모형의 성능 평가방법

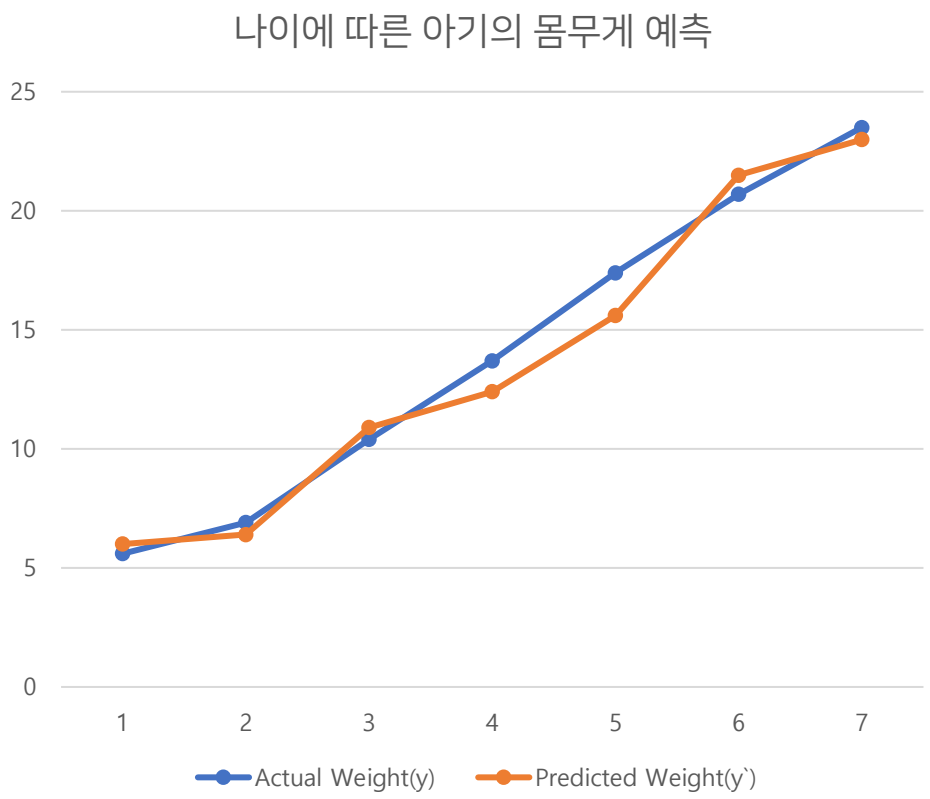
1. 회귀문제 인 경우 MAE, MAPE, MSE, RMSE 등등. 계산 가능

2. 분류문제 인 경우 Precision, Recall 등등

3. ROC, AUC 등등

Ex) 나이에 따른 아기의 몸무게 예측

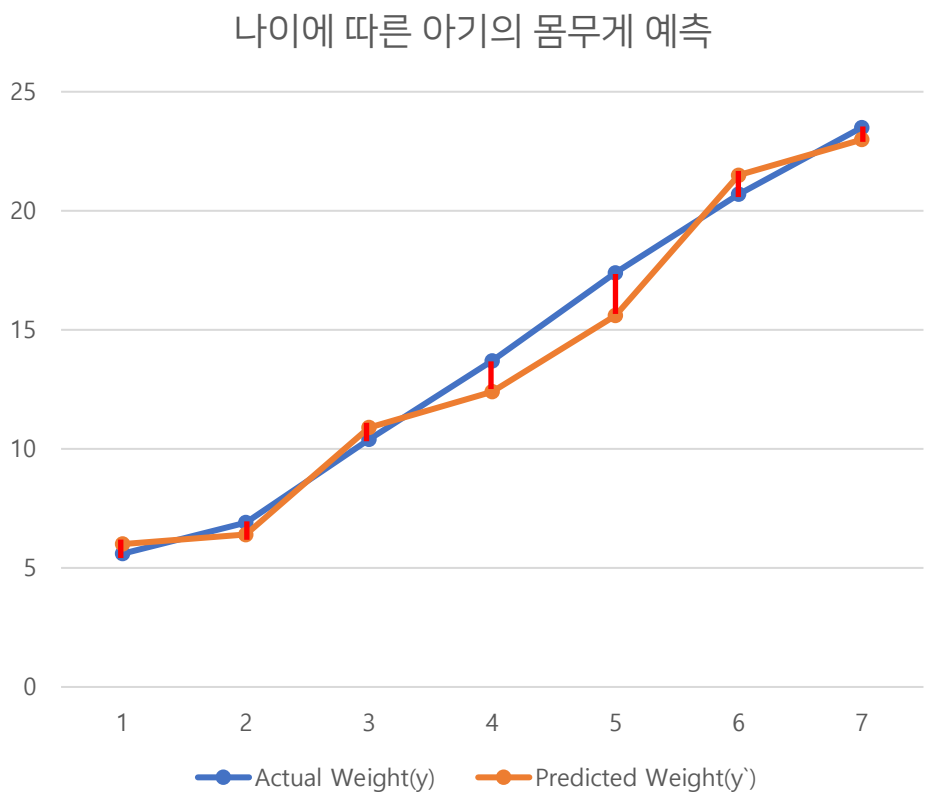
Age	Predicted Weight($\hat{y}$)	Actual Weight(y)
1	6.0	5.6
2	6.4	6.9
3	10.9	10.4
4	12.4	13.7
5	15.6	17.4
6	21.5	20.7
7	23.0	23.5



성능평가: 눈으로 보는 것 이외에도 적절한 기준(지표)이 필요

->잔차($\hat{y} - y$) 활용

Age	Predicted Weight($\hat{y}$)	Actual Weight(y)
1	6.0	5.6
2	6.4	6.9
3	10.9	10.4
4	12.4	13.7
5	15.6	17.4
6	21.5	20.7
7	23.0	23.5



성능평가: 눈으로 보는 것 이외에도 적절한 기준(지표)이 필요

->잔차($\hat{y} - y$) 활용

Age	Predicted Weight($\hat{y}$)	Actual Weight(y)	Residual ($\hat{y} - y$)
1	6.0	5.6	0.4
2	6.4	6.9	-0.5
3	10.9	10.4	0.5
4	12.4	13.7	-1.3
5	15.6	17.4	-1.8
6	21.5	20.7	0.8
7	23.0	23.5	-0.5

성능지표 1: 평균오차(Average error)

- 실제 값에 비해 과대/과소 추정 여부를 판단
- 부호로 인해 잘못된 결론을 내릴 위험이 있음

Average error

$= \frac{1}{n}\sum_{i=1}^{n}(\hat{y} - y)$

= -0.342

Age	Predicted Weight($\hat{y}$)	Actual Weight(y)	Residual ($\hat{y} - y$)
1	6.0	5.6	0.4
2	6.4	6.9	-0.5
3	10.9	10.4	0.5
4	12.4	13.7	-1.3
5	15.6	17.4	-1.8
6	21.5	20.7	0.8
7	23.0	23.5	-0.5

