한국소음진동공학회 AI혁신위원회 주관
2022년 인공지능 강습회

기계이상진단을 위한 인공지능 학습 기법

제 2강 이상진단 성능평가 방법

한국과학기술원 전기및전자공학부

최정우
jwoo@kaist.ac.kr

KAIST EE

강의 목차

- 자기지도 학습의 예
  - 재현 학습법 Reconstruction Task

- 이상진단 성능평가
  - 심층 신경망의 이상진단 성능을 평가하는 방법

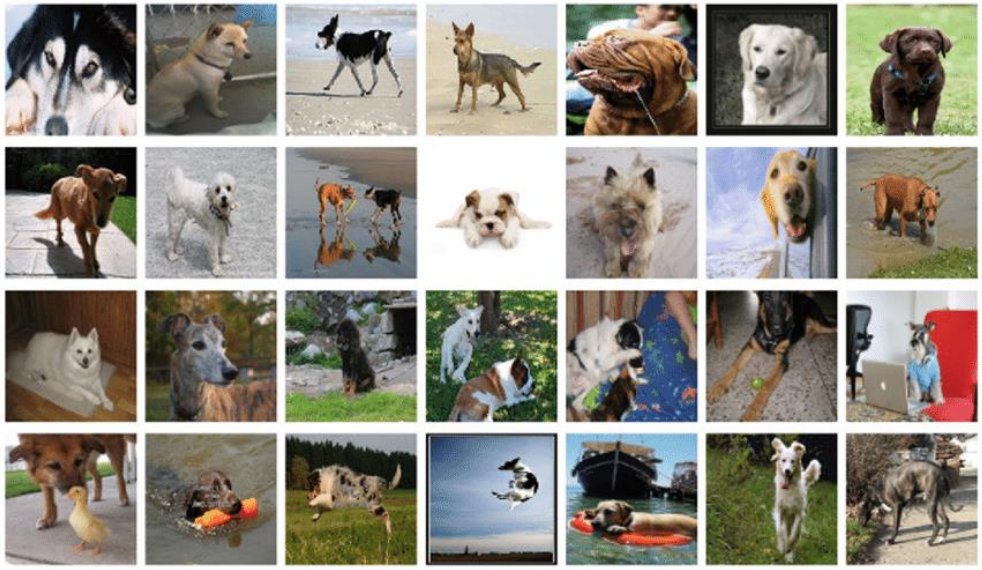
  - Accuracy/TPR/FPR
  - Precision/Recall
  - ROC-AUC

심층 신경망을 사용한 이상진단

- Training only with normal data
  - DNNs only capture the “features” of normal data
  - Some intra-class can be provided for the normal data

- Test & Validate with normal + anomaly data

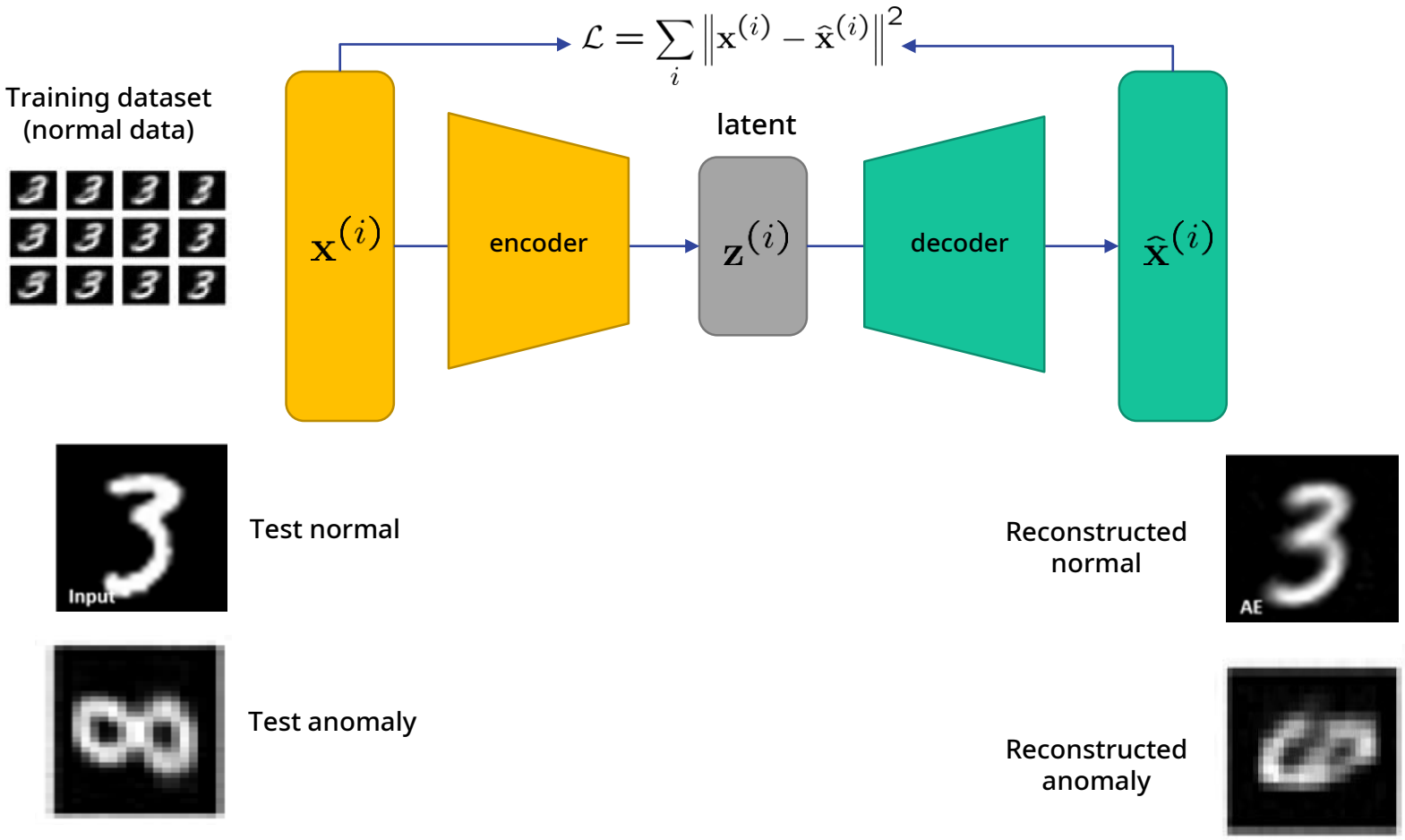
Training data



Test data

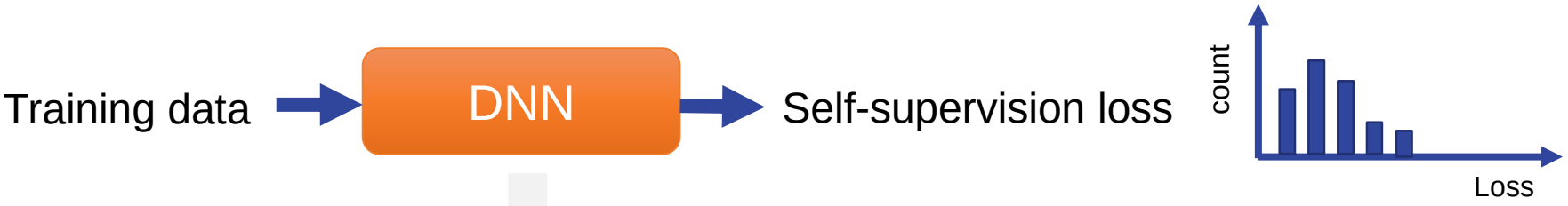
자기지도학습: 재현 훈련 기법

- Autoencoder example

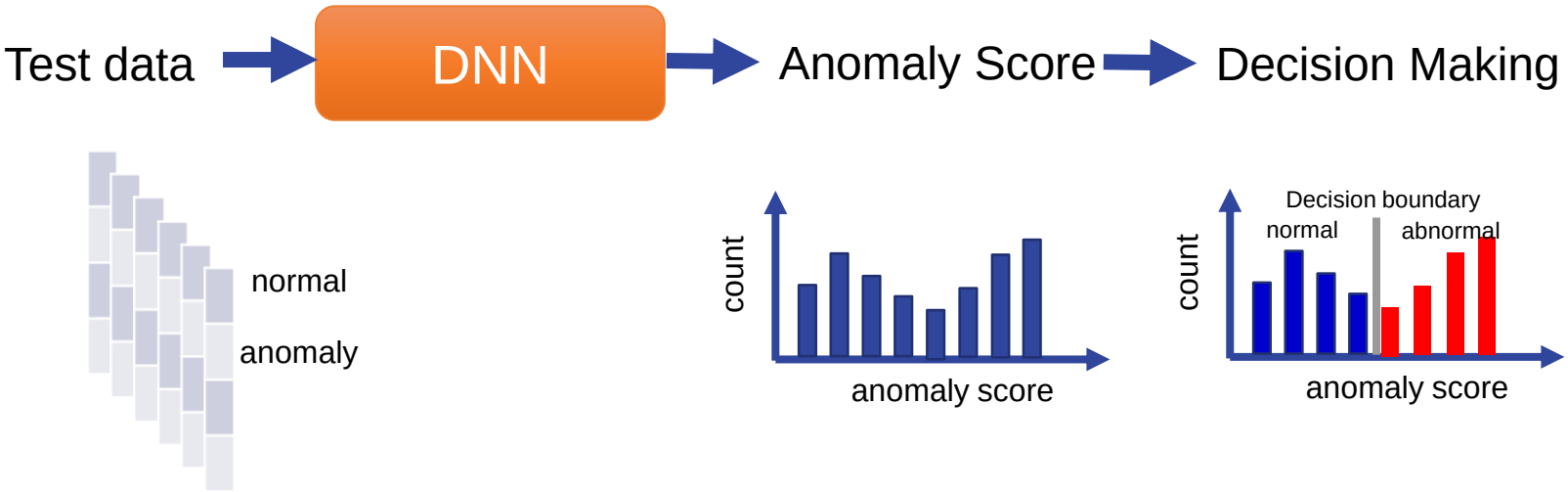


이상 진단 과정

- Pretraining using pretext task

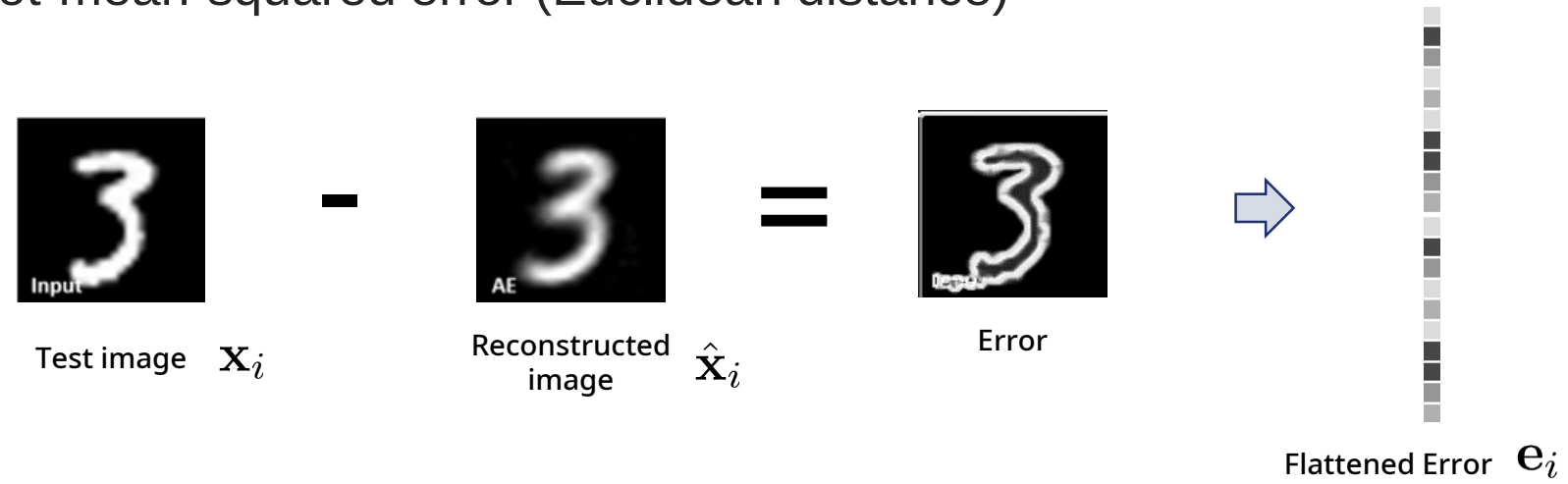


- Downstream task (anomaly detection)



