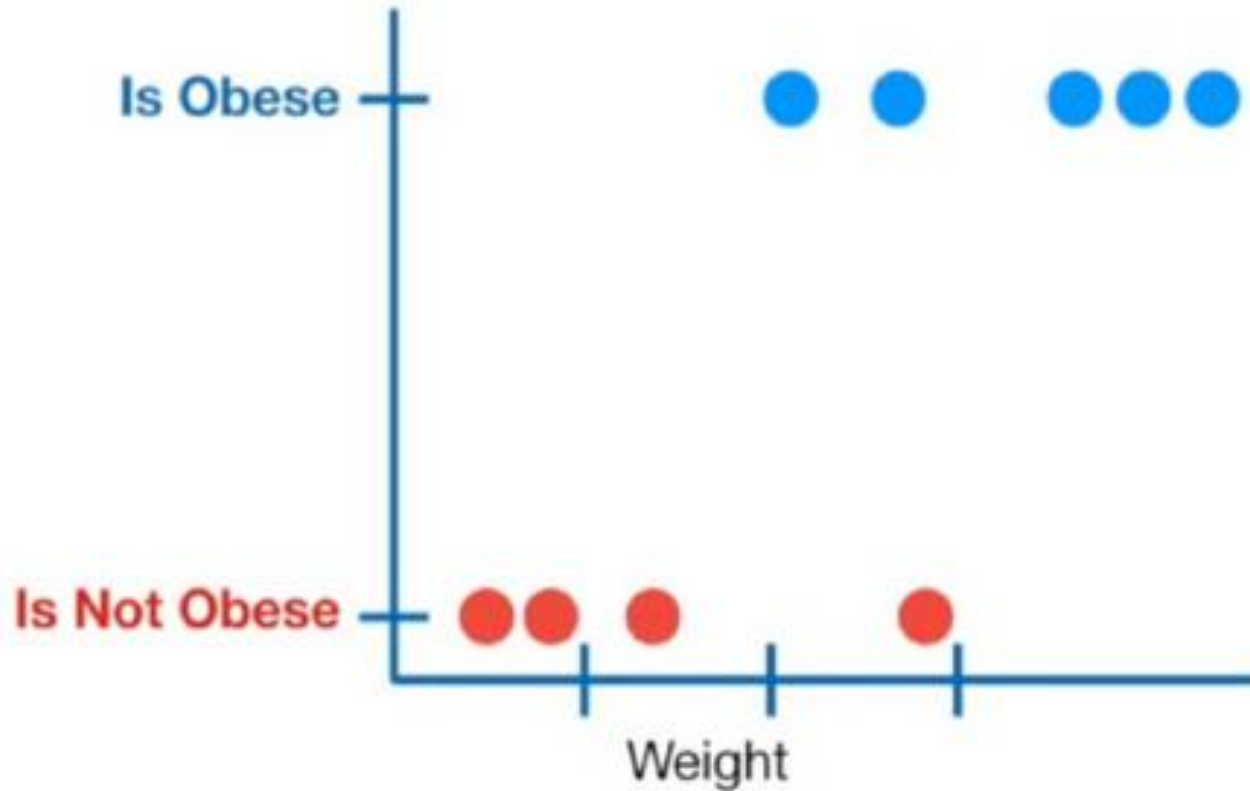


분류의 성과지표를 위한 기본 셋업

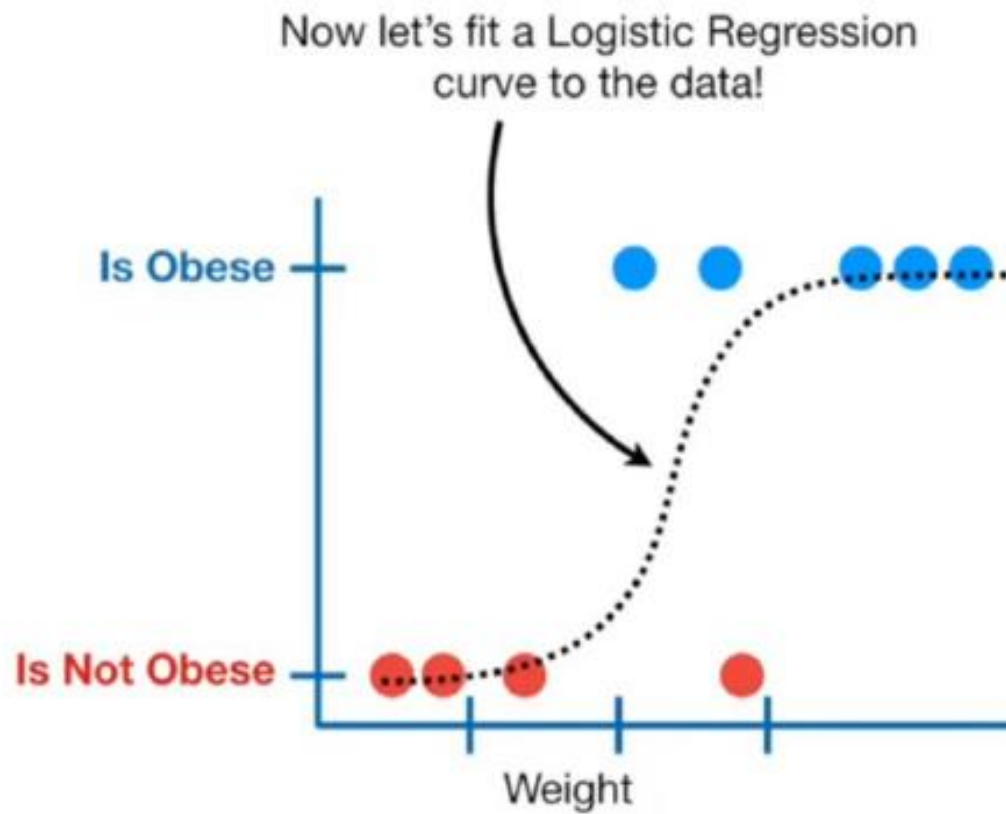
기본 셋업

- 체중을 가지고 비만과 정상 분류 문제를 고려하자.



기본 셋업

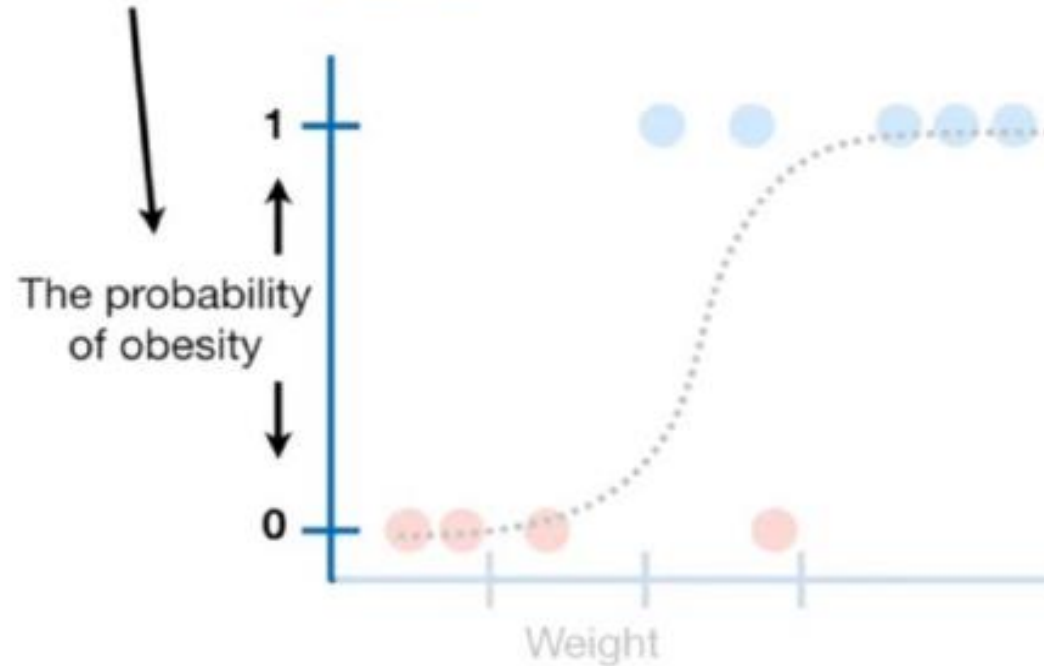
- 로지스틱 함수로 적합화해보자



기본 셋업

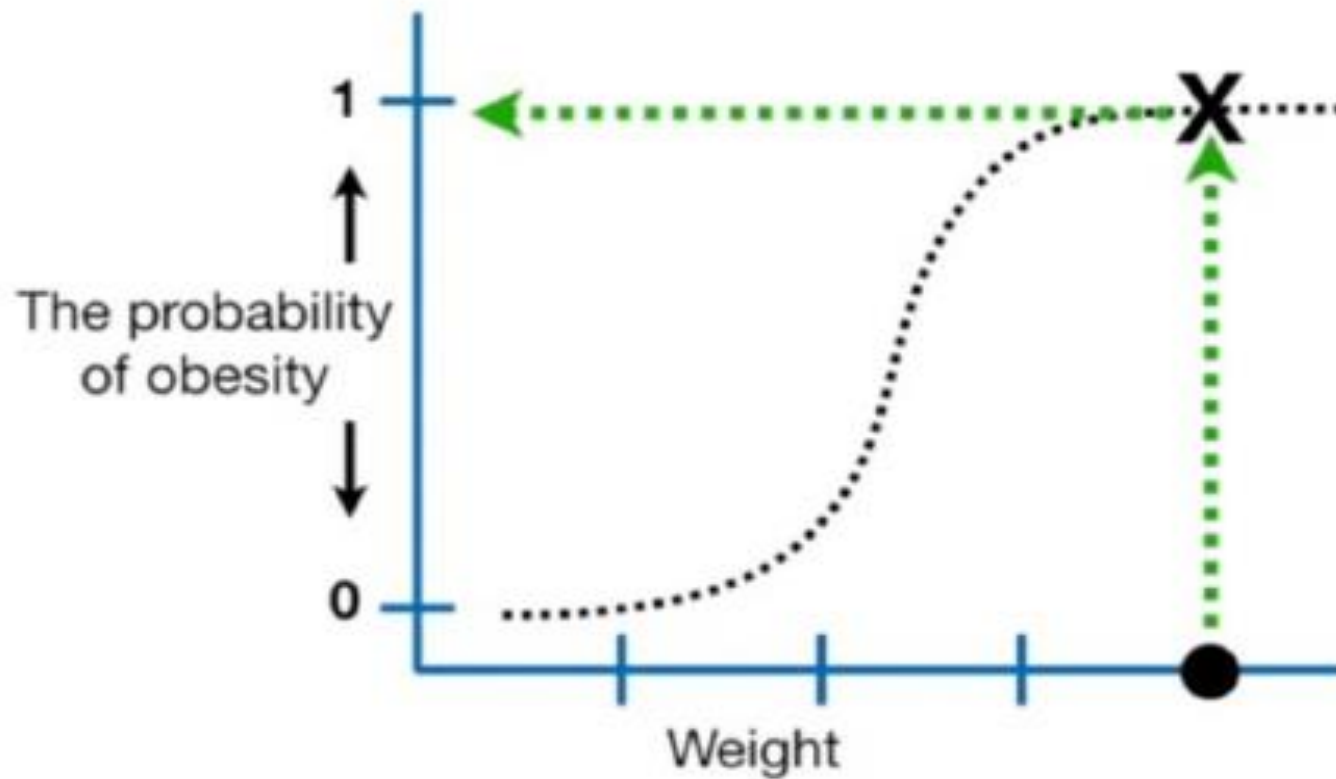
- 로지스틱 회귀를 실행하면 y축은 확률로 생각할 수 있다.

When we're doing Logistic Regression,
the y-axis is converted to the
probability that a mouse **is obese**.



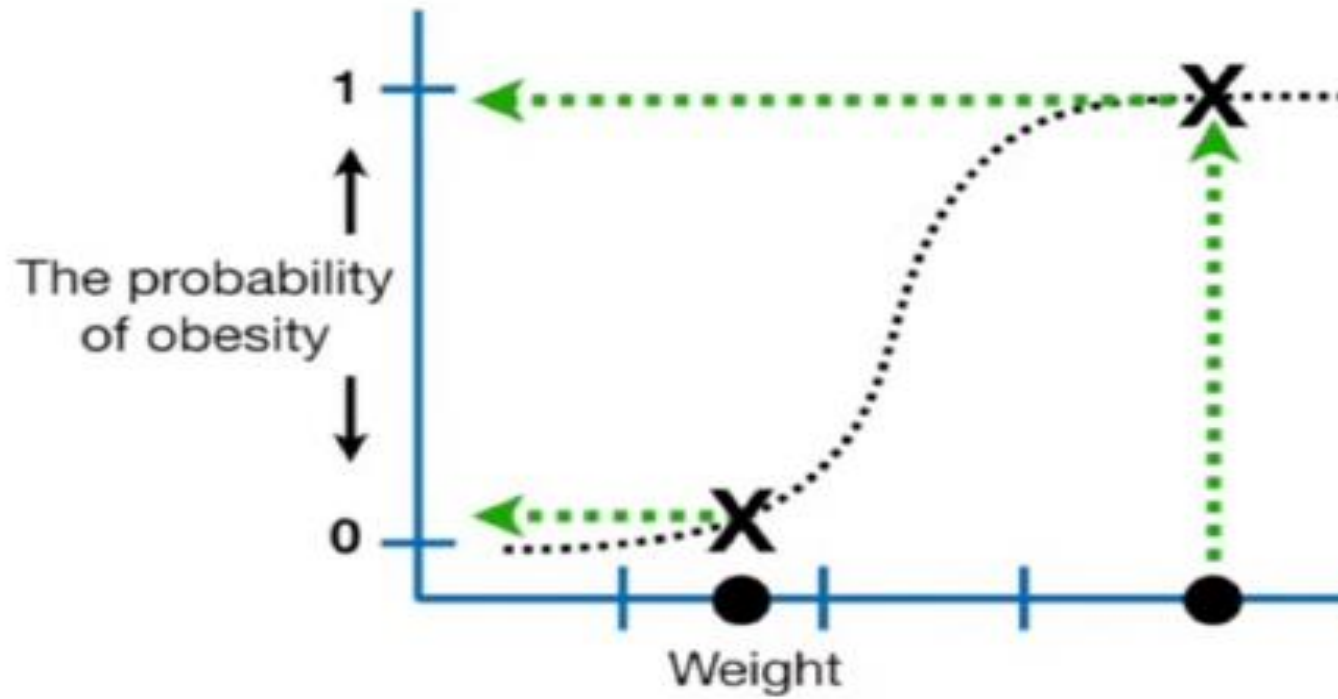
기본 셋업

- 체중이 많이 나가는 경우는 비만일 확률이 높다.



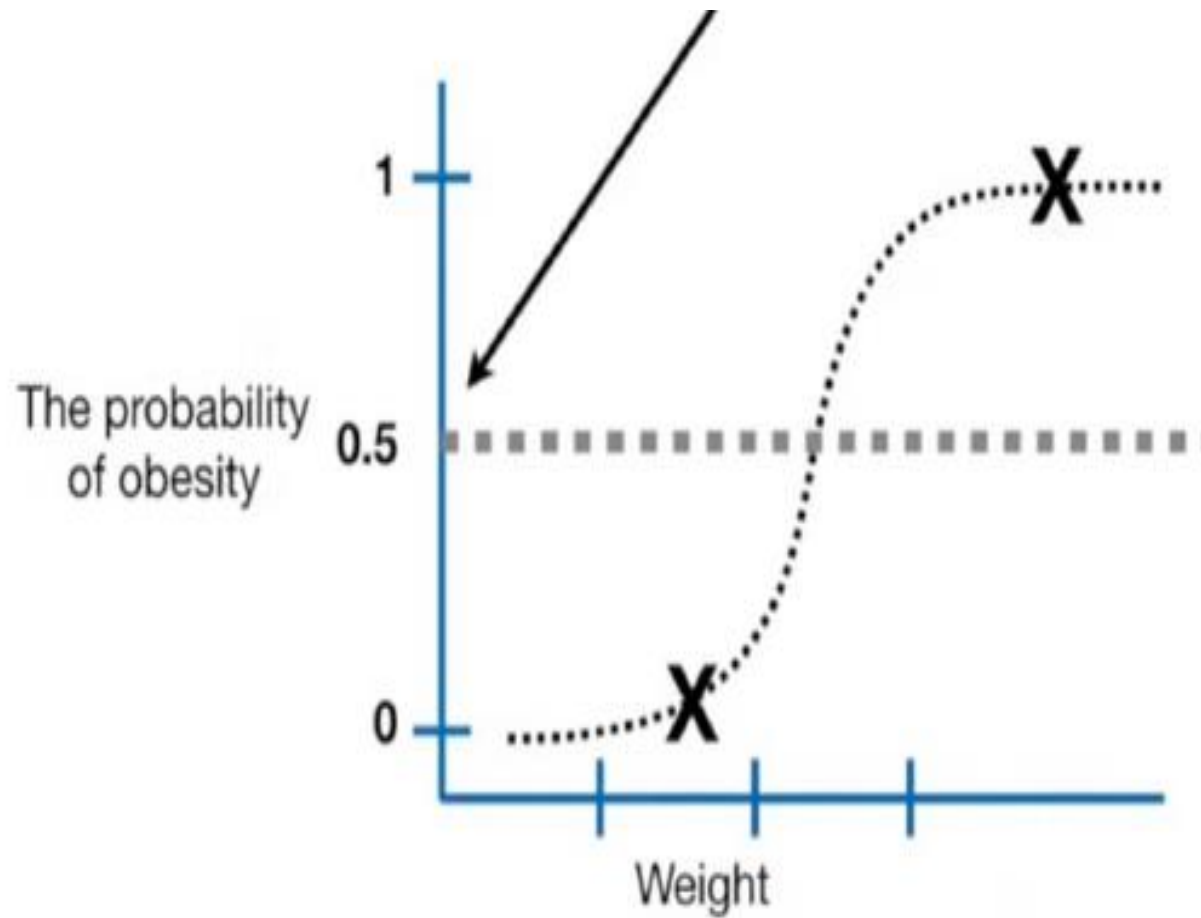
기본 셋업

- 반면, 체중이 덜 나가는 경우는 비만일 확률이 낮다.



