Introduction to Model Evaluation Metrics

김태연

Data Mining & Quality Analytics Lab.

2021.09.17(금)

KOREA UNIVERSITY Data Mining Quality Analytics

발표자 소개

❖ 김태연

- 고려대학교 산업경영공학과
- Data Mining & Quality Analytics Lab. (김성범 교수님)
- M.S. Student (2021.03 ~)

❖ Research Interest

- Machin Learning & Deep Learning
- Open set recognition

❖ Contact

- E-mail : wind0971@korea.ac.kr

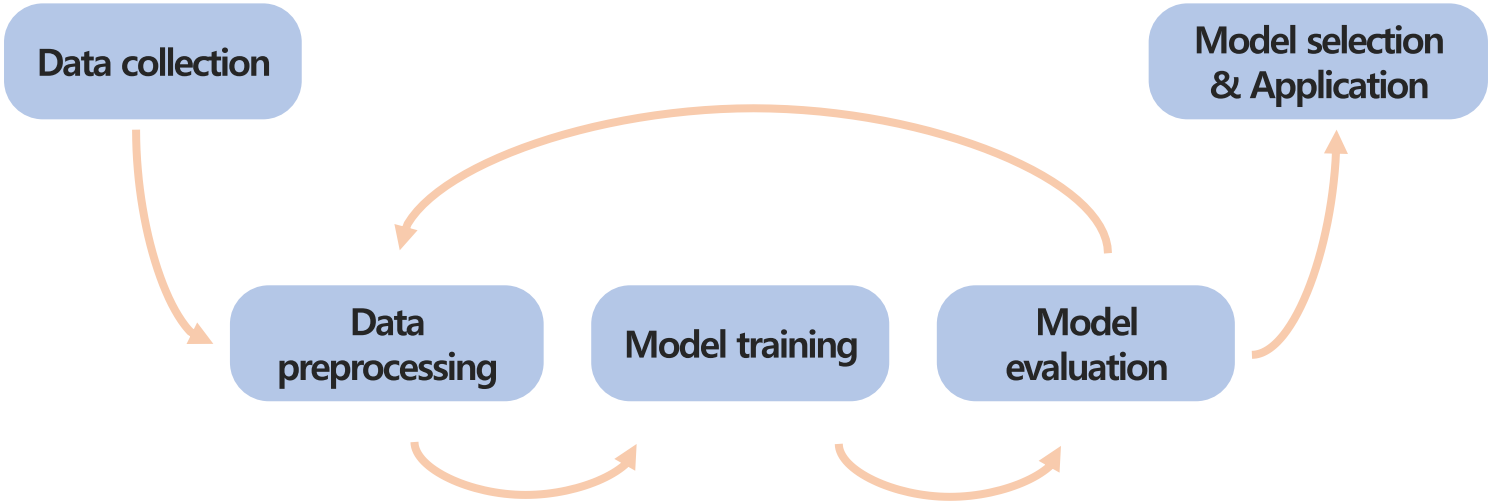
목차

1. Introduction

Introduction

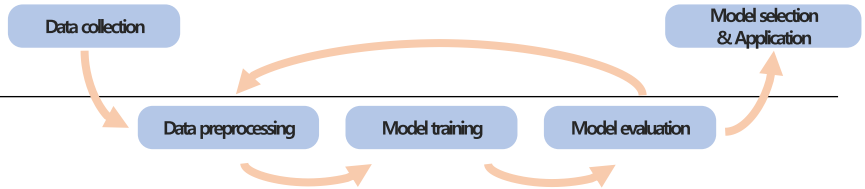
❖ 기계 학습 (Machine learning)

- 기계 학습이란 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야



기계 학습 과정

Introduction



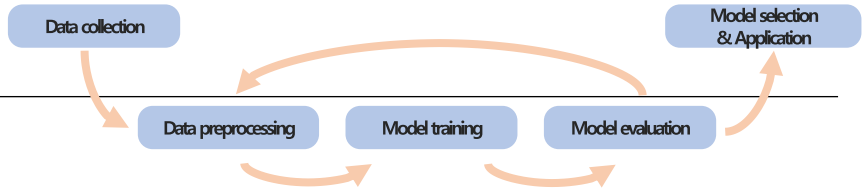
❖ 기계 학습 과정

- 데이터 수집 : 목적에 적합한 다양한 데이터를 모으는 작업
- 데이터 전처리 : 수집된 데이터 중 결측치 혹은 이상치를 정제하는 작업등을 통해 분석에 적합한 형태로 바꿔주는 과정

기온(x_1)	습도(x_2)	풍속(x_3)	일사량(y)
19.3	65.4	3.2	0.18
3.6	Nan	6.3	0.34
15.8	55.3	4.8	0.23
35.1	81.2	Nan	0.42
…	…	…	…

일사량 예측을 위한 데이터 수집

Introduction



❖ 기계 학습 과정

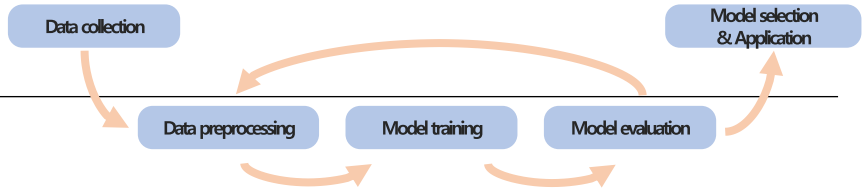
- 데이터 수집 : 목적에 적합한 다양한 데이터를 모으는 작업
- 데이터 전처리 : 수집된 데이터 중 결측치 혹은 이상치를 정제하는 작업등을 통해 분석에 적합한 형태로 바꿔주는 과정

기온(x_1)	습도(x_2)	풍속(x_3)	일사량(y)
19.3	65.4	3.2	0.18
3.6	**Nan**	6.3	0.34
15.8	55.3	4.8	0.23
35.1	81.2	**Nan**	0.42
…	…	…	…