이상 점수 (anomaly score)의 예

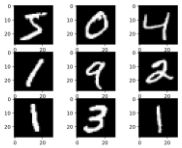
- Root-mean-squared error (Euclidean distance)

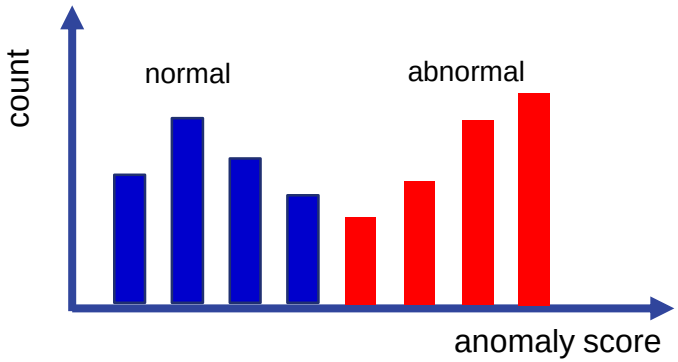


Anomaly score

RMS error $$||\mathbf{e}_i|| = \sqrt{\frac{1}{K}\sum_k |e_i(k)|^2}$$

Test dataset





이상 점수의 예) Mahalanobis distance

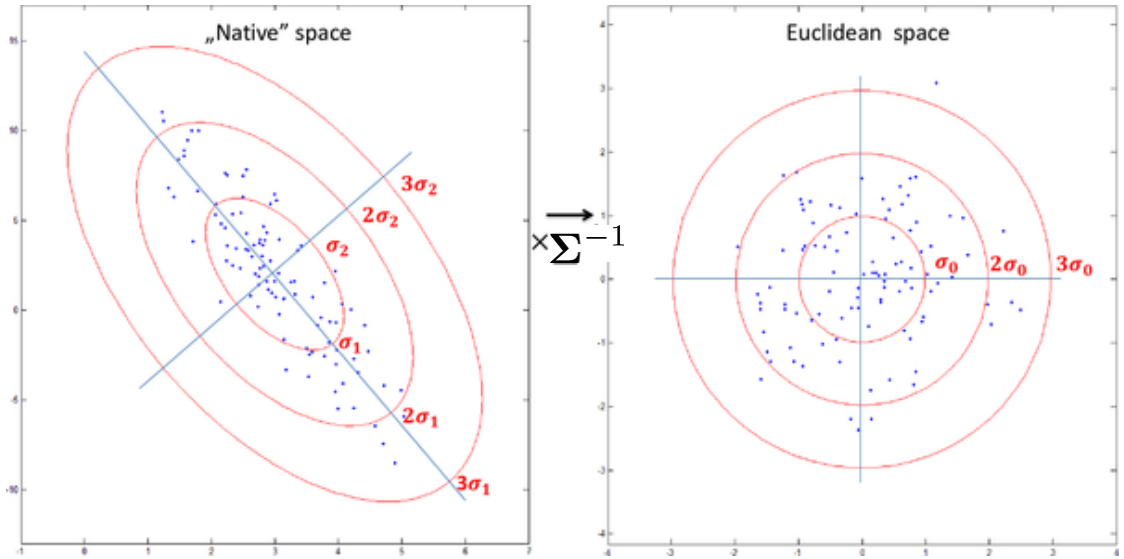
- Distance in a multivariate space normalized by mean and variance

$$d(\mathbf{e}_i) = \sqrt{(\mathbf{e}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{e}_i - \boldsymbol{\mu})}$$

mean: translation

Covariance: rotation & scaling

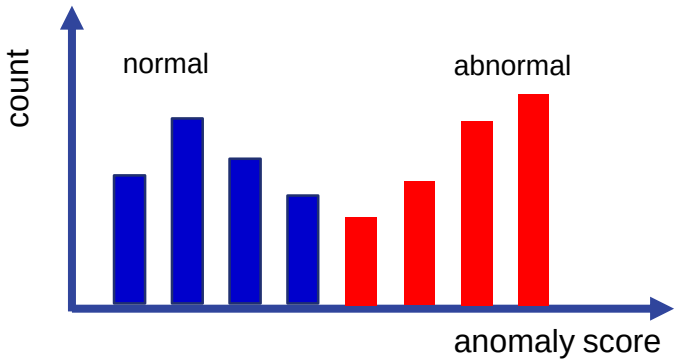
Flattened Error $\mathbf{e}_i = \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$



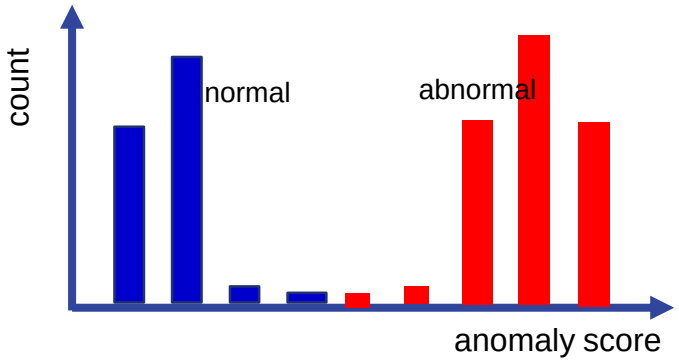
좋은 이상진단기란 ?

- Which model is better?

Model A



Model B



이상진단기 성능평가 지표

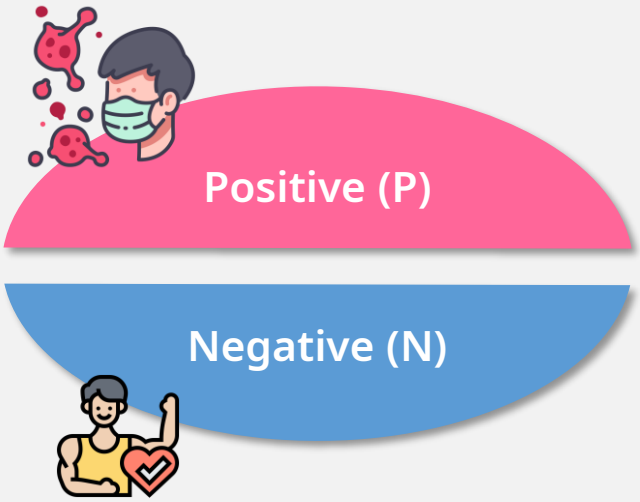
Performance indexes & measures
for Deep Anomaly Detection

이상진단기의 예

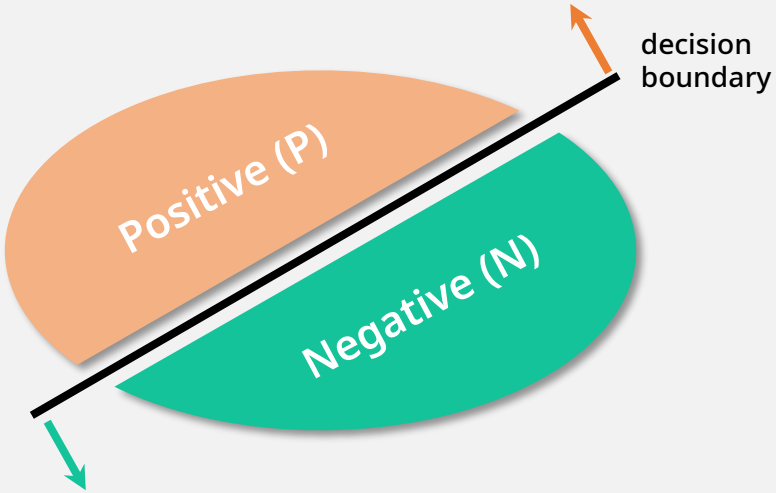
- A coronavirus diagnosis kit
  - How to evaluate the performance of the kit?

이상진단 결과 – 실제와 판단 차이

- Prediction of Positive / Negative using decision boundary



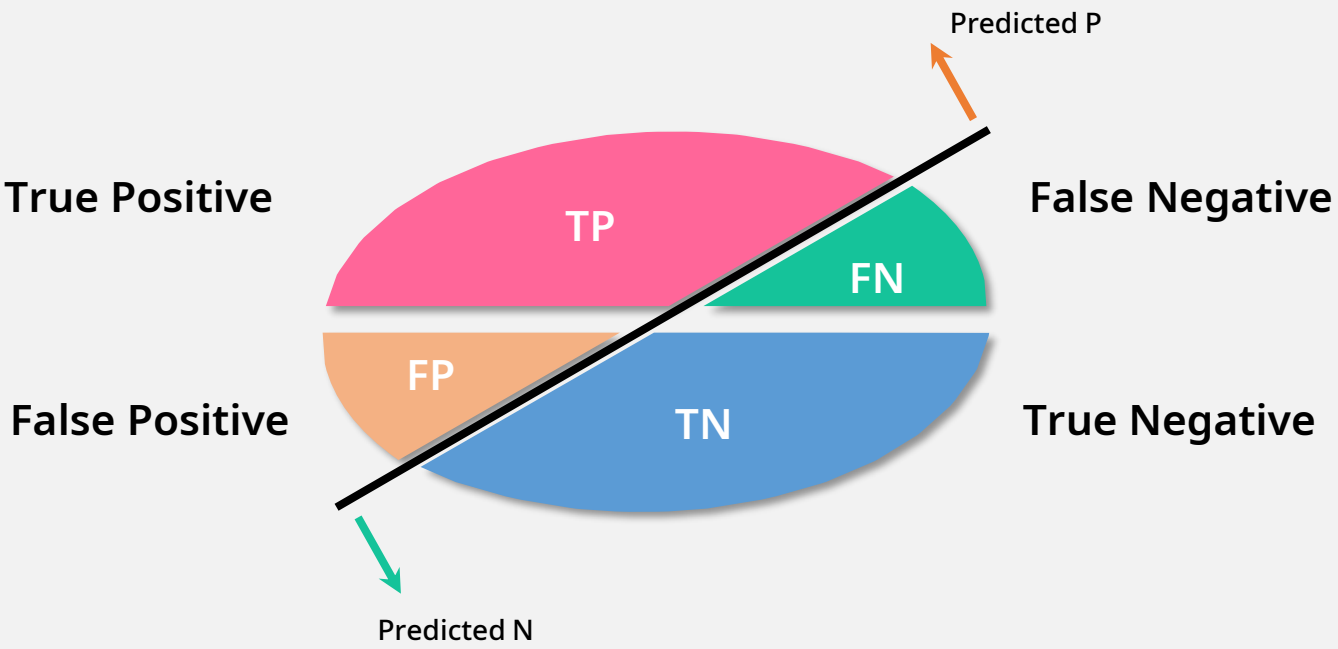
Actual



Prediction

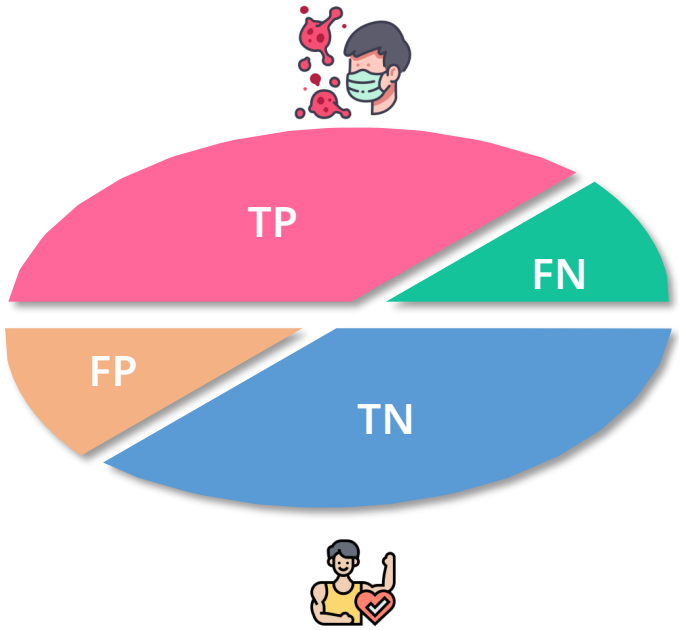
이상진단 결과 – 실제와 판단 차이

- 진양성(TP), 위양성(FP), 진음성(TN), 위음성(FN)

