D&A

ML Session 2차시

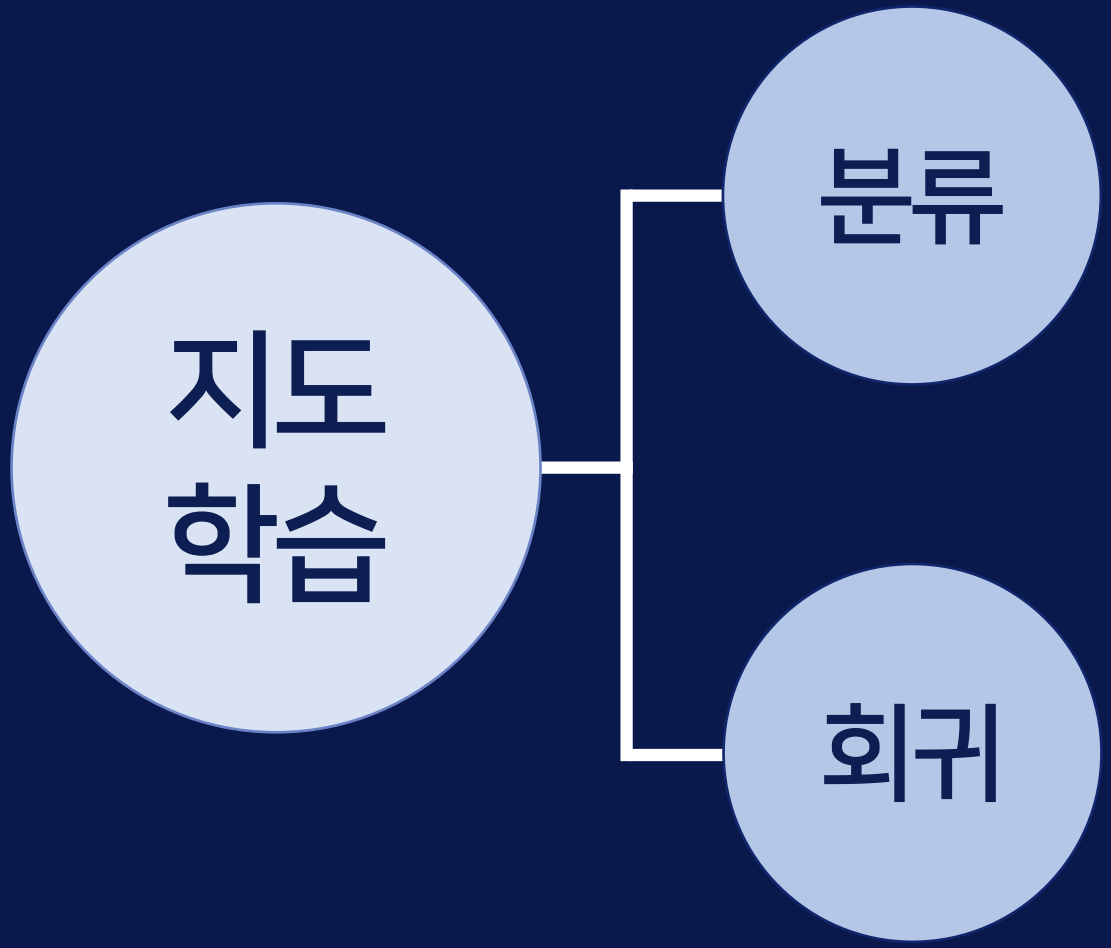
머신러닝 기초2.

2022 / 09 / 13

D&A 운영진 윤경서

CONTENTS.

01. 분류와 회귀



- 종속변수 y : 이름 혹은 문자 (= 범주형 변수)
- 예측결과가 이산값이다.

※ 이산값 : yes/no와 같은 2가지 답으로 나뉘는 이진 분류,
2가지 이상의 답으로 나뉘는 다중 분류
➡ 종류 예측

- 종속변수 y : 숫자 (= 양적 변수)
- 예측결과가 연속성을 지닌다.

