

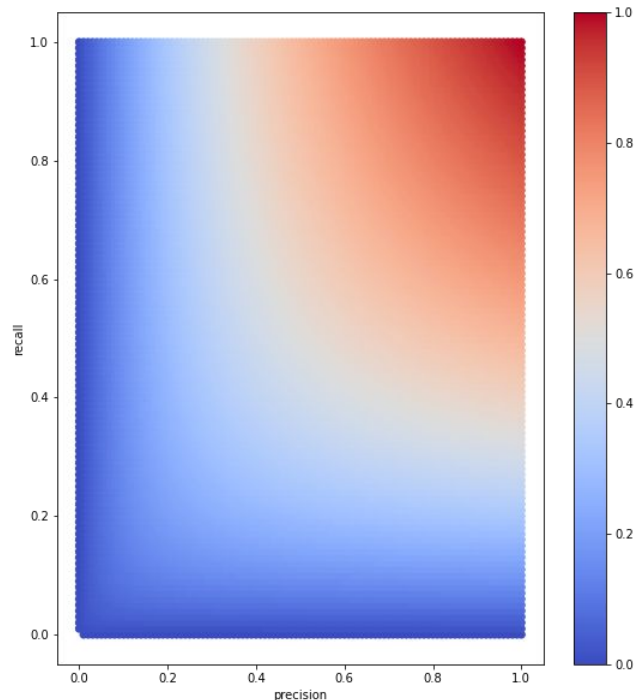
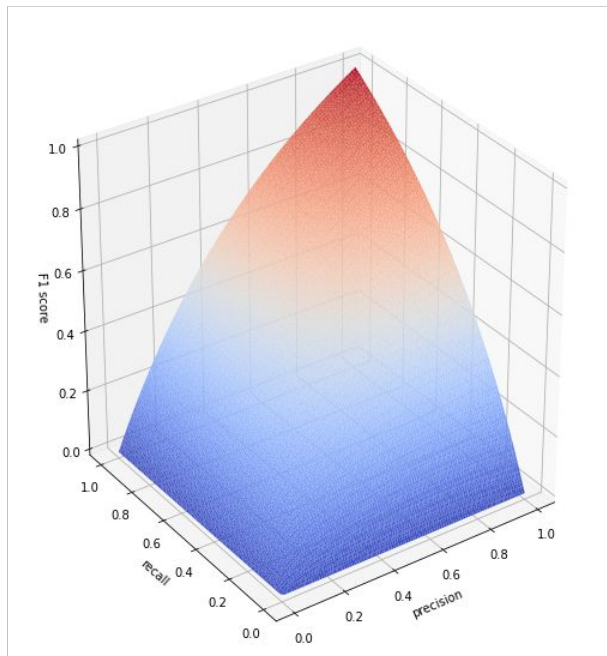
F₁ Score

정밀도와 재현율이 비슷한
분류기에서는 F₁ 점수가
높습니다.

F₁ 점수가 높아질수록
변화율이 점점 작아진다.

조화 평균은 **평균적인
변화율을 구할 때에 주로
사용된다.**

$$F = 2 \times \frac{\text{정밀도} \times \text{재현율}}{\text{정밀도} + \text{재현율}}$$



1) <https://www.mikulskibartosz.name/f1-score-explained/>

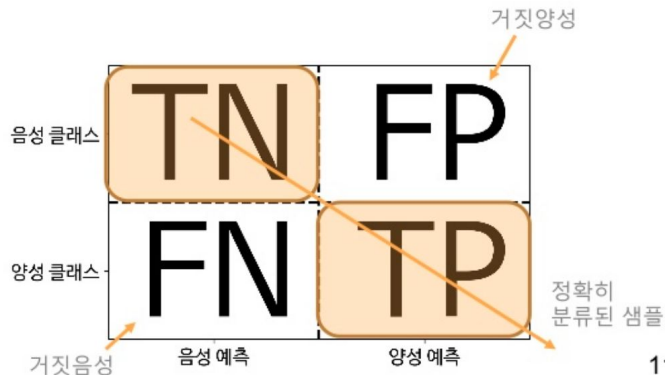
2) <https://www.slideshare.net/RickyPark3/5model-evaluation-and-improvementepoch2-2-87291677>

PR과 ROC 그리고 Accuracy에 대한 고찰

부제: PR에 대한 간략 설명

교재에서는 양성 클래스가 드물거나 거짓 음성보다 거짓 양성이 더 중요할 때 PR 곡선을 사용하고 그렇지 않으면 ROC 곡선을 사용하라고 합니다.