기온(x_1)	습도(x_2)	풍속(x_3)	일사량(y)
19.3	65.4	3.2	0.18
3.6	**35.8**	6.3	0.34
15.8	55.3	4.8	0.23
35.1	81.2	**3.3**	0.42
…	…	…	…

데이터 전처리 작업

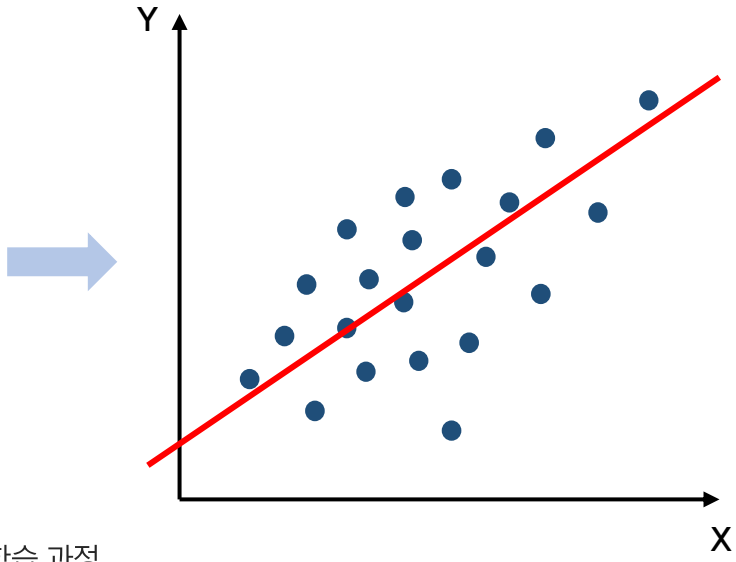
Introduction



❖ 기계 학습 과정

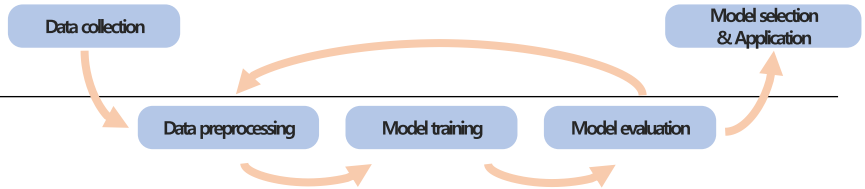
- 모델 학습: 전처리 된 데이터를 통해 올바른 예측이 가능하도록 규칙을 찾는 과정
- 모델 평가 : 학습이 이루어진 모델이 어느 수준의 성능으로 예측이 가능한지 파악하는 과정
- 모델 선택 및 적용 : 모델 평가 과정에서 가장 우수한 성능의 모델을 선택하고, 해당 모델을 활용하여 적용하는 과정

기온(x_1)	습도(x_2)	풍속(x_3)	일사량(y)
19.3	65.4	3.2	0.18
3.6	35.8	6.3	0.34
15.8	55.3	4.8	0.23
35.1	81.2	3.3	0.42
…	…	…	…



모델 학습 과정

Introduction



❖ 기계 학습 과정

- 모델 학습: 전처리 된 데이터를 통해 올바른 예측이 가능하도록 규칙을 찾는 과정
- 모델 평가 : 학습이 이루어진 모델이 어느 수준의 성능으로 예측이 가능한지 파악하는 과정
- 모델 선택 및 적용 : 모델 평가 과정에서 가장 우수한 성능의 모델을 선택하고, 해당 모델을 활용하여 적용하는 과정

기온(x_1)	습도(x_2)	풍속(x_3)	일사량(y)	예측 일사량($\hat{y}$)
21.3	51.2	7.2	0.19	0.15
12.6	41.8	5.1	0.31	0.40
36.5	88.1	6.8	0.20	0.22
27.4	61.2	2.3	0.39	0.23
…	…	…	…	…

모델 평가 과정 및 선택

Introduction

❖ 모델의 평가

- 해결하고자 하는 문제에 대한 명확한 이해를 통한 모델의 평가
- 다양한 모델들 속에서 수행하고자 하는 목적에 가장 부합한 훈련된 모델을 선택하는 과정 필요
- 분류(classification),회귀(regression) 등 각각의 상황에 따라서 사용되는 다양한 평가 지표 존재로 혼란

올바른 모델을 태스크에 적용하기 위해

적합한 평가 지표 선택이 중요

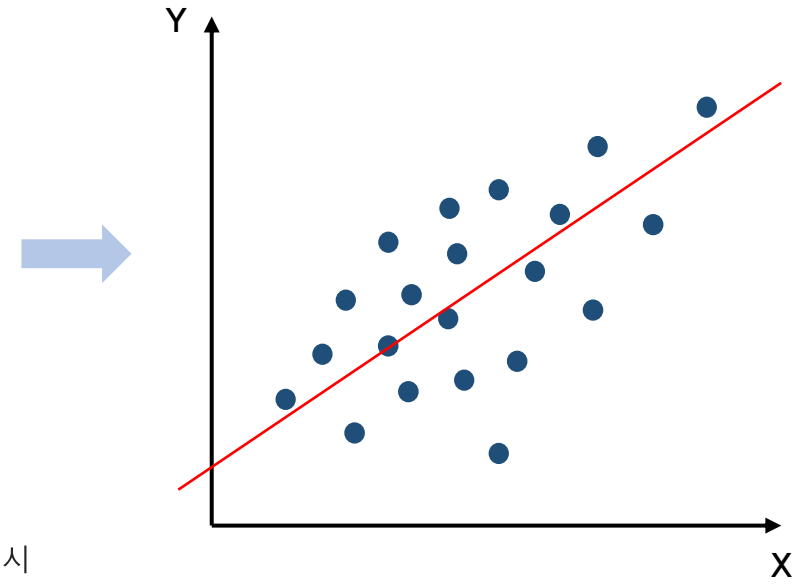
2. Regression model

Regression model evaluation

❖ 회귀 (regression) 모델

- 종속 변수 Y 와 이를 설명하기 위한 독립 변수 X 사이의 관계를 예측하는 모델
- 실제 값(Y)와 모델이 예측한 값($\hat{Y}$) 사이의 차이와 관련된 지표를 통해서 모델을 평가
- 크기 의존적 지표 / 비율을 통한 지표

