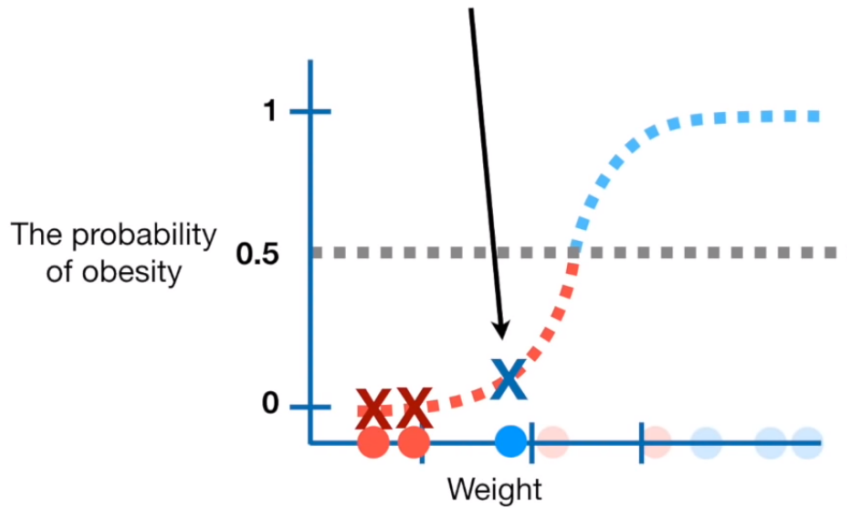


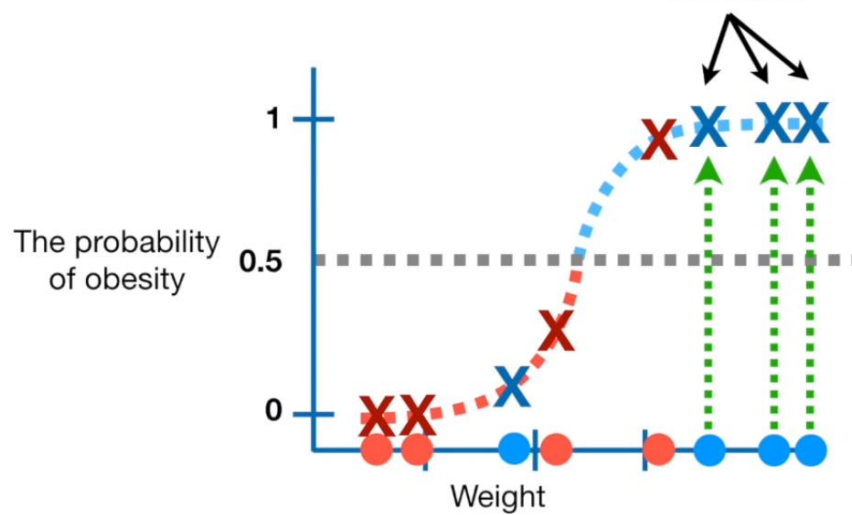
# AUC AND ROC

### A complete explanation with plotting

We know that it is **obese**, but it is classified as **not obese**.

