BE530 – Medical Deep Learning

- Performance Metrics -

Byoung-Dai Lee

Division of AI Computer Science and Engineering

Kyonggi University

KYONGGI UNIVERSITY

MAE vs. RMSE

- Regression 모델 평가
  - MAE (Mean Absolute Error)

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

y_j – the predicted value

$\hat{y}_j$ – ground truth

  - RMSE (Root Mean Squared Error)

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

    - RMSE는 결과적으로 오차가 클수록 더 큰 penalty를 부과함

분류 모델 평가

- **정확도(Accuracy)로만 평가할 경우**

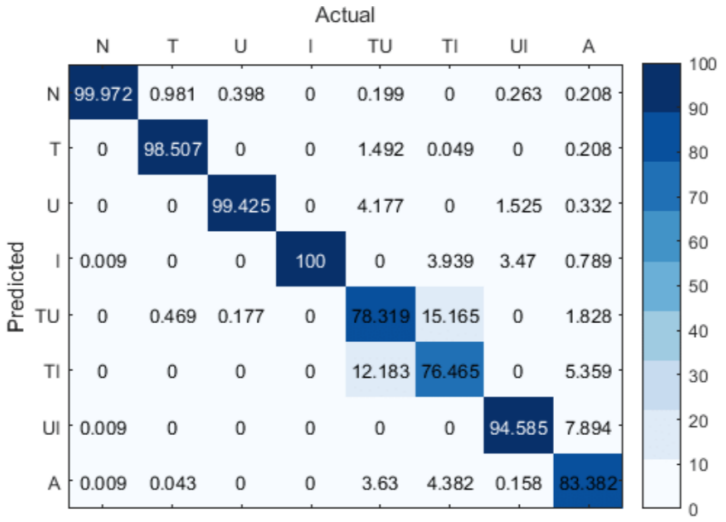
$$\text{정확도}(Accuracy) = \frac{\text{정답과 일치한 수}}{\text{전체 데이터 수}}$$

  - 스팸 메일과 정상 메일을 분류하는 이진 분류 작업
    - 수신된 메일 100건을 사람이 직접 분류해보니 스팸이 60개, 정상 메일이 40개
    - 분류기가 모든 메일을 스팸으로 분류한다면 ➔ **정확도 60%**
  - 분류 대상 클래스의 분포에 영향을 받기 때문에 단순히 정확도만 이용한 모델 평가에는 한계가 존재함
- **평가 지표**
  - 정확도, 정밀도, 재현율, F-점수, ROC, AUROC, ...

혼동행렬 (Confusion Matrix)

- 예측 결과와 실제 결과를 행렬로 표현하는 기법

		Prediction	
		1	0
Actual Class	1	True Positive	False Negative
	0	False Positive	True Negative



- True Positive – Positive로 예측해서 True임 (예측이 맞음)
- False Positive – Positive로 예측해서 False임 (예측이 틀림, 실제값은 Negative)
- False Negative – Negative로 예측해서 False임 (예측이 틀림, 실제값은 Positive)
- True Negative – Negative로 예측해서 True임 (예측이 맞음)

정확도 (Accuracy)

		Prediction	
		1	0
Actual Class	1	True Positive	False Negative
	0	False Positive	True Negative

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

$$\text{오류율}(Error\ Rate) = 1 - Accuracy$$

정밀도 (Precision)

- Positive라고 예측한 것 중에서 얼마나 잘 맞았는지 비율

Precision

		Prediction	
		1	0
Actual Class	1	True Positive	False Negative
	0	False Positive	True Negative

$$\text{정밀도}(Precision) = \frac{TP}{TP + FP}$$

재현율 (Recall)

- 실제 Positive한 것 중에서 얼마나 잘 예측하였는지 비율
  - 재현율 = 민감도 (sensitivity) = True Positive Rate

Recall

		Prediction	
		1	0
Actual Class	1	True Positive	False Negative
	0	False Positive	True Negative

$$재현율(Recall) = \frac{TP}{TP + FN}$$

- Precision과 Recall은 trade-off 관계
  - Recall을 높이기 위해서는 FN를 줄여야 하며, Precision을 높이려면 FP를 줄여야 하기 때문

특이도 (Specificity)

- 실제 Negative한 것 중에서 얼마나 잘 예측하였는지 비율

Specificity

		Prediction	
		1	0
Actual Class	1	True Positive	False Negative
	0	False Positive	True Negative

$$\text{특이도}(Specificity) = \frac{TN}{TN + FP}$$

음성예측도 (Negative Predicted Value)