성능지표2: 평균 절대 오차(Mean absolute error ; MAE)

- 실제 값과 예측 값 사이의 절대적인 오차의 평균을 이용
- 단위(Scale)의 영향

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

$$= 0.829$$

Age	Predicted Weight($\hat{y}$)	Actual Weight(y)	Residual ($\hat{y} - y$)
1	6.0	5.6	0.4
2	6.4	6.9	-0.5
3	10.9	10.4	0.5
4	12.4	13.7	-1.3
5	15.6	17.4	-1.8
6	21.5	20.7	0.8
7	23.0	23.5	-0.5

성능지표3: 평균 절대 비율오차(Mean absolute percentage error ; MAPE)

- 실제 값 대비 얼마나 예측값이 차이가 있는지를 %로 표현

  상대적인 오차를 추정

MAPE

$=100\% \times \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y}-y|}{|y|}$

= 6.43%

Age	Predicted Weight($\hat{y}$)	Actual Weight(y)	Residual ($\hat{y}-y$)
1	6.0	5.6	0.4
2	6.4	6.9	-0.5
3	10.9	10.4	0.5
4	12.4	13.7	-1.3
5	15.6	17.4	-1.8
6	21.5	20.7	0.8
7	23.0	23.5	-0.5

성능지표4&5 : (Root) Mean squared error : MSE & RMSE

- 부호의 영향을 제거하기 위해 절대값이 아닌 제곱을 취한 지표

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

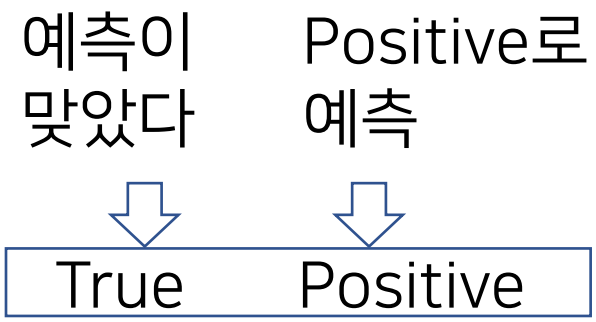
$$= 0.926$$

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

$$= 0.962$$

Age	Predicted Weight($\hat{y}$)	Actual Weight(y)	Residual ($\hat{y}-y$)
1	6.0	5.6	0.4
2	6.4	6.9	-0.5
3	10.9	10.4	0.5
4	12.4	13.7	-1.3
5	15.6	17.4	-1.8
6	21.5	20.7	0.8
7	23.0	23.5	-0.5

		실제 정답	
		Positive	Negative
실험 결과	Positive	True Positive	False Positive (Type I error)
	Negative	False Negative (Type II error)	True Negative



true/false & positive/negative

true/false는 예측한 결과가 맞는지/틀린지를 의미

positive/negative는 분류기가 예측(분류)한 범주를 의미

		실제 정답	
		Positive	Negative
실험 결과	Positive	True Positive	False Positive (Type I error)
	Negative	False Negative (Type II error)	True Negative

true/false & positive/negative

true/false는 예측한 결과가 맞는지/틀린지를 의미

positive/negative는 분류기가 예측(분류)한 범주를 의미

예) positive로 예측을 했는데 실제 negative인 경우(틀린 경우)

: ~~~ positive -> 예측이 틀렸으니까 False Positive(FP)

		실제 정답	
		Positive	Negative
실험 결과	Positive	True Positive	False Positive (Type I error)
	Negative	False Negative (Type II error)	True Negative

true/false & positive/negative

true/false는 예측한 결과가 맞는지/틀린지를 의미

positive/negative는 분류기가 예측(분류)한 범주를 의미

예) negative로 예측을 했는데 실제 negative인 경우(맞은 경우)

: ~~~ negative -> 예측이 맞았으니까 True Negative(TN)

		실제 정답	
		Positive	Negative
실험 결과	Positive	True Positive	False Positive (Type I error)
	Negative	False Negative (Type II error)	True Negative

다시 한 번 강조하지만 Positive/Negetive는 예측한 결과이다.

앞에서 살펴본 False Positive, True Negative 모두

'실제 정답'은 Negative이다.

처음에는 헷갈릴 수 있으니 그때마다 자료를 찾아보자.

혼동 행렬(Confusion Matrix)

		실제 정답	
		Positive	Negative
실험 결과	Positive	True Positive	False Positive (Type I error)
	Negative	False Negative (Type II error)	True Negative

		질병여부	
		유	무
검사결과	양성(질병)	a	b
	음성(정상)	c	d

$$\text{정밀도(Precision)} = \frac{True\ Positive}{(predicted\ \ positive)} = \frac{TP}{TP+FP} = \frac{a}{a+b}$$

예측이 positive인 것의 개수 중 : a+b
실제로 positive 한 것의 개수 : a

: 질병이 있다고 예측한 것 중 실제 질병이 있는 경우

혼동 행렬(Confusion Matrix)

		실제 정답	
		Positive	Negative
실험 결과	Positive	True Positive	False Positive (Type I error)
	Negative	False Negative (Type II error)	True Negative

		질병여부	
		유	무
검사결과	양성(질병)	a	b
	음성(정상)	c	d

$$\text{재현율(Recall)} = \frac{True\ Positive}{(actually\ positive)} = \frac{TP}{TP+FN} = \frac{a}{a+c}$$

= True Positive Rate = 민감도(Sensitivity)

실제로 positive인 것의 개수 중 : a+c
positive로 예측한 것의 개수 : a

:실제 질병이 있는 경우 실시한 검사에서 질병이 있다고 판정할 수 있는 능력

혼동 행렬(Confusion Matrix)

		실제 정답	
		Positive	Negative
실험 결과	Positive	True Positive	False Positive (Type I error)
	Negative	False Negative (Type II error)	True Negative

		질병여부	
		유	무
검사결과	양성(질병)	a	b
	음성(정상)	c	d

$$\text{특이도(Specificity)}=\frac{True\ Negative}{(actually\ negetive)}=\frac{TN}{FP+TN}=\frac{d}{b+d}$$

= True Negative Rate

실제로 negative인 것의 개수 중 : b+d
negative로 예측한 것의 개수 : d

:실제 질병이 없는 경우 실시한 검사에서 질병이 없다고 판정할 수 있는 능력

혼동 행렬(Confusion Matrix)

		실제 정답	
		Positive	Negative
실험 결과	Positive	True Positive	False Positive (Type I error)
	Negative	False Negative (Type II error)	True Negative

		질병여부	
		유	무
검사결과	양성(질병)	a	b
	음성(정상)	c	d

$$\text{특이도(Specificity)} = \frac{True\ Negative}{(actually\ negetive)} = \frac{TN}{FP+TN} = \frac{d}{b+d}$$