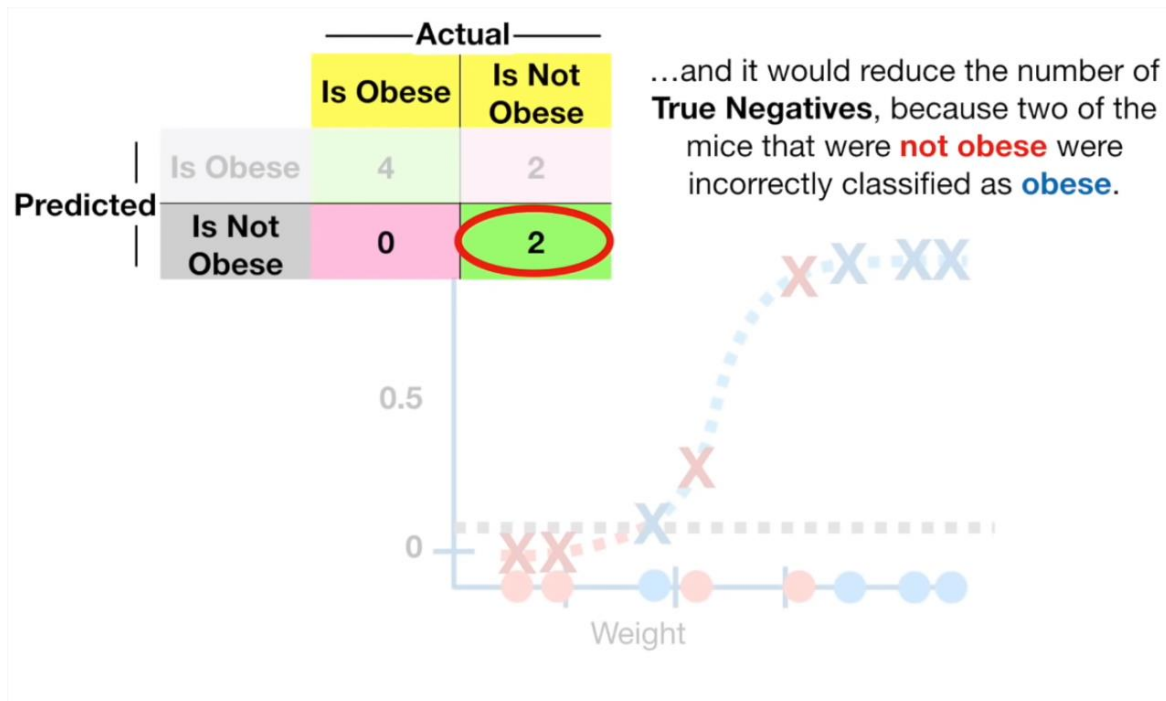


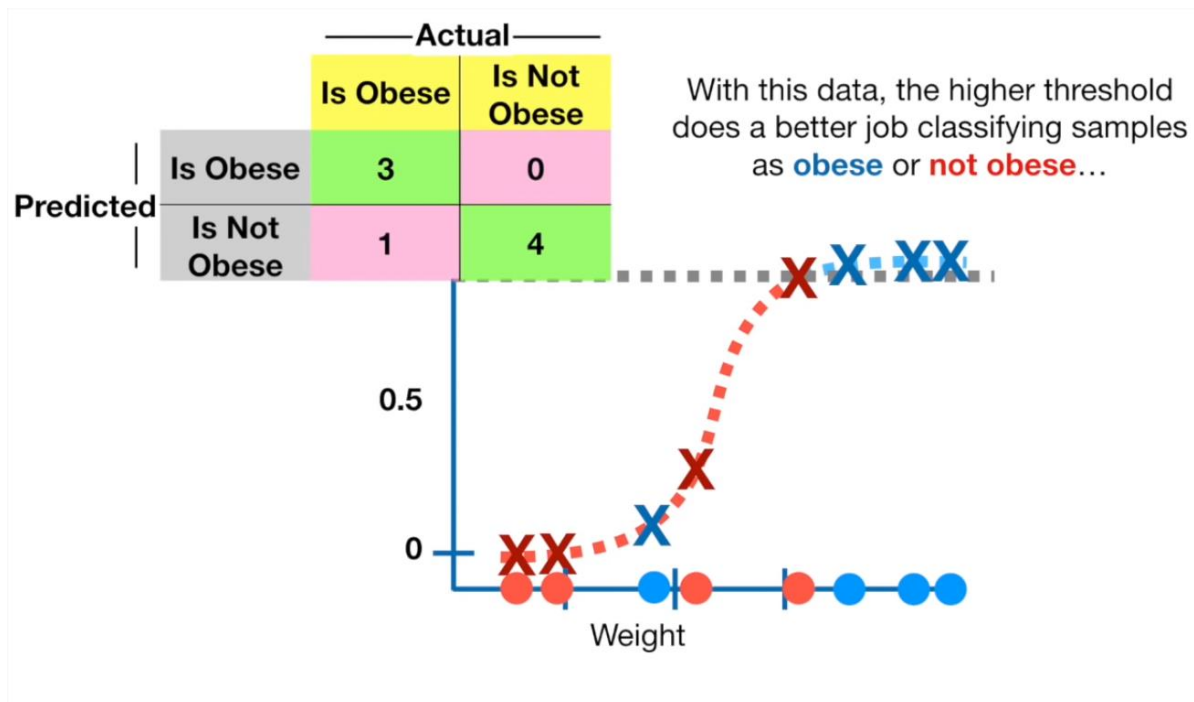
The last three mice are correctly classified.