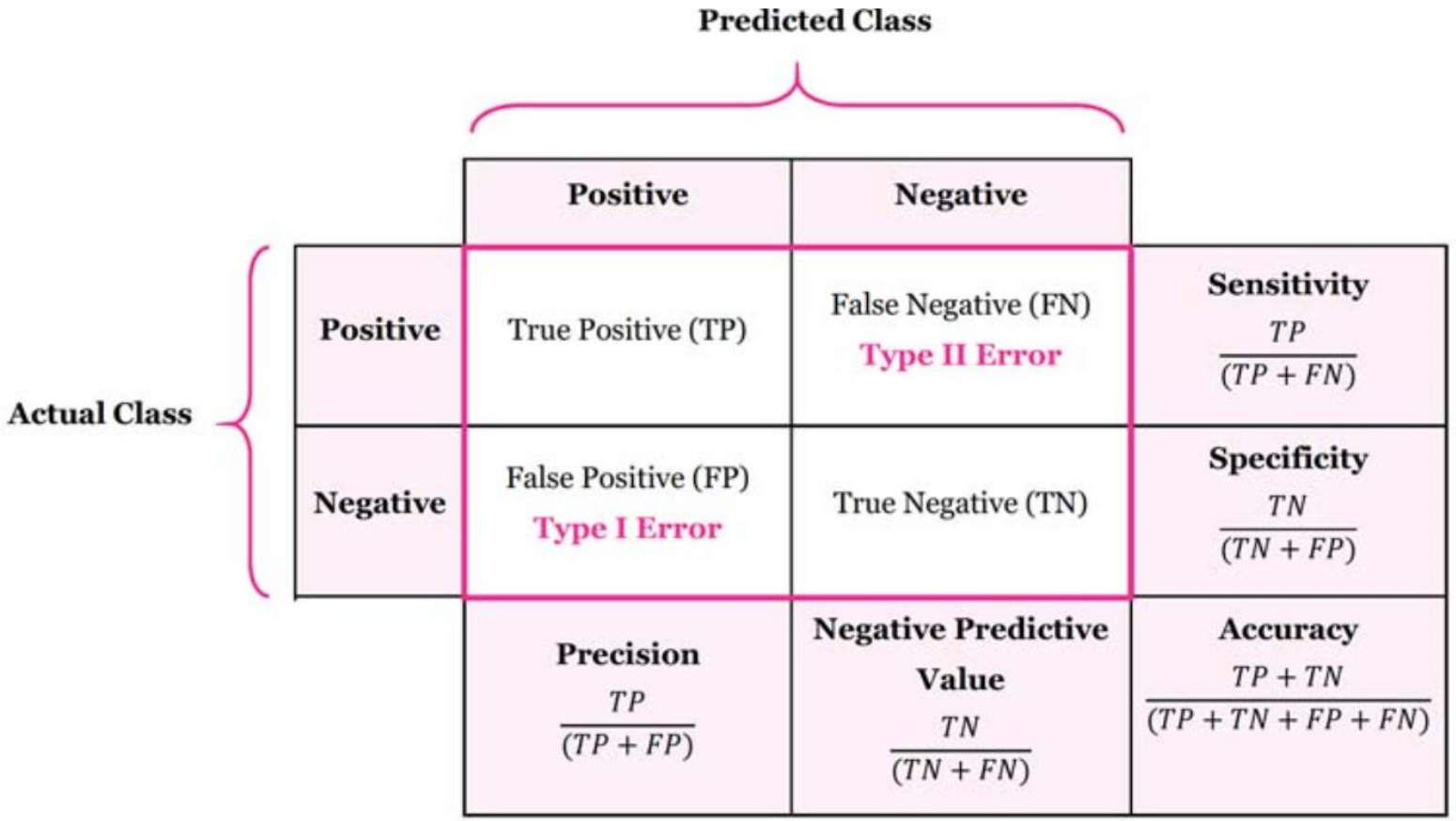
- Negative라고 예측한 것 중에서 얼마나 잘 맞았는지 비율

NPV

		Prediction	
		1	0
Actual Class	1	True Positive	False Negative
	0	False Positive	True Negative

$$\text{음성예측도}(Negative\ Prediced\ Value) = \frac{TN}{TN + FN}$$

분류 모델 성능 평가 지표


Predicted Class
Positive
Negative
Actual Class
Positive
True Positive (TP)
False Negative (FN)
Type II Error
Sensitivity
$\frac{TP}{(TP+FN)}$
Negative
False Positive (FP)
Type I Error
True Negative (TN)
Specificity
$\frac{TN}{(TN+FP)}$
Precision
$\frac{TP}{(TP+FP)}$
Negative Predictive Value
$\frac{TN}{(TN+FN)}$
Accuracy
$\frac{TP+TN}{(TP+TN+FP+FN)}$

