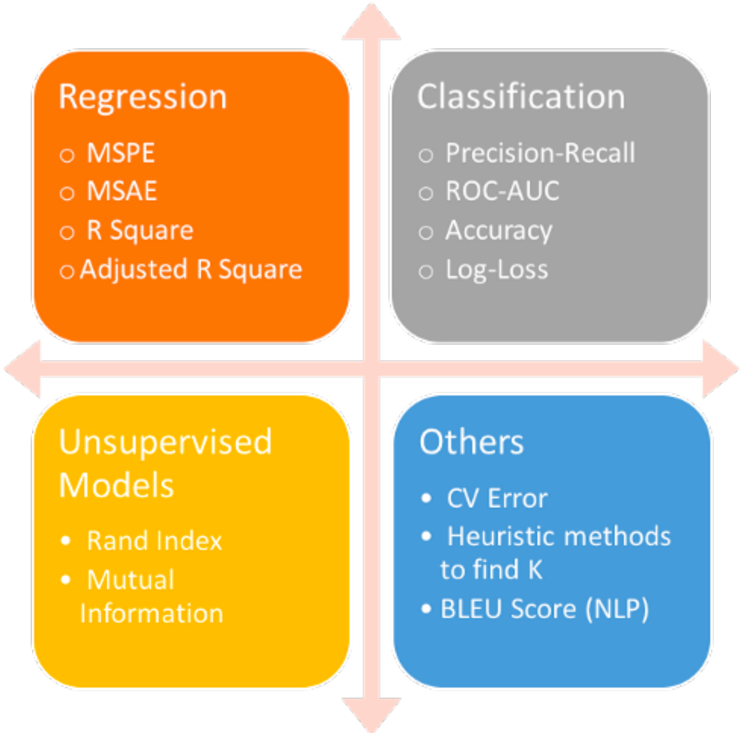
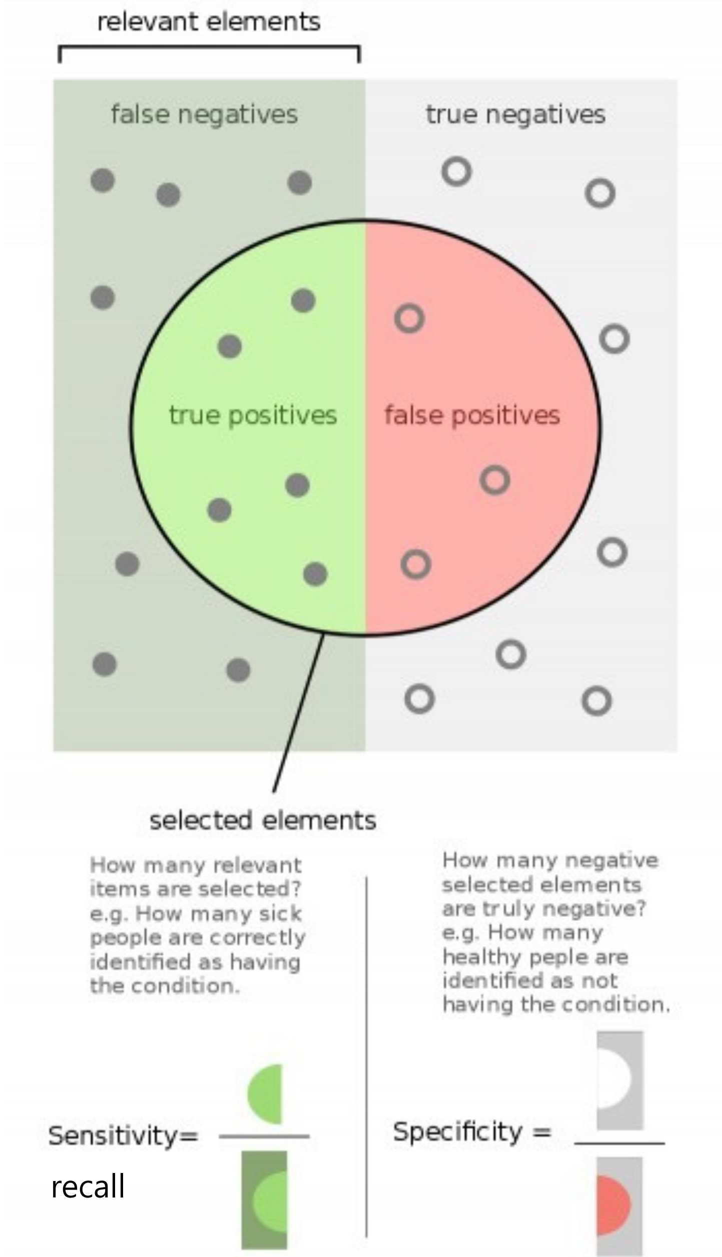
(절대적) 성능 측정은 ML에서 가장 어려운 과제 중 하나



		Predicted	
		True	False
Actual	True	True Positives	False Negatives
	False	False Positives	True Negatives

	Actual = Yes	Actual = No
Predicted = Yes	TP	FP
Predicted = No	FN	TN

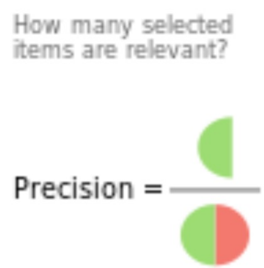


$$\text{Recall} = \frac{\boxed{\mathrm{TP}}}{\boxed{\mathrm{TP}} + \boxed{\mathrm{FN}}}$$

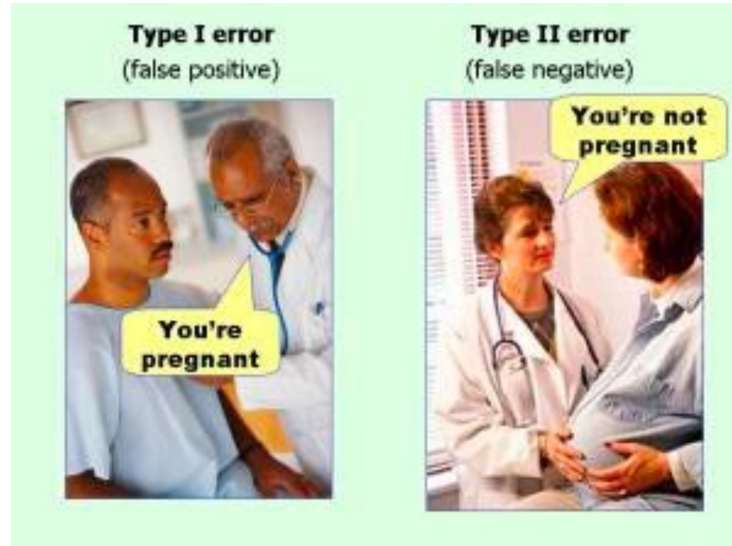
$$\text{Precision} = \frac{\boxed{\mathrm{TP}}}{\boxed{\mathrm{TP}} + \boxed{\mathrm{FP}}}$$

