

PERFORMANCE MEASURES

FOR

LOGISTIC REGRESSION

MODELS

The Confusion Matrix

		Value Predicted by Classifier	
		1	0
Actual Value	1	TP	FN
	0	FP	TN

TN = True Negative

Maximize

FP = False Positive

Minimize

FN = False Negative

TP = True Positive

		Predicted		Total
		1	0	
Actual Value	1	485 tp	15 fn	500
	0	50 fp	450 tn	500
Total		535	465	1000

Goals: Increase tp



Decrease fn

Increase tn



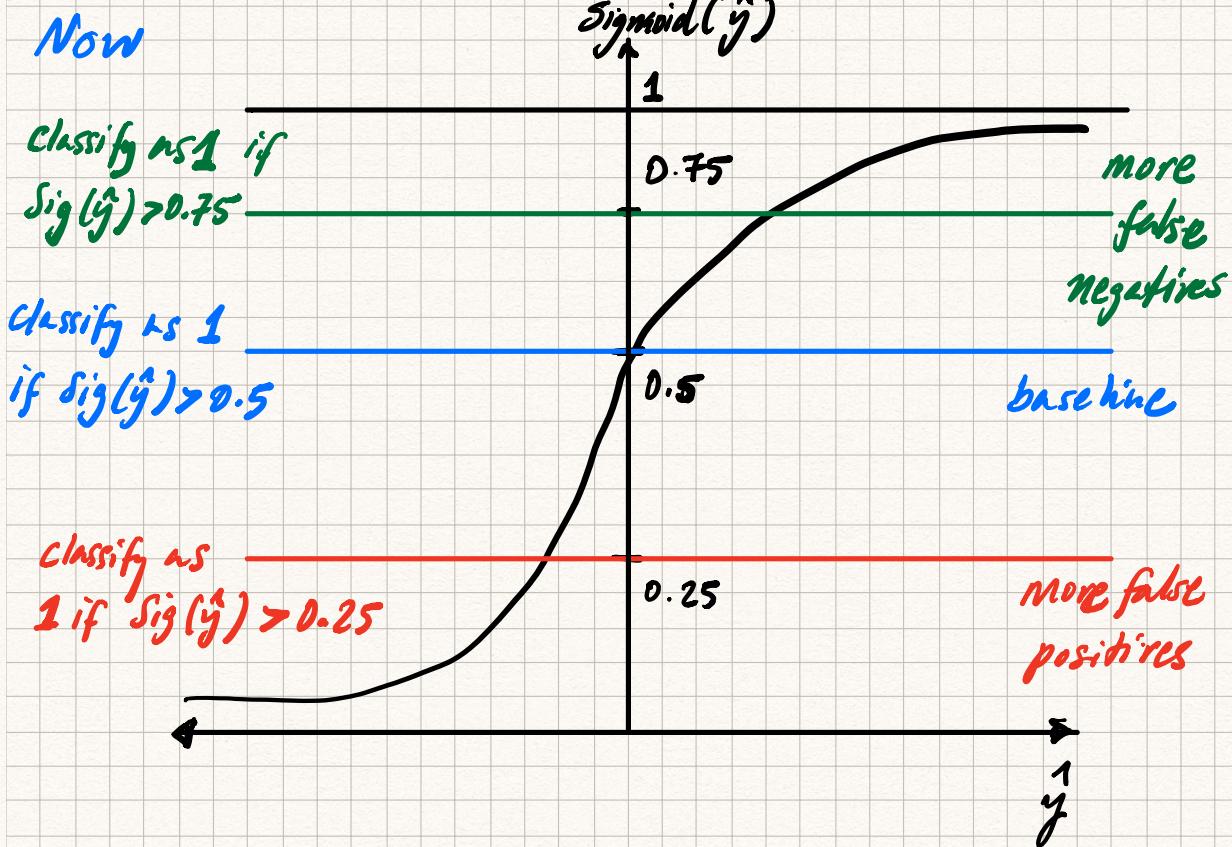
Decrease fp

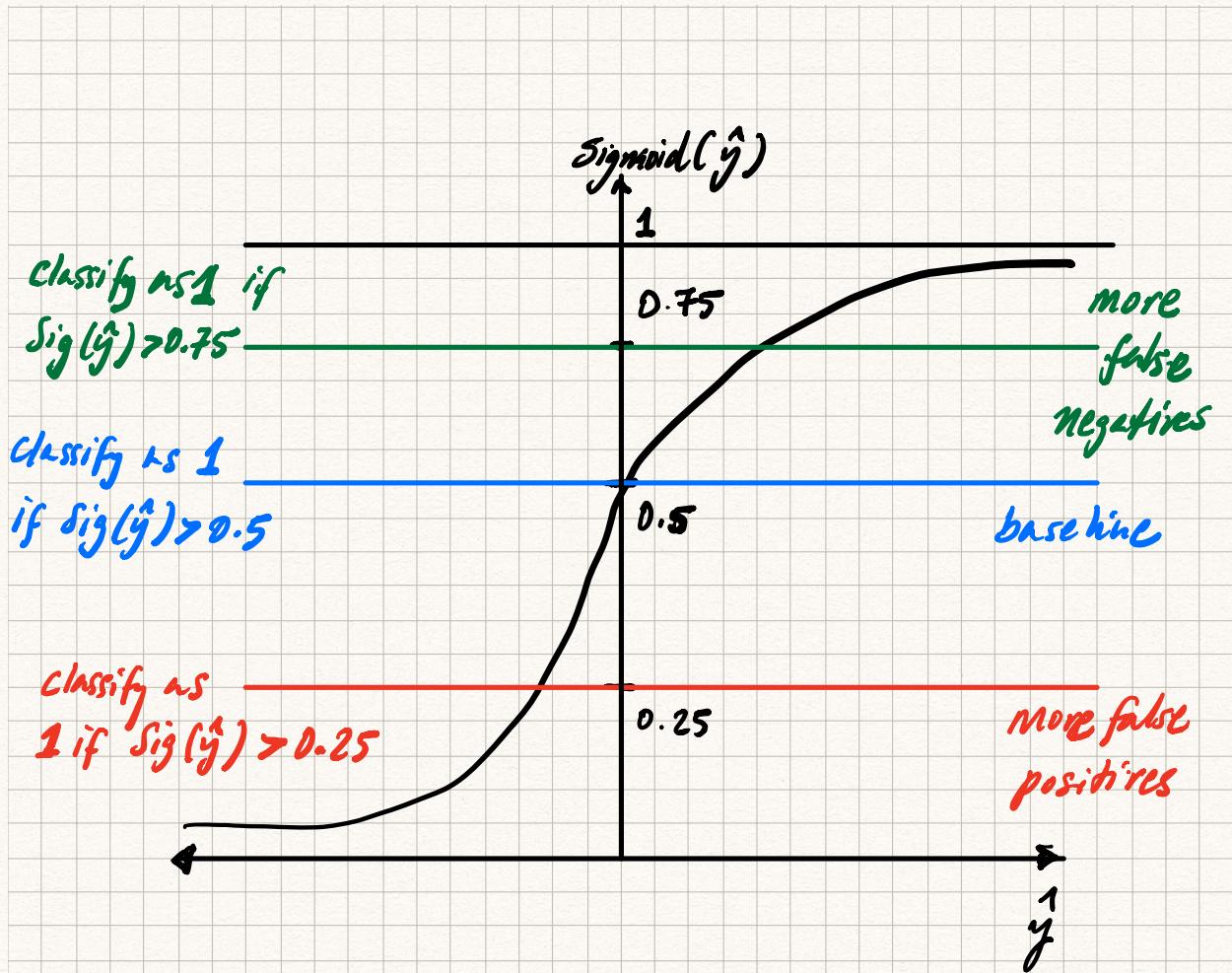