F-measure

모델	Precision	Recall	평균
A	0.6	0.39	0.495
B	0.02	1.0	0.51

- 모델 B
  - 모든 데이터를 Positive로 분류한다면 → Recall = 1.0
  - 그러나 Negative도 Positive로 분류하기 때문에 좋은 분류기(X)

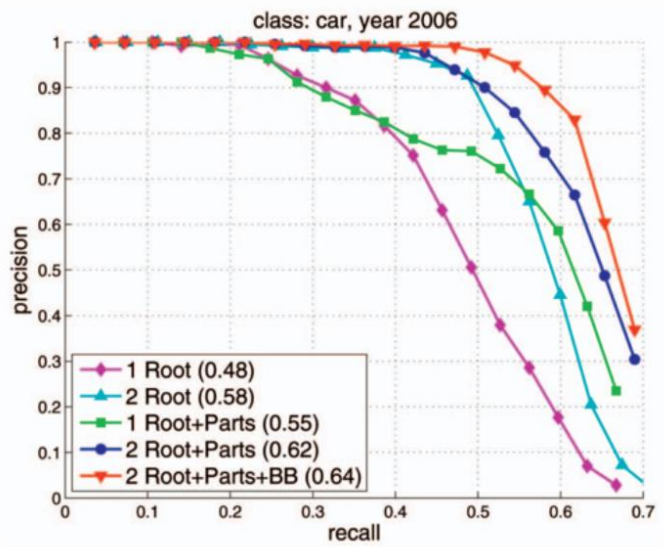
  ➔단순 평균으로 모델이 좋은지 나쁜지 평가하기 힘들 수 있음
- F-measure

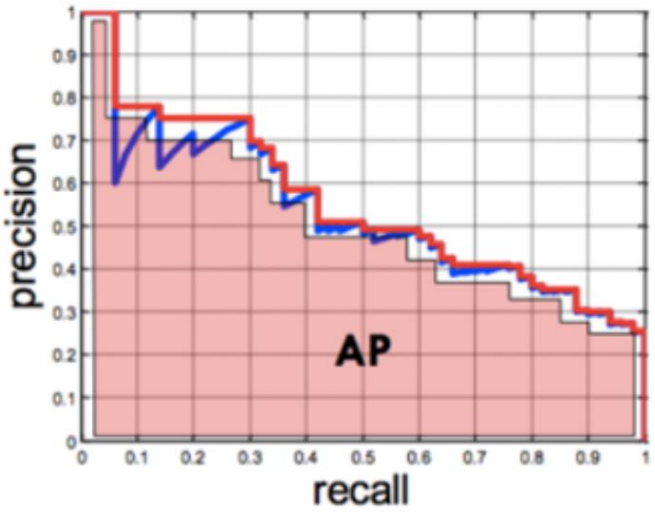
$$F-meaure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

모델	Precision	Recall	평균	F-measure
A	0.6	0.39	0.495	0.472
B	0.02	1.0	0.51	0.039

Average Precision (1)

- Precision-Recall 그래프
  - 알고리즘의 매개변수 조절에 따른 precision과 recall의 변화 표현
  - 알고리즘의 성능을 전반적으로 파악하기에는 좋으나 서로 다른 두 알고리즘을 정량적으로 비교하기에는 불편함
- Average Precision (AP)
  - Precision-Recall 그래피에서 그래프 선 아래쪽의 면적 계산
  - 면적의 값 (AP)가 높을수록 성능이 우세





Average Precision (2)

- 15개의 얼굴이 존재하는 이미지에서 얼굴 검출 알고리즘에 의해 총 10개의 얼굴이 검출되었다고 가정
  - Confidence Score를 0%에서부터 100%까지 모두 고려

Detections	confidences	TP or FP
A	57%	TP
B	78%	TP
C	43%	FP
D	85%	TP
E	91%	TP
F	13%	FP
G	45%	TP
H	68%	FP
I	95%	TP
J	81%	TP

  - Precision (7/10 = 0.7), Recall (7/15 = 0.47)

Average Precision (3)

