

# CSE 416 Section 3!

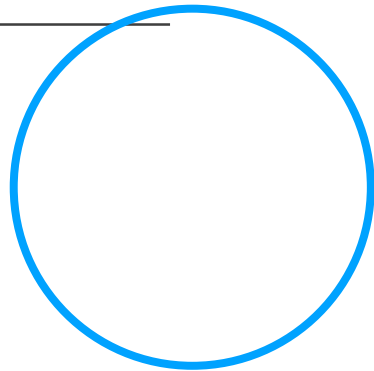
Zoom University – Global Pandemic Summer

~~(Wanted to go outside but afraid of getting infected so stay home and teach machine learning uwu)~~

---

JULY 9, 2020

● HONGJUN JACK WU 😊



This material is made with color blind folks in mind.

If there is anything that is not clear or you cannot distinguish **PLEASE** let us know so we can fix it ASAP.

A decorative graphic consisting of several overlapping circles and lines. A large pink circle is the central element, with a dashed orange circle to its top-left, a dashed green circle to its top-right, and a solid blue circle to its bottom-right. A small yellow dot is on the left side of the pink circle, and a small cyan dot is on the right side. The text is centered within the pink circle.

# Goal for today!

MAIN GOAL:  
LOGISTIC REGRESSION

# Materials of the Day

There are notebooks about sigmoid, logistic regression, and other materials by Ben Evans!

(Thank you Ben 😊)

We will post them on the course website after we are done with all sections, as always.

Very helpful if you want to dig deep into the implementation of algorithms (and how to actually use them in Python) and might benefit you from doing your next assignment!

Me after running an old notebook in the exact same order it was supposed to run



# Index

Part I (5-7) : Sigmoid

Part II (8-11) : Logistic Regression

Part III (12-13) : Breast Cancer Example

Part IV (14-19) : Confusion Matrix

Part V (20-27) : ROC Curve

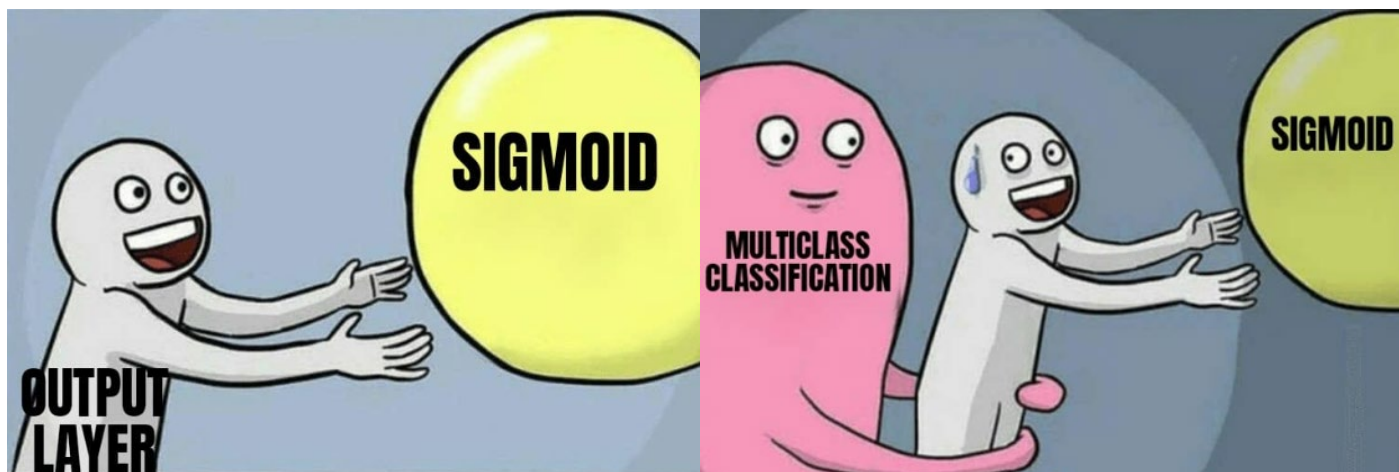
Part VI (28-31) : MLE of  $\beta$  (Optional)

Part VII (32-38) : Section Handout and Solutions

Part VIII (39-42) : Other Stuff (Credit, License, etc)

Tom was the first guy losing his job  
because of Artificial intelligence





# Sigmoid.

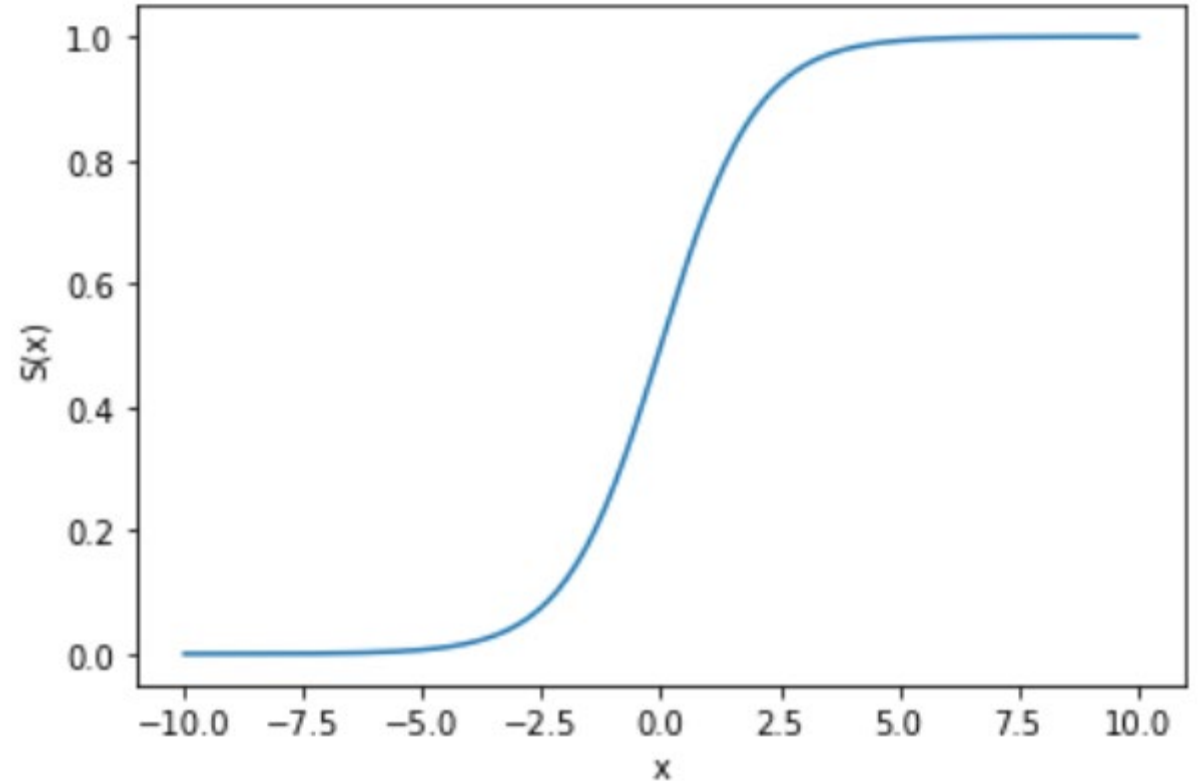
PART I

# SIGMOID

Literally means curved like S  
(Nothing fancy about this) but we  
gave it a fancy name anyway  
because it is actually quite useful  
in machine learning.

In mathematics equation:

$$S(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}$$

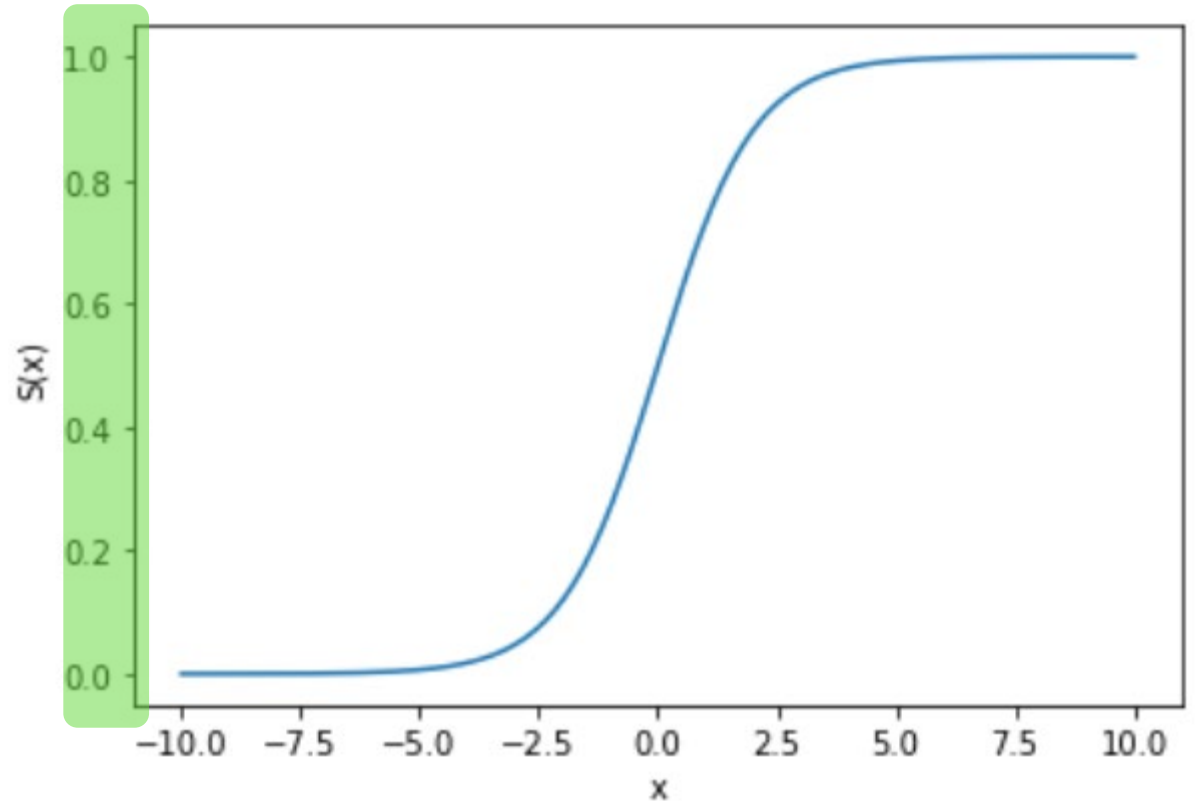


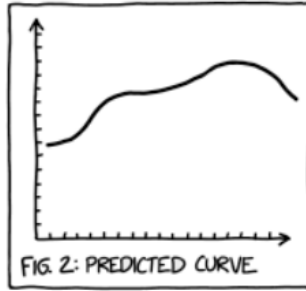
# WHY WE LIKE SIGMOID

Well, because it is nice...

It has two things that are useful:

1. It is monotone increasing in  $x$ .
2. The function value is always between  $[0,1]$ . (perfect for modeling probabilities!)





SCIENCE TIP: IF YOUR MODEL IS BAD ENOUGH, THE CONFIDENCE INTERVALS WILL FALL OUTSIDE THE PRINTABLE AREA.

# Logistic Regression

PART II



# LOGISTIC REGRESSION

Logistic regression is a statistical model that uses a logistic function (aka sigmoid function) to model a binary dependent variable (in human words, 0 or +1).

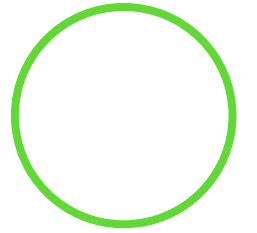
In human language, **logistic regression uses sigmoid function to estimate the probability of  $y_i$  being +1.**

More math (like a lot more) in the notebook for math fans.

Note:

Data:  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  where  $x_i \in \mathbb{R}^k$  and  $y_i \in \{0, +1\}$ .

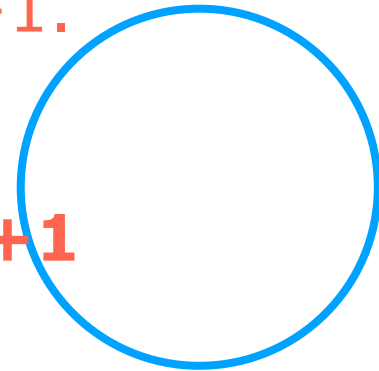
# LOGISTIC REGRESSION AS CLASSIFIER



Enough theory, so how do we use logistic regression???

Well, there are three steps to this problem:

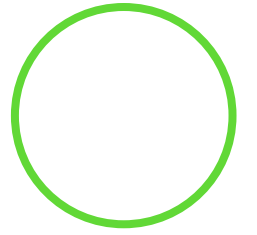
1. Use data to "get" an estimate of parameter  $\beta$ .
2. Given an new input  $x$ , estimate the probability of  $y$  being  $+1$ .
3. Pick a threshold, say  $0.5$ . then if:
  1. **The probability of  $y$  being  $+1 > \text{Threshold} \rightarrow \hat{y} = +1$**
  2. **Otherwise  $\rightarrow \hat{y} = 0$**



Note: **The threshold does not need to be fixed as 0.5.**



# LOGISTIC VS. LINEAR

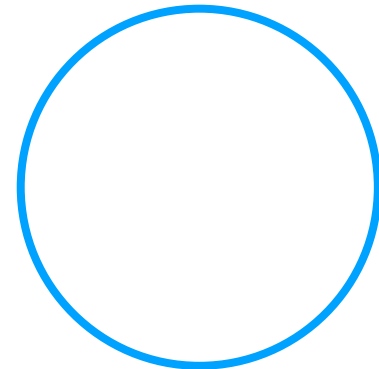


Difference between Linear & Logistic regression in one sentence:

**Logistic regression is used when the dependent variable is binary in nature.**

**Linear regression is used when the dependent variable is continuous and nature of the regression line is linear.**

Yea. That's it. Nothing too complicated.



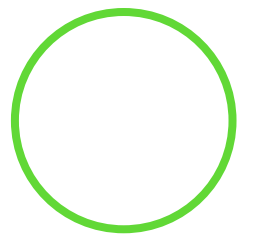


# Breast Cancer Example

PART III



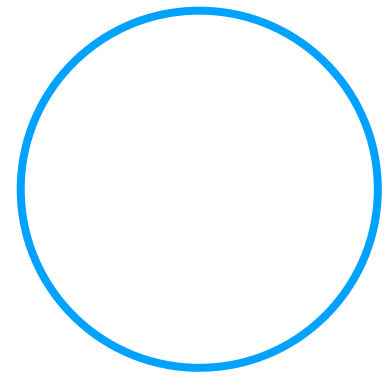
# QUICK NOTE



Goal: Use features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass to predict type of breast cancer.

All that we are doing in a summary:

1. There are two types of output: malignant & benign. (Therefore binary)
2. We split the data, make a logistic regression model and train using training data.
3. All we are trying to get is probability of the datapoint predicted to be benign.
4. We then try a couple of threshold to see which threshold works the best.  
(The prob threshold for the model to call this patient is benign)





\*Exactly what I mean when I see Confusion Matrix

# Confusion Matrix

PART IV

# METRICS FOR EVALUATING A CLASSIFIER

**Goal: We need a way to see how good our classifier is.**

**There are four possible cases when a classifier predict on data:**

**FP:** False Positive (aka Type 1 Error) ✗

- when our model outputs positive when the correct label was negative.

**TN:** True Negative ✓

- when our model outputs negative when the correct label was negative.

**FN:** False Negative (aka Type 2 Error) ✗

- when our model outputs negative when the correct label was positive.

**TP:** True Positive ✓

- when our model outputs positive when the correct label was positive.

Read more: [\[1\]](#)

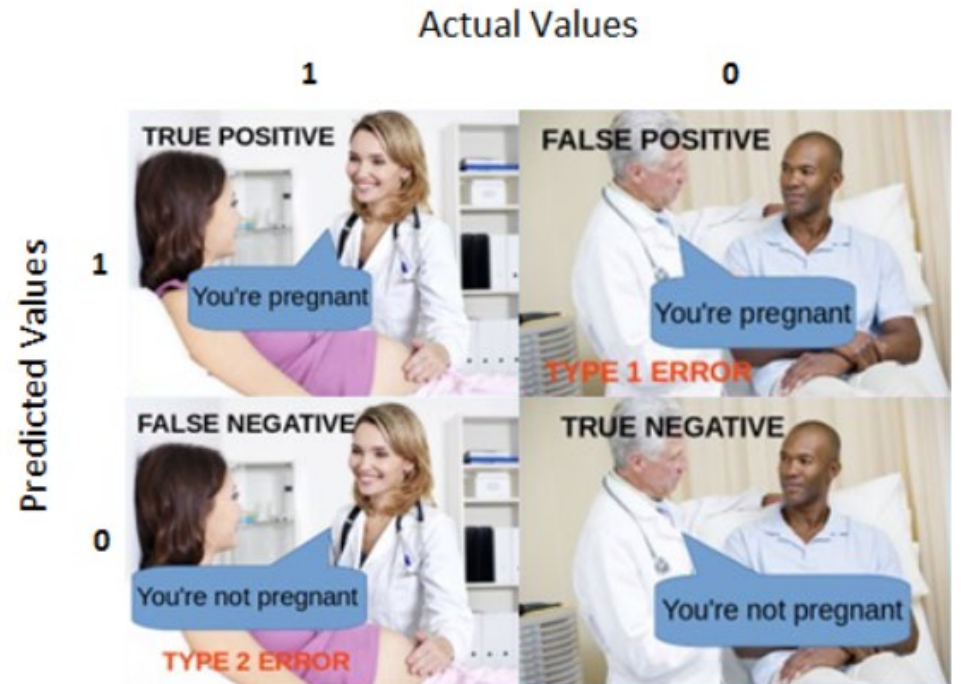
# PLOT IT OUT

Then we plot this out in a 2x2 matrix and called this confusion matrix.