※ 연속성 : 연속하는 값
➡ 연속된 값 예측

01. 분류와 회귀

분류 예시

- 공부시간(x)을 입력 받아 합격 여부(y)를 예측
- 메일 발신인, 제목, 본문 내용(x)을 입력 받아 스팸 메일 여부(y)를 예측
- X-ray 사진과 영상 속 종양의 크기, 두께(x)를 입력 받아 악성 종양 여부(y)를 예측

회귀 예시

- 공부시간(x)을 입력 받아 시험점수(y)를 예측
- 온도(x)를 입력 받아 레모네이드 판매량(y)를 예측
- 자동차 속도(x)를 입력 받아 충돌 시 사망확률(y)를 예측

02. 과적합

과대적합(overfitting)이란?

'학습이 너무 잘 된 상태'를 의미하는 것으로 학습 데이터를 너무 잘 학습하여 학습 데이터가 아닌 새로운 데이터에 대해서는 제대로 예측을 하지 못하는 것을 말한다.

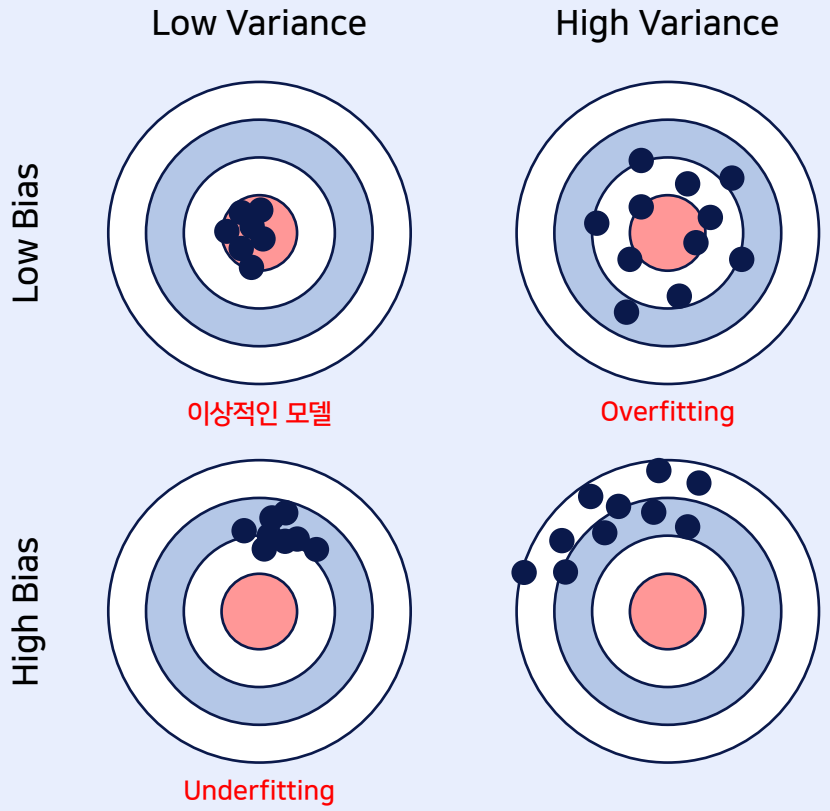
과소적합(underfitting)이란?

모델이 너무 단순하여 학습 데이터조차 제대로 학습하지 못하는 것을 말한다.

