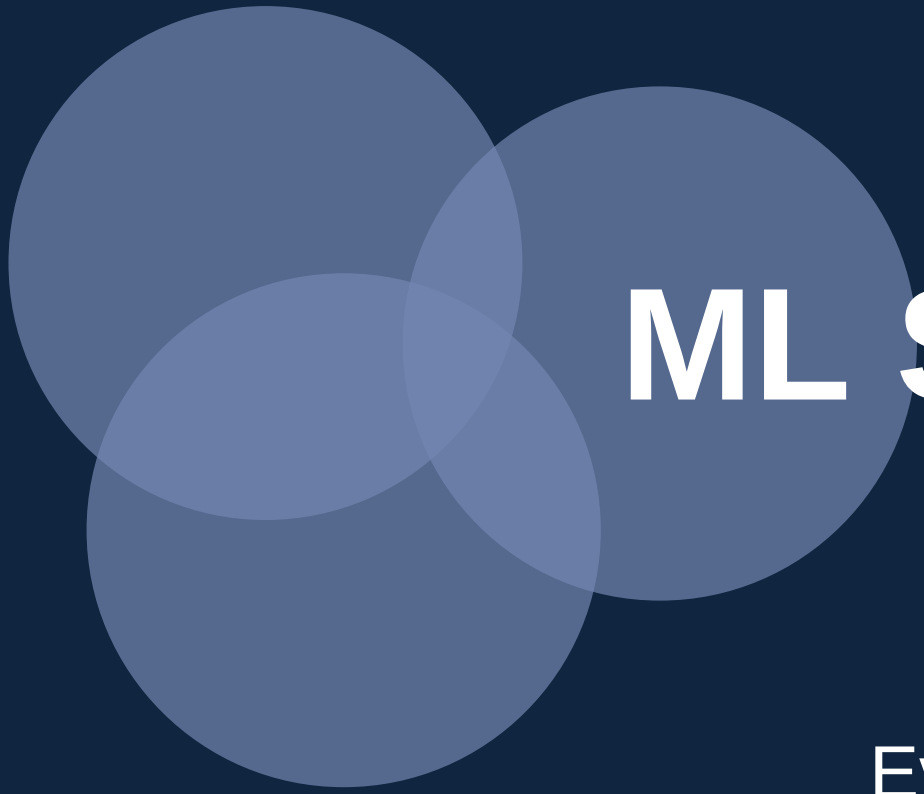
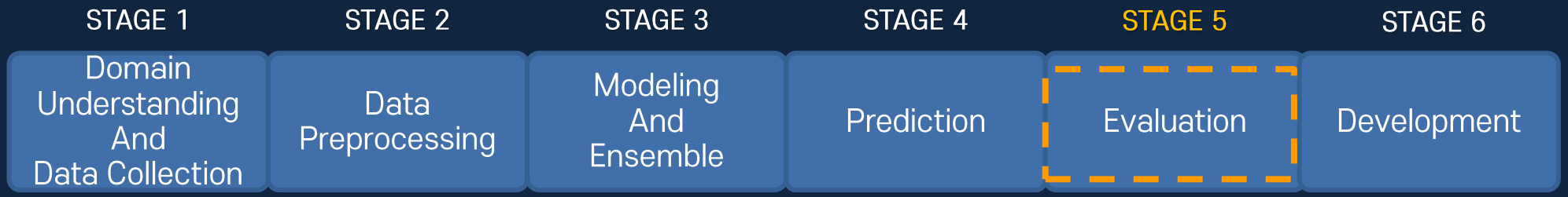


ML SESSION

#03
Evaluation Metrics

INDEX

0 평가지표의 사용단계


STAGE 1
Domain Understanding And Data Collection
STAGE 2
Data Preprocessing
STAGE 3
Modeling And Ensemble
STAGE 4
Prediction
STAGE 5
Evaluation
STAGE 6
Development

1 여러가지 평가지표


머신러닝
지도학습
분류문제
: 정확도, 정밀도, 재현율(민감도), 특이도,
F1 score, ROC-AUC, Log Loss
회귀문제
: MAE, MSE, RMSE, MSLE, RMSLE, R2
비지도학습
군집 분석
: ARI, NMI, 실루엣계수(silhouete coefficient)
차원 축소
: 특별한 평가지표가 존재하지 않는다.

2 분류문제 평가지표

오차행렬

예측 클래스
(Predicted Class)

실제 클래스
(Actual Class)

	Negative(0)	Positive(1)
Negative(0)	TN (True Negative)	FP (False Positive)
Positive(1)	FN (False Negative)	TP (True Positive)

❖ 이때, 긍정과 부정에 상관없이 우리가 관심있는 결과가 Positive가 됨을 명심하자.