= True Negative Rate

$$1\text{-(Specificity)} = \frac{FP}{(negetive)} = \frac{FP}{FP+TN} = \frac{b}{b+d}$$ = False Positive Rate

혼동 행렬(Confusion Matrix)

		실제 정답	
		Positive	Negative
실험 결과	Positive	True Positive	False Positive (Type I error)
	Negative	False Negative (Type II error)	True Negative

		질병여부	
		유	무
검사결과	양성(질병)	a	b
	음성(정상)	c	d

재현율, 민감도, TPR

(Sensitivity, **True Positive Rate**, Recall, Hit Rate) = a/(a+c)

제2종 오류(Type II error, False Negative Rate) = c/(a+c)

혼동 행렬(Confusion Matrix)

		실제 정답	
		Positive	Negative
실험 결과	Positive	True Positive	False Positive (Type I error)
	Negative	False Negative (Type II error)	True Negative

		질병여부	
		유	무
검사결과	양성(질병)	a	b
	음성(정상)	c	d

민감도(Sensitivity, **True Positive Rate**, Recall, Hit Rate) = a/(a+c)

제2종 오류(Type II error, False Negative Rate) = c/(a+c)

특이도(Specificity, True Negative Rate) = d/(b+d)

제1종 오류(Type I error, **False Positive Rate**) = b/(b+d)

혼동 행렬(Confusion Matrix)

		실제 정답	
		Positive	Negative
실험 결과	Positive	True Positive	False Positive (Type I error)
	Negative	False Negative (Type II error)	True Negative

		질병여부	
		유	무
검사결과	양성(질병)	a	b
	음성(정상)	c	d

오분류율(Misclassification error) = (b+c)/(a+b+c+d)

정분류율(**Accuracy** = 1 – misclassification error) = (a+d)/(a+b+c+d)

이 외에도 성능을 평가하는 여러 가지 기준들이 있다.

여러 가지 기준이 필요한 이유

(경우 1) 실제 positive인 경우를 민감하게 positive라고 예측하면 좋은가?

여러 가지 기준이 필요한 이유

(경우 1) 실제 positive인 경우를 민감하게 positive라고 예측하면 좋은가?

-> True Positive Rate(Sensitivity, Recall)가 높을수록 좋은 모형인가?

여러 가지 기준이 필요한 이유

(경우 1) 실제 positive인 경우를 민감하게 positive라고 예측하면 좋은가?

-> True Positive Rate(Sensitivity, Recall)가 높을수록 좋은 모형인가?

-> No! (반드시 그렇지는 않다.)

여러 가지 기준이 필요한 이유

(경우 1) 실제 positive인 경우를 민감하게 positive라고 예측하면 좋은가?

-> True Positive Rate(Sensitivity, Recall)가 높을수록 좋은 모형인가?

-> No! (반드시 그렇지는 않다.)

(반례 1) 주어진 모든 환자들을 암으로 예측하는 엉터리 분류기가 있는 경우

여러 가지 기준이 필요한 이유

(경우 1) 실제 positive인 경우를 민감하게 positive라고 예측하면 좋은가?

-> True Positive Rate(Sensitivity, Recall)가 높을수록 좋은 모형인가?

-> No! (반드시 그렇지는 않다.)

(반례 1) 주어진 모든 환자들을 암으로 예측하는 엉터리 분류기가 있는 경우

-> 모든 환자를 암환자로 분류하기 때문에, 모든 암환자를 100% 잡아낸다.

여러 가지 기준이 필요한 이유

(경우 1) 실제 positive인 경우를 민감하게 positive라고 예측하면 좋은가?

-> True Positive Rate(Sensitivity, Recall)가 높을수록 좋은 모형인가?

-> No! (반드시 그렇지는 않다.)

(반례 1) 주어진 모든 환자들을 암으로 예측하는 엉터리 분류기가 있는 경우

-> 모든 환자를 암환자로 분류하기 때문에, 모든 암환자를 100% 잡아낸다.

-> 단점: 암이 아닌 환자들도 모두 암으로 잘못 예측(False Positive)

여러 가지 기준이 필요한 이유

(반례 1) 주어진 모든 환자들을 암으로 예측하는 엉터리 분류기가 있는 경우

이 경우 True Positive Rate = 1 , False Positive Rate = 1

		질병여부	
		유	무
검사결과	양성(질병)	a	b
	음성(정상)	0	0

$$TPR = \frac{a}{a+0} = 1$$

$$FPR = \frac{b}{b+0} = 1$$

여러 가지 기준이 필요한 이유

비슷하게 주어진 모든 환자들이 암환자가 아니라고 분류하는 경우,
True Positive Rate = 0 , False Positive Rate = 0

		질병여부	
		유	무
검사결과	양성(질병)	0	0
	음성(정상)	c	d

$$TPR = \frac{0}{0 + c} = 0$$

$$FPR = \frac{0}{0 + d} = 0$$

여러 가지 기준이 필요한 이유

비슷하게 주어진 모든 환자들이 암환자가 아니라고 분류하는 경우,
True Positive Rate = 0 , False Positive Rate = 0

바람직한 분류기는 어떤 분류기일까?

여러 가지 기준이 필요한 이유

비슷하게 주어진 모든 환자들이 암환자가 아니라고 분류하는 경우,
True Positive Rate = 0 , False Positive Rate = 0

바람직한 분류기는 어떤 분류기일까?

-> 실제 질병이 있으면 질병이 있다고 예측
실제 질병이 없으면 질병이 없다고 예측

여러 가지 기준이 필요한 이유

비슷하게 주어진 모든 환자들이 암환자가 아니라고 분류하는 경우,
True Positive Rate = 0 , False Positive Rate = 0

바람직한 분류기는 어떤 분류기일까?

-> 실제 질병이 있으면 질병이 있다고 예측
실제 질병이 없으면 질병이 없다고 예측

-> True Positive Rate가 1에 가깝고(= Sensitivity가 1에 가까운),
False Positive Rate가 0에 가까운(= Specificity가 1에 가까운) 분류기

여러 가지 기준이 필요한 이유

바람직한 분류기는 어떤 분류기일까?

-> (True Positive Rate) ≈ 1 , (False Positive Rate) ≈ 0

여러 가지 기준이 필요한 이유

바람직한 분류기는 어떤 분류기일까?

-> (True Positive Rate) ≈ 1 , (False Positive Rate) ≈ 0

현실적으로 둘 다 만족하기 힘들 수 있다.

-> 극단적이지 않게 '적당히' Positive, Negative를 예측하고 기준에 따라 판단

여러 가지 기준이 필요한 이유

바람직한 분류기는 어떤 분류기일까?

-> (True Positive Rate) ≈ 1 , (False Positive Rate) ≈ 0

현실적으로 둘 다 만족하기 힘들 수 있다.