편향-분산 트레이드 오프 (Bias-Variance Trade off)

- 편향(Bias) : 예측값이 정답과 얼마나 멀리 떨어져 있는지
- 분산(Variance) : 예측값끼리의 차이

Low Bias / Low Variance : 이상적인 모델
Low Bias / High Variance : Overfitting
High Bias / Low Variance : Underfitting



02. 과적합

과적합 해결방법

- 학습 데이터의 양을 늘리는 것
- 데이터 정규화, Dropout, 앙상블, 교차검증,,,,,

기계학습 모델 성능 평가 단계

기계학습에서 일반적으로 train set으로 모델을 훈련, test set으로 모델 검증을 진행

전체 데이터 전부를 훈련에 이용할 수 없음 & 만일 노이즈 값이 큰 데이터들이 한 쪽에 쏠린다면
➡ 훈련 및 검증이 제대로 되지 않는 문제가 있음

고정된 train set과 test set을 사용하다 보면
➡ test set에 최적의 성능을 발휘하도록 과적합이 발생할 수 있음

03. 교차검증

교차검증 idea

Train set과 Test set을 변경해보자!
(= 즉, 모든 데이터셋을 훈련과 평가에 모두 활용하자)

장점

- 통계적인 평가방법으로 일반화 성능을 가능
- 모든 데이터셋을 학습에 사용함으로써 정확도 향상
- 특정 데이터셋에 대해 과적합 방지
- 데이터 규모가 작다면 과소적합 방지

※ 일반화 성능
: '이전에 본적 없는 데이터'에 대해서도 잘 수행하는 능력

단점

- 모델 훈련 및 평가 소요시간 증가

교차검증 종류

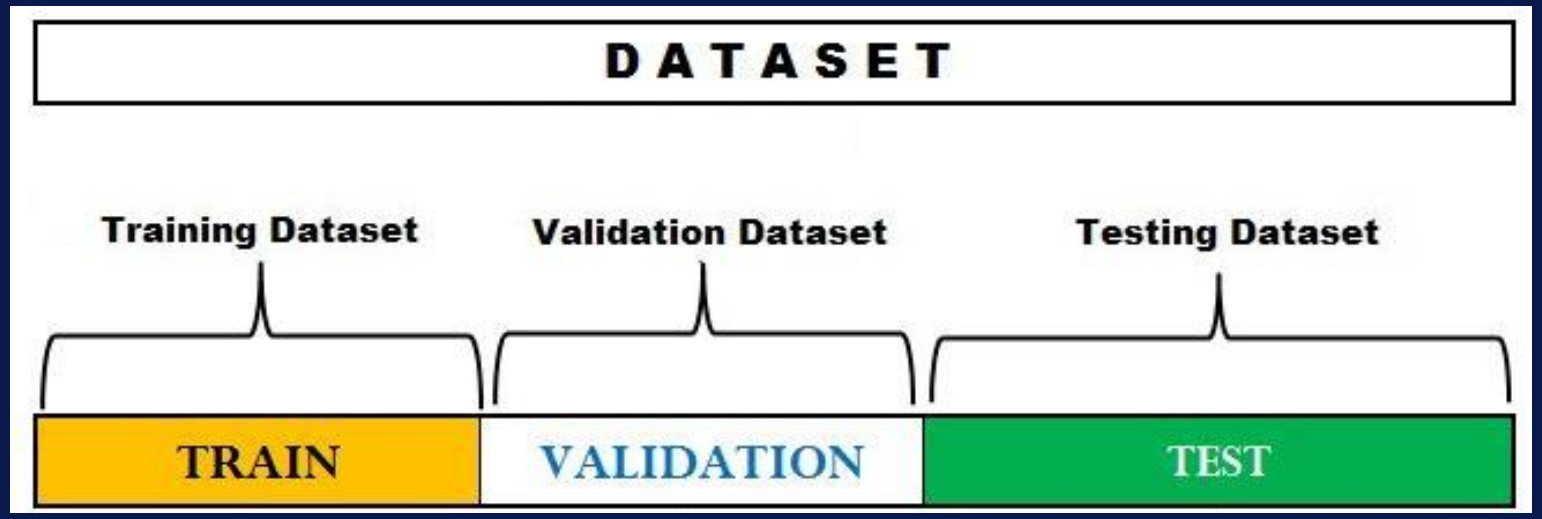
- Hold-out Cross Validation
- K-fold Cross Validation(k-겹 교차검증)
- Stratified K-fold Cross Validation(계층별 k-겹 교차검증)
- Leave-One-Out Cross Validation(LOOCV)

03. Hold-Out Cross Validation

Hold-Out Cross Validation이란?