독립 변수(X)			종속 변수(Y)
기온(x_1)	습도(x_2)	풍속(x_3)	일사량(y)
19.3	65.4	3.2	0.18
3.6	35.8	6.3	0.34
15.8	55.3	4.8	0.23
35.1	81.2	3.3	0.42
…	…	…	…



회귀 모델 예시

Regression model evaluation

❖ Mean absolute error (MAE)

- 실제 값과 모델이 예측한 값 차이의 절대값들의 평균 값
- 절대값을 통해 양의 에러와 음의 에러가 상쇄되는 효과 제거
- 실제 데이터와의 통일된 단위로 분석이 직관적이라는 장점
- 실제 값과 비교해 예측 값이 보다 크거나 작은지 파악의 어려움

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y_i}|$$

N : 데이터의 개수, y_i : i 번째실제 관측 값, $\hat{y_i}$: i 번째 예측 값

Regression model evaluation

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

❖ Mean absolute error (MAE)

- 실제 값과 모델이 예측한 값 차이의 절대값들의 평균 값
- 절대값을 통해 양의 에러와 음의 에러가 상쇄되는 효과 제거
- 실제 데이터와의 통일된 단위로 분석이 직관적이라는 장점
- 실제 값과 비교해 예측 값이 보다 크거나 작은지 파악의 어려움

학생	시험 점수(y)	예측 점수($\hat{y}$)	$y - \hat{y}$
1	83	86	-3
2	86	76	10
3	100	92	8
4	75	79	-4
5	92	89	3

$$MAE : \frac{3 + 10 + 8 + 4 + 3}{5} = 5.6$$

시험 점수를 예측하는 모델이

평균적으로 5.6점 잘못 예측

MAE 평가 지표 예시

Regression model evaluation

❖ Mean squared error (MSE)

- 실제 값과 모델이 예측한 값 차이를 제곱한 값들의 평균 값
- 실제 값과 예측 값의 차이 (=에러)를 제곱하여서 1미만의 에러는 보다 작아지며, 1이상의 에러는 보다 커지게 하여서 값의 왜곡 발생으로 이상치에 보다 민감함
- 실제 데이터와의 단위가 통일 되어 있지 않아 추가적인 해석이 필요하며, 실제 값과 비교해 예측 값이 보다 크거나 작은지 파악의 어려움

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

N : 데이터의 개수, y_i : i 번째실제 관측 값, $\hat{y}_i$: i 번째 예측 값

Regression model evaluation

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

❖ Mean squared error (MSE)

- 실제 값과 모델이 예측한 값 차이를 제곱한 값들의 평균 값
- 실제 값과 예측 값의 차이 (=에러)를 제곱하여서 1미만의 에러는 보다 작아지며, 1이상의 에러는 보다 커지게 하여서 값의 왜곡 발생으로 이상치에 보다 민감함
- 실제 데이터와의 단위가 통일 되어 있지 않아 추가적인 해석이 필요하며, 실제 값과 비교해 예측 값이 보다 크거나 작은지 파악의 어려움

학생	시험 점수(y)	예측 점수($\hat{y}$)	$(y_i - \hat{y}_i)^2$
1	83	86	9
2	86	76	100
3	100	92	64
4	75	79	16
5	92	89	9

$$MSE : \frac{9 + 100 + 64 + 16 + 9}{5} = 39.6$$

시험 점수를 예측하는 모델이

평균적으로 $6.3(\fallingdotseq \sqrt{39.6})$점 잘못 예측

MSE 평가 지표 예시

Regression model evaluation

❖ Root mean squared error (RMSE)

- MSE에 루트를 씌운 값
- MSE에서 에러의 제곱을 통한 왜곡 발생을 방지하며 실제 오류 값들 보다 더욱 커지는 특성을 방지
- 실제 데이터와의 단위 통일로 해석이 용이함
- 실제 값과 비교해 예측 값이 보다 크거나 작은지 파악의 어려움

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

N : 데이터의 개수, y_i : i 번째실제 관측 값, $\hat{y}_i$: i 번째 예측 값

Regression model evaluation

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

❖ Root mean squared error (RMSE)

- MSE에 루트를 씌운 값
- MSE에서 에러의 제곱을 통한 왜곡 발생을 방지하며 실제 오류 값들 보다 더욱 커지는 특성을 방지
- 실제 데이터와의 단위 통일로 해석이 용이함
- 실제 값과 비교해 예측 값이 보다 크거나 작은지 파악의 어려움

학생	시험 점수(y)	예측 점수($\hat{y}$)	$(y_i - \hat{y}_i)^2$
1	83	86	9
2	86	76	100
3	100	92	64
4	75	79	16
5	92	89	9

$$RMSE : \sqrt{\frac{9 + 100 + 64 + 16 + 9}{5}} \fallingdotseq 6.3$$

시험 점수를 예측하는 모델이

평균적으로 $6.3(\fallingdotseq \sqrt{39.6})$점 잘못 예측

RMSE 평가 지표 예시

Regression model evaluation

❖ Root mean squared log error (RMSLE)

- RMSE에 로그를 적용해준 지표
- 직관적이지 않은 값으로 에러를 통한 해석에는 어려움
- RMSE에서 값의 절대적 크기에 영향을 많이 받는 단점을 해결하기 위한 상대적인 에러
- 실제 값보다 작게 예측을 진행하는 경우 더욱 큰 패널티를 부여하는 지표

$$RMSLE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[\log(\hat{y}_i + 1) - \log(y_i + 1)]^2}$$