We want as many TP & TN and as little FP & FN as possible.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

If this helps you learn then here u go ↓







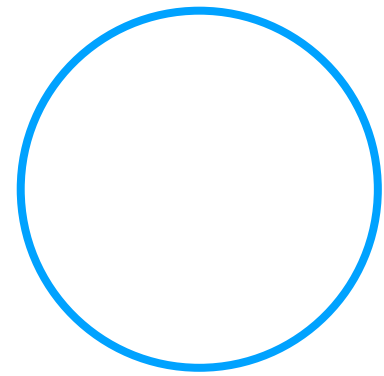
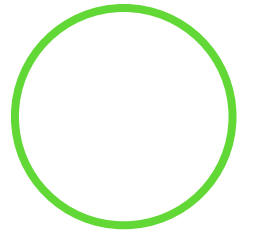
# RECALL

**Recall:**

**Out of all the positive classes, how much we predicted correctly?**

It should be as high as possible.

$$\text{recall} = \frac{TP}{TP + FN}$$



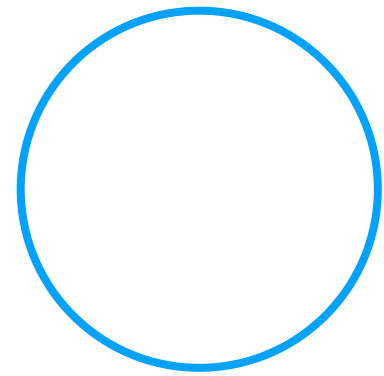
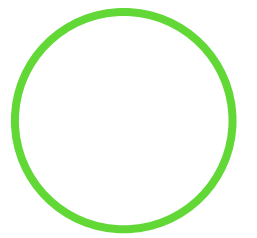


# PRECISION

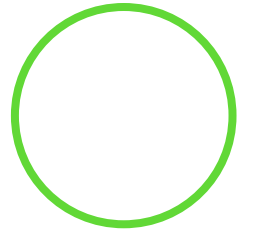
## Precision:

Out of all the positive classes we have predicted correctly, how many are actually positive?

$$\text{precision} = \frac{TP}{TP + FP}$$



# F1 SCORE (F-MEASURE)



## **F1 Score (F-Measure):**

It is difficult to compare two models with low precision and high recall or vice versa.

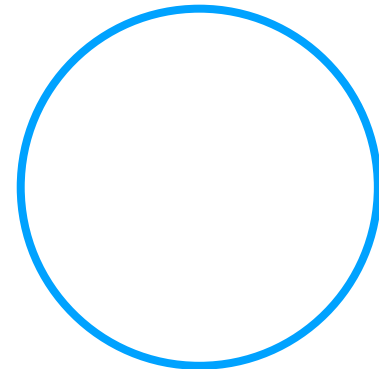
So to make them comparable, we use F1-Score.

**F1-score helps to measure Recall and Precision at the same time.**



It uses Harmonic Mean in place of Arithmetic Mean by punishing the extreme values more.

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$





Linear Regression

# ROC Curve

PART V



# RECEIVER OPERATING CHARACTERISTICS CURVE



This name is wayyyyyyyyyyyyyyy too long. We usually just call it ROC curve.

Don't worry about remembering the name. ~~Nobody memorizes it anyway :P~~

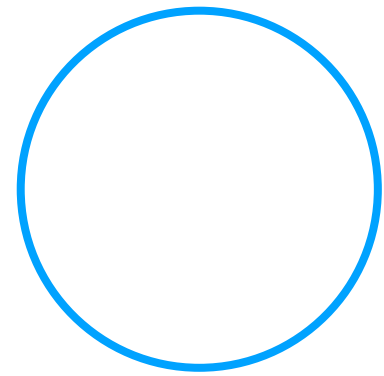
What it is:

ROC curve is a graphical plot that **illustrates the diagnostic ability of a binary classifier** as its discrimination threshold is varied.

How to make it:

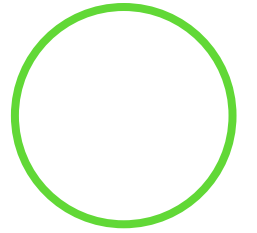
ROC curve is created by **plotting the true positive rate (TPR) against the false positive rate (FPR)** at various threshold settings.

Read more: [\[2\]](#)





# TPR AND FPR



True Positive Rate (TPR) / Recall / Sensitivity:

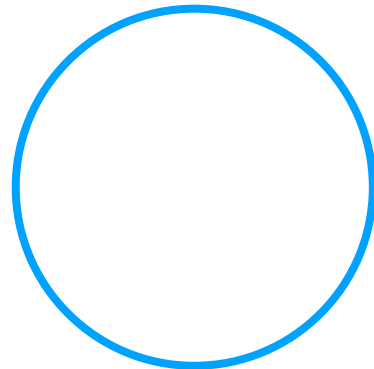
Of all the things that are truly positive, how many of them did I correctly predict as positive?

$$\text{TPR} = \frac{TP}{TP + FN}$$

False Positive Rate(FPR):

Of all the things that are truly negative, how many of them did I falsely predict as positive?

$$\text{FPR} = \frac{FP}{FP + TN}$$



# ROC CURVE

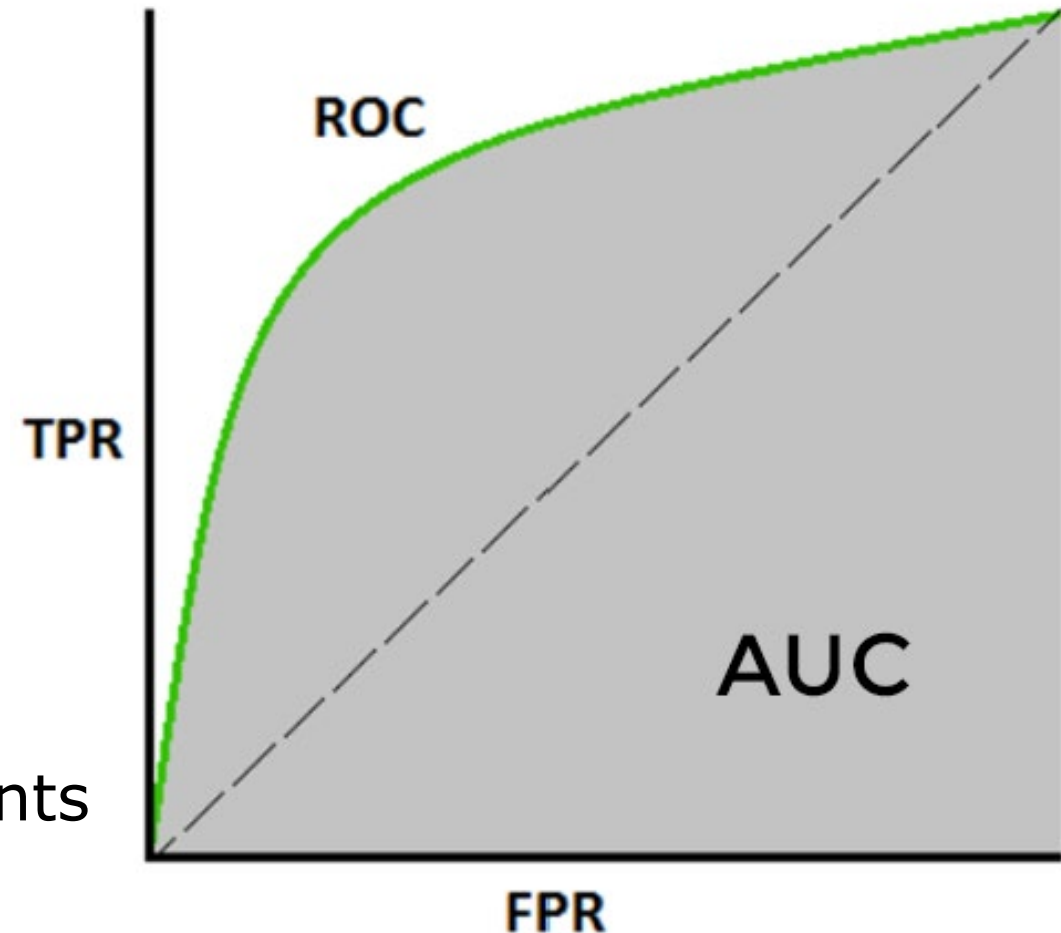
Then we just plot a curve using TPR and FPR.

