Classification Metrics

- Classification Accuracy
- Confusion matrix
- Logarithmic Loss
- Area under curve (AUC)
- F-Measure or F1 Score

Classification Accuracy

- Accuracy is a common evaluation metric for classification problems
- Number of correct predictions made as a ratio of all predictions made

$$\text{Accuracy} = \frac{\text{No. of Correct Predictions}}{\text{Total No. of Predictions Made}}$$

- Disadv: Accuracy will yield misleading results if the data set is unbalanced
- Eg: If there were 95 cats and only 5 dogs in the data, a particular classifier might classify all the observations as cats
  - Overall accuracy would be 95%, but in more detail the classifier would have a 100% recognition rate (sensitivity) for the cat class but a 0% recognition rate for the dog class

Confusion Matrix

- TP, TN, FP, FN

- Classification system trained to distinguish between cats and dogs

- Sample of 13 animals —> 8 cats and 5 dogs

-> Actual = [1,1,1,1,1,1,1,1, 0,0,0,0,0]

-> Prediction = [0,0,0,1,1,1,1,1, 0,0,0,1,1]

	Actual Class	
		Cat	**Dog**
Predicted Class	**Cat**	**5**	2
	Dog	3	**3**

	Actual Class	
		Cat	**Non-cat**
Predicted Class	**Cat**	**5 TP**	2 FP
	Non-cat	3 FN	**3 TN**

Accuracy

- **Accuracy:** Measure of the effectiveness of the algorithm

$$Accuracy(ACC) = \frac{\sum TP + \sum TN}{\sum TP + TN + FP + FN} = \frac{5+3}{5+3+2+3} = 61.5\%$$

	Actual Class	
Predicted Class		**Cat**	**Dog**
	Cat	5	2
	Dog	3	3

	Actual Class	
Predicted Class		**Cat**	**Non-cat**
	Cat	5 TP	2 FP
	Non-cat	3 FN	3 TN

Drawback of Accuracy

$$Accuracy(ACC) = \frac{\sum TP + \sum TN}{\sum TP + TN + FP + FN} = \frac{5+3}{5+3+2+3} = 61.5\%$$

	Actual Class	
Predicted Class	10 TP	15 FN
	25 FP	100 TN

- No. of Samples = 150
- Class 1 = 35 *(Aim to detect)*
- Class 0= 115

- Accuracy = (10 + 100) / (10 + 100 + 25 + 15) = 73.3%

	Actual Class	
Predicted Class	0 TP	25 FN
	0 FP	125 TN

	Actual Class	
Predicted Class	0 TP	35 FN
	0 FP	115 TN

- Accuracy = (0 + 115) / (0 + 115 + 0 + 35) = 76.7%

Precision & Recall

- **Precision:** Ratio of correctly predicted positive values to the total predicted positive values

'how much the model is right when it says it is right'

$$Precision = \frac{TP}{TP+FP} = \frac{5}{5+2} = 71.4\%$$

- **Recall/Sensitivity/True Positive Rate:** Percentage of total relevant results correctly classified by the algorithm

'how much extra right ones, the model missed when it showed the right ones'

$$Recall = \frac{TP}{TP+FN} = \frac{5}{5+3} = 62.5\%$$

	Actual Class	
Predicted Class	**Cat**	**Dog**
Cat	5	2
Dog	3	3

Precision and Recall are used when class imbalance is present and also the detection of positive classes is very important

	Actual Class	
Predicted Class	**Cat**	**Non-cat**
Cat	5 TP	2 FP
Non-cat	3 FN	3 TN

F1 Score

- **F1 Score:** Weighted harmonic mean of Precision and Recall

- Measure of a test's accuracy that considers both precision and recall of the test to compute the score

$$F1\,Score = 2 * \frac{Recall * Precision}{Recall + Precision} = 2 * \frac{0.714 * 0.625}{0.714 + 0.625} = 66.6\%$$

	Actual Class	
Predicted Class	**Cat**	**Dog**
Cat	5	2
Dog	3	3

	Actual Class	
Predicted Class	**Cat**	**Non-cat**
Cat	5 TP	2 FP
Non-cat	3 FN	3 TN

Specificity

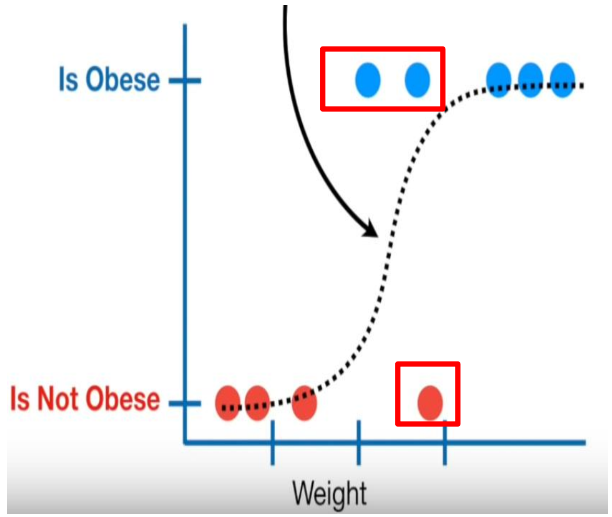
- Percentage of negative instances out of the *total actual negative* instances