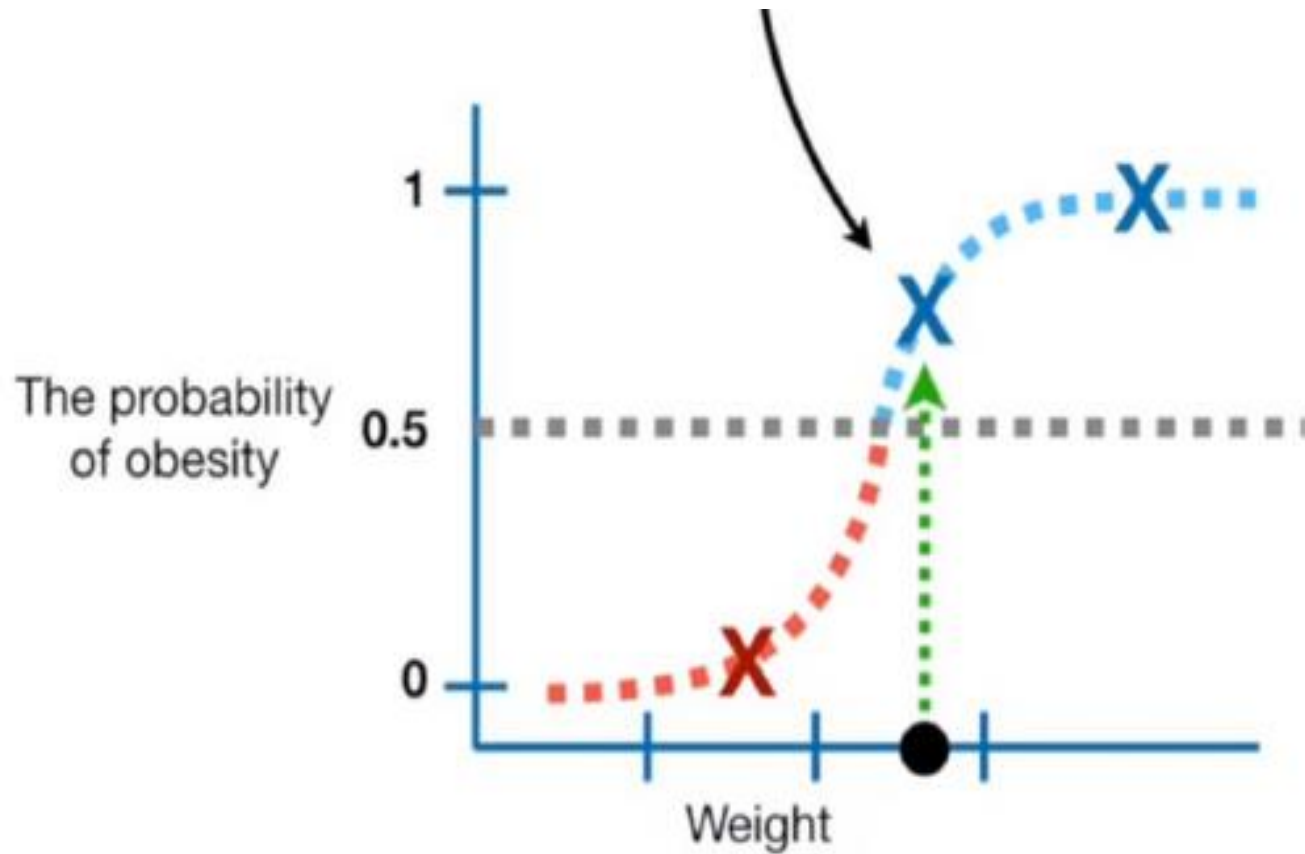
기본 셋업

임계값을 0.5로 지정해
비만/정상을 분류할 수 있다.



기본 셋업

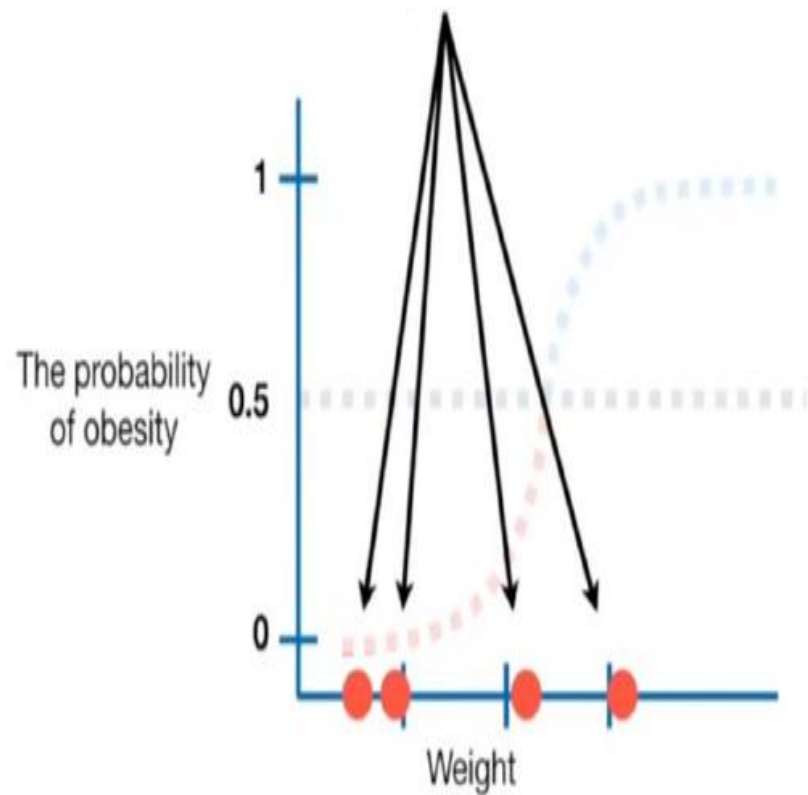
이경우 비만으로 분류될 것이다.



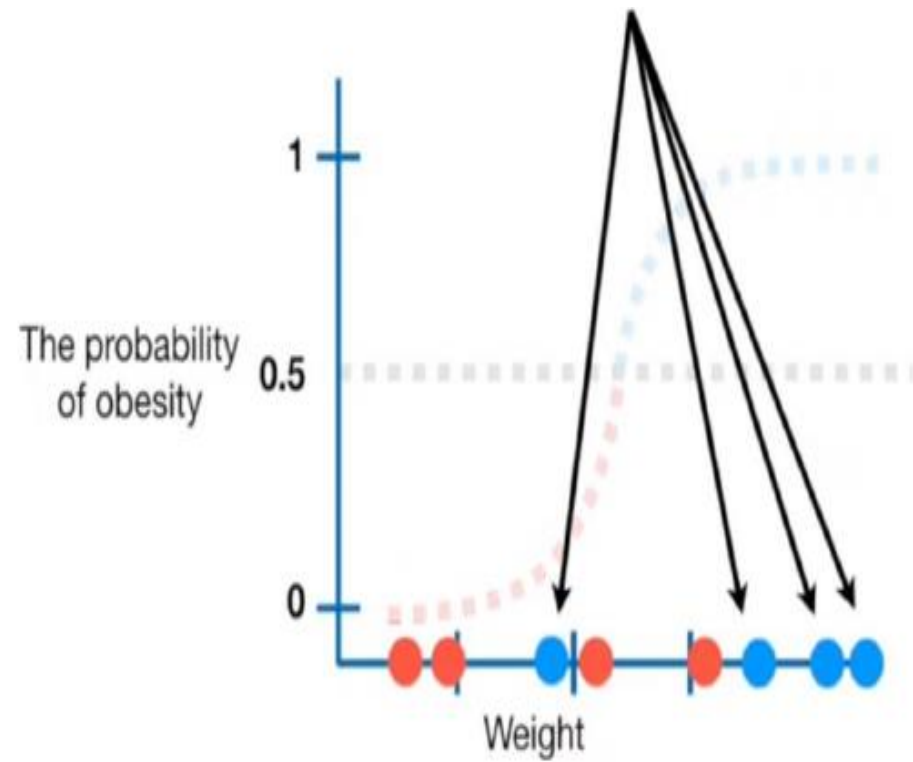
기본 셋업

- 빨간점은 정상을 나타내고 파란점을 비만을 나타낸다.

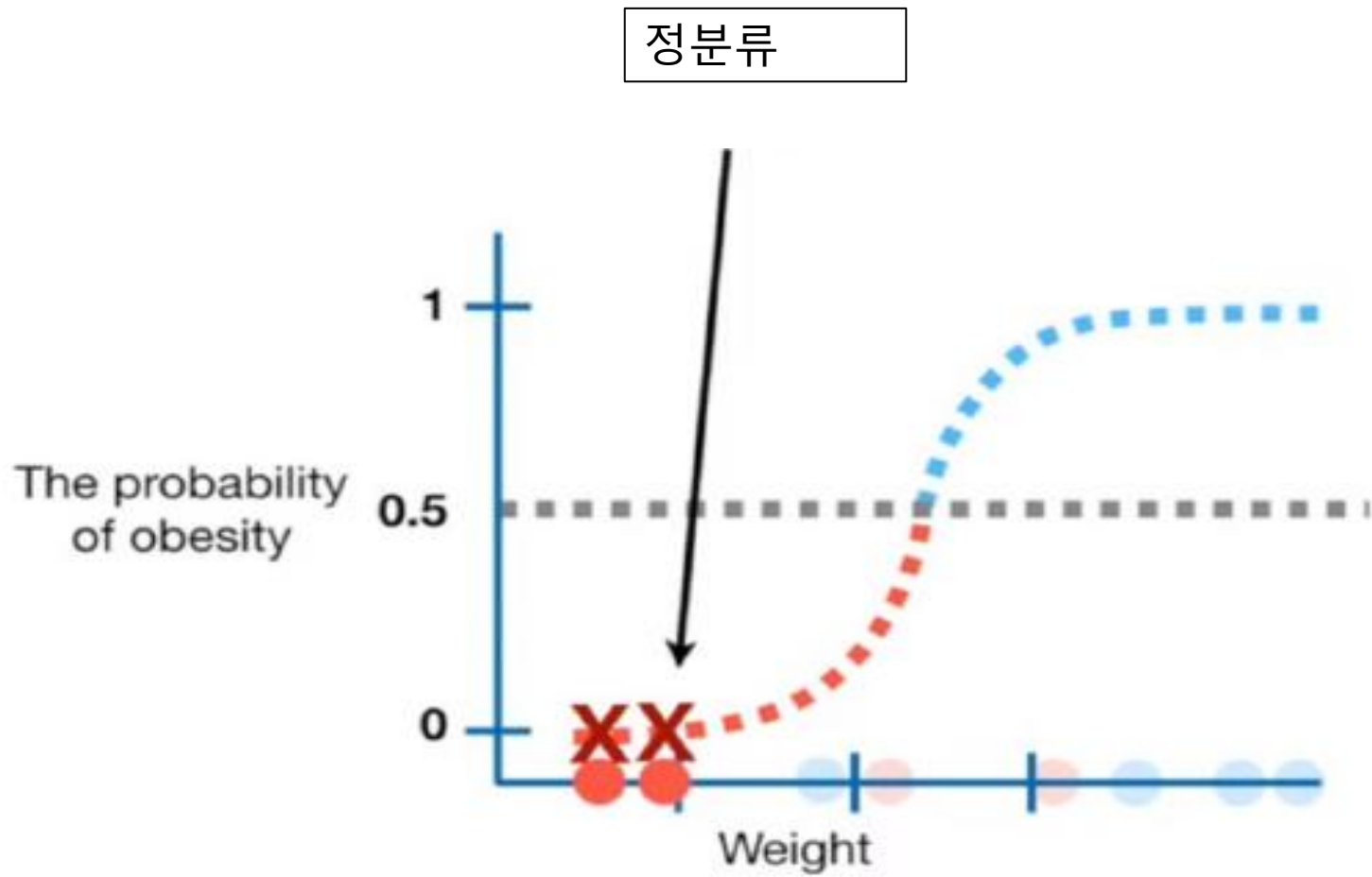
정상.



