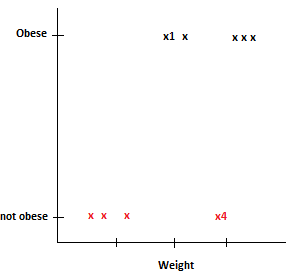
Receiver Operating Characteristic (ROC) and Area Under ROC (AUROC)

Using a logistic regression example.

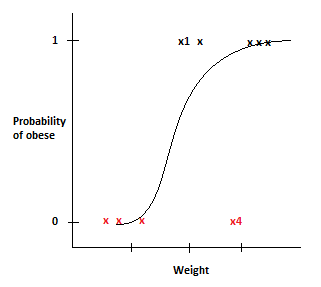
# ROC



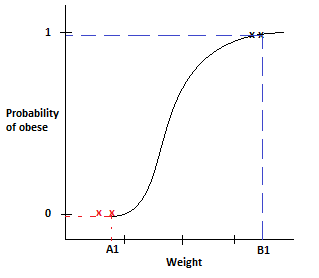
Refer to the fig above.

The black points represent **‘obese’** people and the red points represent people who are **‘not obese’.** Point x4 is not obese even though it weighs a lot (maybe it is muscle mass) and similarly, point x1 is obese even though it weighs comparatively less.

Let us fit a logistic regression. When we fit a logistic regression, the y-axis gets converted into a probability that the person is obese. 0 being less probability and 1 being highest.



Using this logistic regression, given the weight of a person, we can determine the probability of being obese.

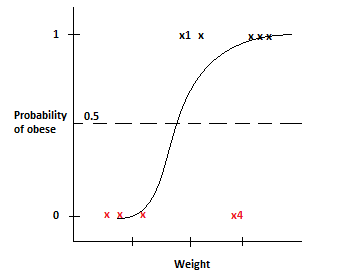


For e.g. a person weighing B1 has a high probability of being obese than a person who weighs A1 who is probably not obese.

Now, if we want to classify the person as obese or not obese, we have to find a way to convert the probabilities into classifications.

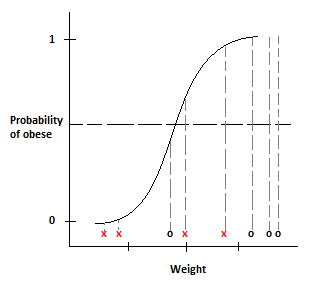
1. One way to classify is by setting a threshold. (say, 0.5 is the threshold, so any point above 0.5 is obese and any point below 0.5 is not obese.

Therefore, if we consider 0.5 as a threshold, x4 becomes not obese and x1 becomes obese.



Let consider known data points to test our logistic regression.

**x points** are not obese, and **o points** are obese. Now, according to the 0.5 threshold, two **x points** and one **o point** is wrongly classified.

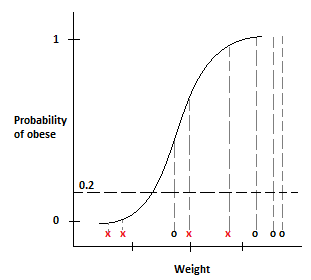


Confusion matrix would be as below:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Predicted | |
|  | Obese | Not obese |
| Actual | Obese | 3 | 1 |
| Not obese | 2 | 2 |

Now let us see what happens when we change the threshold to 0.2

**x points** indicate not obese, and **o points** indicate obese. This time, any point < 0.2 is not obese and > 0.2 is obese.



Our new confusion matrix would be

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Predicted | |
|  | Obese | Not obese |
| Actual | Obese | 4 | 0 |
| Not obese | 2 | 2 |

This is useful when it is absolutely important to classify every obese person as obese. Lowering the threshold results is correct classification for all 4 obese points.

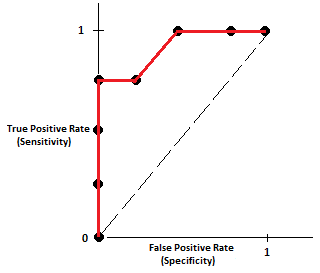
Similarly, if we increase the threshold to 0.9, we would get a new confusion matrix like

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Predicted | |
|  | Obese | Not obese |
| Actual | Obese | 3 | 1 |
| Not obese | 0 | 4 |

This is useful when it is absolutely important to classify every not obese person as not obese. Increasing the threshold results is correct classification for all 4 not obese points.