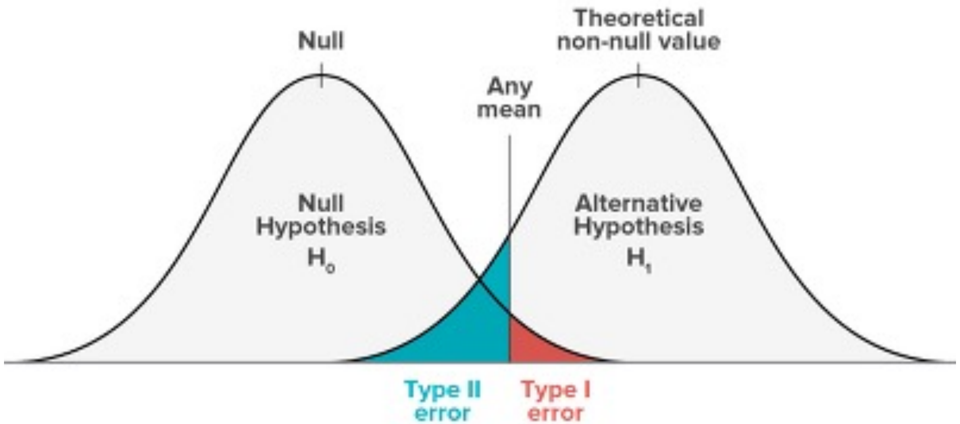
$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

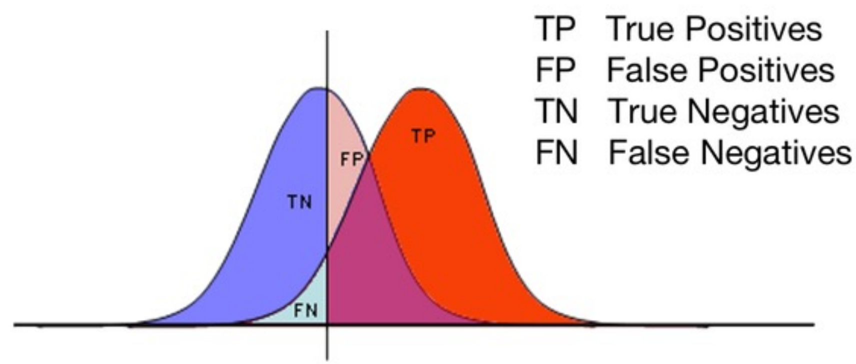
$$= \frac{\boxed{\mathrm{TP}}}{\boxed{\mathrm{TP}} + \frac{1}{2}\left(\boxed{\mathrm{FP}} + \boxed{\mathrm{FN}}\right)}$$



		Predicted Class		
		Positive	**Negative**	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$ recall
	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** $\frac{TP+TN}{(TP+TN+FP+FN)}$







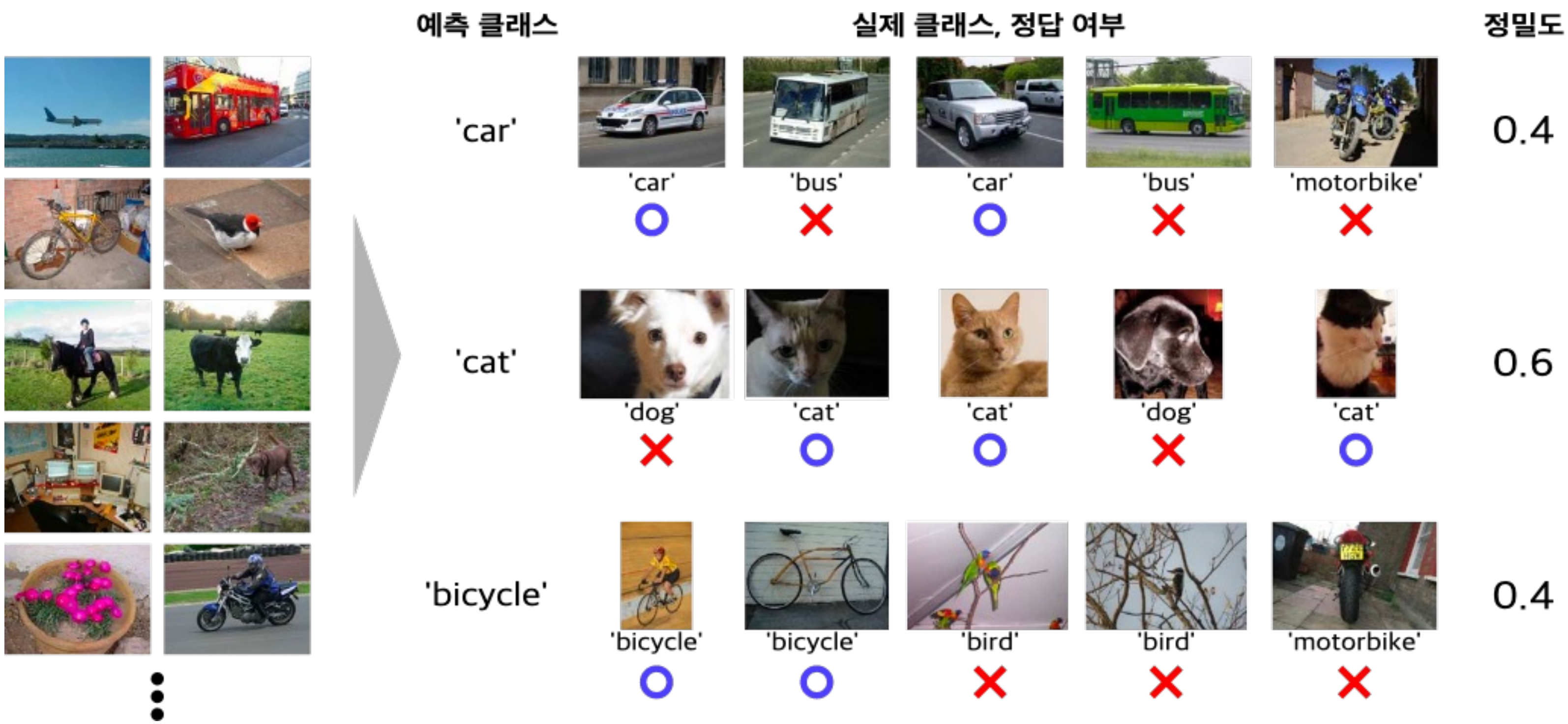
$$\text{Accuracy} = \frac{\text{올바르게 분류한 이미지 수}}{\text{전체 이미지 수}}$$

	예측 클래스	실제 클래스	정답 여부
	'aeroplane'	'aeroplane'	O
	'car'	'bicycle'	X
	'horse'	'horse'	O
	'train'	'TV/monitor'	X
	'potted plant'	'potted plant'	O
	'bus'	'bus'	O
	'bird'	'bird'	O
	'cow'	'cow'	O
	'dog'	'dog'	O
	'bicycle'	'motorbike'	X

$$\text{Precision c} = \frac{\text{올바르게 분류한 클래스 c 이미지 수}}{\text{클래스 c일 것으로 예측한 이미지 수}}$$

$$\text{Recall c} = \frac{\text{올바르게 분류한 클래스 c 이미지 수}}{\text{전체 클래스 c 이미지 수}}$$

Sensitivity, hit rate



평균 정밀도

(0.4+0.6+0.4)/3=0.47(47%)