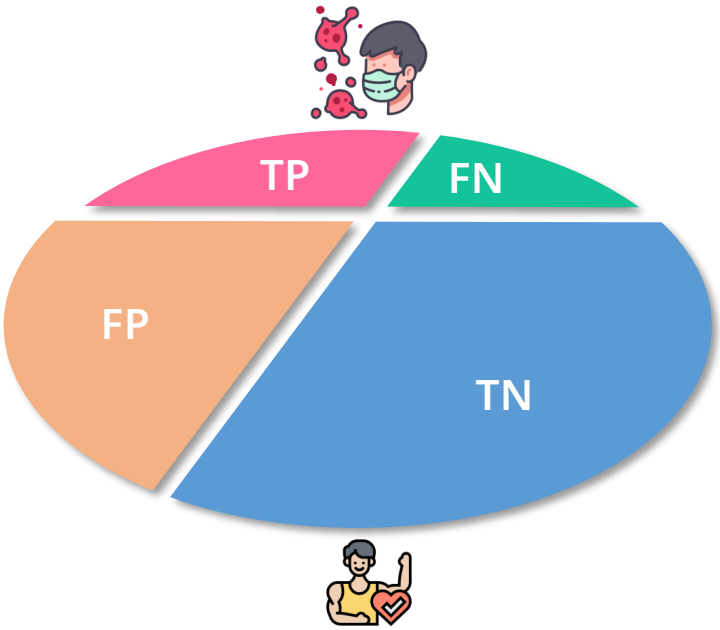


이상진단 결과 – 실제와 판단 차이

- Data imbalance problem
  - Metrics can be distorted when the number of Ps and Ns are very different



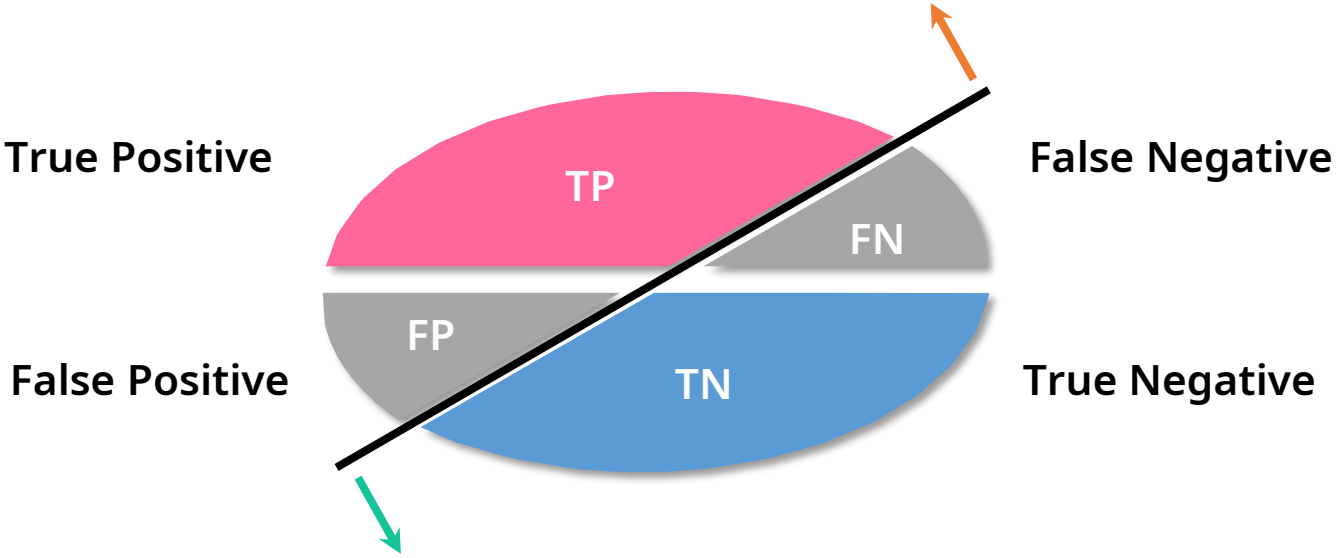
균형 데이터 **(P≈N)**



불균형 데이터 **(P<<N)**

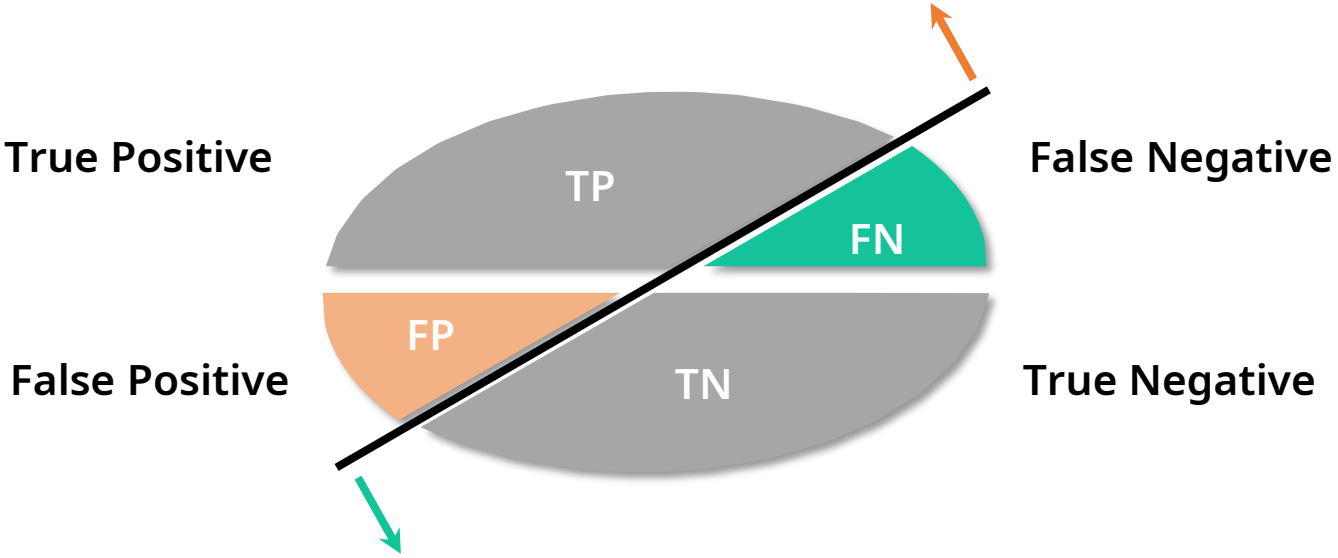
Metrics - 정확도

- **Accuracy**
  - The ratio of correct prediction to incorrect prediction
  - **(TP+TN)/(All)**



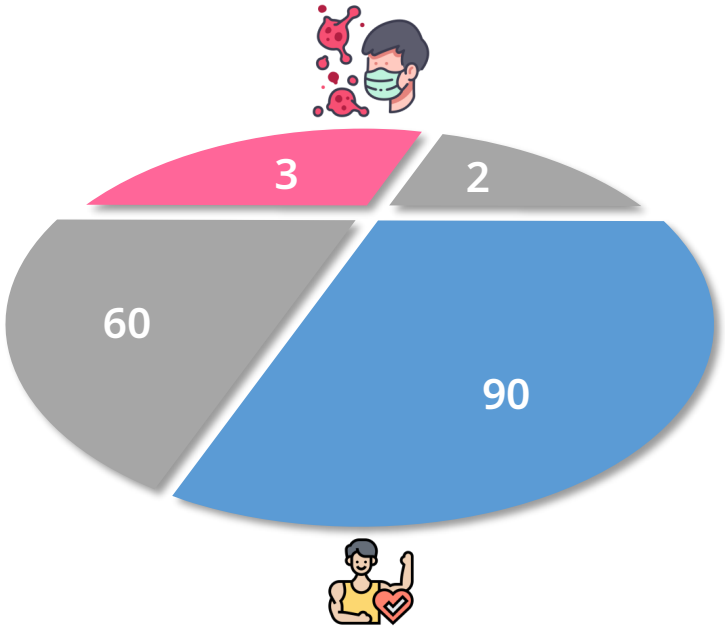
Metrics - 에러율

- **Error rate (ERR)**
  - 1 - Accuracy
  - **(FP+FN)/(All)**



정확도의 문제점

- **불균형 데이터**
  - Accuracy can be biased for imbalanced data
  - For less number of POSITVE data,
    metric is mostly determined by TN data (90)
  - Accuracy = (3+**90**)/(3+2+90+60)
  - What is the problem?
    - When normal (negative) dataset is large,
      the detection performance depends on the filtering of
      true normal data, irrespective of the detection of abnormal (positive) data



정확도의 문제점

- **불균형 데이터**
  - Corona virus kit

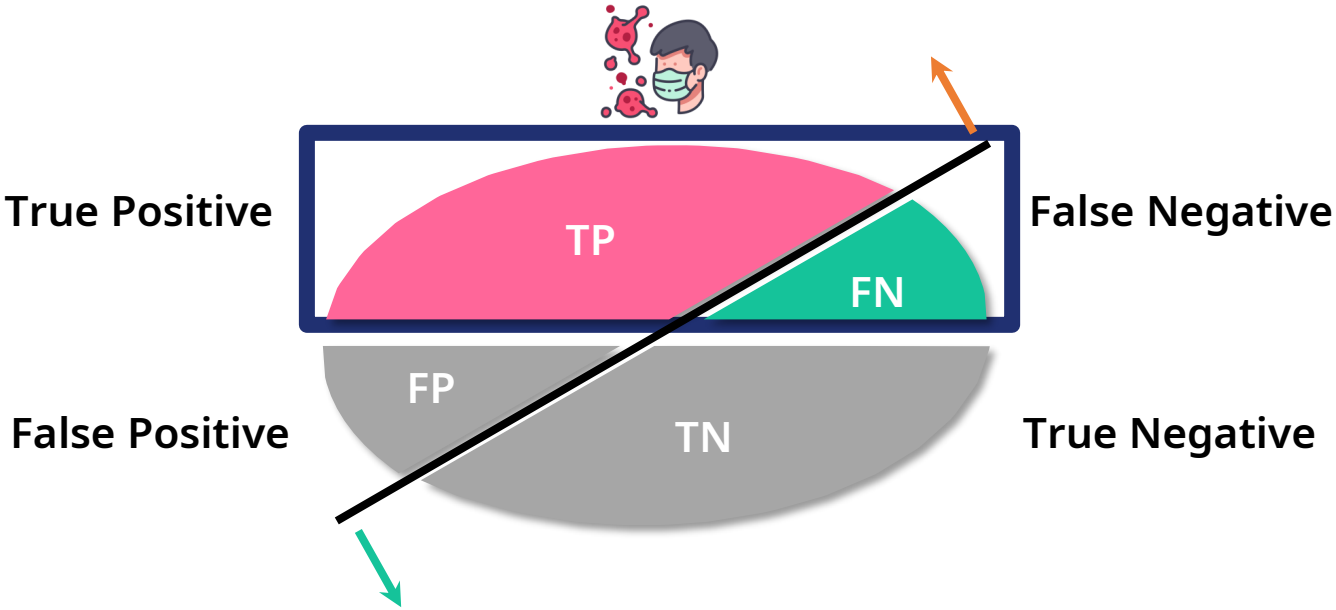
Prediction result (진단 결과)

Status of patients (실제 정답)	**Positive**	**Negative**	Marginal Statistics
Positive	(TP) 20	(FN) 100 (type II error)	120 (infected)
Negative	(FP) 40 (type I error)	(TN) 8000	8040 (uninfected)
Marginal Statistics	60 (predicted as infected)	8100 (predicted as uninfected)	

Accuracy: (8000+20)/(8160)= 98%, even though the kit cannot detect 100 patients !