비만

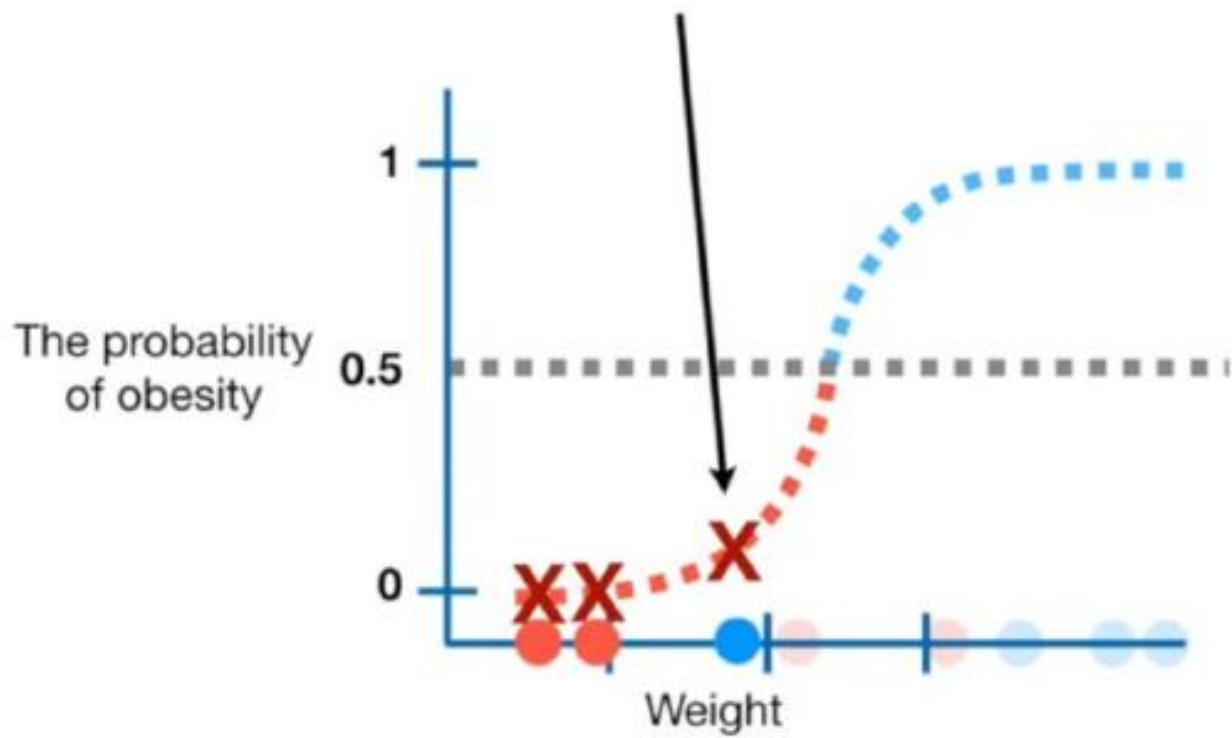


기본 셋업



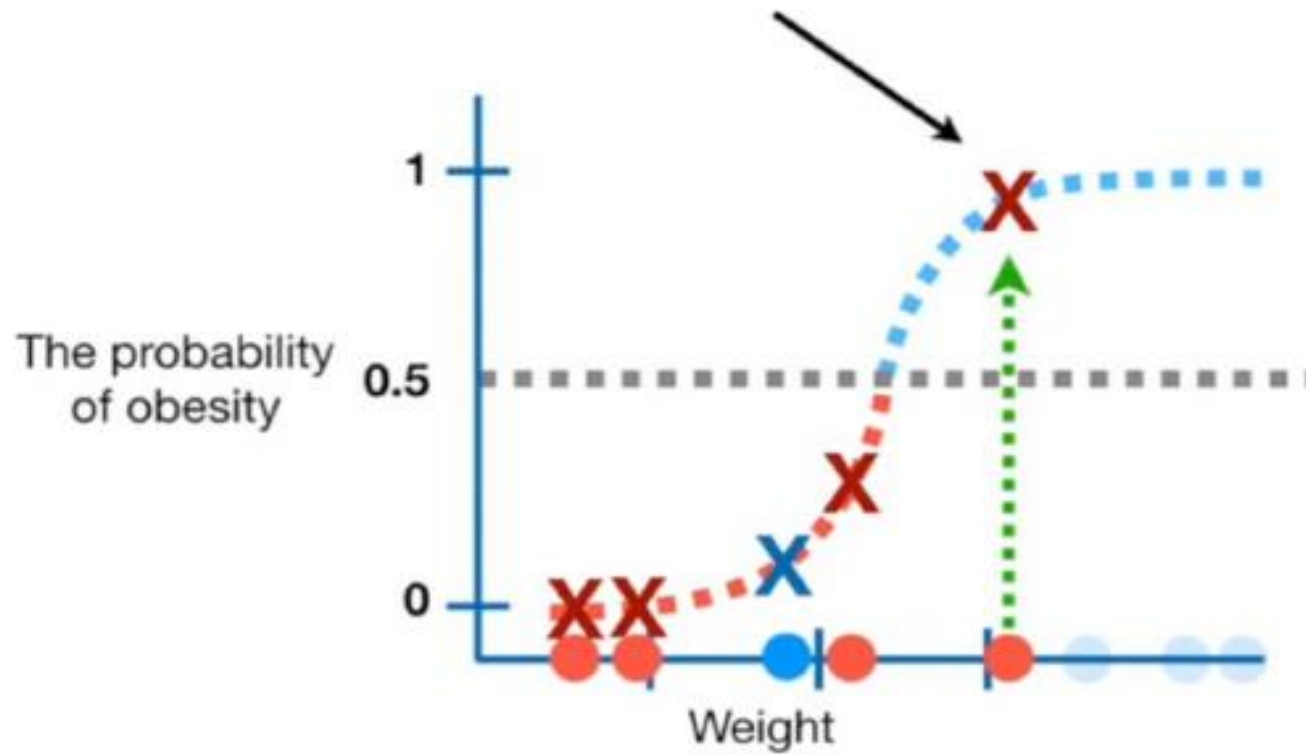
기본 셋업

오분류



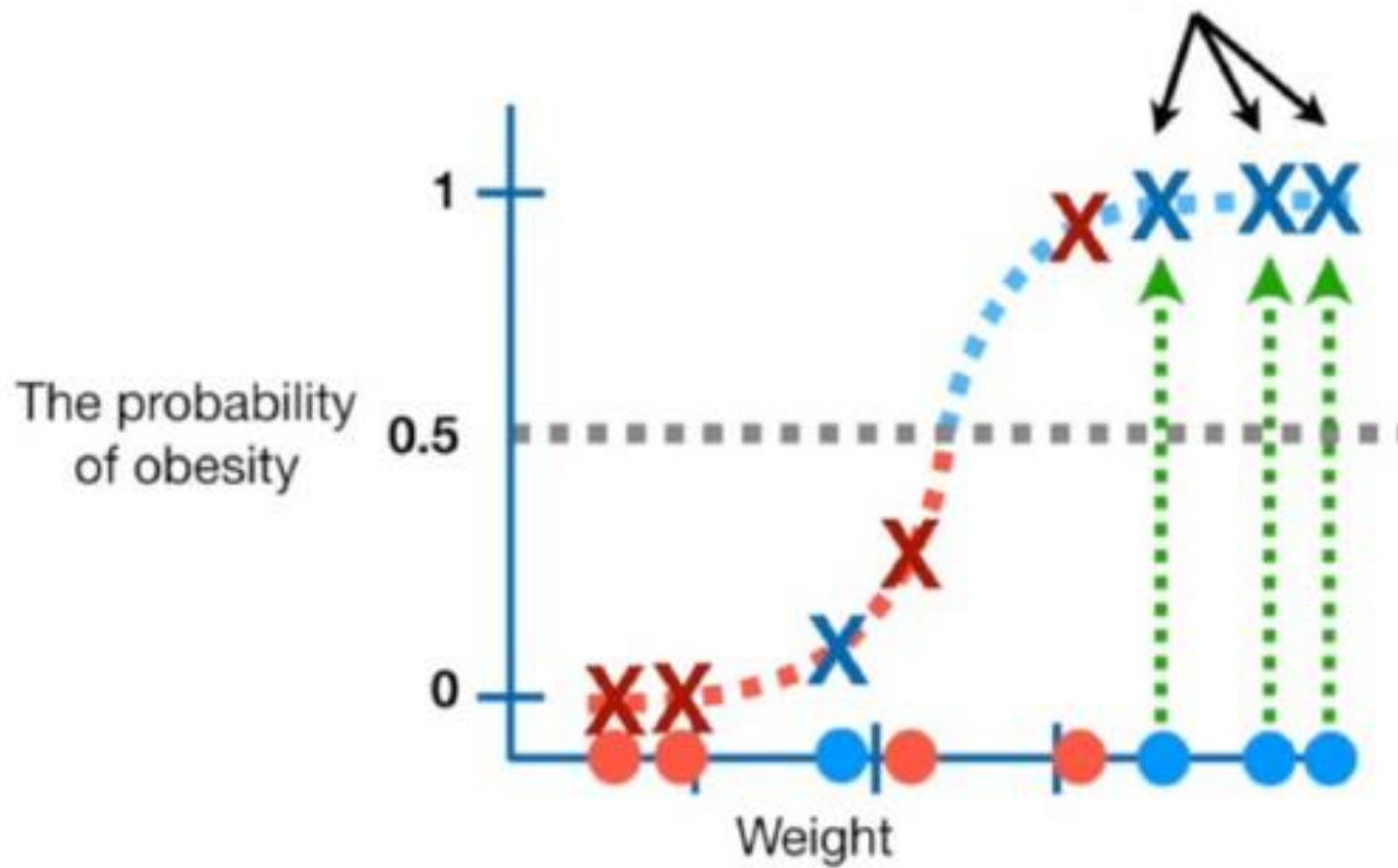
기본 셋업

오분류



기본 셋업

정분류



회동행렬

혼동행렬

- 혼동 행렬(Predicted와 Actual이 바뀌는 경우도 있으니 주의.
내용은 변화없음)

		Actual	
		Is Obese	Is Not Obese
Predicted	Is Obese	3	1
	Is Not Obese	1	3

Heart Disease

- Heart Disease

		Actual	
		Has Heart Disease	Does Not Have Heart Disease
Predicted	Has Heart Disease	True Positives	False Positives
	Does Not Have Heart Disease	False Negatives	True Negatives

혼동행렬

- 분류 성과지표 기본 개념

1. 정확도(Accuracy) = $(TP + TN) / (\text{전체 샘플})$
2. 재현율(Recall) = $TP / (TP + FN)$
3. 정밀도(Precision) = $TP / (TP + FP)$

		Actual	
		Has Heart Disease	Does Not Have Heart Disease
Predicted	Has Heart Disease	True Positives	False Positives
	Does Not Have Heart Disease	False Negatives	True Negatives

분류성과지표 예제

- 테러리스트를 무조건 잡아라는 특명에 모두 테러리스트로 예측

	실제로 테러리스트	실제는 테러리스트 아님	합계
테러리스트 이라고 예측	1	99	100
테러리스트 아니라고 예측	0	0	0
합계	1	99	100

- $\text{Recall} = 1 / (1+0) = 1 = 100\%$
- $\text{Precision} = 1 / (1 + 99) = 1/100 = 1\%$
- ⇒ 평균은 여전히 50%로 높음
- ⇒ 바람직하지 않은 결과
- ⇒ 두척도를 다 봐야 함
- ⇒ 제3의 척도 예를 들면 F1 점수의 필요성

분류성과지표 예제

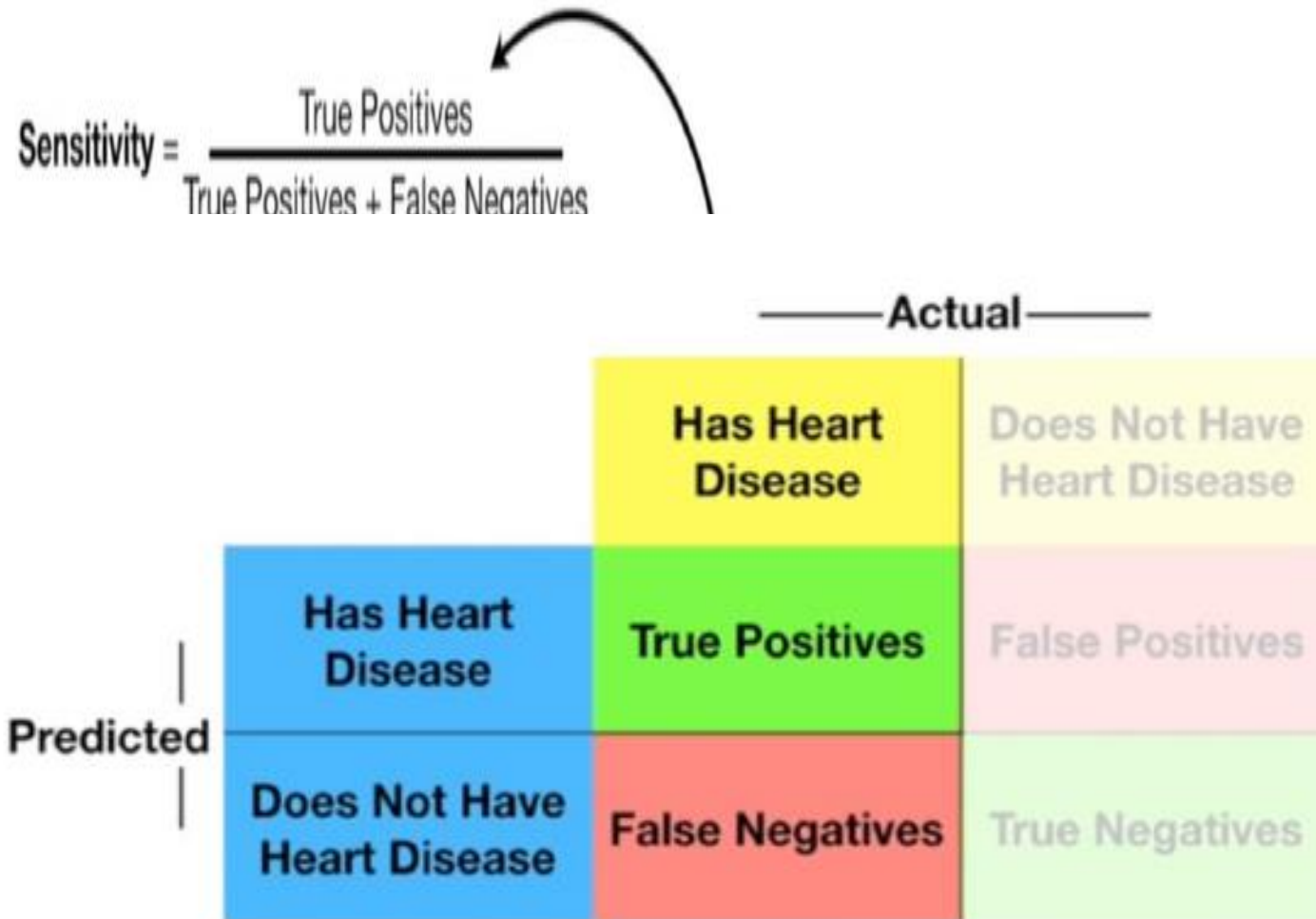
- F1 점수 (F1 Score)은 precision과 recal의 조화평균

$$F! = 2 \times (P \times R) / (R + P) = 2PR/(R+P)$$

⇒ 하나가 0이면 F1이 0에 가까워지므로, 높은 F1을 얻기 위해서는 양쪽이 다 높아야 한다.

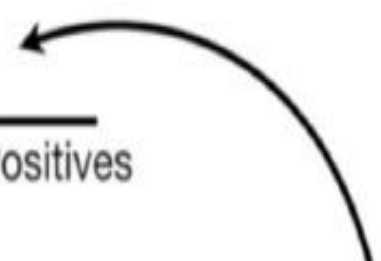
민감도

- 민감도 (Sensitivity) : 재현율(Recall)



특이도

- 특이도(Specificity)

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$


		Actual	
		Has Heart Disease	Does Not Have Heart Disease
Predicted	Has Heart Disease	True Positives	False Positives
	Does Not Have Heart Disease	False Negatives	True Negatives

민감도와 특이도

- 민감도와 특이도 구하기

	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	139	20
Does Not Have Heart Disease	32	112

민감도와 특이도

- 민감도와 특이도 구하기

	Has Heart Disease	Does Not Have Heart Disease
Has Heart Disease	139	20
Does Not Have Heart Disease	32	112

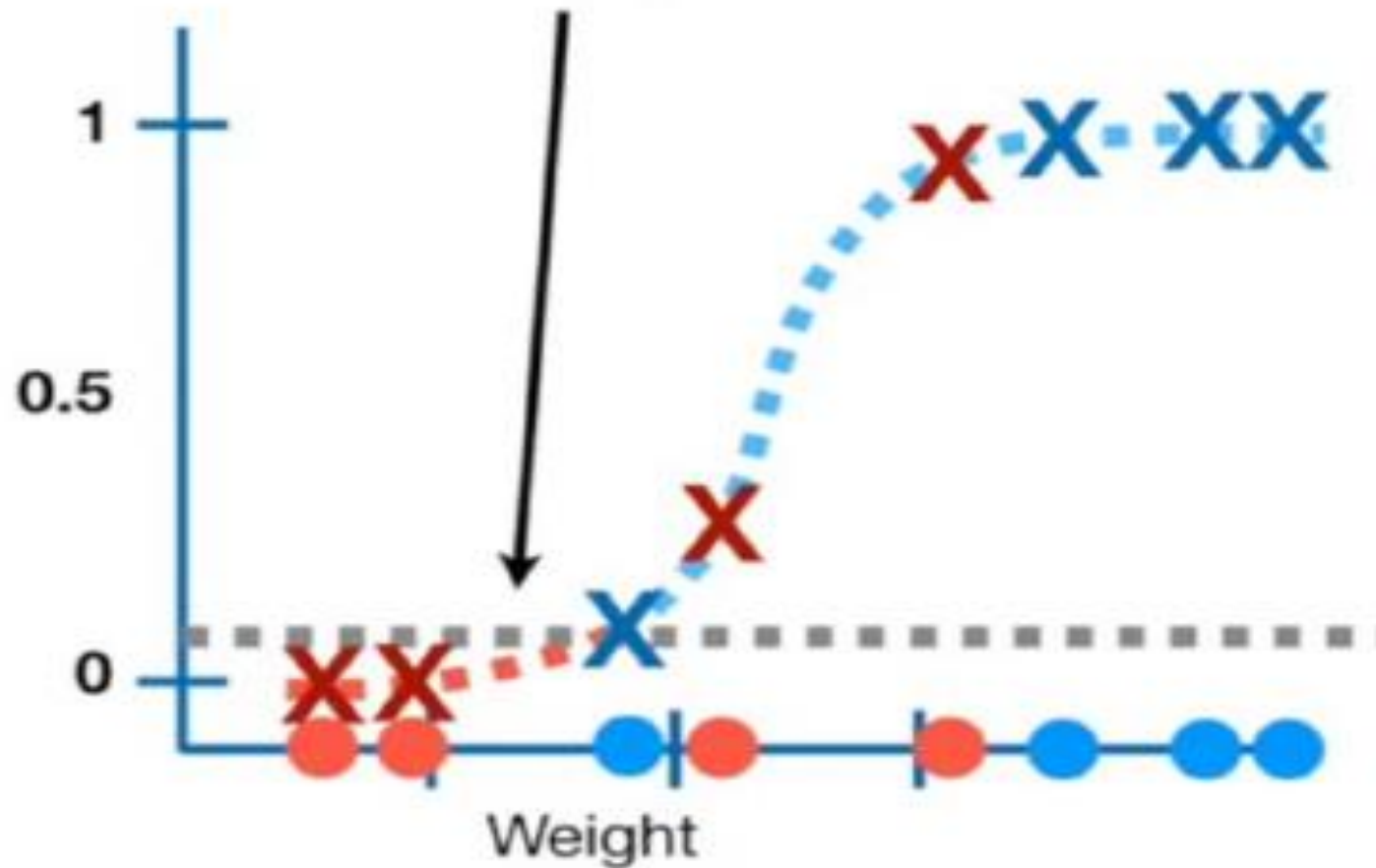
민감도 =
 $139 / (139 + 32) = 0.81$

특이도 =
 $112 / (112 + 20) = 0.85$

ROC 곡선

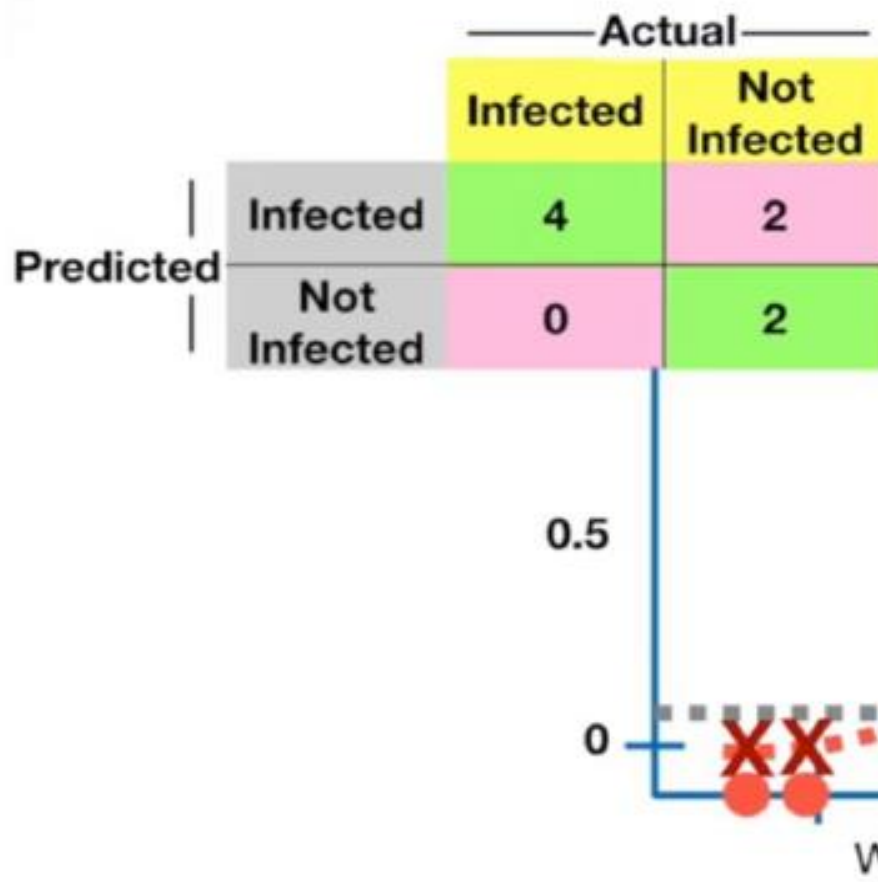
ROC 곡선

- 이제 임계값을 0.1로 설정해보자. (모든 비만을 잡을 수 있을 것이다.)