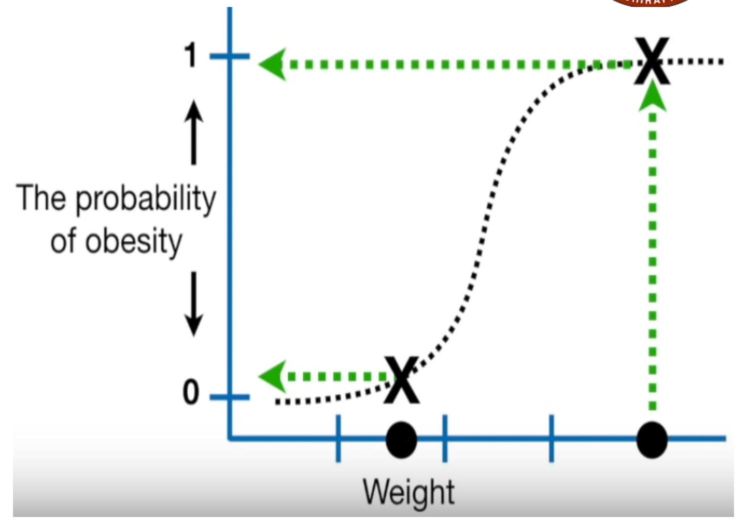
Like finding out how many healthy patients were not having cancer and were told they don't have cancer

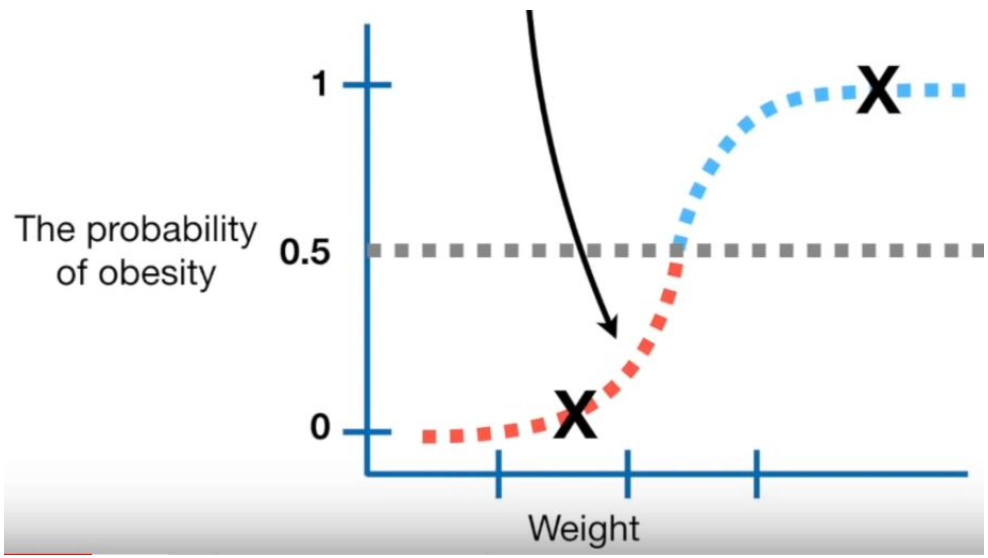
$$\textit{Specificity} = \frac{TN}{TN + FP}$$

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

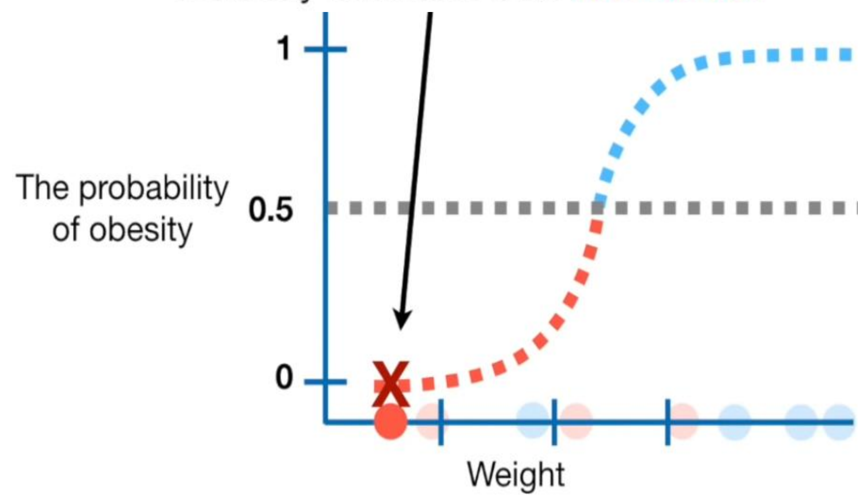
Receiver Operator Characteristic (ROC)