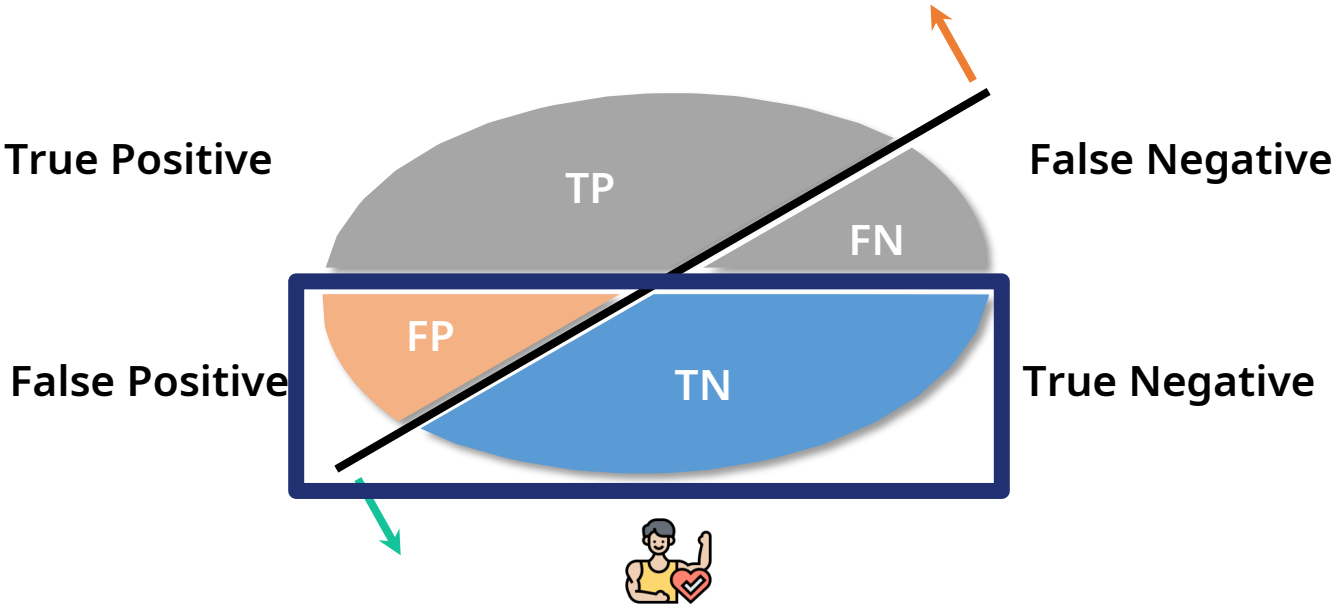
재현율 (민감도)

- **Recall** (True Positive Rate; **TPR, sensitivity**)
  - How many true POSITIVES among real POSITIVE data ?
  - TP / (TP+FN)
  - 환자를 얼마나 잘 찾는가



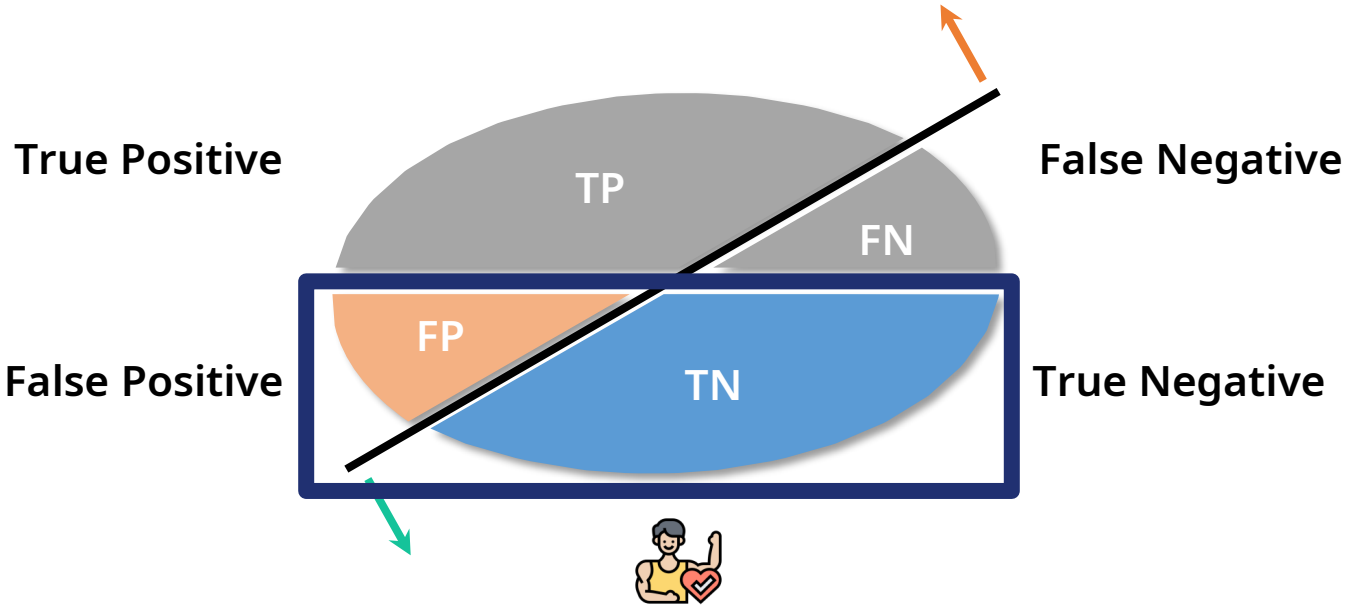
특이도

- **Specificity (TNR)**
  - Recall for NEGATIVE data : TN / (TN+FP)
  - How many negative data I predicted among true NEGATIVES ?
  - 정상을 얼마나 잘 찾는가



위양성율

- **False positive rate (FPR)**
  - 1 – specificity: FP / (TN+FP)



Metrics

- Summary
  - Corona virus kit

Prediction result

Status of patients	Positive	Negative	Marginal Statistics
Positive (TPR (Recall), FNR)	(TP) 20	(FN) 100	120 (infected)
Negative (TNR (Specificity), FPR)	(FP) 40	(TN) 8000	8040 (uninfected)
Marginal Statistics	60 (predicted as infected)	8100 (predicted as uninfected)	

TPR-FPR

Marginal probability of Positives
P (P)= 120/(8040+120)

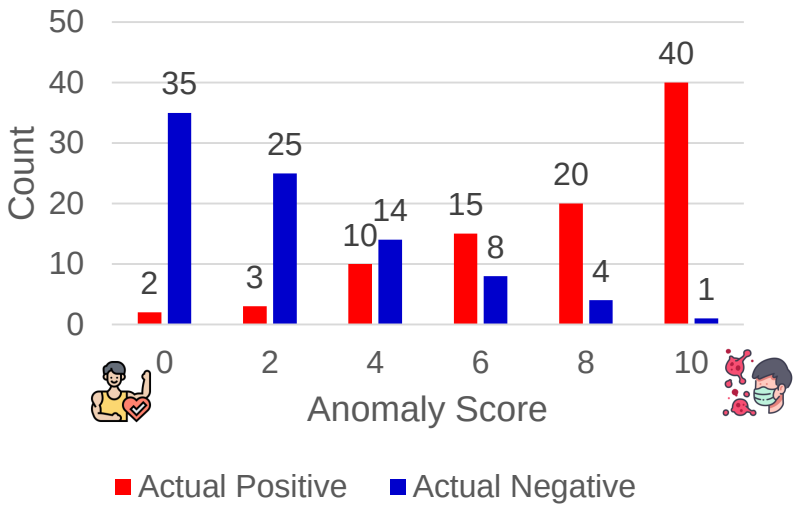
Prediction result

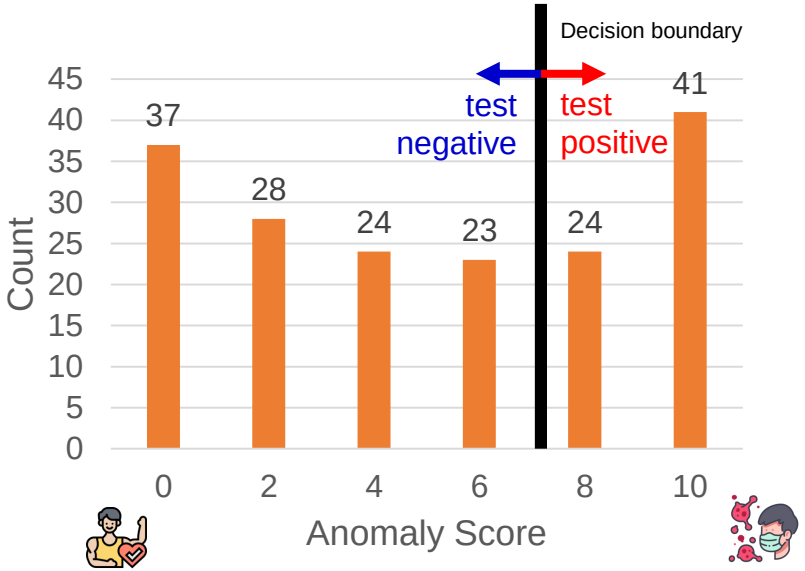
Status of patients	Positive	Negative	Marginal Statistics
Positive	(TP) 20	(FN) 100	120 (infected)
Negative	(FP) 40	(TN) 8000	8040 (uninfected)
Marginal Statistics	60 (predicted as infected)	8100 (predicted as uninfected)	

TPR(Recall): 20/120 = 20%
FPR: 40/8040 = 0.5%

히스토그램 예

- COVID-19 Kit test results
  - Suppose that the kit evaluate score for each patient (0~10)
    - 0: likely negative (healthy)
    - 10: likely positive (infected)

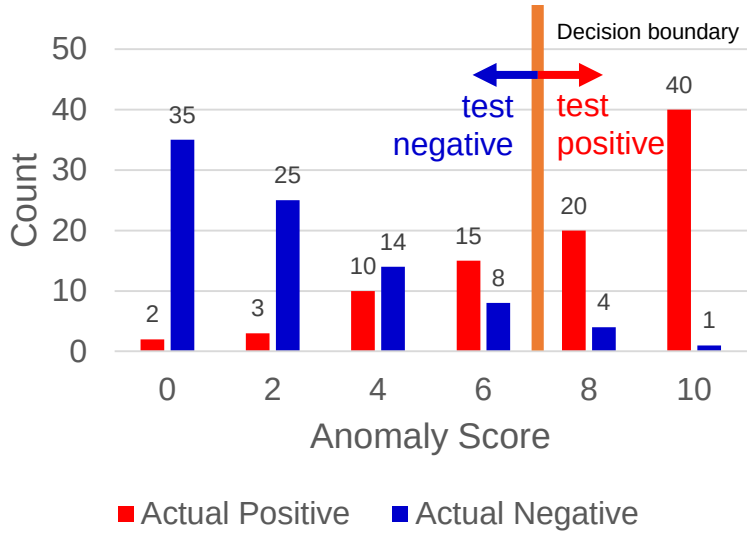


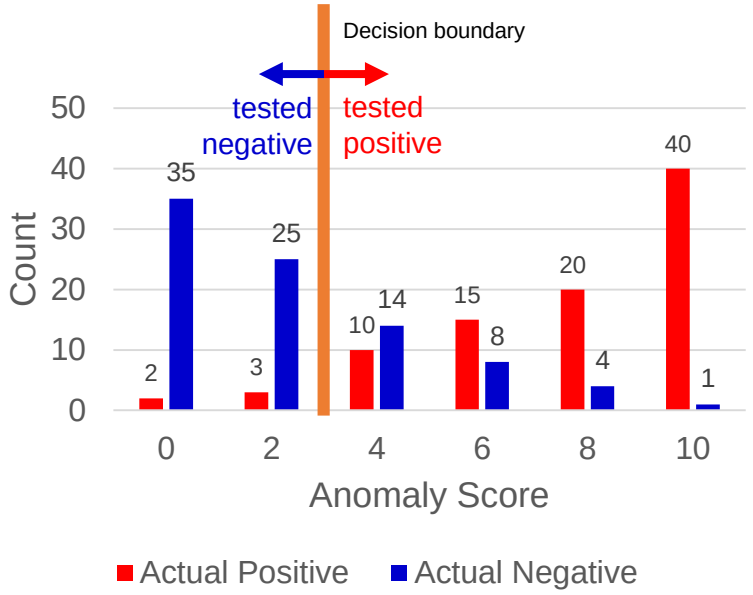


결정 경계

- Decision boundary
  - A boundary to discriminate positives and negatives

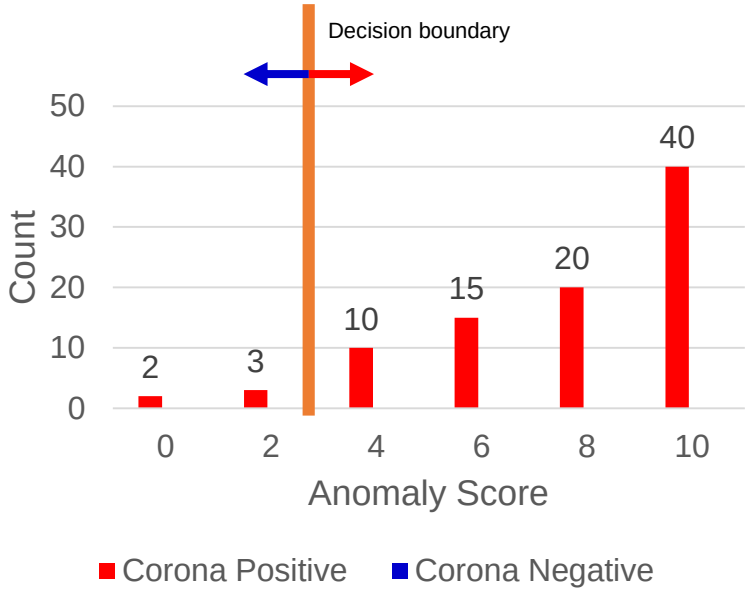
- Performance of diagnosis kit
  - Varies with the position of decision boundary