- Confidence score에 따른 재정렬

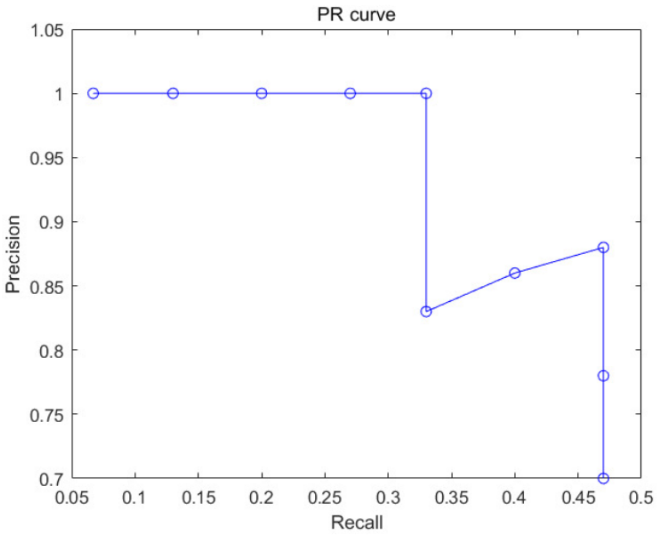
Detections	confidences	TP or FP
I	95%	TP
E	91%	TP
D	85%	TP
J	81%	TP
B	78%	TP
H	68%	FP
A	57%	TP
G	45%	TP
C	43%	FP
F	13%	FP

- Confidence score에 따른 Precision-Recall

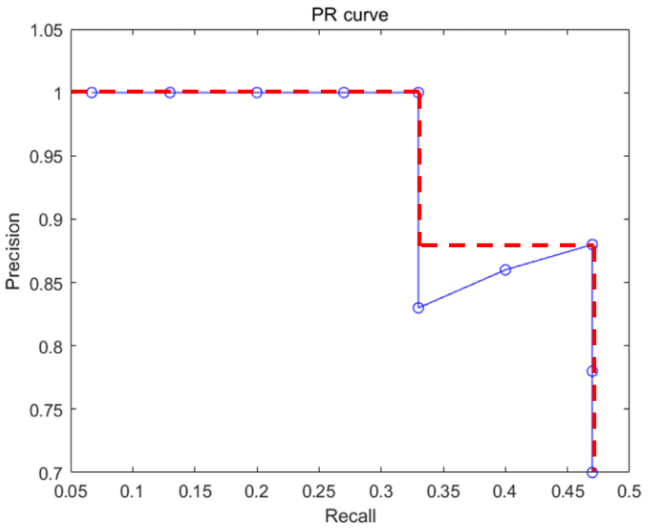
Detections	confidences	TP or FP	누적 TP	누적 FP	Precision	Recall
I	95%	TP	1	0	1/1=1	1/15=0.067
E	91%	TP	2	0	2/2=1	2/15=0.13
D	85%	TP	3	0	3/3=1	3/15=0.2
J	81%	TP	4	0	4/4=1	4/15=0.27
B	78%	TP	5	0	5/5=1	5/15=0.33
H	68%	FP	5	1	5/6=0.83	5/15=0.33
A	57%	TP	6	1	6/7=0.86	6/15=0.4
G	45%	TP	7	1	7/8=0.88	7/15=0.47
C	43%	FP	7	2	7/9=0.78	7/15=0.47
F	13%	FP	7	3	7/10=0.7	7/15=0.47

Average Precision (4)