N : 데이터의 개수, y_i : i 번째실제 관측 값, $\hat{y}_i$: i 번째 예측 값

Regression model evaluation

$$RMSLE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[\log(\hat{y}_i + 1) - \log(y_i + 1)]^2}$$

❖ Root mean squared log error (RMSLE)

- RMSE에 로그를 적용해준 지표
- 직관적이지 않은 값으로 에러를 통한 해석에는 어려움
- RMSE에서 값의 절대적 크기에 영향을 많이 받는 단점을 해결하기 위한 상대적인 에러
- 실제 값보다 작게 예측을 진행하는 경우 더욱 큰 패널티를 부여하는 지표

학생	시험 점수(y)	예측 점수($\hat{y}$)	$[\log\hat{y}_i + 1) - \log(y_i + 1)]^2$
1	83	86	$[\log(87) - \log(84)]^2$
2	86	76	$[\log(77) - \log(87)]^2$
3	100	92	$[\log(93) - \log(101)]^2$
4	75	79	$[\log(80) - \log(76)]^2$
5	92	89	$[\log(89) - \log(93)]^2$

$$RMSLE : \sqrt{\frac{0.0267}{5}} \fallingdotseq 0.073$$

RMSLE 평가 지표 예시

Regression model evaluation

$$RMSLE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[\log(\hat{y}_i + 1) - \log(y_i + 1)]^2}$$

❖ Root mean squared log error (RMSLE)

- RMSE에 로그를 적용해준 지표
- 직관적이지 않은 값으로 에러를 통한 해석에는 어려움
- RMSE에서 값의 절대적 크기에 영향을 많이 받는 단점을 해결하기 위한 상대적인 에러
- 실제 값보다 작게 예측을 진행하는 경우 더욱 큰 패널티를 부여하는 지표

가격(y)	예측 가격($\hat{y}$)
90	100
9000	10000

RMSE	RMSLE
10	0.10
1000	0.10

RMSLE 평가 지표 예시

Regression model evaluation

$$RMSLE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[\log(\hat{y}_i + 1) - \log(y_i + 1)]^2}$$

❖ Root mean squared log error (RMSLE)

- RMSE에 로그를 적용해준 지표
- 직관적이지 않은 값으로 에러를 통한 해석에는 어려움
- RMSE에서 값의 절대적 크기에 영향을 많이 받는 단점을 해결하기 위한 상대적인 에러
- 실제 값보다 작게 예측을 진행하는 경우 더욱 큰 패널티를 부여하는 지표

소요 시간(y)	예측 소요 시간($\hat{y}$)
30	20
30	40

RMSE	RMSLE
10	0.39
10	0.28

작게 예측을 진행하는 경우, 보다 큰 문제가 발생하는 상황에서 해당 지표 사용

RMSLE 평가 지표 예시

Regression model evaluation

❖ Mean absolute percentage error (MAPE)

- MAE를 비율로 변환하여 표현한 값
- 크기 의존적인 에러의 문제점인 단위의 차이가 크게 나타나는 경우 해석이 어려웠던 문제점을 해결
- 실제 값에 0이 포함된 경우에는 사용이 불가능

$$MAPE = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y_i}}{y_i}\right|$$

N : 데이터의 개수, y_i : i 번째실제 관측 값, $\hat{y_i}$: i 번째 예측 값

Regression model evaluation

$$MAPE = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

❖ Mean absolute percentage error (MAPE)

- MAE를 비율로 변환하여 표현한 값
- 크기 의존적인 에러의 문제점인 단위의 차이가 크게 나타나는 경우 해석이 불가능했던 문제점을 해결
- 실제 값에 0이 포함된 경우에는 사용이 불가능

학생	시험 점수(y)	예측 점수($\hat{y}$)	$\left\|{}^{y_i - \hat{y}_i}/_{y_i}\right\|$
1	83	86	0.04
2	86	76	0.12
3	100	92	0.08
4	75	79	0.05
5	92	89	0.03

$$MAPE : \frac{100}{5} \times 0.32 \fallingdotseq 6.36$$

시험 점수를 예측하는 모델이

6.36% 차이가 나도록 예측

MAPE 평가 지표 예시

Regression model evaluation

❖ 결정계수(R^2)

- 데이터의 분산을 기반으로 한 평가 지표로 결정 계수라고 불림
- 데이터의 크기에 따라 영향을 받지 않는 평가지표로, 다른 지표들과는 큰 값을 가질수록 좋은 성능
- 모델이 예측을 모두 평균으로 하는 것보다 큰 오차를 보인다면 음(-)의 값을 보임

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

N : 데이터의 개수, y_i : i 번째실제 관측 값, $\hat{y}_i$: i 번째 예측 값 , $\bar{y}$: 관측 값의 평균

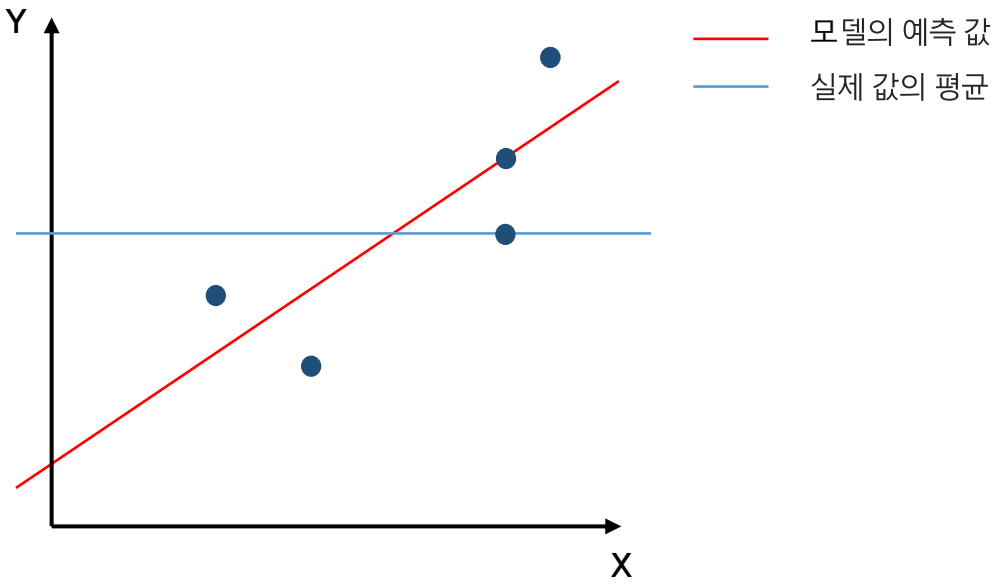
Regression model evaluation

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

❖ 결정계수(R^2)