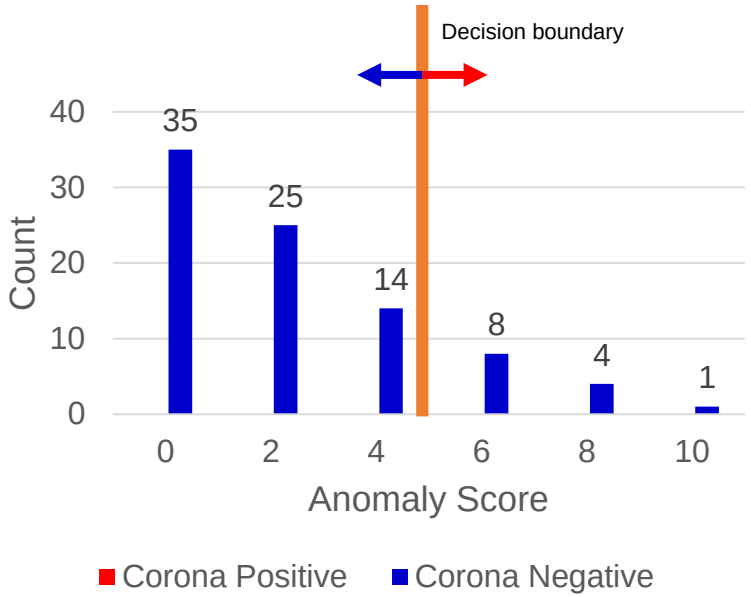
TPR-FPR 커브

- TPR
  - Ratio of positive data on the left/right of the decision boundary
    - In this example, TPR = (15+20+40)/(90)=83%
  - TPR changes as we move the decision boundary
    - TPR = (10+15+20+40)/90 = 94%
  - TPR increases as we move the decision boundary to the left



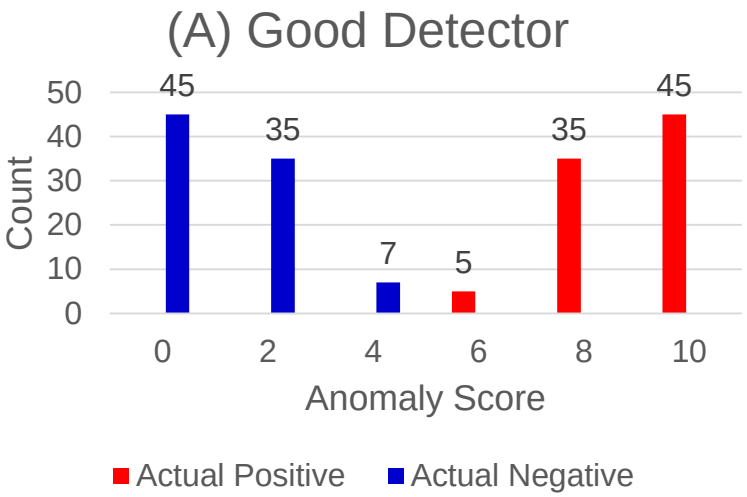
TPR-FPR 커브

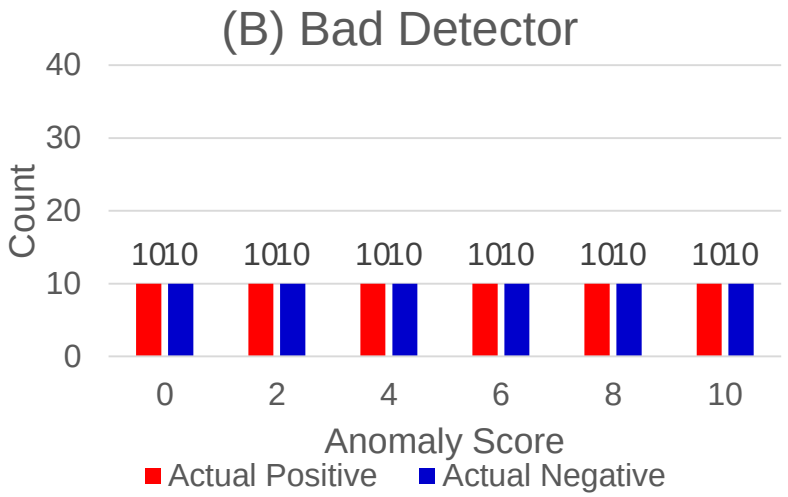
- FPR
  - Ratio of positive data on the left/right of the decision boundary
    - In this example, FPR = (8+4+1)/(87)=15%
  - FPR changes as we move the decision boundary
    - TPR = (14+8+4+1)/87 = 31%
  - FPR increases as we move the decision boundary to the left



무엇이 좋은 진단기인가?

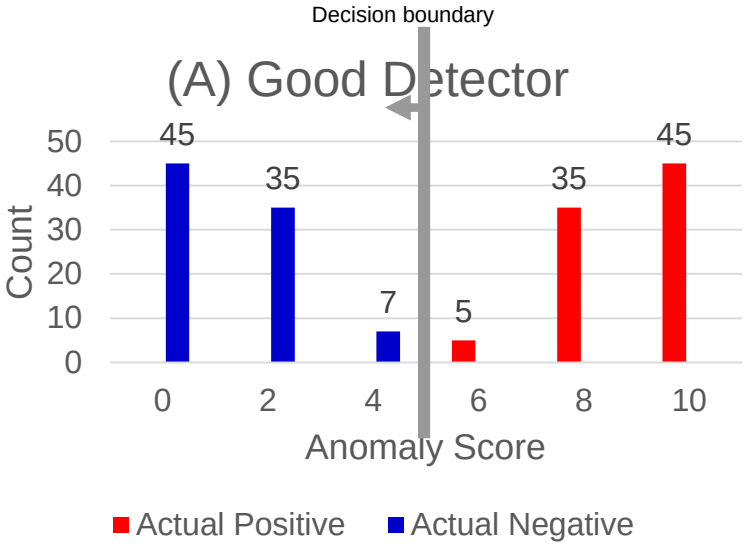
- How can we measure the goodness of “distribution”?
  - Independent of decision boundary

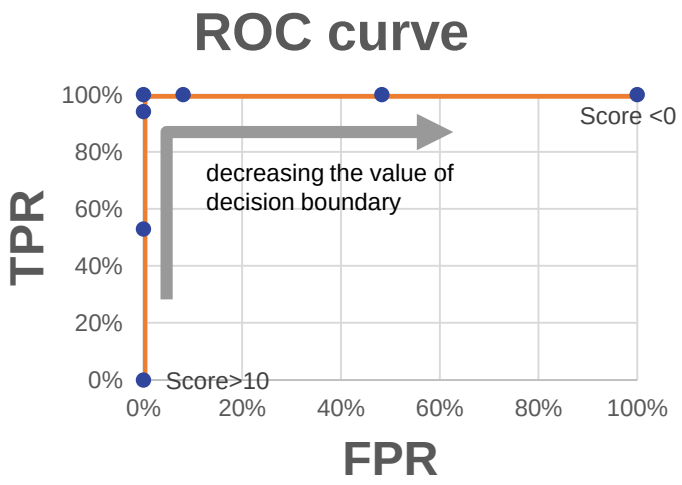




TPR-FPR 커브

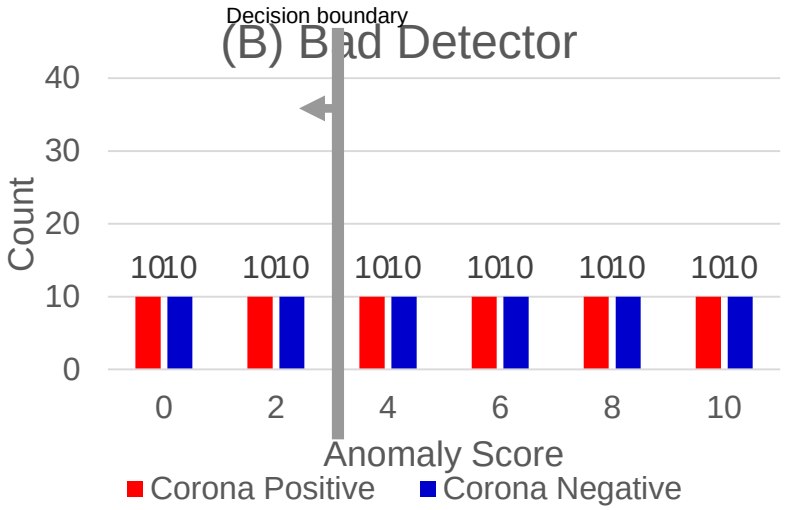
- Change of TPR & FPR with respect to decision boundary

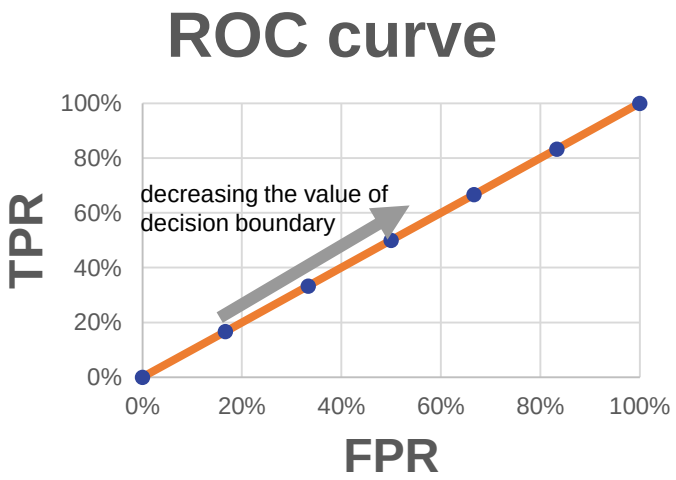




TPR-FPR 커브

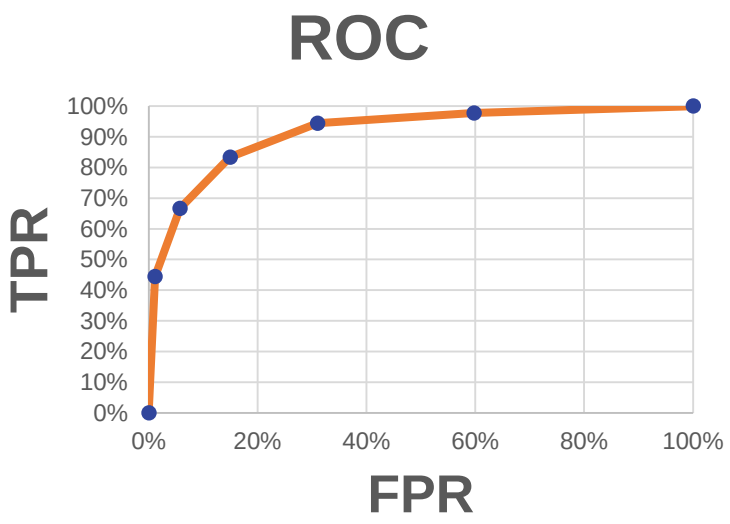
- Change of TPR & FPR with respect to decision boundary