ROC 곡선

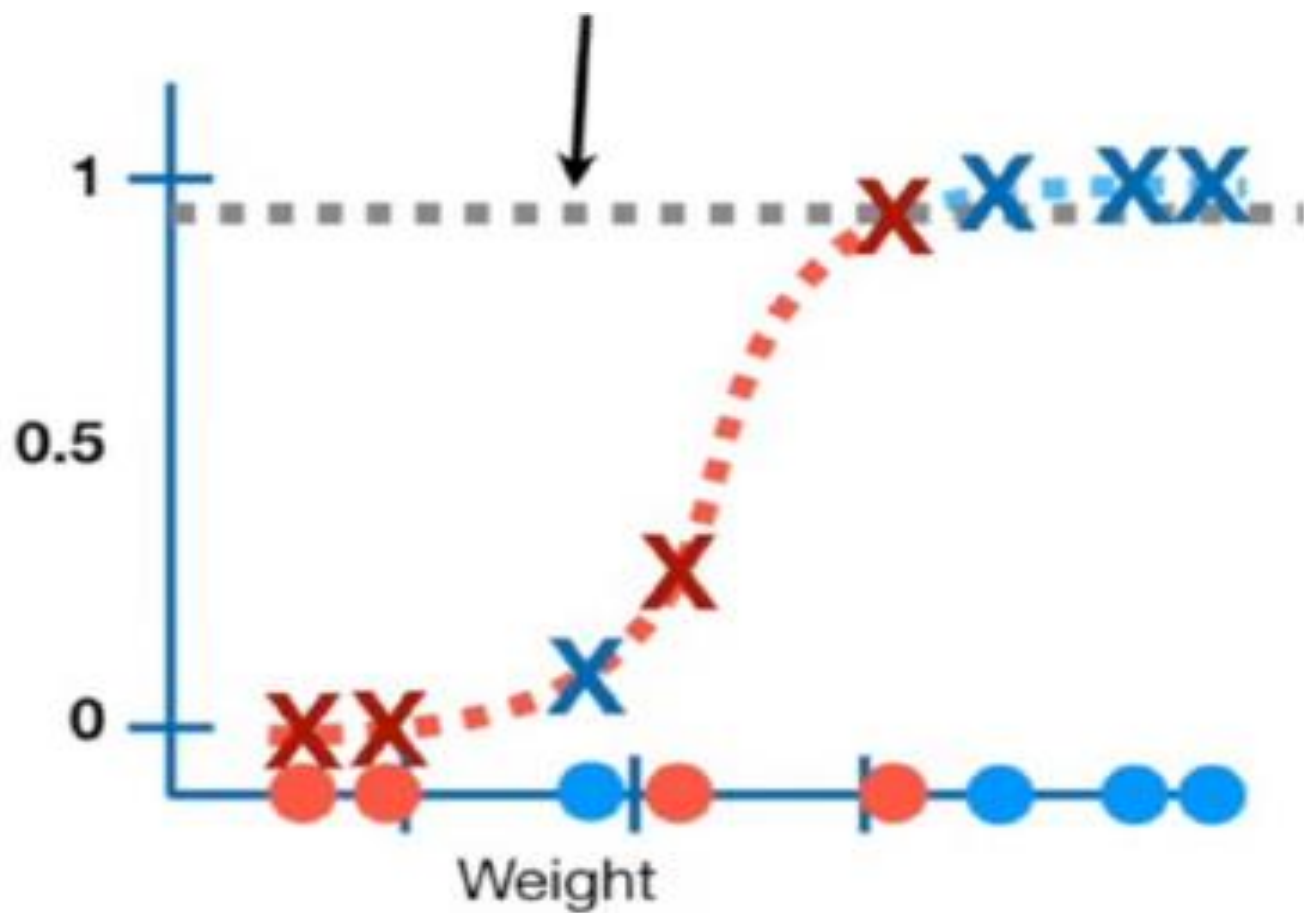
- 이제 임계값을 0.1로 설정해보자.



임계값을 낮추면 거짓양성(비만으로 예측했는데 틀린것)은 높아지지만 양성은 다 잡아낸다.

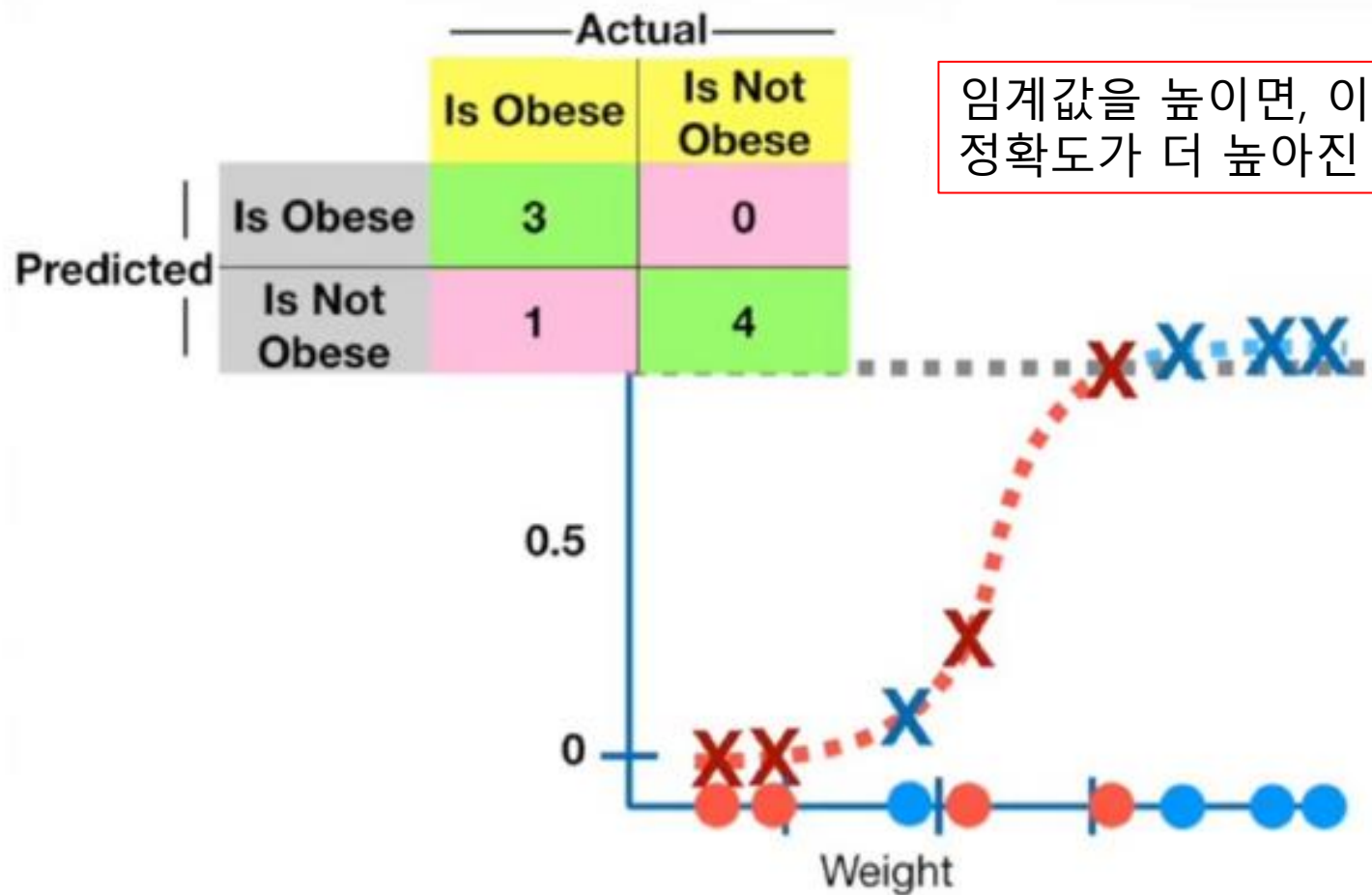
ROC 곡선

- 이제 임계값을 0.9로 설정해보자.



ROC 곡선

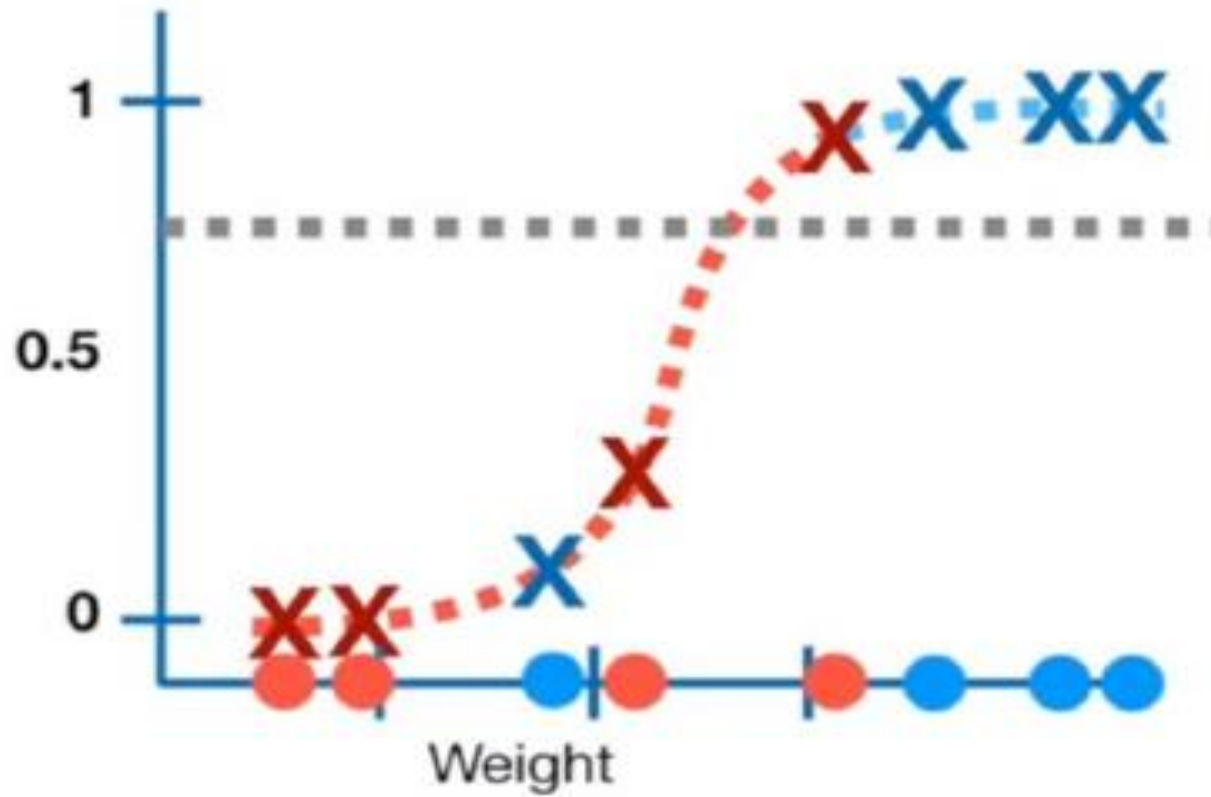
- 이제 임계값을 0.9로 설정해보자.



임계값을 높이면, 이 경우 분류 정확도가 더 높아진 것을 알 수 있다.

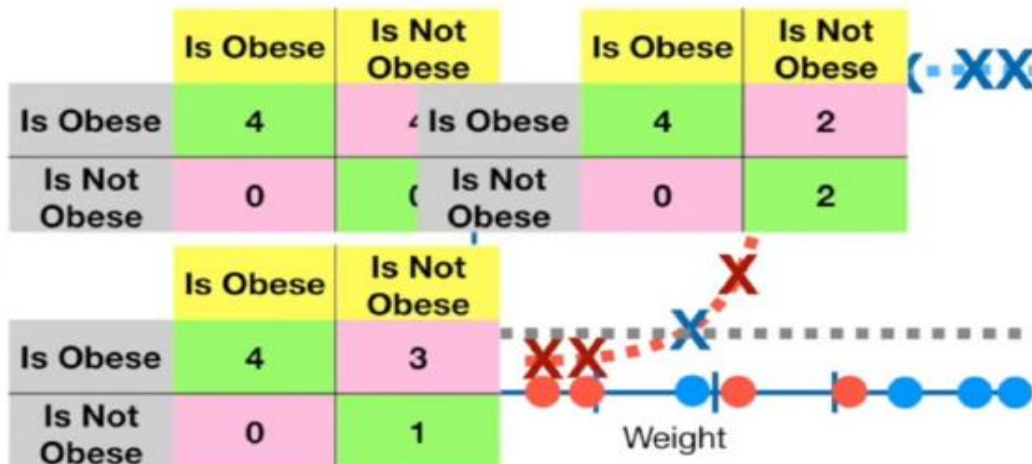
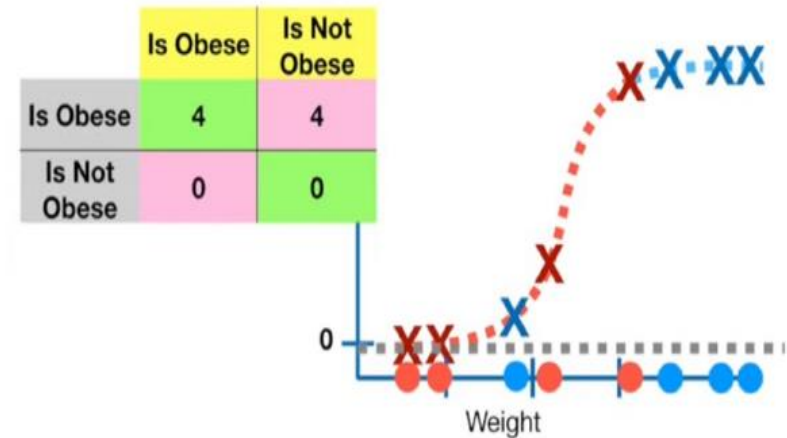
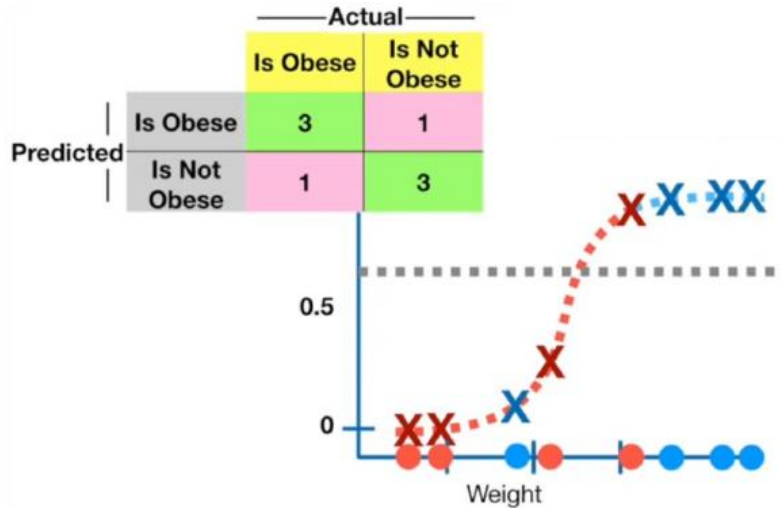
ROC 곡선

- 어떻게 임계값을 정할까?



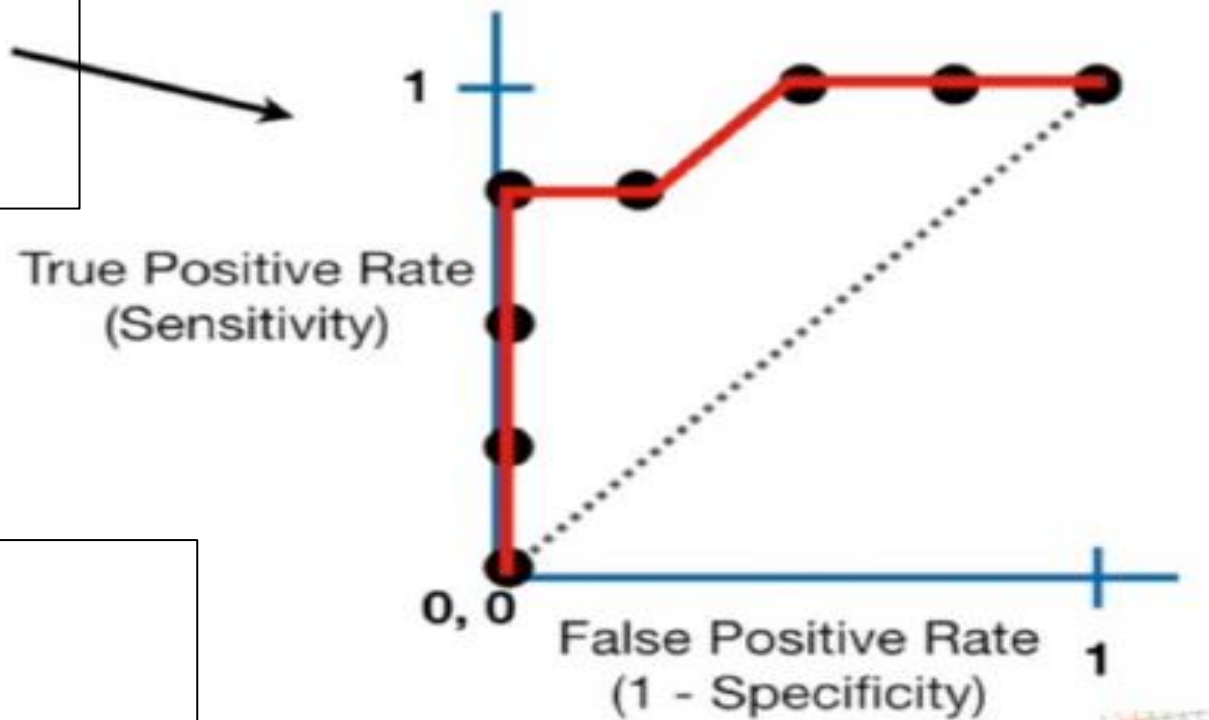
ROC 곡선

- 여러 임계값을 시도하면서 혼동행렬을 그려본다. 하지만, 이는 너무 복잡한다.



ROC 곡선

일목요연하게 볼 수
있는 더 좋은 방법이
소위 ROC(Receiver
Operator
Characteristic)
그래프를 그리는
것이다.