The threshold could be set to anywhere between 0 to 1. How do we determine which threshold is the best?

If we calculate a confusion matrix for all possible threshold values we would end up with a lot of tables. So, instead of being confused with confusion matrices, Receiver Operating Characteristic (ROC) graphs provide a simple way to summarize all the information.



|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Predicted | |
|  | Obese | Not obese |
| Actual | Obese | True Positive | False Negative |
| Not obese | False Positive | True Negatives |

True positive Rate = Sensitivity =

True positive rate tells us what proportion of obese samples

were correctly classified as obese..

False positive Rate = Specificity =

False positive rate tells us what proportion of non-obese samples were incorrectly classified as obese..

Now let’s see how a ROC is plotted.

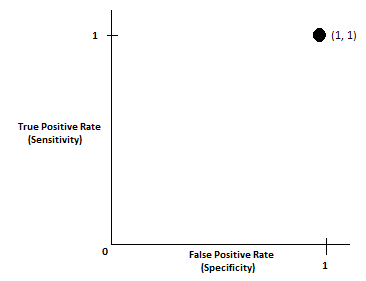
For a threshold of 0, which will classify all the points as obese, we will get this confusion matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Predicted | |
|  | Obese | Not obese |
| Actual | Obese | 4 | 0 |
| Not obese | 4 | 0 |

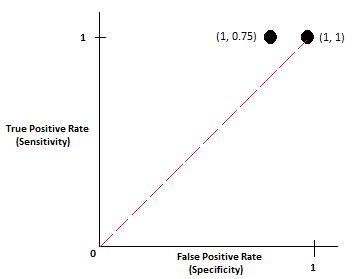
Now calculate the TPR and FPR

TPR = = 1 and FPR =

Now plot a point on the TPR vs FPR axis

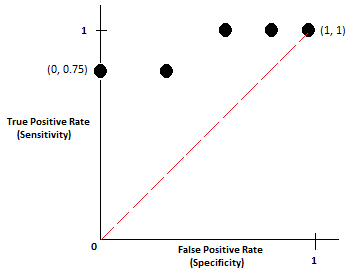


Now as we increase the threshold, and calculate the confusion matrix and our new TPR and FPR values we would get our second point as below



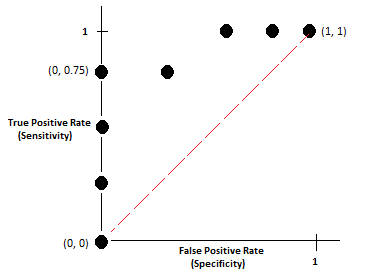
Since the new point (0.75, 1) is to the left of the dotted line, the proportion of true positives is greater than the proportion of false positives. ( we can always confirm that with the confusion matrix)

Continuing with these steps



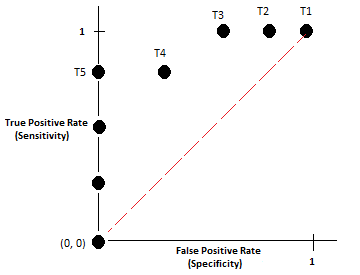
The threshold represented by the new point (0, 0.75) correctly classified 75% of the obese samples and 100% of the samples that were not obese. i.e. it resulted in No False Positives

Continuing with increasing the threshold



The point (0, 0) represents a threshold value where there are 0 True Positives and 0 False Positives.

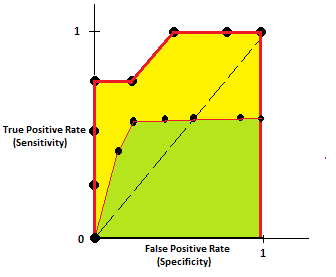
Therefore, without calculating and sorting through confusion matrices, we can easily say that each threshold point on the left is better than it’s consecutive threshold point on its right. (i.e. T2 is better than T1)



Depending up on how many False Positives we are willing to accept, the optimal threshold point is either T3 or T5.

# Area Under ROC (AUROC)

The AUC makes it easy to compare one ROC curve to another.



The AUC for the yellow ROC curve is greater than the AUC for the underlying green ROC curve, suggesting that the yellow curve is better.

i.e. if the yellow ROC curve represented a logistic regression and the blue ROC curve represented a Random Forest, we would choose Logistic Regression..

# Conclusion

ROC curves make it easy to identify the best threshold for making a decision..

AUROC helps us decide which categorization method is better.