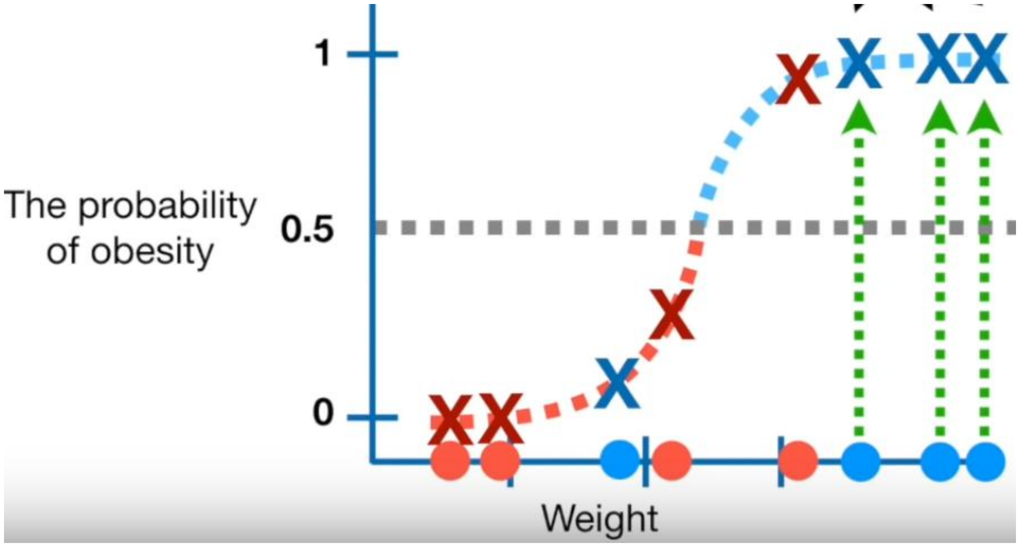




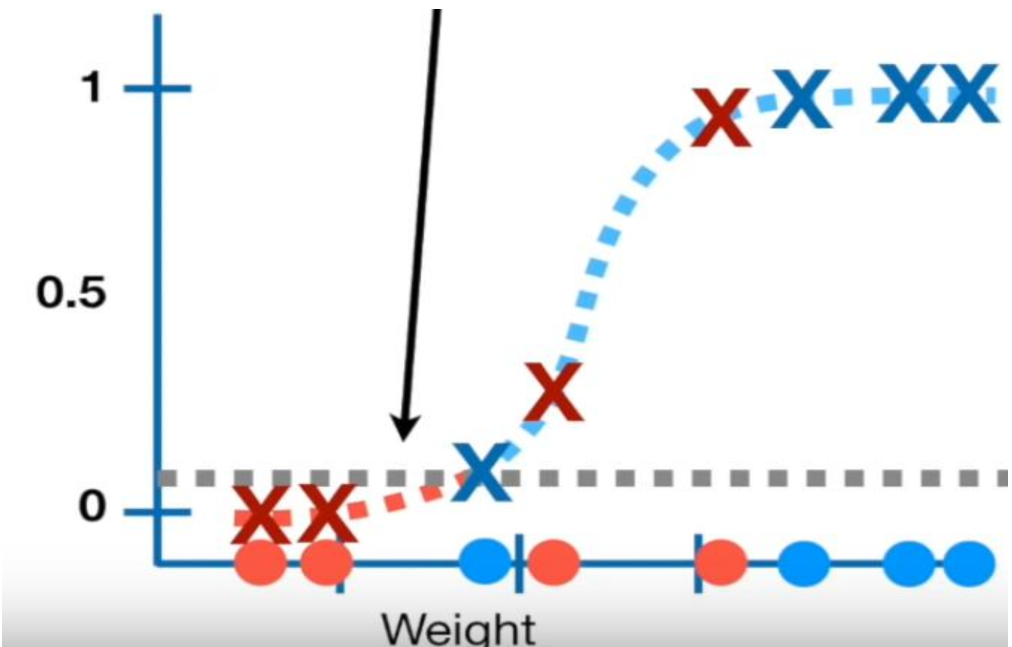
…and the Logistic Regression, with the classification threshold set to **0.5**, correctly classifies it as **not obese**.