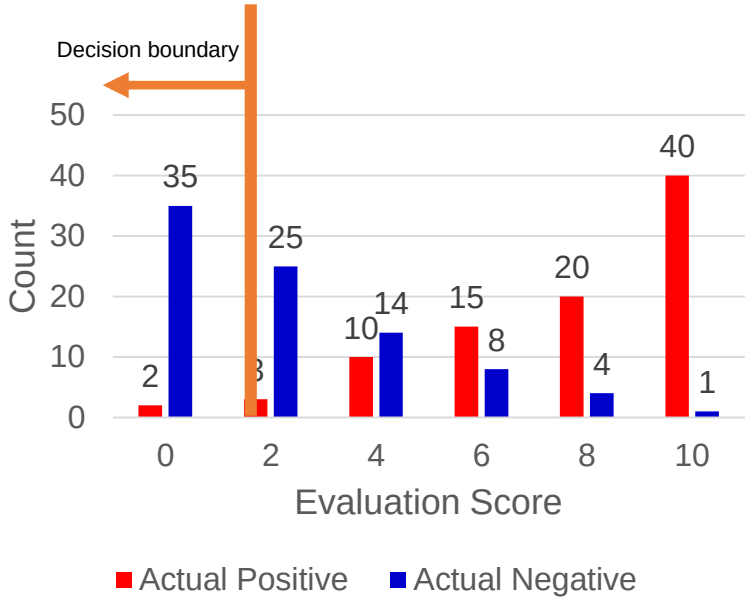




TPR-FPR 커브

- Curves of TPR vs. FPR for moving decision boundary
  - From right to left

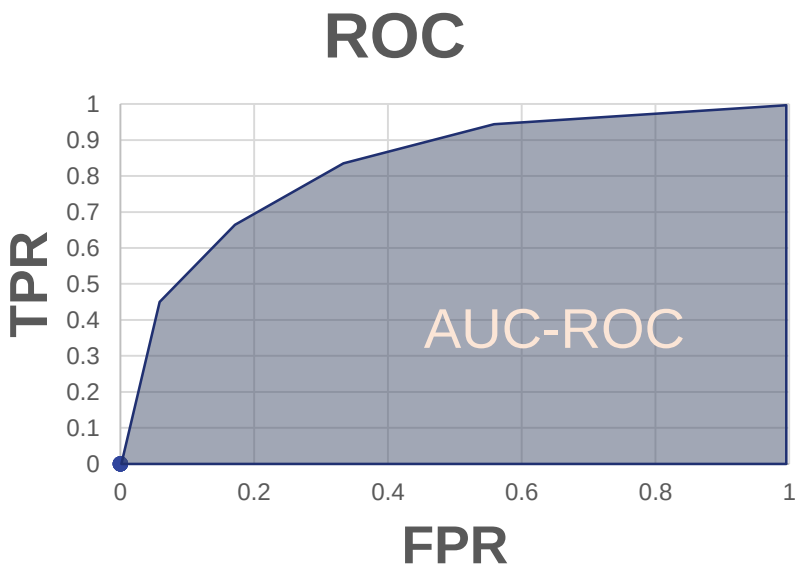




AUC-ROC (or ROC-AUC)

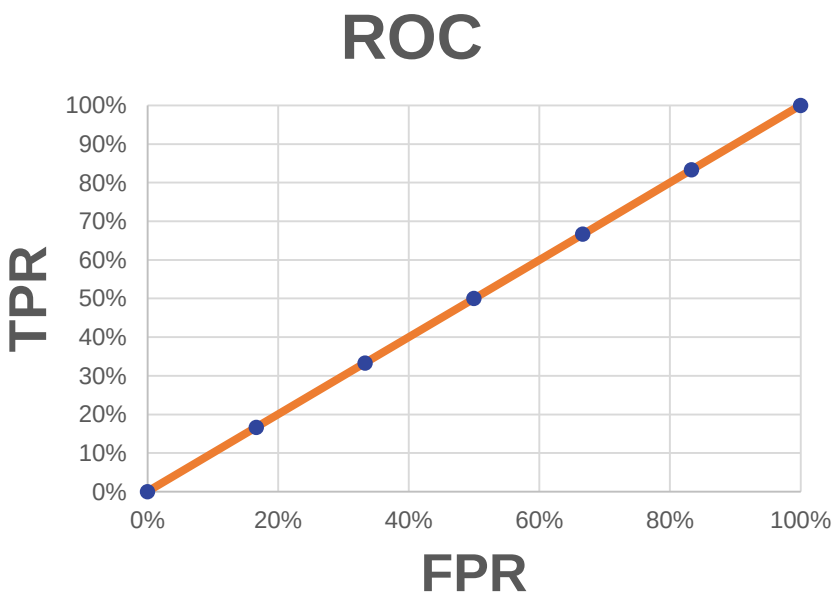
- Receiver Operating Characteristic (ROC)
  - Projection of TPR-FPR for various decision thresholds

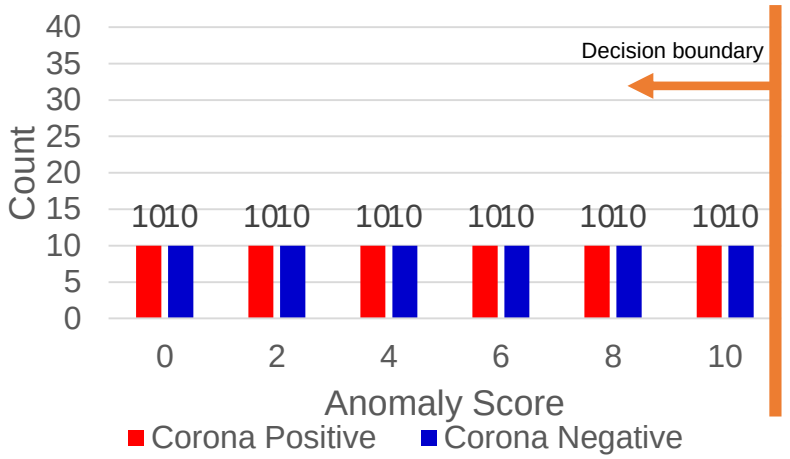
- AUC-ROC
  - Area under the ROC curve
  - Measure of how well the positive/negative distributions are separated



AUC-ROC

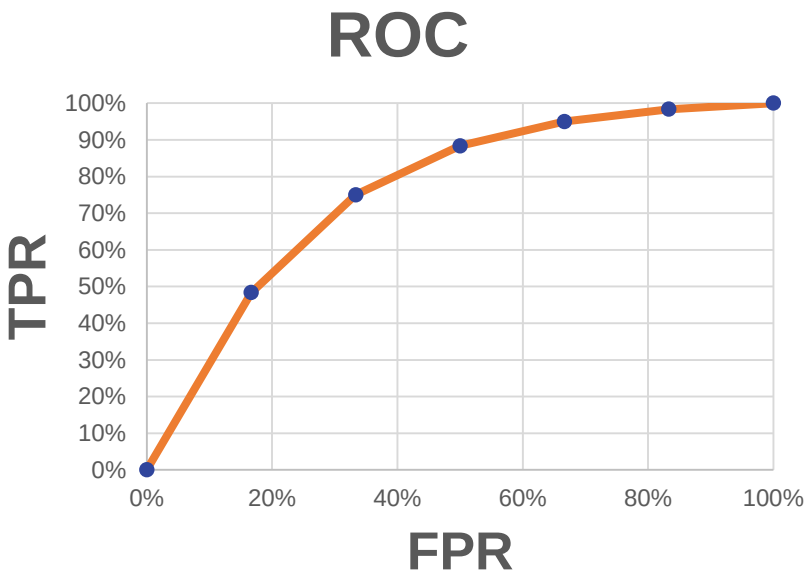
- Suppose that both data has a uniform distribution
  - Distributions are meaningless (50% correct, 50% incorrect)

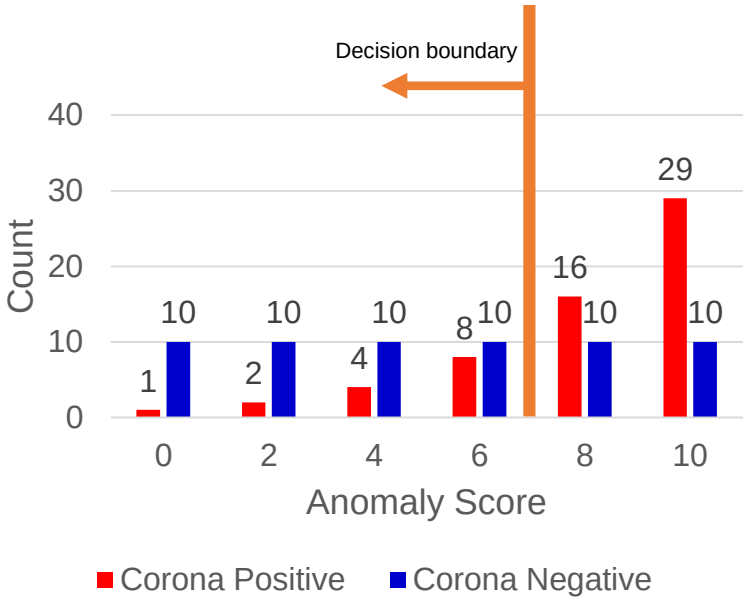




AUC-ROC

- When we have distinguishable distributions
  - Rate of increase in TPR is faster than that of FPR





불균형 데이터 문제

- Example
  - Corona virus kit

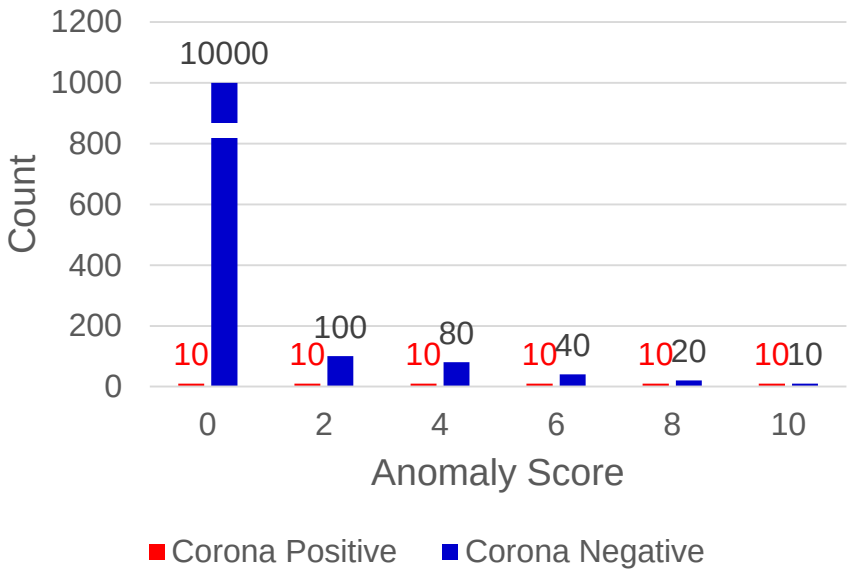
Prediction result

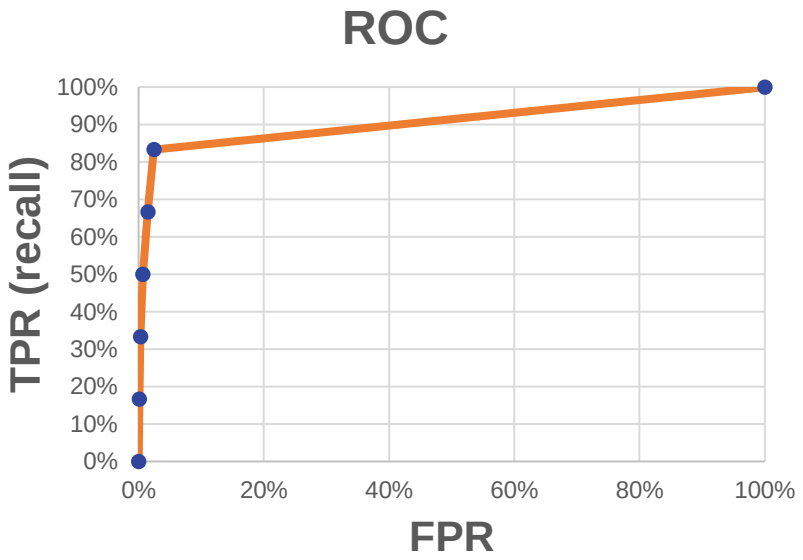
Status of patients	Positive	Negative	Marginal Statistics
Positive	(TP) 20	(FN) 2	22 (infected)
Negative	(FP) 40	(TN) 8000	8040 (uninfected)
Marginal Statistics	60 (predicted as infected)	8002 (predicted as uninfected)	

TPR(Recall, sensitivity): 20/22 = 91% → The kit has good recall performance
FPR: 40/8040 = 0.5% → The kit has good specificity (99.5%)