Make the classifier
better at detecting
1s

Make the
& classifier
better at
detecting 0s

The ways to do this :

- Later {
- ▢ get more data
 - ▢ create "better" features
 - ▢ tune the model's hyperparameters
 - ▢ change the sigmoid threshold



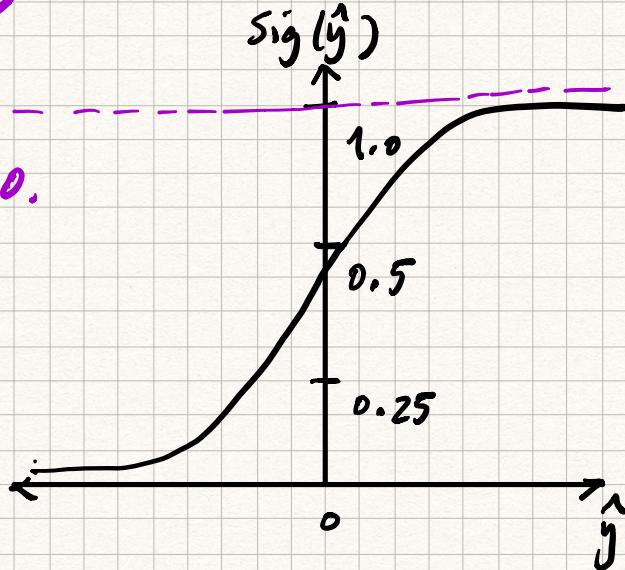


		Predicted		
		1	0	
Actual	1	$tp \downarrow$	$fn \uparrow$	$tp + fn$
	0	$fp \uparrow$	$tn \downarrow$	$fp + tn$
		$tp + fp$	$fn + tn$	

Think of moving the threshold down
starting at $\text{sig}(\hat{y}) = 1$.