If we change the threshold frequency:





Our main aim is to show the best threshold so that, we would don't get Type errors.

So making confusion matrix for each threshold is not valid like below:

But even if we made one confusion matrix for each threshold that mattered, it would result in a confusingly large number of confusion matrices.

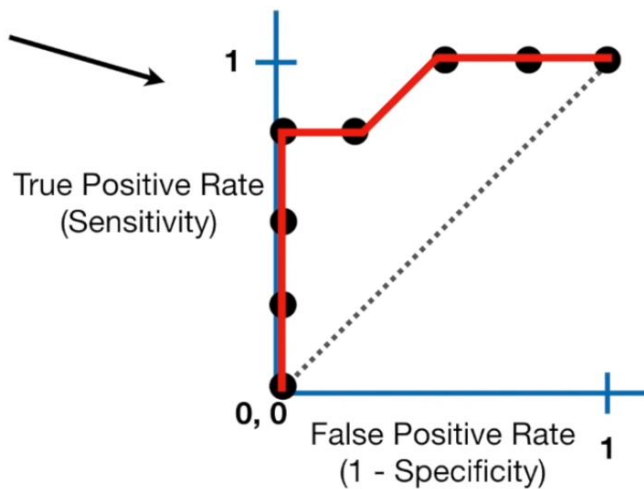
	Is Obese	Is Not Obese		Is Obese	Is Not Obese		Is Obese	Is Not Obese
Is Obese	4	1	Is Obese	4	2	Is Obese	3	1
Is Not Obese	0	1	Is Not Obese	0	2	Is Not Obese	1	3

	Is Obese	Is Not Obese		Is Obese	Is Not Obese		Is Obese	Is Not Obese
Is Obese	4	2	Is Obese	3	2	Is Obese	4	0
Is Not Obese	0	1	Is Not Obese	1	2	Is Not Obese	0	4

Rather than doing all the calculations, we for a technique called **ROC, AUC**.

So instead of being overwhelmed with confusion matrices, **Receiver Operator Characteristic (ROC)** graphs provide a simple way to summarize all of the information.



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

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

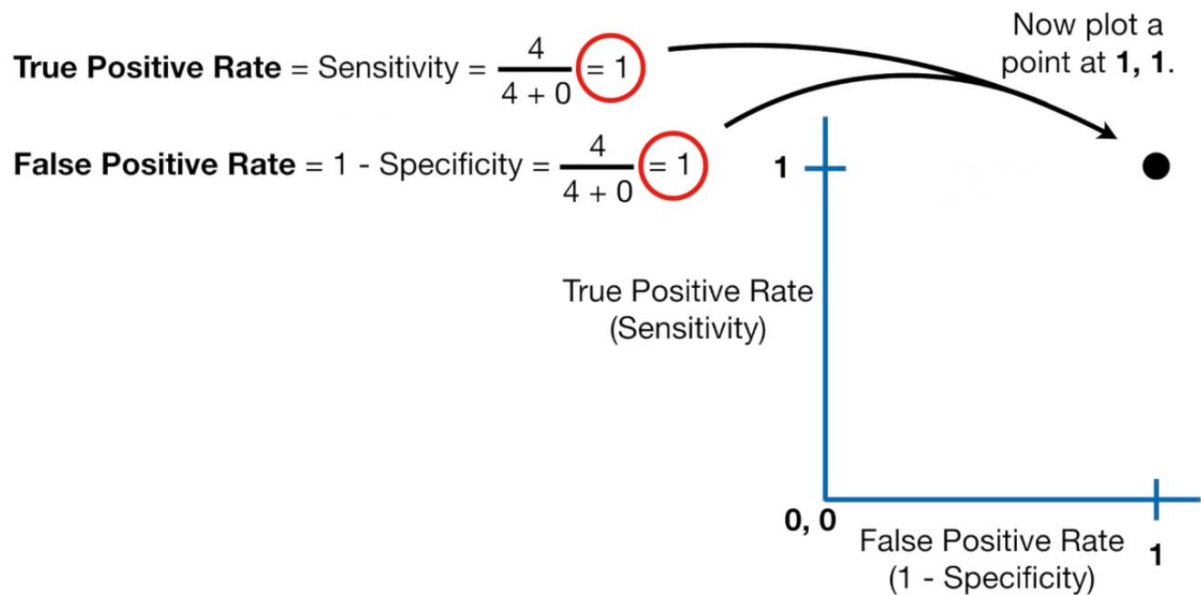
$$\text{True Positive Rate} = \text{Sensitivity} = \frac{4}{4 + 0} = 1$$