**AUC: Area Under The Curve (☹).**

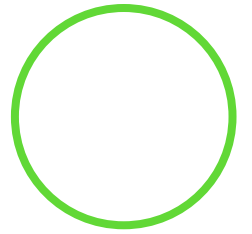
It tells how much model is capable of distinguishing between classes.

Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s.

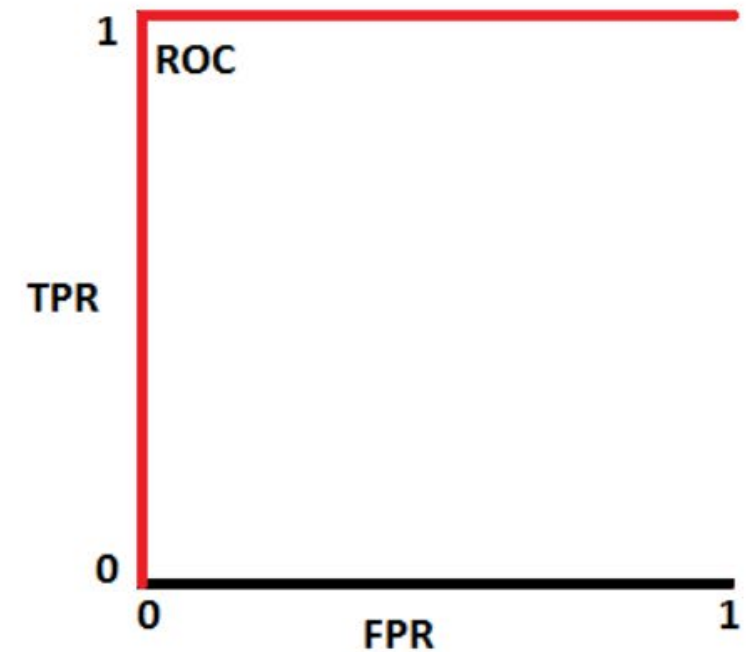
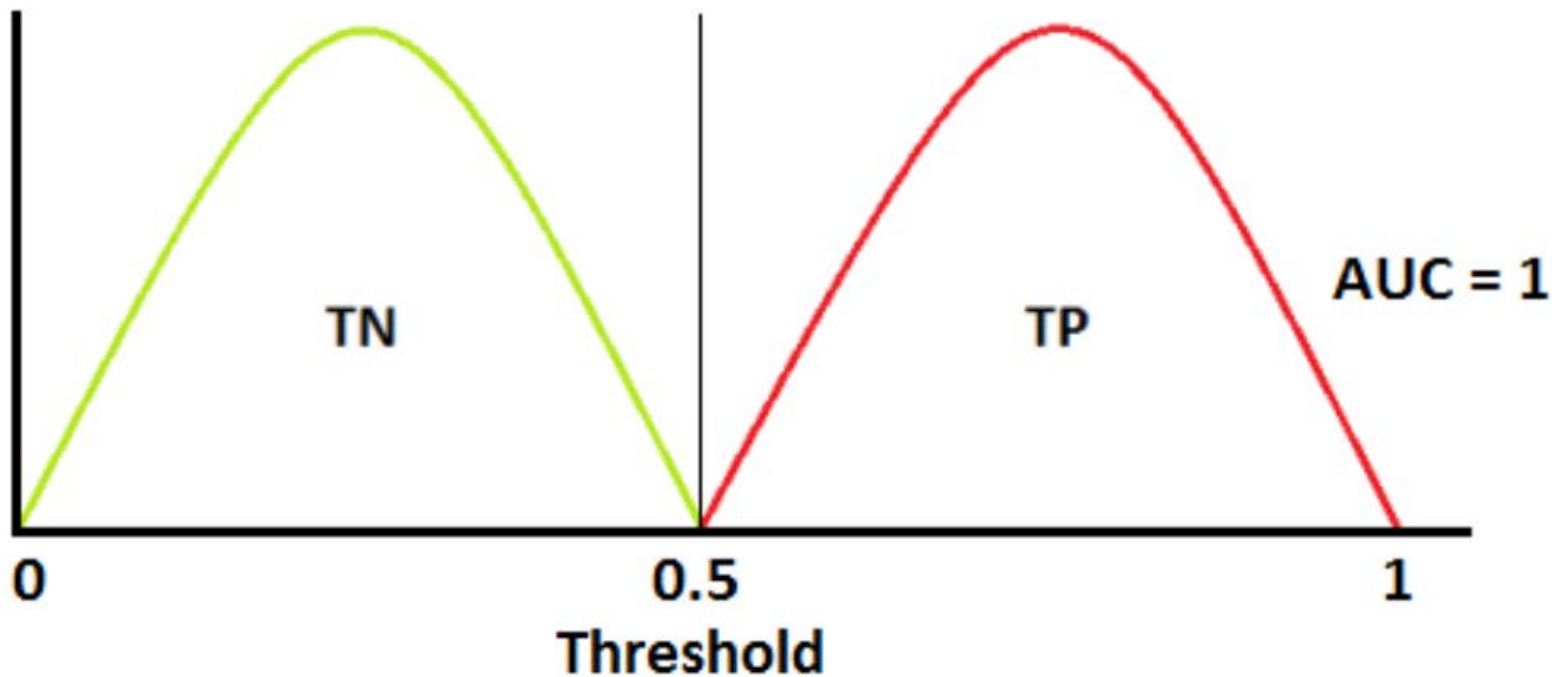
By analogy, Higher the AUC, better the model is at distinguishing between patients with disease and no disease.



# ROC CURVE PERFORMANCE

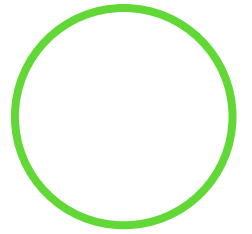


An **excellent model has AUC near to the 1** which means it has good measure of separability.



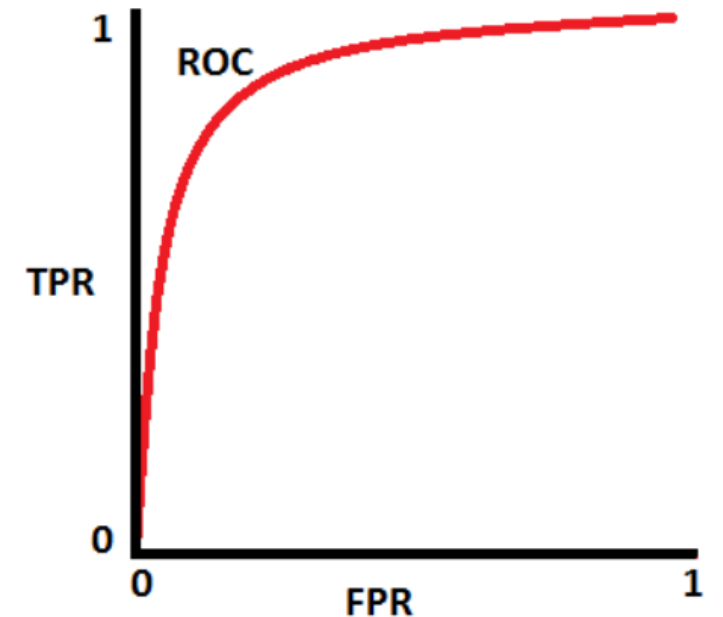
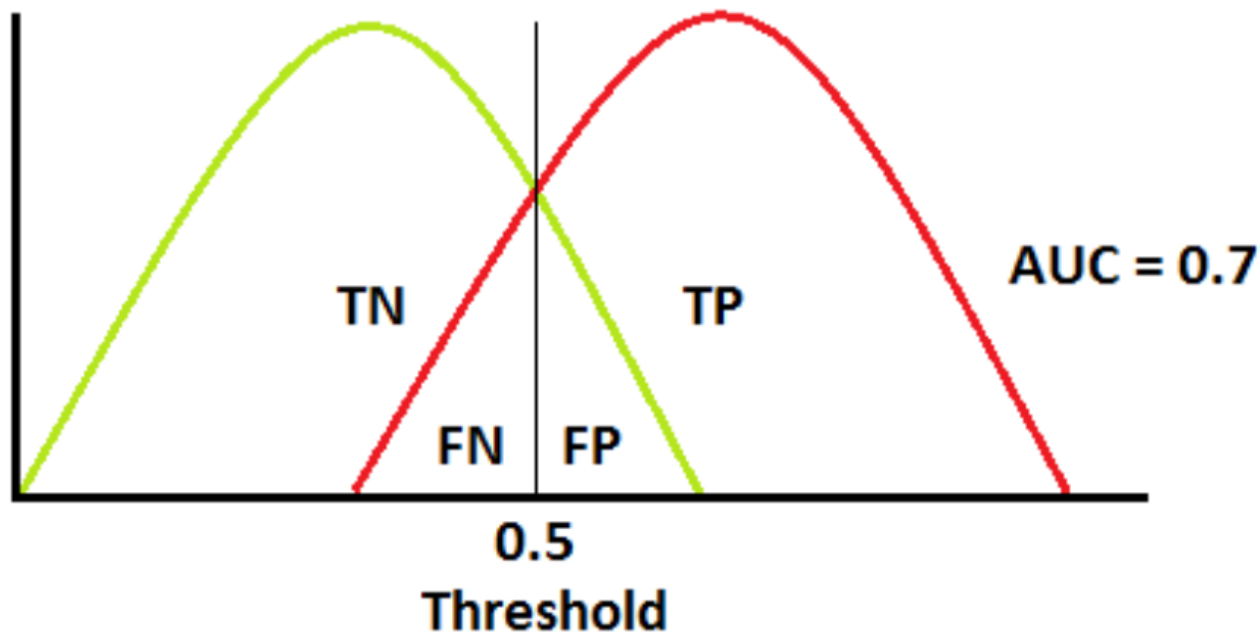