데이터 불균형 상황의 AUC-ROC

- AUC-ROC doesn't work well with imbalanced data

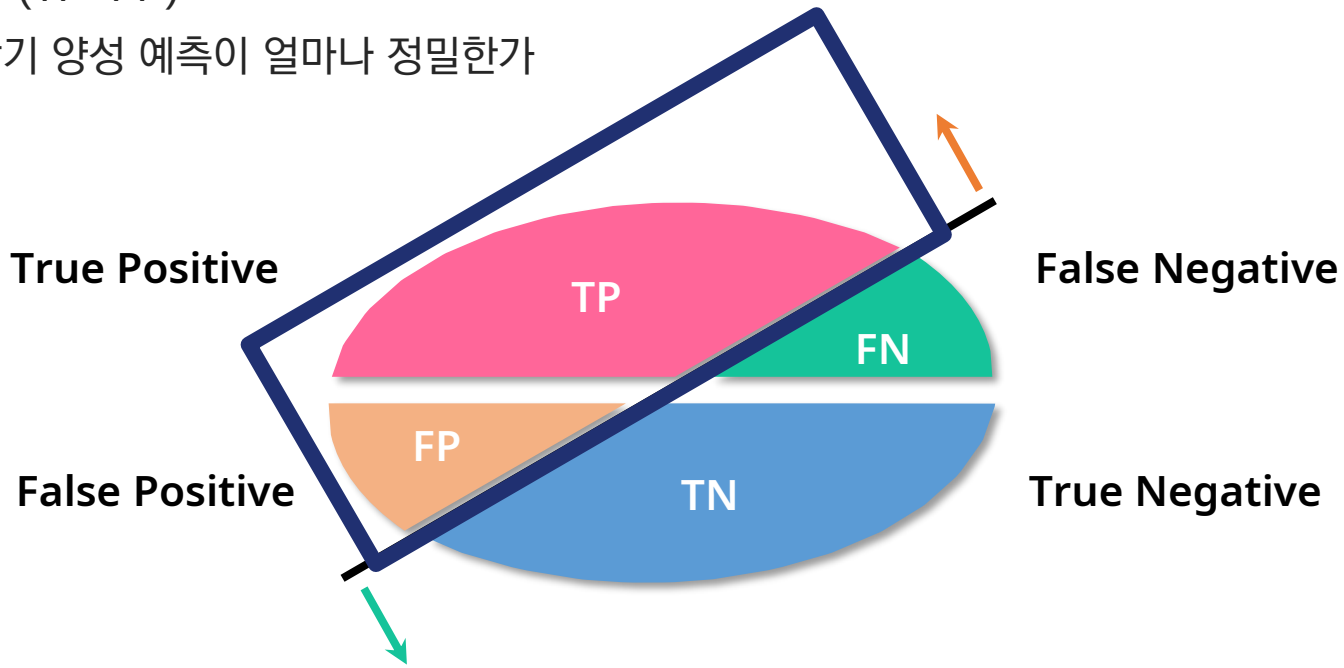




Concentrated distribution of FPR increases AUC-ROC

정밀도

- Precision
  - How many true POSITIVES among my POSITIVE prediction ?
  - TP / (TP+FP)
  - 진단기 양성 예측이 얼마나 정밀한가



데이터 불균형 상황의 정밀도

Prediction result

Status of patients	Positive	Negative	Marginal Statistics
Positive	(TP) 20	(FN) 2	22 (infected)
Negative	(FP) 40	(TN) 8000	8040 (uninfected)
Marginal Statistics	60 (predicted as infected)	8002 (predicted as uninfected)	

Precision: 20/60 = 33% → The kit has bad precision

Antigen Rapid test

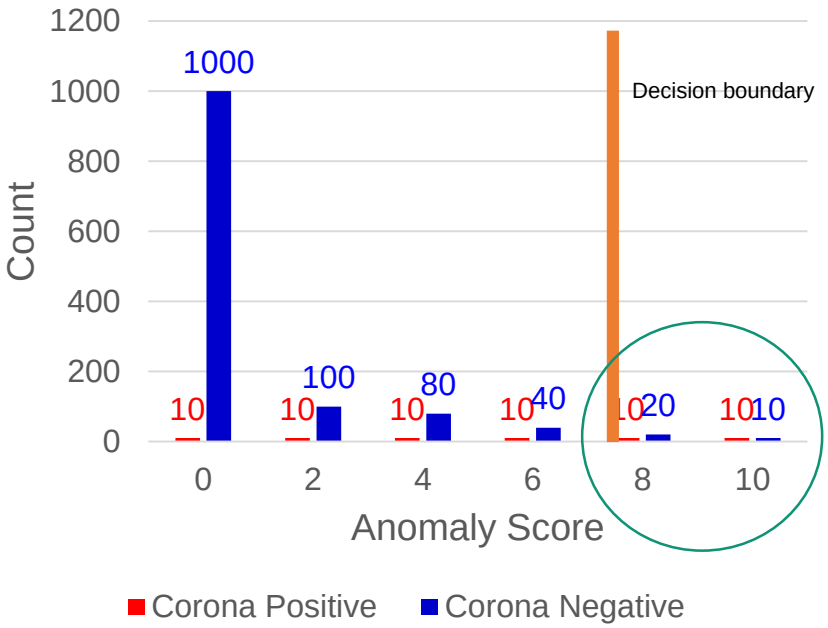
Prediction result

Status of patients	Positive	Negative	Marginal Statistics
Positive	(TP) 20	(FN) 10	30 (infected)
Negative	(FP) 1	(TN) 8000	8001 (uninfected)
Marginal Statistics	21 (predicted as infected)	8010 (predicted as uninfected)	

Precision: 20/21 = 95% → The kit has good precision
Recall: 66% (bad sensitivity)

데이터 불균형 문제의 해결법

- Using ‘Precision’ instead of FPR



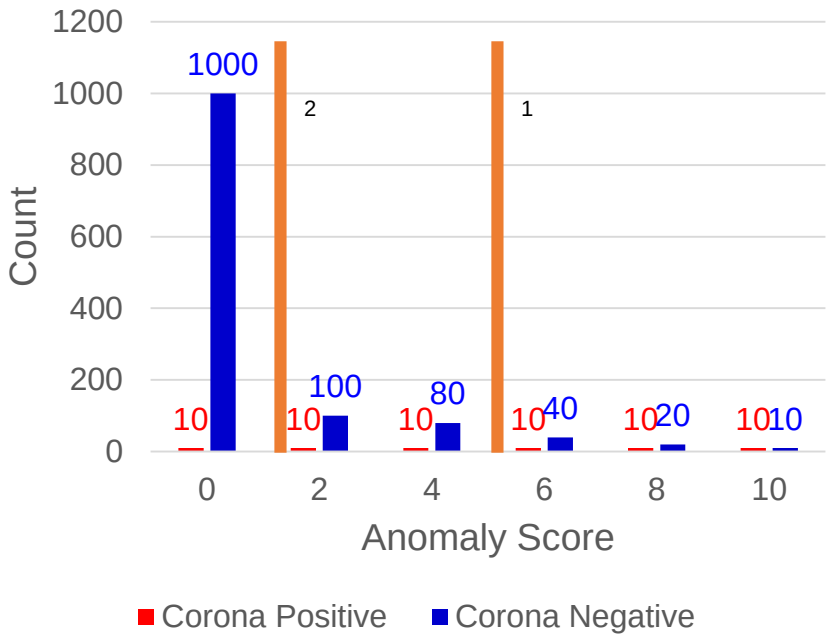
- Precision = TP / (TP + FP) = (10+10)/(10+10+20+10)
- As the decision boundary moves to the left, precision converges to Real (Positive / Negative) ratio.

  P / (P + N)

- Positive/(Positive + Negative) is very small for imbalanced data (huge number of negatives)
- Therefore, precision will decrease with the movement of decision boundary

How to resolve imbalance problem?

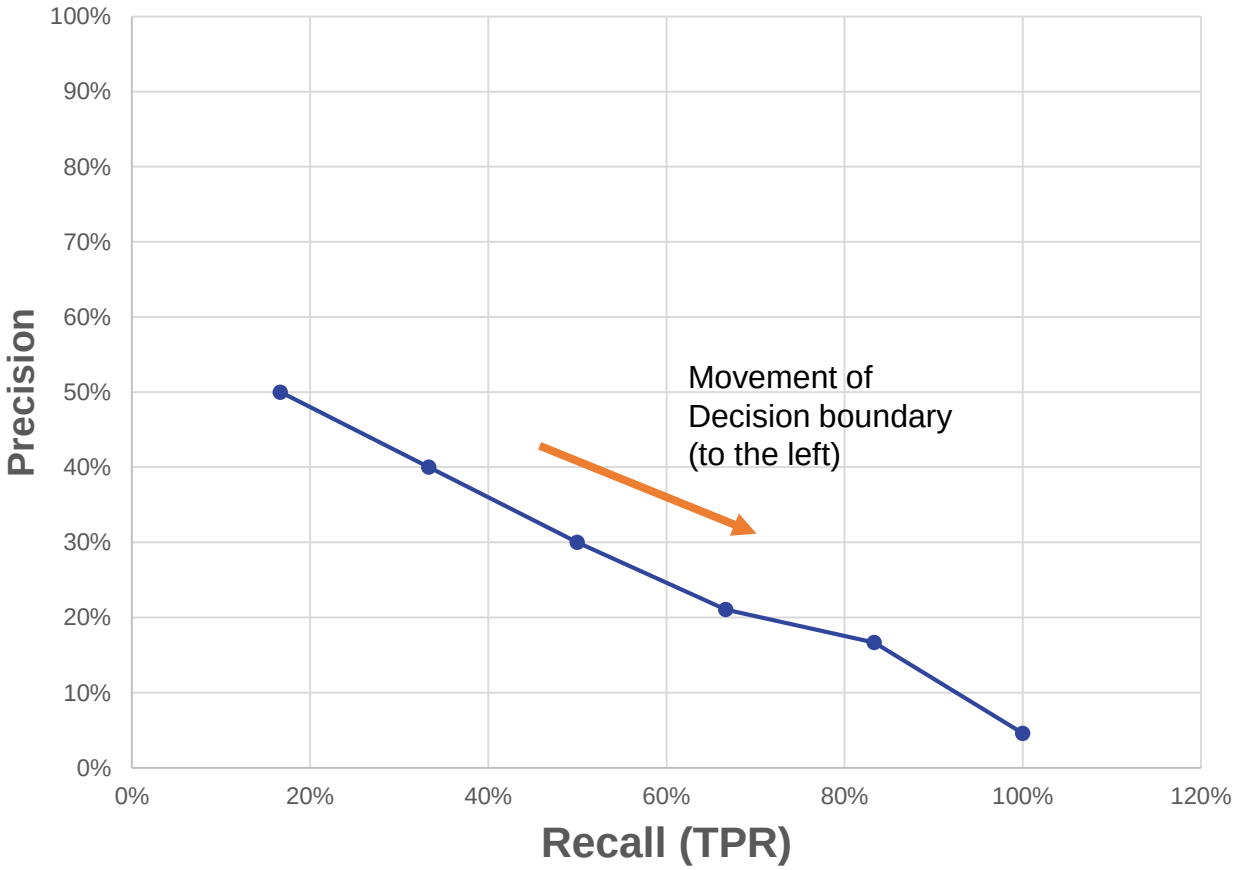
- Using ‘Precision’ instead of FPR



- Example
- Precision 1: (10 + 10 + 10)/(10*3 + 40+20+10)
  = 30/100 = 30%
- Precision 2: (10*5)/(10*5+100+80+40+20+10) = 17%