-> 극단적이지 않게 적당히 Positive, Negative를 예측하고 기준에 따라 판단

-> 기준 1) 문제 상황에 맞게 우선순위를 두기

기준 2) 오분류율 등을 이용

기준 3) AUC(AUROC) 이용

기준 1) 문제 상황에 맞게 우선순위를 두기

예 1) 암과 같은 중요도가 높은 질병의 경우 True Positive Rate가 중요

일단 병이 있다고(양성,positive) 진단하고 추후에 정밀진단을 하면 됨

- 민감도(Sensitivity) = True Positive Rate

진짜 암 환자 중 검사로 얼마나 암 환자를 잘 골라내는가?

- 1-특이도(1-Specificity) = False Positive Rate

진짜 정상인 사람 중 검사로 얼마나 암 환자로 오진했는가?

기준 1) 문제 상황에 맞게 우선순위를 두기

예 2) 살인 사건의 용의자에게 선고를 하는 경우 False Positive Rate가 중요

무죄인 사람을 처벌하는 경우 평생 고통받을 것이고, 돌이킬 수 없음

- 민감도(Sensitivity) = True Positive Rate

진짜 범인 중 검사로 얼마나 범인을 잘 골라내는가?

- 1-특이도(1-Specificity) = False Positive Rate

진짜 무죄인 사람 중 검사로 얼마나 범인으로 오진했는가?

기준 2) 오분류율, F1-measure 등을 이용

		실제 정답	
		Positive	Negative
실험 결과	Positive	True Positive	False Positive (Type I error)
	Negative	False Negative (Type II error)	True Negative

		질병여부	
		유	무
검사결과	양성(질병)	a	b
	음성(정상)	c	d

- 오분류율(Misclassification error) = (b+c)/(a+b+c+d)
- 정분류율(Accuracy = 1 – misclassification error) = (a+d)/(a+b+c+d)

- 위 지표는 True Positive, False Positive를 동시에 고려

기준 3) ROC, AUC(AUROC) 이용

- 일반적인 분류기는 특정 범주에 속할 확률(probability)이나 가능도를 생성

- 동일한 확률값 하에서도 Cut-off를 어떻게 설정하는지에 따라 분류 성능이 크게 좌우되는 상황이 발생할 수 있다.

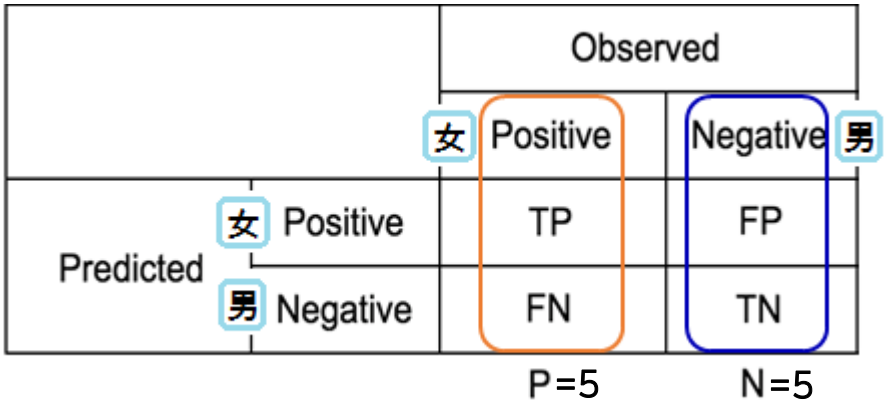
- receiver operating characteristic(ROC) curve 등이 성능 평가 기준으로 사용된다.

  (뒤 예시를 통해 확인)

예시: 성별 분류

체지방률이 θ 보다 크면 여성(positive)으로, 작으면 남성(negative)으로 분류

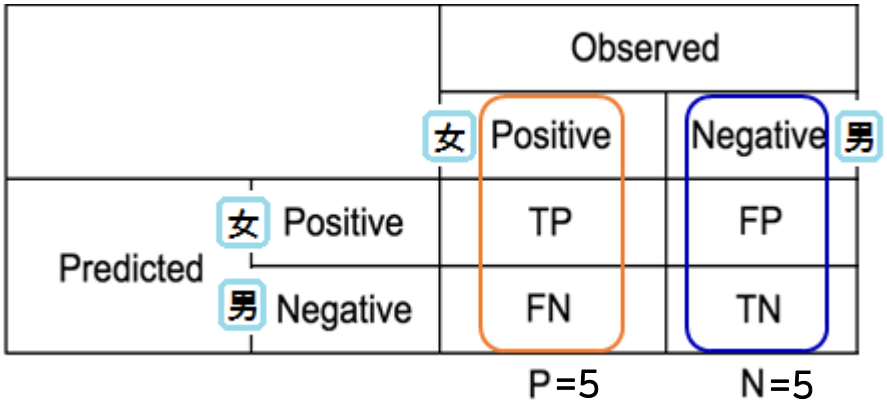
No.	체지방률	성별
1	28.6	여자
2	25.4	남자
3	24.2	여자
4	23.6	여자
5	22.7	여자
6	21.5	남자
7	19.9	여자
8	15.7	남자
9	10.0	남자
10	8.9	남자



예시: 성별 분류

체지방률이 θ 보다 크면 여성(positive)으로, 작으면 남성(negative)으로 분류

No.	체지방률	성별
1	28.6	여자
2	25.4	남자
3	24.2	여자
4	23.6	여자
5	22.7	여자
6	21.5	남자
7	19.9	여자
8	15.7	남자
9	10.0	남자
10	8.9	남자



Misclassification error: (FN+FP)/(전체)

(Accuracy: (TP+TN)/(전체))

True Positive Rate = TP/(TP+FN)

False Positive Rate = FP/(FP+TN)

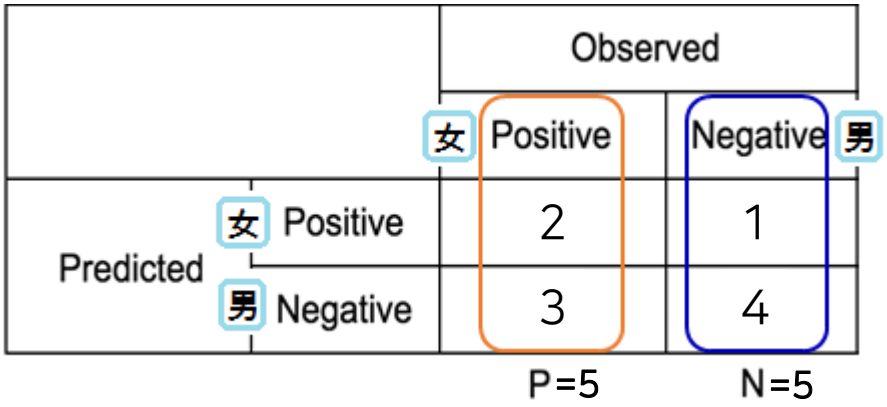
분류 알고리즘의 Cut-off 설정

다양한 Cut-off에 따른 분류 성능 비교

No.	체지방률	성별
1	28.6	여자
2	25.4	남자
3	24.2	여자
4	23.6	여자
5	22.7	여자
6	21.5	남자
7	19.9	여자
8	15.7	남자
9	10.0	남자
10	8.9	남자