평균 재현율은 (0.6+1.0+0.8)/3=0.8(80%)

- **Accuracy**
  - Percentage of total items classified correctly- (TP+TN)/(N+P)
- **Recall = Sensitivity = TPR (True Positive Rate)**
  - Number of items correctly identified as positive out of total true positives- TP/(TP+FN)
- **Specificity = TNR (True Negative Rate)**
  - Number of items correctly identified as negative out of total negatives- TN/(TN+FP)
- **Precision**
  - Number of items correctly identified as positive out of total items identified as positive- TP/(TP+FP)
- **False Positive Rate = Type I Error = Fall-out = 1 - specificity**
  - Number of items wrongly identified as positive out of total true negatives- FP/(FP+ TN)
- **False Negative Rate = Type II Error**
  - Number of items wrongly identified as negative out of total true positives- FN/(FN+TP)
- **F1 Score**
  - It is a harmonic mean of precision and recall given by- F1 = 2*Precision*Recall/(Precision + Recall)

microaveraging

		gold labels			
		urgent	normal	spam	
system output	urgent	8	10	1	**precision**$_u = \frac{8}{8+10+1}$
	normal	5	60	50	**precision**$_n = \frac{60}{5+60+50}$
	spam	3	30	200	**precision**$_s = \frac{200}{3+30+200}$
		recall$_u = \frac{8}{8+5+3}$	**recall**$_n = \frac{60}{10+60+30}$	**recall**$_s = \frac{200}{1+50+200}$	

macroaveraging

Class 1: Urgent

	true urgent	true not
system urgent	8	11
system not	8	340

$$\text{precision} = \frac{8}{8+11} = .42$$

Class 2: Normal

	true normal	true not
system normal	60	55
system not	40	212

$$\text{precision} = \frac{60}{60+55} = .52$$

Class 3: Spam

	true spam	true not
system spam	200	33
system not	51	83

$$\text{precision} = \frac{200}{200+33} = .86$$

Pooled

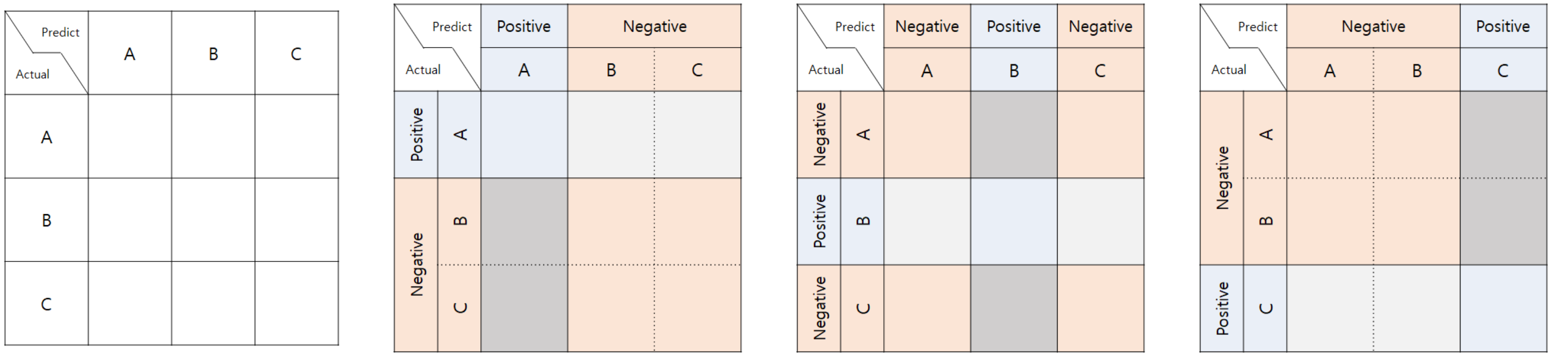
	true yes	true no
system yes	268	99
system no	99	635

$$\begin{array}{c}\text{microaverage}\\ \text{precision}\end{array} = \frac{268}{268+99} = \mathbf{.73}$$

$$\begin{array}{c}\text{macroaverage}\\ \text{precision}\end{array} = \frac{.42+.52+.86}{3} = \mathbf{.60}$$

A **microaverage** is dominated by the more frequent class, since the counts are pooled. The **macroaverage** better reflects the statistics of the smaller classes, and so is more appropriate when performance on all the classes is equally important.

One vs Rest (=> multi class)



TP	FN
FP	TN

$$\mathrm{TP}_{\mathrm{total}} = \sum_{i=1}^{c} \mathrm{TP}_i$$

$$\mathrm{FN}_{\mathrm{total}} = \sum_{i=1}^{c} \mathrm{FN}_i$$

$$\mathrm{FP}_{\mathrm{total}} = \sum_{i=1}^{c} \mathrm{FP}_i$$

$$\mathrm{TN}_{\mathrm{total}} = \sum_{i=1}^{c} \mathrm{TN}_i$$

predictions
(output)

actual
class
(input)

	A	B	C	D
A	9	1	0	0
B	1	15	3	1
C	5	0	24	1
D	0	4	1	15

TP(True Positive)

	A	B	C	D
A	9	1	0	0
B	1	15	3	1
C	5	0	24	1
D	0	4	1	15

TN_A(True Negative for A)

	A	B	C	D
A	9	1	0	0
B	1	15	3	1
C	5	0	24	1
D	0	4	1	15

TN_D(True Negative for D)

	A	B	C	D
A	9	1	0	0
B	1	15	3	1
C	5	0	24	1
D	0	4	1	15

FP_A(False Positive for A)

	A	B	C	D
A	9	1	0	0
B	1	15	3	1
C	5	0	24	1
D	0	4	1	15

FP_B(False Positive for B)

	A	B	C	D
A	9	1	0	0
B	1	15	3	1
C	5	0	24	1
D	0	4	1	15

FN_A(False Negative for A)

	A	B	C	D
A	9	1	0	0
B	1	15	3	1
C	5	0	24	1
D	0	4	1	15

FN_D(False Negative for D)

balanced

Balanced Data

Model 1

predictions (output) →

actual class (input) ↓	A	B	C	D
A	10	0	0	0
B	0	5	3	2
C	0	1	8	1
D	0	1	0	9

Accuracy = sum(TP) / total data set #
(10+5+8+9) / 40 = 0.8

Model 2

predictions (output) →

actual class (input) ↓	A	B	C	D
A	8	2	0	0
B	1	7	0	2
C	0	0	9	1
D	2	3	0	5

Accuracy = sum(TP) / total data set #
(8+7+9+5) / 40 = 0.725

imbalanced

Accuracy = sum(TP)/Total Dataset #
Precision = TP/(TP+FP)
(macro)average precision = sum(precision) / (the number of classes)
Recall = TP/(TP+FN)
(macro)average recall = sum(recall) / (the number of classes)
F1 Score = 2 * ((precision * recall) / (precision+recall))