[용어정리]

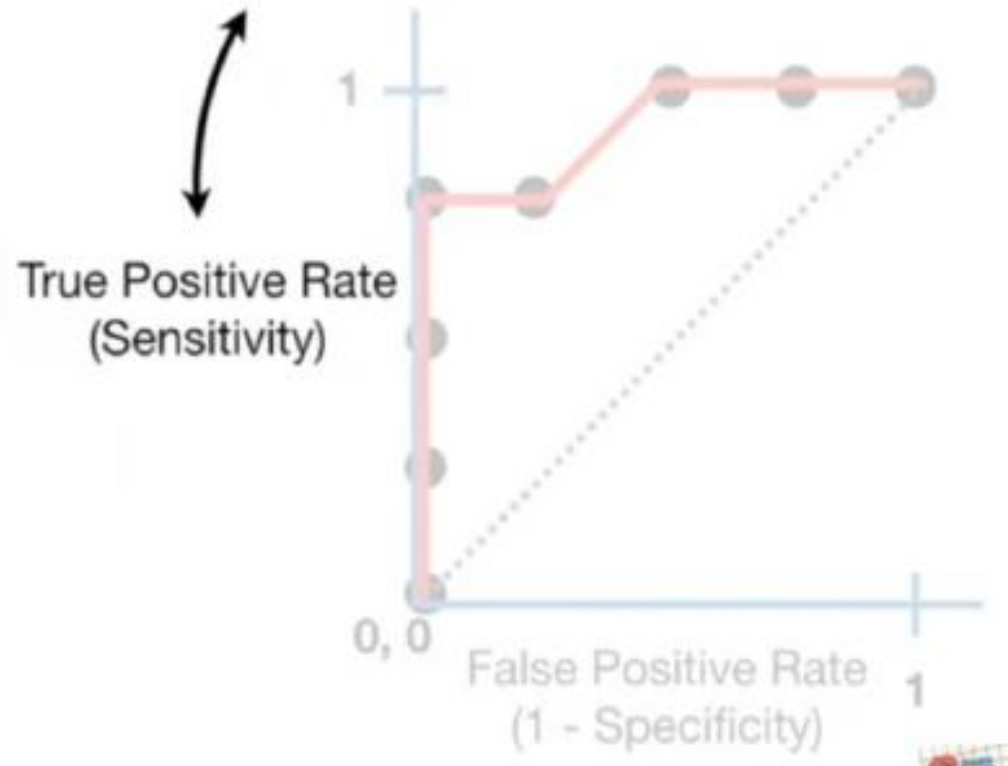
민감도(Sensitivity)

특이도(Specificity)

ROC 곡선

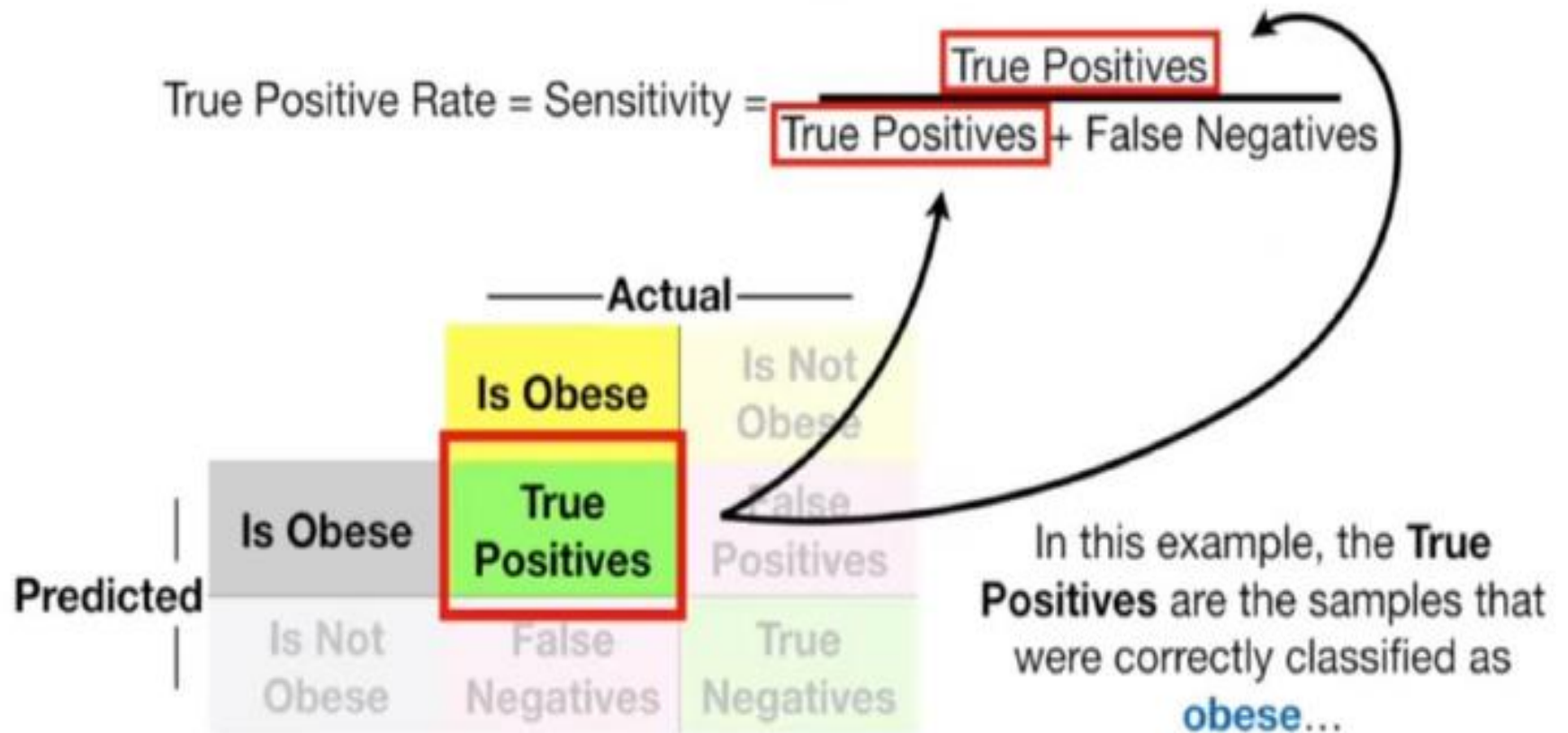
- 참양성율 (민감도)

$$\text{True Positive Rate} = \text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$



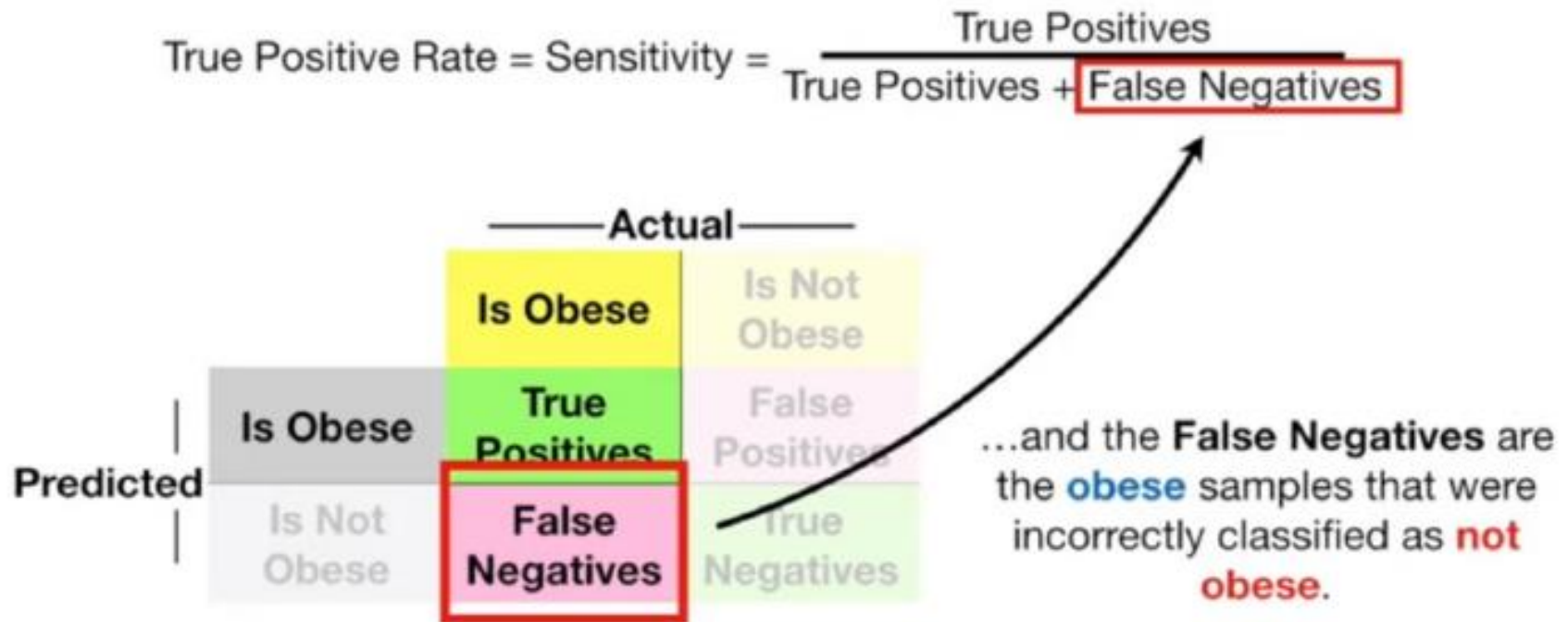
ROC 곡선

- 참양성율 (민감도)



ROC 곡선

- 참양성율 (민감도)



ROC 곡선

- 참양성율 (민감도)

$$\text{True Positive Rate} = \text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

		Actual	
		Is Obese	Is Not Obese
Predicted	Is Obese	True Positives	False Positives
	Is Not Obese	False Negatives	True Negatives

The **True Positive Rate** tells you what proportion of **obese** samples were correctly classified.

ROC 곡선

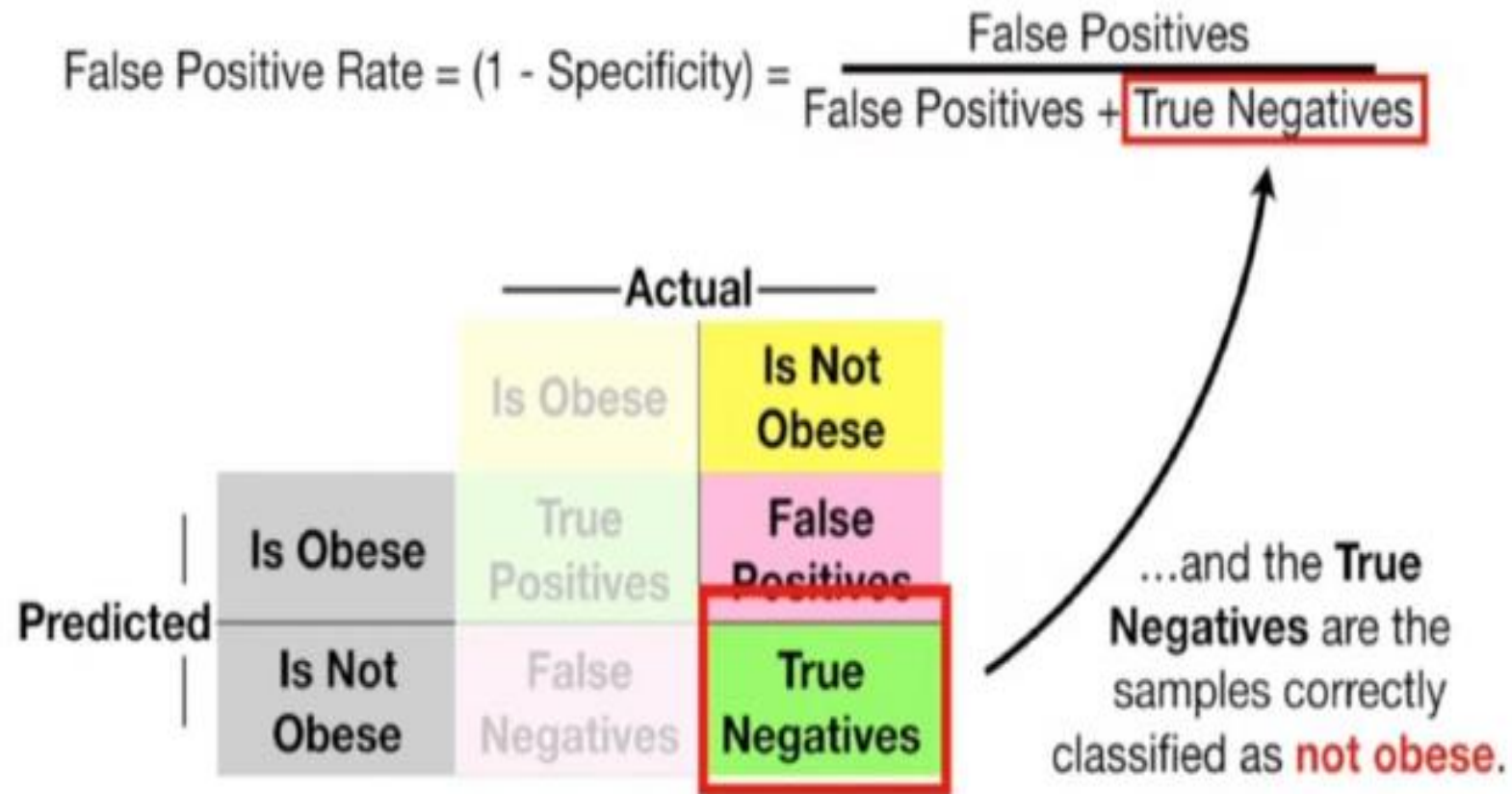
- 거짓양성율 (1-특이도)

$$\text{False Positive Rate} = (1 - \text{Specificity}) = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}$$



ROC 곡선

- 거짓양성율 (1-특이도)



ROC 곡선

- 거짓양성율 (1-특이도)

$$\text{False Positive Rate} = (1 - \text{Specificity}) = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}$$

		Actual	
		Is Obese	Is Not Obese
Predicted	Is Obese	True Positives	False Positives
	Is Not Obese	False Negatives	True Negatives

The **False Positive Rate** tells you the proportion of **not obese** samples that were incorrectly classified and are **False Positives**.

- 임계값이 낮아서 모든 비만으로 예측하는 경우 ROC 곡선을 그려보자.



ROC 곡선

- 임계값이 낮아서 모든 비만을 찾아내는 예에서 ROC 곡선을 그려보자.

$$\text{True Positive Rate} = \text{Sensitivity} = \frac{4}{4+0} = 1$$

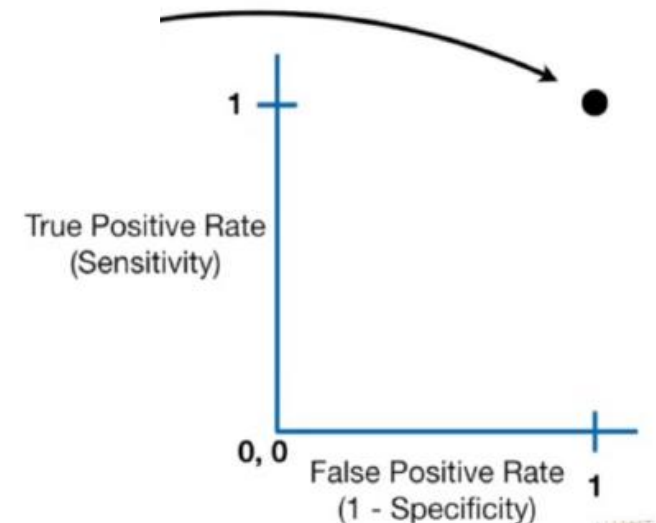
The **True Positive Rate**, when the threshold is so low that every single sample is classified as **obese**, is 1.

		Actual	
		Is Obese	Is Not Obese
Predicted	Is Obese	4	4
	Is Not Obese	0	0

$$\text{False Positive Rate} = 1 - \text{Specificity} = \frac{4}{4+0} = 1$$

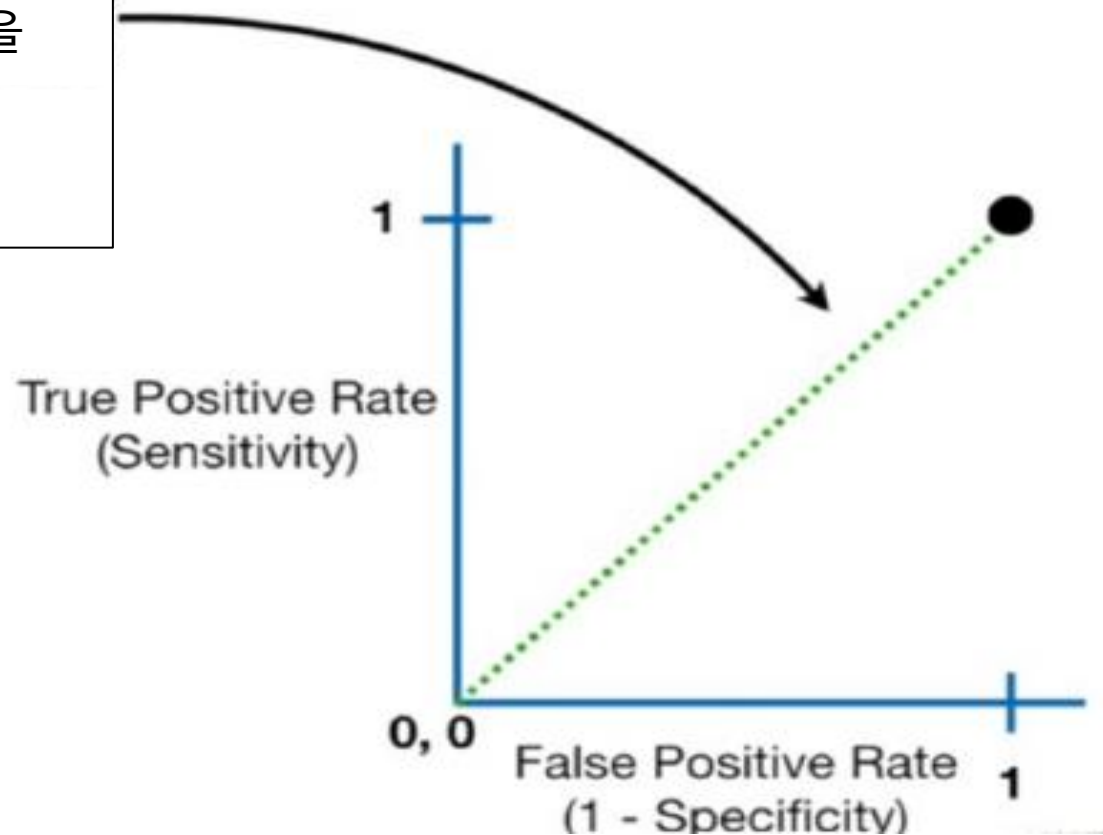
This means that every single sample that was **not obese** was *incorrectly* classified as **obese**.