女 (2)

男 (8)



If θ = 24,

Misclassification error: 4/10=0.4

(Accuracy: 0.6)

True Positive Rate = 2/(2+3)=2/5=0.4

False Positive Rate = 1/(1+4)=4/5=0.2

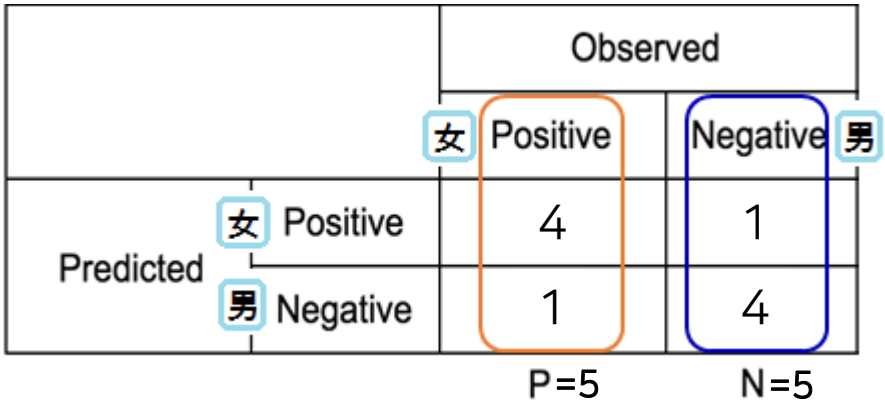
분류 알고리즘의 Cut-off 설정

다양한 Cut-off에 따른 분류 성능 비교

No.	체지방률	성별
1	28.6	여자
2	25.4	남자
3	24.2	여자
4	23.6	여자
5	22.7	여자
6	21.5	남자
7	19.9	여자
8	15.7	남자
9	10.0	남자
10	8.9	남자

女 (2)

男 (8)



If θ = 22,

Misclassification error: 2/10=0.2

(Accuracy: 0.8)

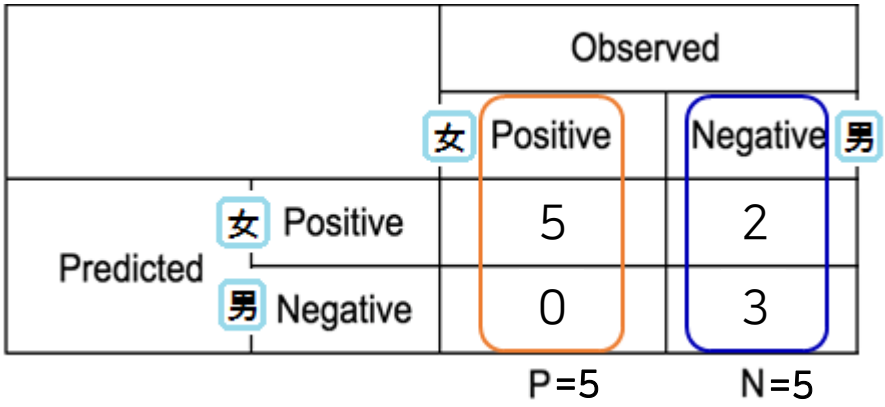
True Positive Rate = 4/(4+1)=4/5=0.8

False Positive Rate = 1/(1+4)=4/5=0.2

분류 알고리즘의 Cut-off 설정

다양한 Cut-off에 따른 분류 성능 비교

	No.	체지방률	성별
女	1	28.6	여자
	2	25.4	남자
	3	24.2	여자
	4	23.6	여자
	5	22.7	여자
	6	21.5	남자
	7	19.9	여자
男	8	15.7	남자
	9	10.0	남자
	10	8.9	남자



If θ = 18,

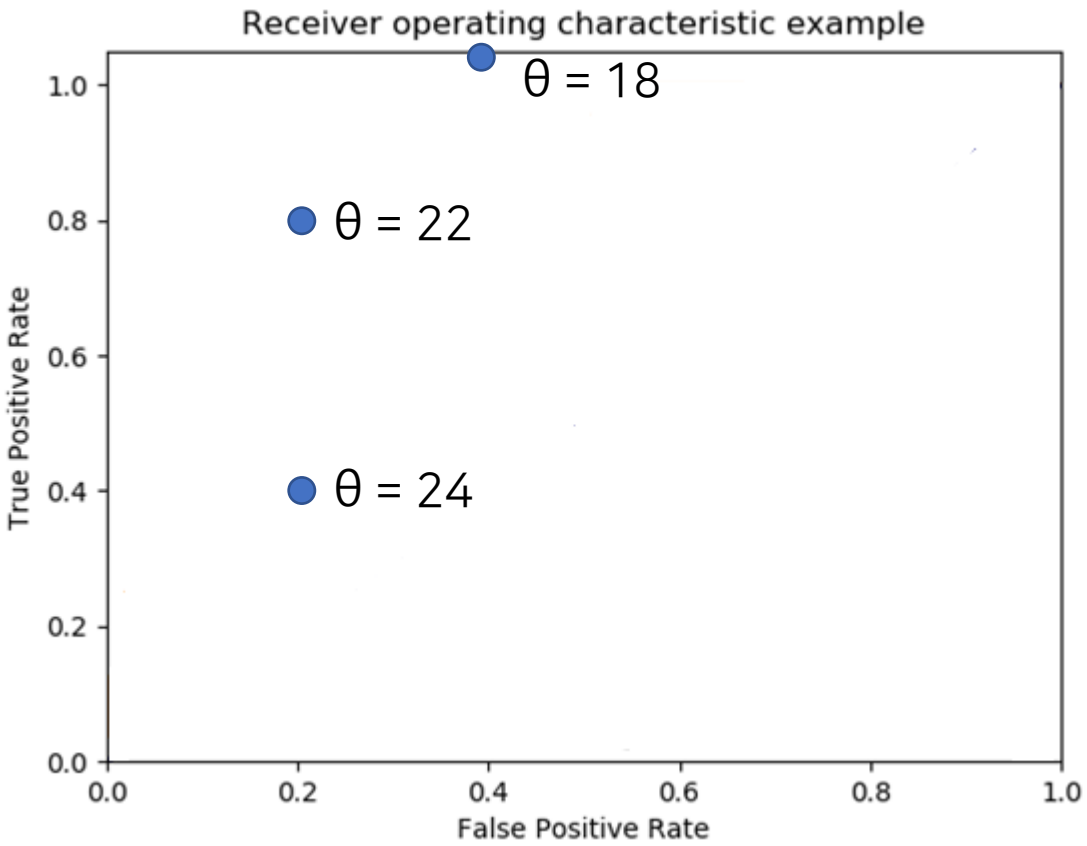
Misclassification error: 2/10=0.2

(Accuracy: 0.8)