2 분류문제 평가지표

실제 클래스 (Actual Class) \ 예측 클래스 (Predicted Class)	Negative(0)	Positive(1)
Negative(0)	TN (True Negative)	FP (False Positive)
Positive(1)	FN (False Negative)	TP (True Positive)

○ TP(True Positive) : 실제 Positive인 정답을 Positive라고 예측 (True)

○ TN(True Negative) : 실제 Negative인 정답을 Negative라고 예측 (True)

○ FP(**False Positive**): 실제 Negative인 정답을 Positive라고 예측 (False) – *Type I error*

○ FN(**False Negative**): 실제 Positive인 정답을 Negative라고 예측 (False) – *Type II error*

2 분류문제 평가지표

Accuracy (정확도)

실제 클래스 (Actual Class) \ 예측 클래스 (Predicted Class)	Negative(0)	Positive(1)
Negative(0)	TN (True Negative)	FP (False Positive)
Positive(1)	FN (False Negative)	TP (True Positive)

$$\frac{TP + TN}{TP + TN + FP + FN}$$

❖ 정의 : 전체 예측 건수에서 정답을 맞춘 건수의 비율

❖ 단점 : 데이터의 불균형이 심할 경우 정확한 성능을 측정하기 어려움

→ 만약, 성별분류 문제에서 데이터의 정답이 90%가 남성이고 10%가 여성일 경우 모델이 모든 데이터를 남성으로 예측한다고 해도 정확도가 90%이므로 제대로 된 평가를 할 수 없다.

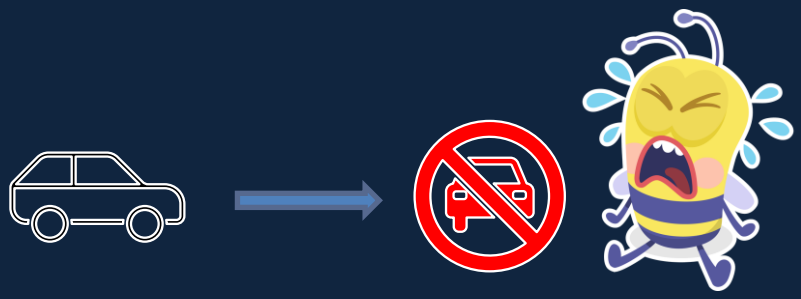
2 분류문제 평가지표

Type I error

예측 클래스
(Predicted Class)

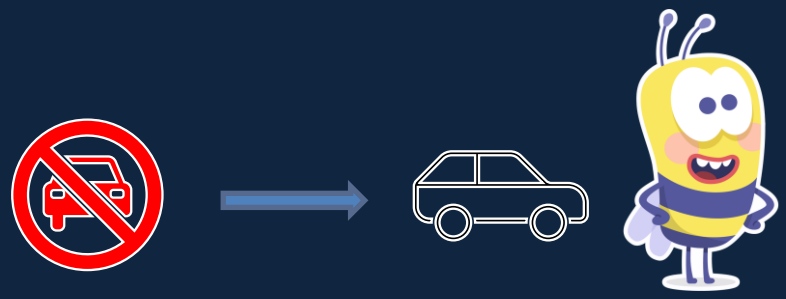
실제 클래스 (Actual Class)	Negative(0)	Positive(1)
Negative(0)	TN (True Negative)	FP (False Positive)
Positive(1)	FN (False Negative)	TP (True Positive)

* 실제 Negative인 정답을 Positive라고 예측 (False)



예측 클래스 : Positive
-> 좋은 자동차

실제 클래스 : Negative
-> 나쁜 자동차



예측 클래스 : Negative
-> 나쁜 자동차

실제 클래스 : Positive
-> 좋은 자동차

2 분류문제 평가지표

Type II error

예측 클래스
(Predicted Class)

실제 클래스 (Actual Class)	Negative(0)	Positive(1)
Negative(0)	TN (True Negative)	FP (False Positive)
Positive(1)	FN (False Negative)	TP (True Positive)

* 실제 Positive인 정답을 Negative라고 예측 (False)

2 분류문제 평가지표

Recall (재현율,민감도)

예측 클래스
(Predicted Class)

실제 클래스 (Actual Class)	Negative(0)	Positive(1)
Negative(0)	TN (True Negative)	FP (False Positive)
Positive(1)	FN (False Negative)	TP (True Positive)

$$\frac{TP}{TP + FN}$$

❖ 정의 : 실제로 정답이 True 인 것들 중 분류기가 True로 예측한 비율

→ 데이터에서 True가 발생하는 확률이 적을 때 사용하면 좋은 평가지표

❖ 단점 : 남성 10%, 여성 90%의 불균형 데이터의 분류 문제에서 만약 True로만 답하는 분류기가 있다고 할 때 , 남성의 경우 Accuracy는 낮지만 Recall은 1이 된다.