- Precision-Recall 그래프



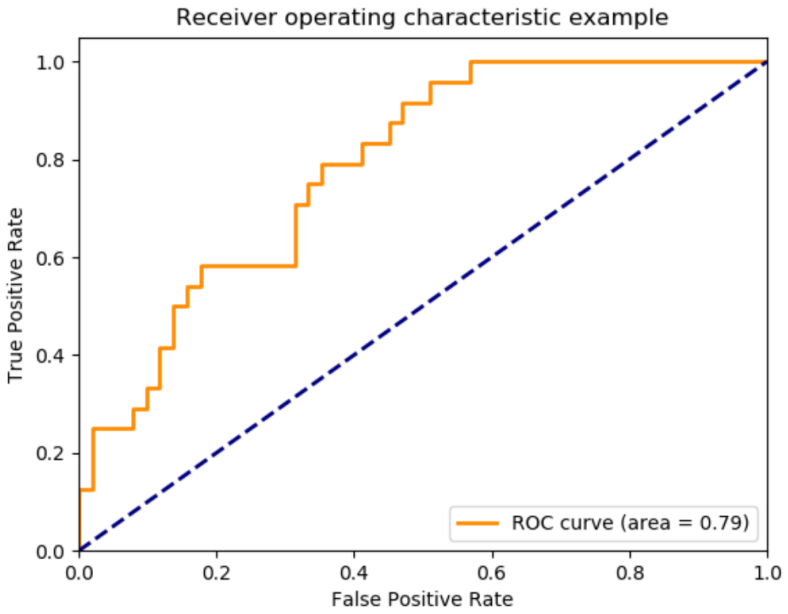
단조적으로
감소하는
그래프로 수정

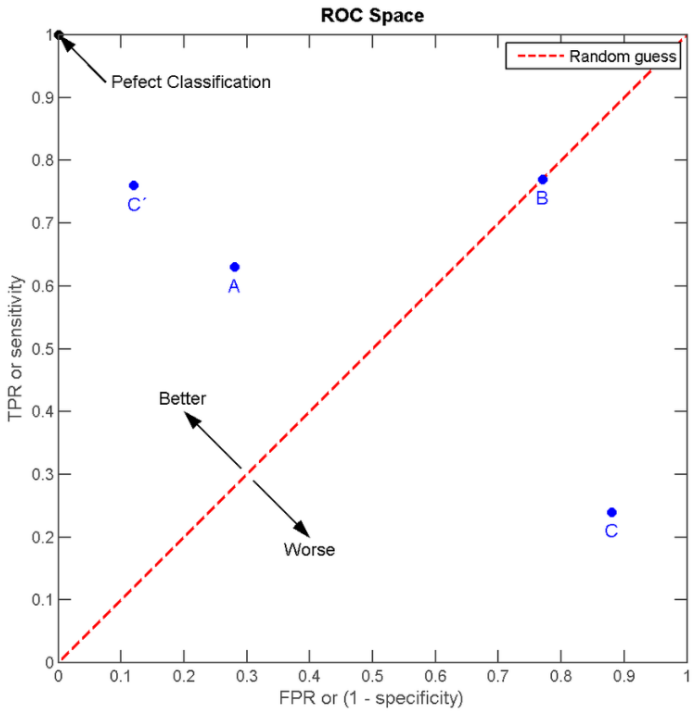


AP = 1×0.33+0.88×(0.47−0.33) = 0.4532

ROC Curve (1)

- Receiver Operating Characteristics (ROC) Curve
  - x축을 FP rate(1-specificity), y축을 TP rate으로 하여 시각화한 그래프
    - TPR과 FPR은 비례 관계
    - 모든 환자를 암이라고 한다면 → TPR=1, FPR=1
    - 모든 환자를 정상이라고 한다면 → TPR=0, FPR=0
  - 이진 분류 또는 의료 분야에서 많이 사용되는 성능 지표





ROC Curve (2)

- 모델의 예측 확률을 기준으로 오름차순 정렬

실제 정답	예측 확률
P	0.6
N	0.7
P	0.4
N	0.2

실제 정답	예측 확률
N	0.7
P	0.6
P	0.4
N	0.2

- 예측 확률에 따른 FPR/TPR 계산

FPR = AP = 1×0.33+0.88×(0.47−0.33) = 0.4532

Threshold = 0.7

모델 예측 결과 및 실제 정답 데이터

실제 정답	예측 확률	모델의 예측
N	0.7	P
P	0.6	N
P	0.4	N
N	0.2	N

$$False\ Positive\ Rate = \frac{\sum False\ Positive\ 수}{\sum 정답의\ Negative\ 수} = \frac{1}{2}$$

$$True\ Positive\ Rate = \frac{\sum True\ Positive\ 수}{\sum 정답의\ Positive\ 수} = \frac{0}{2}$$

$$(\frac{1}{2}, 0)$$

Threshold = 0.6

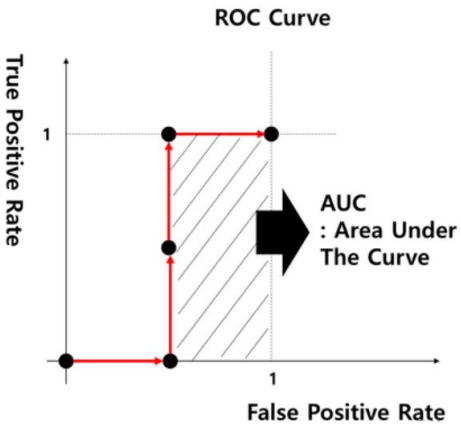
모델 예측 결과 및 실제 정답 데이터

실제 정답	예측 확률	모델의 예측
N	0.7	P
P	0.6	P
P	0.4	N
N	0.2	N

$$True\ Positive\ Rate = \frac{\sum True\ Positive\ 수}{\sum 정답의\ Positive\ 수} = \frac{1}{2}$$