		Actual	
		Is Obese	Is Not Obese
Predicted	Is Obese	4	4
	Is Not Obese	0	0



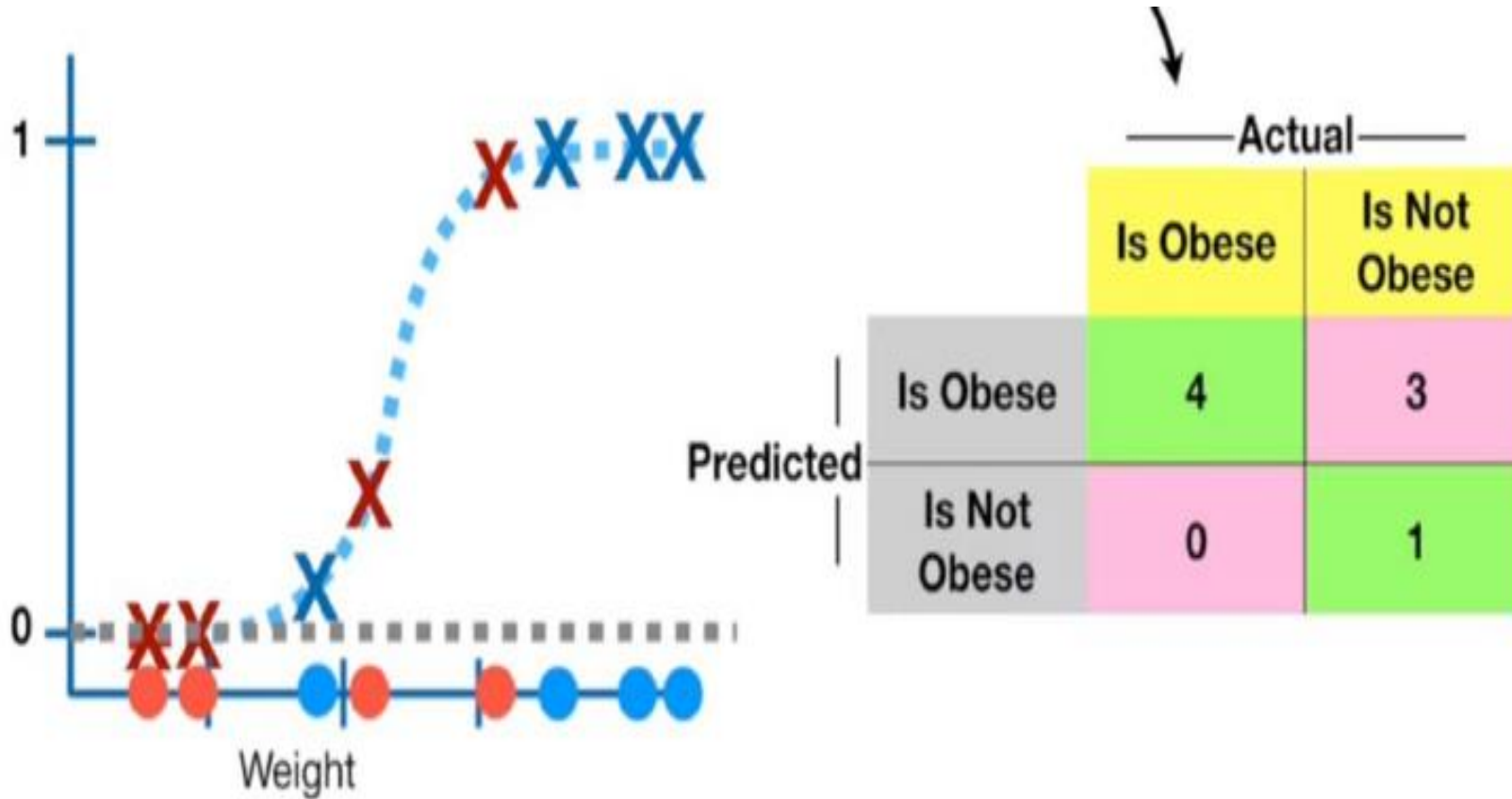
ROC 곡선

- 초록색선은 참양성율과 거짓양성율이 같은 경우를 표시하며, 예에서 비만을 비만으로 정확하게 분류하는 비율과 정상을 비만으로 오분류하는 비율이 같은 경우를 의미한다.



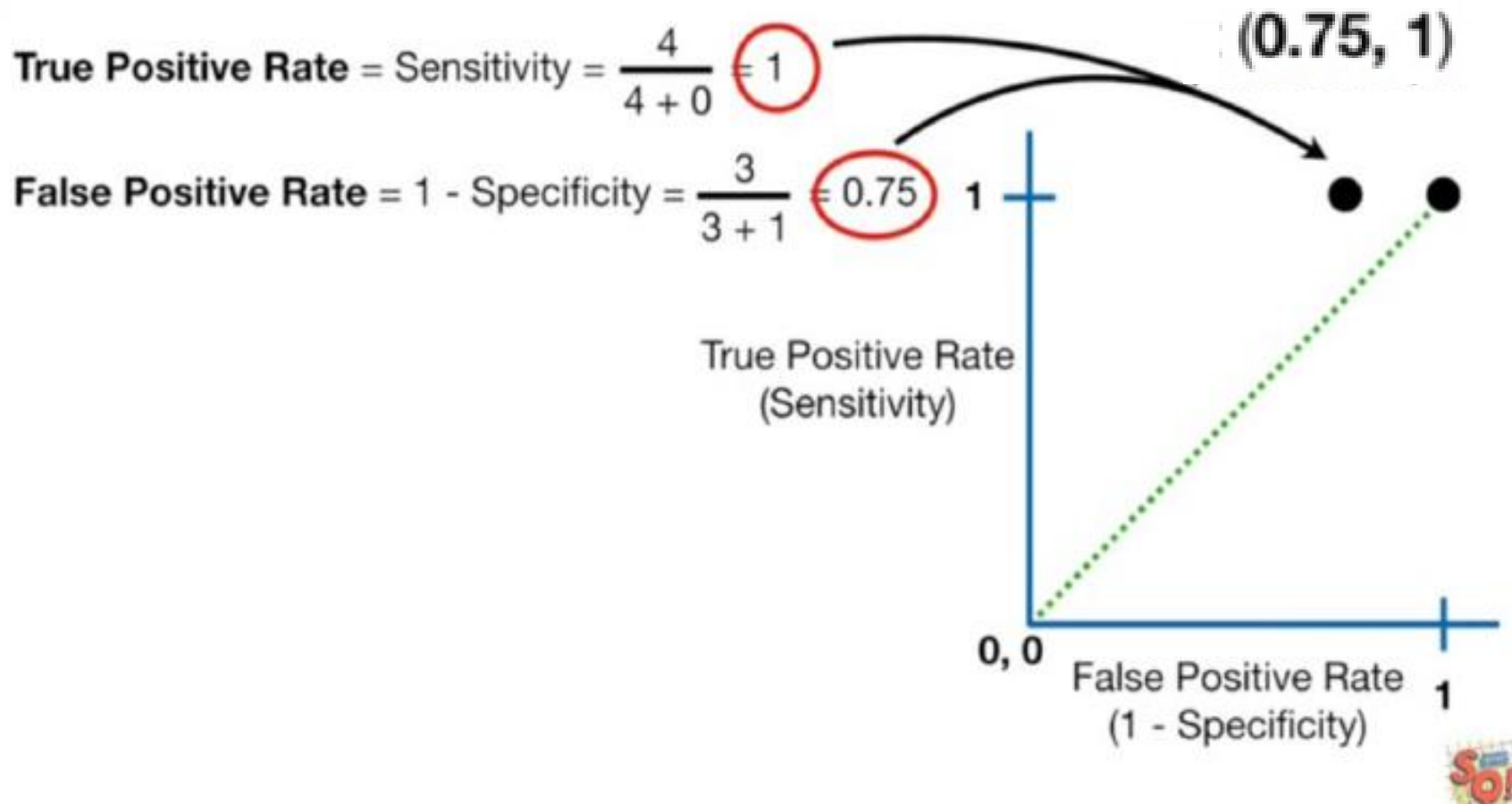
ROC 곡선

- 이제 임계값을 조금 올려서 하나만 정상이라고 하고 나머지를 모두 비만이라고 하는 경우 ROC 곡선을 그려보자.



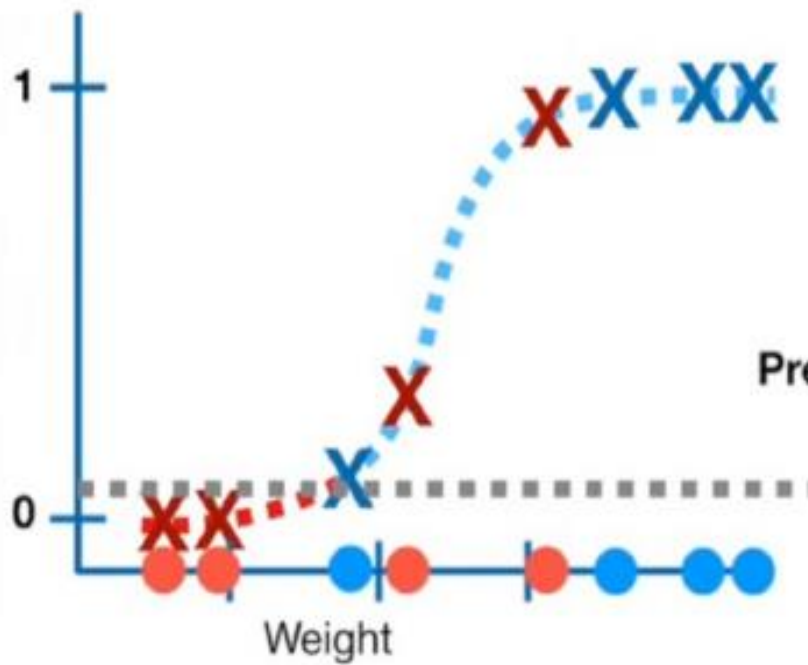
ROC 곡선

- ROC 곡선 상의 (0.75, 1)은 초록선의 왼쪽에 있으므로 참양성율(비만을 정확하게 분류하는 비율)이 거짓양성(잘못 비만으로 분류하는 비율)보다 큰 것을 알 수 있다.
-> 샘플이 비만인지 아닌지를 결정하는 새로운 임계값이 첫번째보다 더 낮다.



ROC 곡선

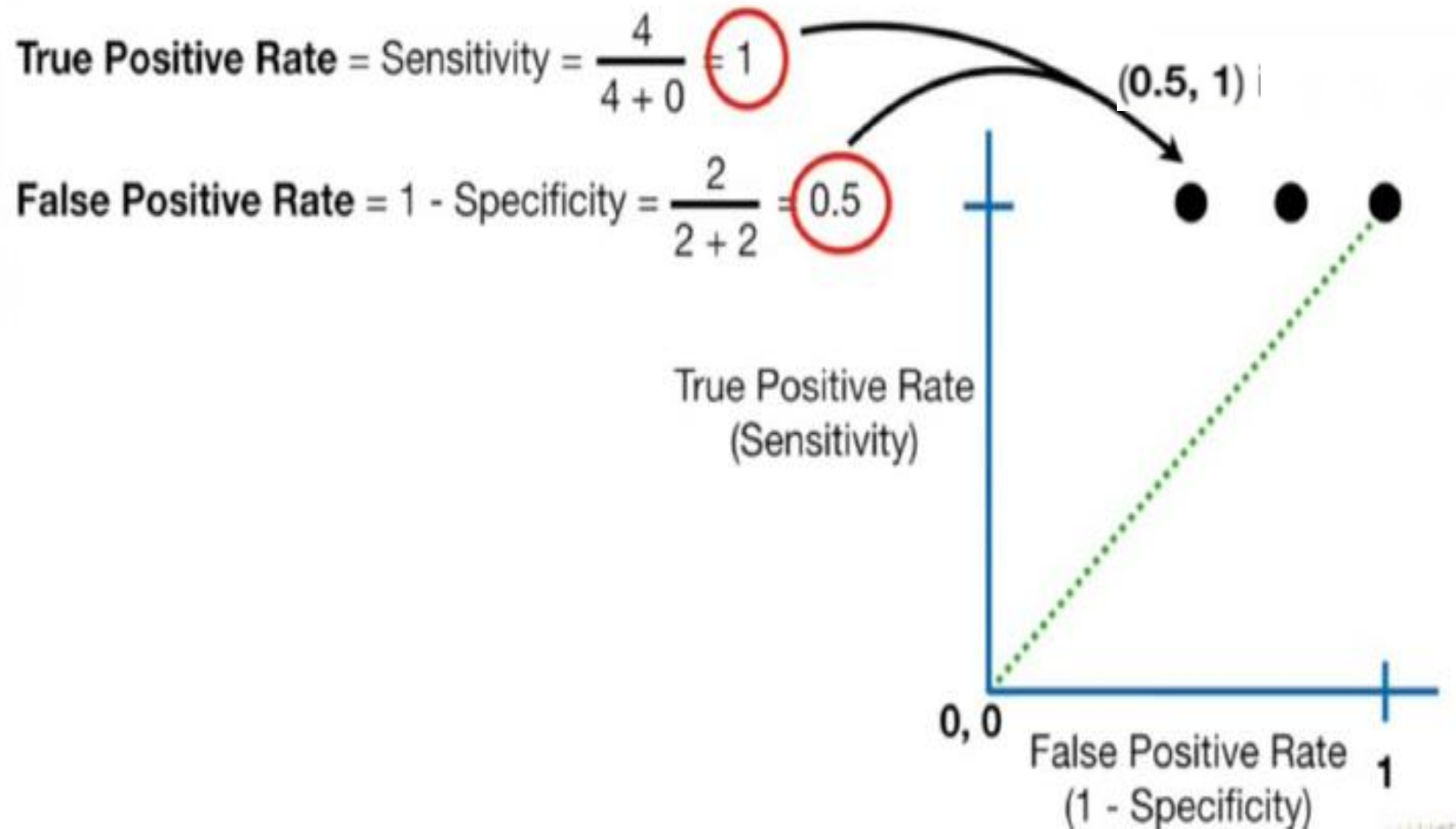
- 이제 임계값을 조금 더 올려보자.



	Actual	
	Is Obese	Is Not Obese
Is Obese	4	2
Is Not Obese	0	2

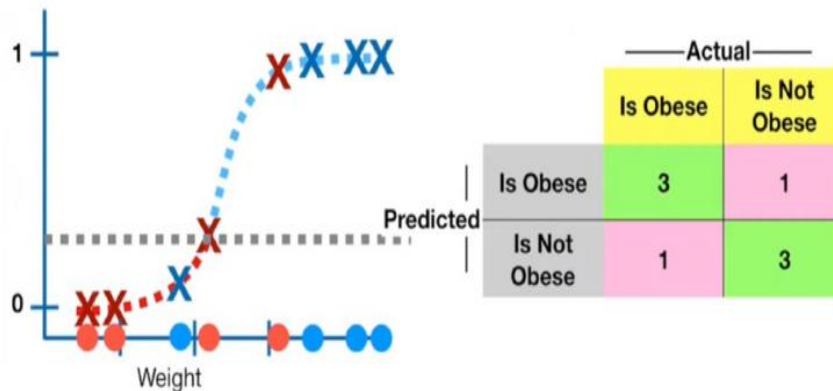
ROC 곡선

- 새로운 점을 찍어보자.



ROC 곡선

- 비만과 정상 분류 문제를 고려하자.



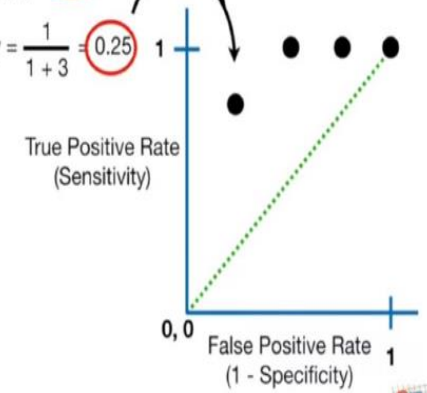
True Positive Rate = Sensitivity = $\frac{3}{3+1} = 0.75$...calculate the True Positive Rate and the False Positive Rate...

False Positive Rate = 1 - Specificity = $\frac{1}{1+3} = 0.25$

		Actual	
		Is Obese	Is Not Obese
Predicted	Is Obese	3	1
	Is Not Obese	1	3

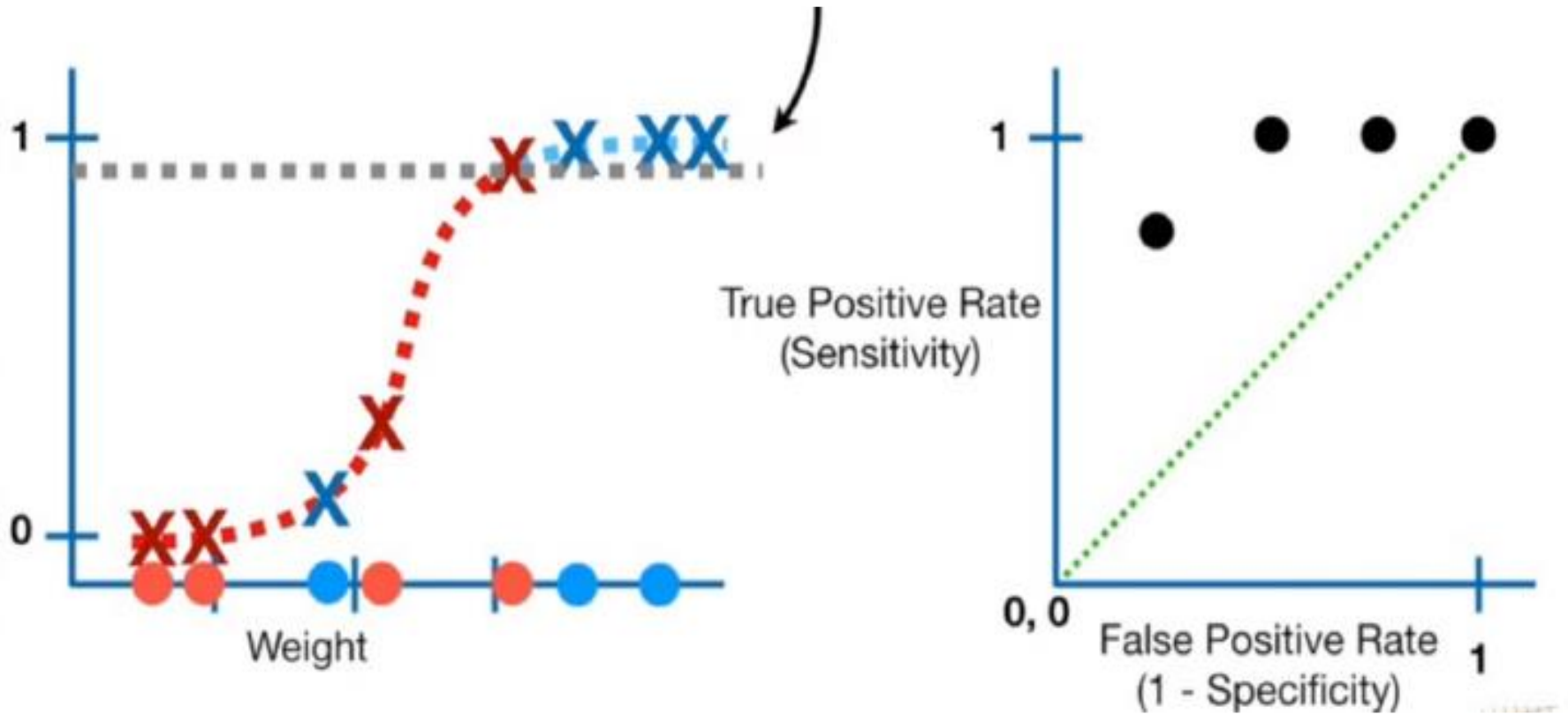
True Positive Rate = Sensitivity = $\frac{3}{3+1} = 0.75$...and plot the point.