데이터셋을 train set, validation set, test set 세 개로 나누어 train set으로 모델을 훈련시키고, validation set으로 성능을 평가한 후, test set을 이용하여 모델의 일반화 성능을 추정하는 교차 검증 방법

- 장점 : Test set을 넣기 전 이미 validation set으로 성능을 평가하기 때문에 모델의 예측 성능을 측정할 수 있다
- 한계 : validation set으로 쓸 데이터는 훈련에 쓰이지 않기 때문에 데이터 자원을 낭비한다.

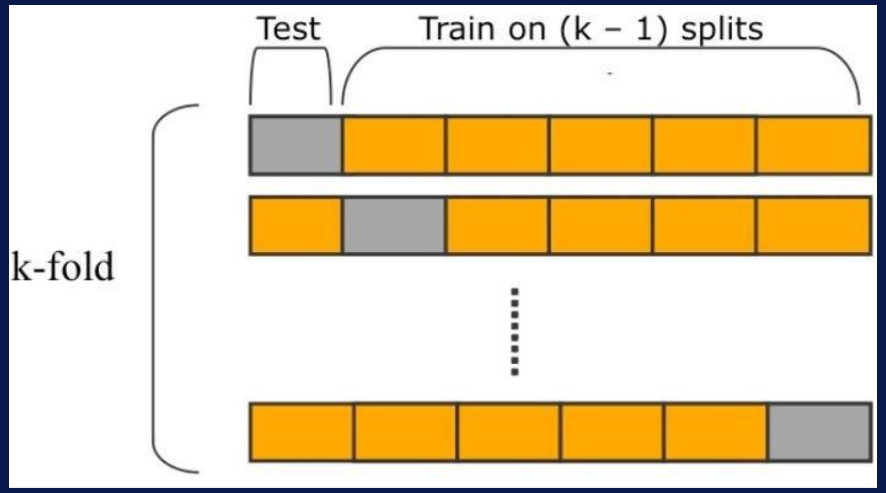


03. K-fold Cross Validation

K-fold cross validation이란?

데이터셋을 k개로 나누어 하나씩 test set으로 사용하고 나머지를 train set으로 사용하는 교차검증 방법

- 장점 : 모든 데이터셋을 train set으로도, test set으로도 활용 가능
- 단점 : 여전히 데이터가 편향되어 있을 경우, 편향된 데이터가 분할되지 못하고 몰릴 수 있음



K-fold 단계

1) 데이터 집합을 k개의 데이터 Fold로 나눈다.
2) (k-1)개의 fold는 train fold, 나머지 1개는 test fold로 지정한다.
3) Train fold를 이용하여 모델을 훈련 시키고, test fold를 이용하여 정확도를 측정한다.
4) 2-3번 과정을 k번 반복한다.
   이때, 한 번 선정했던 test fold는 다시 test fold로 선택하지 않는다.
5) 총 k개의 성능 결과가 나오고, 이 k개의 평균을 학습 모델 성능으로 본다.