1. TN(모델이 음성 클래스라고 옳게 예측한 값)은 전혀 중요하지 않다.
-> 이것 중요하게 생각한다면 Accuracy
2. Threshold 값을 통과하지 못하는 데이터들은 정밀도(Precision)에 영향을 끼치지 않는다.
3. 거짓 양성이 더 중요할 때는 정밀도가 굉장히 중요함을 의미한다.
(Youtube Kids 동영상 예)



11

$$\text{정밀도} = \frac{TP}{TP + FP} \quad \text{재현율} = \frac{TP}{TP + FN}$$

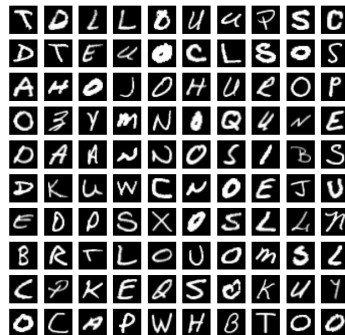
PR과 ROC 그리고 Accuracy에 대한 고찰

부제: 5 분류기의 반대는 Not 5 분류기가 아니다.



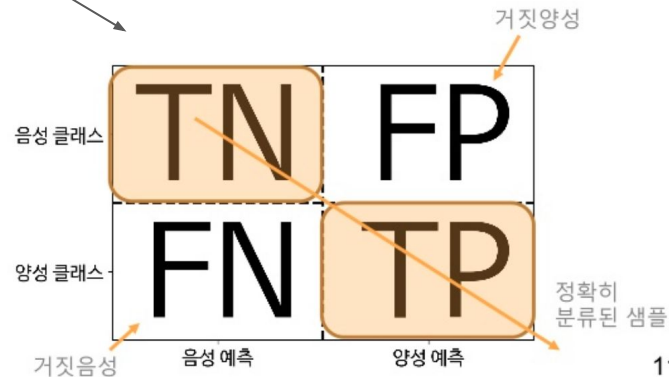
PR과 ROC 그리고 Accuracy에 대한 고찰

부제: PR을 측정 지표로 쓴다는 것은?



5 분류기

속하는 곳은 TN 아니면 FP



11

ROC나 Accuracy를 썼다면?

PR과 ROC 그리고 Accuracy에 대한 고찰

부제: ROC에 대한 간략 설명

ROC 곡선은 거짓 양성 비율에 대한 진짜 양성 비율(재현율)의 곡선이다.

TP + FN : 전체 양성 클래스의 수

TN + FP : 전체 음성 클래스의 수



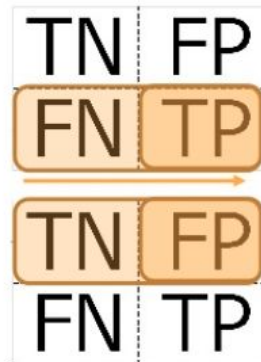
1. **FPR**이 높을수록 **TPR**도 늘어난다. (임계값이 점점 낮아진다는 것을 의미한다.)
→ 같이 증가하는 것은 아니다!
→ 같은 방향으로 다르게 걸어가는 두 사람 (독립적)

2. 임계값을 조정하면 $TP \leftrightarrow FN$, $FP \leftrightarrow TN$ 변화가 일어난다.
 $TP \rightarrow FN$: TPR 감소, $TP \leftarrow FN$: TPR 증가
 $FP \rightarrow TN$: FPR 감소, $FP \leftarrow TN$: FPR 증가

3. 클래스 분포가 **균일하지 않아도** 성능 측정에 영향을 받지 않는다.

$$\text{재현율} = \frac{TP}{TP + FN}$$

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



1) <https://www.slideshare.net/RickyPark3/5model-evaluation-and-improvementepoch2-2-87291677>

2) [https://marco0332.github.io/categories/study/2019-09-19-MACHINE_LEARNING-%EB%AA%A8%EB%8D%B8-%EC%84%B1%EB%8A%A5-%EC%B8%A1%EC%A0%95-%EC%A7%80%ED%91%9C\(metrics\)/](https://marco0332.github.io/categories/study/2019-09-19-MACHINE_LEARNING-%EB%AA%A8%EB%8D%B8-%EC%84%B1%EB%8A%A5-%EC%B8%A1%EC%A0%95-%EC%A7%80%ED%91%9C(metrics)/)