- 데이터의 분산을 기반으로 한 평가 지표로 결정 계수라고 불림
- 데이터의 크기에 따라 영향을 받지 않는 평가지표로, 다른 지표들과는 큰 값을 가질수록 좋은 성능
- 모델이 예측을 모두 평균으로 하는 것보다 큰 오차를 보인다면 음(-)의 값을 보임

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$



R^2 평가 지표 예시

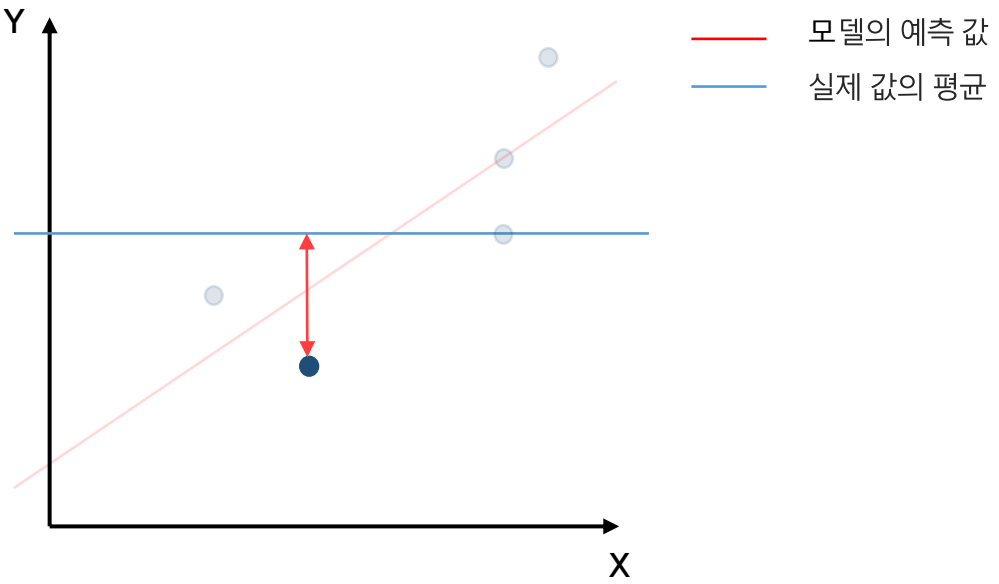
Regression model evaluation

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

❖ 결정계수(R^2)

- 데이터의 분산을 기반으로 한 평가 지표로 결정 계수라고 불림
- 데이터의 크기에 따라 영향을 받지 않는 평가지표로, 다른 지표들과는 큰 값을 가질수록 좋은 성능
- 모델이 예측을 모두 평균으로 하는 것보다 큰 오차를 보인다면 음(-)의 값을 보임

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$



R^2 평가 지표 예시

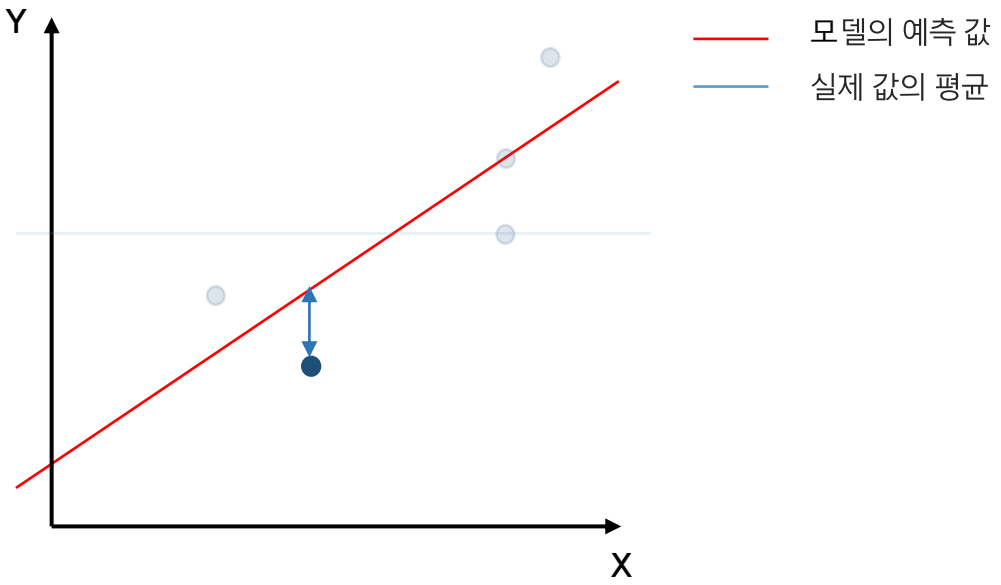
Regression model evaluation

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

❖ 결정계수(R^2)

- 데이터의 분산을 기반으로 한 평가 지표로 결정 계수라고 불림
- 데이터의 크기에 따라 영향을 받지 않는 평가지표로, 다른 지표들과는 큰 값을 가질수록 좋은 성능
- 모델이 예측을 모두 평균으로 하는 것보다 큰 오차를 보인다면 음(-)의 값을 보임