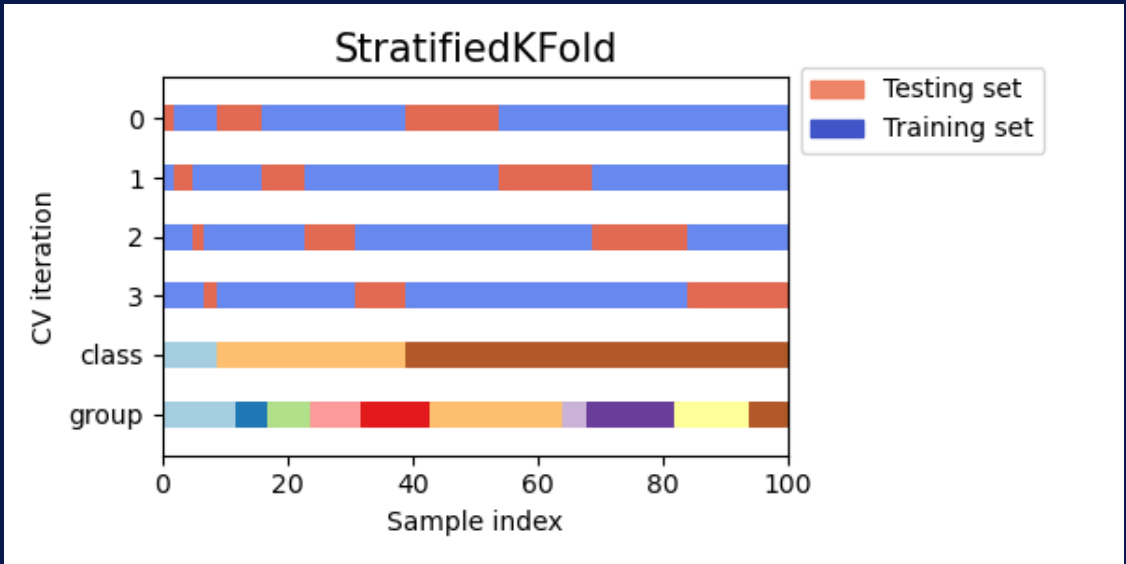
03. Stratified K-fold Cross Validation

K-fold의 경우 데이터를 일정한 간격으로 잘라 사용하기 때문에 target값이 편향되면 학습에 어려움이 생긴다.

```
학습 레이블 데이터 분포 :
 Iris-versicolor     50
Iris-virginica       50
검증 레이블 데이터 분포 :
 Iris-setosa         50
```

Stratified k-fold cross validation이란?

위와 같은 k-fold의 단점을 보완하기 위해 target 속성값의 개수를 동일하게 하여 데이터가 한 곳으로 몰리는 것을 방지하는 교차검증 방법
(※ 다만 회귀의 경우 target값이 연속적인 값이기 때문에 회귀에서는 지원되지 않는다.)