PR과 ROC 그리고 Accuracy에 대한 고찰 부제: ROC 곡선이 왜 저렇게 그려질까?

thresholds

```
array([ 49442.43765905,  49441.43765905,  36801.60697028, ...,
       -105763.22240074, -106527.45300471, -146348.56726174])
```

thresholds 값이 낮아짐에 따라서 TPR과 FPR이
증가하게 된다. (독립적으로 증가한다)

즉, 저 그래프는 함수라고 볼 수 없음.

```
from pprint import pprint
np.set_printoptions(precision=6, suppress=True)

fpr_tpr = ["%.10f %.10f" %(f, t) for f, t in zip(fpr[:10], tpr[:10])]

pprint(fpr_tpr)
```

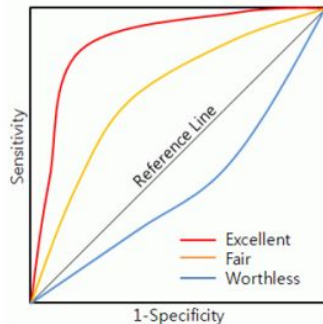
```
['0.0000000000 0.0000000000',
 '0.0000000000 0.0001844678',
 '0.0000000000 0.0009223391',
 '0.0000183221 0.0009223391',
 '0.0000183221 0.0012912747',
 '0.0000366441 0.0012912747',
 '0.0000366441 0.0023980815',
 '0.0000549662 0.0023980815',
 '0.0000549662 0.0038738240',
 '0.0000916103 0.0038738240']
```

AUC

- Area Under the Curve
- ROC curve의 밑면적을 계산한 값
- AUC=1.0 → 가장 완벽한 검사 방법
→ 민감도 및 특이도가 모두 100%

- ✓ 0.90 - 1.00 = Excellent
- ✓ 0.80 - 0.90 = Good
- ✓ 0.70 - 0.80 = Fair
- ✓ 0.60 - 0.70 = Poor
- ✓ 0.50 - 0.60 = Fail

AUC



Muller, Matthew P., et al. "Can routine laboratory tests discriminate between severe acute respiratory syndrome and other causes of community-acquired pneumonia?." *Clinical infectious diseases* 40.8 (2005): 1079-1086.

random_state 매개변수 ?

random_state : 데이터가 분할될 때 이루어지는 무작위성에도 안정적인 데이터 분할을 위한 시드값

‘ data set이 업데이트 되어도 안정적인 training/test data 분할을 위해 사용 ‘

시드값?

np.random.seed() 는 난수를 예측가능하도록 만든다.

시드값들은 고유의 유사난수 생성기를 만든다.

random_state 매개변수 ?

어떤 수를 넣어야 좋은 것일까? 책의 예제에 있는 42와 0의 의미가 있는가?

중요한 문제가 아니다 = 아무거나 상관없다 = 의미도 없다
대부분의 경우 파라미터로 어떤 값을 넣든지 별 차이가 없다.

서로 다른 시드는 서로 다른 유사난수를 생성하게 한다는 점이 중요 !

```
1 np.random.seed(0)
2 np.random.randint(99, size = 5)
```

array([44, 47, 64, 67, 67])

```
1 np.random.seed(1)
2 np.random.randint(99, size = 5)
```

array([37, 12, 72, 9, 75])

```
1 import numpy as np
2 np.random.seed(0) ; np.random.rand(4)
```

array([0.5488135 , 0.71518937, 0.60276338, 0.54488318])

```
np.random.seed(0) ; np.random.rand(4)
```


PR과 ROC 그리고 Accuracy에 대한 고찰

부제: 양성과 음성의 분포가 균일하지 않을 경우 발생하는 일 (클래스 분포량은 살짝 다름)