True Positive Rate = 5/(5+0)=5/5=1

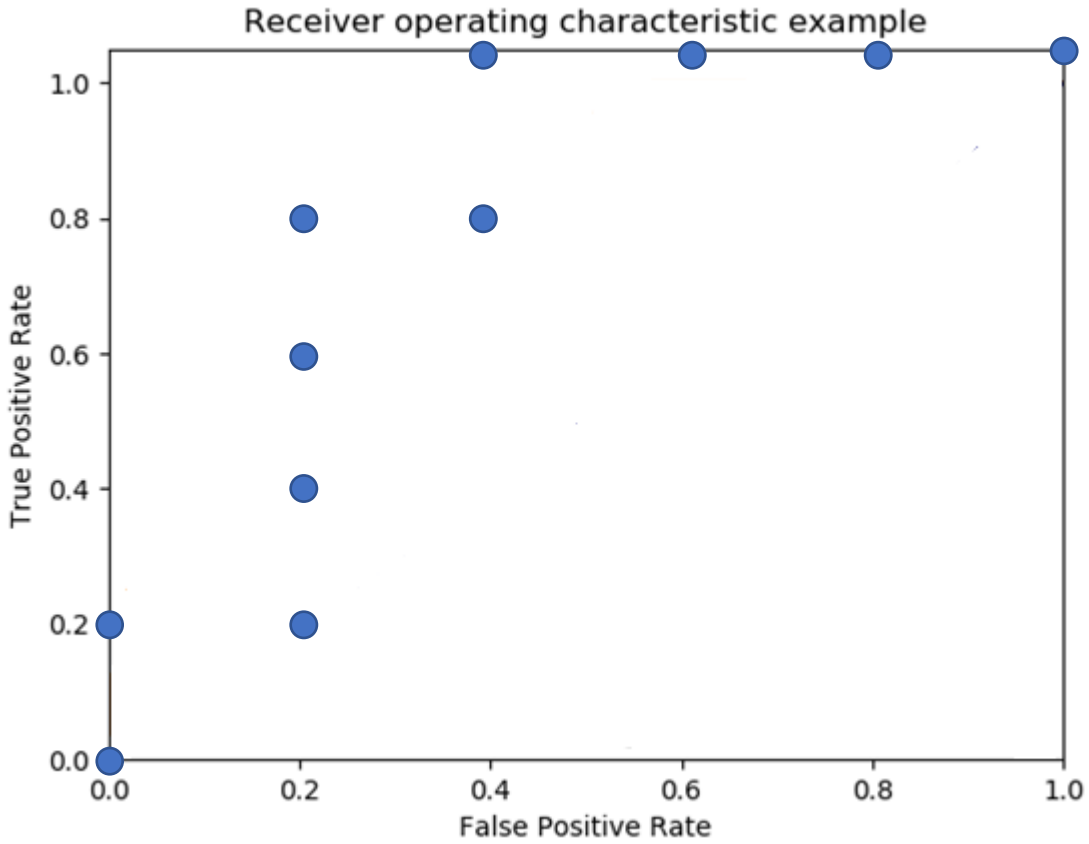
False Positive Rate = 2/(2+3)=2/5=0.4

ROC Curve



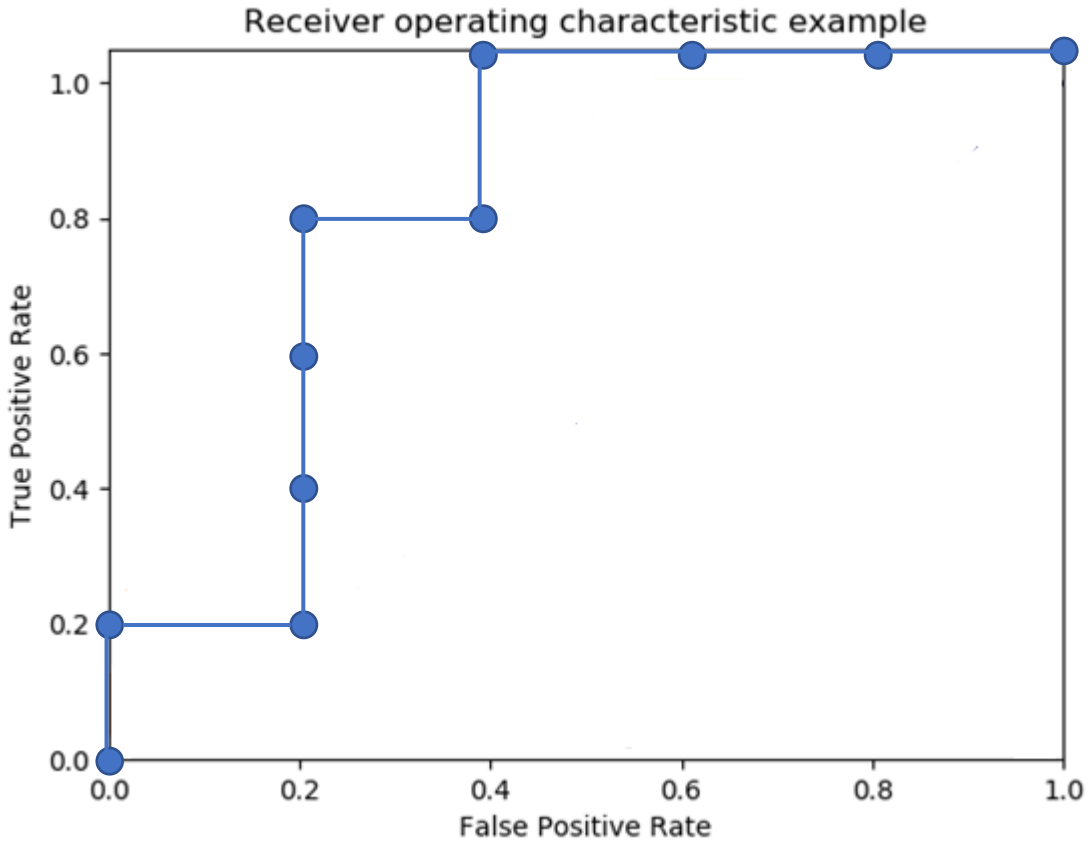
- X축이 False Positive Rate, Y축이 True Positive Rate가 되는 2차원 그래프