ROC

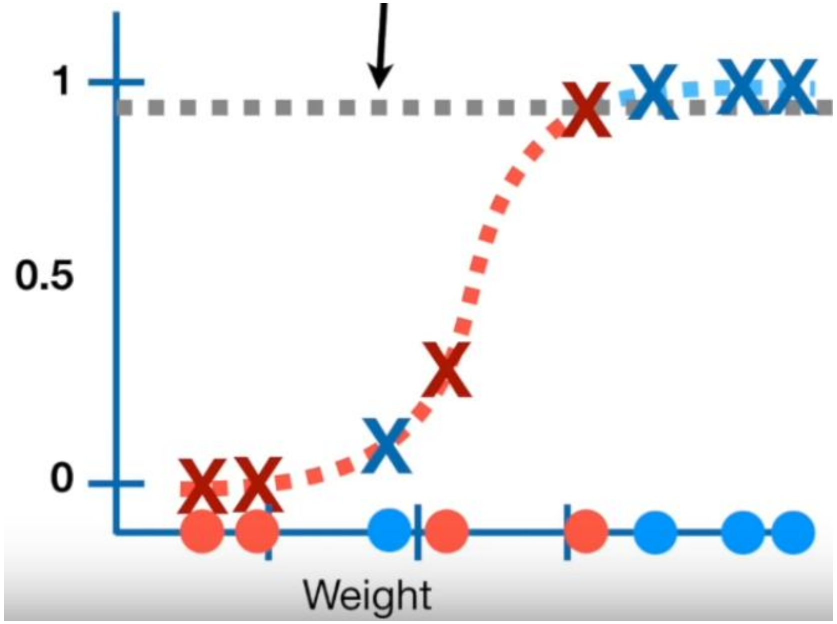


		Actual: Is Obese	Actual: Is Not Obese
Predicted	Is Obese	3	1
	Is Not Obese	1	3



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

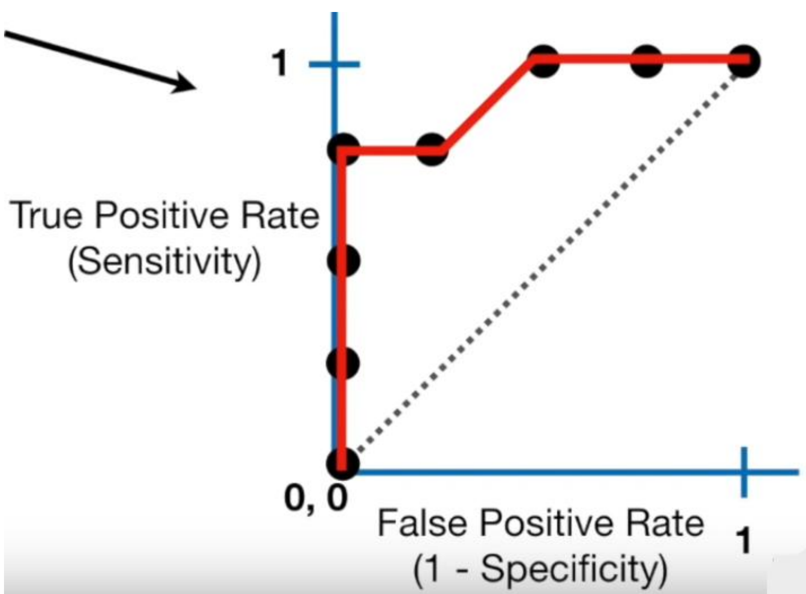
ROC



Instead of being overwhelmed with Confusion Matrices, ROC graphs provide a simple way to summarize all of the information

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

$$Recall = \frac{TP}{TP+FN} = \frac{5}{5+3} = 62.5\%$$



TPR tells you what proportion of obese samples that were correctly classified

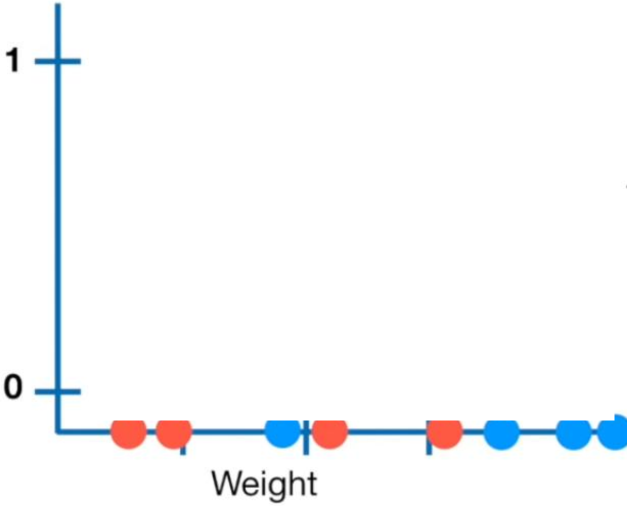
		Actual	
		Is Obese	Is Not Obese
Predicted	Is Obese	True Positives	False Positives
	Is Not Obese	False Negatives	True Negatives

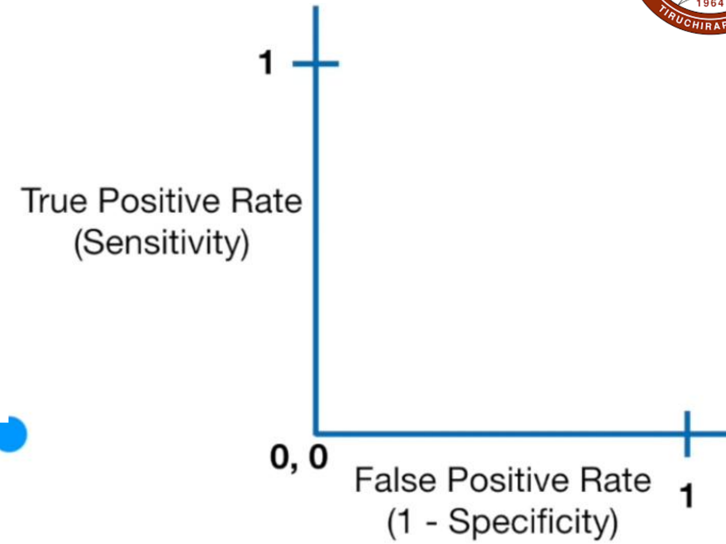
ROC

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

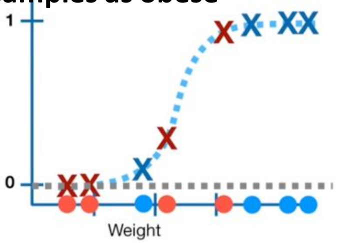
FPR tells you the proportion of not obese samples that were incorrectly classified and are FP