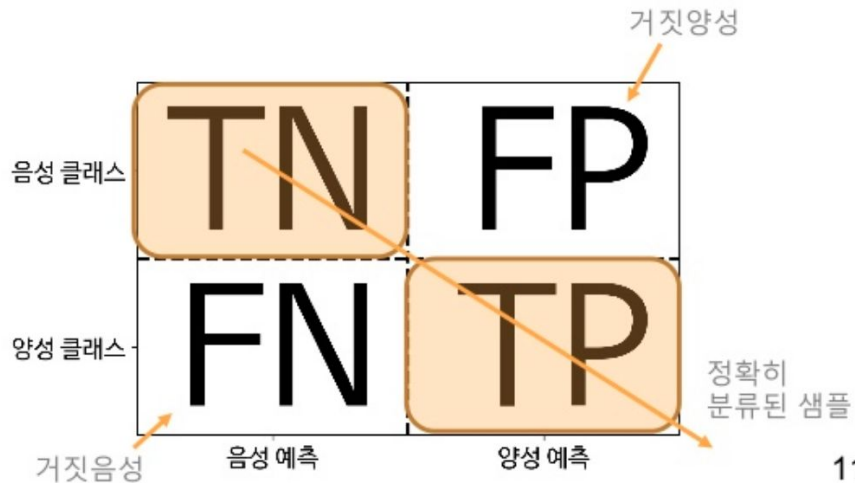
양성 샘플이 많을 경우

음성 샘플이 많을 경우



분포에 상관 없이

모델의 성능에 알맞는 값이 도출 될 것 같다.



PR과 ROC 그리고 Accuracy에 대한 고찰

부제: 양성과 음성의 분포가 균일하지 않을 경우 발생하는 일 (클래스 분포랑은 살짝 다름)

클래스 3개 [0, 1, 2]라고 할때

클래스 0이 50%, 클래스 1이 25%, 클래스 2가 25% 비율일 때

클래스 0을 분류하는 문제에서는 양성 분포 50%, 음성분포 50%가 된다

클래스 1을 분류하는 문제에서는 양성분포 25%, 음성분포 75%가 된다

양성샘플 : FN + TP, 아무리 양성 클래스가 많이 들어와도 비율로 계산된다

음성 클래스도 마찬가지

분포에 상관없이 모델의 성능에 알맞는 값이 도출될 것 같다는 생각

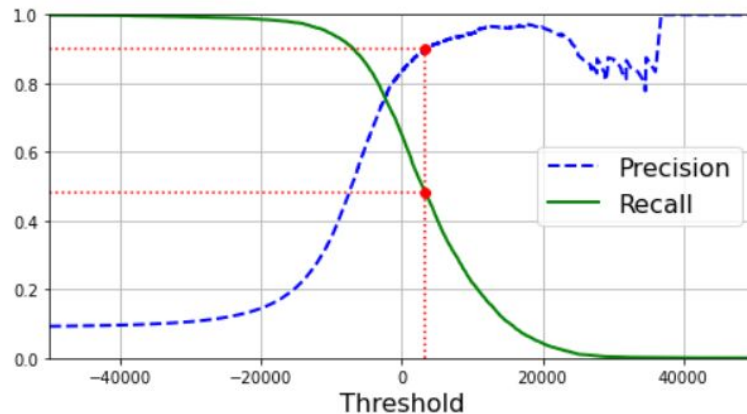
정밀도(precision)/재현율(recall) tradeoff

p.137

decision function, decision threshold 개념이 잘 이해가 되지 않습니다.

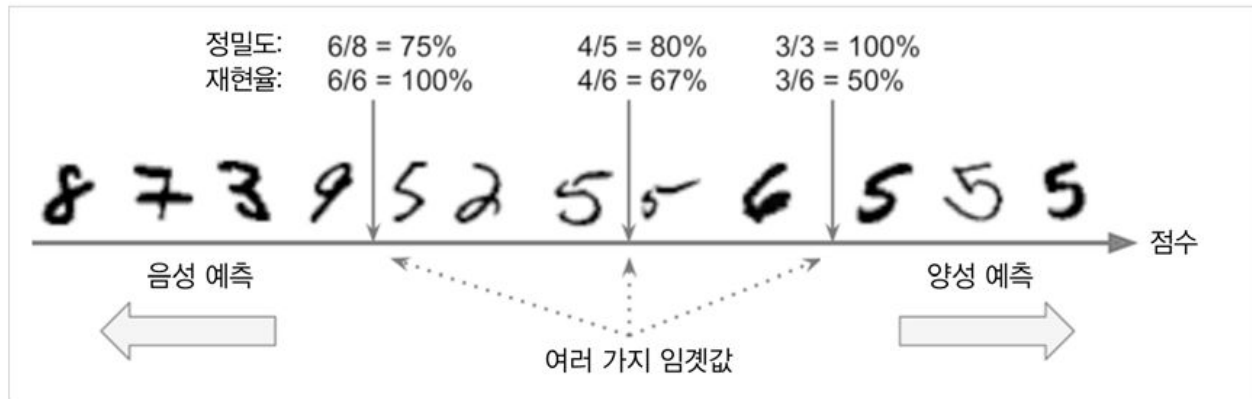
decision function은 그냥 모델에서 나온 점수이고, 이 점수로 기준을 나눠 classification 하기 위한 값이 threshold라고 생각하면 되는걸까요?