ROC Curve



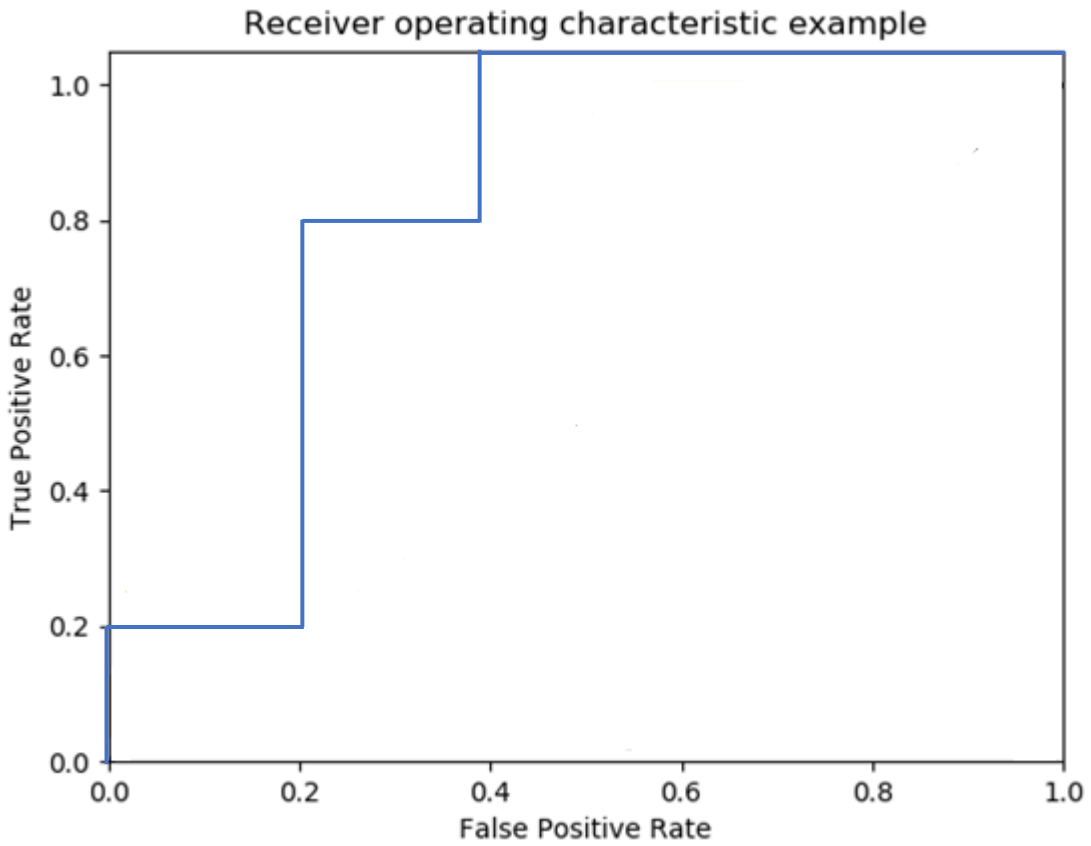
- X축이 False Positive Rate, Y축이 True Positive Rate가 되는 2차원 그래프