Model 1

predictions (output)

actual class (input)

	A	B	C	D
A	100	80	10	10
B	0	9	0	1
C	0	1	8	1
D	0	1	0	9

Precision

Recall

Accuracy = (100+9+8+9) / 230 = 0.547

- Precision A =100 / (100+0) = 1
- Precision B = 9 / (9+82) = 9/91
- Precisian C = 8 / (8+10) = 8/18
- Precision D = 9 / (9+12) = 9/21

average precision = 0.492

- recall A =100 / (100+100) = 100/200
- recall B = 9 / (9+1) = 9/10
- recall C = 8 / (8+2) = 8/10
- recall D = 9 / (9+1) = 9/10

average recall = 0.775

F1 Score = 2 * ((0.492*0.775) / (0.492+0.775)) = 0.601

Model 2

predictions (output)

actual class (input)

	A	B	C	D
A	198	2	0	0
B	7	1	0	2
C	0	8	1	1
D	2	3	4	1

Accuracy = (198+1+1+1) / 230 = 0.87

- Precision A =198 / (198+9) = 198/207
- Precisian B = 1 / (1+13) = 1/14
- Precisian C = 1 / (1+4) = 1/5
- Precision D = 1 / (1+3) = 1/4

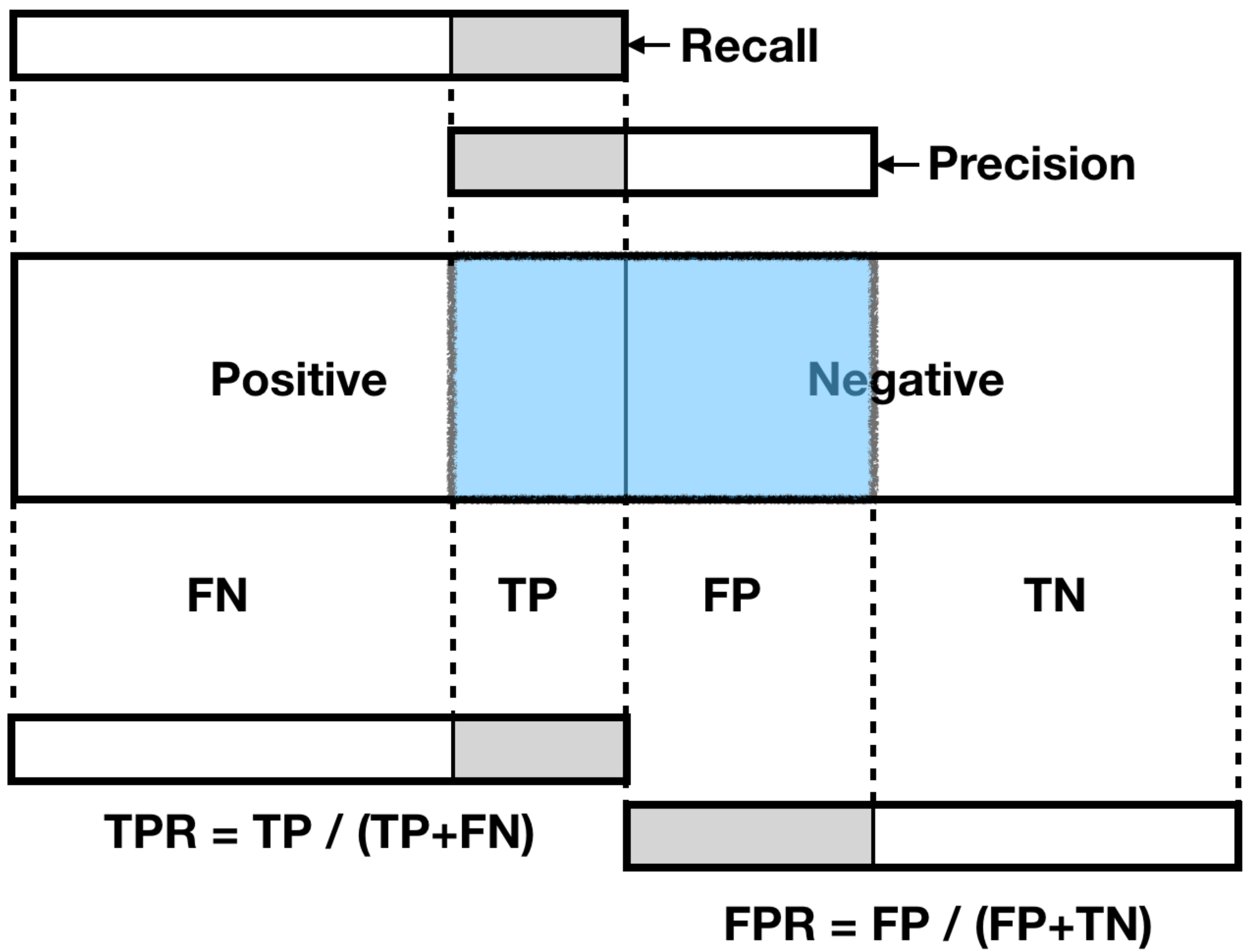
average precision = 0.369

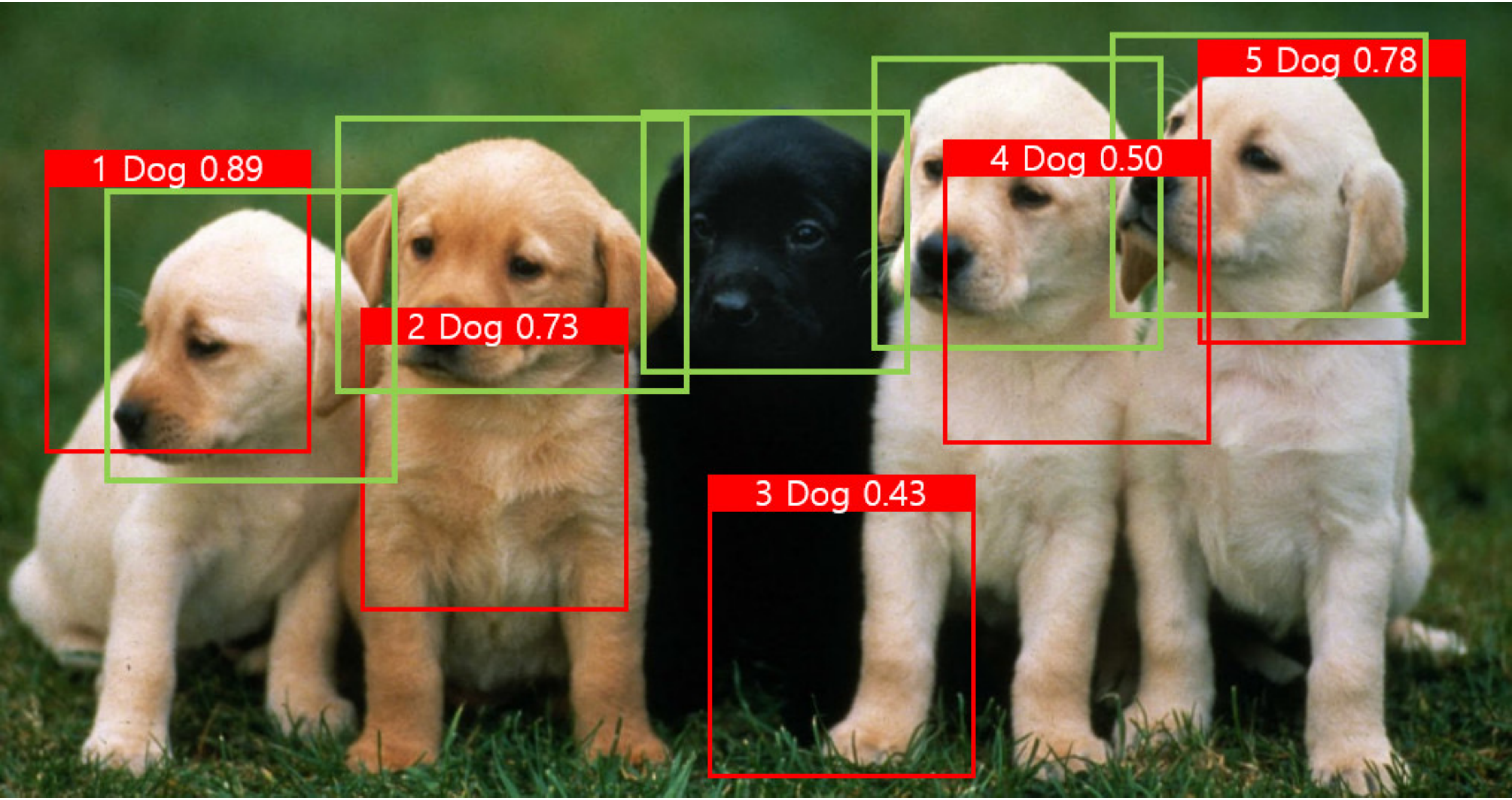
- recall A =198 / (198+2) = 198/200
- recall B = 1 / (1+9) = 1/10
- recall C = 1 / (1+9) = 1/10
- recall D = 1 / (1+9) = 1/10

average precision = 0.323

F1 Score = 2 * ((0.492*0.775) / (0.492+0.775)) = 0.344

Accuracy에 의해 모델을 선택하면 Model2가 더 좋으나 A에만 좋은 모델 B,C,D에 대한 예측은 떨어지는 문제 발생,
Imbalanced data인 경우에는 F1 Score를 평가 지표로 이용


Recall
Precision
Positive
Negative
FN
TP
FP
TN
TPR = TP / (TP+FN)
FPR = FP / (FP+TN)

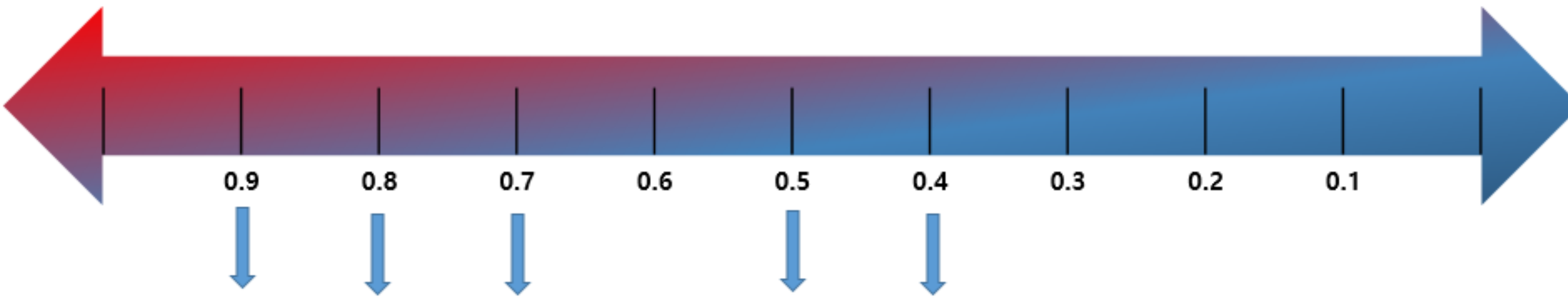

1 Dog 0.89
2 Dog 0.73
3 Dog 0.43
4 Dog 0.50
5 Dog 0.78

Intersection over unit (IOU)

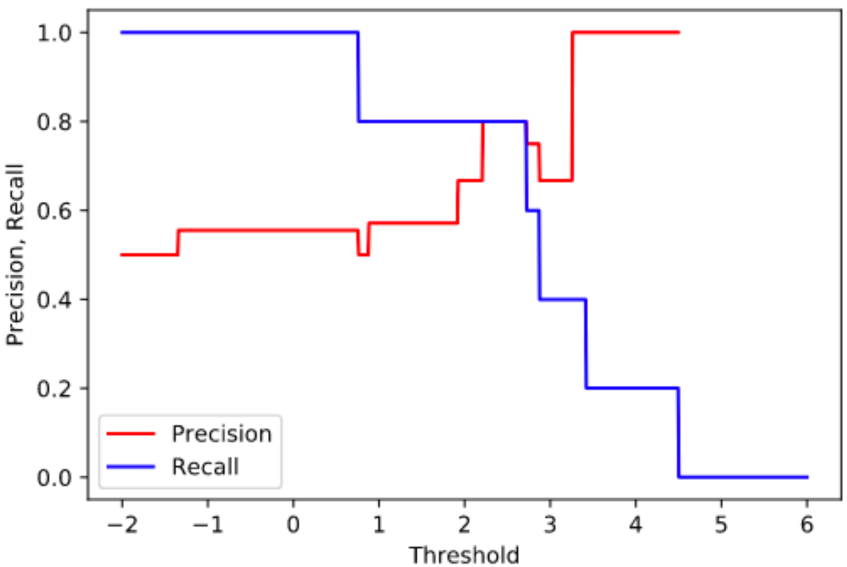
$$IoU = \frac{\text{area of overlap}}{\text{area of union}}$$