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$



R^2 평가 지표 예시

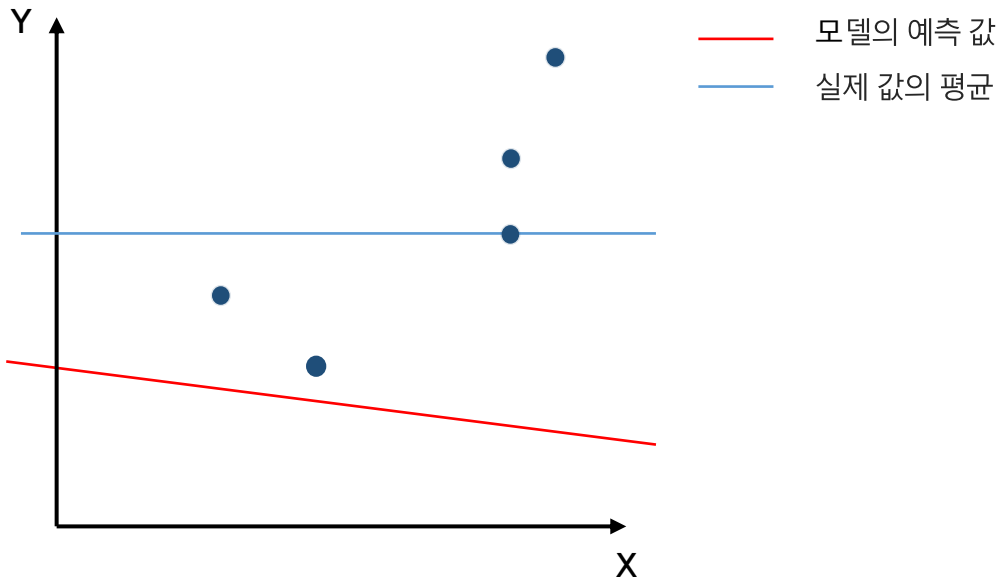
Regression model evaluation

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

❖ 결정계수(R^2)

- 데이터의 분산을 기반으로 한 평가 지표로 결정 계수라고 불림
- 데이터의 크기에 따라 영향을 받지 않는 평가지표로, 다른 지표들과는 큰 값을 가질수록 좋은 성능
- 모델이 예측을 모두 평균으로 하는 것보다 큰 오차를 보인다면 음(-)의 값을 보임

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$



R^2 평가 지표 예시

Regression model evaluation

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

❖ 결정계수(R^2)

- 데이터의 분산을 기반으로 한 평가 지표로 결정 계수라고 불림
- 데이터의 크기에 따라 영향을 받지 않는 평가지표로, 다른 지표들과는 큰 값을 가질수록 좋은 성능
- 모델이 예측을 모두 평균으로 하는 것보다 큰 오차를 보인다면 음(-)의 값을 보임

학생	시험 점수(y)	예측 점수($\hat{y}$)	$y_i - \hat{y}_i$	$y_i - \bar{y}$
1	83	86	-3	-4.2
2	86	76	10	-1.2
3	100	92	8	12.8
4	75	79	-4	-12.2
5	92	89	3	4.8

$\bar{y} = 87.2$

$$R^2 : 1 - \frac{198}{354.8} \fallingdotseq 0.44$$

시험 점수를 예측하는 모델이 44%의 설명력을 가지며 예측

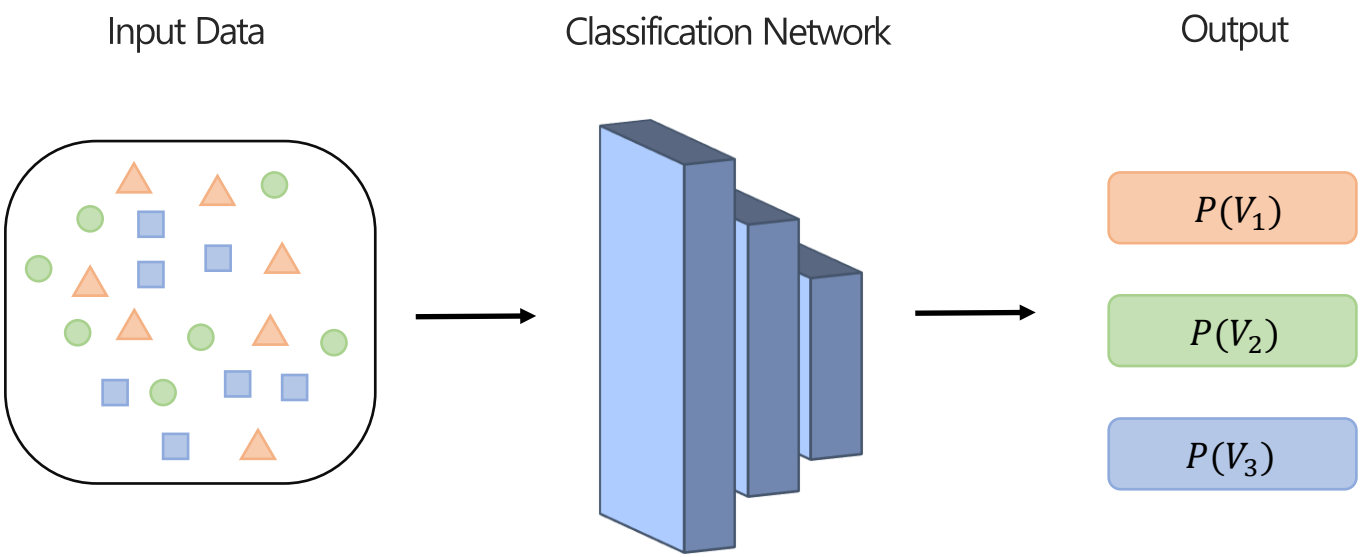
R^2 평가 지표 예시

3. Classification model

Classification model evaluation

❖ 분류(Classification) 문제

- 주어진 데이터들을 각각 속하는 클래스로 나누어주는 문제
- 실제 데이터의 클래스와 모델이 예측한 클래스가 일치하는 비율을 통해서 모델을 평가



분류 모델 예시

Classification model evaluation

❖ 혼동 행렬 (Confusion matrix)