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

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

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

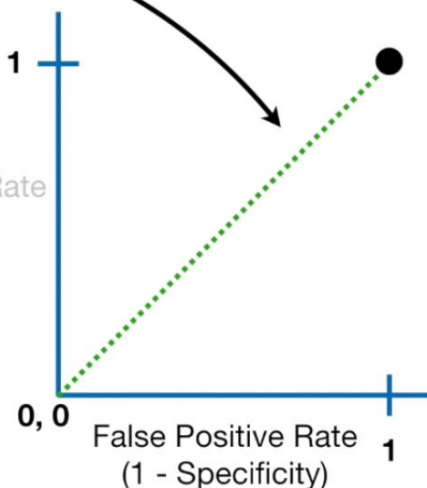
Now let's go to plot the point (1, 1)



This **green diagonal line** shows where the True Positive Rate = False Positive Rate

Any point on this **line** means that the proportion of **correctly** classified **obese** samples is the same as the proportion of **incorrectly** classified samples that are **not obese**.

True Positive Rate (Sensitivity)



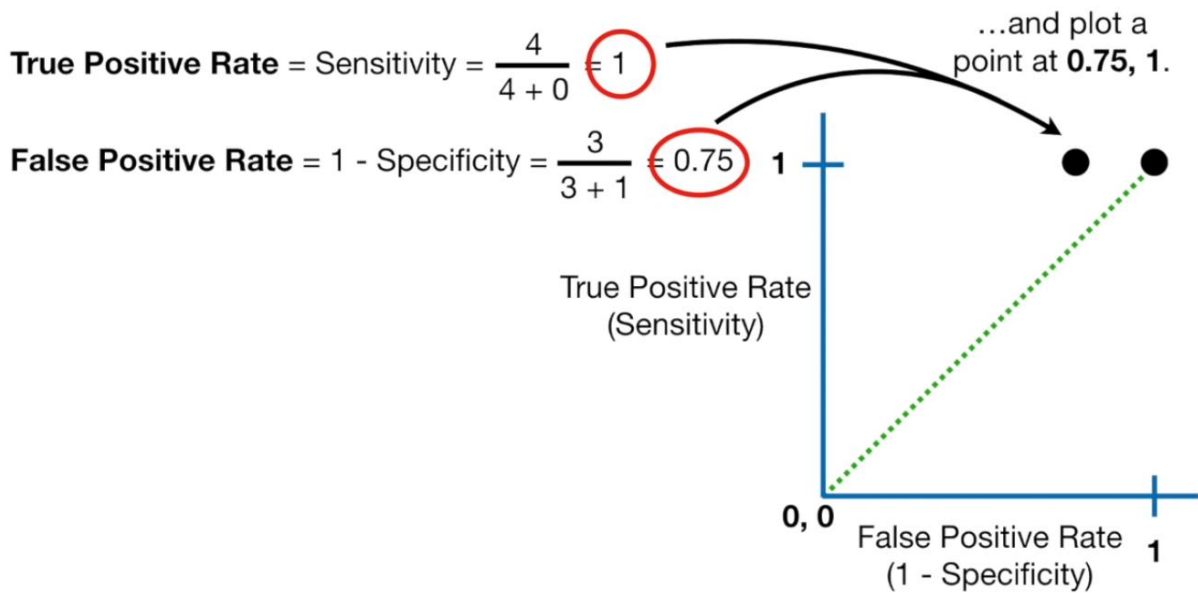
For Another confusion matrix of different threshold:

**True Positive Rate = Sensitivity** =  $\frac{4}{4+0} = 1$  ...and the **False Positive Rate...**

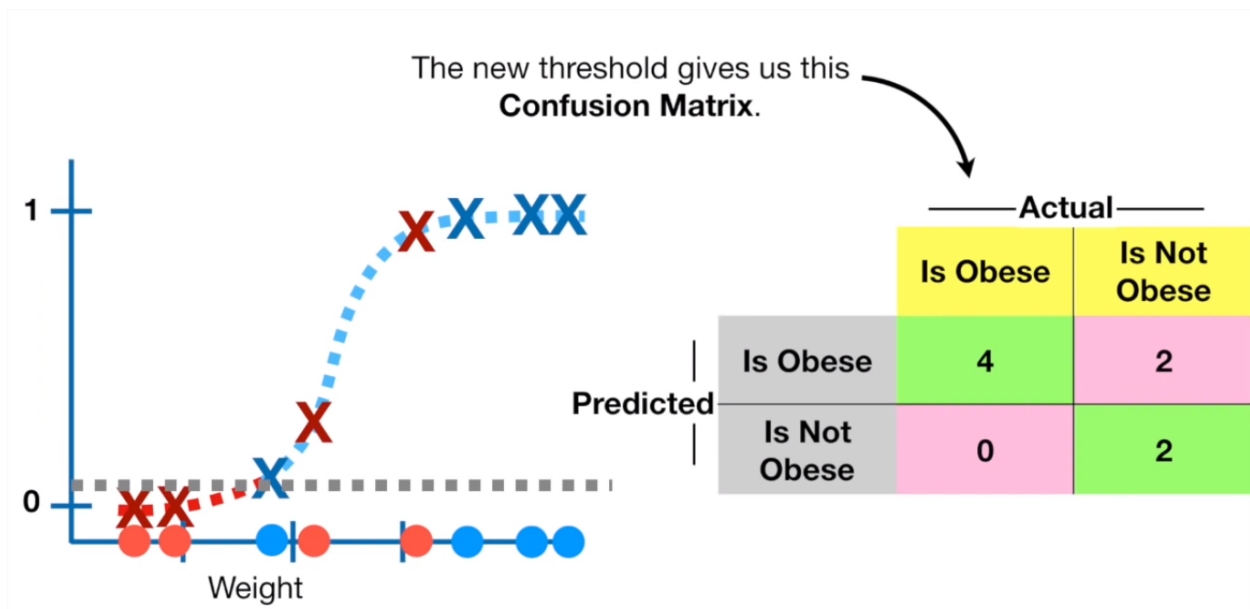
**False Positive Rate** =  $1 - \text{Specificity} = \frac{3}{3+1} = 0.75$

		Actual	
		Is Obese	Is Not Obese
Predicted	Is Obese	4	3
	Is Not Obese	0	1

Let's again plot it on the graph (0.75, 1):



Again let's set new threshold matrix and see the confusion matrix:



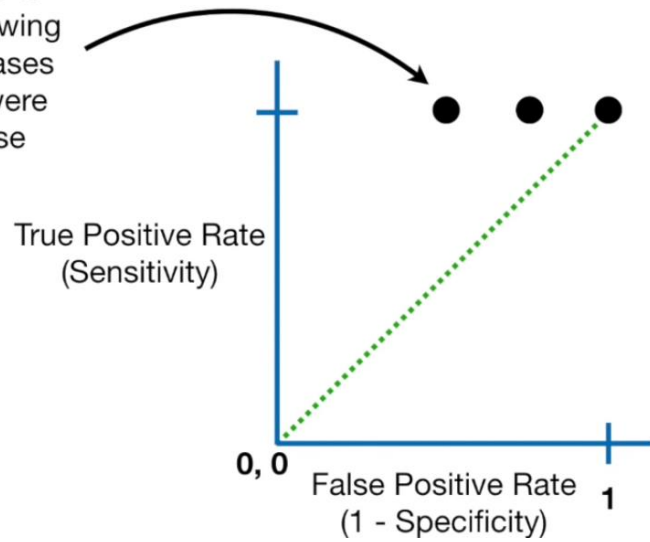
**True Positive Rate** = Sensitivity =  $\frac{4}{4 + 0} = 1$  ...and the **False Positive Rate**...