2 분류문제 평가지표

Precision (정밀도)

예측 클래스
(Predicted Class)

실제 클래스 (Actual Class)	Negative(0)	Positive(1)
Negative(0)	TN (True Negative)	FP (False Positive)
Positive(1)	FN (False Negative)	TP (True Positive)

$$\frac{TP}{TP + FP}$$

❖ 정의 : True 로 예측한 것들 중 실제 정답이 True인 비율

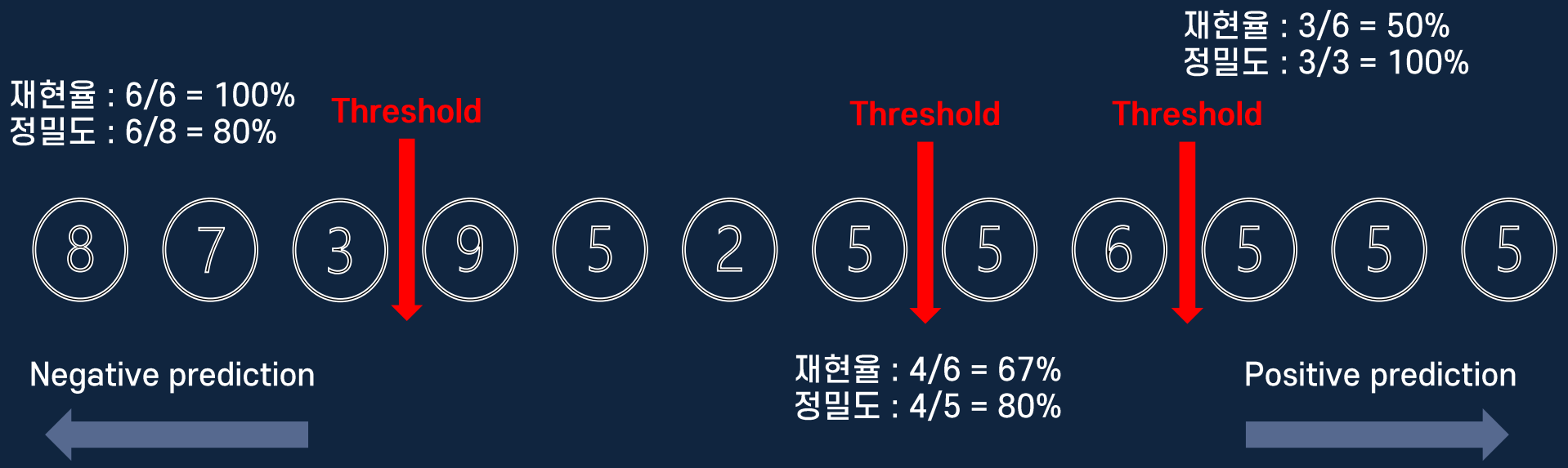
→ 즉, 언제나 True만 답하는 분류기는 Recall은 1이지만 precision은 0에 가깝다.

❖ 단점 : Recall의 장점이 Precision의 단점이 된다. 두 지표는 서로 반대의 개념이기 때문이다.

2 분류모델 평가지표

Threshold

* 재현율(Recall) : 실제로 정답이 True 인 것들 중 분류기가 True로 예측한 비율
* 정밀도(Precision) : True 로 예측한 것들 중 실제 정답이 True인 비율



❖ Threshold : 모델이 0,1을 나누는 기준

→ 예를들어, 모델의 Threshold가 0.5이면 0.5보다 작으면 0, 크면 1로 예측한다.

2 분류문제 평가지표

F1 Score

$$2 \times \frac{Precision \times Recall}{Precision + Recall}$$

❖ 정의 : Precision과 Recall의 조화평균

→ 반대 개념인 두 지표를 모두 균형 있게 반영하기 위함이다. 이 지표는 어느 한쪽에 치우치지 않았을 때 높은 값이 가지게 된다.

❖ Example

: Recall 1, Precision 0.01의 값을 가지는 분류기가 있다고 하자.