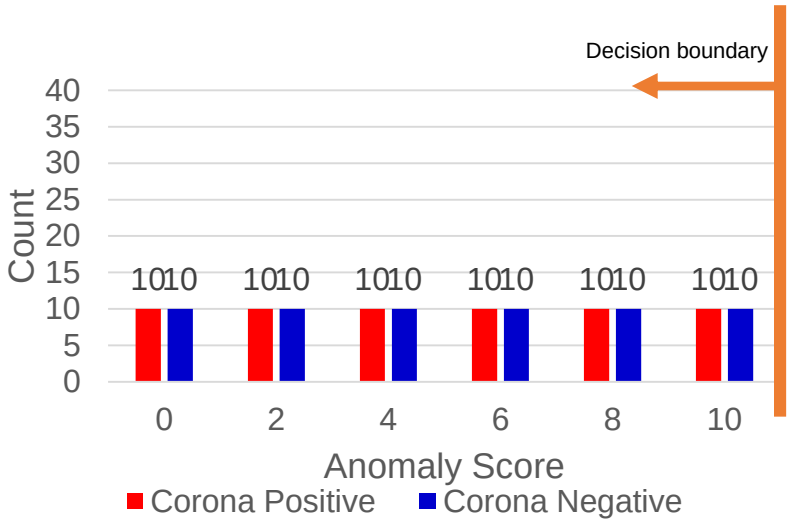
Precision-Recall curve

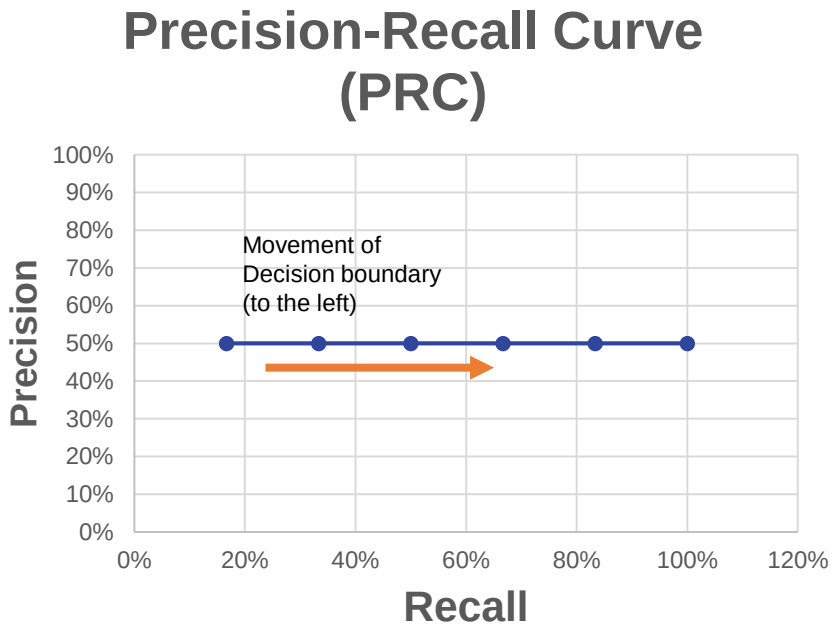
- Alternative to ROC



Precision-Recall curve

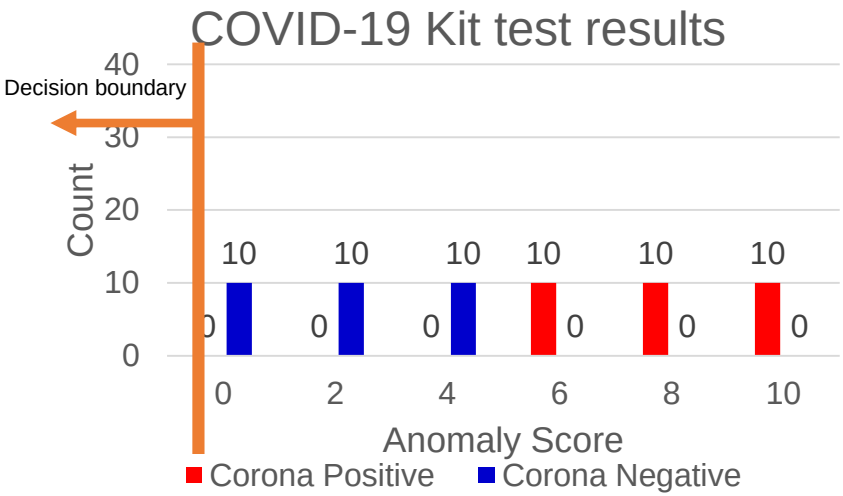
- For uniform distribution

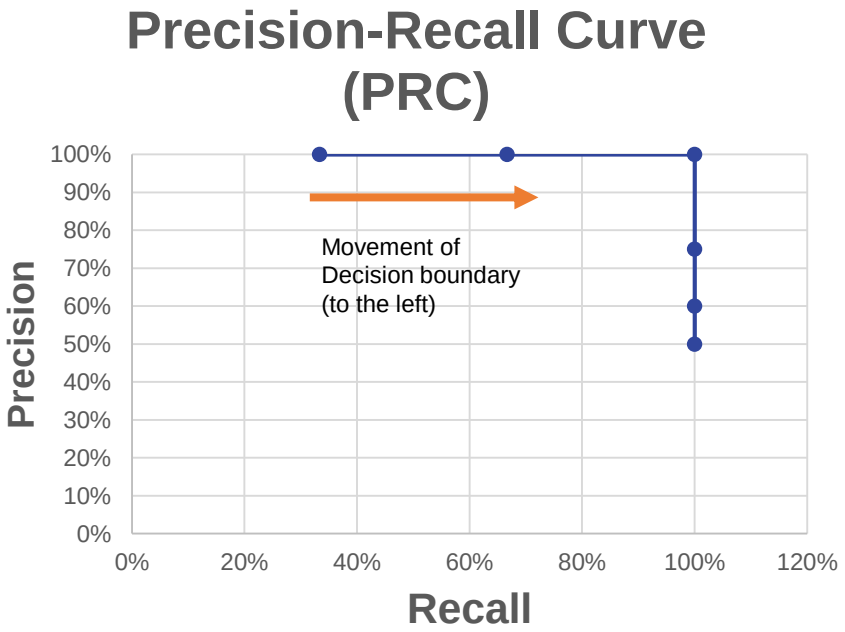




Precision-Recall Curve (PRC)

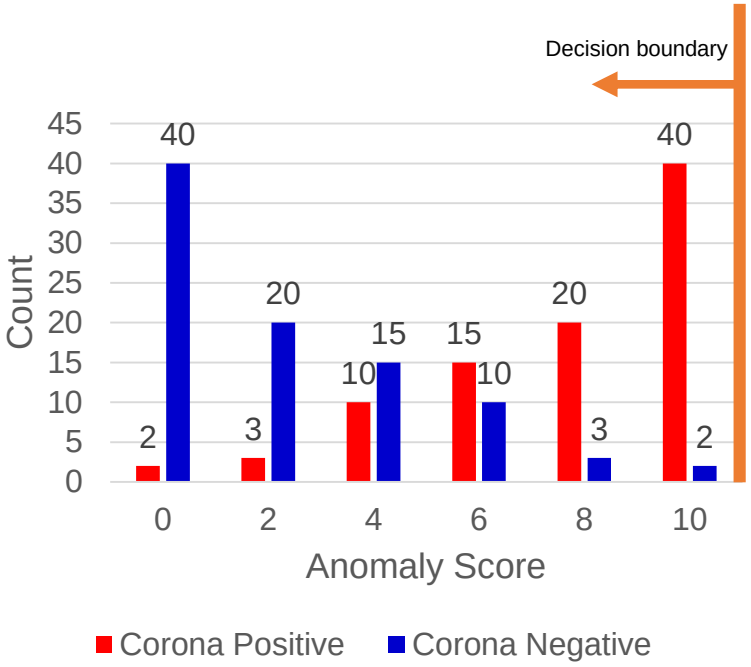
- For perfectly separated distribution

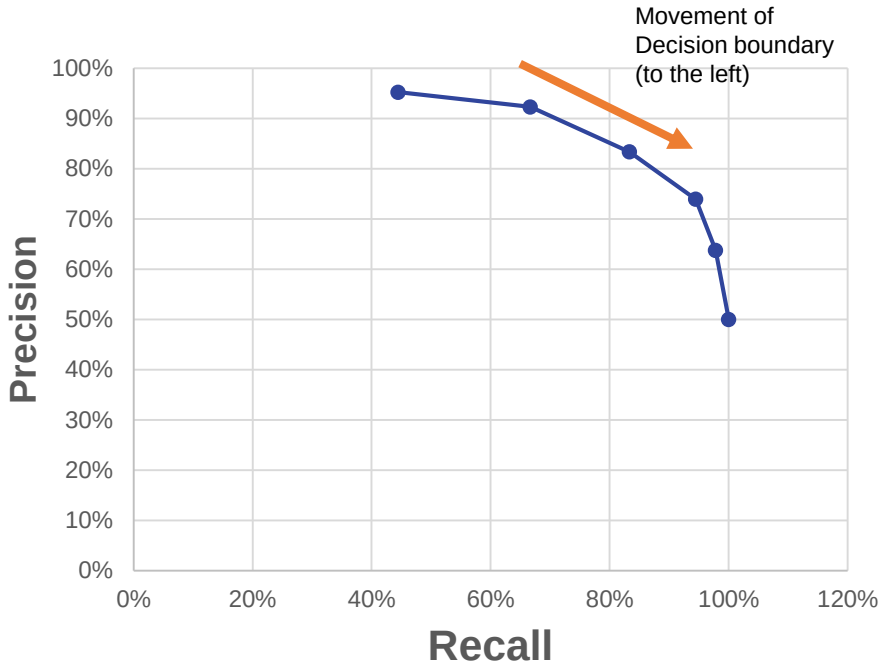




Precision-Recall curve

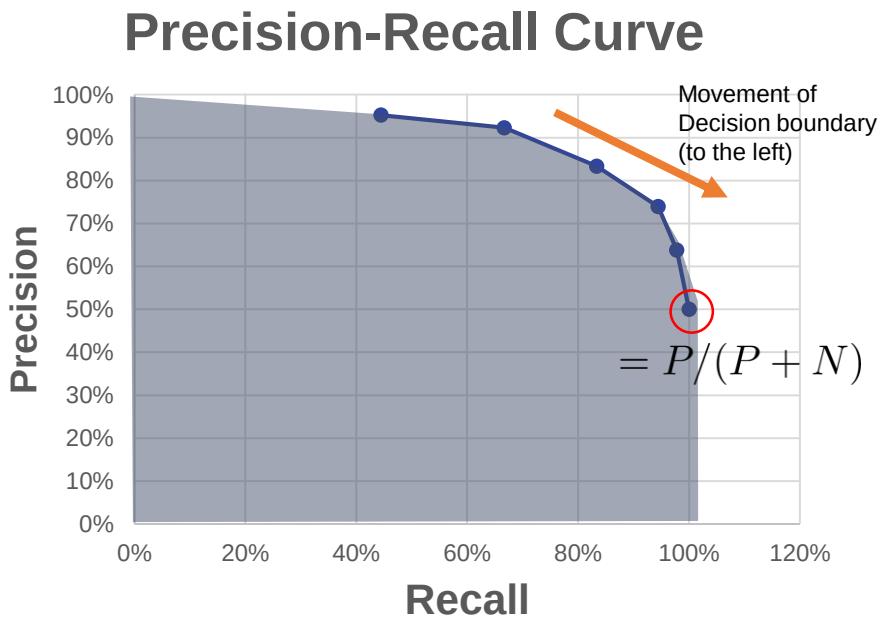
- For good distribution





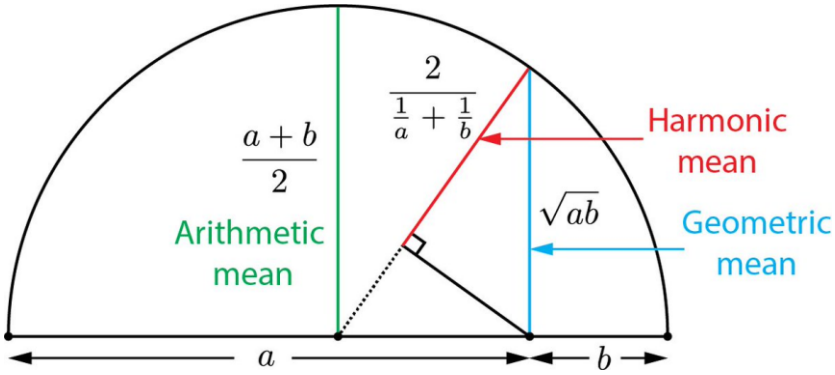
PR-AUC

- Area under the PRC curve
- Advantage
  - Imbalance is reflected in the performance
- Disadvantage
  - Area is mainly influenced by the marginal probability = P/(P+N)



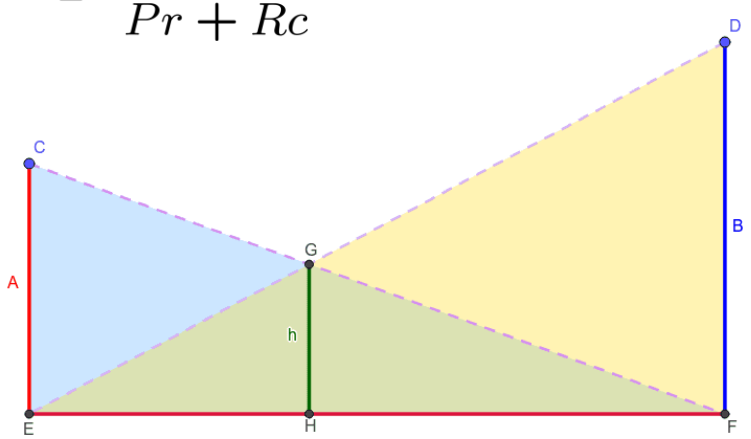
F1-Score

- Precision & Recall 의 조화평균 (harmonic mean)
- Closer to the smaller one among Prec. & Recall



https://www.geogebra.org/m/dpyumqUZ

$$F_1 = 2 \cdot \frac{Pr \cdot Rc}{Pr + Rc}$$



https://www.geogebra.org/m/EM2bxp3A

요약

- **데이터 재현 태스크**
  - DNN is trained to compress/decompress normal images
  - Compress the input image using a few latent variables
  - Decompressed image includes some error compared to the original
  - For anomalous data, the reconstruction error would be greater than normal images' (anomaly score)

- **이상진단 점수와 성능평가 지표**
  - From the histogram of anomaly scores, P and N are separated
  - Decision boundary
  - Quality of distribution determines the model's AD performance
  - ROC-AUC
  - PR-AUC
  - F1-score