# ROC CURVE PERFORMANCE

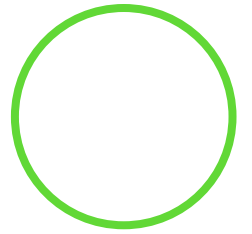


When two curves don't overlap at all means model has an ideal measure of separability. (Usually this is the case)

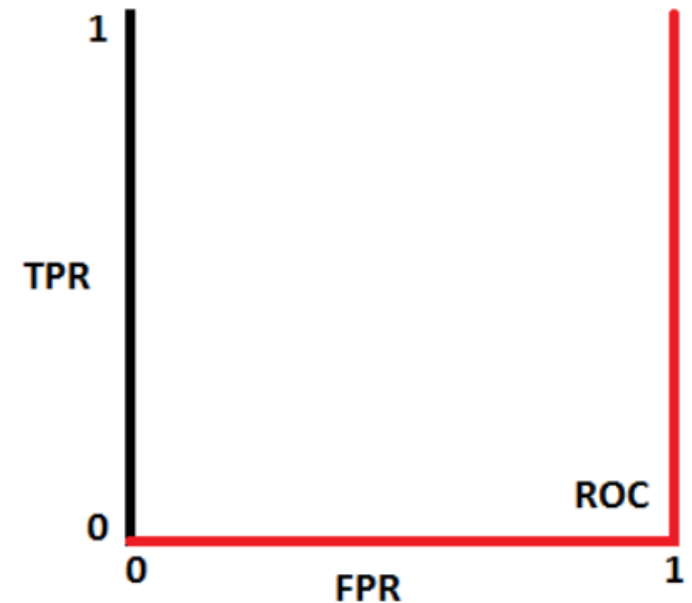
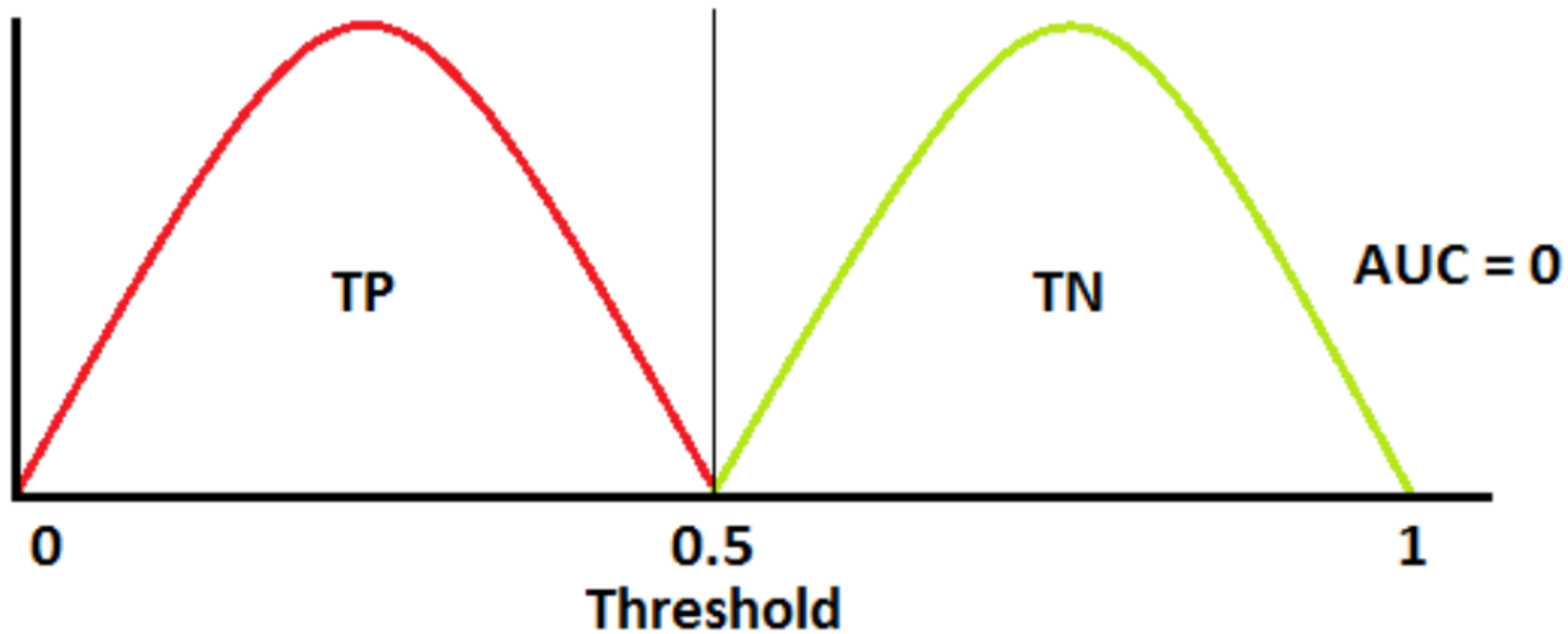
It is perfectly able to distinguish between positive class and negative class.



# ROC CURVE PERFORMANCE

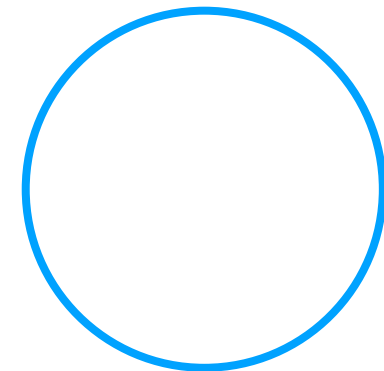
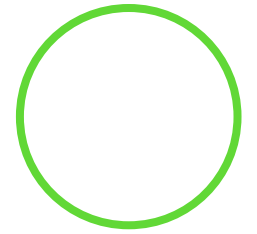
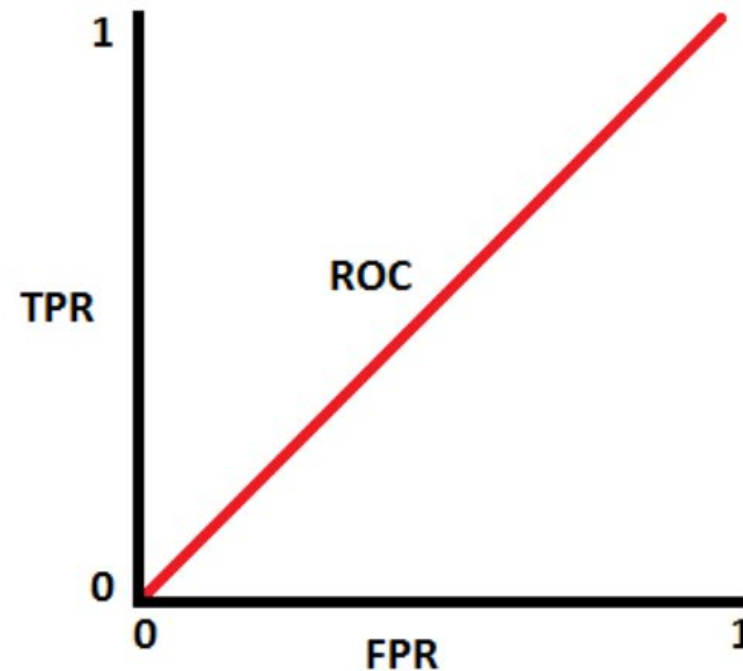
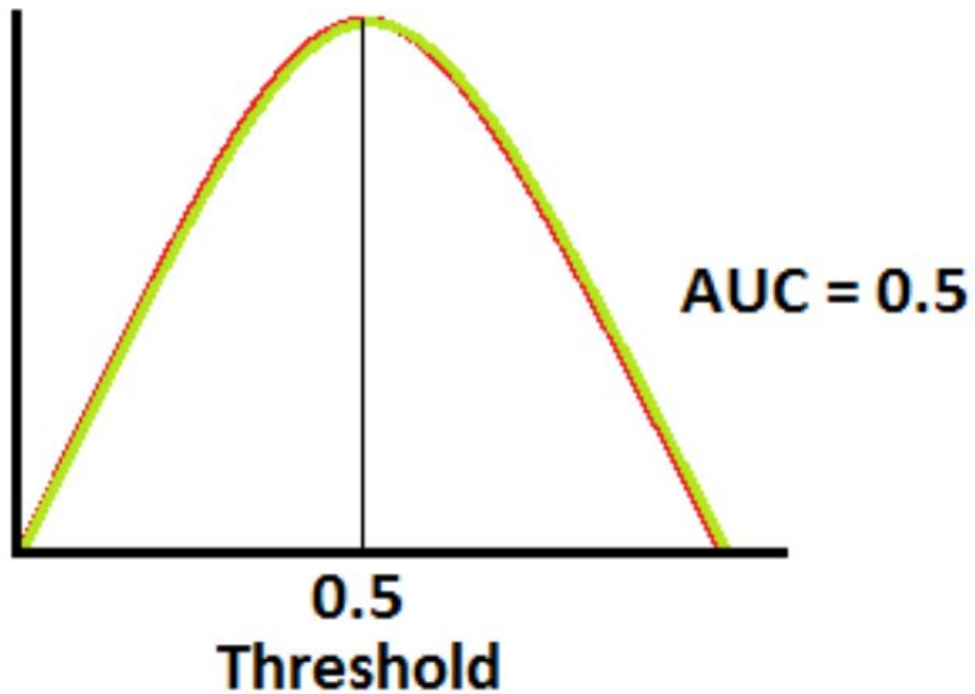