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



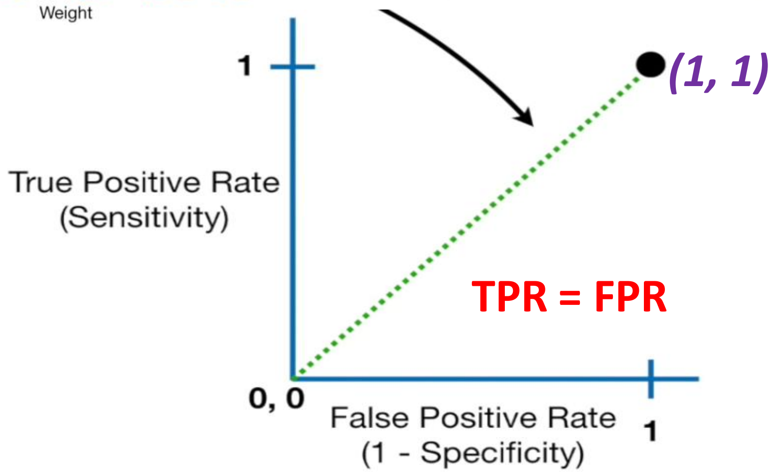


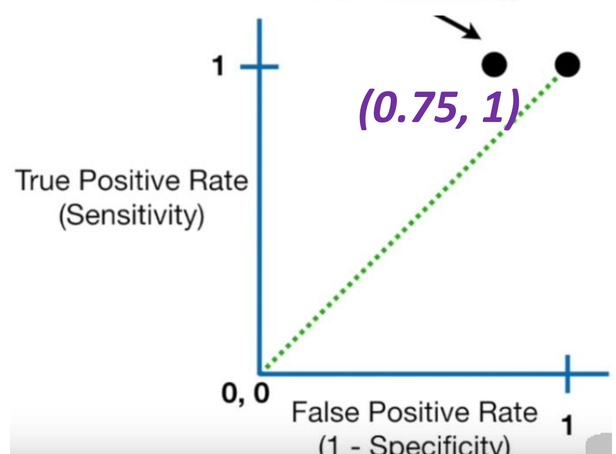
Choose a threshold that classifies all the samples as obese



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

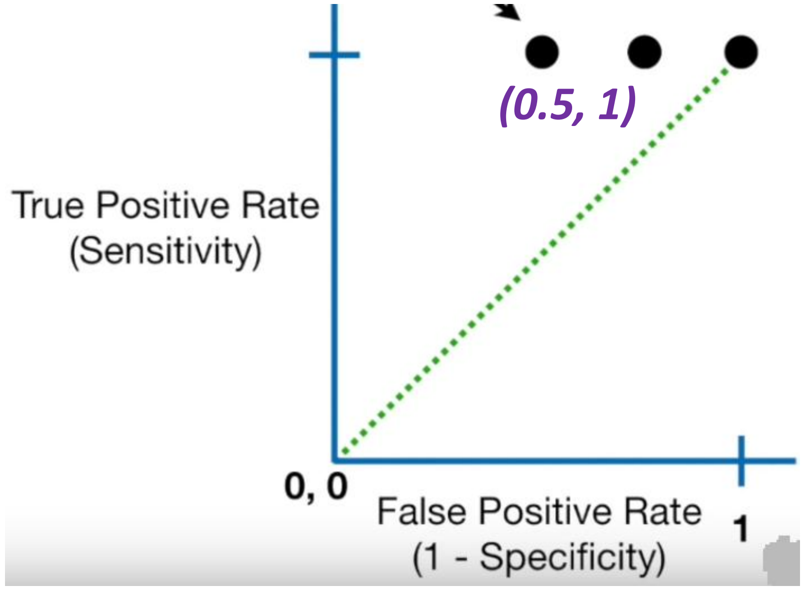
		Actual: Is Obese	Actual: Is Not Obese
Predicted	Is Obese	4	3
	Is Not Obese	0	1