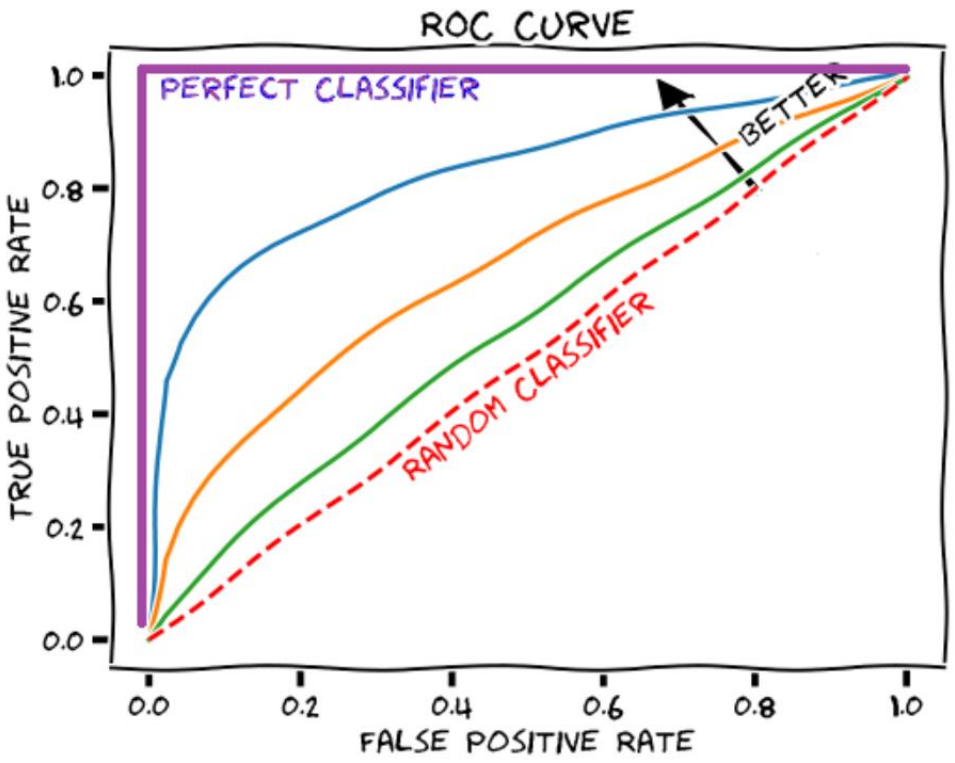
$$False\ Positive\ Rate = \frac{\sum False\ Positive\ 수}{\sum 정답의\ Negative\ 수} = \frac{1}{2}$$