- 실제 정답과 모델의 예측 결과를 행렬의 형태로 표기한 것
- True / False : 실제 정답과 모델의 예측 결과가 동일한지 / 동일하지 않은지를 표현
- Positive / Negative : 모델의 예측 결과에 대한 표현

True Positive

예측 결과

실제 결과	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

이진 분류 결과표

Classification model evaluation

예측 결과

실제 결과	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

❖ 정확도 (Accuracy)

- 전체 데이터 중 올바르게 분류가 이루어진 데이터의 비율
- 해석에 있어서 가장 직관적인 평가 지표지만, 데이터 사이의 불균형의 문제가 있을 경우 적절하지 못한 평가지표

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

예측 결과

실제 결과	고양이	강아지
고양이	46	6
강아지	5	43

$$\text{Accuracy} : \frac{\quad}{46+6+5+43} =$$

Classification model evaluation

예측 결과

실제 결과	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

❖ 정확도 (Accuracy)

- 전체 데이터 중 올바르게 분류가 이루어진 데이터의 비율
- 해석에 있어서 가장 직관적인 평가 지표지만, 데이터 사이의 불균형의 문제가 있을 경우 적절하지 못한 평가지표

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

예측 결과

실제 결과	고양이	강아지
고양이	46	6
강아지	5	43

$$\text{Accuracy} : \frac{46+43}{46+6+5+42} = 0.89$$

해당 분류 모델이

89%의 정확도를 가지며 분류

Classification model evaluation

예측 결과

실제 결과	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

❖ 정밀도 (Precision)

- 모델이 positive로 분류를 진행한 데이터 중에서, 실제 결과도 positive인 데이터의 비율
- 모델이 positive로 예측하였지만 실제 결과가 negative일 때 큰 문제가 생기는 상황에서 사용되는 지표

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

예측 결과

실제 결과	스팸 메일	정상 메일
스팸 메일	55	23
정상 메일	45	37

$$\text{Precision} : \frac{}{55+45} =$$

Classification model evaluation

예측 결과

실제 결과	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

❖ 정밀도 (Precision)

- 모델이 positive로 분류를 진행한 데이터 중에서, 실제 결과도 positive인 데이터의 비율
- 모델이 positive로 예측하였지만 실제 결과가 negative일 때 큰 문제가 생기는 상황에서 사용되는 지표

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

예측 결과

실제 결과	스팸 메일	정상 메일
스팸 메일	55	23
정상 메일	45	37

$$\text{Precision} : \frac{55}{55+45} = 0.55$$

해당 분류 모델이

55%의 정밀도로 스팸메일을 분류

Classification model evaluation

예측 결과

실제 결과	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

❖ 재현율 (Recall)

- 실제 결과가 positive인 데이터 중에서, 모델이 positive로 분류한 데이터의 비율
- 실제 값이 positive일 때, 모델의 예측 값도 positive인 경우가 중요한 상황일 때 사용되는 지표

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

예측 결과

실제 결과	암 환자	정상 환자
암 환자	15	85
정상 환자	25	9875

$$\text{Recall} : \frac{}{15+85} =$$

Classification model evaluation

예측 결과

실제 결과	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

❖ 재현율 (Recall)

- 실제 결과가 positive인 데이터 중에서, 모델이 positive로 분류한 데이터의 비율
- 실제 값이 positive일 때, 모델의 예측 값도 positive인 경우가 중요한 상황일 때 사용되는 지표

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

예측 결과

실제 결과	암 환자	정상 환자
암 환자	15	85
정상 환자	25	9875

$$\text{Recall} : \frac{15}{15+85} = 0.15$$

해당 분류 모델이

15%의 재현율로 암환자를 분류

Classification model evaluation

예측 결과

실제 결과	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

❖ F_β - score

- 정밀도와 재현율의 trade-off 관계로 두 값을 동시에 높게 하는 것이 어려움
- 정밀도와 재현율의 가중조화평균을 활용하여 정밀도와 재현율의 값을 하나의 값으로 표현한 지표
- 정밀도에 주어진 가중치를 β라 하며, β = 1인 경우를 F_1 - score를 의미

$$F_\beta \text{ - score} = \frac{(1+\beta^2)(precision \times recall)}{\beta^2 precision + recall}$$

예측 결과

실제 결과	암 환자	정상 환자
암 환자	15	85
정상 환자	25	9875

$$F_1 \text{ - score} : \frac{0.38 \times 0.15}{0.38 + 0.15} \fallingdotseq 0.11$$

Classification model evaluation

예측 결과

실제 결과	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

❖ F_β - score

- 정밀도와 재현율의 trade-off 관계로 두 값을 동시에 높게 하는 것이 어려움
- 정밀도와 재현율의 가중조화평균을 활용하여 정밀도와 재현율의 값을 하나의 값으로 표현한 지표
- 정밀도에 주어진 가중치를 β라 하며, β = 1인 경우를 F_1 - score를 의미

$$F_\beta \text{ - score} = \frac{(1+\beta^2)(precision \times recall)}{\beta^2 precision + recall}$$