**False Positive Rate** = 1 - Specificity =  $\frac{2}{2 + 2} = 0.5$

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

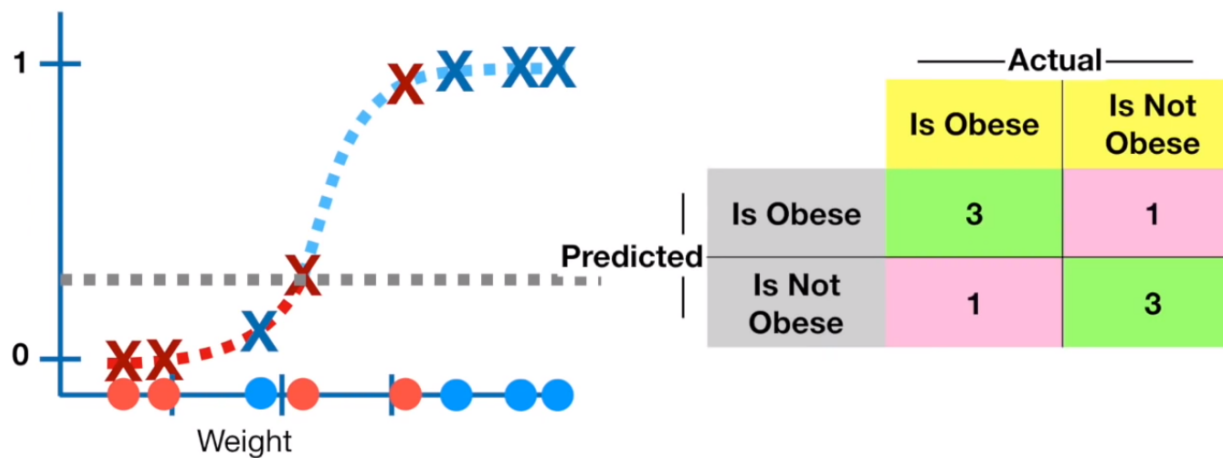
Let's again plot points on (0.5, 1):

The new point (0.5, 1) is even further to the left of the **dotted green line**, showing that the new threshold further decreases the proportion of the samples that were *incorrectly* classified as **obese** (false positives).



Again we change the threshold and calculate the confusion matrix:

...create a confusion matrix...



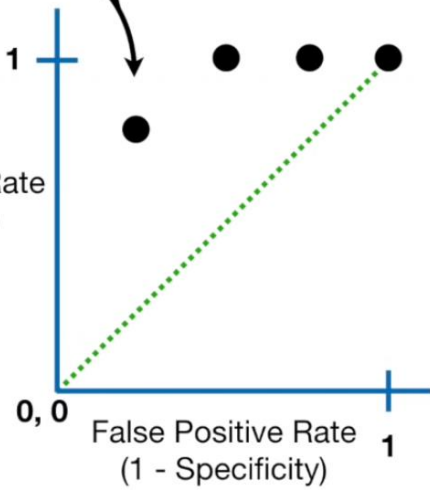


**True Positive Rate** = Sensitivity =  $\frac{3}{3+1} = 0.75$

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

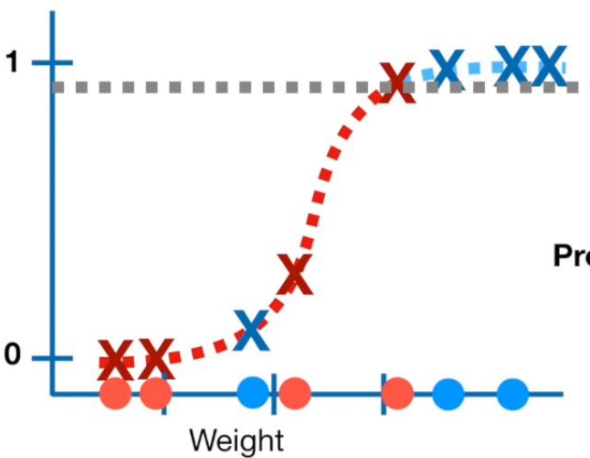
...and plot the point.