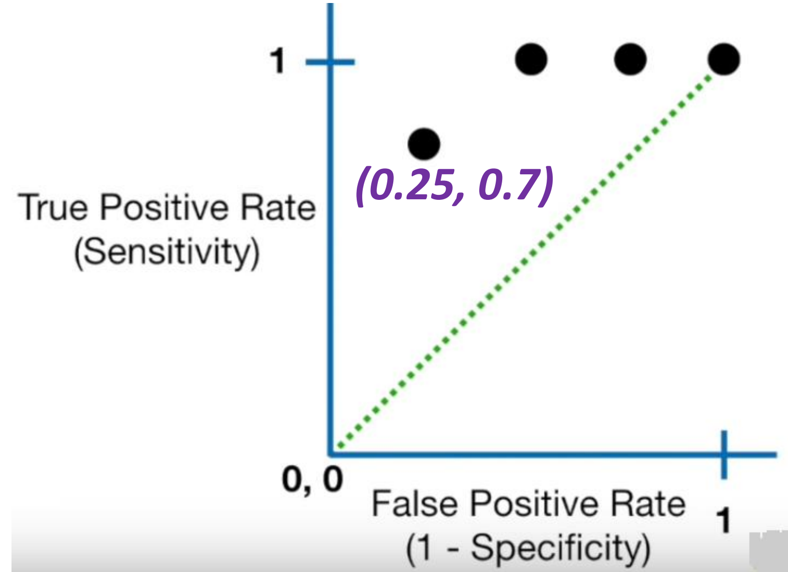


ROC

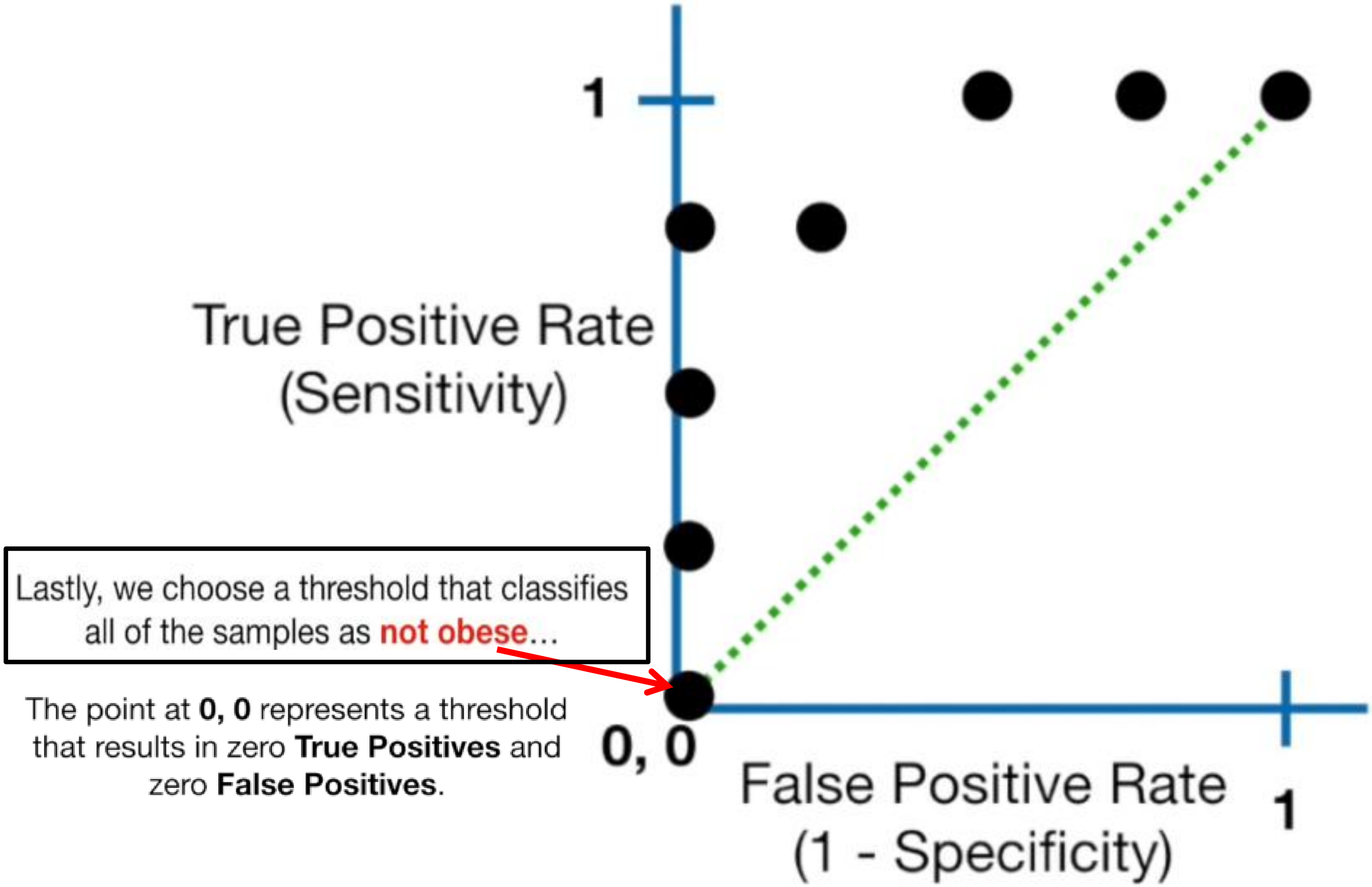
		Actual	
		Is Obese	Is Not Obese
Predicted	Is Obese	4	2
	Is Not Obese	0	2