→ 산술평균 : (1 + 0.01) / 2 = 0.505

→ 조화평균 : 2 * (1 * 0.01) / (1 + 0.01) = 0.019

2 분류문제 평가지표

예측 클래스
(Predicted Class)

실제 클래스 (Actual Class)	Negative(0)	Positive(1)
Negative(0)	TN (True Negative)	FP (False Positive)
Positive(1)	FN (False Negative)	TP (True Positive)

- True Positive (truth = 1, guess = 1)
- True Negative (truth = 0, guess = 0)
- False Positive (truth = 0, guess = 1) *- Type I error*
- False Negative (truth = 1, guess = 0) *- Type II error*

▷ **Accuracy** : 분류기가 정답을 맞춘 비율

▷ **Recall** : $\frac{TP}{TP+FN}$ → 즉, Type II error 가 더 중요할 경우에 살펴봐야 할 평가지표이다.

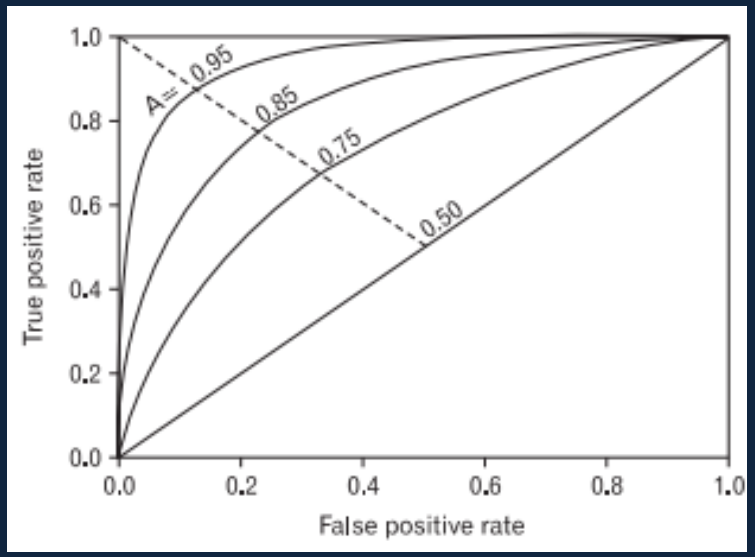
▷ **Precision** : $\frac{TP}{TP+FP}$ → 즉, Type I error 가 더 중요할 경우에 살펴봐야 할 평가지표이다.

▷ **F1 Score** : 재현율과 정밀도 모두를 조화롭게 살펴볼 때 측정하는 평가지표이다.

2 분류문제 평가지표

ROC-AUC : ROC curve

: 보통 binary classification (이진분류) 나 medical application 에서 많이 쓰는 평가 지표이다.



▷ x축 : 1-특이도(Specificity)

▷ y축 : 재현율 (Recall)

* 특이도란?

	Negative(0)	Positive(1)
Negative(0)	TN (True Negative)	FP (False Positive)
Positive(1)	FN (False Negative)	TP (True Positive)

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

→ Threshold를 다르게 했을 때 성능변화를 그래프로 그린 것으로 모델의 성능을 평가하거나 최적의 Threshold를 찾을 수 있다.

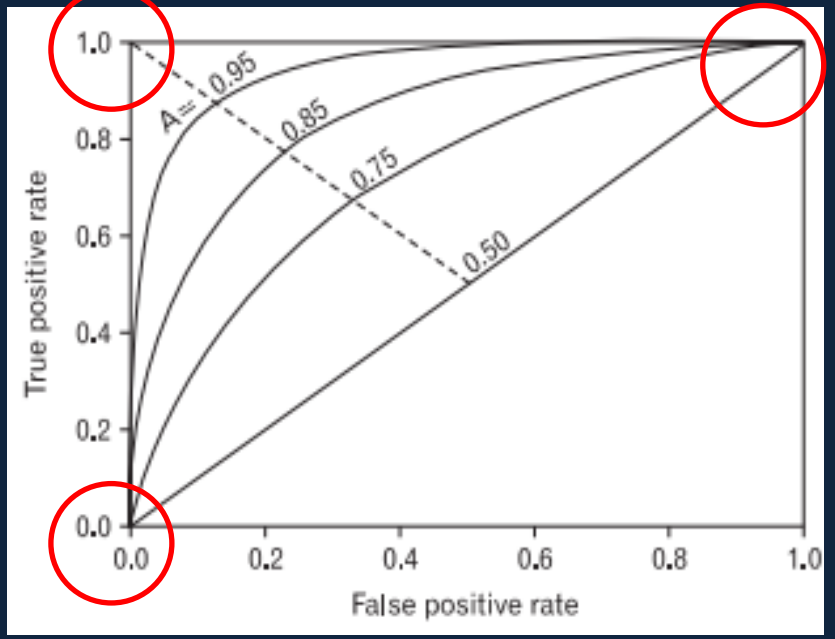
2 분류문제 평가지표

ROC-AUC

→ ROC curve 아래 면적을 AUC(Area Under the Curve) 라고 부른다.
→ 아래 면적이 넓을 수록 좋은 성능을 가진 모델

잘못 예측한 것 없이 모두 맞춘 경우로 완벽한 모델인 경우

모든 예측을 Positive(1) 으로 한 경우



모든 예측을 Negative(0) 으로 한 경우

2 분류문제 평가지표

그 외 평가지표

❖ Log Loss

:모델이 예측한 확률 값을 음의 log함수에 넣어 변환시킨 값으로 평가를 진행한다.