Precision/Recall Trade-off

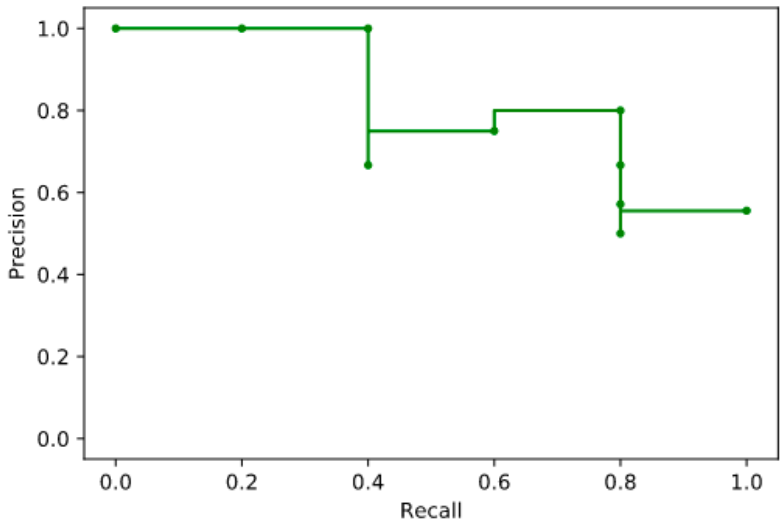
Confidence score Threshold



Bounding box	0개	1개	3개	4개	5개
Box index	X	1	1,2,5	1,2,4,5	1,2,3,4,5
Precision	0	1/1=1	3/3=1	¾=0.75	3/5=0.6
Recall	0	1/5=0.2	3/5=0.6	5/4=0.8	5/5=1



1.0
0.8
0.6
0.4
0.2
0.0
Precision, Recall
−2
−1
0
1
2
3
4
5
6
Threshold
Precision
Recall



1.0
0.8
0.6
0.4
0.2
0.0
Precision
0.0
0.2
0.4
0.6
0.8
1.0
Recall

F1 Score

No	Actual	Predicted	Match
1	Airplane	Airplane	✓
2	Car	Boat	X
3	Car	Car	✓
4	Car	Car	✓
5	Car	Boat	X
6	Airplane	Boat	X
7	Boat	Boat	✓
8	Car	Airplane	X
9	Airplane	Airplane	✓
10	Car	Car	✓

imbalanced

		Predicted		
		Airplane	Boat	Car
Actual	Airplane	2	1	0
	Boat	0	1	0
	Car	1	2	3

Label	True Positive (TP)	False Positive (FP)	False Negative (FN)
Airplane	2	1	1
Boat	1	3	0
Car	3	0	3

One-vs-Rest (OvR)

Label	True Positive (TP)	False Positive (FP)	False Negative (FN)	Precision	Recall	F1 Score
Airplane	2	1	1	0.67	0.67	2 * (0.67 * 0.67) / (0.67 + 0.67) = **0.67**
Boat	1	3	0	0.25	1.00	2*(0.25 * 1.00) / (0.25 + 1.00) = **0.40**
Car	3	0	3	1.00	0.50	2 * (1.00 * 0.50) / (1.00 + 0.50) = **0.67**

<table>
<tr><th>Label</th><th>Per-Class F1 Score</th><th>Macro-Averaged F1 Score</th></tr>
<tr><td>Airplane</td><td>0.67</td><td rowspan="3">$\frac{0.67 + 0.40 + 0.67}{3}$
= 0.58</td></tr>
<tr><td>Boat</td><td>0.40</td></tr>
<tr><td>Car</td><td>0.67</td></tr>
</table>

Label	Per-Class F1 Score	Support	Support Proportion	Weighted Average F1 Score
Airplane	0.67	3	0.3	$(0.67 * 0.3) + (0.40 * 0.1) + (0.67 * 0.6)$ **= 0.64**
Boat	0.40	1	0.1	
Car	0.67	6	0.6	
Total	-	10	1.0	

Label	True Positive (TP)	False Positive (FP)	False Negative (FN)	Micro-Averaged F1 Score
Airplane	2	1	1	$\frac{\text{TP}}{\text{TP}+\frac{1}{2}(\text{FP}+\text{FN})} = \frac{6}{6+\frac{1}{2}(4+4)}$ $= 0.60$
Boat	1	3	0	
Car	3	0	3	
TOTAL	**6**	**4**	**4**	

Label	True Positive (TP)	False Positive (FP)	False Negative (FN)	Micro-Averaged Values
Airplane	2	1	1	Precision = $\frac{6}{6+4}$ = **0.60**
Boat	1	3	0	
Car	3	0	3	Recall = $\frac{6}{6+4}$ = **0.60**
TOTAL	**6**	**4**	**4**	F1 Score = $\frac{6}{6+\frac{1}{2}(4+4)}$ = **0.60**

micro-F1

= accuracy = micro-precision = micro-recall

	precision	recall	f1-score	support
Aeroplane	0.67	0.67	0.67	3
Boat	0.25	1.00	0.40	1
Car	1.00	0.50	0.67	6
accuracy			0.60	10
macro avg	0.64	0.72	0.58	10
weighted avg	0.82	0.60	0.64	10