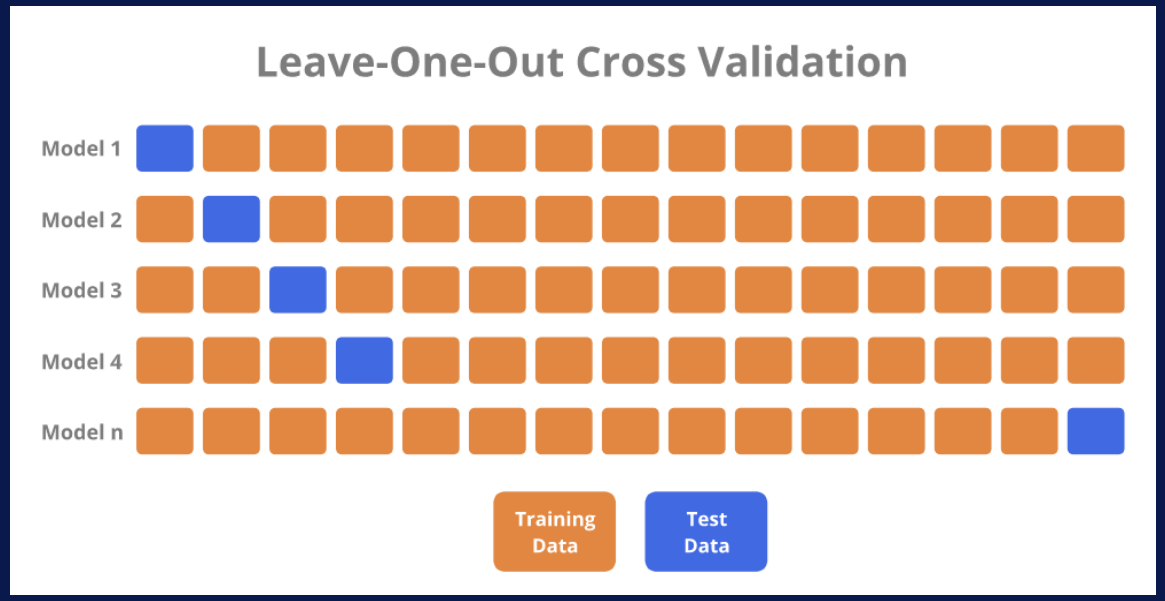


03. Leave-One-Out Cross Validation

Leave-One-Out Cross validation 이란?

교차 검증을 극단적으로 사용한 것으로 n개의 데이터 샘플에서 한 개의 데이터 샘플을 test set으로 하고,
그 1개를 뺀 나머지를 train set으로 두고 모델을 검증하는 교차 검증 방법

- 장점 : 훈련에 거의 데이터셋의 전부를 사용하기 때문에 모델 성능에 대한 신뢰를 할 수 있고 편향되지 않은 추정치를 제공한다.
- 단점 : 계산 비용이 많이 든다.



04. 회귀 평가지표

회귀 모델의 target값은 연속적인 값 ➡ 실제값과 예측값의 차이가 작을수록 해당 모델의 성능이 좋다는 것을 의미한다.

1. MSE(Mean Squared Error) : 평균 제곱 오차

- 실제값과 예측값의 차이를 제곱하여 평균한 것
- 장점 : 잔차의 값이 음수가 될 수 있는 경우를 방지하고, 오차의 민감도를 높였다.
- 단점 : 예측 변수와 단위가 다르며, 잔차를 제곱하기 때문에 이상치에 민감하다.

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

2. RMSE(Root Mean Squared Error) : 평균 제곱근 오차

- MSE에 루트를 씌운 것
- 장점 : 직관적이고 예측변수와 단위가 같으며, 제곱 값을 루트로 풀어주기 때문에 잔차를 제곱해서 생기는 값의 왜곡이 덜하다.
- 단점 : 실제 값에 대해 underestimates인지 overestimates인지 파악하기 힘들다.

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

04. 회귀 평가지표

3. MAE(Mean Absolute Error) : 평균 절대 오차

- 실제값과 예측값의 차이를 절댓값으로 변환해 평균한 것
- 장점 : 잔차의 값이 음수가 될 수 있는 경우를 방지하고, 예측변수와 단위가 같다.
- 단점 : 잔차에 절댓값을 씌우기 때문에 실제 값에 대해 underestimates인지 overestimates인지 파악하기 힘들다.

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

4. MAPE(Mean Absolute Percentage Error) : 평균 절대 비율 오차

- MAE를 비율(%)로 표현한 것
- 장점 : 직관적이고, 비율 변수이기 때문에 다른 평가지표에 비해 비교에 용이하다.
- 단점 : 실제값에 대해 underestimates인지 overestimates인지 파악하기 힘들다.

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{f}(x_i)}{y_i}\right|$$

05. 분류 평가지표

분류 모델의 target값은 종류 ➡ 실제값과 예측값이 일치하는 수가 많을수록 모델의 성능이 좋다는 것을 의미한다.