$$(\frac{1}{2}, \frac{1}{2})$$



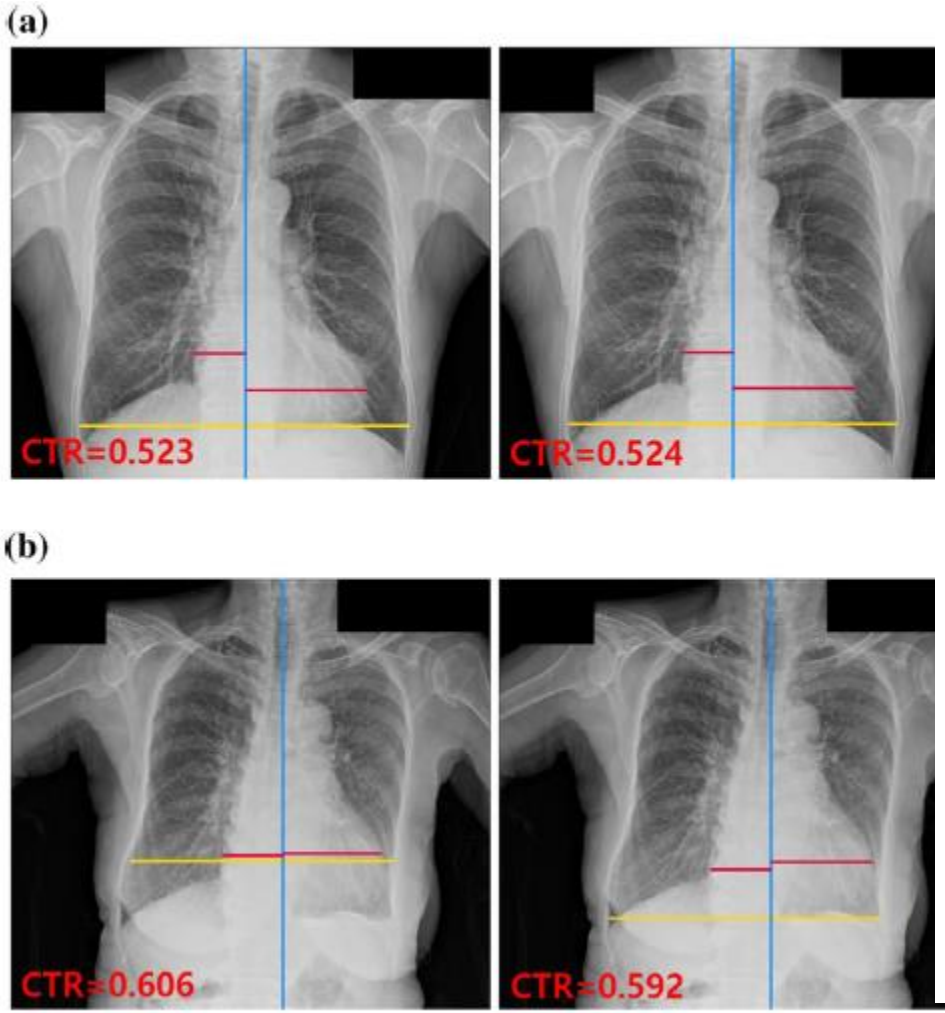
AUROC

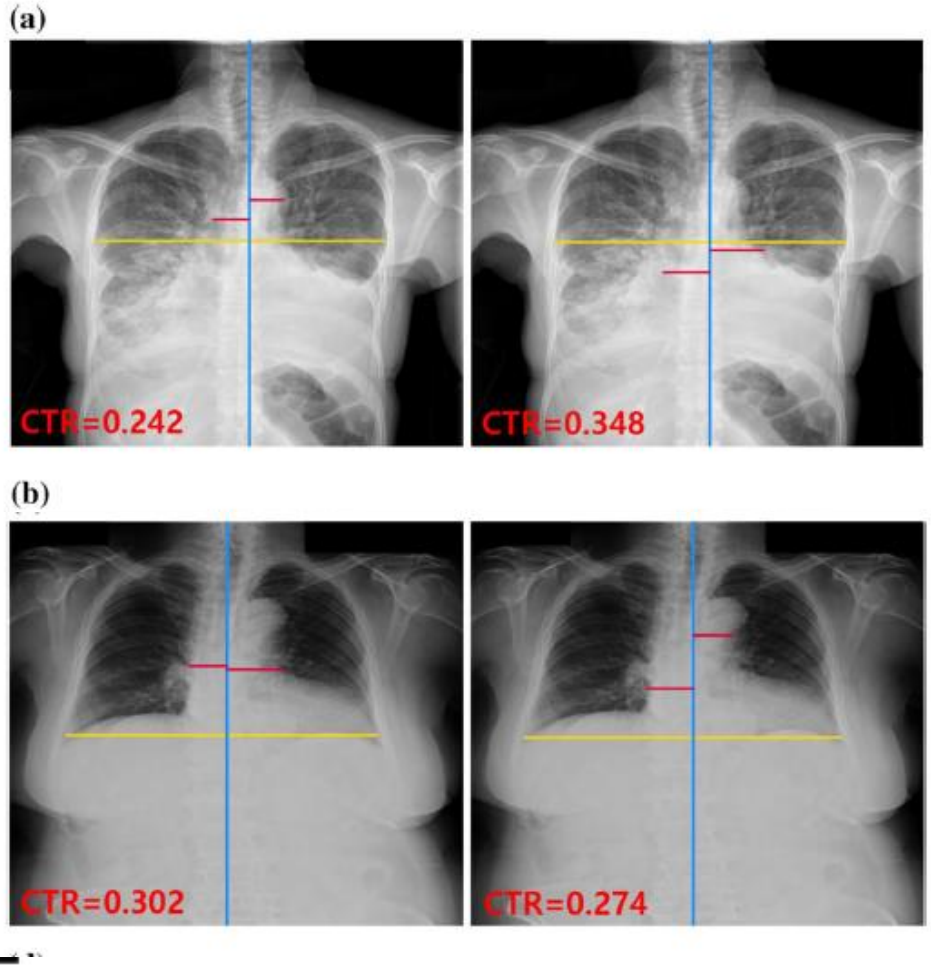
- AUROC
  - Area Under the ROC Curve



AUROC value	Interpretation
0.5	The worst
1.0	The best
0.50 ~ 0.70	Sub-optimal
0.70 ~ 0.80	Good performance
> 0.80	Excellent performance

An Example


(a)
CTR=0.523
CTR=0.524
(b)
CTR=0.606
CTR=0.592


(a)
CTR=0.242
CTR=0.348
(b)
CTR=0.302
CTR=0.274

An Example (cont.)