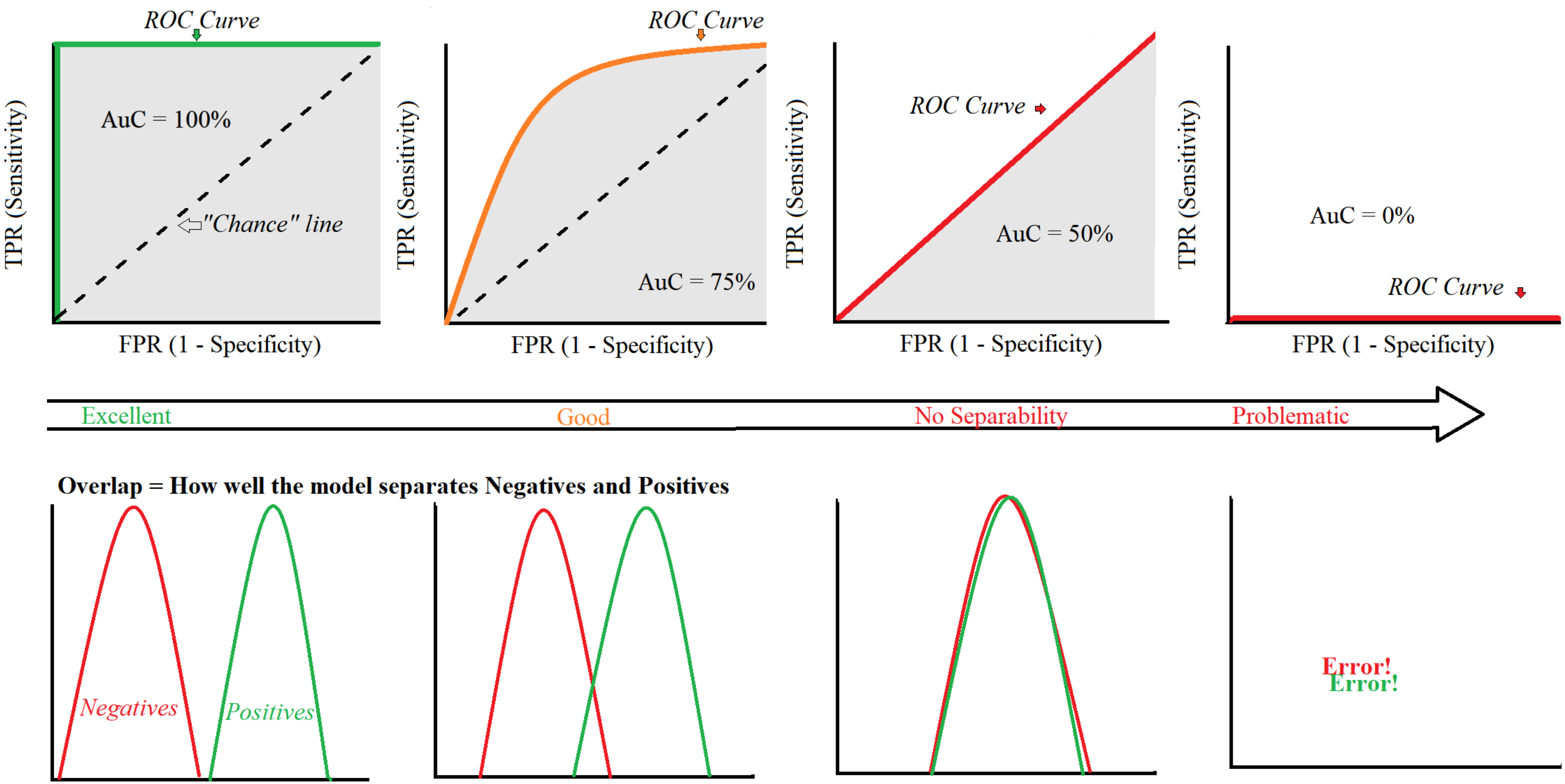
Per-Class F1 scores

Average F1 scores

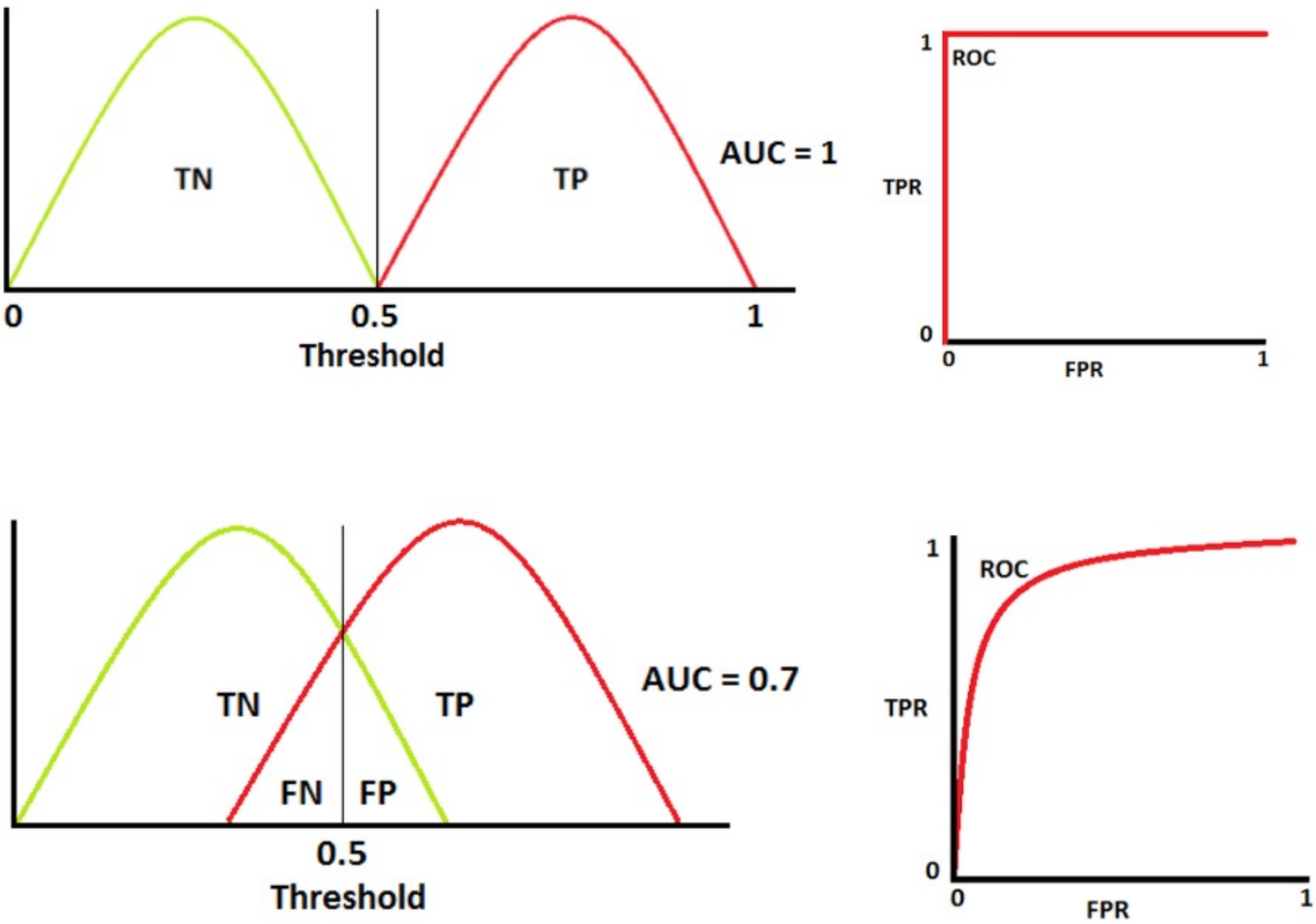
- In general, if you are working with an imbalanced dataset where all classes are equally important, using the **macro** average would be a good choice as it treats all classes equally.
- If you have an imbalanced dataset but want to assign greater contribution to classes with more examples in the dataset, then the **weighted** average is preferred. This is because, in weighted averaging, the contribution of each class to the F1 average is weighted by its size.
- Suppose you have a balanced dataset and want an easily understandable metric for overall performance regardless of the class. In that case, you can go with accuracy, which is essentially our **micro** F1 score.

ROC (Receiver Operating Characteristic)

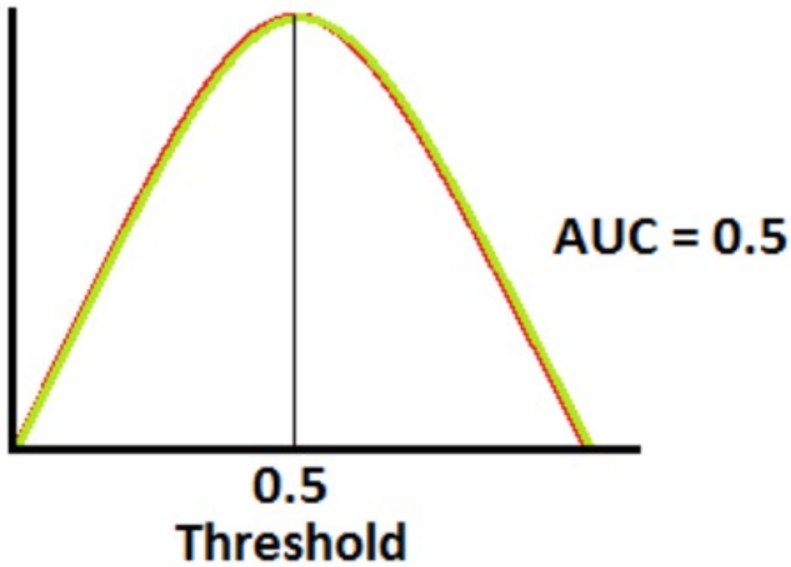
AUC (Area Under the Curve)

