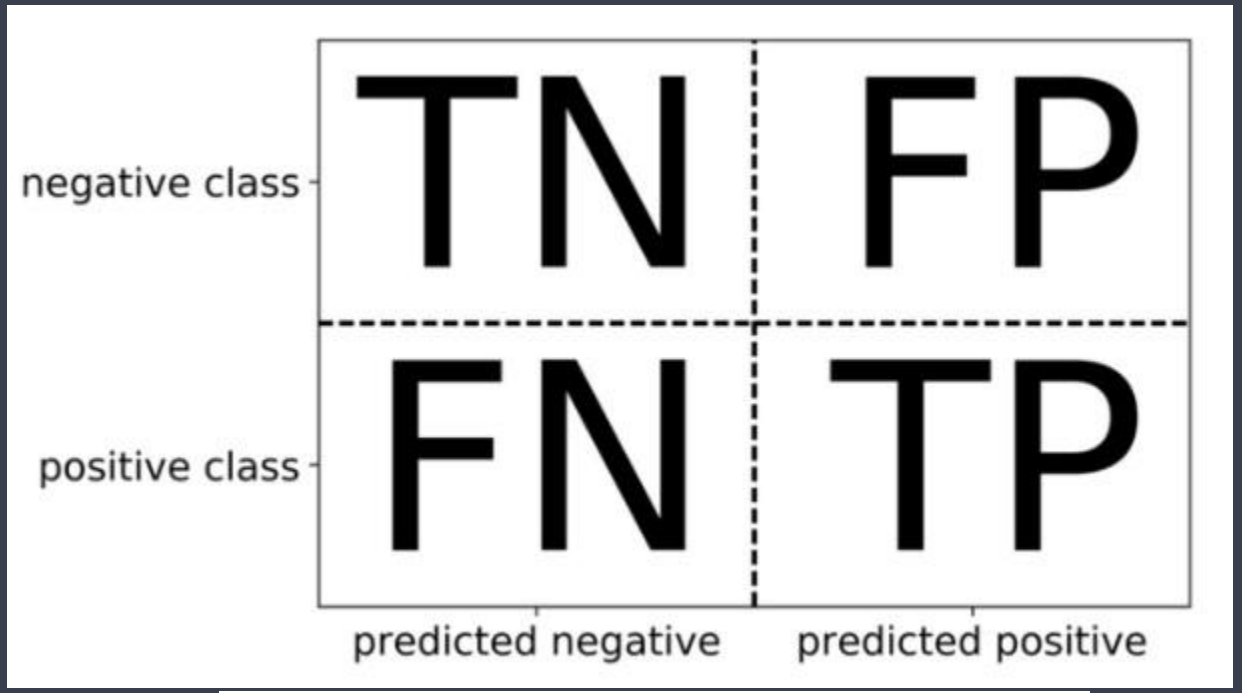
100명 중 암 환자는 5명

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

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

분류 평가 지표

Confusion_matrix



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

F1 - score

정밀도와 재현율의 조화평균

분류 평가 지표

낮은 재현율보다 높은 정밀도를 선호하는 경우

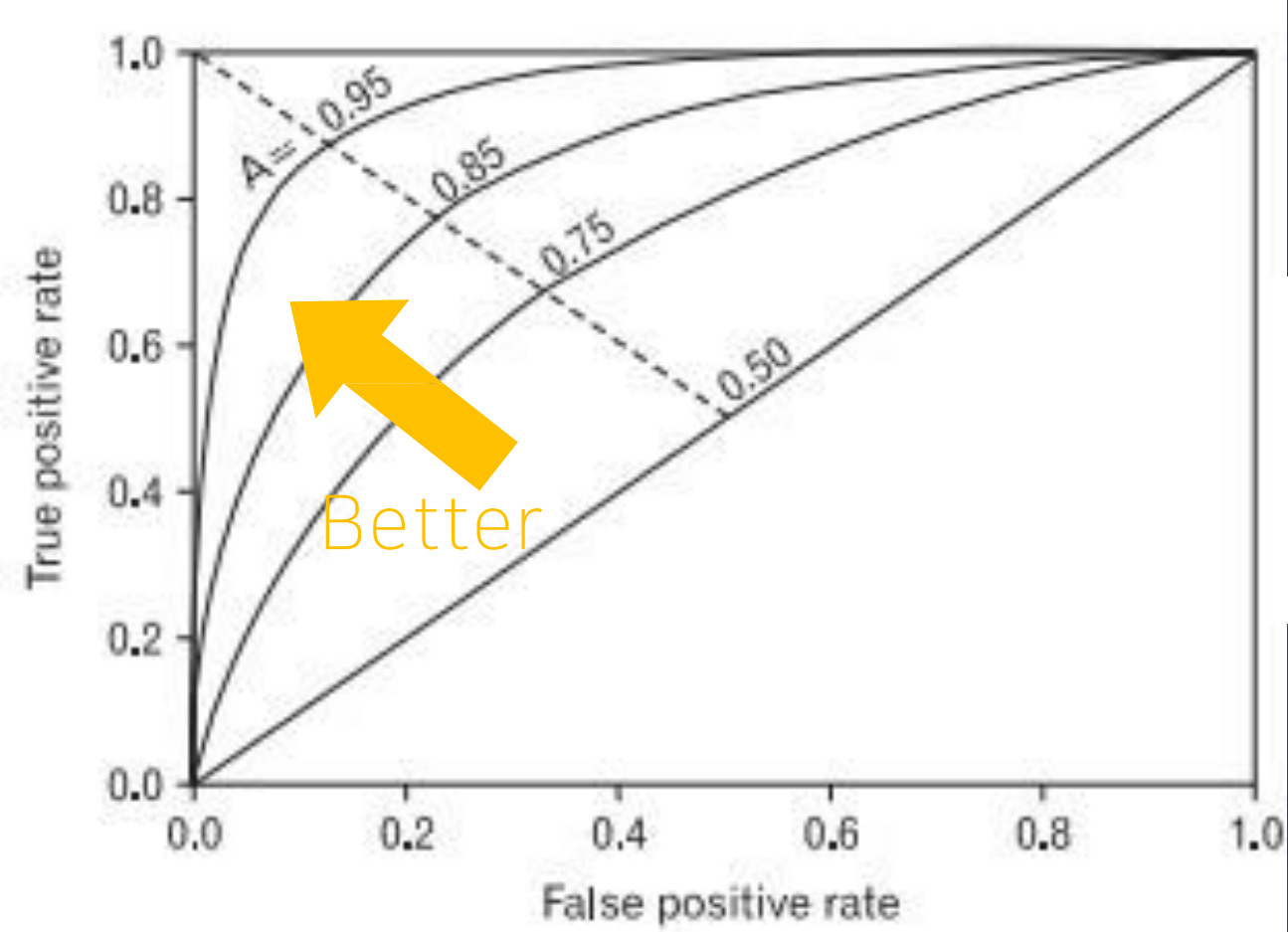
어린아이에게 안전한 동영상(양성)을 걸러내는 분류기를 훈련시킬 경우 좋은 동영상이 많이 제외되더라도(낮은 재현율) 안전한 것들만 노출시키는(높은 정밀도) 분류기가 더 좋다.

낮은 정밀도보다 높은 재현율을 선호하는 경우

감시 카메라로 좀도둑(양성)을 잡아내는 분류기를 훈련시킬 경우 경비원이 잘못된 호출을 종종 받지만(낮은 정밀도) 거의 모든 좀도둑을 잡는(높은 재현율) 분류기가 더 좋다.

분류 평가 지표

ROC curve



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

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

Hand-written digit data classification : ex08 추가

분류 결과의 불확실성 및 분류결과에 대한
평가지표를 확인해보자.