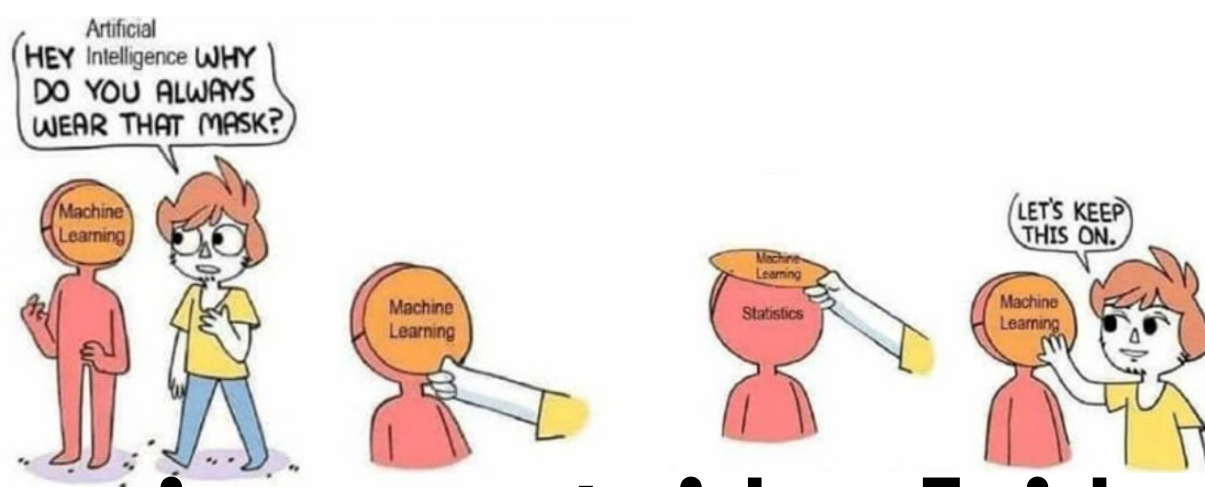
A **poor model has AUC near to the 0** which means it has worst measure of separability. In fact it means it is reciprocating the result. It is predicting 0s as 1s and 1s as 0s.



# ROC CURVE PERFORMANCE

And when AUC is 0.5, it means **model has no class separation capacity** whatsoever.





# Maximum Likelihood Estimation (MLE) of $\beta$

OPTIONAL PART VI

I DOUBT WE WILL HAVE TIME FOR THIS...

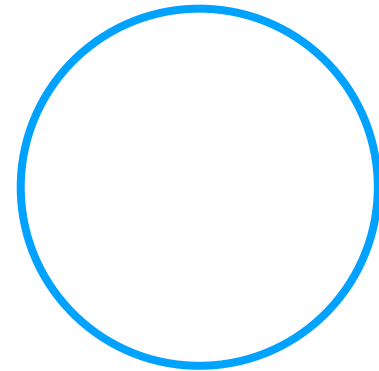
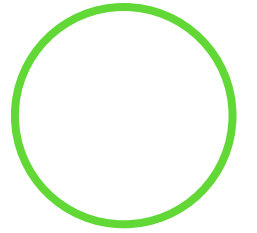
# WHAT THE HECK IS MLE ANYWAY?

Essentially what MLE wants to solve is:

Which are the best parameters/coefficients for my model?

In linear regression, we minimized the RSS.

With logistic regression (a probabilistic model), we could use maximum likelihood estimation to estimate the parameter  $\beta$ .





# LONG DEFINITION OF MLE

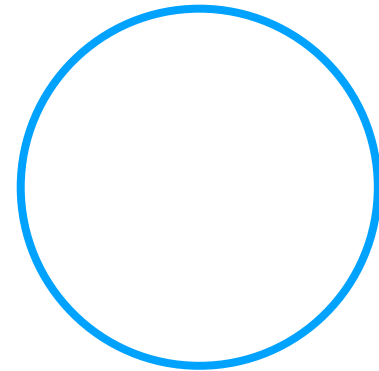
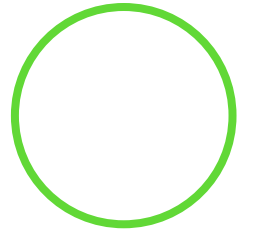
Maximum Likelihood Estimation (MLE) :

A method of estimating the parameters of a probability distribution by maximizing a likelihood function.

So that under the assumed statistical model the observed data is most probable.

In human language, a method used in probability models to find the best model.

Read more: [\[3\]](#)






# MLE IN LOGISTIC REGRESSION



For logistic regression, given a data point  $(x_i, y_i)$ , assuming that  $y_i = 1$ , then likelihood for this single data point would be:

$$l(\beta) = P(y = 1|x; \beta) = \frac{1}{1 + \exp(-x^T \beta)}$$


The idea for maximum likelihood estimation is that we would like our parameter  $\beta$  to maximize the above probability.

A larger probability indicates a better fit of our model.





Brute Force



*Non-Convex  
optimization using  
Gradient Descent*

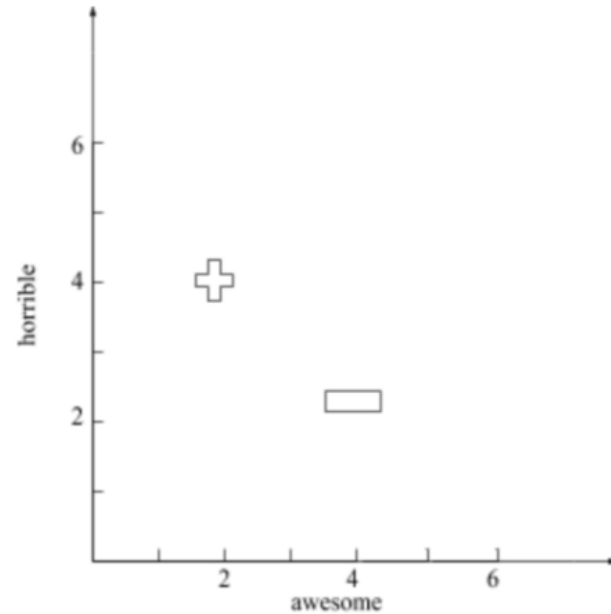
# Handout & Solutions

PART VII



# QUESTION 1

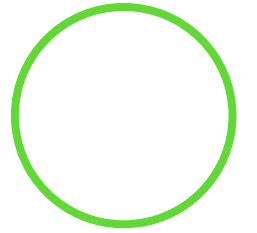
Suppose your work for Amazon and are tasked with classifying product reviews as positive or negative. You aren't very familiar with "big data", so your boss assigns you a small dataset with only two points. Your job is to create a classifier that correctly classifies the two points.



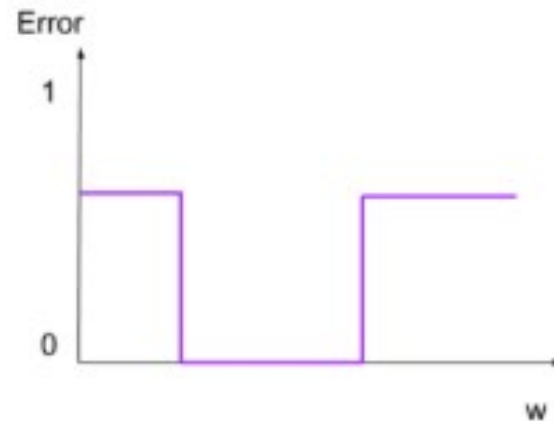
You develop the following procedure: you fix the weight of each count of "horrible" to be 1 and decide to vary the weight of "awesome". Create a graph below that shows how the classification error changes as you move through different weights for "awesome". The graph does not need to have correct values for the coefficients when the error changes; just show how the error changes as you change the weights for "awesome".