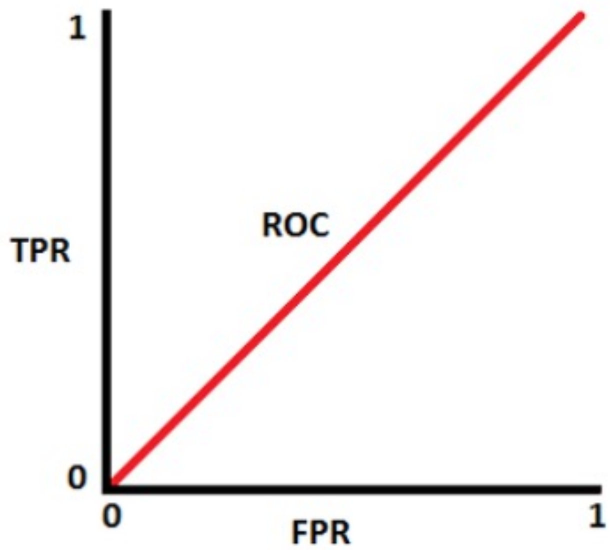


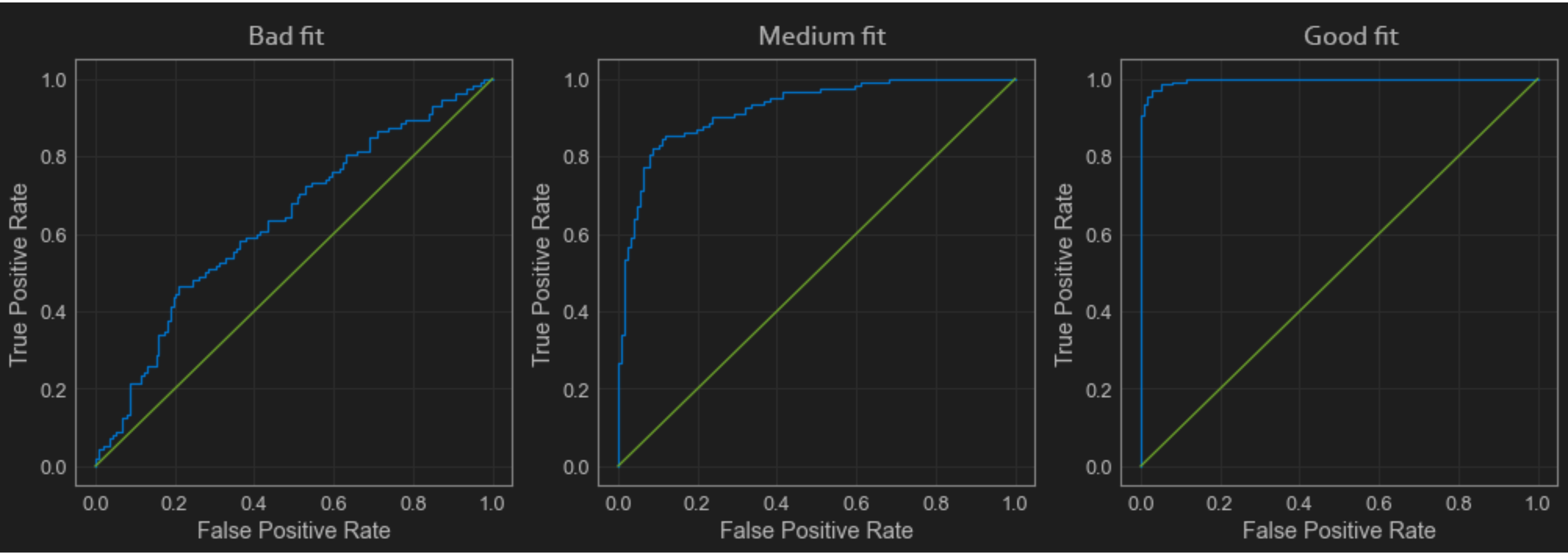
[What's WRONG with Metrics?. For any kind of machine learning… | by Amine Aoullay | Towards Data Science]

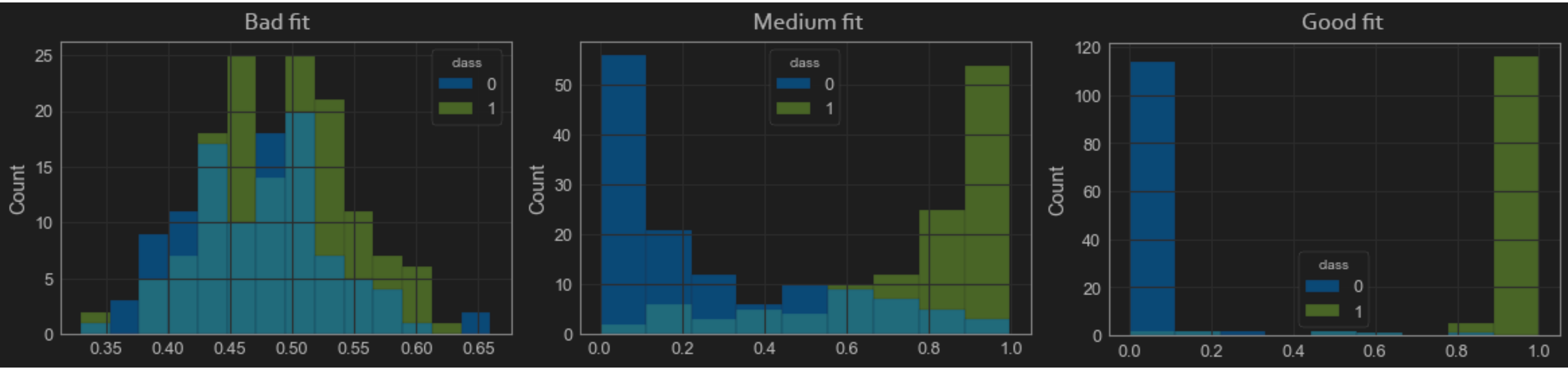

TN
TP
AUC = 1
0
0.5
1
Threshold
1
ROC
TPR
0
0
FPR
1
TN
TP
FN
FP
AUC = 0.7
0.5
Threshold
1
ROC
TPR
0
0
FPR
1

?


AUC = 0.5
0.5
Threshold


1
TPR
ROC
0
0
FPR
1


Bad fit
Medium fit
Good fit
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


Bad fit
Medium fit
Good fit
Count
class
0
1

Object Detection

imbalanced