False Positive Rate = 1 - Specificity = $\frac{1}{1+3} = 0.25$



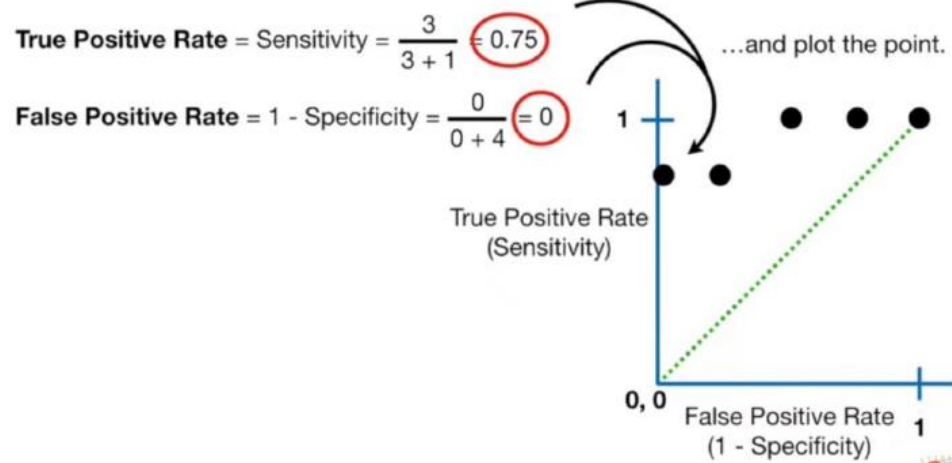
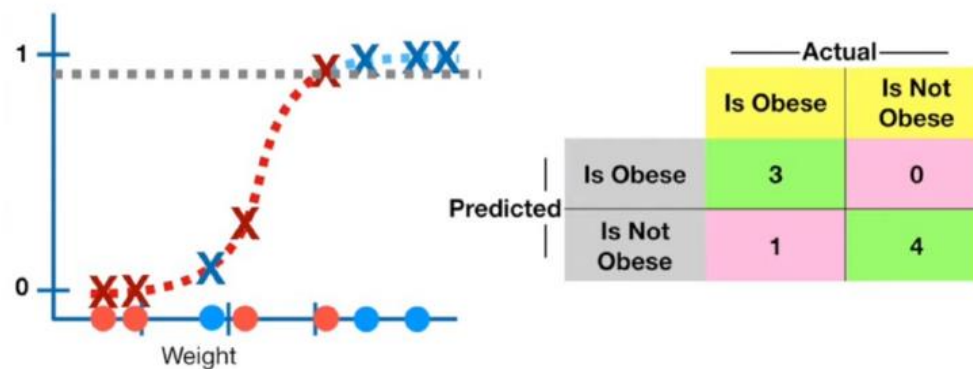
ROC 곡선

- 임계값을 더 증가해보자.



ROC 곡선

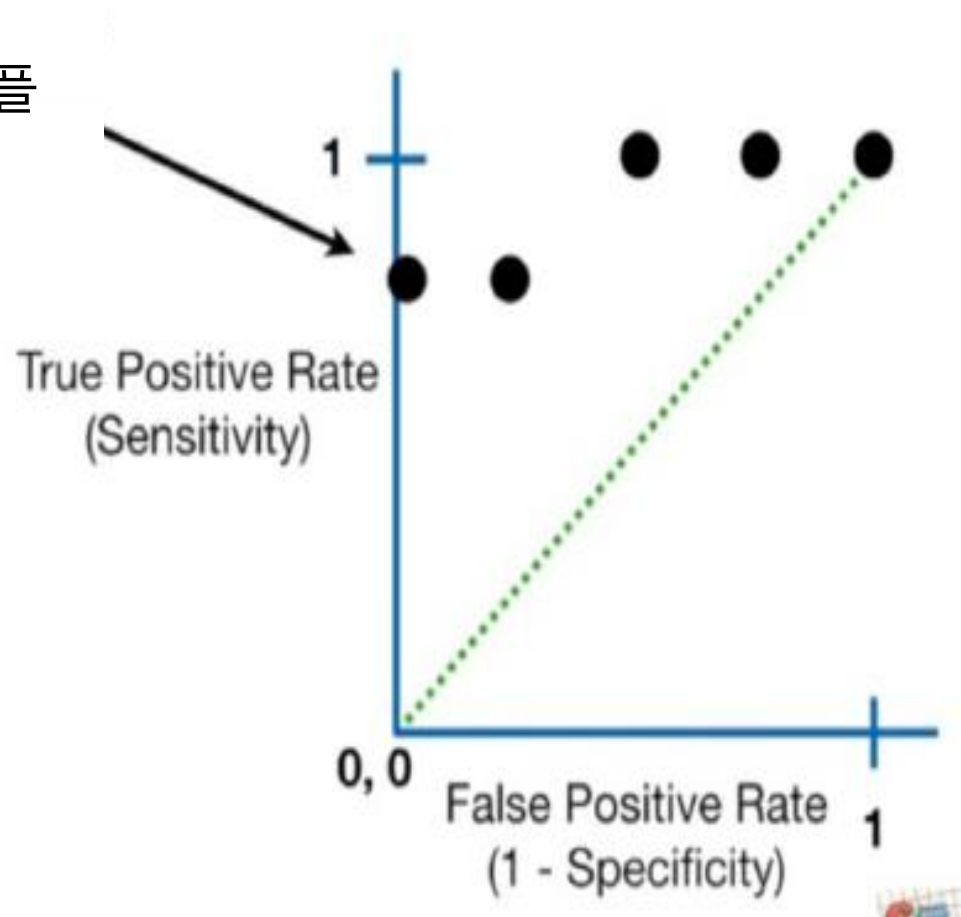
- 임계값을 더 증가해보자.



ROC 곡선

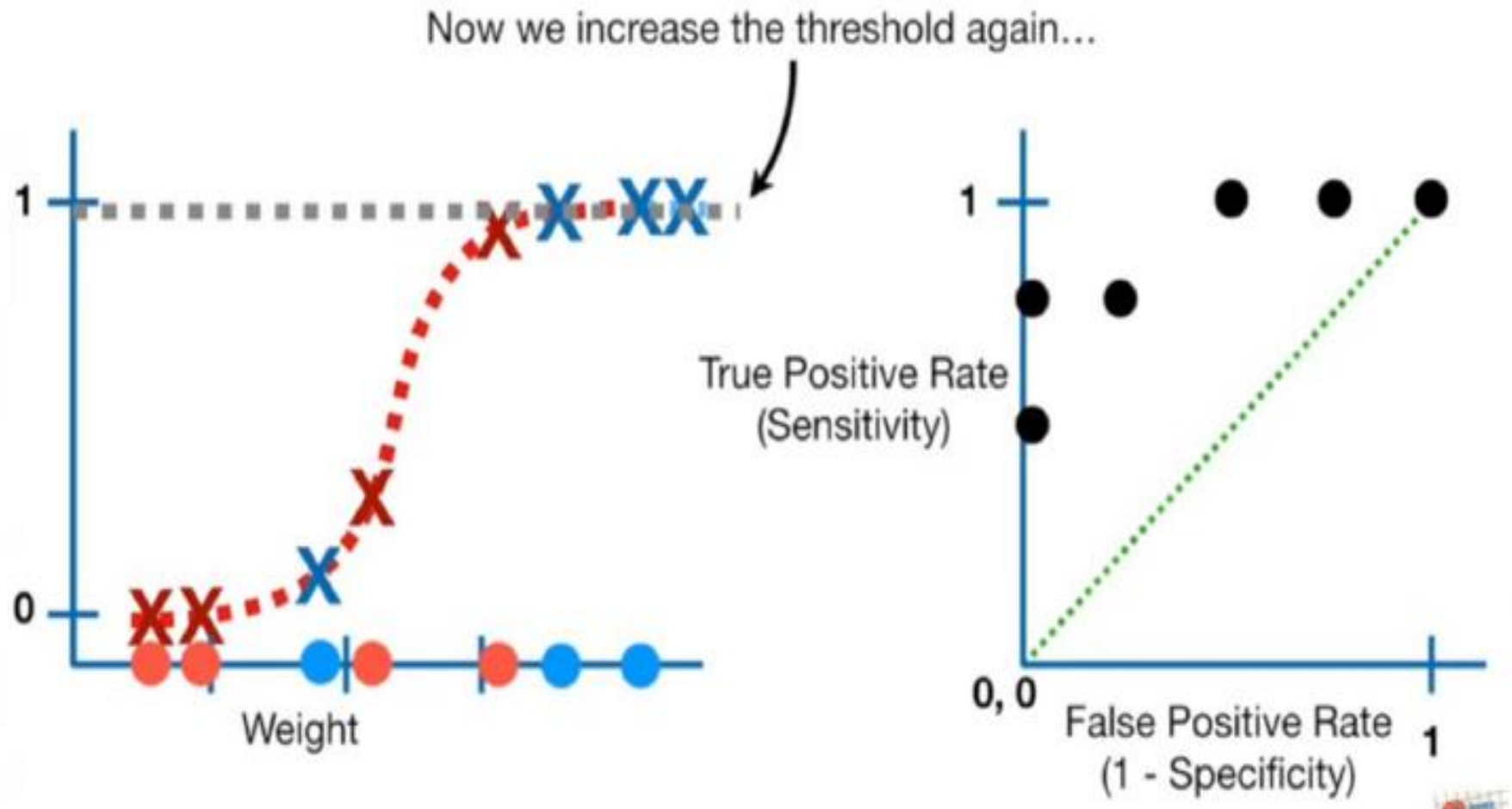
- 비만과 정상 분류 문제를 고려하자.

새로운 점 $(0, 0.75)$ 에 의해
표현되는 임계값은 75%의
비만샘플과 100%의 정상샘플
정확하게 분류한다.



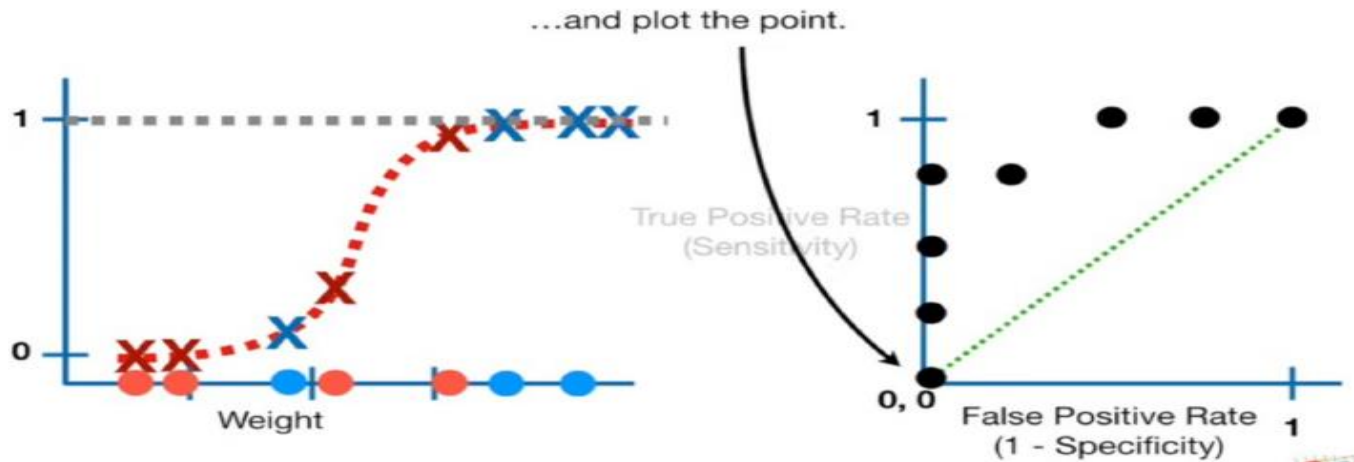
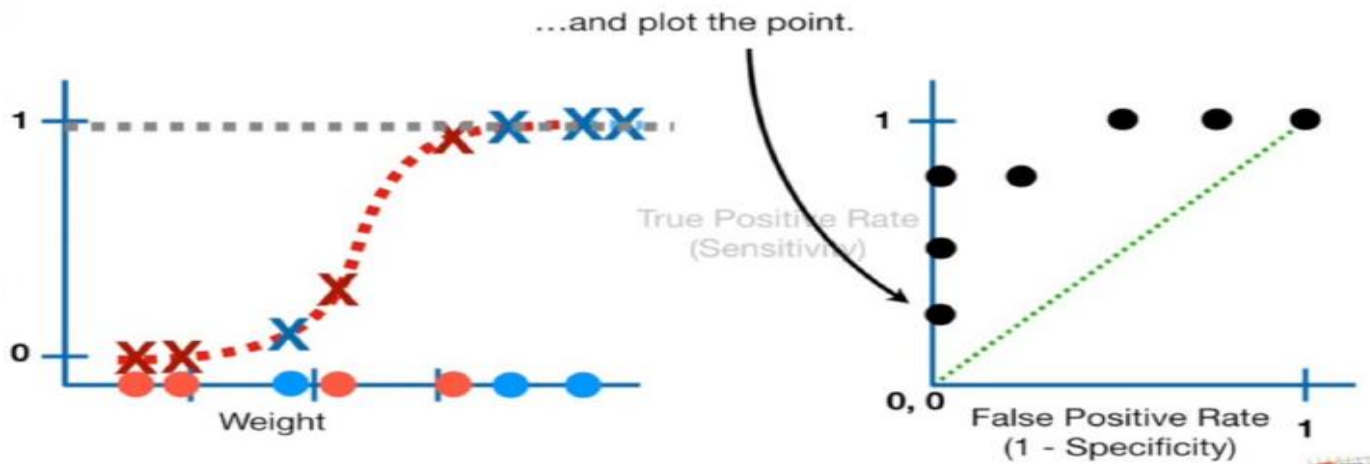
ROC 곡선

- 비만과 정상 분류 문제를 고려하자.



ROC 곡선

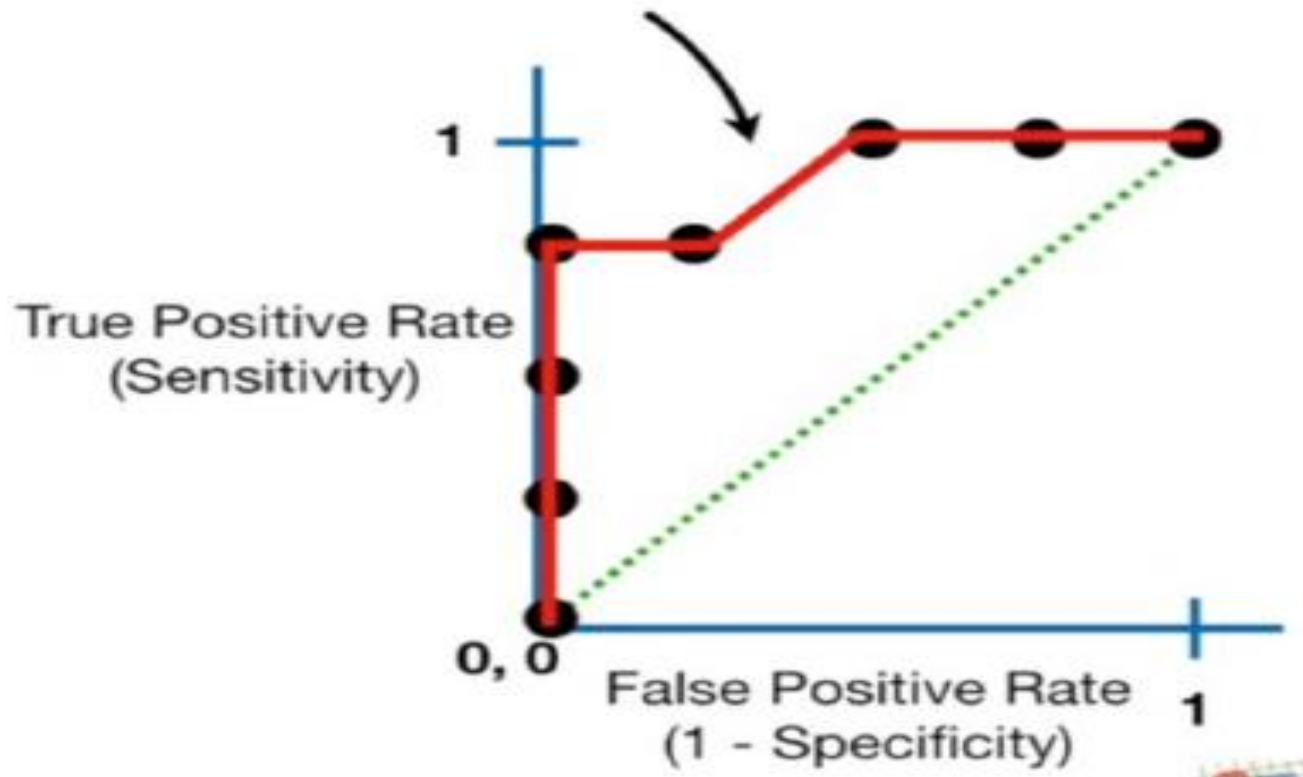
- 비만과 정상 분류 문제를 고려하자.



ROC 곡선

- 비만과 정상 분류 문제를 고려하자.