: 잘못된 분류에 패널티를 적용하여 모델의 정확도를 향상시킨다.

즉, 0에 가까울 수록 좋은 모델이다.

다중분류 평가지표

❖다중분류 평가지표 기본편

: Accuracy, Recall, Precision, F-score , Fall-out

https://moons08.github.io/datascience/classification_score_basic/

❖다중분류 평가지표 ROC_AUC 편

https://moons08.github.io/datascience/classification_score_roc_auc/

3 회귀문제 평가지표

MAE (Mean Absolute Error)

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

→ 예측값과 실제값의 차이의 절댓값을 모두 더해 평균한 값

→ 에러의 절댓값 그 자체이기 때문에 낮을수록 좋은 값

MSE (Mean Squared Error)

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

→ 예측값과 실제값을 제곱하여 평균한 값

→ 제곱을 하기 때문에 큰 오차가 강조되는 효과를 가진다.

▷ 단점

→ 에러를 제곱하기 때문에 1 미만의 에러는 작아지고, 그 이상의 에러는 커지는 값의 왜곡이 발생

→ MAE에 비해 이상치에 민감하다.

3 회귀문제 평가지표

RMSE (Root Mean Squared Error)

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

→ MSE에 루트를 씌운 평가 지표

→ 루트로 인해 에러를 제곱하여 생기는 값의 왜곡이 완화

RMSLE (Root Mean Squared Log Error)

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

→ RMSE에 로그를 추가한 평가지표

→ RMSLE는 예측 값이 실제 값보다 작을 경우 더 높은 패널티가 주어진다.

3 회귀문제 평가지표

MAPE (Mean Absolute Percentage Error)

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

→평균 절대 백분율 오차

→ A_t 는 실제값, F_t 는 예측값이다.

▷ 단점

→ 실제 값이 1보다 작을 경우 무한대로 값이 발산할 수 있다.

→ 실제 값이 0이라면 해당 지표는 사용할 수 없다.

R-Squared

→ 실제 값의 분산 대비 예측값의 분산 비율을 지표로 평가하며 1에 가까울 수록 예측 정확도가 높다.

→ 다른 회귀 성능지표인 RMSE나 MAE는 데이터의 Scale에 따라 값이 매우 다르지만, r-squared의 경우 상대적인 성능이기 때문에 직관적으로 성능을 판단할 수 있다.

3 회귀문제 평가지표

※ cross_val_score 와 같은 sklearn의 Scoring 함수에 회귀 평가 적용시 주의할 점

scoring 함수에 회귀 평가 지표를 적용 할 때는

MAE : 'neg_mean_absolute_error'
MSE : 'neg_mean_squared_error'
R-Square : 'r2'

로 scoring 파라미터에 적어주어야 함.

▷ Sklearn에서는 score값이 클수록 좋은 평가 결과로 자동 평가해버리기 때문에 -을 곱해서 반대의 의미로 만들어 적용된다.

4 비지도학습의 평가

타깃 값으로 군집 평가하기 : ARI, NMI

: 위 두가지 평가지표는 실제 정답 클러스터를 알고 있는 경우 사용할 수 있다.

→ 최적의 분류일 경우 1, 무작위적인 분류일 경우 0에 가까운 값으로 나타난다.

타깃 값 없이 군집 평가하기 : 실루엣 계수

: 실루엣 계수는 클러스터의 밀집 정도를 계산하는 것으로 1에 가까운 값일 수록 좋다.

→ 하지만 실제로 잘 동작하지는 않는다.

5 과제

1. 하이퍼파라미터 튜닝(Grid search, Random search) 과 CV를 사용하여 성능 올리기

→ Submission은 각 조 멘토에게 두 번까지 제출할 수 있습니다.

→ 제출양식과 기한은 지난주와 동일합니다.

2. 여러가지 분류 평가지표를 사용하여 성능을 검증해보고 현재 데이터에 가장 적합한 평가지표가 무엇인지 토론해보기

THANK YOU