Confusion Matrix(오차행렬)

- True Positive(TP) : 실제 True인 정답을 True라고 예측 (정답)
- False Positive(FP) : 실제 False인 정답을 True라고 예측 (오답)
- False Negative(FN) : 실제 True인 정답을 False라고 예측 (오답)
- True Negative(TN) : 실제 False인 정답을 False라고 예측 (정답)

		Predicted	
		Positive	Negative
Actual	Positive	TP (True Positive)	FN (False Negative)
	Negative	FP (False Positive)	TN (True Negative)

05. 분류 평가지표

1) 정확도(accuracy)

- 전체 데이터 중, 정확하게 예측한 데이터의 비율

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

- 불균형 데이터의 경우 정확한 평가지표가 될 수 없다.
  ex) 양성과 음성의 비율이 1:9인 경우 모두 음성이라고 모델이 예측한다면 정확도는 90%가 된다.

2) 특이도 (Specificity)

- Negative로 예측한 것 중, 진짜 Negative의 비율

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

- Negative에 집중한 평가지표이다.

		Predicted	
		Positive	Negative
Actual	Positive	TP (True Positive)	FN (False Negative)
	Negative	FP (False Positive)	TN (True Negative)

05. 분류 평가지표

3) 정밀도(precision)

- Positive로 예측한 것 중, 진짜 Positive 의 비율

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

		Predicted	
		Positive	Negative
Actual	Positive	TP (True Positive)	FN (False Negative)
	Negative	FP (False Positive)	TN (True Negative)

- Positive에 집중한 평가지표이다.
- 실제 Negative 데이터를 Positive로 잘못 판단하면 업무상 큰 영향이 있는 경우에 사용한다.
  ex) 스팸메일 판정(스팸 메일로 예측한 것 중 스팸메일의 비율)

4) 재현율(recall) = 민감도(sensitivity)

- 진짜 Positive인 것들 중, 올바르게 Positive로 예측한 비율

$$Recall\ (= Sensitivity) = \frac{TP}{TP + FN}$$

		Predicted	
		Positive	Negative
Actual	Positive	TP (True Positive)	FN (False Negative)
	Negative	FP (False Positive)	TN (True Negative)

- Positive에 집중한 평가지표이다.
- 실제 Positive 데이터를 Negative로 잘못 판단하면 업무상 큰 영향이 있는 경우에 사용한다.
  ex) 암환자 판정(실제 암환자 중에 양성이라고 예측한 비율)

05. 분류 평가지표

5) F1-score

- Precision과 Recall을 이용하여 조화평균을 구한 지표

$$F1\ Score = 2 \times \frac{recall \times precision}{recall + precision}$$

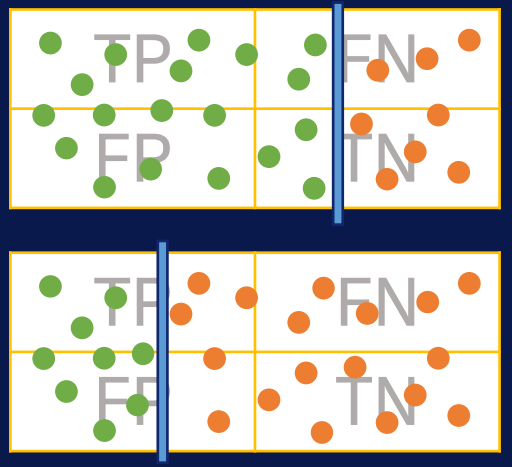
- 정밀도와 재현율이 어느 한쪽으로 치우치지 않는 수치를 나타낼 때 높은 값을 갖는다.

※ 임계값 변경에 따른 정밀도와 재현율 변화관계

- 임계값을 높일 경우 : 양성으로 예측하는 기준이 엄격해진다.
  (= 음성으로 예측되는 샘플이 많아진다.)
  ⇒ 정밀도 : 높아짐, 재현율 : 낮아짐

- 임계값을 낮출 경우 : 양성으로 예측하는 기준이 낮아진다.
  (= 양성으로 예측되는 샘플이 많아진다.)
  ⇒ 정밀도 : 낮아짐, 재현율 : 높아짐