One extreme

everything
is classified 0.



Precision

tolerance for
false negatives
at the cost
of fewer
true positives

Predicted

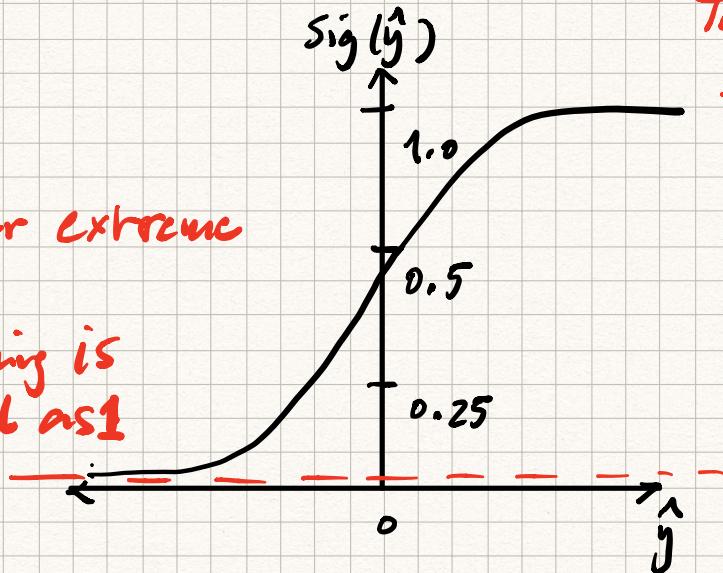
		1	0	
		t_p	f_n	$t_p + f_n$
Actual	1	t_p	f_n	$t_p + f_n$
	0	f_p	t_n	$f_p + t_n$
		$t_p + f_p$	$f_n + t_n$	

Recall

Tolerance for
false positives
at the cost
of fewer
true negatives

The other extreme

everything is
classified as 1



Predicted

		1	0	
		tp	fn	tp + fn
Actual	1	tp	fn	tp + fn
	0	fp	tn	fp + tn

$tp + fp$

$fn + tn$

Measuring Classifier Performance

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

"In simple terms, high precision means that an algorithm returns substantially more relevant results than irrelevant ones, while high recall means an algorithm returns most of the relevant results."

- Wikipedia (https://en.wikipedia.org/wiki/Precision_and_recall)

Measuring classifier performance (Contd.)

F1 is a measure combining precision & recall

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

And finally,

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

Accuracy can be tricky, especially
when the dataset's target feature
is imbalanced.

When the incidence rate is
very low or very high, accuracy is not a
good measure of classifier performance.

Ready for a surefire scheme?

...

Idea Make a TSA detector

that detects if a person is
a terrorist. Very high accuracy!
 $> 99.99\%$

Going to sell it to LAX
and make a lot of money!

I know (lets' assume)

- 1 in 1 million passengers who
fly out of LAX are terrorists.
(very low incidence rate)
- 10 million passengers fly out of
LAX every year.

Here's how I'm going to build my detector.

		Predicted		Total
		0	1	
Actual	0	6	4	10
	1	9,999,980	10	9,999,990
Total	14	9,999,986	10,000,000	

Predict
very few terrorists - } This is the
here 14 in all } brilliant
Scheme !

$$\begin{aligned}
 \text{Accuracy} &= \frac{tp + tn}{tp + tn + fp + fn} \\
 &= \frac{4 + 9,999,980}{10,000,000} \\
 &= 0.9999984 = 99.999\%
 \end{aligned}$$

But notice that precision

and recall will be low

and point out the silliness

of this scheme.

$$\text{Precision} = \frac{tp}{tp+fp} = \frac{4}{4+10} = 0.286 \\ = 28.6\%$$

$$\text{Recall} = \frac{tp}{tp+fn} = \frac{4}{4+6} = \frac{4}{10} \\ = 40\%$$

SUMMARY

- The performance of a model depends on a number of factors
 - the amount of data
 - the features of the dataset
 - the model's hyperparameters
 - the decision threshold
(where to set the sigmoid value to classify into one group or the other)
- Try a number models (with default hyperparameters) on a given dataset.

- Measure the performance of logistic regression models using Precision, Recall, and/or F Scores. (The higher the better.)
- Get insight into where a model is doing badly by using the confusion matrix.
- Measure and compare model performance using the training dataset.

5 Make predictions on the
test dataset.