True Positive Rate  
(Sensitivity)



Again we change the threshold with calculations:

...create a confusion matrix...



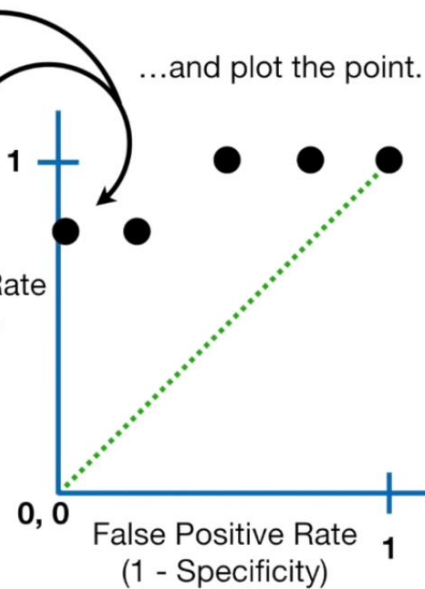
		Actual	
		Is Obese	Is Not Obese
Predicted	Is Obese	3	0
	Is Not Obese	1	4

**True Positive Rate = Sensitivity** =  $\frac{3}{3+1} = 0.75$

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

...and plot the point.

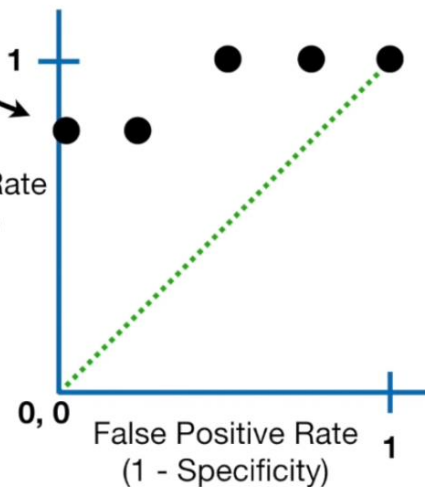
True Positive Rate  
(Sensitivity)



The threshold represented by the new point **(0, 0.75)** correctly classified **75%** of the **obese** samples and **100%** of the samples that were **not obese**.

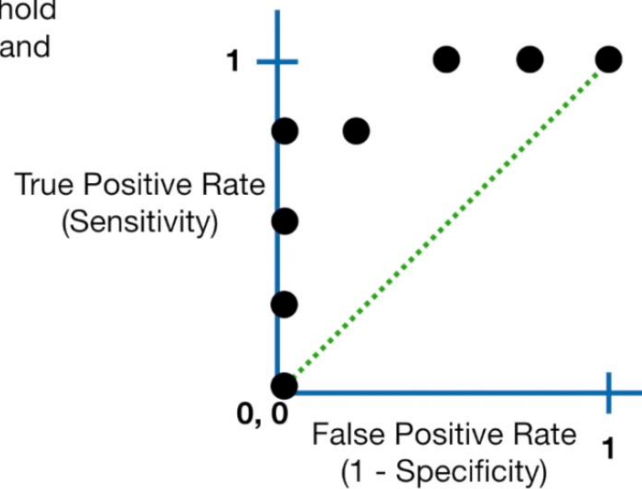
In other words, this threshold resulted in no **False Positives**.

True Positive Rate  
(Sensitivity)



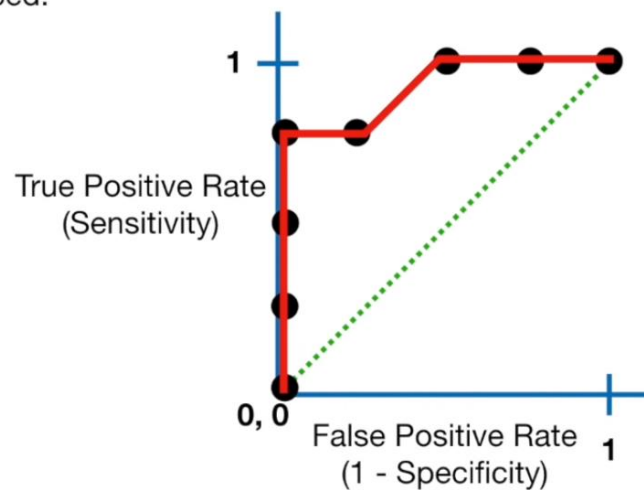
If we get all False-Positive (Error) = 0, we got touched to Sensitivity. But if we increase the threshold again going to decrease our accuracy. And we touch the origin.