decision function으로 인해 두 그래프가 **trade-off** 관계를 나타내는데, 왜 이렇게 그려지는지도 잘 이해가 되지 않습니다 πππ



정밀도(precision)/재현율(recall) tradeoff

decision function은 그냥 모델에서 나온 점수이고, 이 점수로 기준을 나눠 classification 하기 위한 값이 threshold
Tradeoff 관계는 아래 그림의 precision, recall을 직접 계산해보면 이해가 더 잘 될 것



SGDClassifier가 선형모델?

p.151

SGD는 mini-batch만큼의 데이터를 이용해서 gradient를 update하는 방법으로 알고 있는데,
선형 모델이라는 설명이 맞는지 궁금합니다.

SGDClassifier가 선형모델?

SGD는 stochastic gradient descent의약자로, gradient descent 방법이다.
SGDClassifier라고 나타낸 것은, sklearn에서 SGDClassifier가
sklearn.linear_model.SGDClassifier로 linear model에 포함되기 때문에
linear model + SGD 방법을 합쳐서 SGDClassifier라고 나타낸 것 같다.

분류기의 성능 측정 지표로 언제 정확도(Accuracy)를 써야 할까?

굳이 정확도를 쓸 필요는 없지만 만약에 **정확도**를 사용할 수 있는 케이스가 있다면 아래 나열된 조건들

- 이진 분류기일 경우 (2개의 클래스가 존재)
- 클래스의 비율이 균일한 경우
- 문제의 정의에서 label 값이 서로 상반되는 경우 (기냐, 안기냐가 확실히 다른 경우)

예를 들어서 '5 아님' 분류기에서 {5, 5, 5, 6, 6, 6} 의 데이터인 경우 정확도 사용 가능

{5, 5, 5, 2, 1, 6} 이런 경우는 서로 상반되는 경우는 아님. (사용은 가능하지만 **안기냐의 정확도**가 의미가 없음.)

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

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

1) <https://flonelin.wordpress.com/2017/03/29/novelty%EC%99%80-outlier-detection/>

2) https://scikit-learn.org/stable/modules/outlier_detection.html

분류기의 성능 측정 지표로 언제 정확도(Accuracy)를 써야 할까?

굳이 정확도를 쓸 필요는 없지만 만약에 **정확도**를 사용할 수 있는 케이스가 있다면 아래 나열된 조건들을 만족하는 상황일 것 같습니다.

- 이진 분류기일 경우 (2개의 클래스가 존재)
- 클래스의 비율이 균일한 경우
- 문제의 정의에서 label 값이 서로 상반되는 경우 (**기냐**, **안기냐**가 확실히 다른 경우)

예를 들어서 '5 아님' 분류기에서 {5, 5, 5, 6, 6, 6} 의 데이터인 경우 정확도 사용 가능

{5, 5, 5, 2, 1, 6} 이런 경우는 서로 상반되는 경우는 아님. (사용은 가능하지만 **안기냐의 정확도**가 의미가 없음.)

위의 세 가지 조건을 모두 만족하는 경우 accuracy를 사용할 수 있을 것 같다.

책에서는 분류기를 못통과한 것들까지 포함해 accuracy를 재는 것 같다.

다중 레이블 분류에서 클래스 별로 가중치를 다르게 주는 기준이 왜 크기인가?

Page 153

특히 앨리스 사진이 밥이나 찰리 사진보다 훨씬 많다면 앨리스 사진에 대한 분류기의 점수에 더 높은 가중치를 둘 것입니다.

간단한 방법은 레이블의 클래스의 지지도(즉, 타깃 레이블에 속한 샘플 수)를 가중치로 주는 것입니다.

Q. 앨리스 분류에 특화됐으니깐 가중치를 많이 주는 건가요?

Q. 분류기의 F1 점수가 의미가 있을려면 테스트셋의 클래스 분포가 훈련 분포와 동일해야겠죠?

온라인(점진적) 학습에는 못 쓰는 평가점수라고 생각하면 될까요?