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

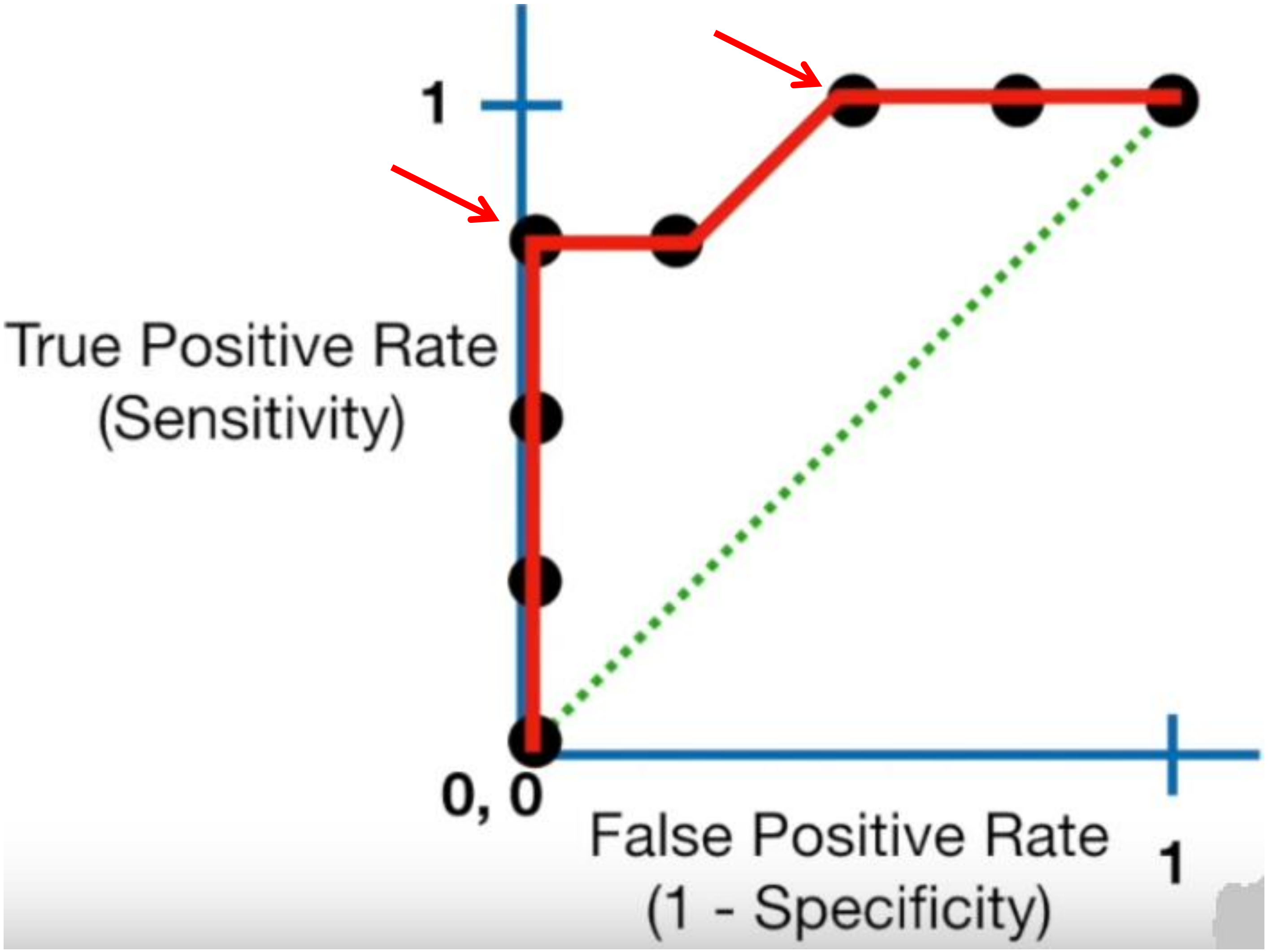


ROC



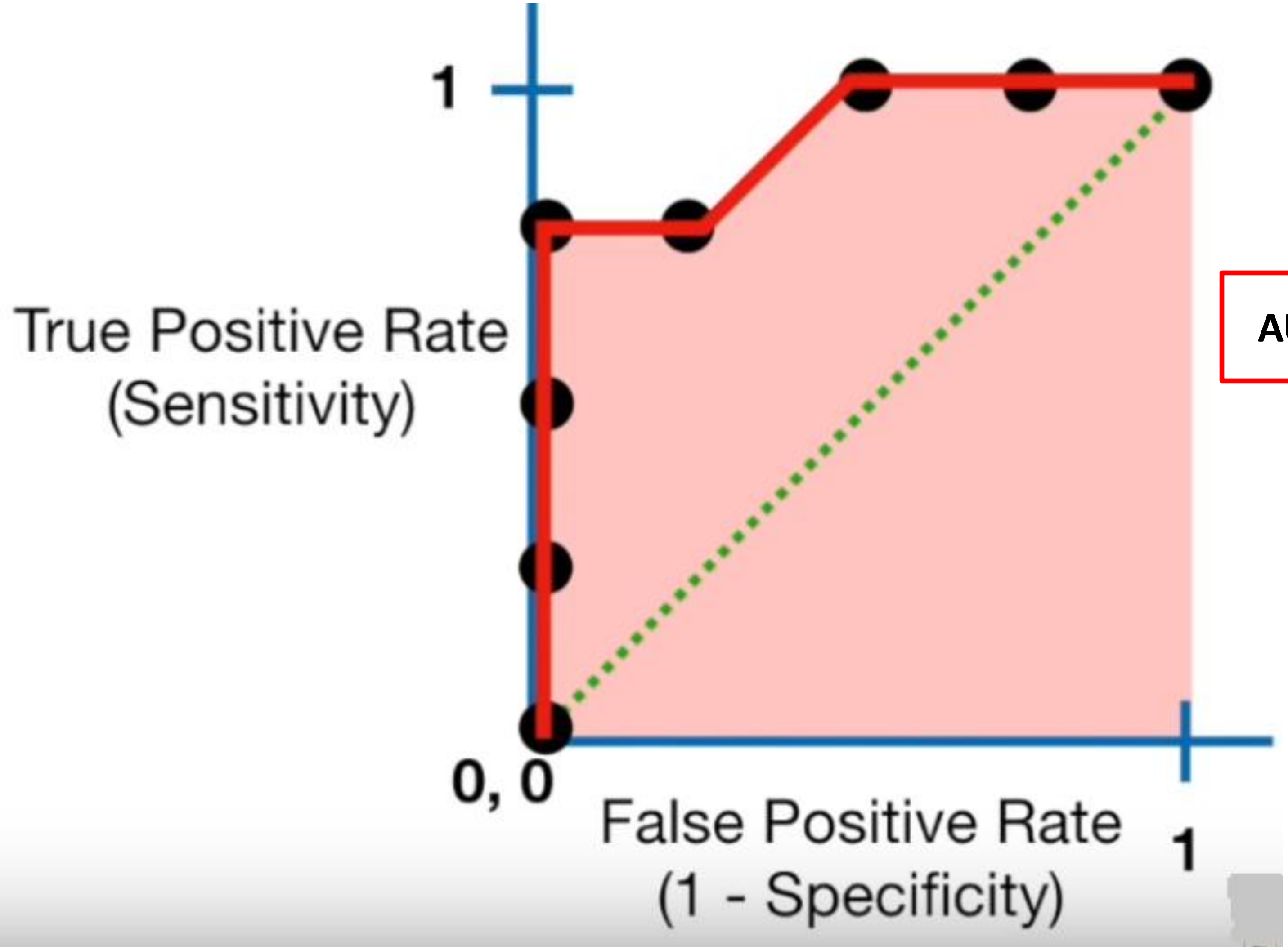
1
True Positive Rate
(Sensitivity)
Lastly, we choose a threshold that classifies all of the samples as **not obese**...
The point at **0, 0** represents a threshold that results in zero **True Positives** and zero **False Positives**.
0, 0
False Positive Rate
(1 - Specificity)
1