How is this graph similar or different to error graphs we have seen so far in this class (i.e. regression)? Can we use the same techniques to optimize classification error as we can RSS? Why or why not?

# QUESTION 1 SOLUTION



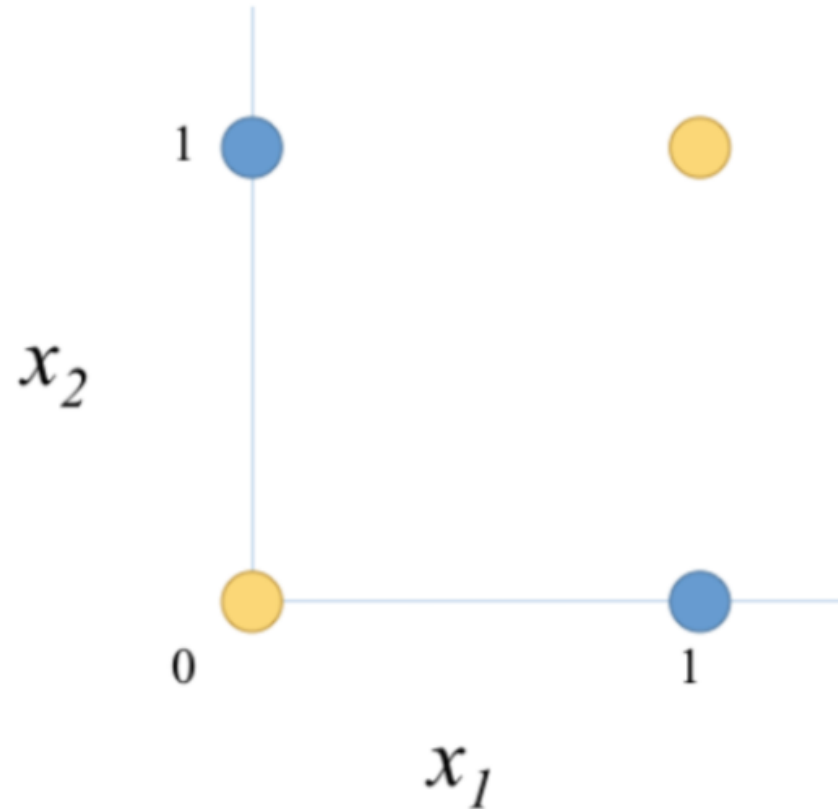
Answer:



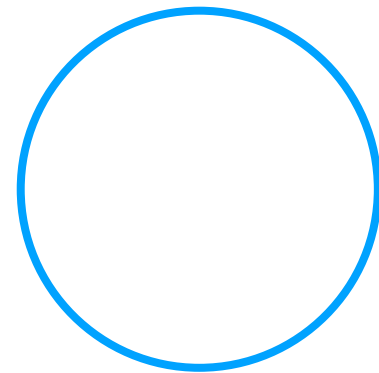
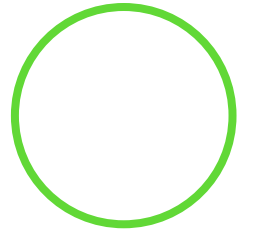
This graph is similar because it has the general shape of decreasing and then increasing, meaning there is a optimal point. It's different because it has flat sections and is not continuous. We can't use gradient descent here the function is not differentiable everywhere and has 0 slope in most places; this means we would never take a step since there is no indication of up or down.

# QUESTION 2

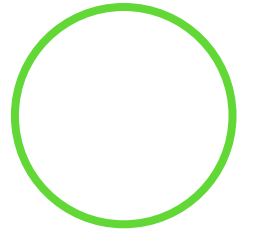
Consider the dataset below with two data inputs  $x_1$  and  $x_2$  and labels (blue) and (yellow)



- Can you use logistic regression with linear features  $\text{Score}(x) = w_0 + w_1x_1 + w_2x_2$  to perfectly classify this dataset? If it is possible, show which weights  $w$  can make the correct predictions and if it is not possible, explain why.
- What if we added a third feature  $x_1x_2$  to the model? If it is possible, show which weights  $w$  can make the correct predictions and if it is not possible, explain why.



# QUESTION 2 SOLUTION



- a. Can you use logistic regression with linear features  $Score(x) = w_0 + w_1x_1 + w_2x_2$  to perfectly classify this dataset? If it is possible, show which weights  $w$  can make the correct predictions and if it is not possible, explain why.

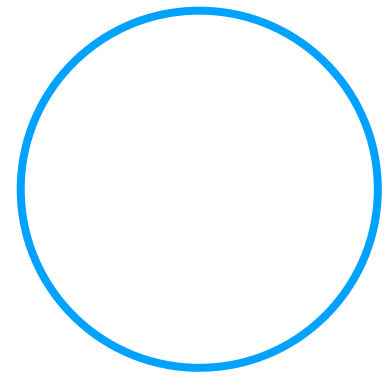
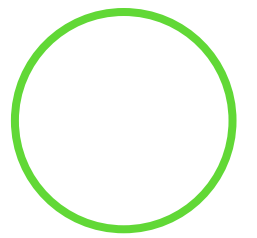
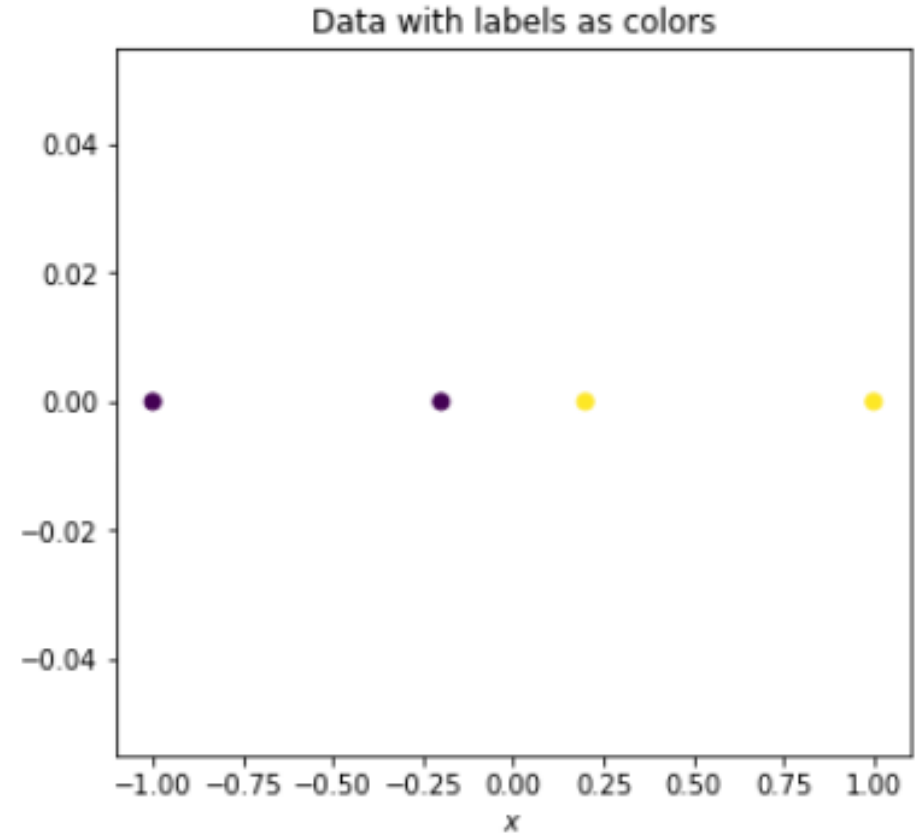
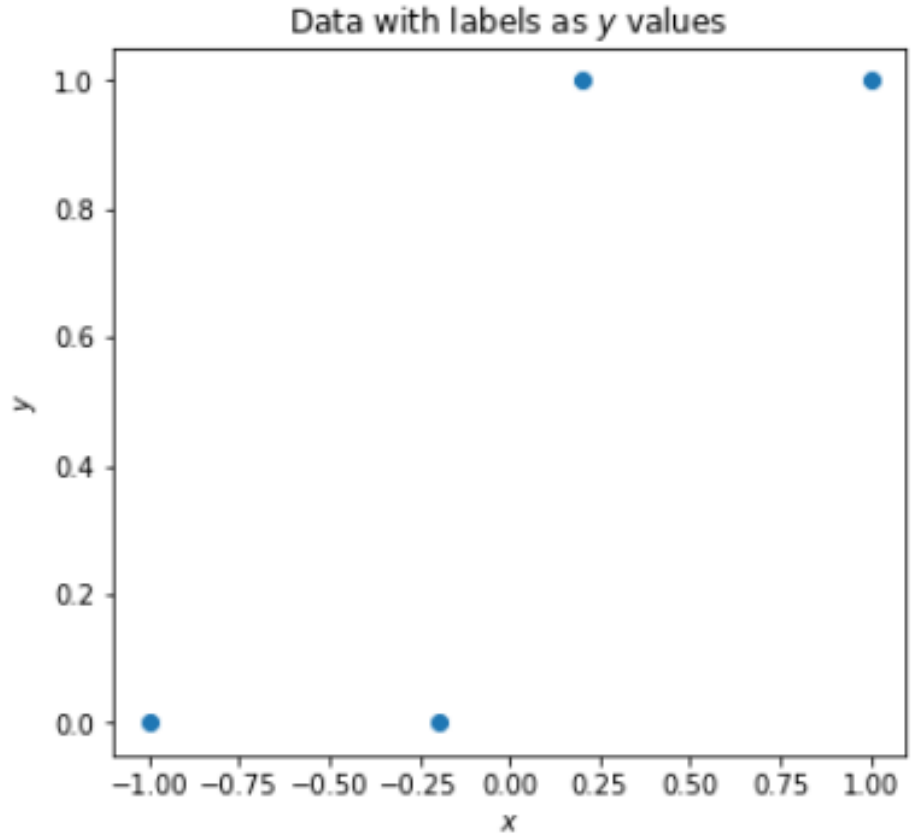
**Answer:** No, this model can only learn linear decision boundaries and it is not possible to draw a line that separates the blue points and the yellow points.