The point at **0, 0** represents a threshold that results in zero **True Positives** and zero **False Positives**.



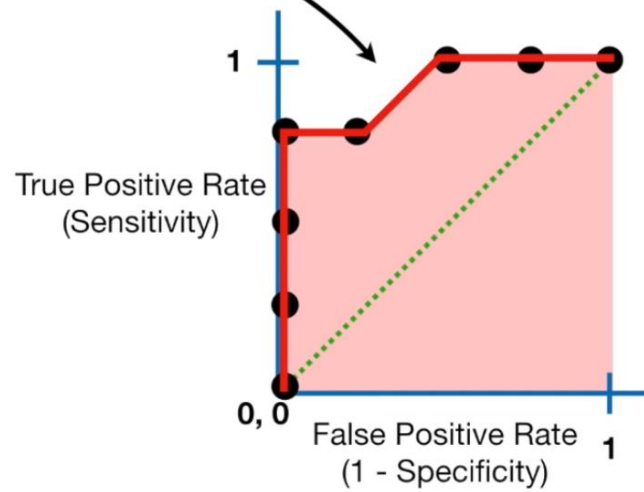
If we connect these lines we get ROC (Receiver Operator Characteristics) graph.

The **ROC** graph summarizes all of the confusion matrices that each threshold produced.

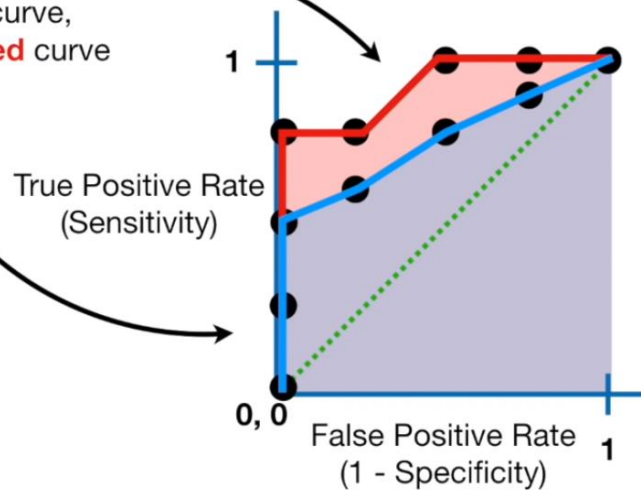


Now what is AUC (Area Under the Curve)?

The **AUC** (Area Under the Curve) is **0.9**

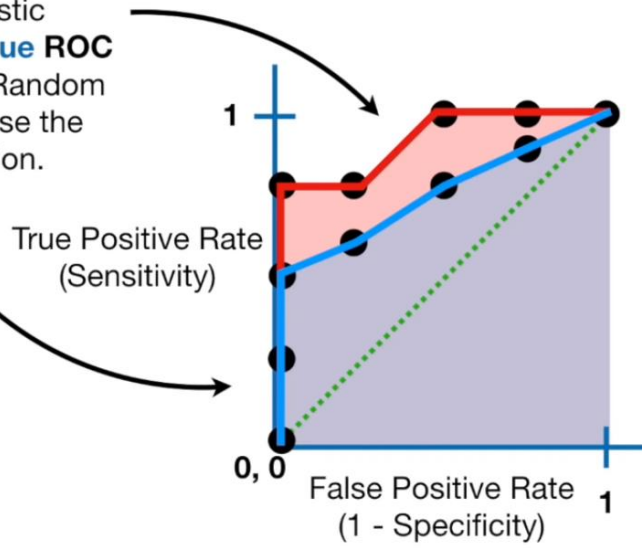


The **AUC** for the **red ROC** curve is greater than the **AUC** for the **blue ROC** curve, suggesting that the **red** curve is better.



**Conclusion:**

So if the **red ROC** curve represented Logistic Regression and the **blue ROC** curve represented a Random Forest, you would use the Logistic Regression.



ASAD ASHRAF KAREL