1) <https://flonelin.wordpress.com/2017/03/29/novelty%EC%99%80-outlier-detection/>

2) https://scikit-learn.org/stable/modules/outlier_detection.html

다중 레이블 분류에서 클래스 별로 가중치를 다르게 주는 기준이 왜 크기인가?

Page 153

특히 앨리스 사진이 밥이나 찰리 사진보다 훨씬 많다면 앨리스 사진에 대한 분류기의 점수에 더 높은 가중치를 둘 것입니다.

간단한 방법은 레이블의 클래스의 지지도(즉, 타깃 레이블에 속한 샘플 수)를 가중치로 주는 것입니다.

Q. 앨리스 분류에 특화됐으니깐 가중치를 많이 주는 건가요?

앨리스 비중이 많기 때문에, 앨리스를 많이 맞추는 것에 가중치를 줄 수 있을 것 같다.

Training set 비율에 따라 가중치를 줄 수 있을 것 같다.

Q. 분류기의 F1 점수가 의미가 있을려면 테스트셋의 클래스 분포가 훈련 분포와 동일해야겠죠?

온라인(점진적) 학습에는 못 쓰는 평가점수라고 생각하면 될까요?

여기에 대한 논의는 별로 되지 않았습니다.

p. 133

불균형한 데이터셋을 다룰 때(즉, 어느 클래스가 다른 것보다 월등히 많은 경우)에 더욱 정확도를 분류기의 성능 측정 지표로 사용하지 않는 것을 선호한다.

정확도를 분류기의 성능 측정 지표로 사용하지 않는 것을 선호하는 것은 이해가 가는데, 불균형한 데이터셋에서 특히 선호하지 않는 이유가 궁금합니다.

p. 133

불균형한 데이터셋을 다룰 때(즉, 어느 클래스가 다른 것보다 월등히 많은 경우)에 더욱 정확도를 분류기의 성능 측정 지표로 사용하지 않는 것을 선호한다.

정확도를 분류기의 성능 측정 지표로 사용하지 않는 것을 선호하는 것은 이해가 가는데, 불균형한 데이터셋에서 특히 선호하지 않는 이유가 궁금합니다.

불균형한 데이터셋과 정확도의 관계가 무엇일까에 대한 질문
앞에서 정확도에 대한 논의를 했으므로 그 부분을 참고하기로 함

F score의 일반화된 조화 평균 식은 어디에 쓰일까요?

$$F_{\beta} = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}.$$

F score의 일반화된 조화 평균 식은 어디에 쓰일까요?

다들 이유를 잘 몰랐습니다.

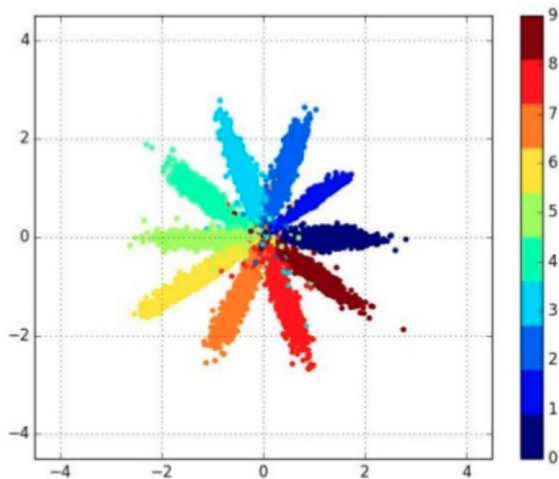
조화평균이 어떤 의미인지에 대해서만 이해하고 넘어갔습니다.

직접 찾아서 설명 부분에 올리는 것을 추천.

이 데이터셋은 균형있는 데이터셋일까요? 정확도를 사용 할 수 있을까요?

클래스 분포가 균일한 데이터셋에서는 'Accuracy'를 사용 할 수 있다고 합니다.

그렇다면 이 데이터셋은 분포가 균일한 데이터셋으로 봐도 될까요?




각 데이터셋에 대해서 어떤 분류 측정 지표를 사용 할지 토의해보세요.