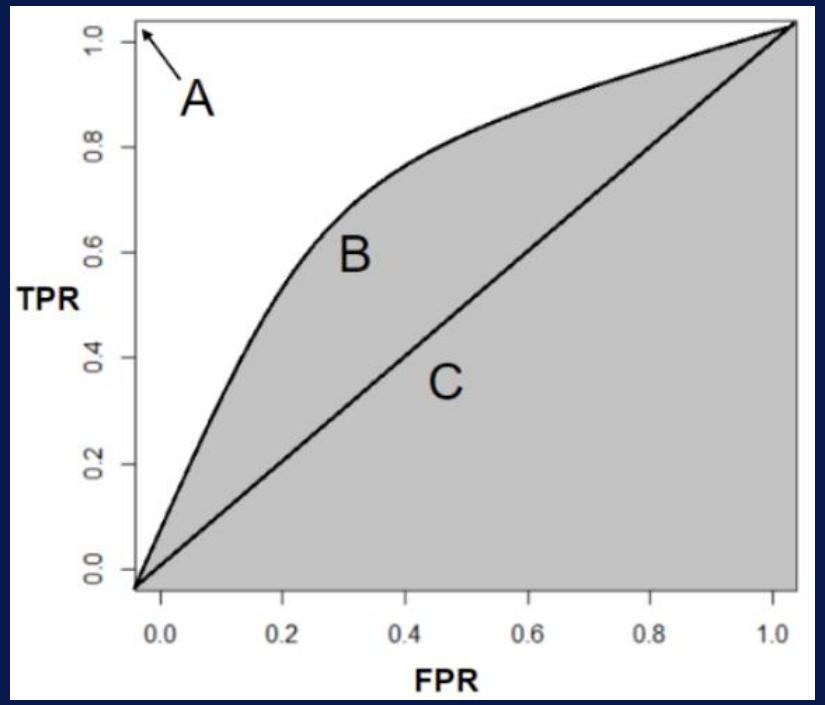
➡ 임계값을 변화시켰을 때 재현율과 정밀도는 음의 상관관계를 가진다.



05. 분류 평가지표

6) ROC / AUC

- FPR (False Positive Rate) : 1 – 특이도(specificity)
  (= 실제 음성 중 양성으로 잘못 예측한 비율)
- TPR (True Positive Rate) : 재현율(recall)
  (= 실제 양성 중 양성으로 맞게 예측한 비율)



- ROC (Receiver Operating Characteristic)
  : 모든 임계값에서 분류 모델의 성능을 보여주는 그래프
- AUC (Area Under the Curve)
  : ROC 곡선 아래의 영역

AUC가 높다는 것은 클래스를 구별하는 모델의 성능이 좋다는 것을 의미한다.
※ AUC는 0~1 사이 값을 갖는다.

A : 완벽한 모델 B : 좋은 모델 C : 평균/기본 모델 회색 : AUC

Reference

03. 교차검증

교차검증_ K-fold Cross Validation 이미지 출처
: https://www.researchgate.net/profile/B-Aksasse/publication/326866871/figure/fig2/AS:669601385947145@1536656819574/K-fold-cross-validation-In-addition-we-outline-an-overview-of-the-different-metrics-used.jpg
교차검증_ Stratified K-fold Cross Validation 이미지 출처
: https://scikit-learn.org/stable/_images/sphx_glr_plot_cv_indices_009.png
교차검증 _ LOOCV 이미지 출처
: https://miro.medium.com/max/1050/0*oHrfoOeToTpdkmHX.png
교차검증 _ Hold-Out Cross Validation 이미지 출처
: https://www.datavedas.com/wp-content/uploads/2018/04/image003.jpg

05. 분류 평가지표

분류 평가지표_ 오차행렬 이미지 출처
: http://here.deepplus.co.kr/?p=24
분류 평가지표 : ROC/AUC
: https://ysyblog.tistory.com/72

D&A

ML Session 2차시 머신러닝 기초 2

Thank You.

2022 / 09 / 13

D&A 운영진 윤경서