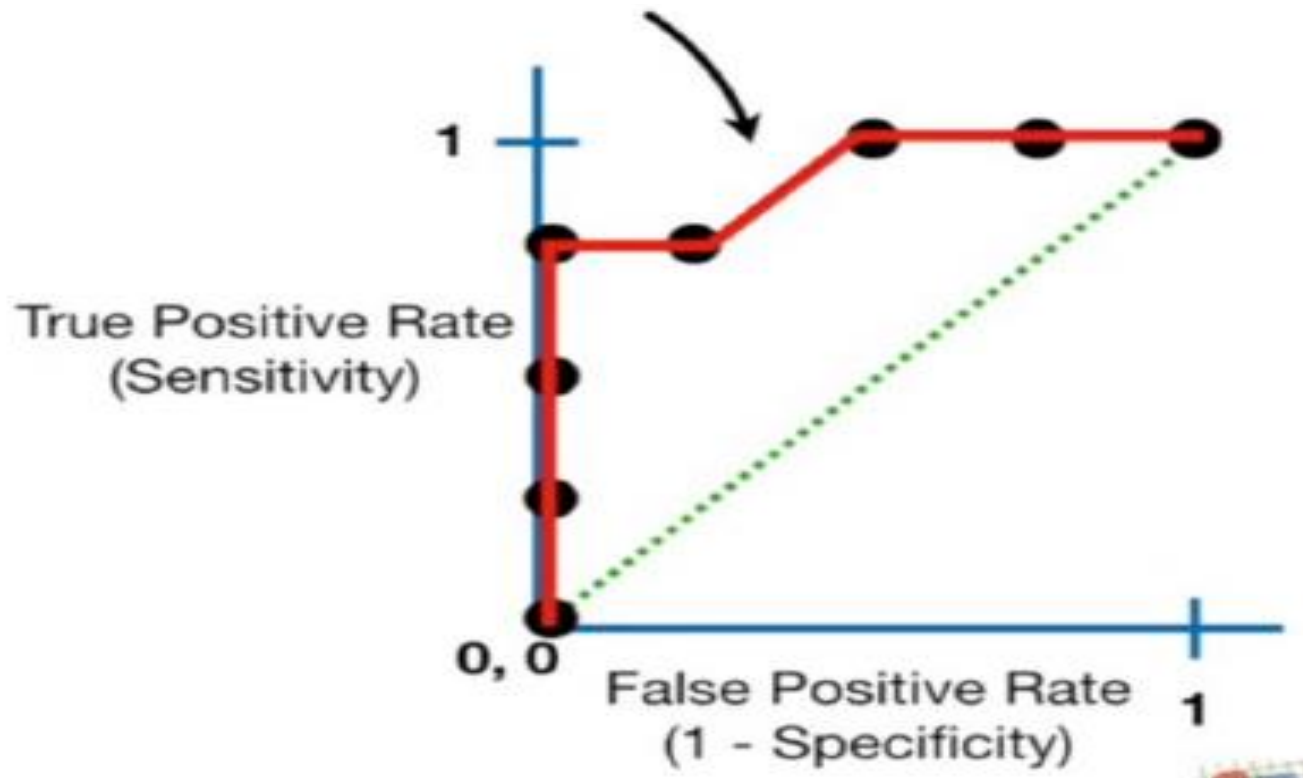
t, we can connect the dots...



ROC 곡선

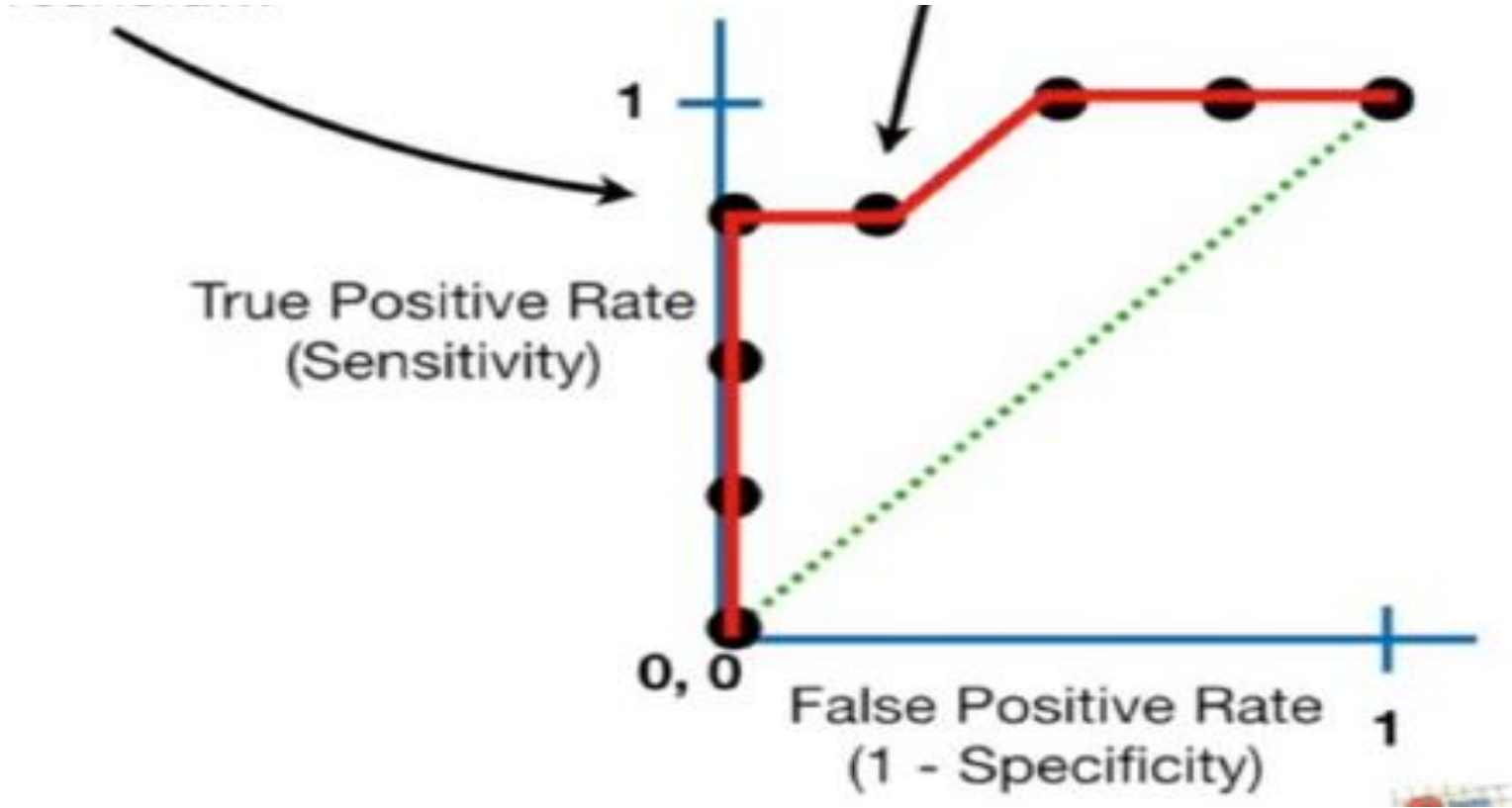
- 비만과 정상 분류 문제를 고려하자.

선을 이어보자



ROC 곡선

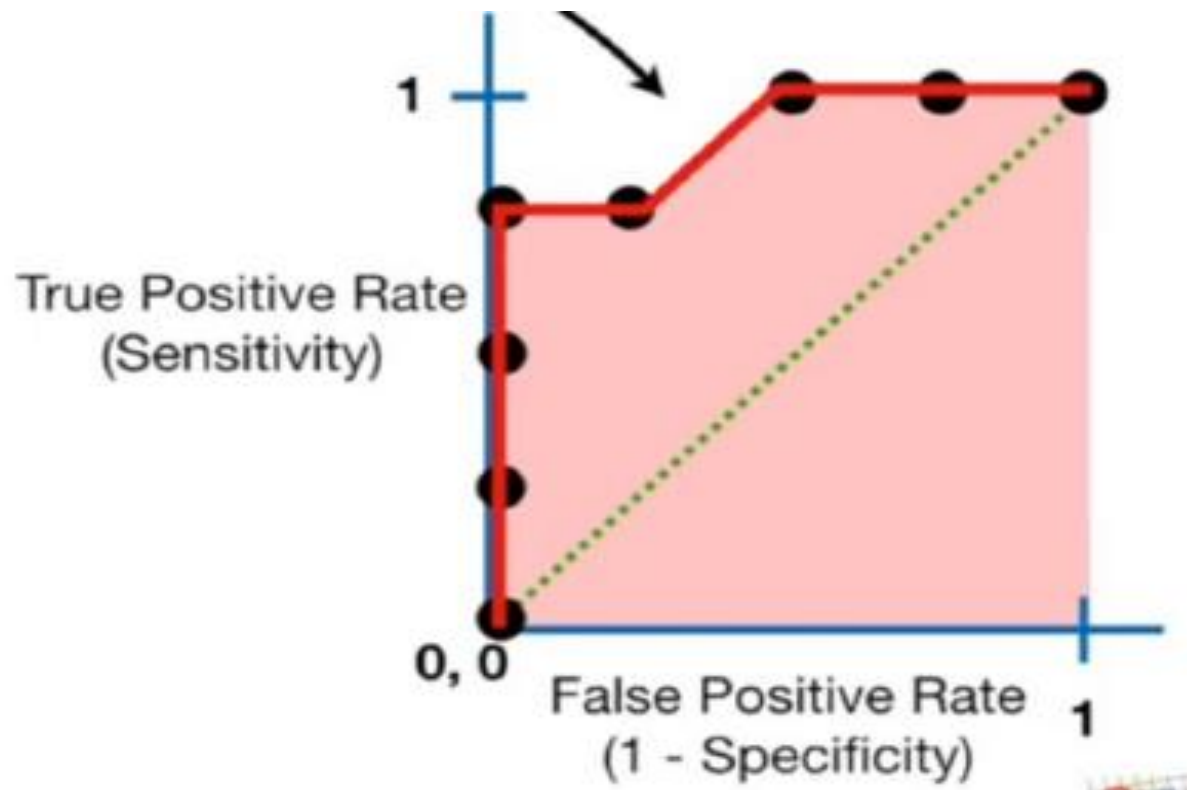
왼쪽의 임계값이 이 임계값보다 나음을 알 수 있다.



ROC AUC

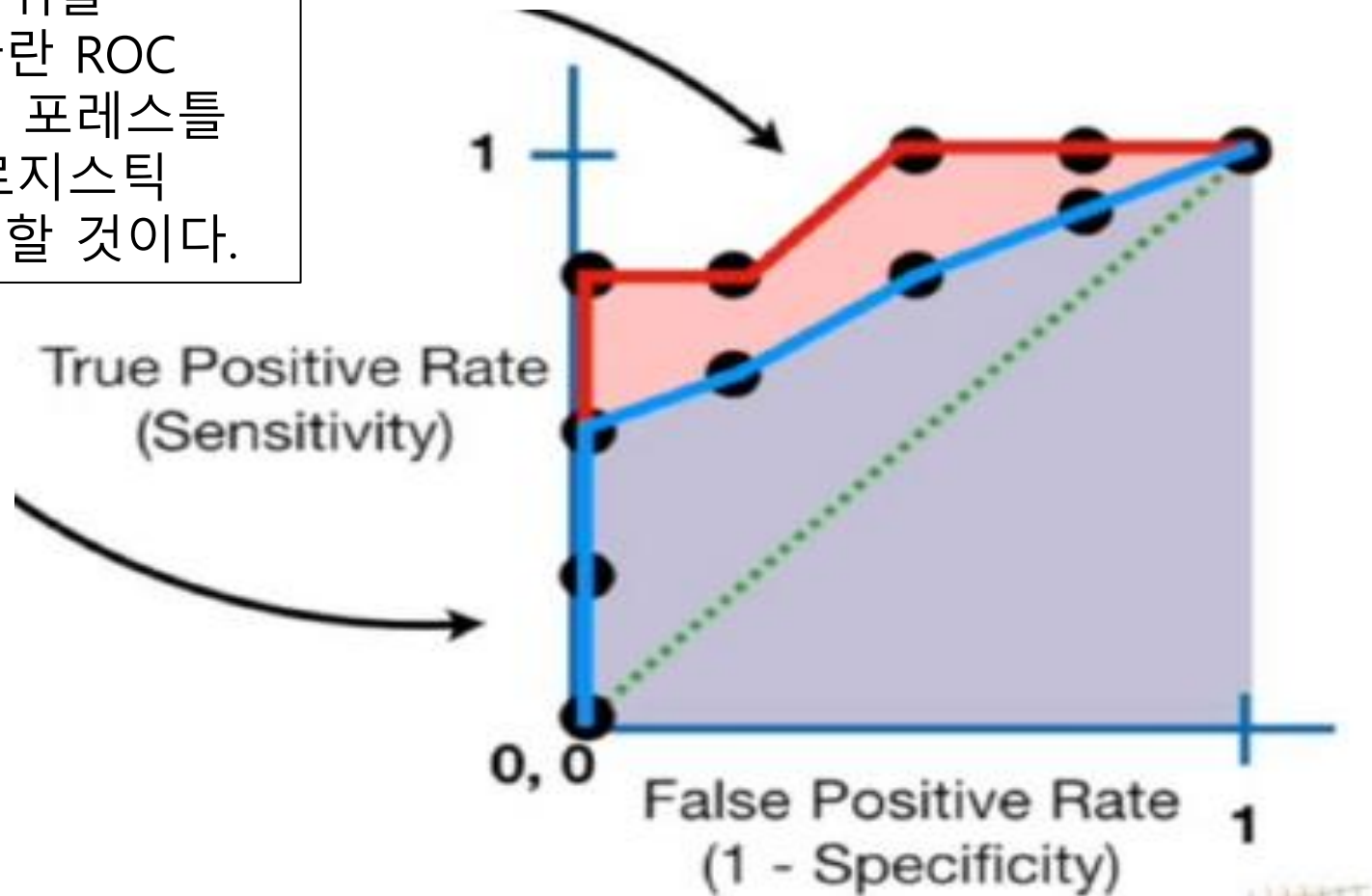
ROC AUC

AUC(곡선 아래의 면적)= 0.9



ROC AUC

빨간 ROC 곡선이
로지스틱 회귀를
나타내고 파란 ROC
곡선이 랜덤 포레스트를
나타내면, 로지스틱
회귀를 선택할 것이다.



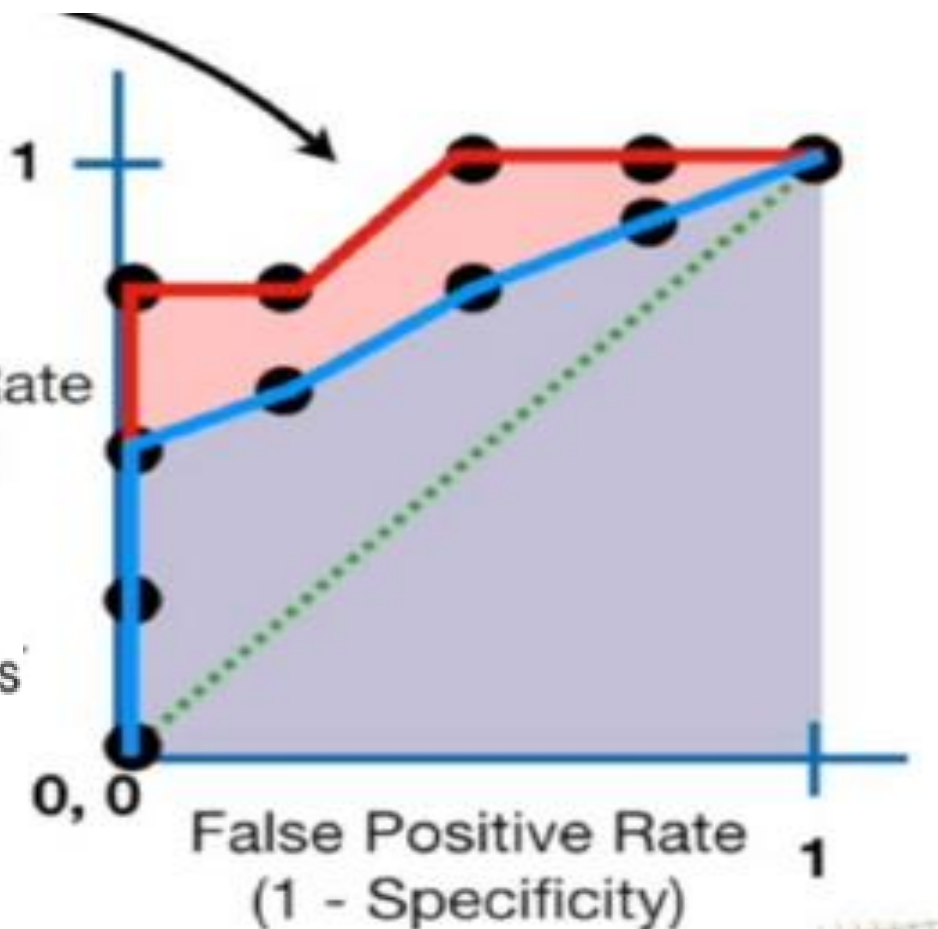
재현율과 정밀도 곡선

FPR 대신 정밀도(precision)를 사용할 수도 있다. 참음성을 계산에서 사용하지 않으므로 불균형에 영향을 받지 않는다.

True Positive Rate
(Sensitivity)

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

		Actual	
		Is Obese	Is Not Obese
Predicted	Is Obese	True Positives	False Positives
	Is Not Obese	False Negatives	True Negatives



분류 문제의 평가

분류문제의 평가

- 정확도와 혼동행렬을 구해보라.

Idx	y	y_hat
1	1	1
2	1	1
3	1	0
4	1	1
5	1	1
6	0	0
7	0	0
8	0	1
9	0	1
10	0	0

분류문제의 평가

□ 정확도

Idx	y	y_hat
1	1	1
2	1	1
3	1	0
4	1	1
5	1	1
6	0	0
7	0	0
8	0	1
9	0	1
10	0	0

Error = 3

Error rate

= Misclassification rate = $3/10 = 0.3$

Accuracy = $1 - \text{misclassification rate} = 0.7$