운동 동작 분류 AI 경진대회
월간 데이터 11 | 헬스 데이터 | Logloss | 분류

🏆 상금 : 100만원
📅 2021.01.11 ~ 2021.02.22 17:59 [+ Google Calendar](#)
👤 579팀 📅 D-21

참여



[주제 및 배경]

- 운동 동작 인식 알고리즘 개발
- 스마트 헬스케어 산업에 적용 가능한 데이터 분석 방법

"데이터, 문화가 된다 : League1"
AI야, 진짜 뉴스를 찾아줘!
금융 | NH투자증권 | 텍스트 분류 | Accuracy + Time | 중복 참가 불가, 대학 과제에 반영 가능

🏆 상금 : 총 5,000만원(League1,2 통합)
📅 2020.11.23 ~ 2020.12.31 17:59 [+ Google Calendar](#)
👤 544팀 📅 마감

마감



딥페이크 변조 영상 탐지 AI 경진대회
서울대 | 영상 | GAN | 분류 | Accuracy

🏆 상금 : 1,000만원
📅 2020.10.19 ~ 2020.11.19 17:59 [+ Google Calendar](#)
👤 331팀 📅 마감

마감



소셜 작가 분류 AI 경진대회
월간 데이터 9 | 소셜 문제 | NLP | Logloss

🏆 상금 : 100만원+애플워치
📅 2020.10.29 ~ 2020.12.04 17:59 [+ Google Calendar](#)
👤 738팀 📅 마감

2. 배경

- a. 작가의 글을 분석하여 특징 도출
- b. 취향 추천 시스템 활용 / 대필, 유사작 탐지

3. 대회 설명

- 소셜 속 문장문치 분석을 통한 저자 예측

1) <https://dacon.io/main>

2) <https://seoyounggh.github.io/machine-learning/ml-logloss/>

분류기 마다의 장점을 파악해서 그 장점이 해당하는 영역을 나눠서 학습할 수 있을까요?

예를들어 특정 A 분류기가 MNIST 자료에서 3을 잘 분류하고 .. B라는 분류기는 다중분류기로 전체적으로 어느정도(90%) 분류를 할 수 있다면

B(다중분류기)로 분류된 데이터를 제외하고서 남은 데이터의 대해서 A 분류기를 사용하거나

추가적으로 소수의 데이터를 레이블링 처리 하면 좋을까요?

분류기 마다의 장점을 파악해서 그 장점이 해당하는 영역을 나눠서 학습할 수 있을까요?

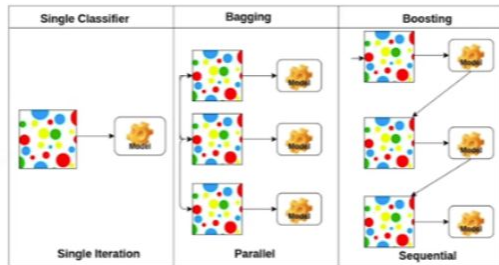
Q. 예를들어 특정 A 분류기가 MNIST 자료에서 3을 잘 분류하고 .. B라는 분류기는 다중분류기로 전체적으로 어느정도(90%) 분류를 할 수 있다면

B(다중분류기)로 분류된 데이터를 제외하고서 남은 데이터의 대해서 A 분류기를 사용하거나

추가적으로 소수의 데이터를 레이블링 처리 하면 좋을까요?

A. 수업에서 배운 Boosting 방법!

Bagging vs. Boosting



마음껏 의견 질러주세요! 컨텐츠는 제가 만들겠습니다...! :')

2:2 팀으로 진행할까 하는데 어때요?

Mnist 데이터셋으로 분류기를 만들어 테스트 세트에서 97%

정확도를 달성해보세요. (예제 1번)

(4:00 ~ 4:40) : 40분간 코드 작성

(4:40 ~ 5:00) : 20분간 설명

Mission 1

- PipeLine 구축

Mission 2

- ROC, PR 곡선 그리기

Mission 3

- 오차 행렬 시각화 하기

Mission 4

- AUC 넓이 구하기

4:00

5:00

6:00

7:00

타이타닉 데이터셋에 도전해보기! (예제 2번)

(5:10 ~ 6:40) : 1시간 30분간 코드 작성

(6:40 ~ 7:00) : 20분간 설명

사용 모델

RandomForest, LinearRegression, DecisionTree, SGD, SVM

Mission 1

- Age, Cabin Null 값 채우든가 삭제하는 추정기/변환기 만들기

Mission 2

- Feature 합쳐서 새로운 특성 만드는 변환기 (한개 이상) 만들기

Mission 3

- Train, Validation 데이터 나눠서 훈련하기