Method	Category	n	Mean ± SD (RS)	Mean ± SD (predicted)	95% CI for mean differences	T-value	*p*-value	MAE ± SD
Segmentation-based method 1	All cases	1000	0.506 ± 0.080	0.505 ± 0.080	(−0.00195, 0.00310)	0.45	0.655	0.023 ± 0.033
	Subgroup 1	123	0.449 ± 0.068	0.455 ± 0.069	(−0.01156, −0.00190)	−2.76	0.007	0.020 ± 0.020
	Subgroup 2	202	0.530 ± 0.083	0.522 ± 0.092	(0.00026, 0.01713)	2.03	0.043	0.035 ± 0.050
	Subgroup 3	83	0.543 ± 0.088	0.514 ± 0.100	(0.01004, 0.04909)	3.01	0.003	0.058 ± 0.074
	Subgroup 4	63	0.493 ± 0.078	0.493 ± 0.081	(−0.00596, 0.01459)	0.84	0.404	0.025 ± 0.032
	Subgroup 5	652	0.505 ± 0.075	0.508 ± 0.075	(−0.00456, 0.00019)	−1.81	0.071	0.019 ± 0.024
Segmentation-based method 2	All cases	1000	0.506 ± 0.080	0.506 ± 0.078	(−0.00263, 0.00236)	−0.10	0.917	0.024 ± 0.032
	Subgroup 1	123	0.449 ± 0.068	0.457 ± 0.069	(−0.01234, −0.00346)	−3.52	0.001	0.021 ± 0.015
	Subgroup 2	202	0.530 ± 0.083	0.524 ± 0.094	(−0.00194, 0.01509)	1.52	0.130	0.038 ± 0.049
	Subgroup 3	83	0.543 ± 0.088	0.514 ± 0.097	(0.01113, 0.04786)	3.20	0.002	0.059 ± 0.066
	Subgroup 4	63	0.493 ± 0.078	0.496 ± 0.081	(−0.01237, 0.00653)	−0.62	0.539	0.025 ± 0.028
	Subgroup 5	652	0.505 ± 0.075	0.507 ± 0.072	(−0.00443, 0.00028)	−1.73	0.084	0.019 ± 0.023

Method	Category	No. of samples					Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
		TP	FP	TN	FN	n				
Segmentation-based method 1	All cases	466	66	444	24	1000	91 (89, 93)	95 (94, 96)	87 (85, 89)	0.96 (0.94, 0.97)
	Subgroup 1	25	3	91	4	123	94 (90, 98)	86 (80, 92)	97 (94, 100)	0.96 (0.92, 0.99)
	Subgroup 2	107	22	68	5	202	87 (82, 92)	96 (93, 99)	76 (70, 82)	0.92 (0.86, 0.97)
	Subgroup 3	43	17	22	1	83	78 (69, 87)	98 (95, 100)	56 (45, 67)	0.86 (0.76, 0.98)
	Subgroup 4	25	6	31	1	63	89 (81, 97)	96 (91, 100)	84 (75, 93)	0.93 (0.85, 1.00)
	Subgroup 5	314	34	288	16	652	92 (90, 94)	95 (93, 97)	89 (87, 91)	0.97 (0.95, 0.98)
Segmentation-based method 2	All cases	470	61	436	32	1000	91 (89, 93)	94 (93, 95)	88 (86, 90)	0.95 (0.94, 0.97)
	Subgroup 1	25	3	91	4	123	94 (90, 98)	86 (80, 92)	97 (94, 100)	0.98 (0.96, 1.00)
	Subgroup 2	105	24	62	11	202	83 (78, 88)	91 (87, 95)	72 (66, 78)	0.89 (0.84, 0.95)
	Subgroup 3	43	17	19	4	83	75 (66, 84)	91 (85, 97)	53 (42, 64)	0.81 (0.69, 0.93)
	Subgroup 4	28	3	29	3	63	90 (83, 97)	90 (83, 97)	91 (84, 88)	0.95 (0.88, 1.00)
	Subgroup 5	317	31	288	16	652	93 (91, 95)	95 (93, 97)	90 (88, 92)	0.97 (0.96, 0.99)

References

- 머신러닝에서 사용되는 평가 지표, https://gaussian37.github.io/ml-concept-ml-evaluation/
- mAP (mean Average Precision) for Object Detection, https://medium.com/@jonathan_hui/map-mean-average-precision-for-object-detection-45c121a31173
- 물체 검출 알고리즘 성능 평가 방법 AP(Average Precision)의 이해, https://bskyvision.com/465
- Receiver Operating Characteristics (ROC), https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html
- ROC Curve란? ROC 커브 해석, https://losskatsu.github.io/machine-learning/stat-roc-curve/#2-1-%EC%A0%95%EB%B0%80%EB%8F%84precision%EC%99%80-%EB%AF%BC%EA%B0%90%EB%8F%84recall
- Measuring Performance: AUC(AUROC), https://glassboxmedicine.com/2019/02/23/measuring-performance-auc-auroc/
- ROC 완벽 정리, https://m.blog.naver.com/PostView.nhn?blogId=sw4r&logNo=221015817276&proxyReferer=https%3A%2F%2Fwww.google.com%2F


ANY
QUESTIONS?