ROC



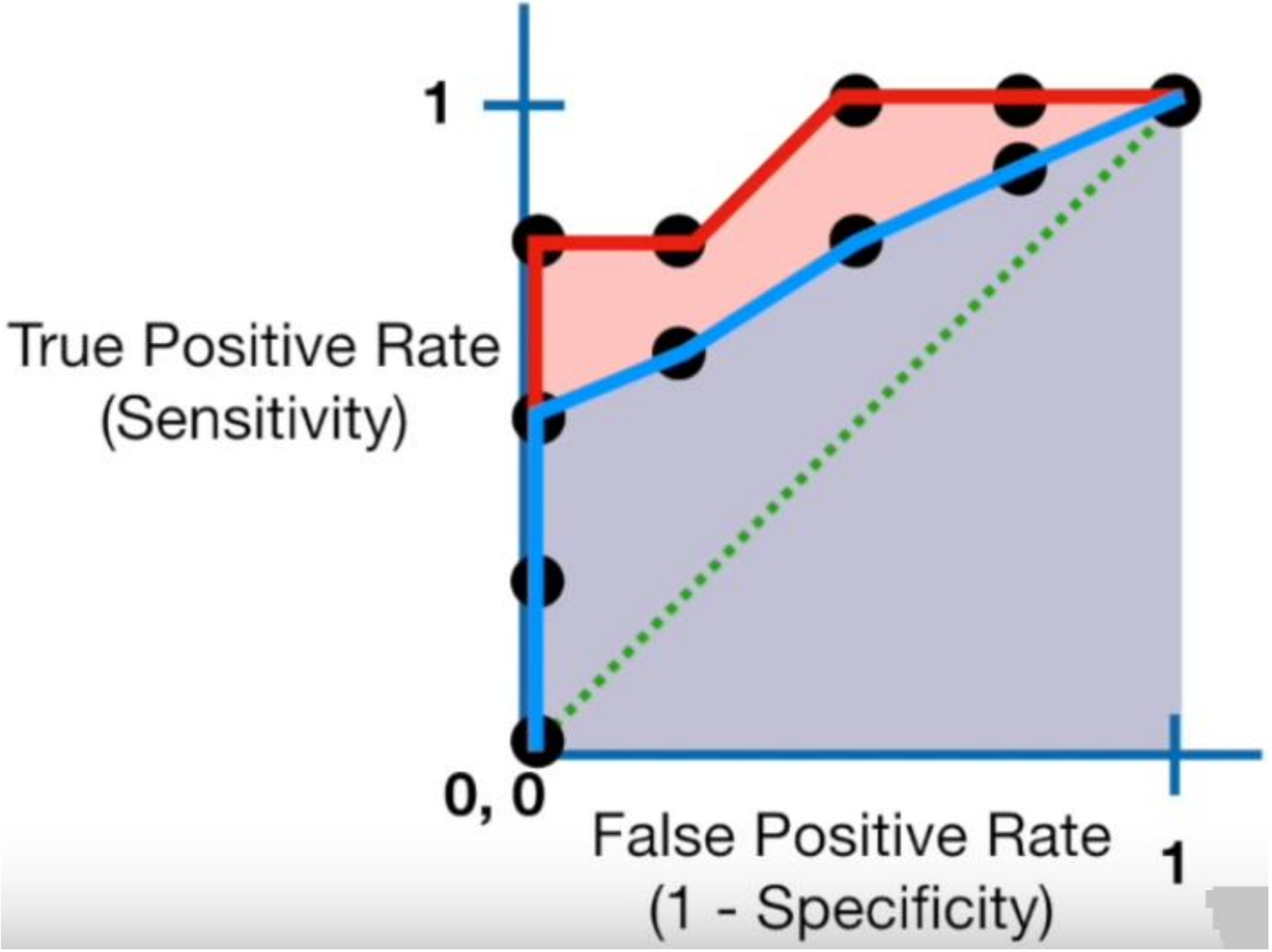
1
True Positive Rate
(Sensitivity)
0, 0
False Positive Rate
(1 - Specificity)
1

Area Under the Curve (AUC)



AUC = 0.9

AUC



1
True Positive Rate
(Sensitivity)
0, 0
False Positive Rate
(1 - Specificity)
1

Discussion

Precision is the proportion of positive results that were correctly classified

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{5}{5+2} = 71.4\%$$

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

If there were lots of samples that were **not obese** relative to the number of **obese** samples, then **Precision** might be more useful than the **False Positive Rate**.

PR Curve

This is because **Precision** does not include the number of **True Negatives** in its calculation, and is not effected by the imbalance.

In practice, this sort of imbalance occurs when studying a rare disease. In this case, the study will contain many more people without the disease than with the disease.

ROC curve makes it easy to identify the best threshold for making a decision and AUC can help you decide which categorization method is better

- *https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/*
- *https://heartbeat.fritz.ai/introduction-to-machine-learning-model-evaluation-fa859e1b2d7f*

Thank You