ROC Curve



- X축이 False Positive Rate, Y축이 True Positive Rate가 되는 2차원 그래프

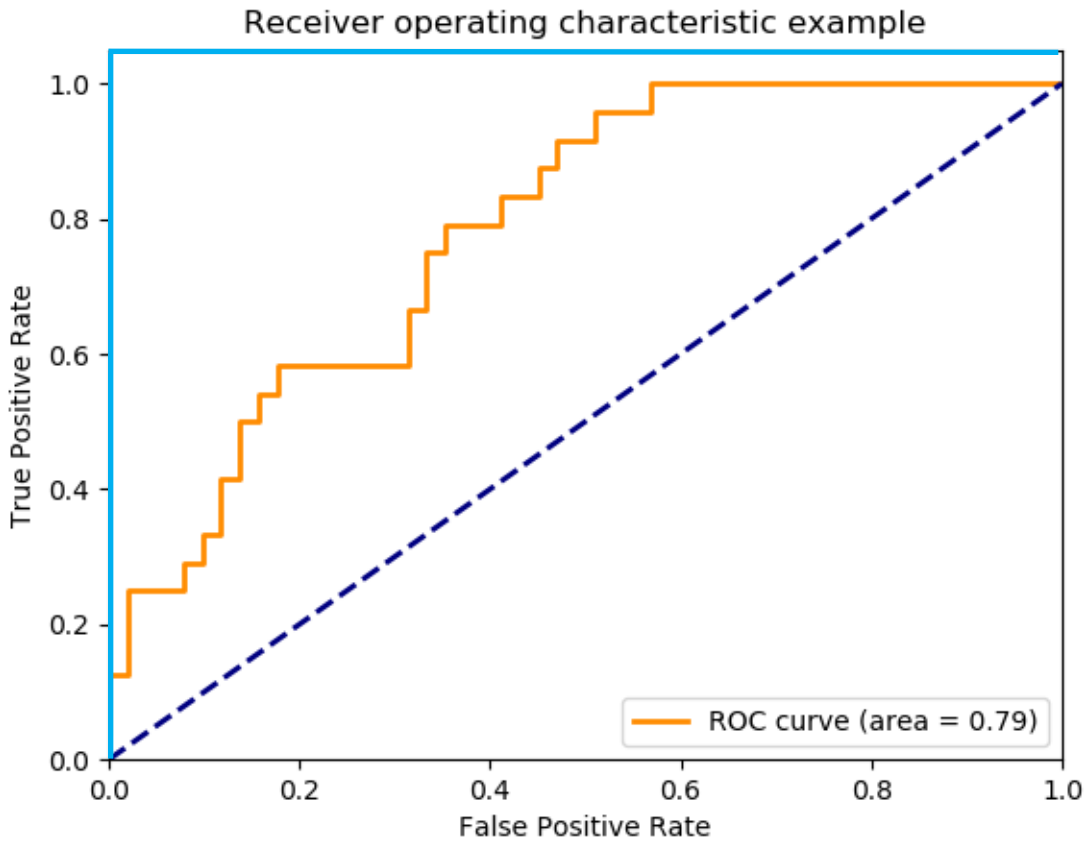
ROC Curve



- X축이 False Positive Rate, Y축이 True Positive Rate가 되는 2차원 그래프

ROC Curve

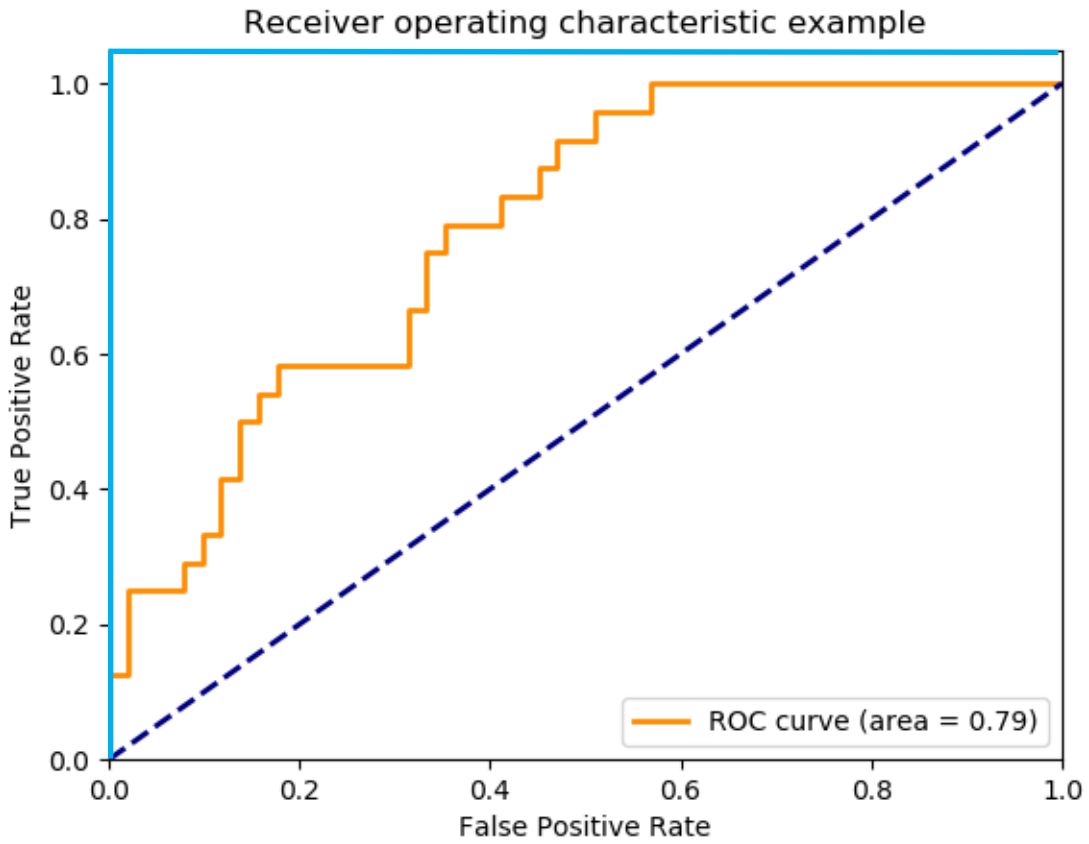
이상적 분류기의 ROC Curve
일반적인 알고리즘의 ROC Curve
random 분류기의 ROC Curve



- 이상적 분류기: True Positive Rate ≈ 1 , False Positive Rate ≈ 0

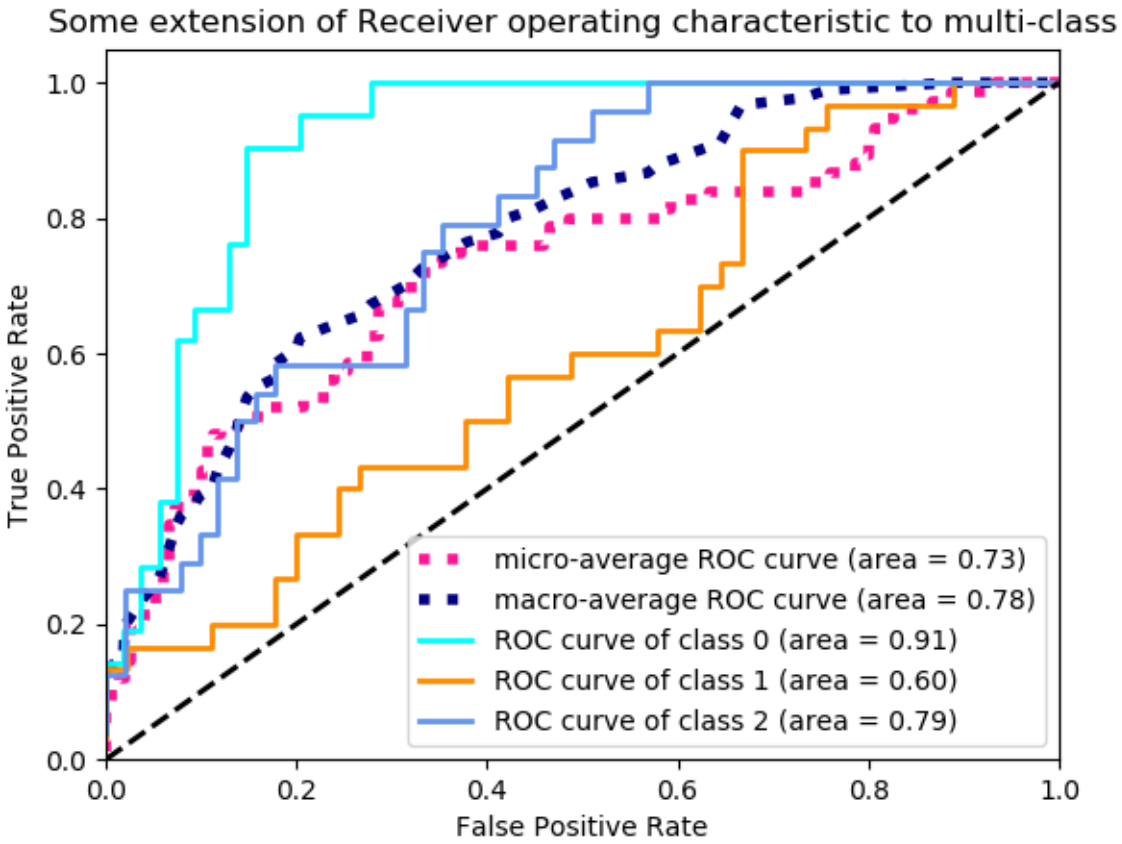
ROC Curve

이상적 분류기의 ROC Curve
일반적인 알고리즘의 ROC Curve
random 분류기의 ROC Curve



- 랜덤 분류기: 실제 positive, negative 비율만 판단 근거로 해서 임의로(randomly) positive, negative 판단

Area Under ROC curve (AUC, AUROC)



- ROC curve 아래의 면적, **cut-off value에 무관**하게 구할 수 있는 값
- 이상적 분류기는 1, Random 분류기는 0.5의 값을 가짐

로지스틱 회귀분석에서의 ROC Curve

로지스틱 회귀분석을 비롯한 많은 분류분석기는 $\hat{y}_i = 0,1$을 직접 산출하지 않는다. 대신에 [0,1] 구간 사이의 값을 가진 예측 확률값 $\hat{p}_i$을 계산한다.

이처럼 예측이 성공확률 값으로 주어졌을 때 분계점(threshold, cutoff)을 바꿔감에 따라 다른 최종 예측값을 얻게 된다. 즉, 다음처럼 계산한다.

$$\hat{y}_i = 1, \text{ if } \hat{p}_i > \text{분계점}$$

$$\hat{y}_i = 0, \text{ if } \hat{p}_i <= \text{분계점}$$

ROC Curve는 이처럼 분계점을 바꿔하면서 TPR과 FPR을 그린 곡선이다.