	Model A	Model B
정밀도	0.9	0.5
재현율	0.1	0.5
F_1 - score	0.18	0.5
Mean	0.5	0.5

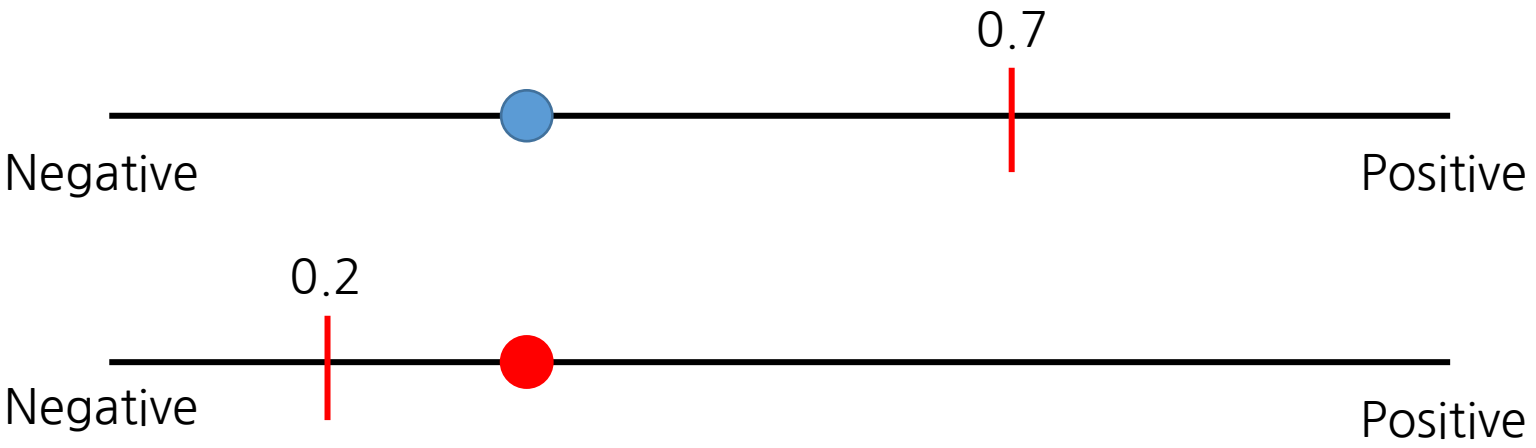
정밀도와 재현율이 극단값을 나타내지 않을 때 보다 높은 성능으로 표기하는 지표를 표현하기 위함

Classification model evaluation

실제 결과 \ 예측 결과	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

❖ 임계값 (Threshold)

- 특정 클래스로 분류하기 위한 기준치로 이진 분류에서는 기본적으로 0.5로 설정
- 임계값 설정에 따라서 정밀도와 재현율의 값이 달라지는데, 정밀도와 재현율을 균형 있게 예측하는 적절한 임계값을 설정



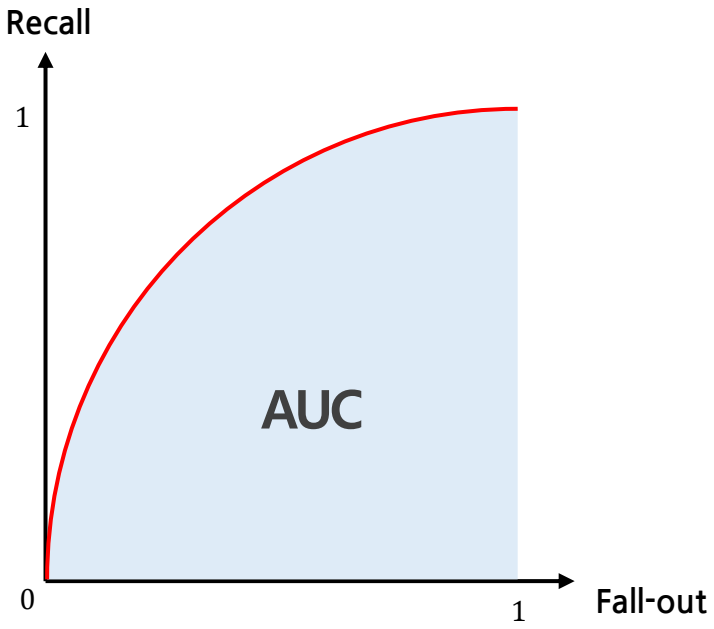
임계값 변화에 따른 예측 결과의 변화

Classification model evaluation

실제 결과 \ 예측 결과	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

❖ ROC curve & AUC

- ROC curve : X축을 Fall-out 지표, Y축을 Recall 지표로 하는 그래프로 표현하는 방식
- Fall-out : 실제 결과가 negative인 데이터 중에서, 모델이 positive로 분류한 데이터의 비율
- AUC : ROC curve를 수치화한 값으로, ROC curve의 면적을 의미



Classification model evaluation

❖ Overconfidence in deep learning

- Overconfidence : 모델에서 특정 클래스로 예측하는 확률이 실제 accuracy 보다 높게 반영된 것
- 다양한 모델에서는 overconfidence한 현상이 존재
- 분류가 명확하게 이루어지는 것도 중요하지만, 모델의 잘못된 예측에 대한 이해도 중요

$$|confidence\ score\ - accuracy|$$

$$confidence\ score = accuracy$$

4. Conclusion

Conclusion

❖ 다양한 상황 속에서 사용되는 평가 지표의 다양함

- 알맞은 평가 지표를 사용하지 않고 모델 선택 시 잘못된 모델이 적용
- 문제 상황에 적합한 평가 지표를 사용하여 모델의 평가 진행

❖ 잘못된 예측에 대한 이해

- 잘못된 예측이 큰 문제로 다가오는 사례가 잦아지며 uncertainty에 대한 이해가 중요
- 모델의 예측 확률 값을 통해 계산된 모델의 신뢰도를 활용하여 모델 평가

감사합니다

Reference

1. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017, July). On calibration of modern neural networks. In International Conference on Machine Learning (pp. 1321-1330). PMLR.
2. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., ... & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. arXiv preprint arXiv:1906.02530.
3. Lee, K., Lee, H., Lee, K., & Shin, J. (2017). Training confidence-calibrated classifiers for detecting out-of-distribution samples. arXiv preprint arXiv:1711.09325.
4. Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. arXiv preprint arXiv:1811.12808.