- b. What if we added a third feature to the model  $h_3(x) = x_1x_2$ ? If it is possible, show which weights  $w$  can make the correct predictions and if it is not possible, explain why.

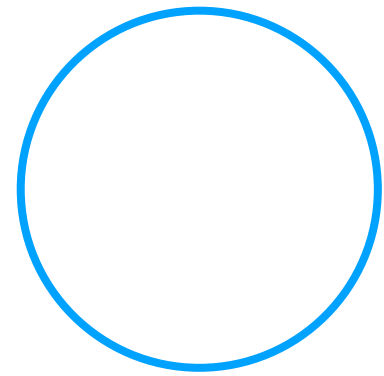
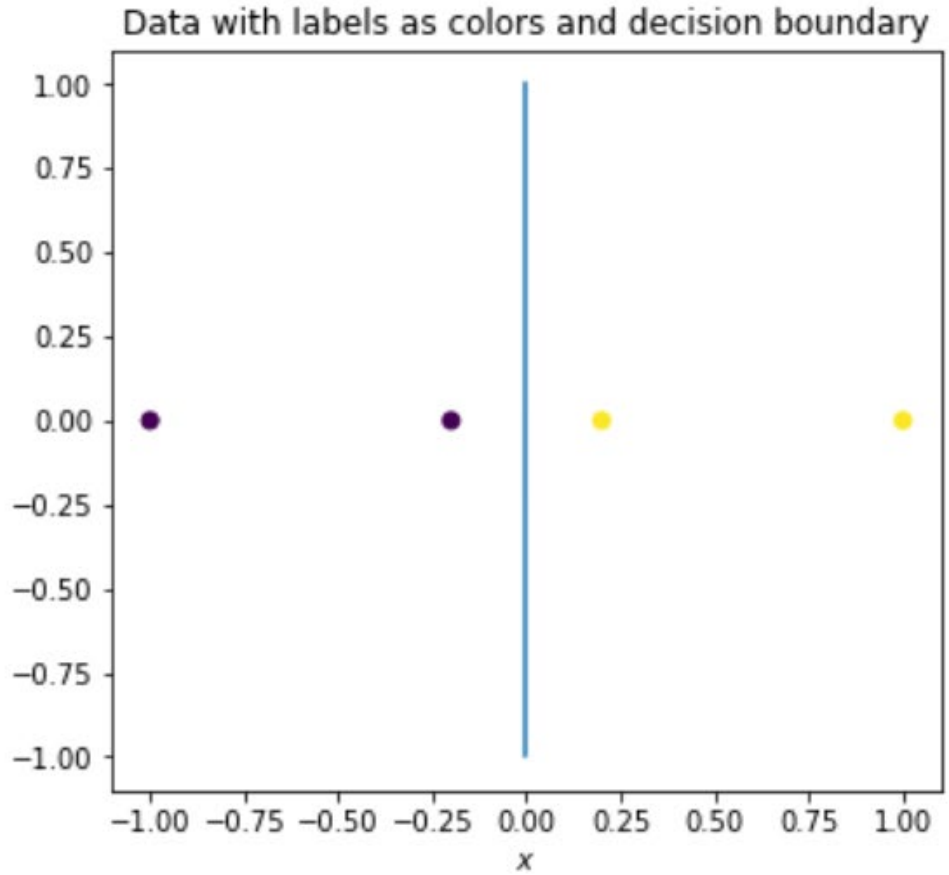
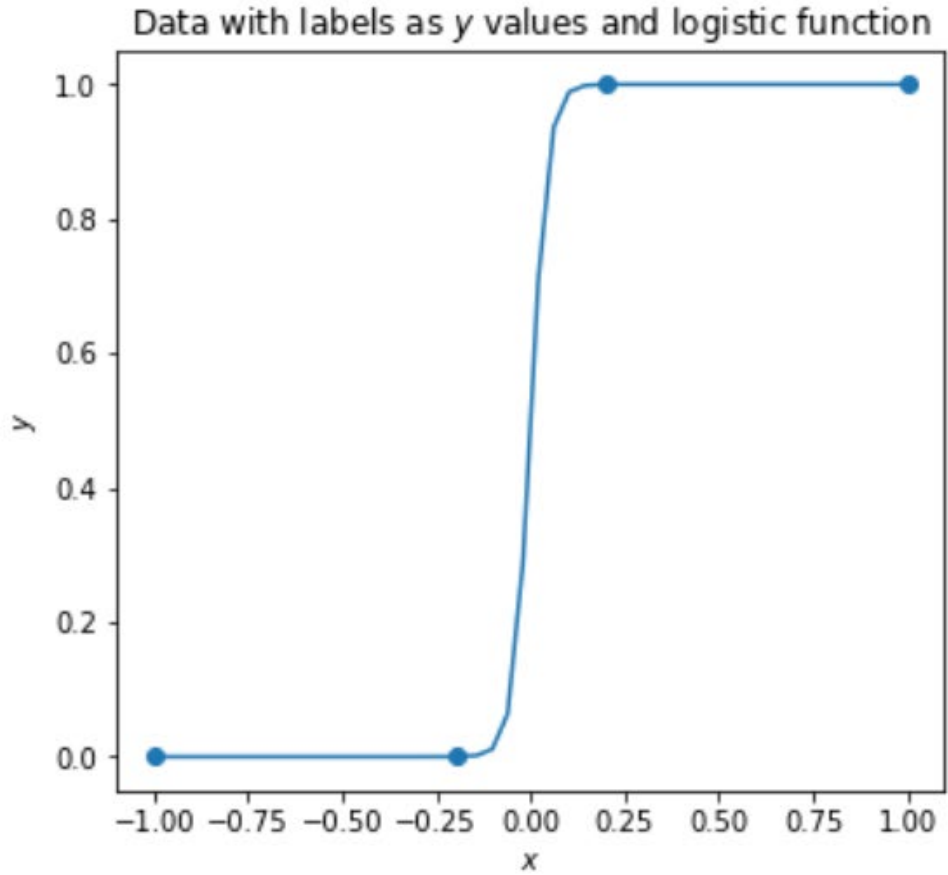
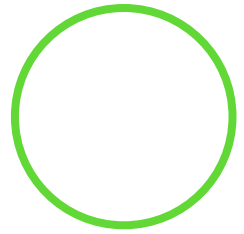
**Answer:** One possible answer  $w_0 = -0.5, w_1 = 1, w_2 = 1, w_3 = -2$ .

# QUESTION 3

Consider the dataset in the notebook. Where do we expect the decision boundary to be? What do we expect  $\beta_0$  and  $\beta_1$  to be?



# QUESTION 3 SOLUTION



Mathematician  
studying deep  
learning



Boy, am I glad I studied linear algebra,  
numerical optimisation, measure theory,  
and convex optimisation during my  
undergrad. This sure would be difficult  
without it

Deep learner  
studying the maths



Wot's chaim rool?

# Other Stuff

PART IDK

# MEMES

Tbh the most fun thing (at least for me) after taking 416 is you start to understand memes about machine learning...

Here's my source of memes lol as the quarter goes you'll understand these memes more and more!

<https://www.facebook.com/groups/1638417209555402>

**(AI Memes) AI & Deep Learning Memes For Back-propagated Poets**

Public group · 76.2K members

when they mention AI in a movie

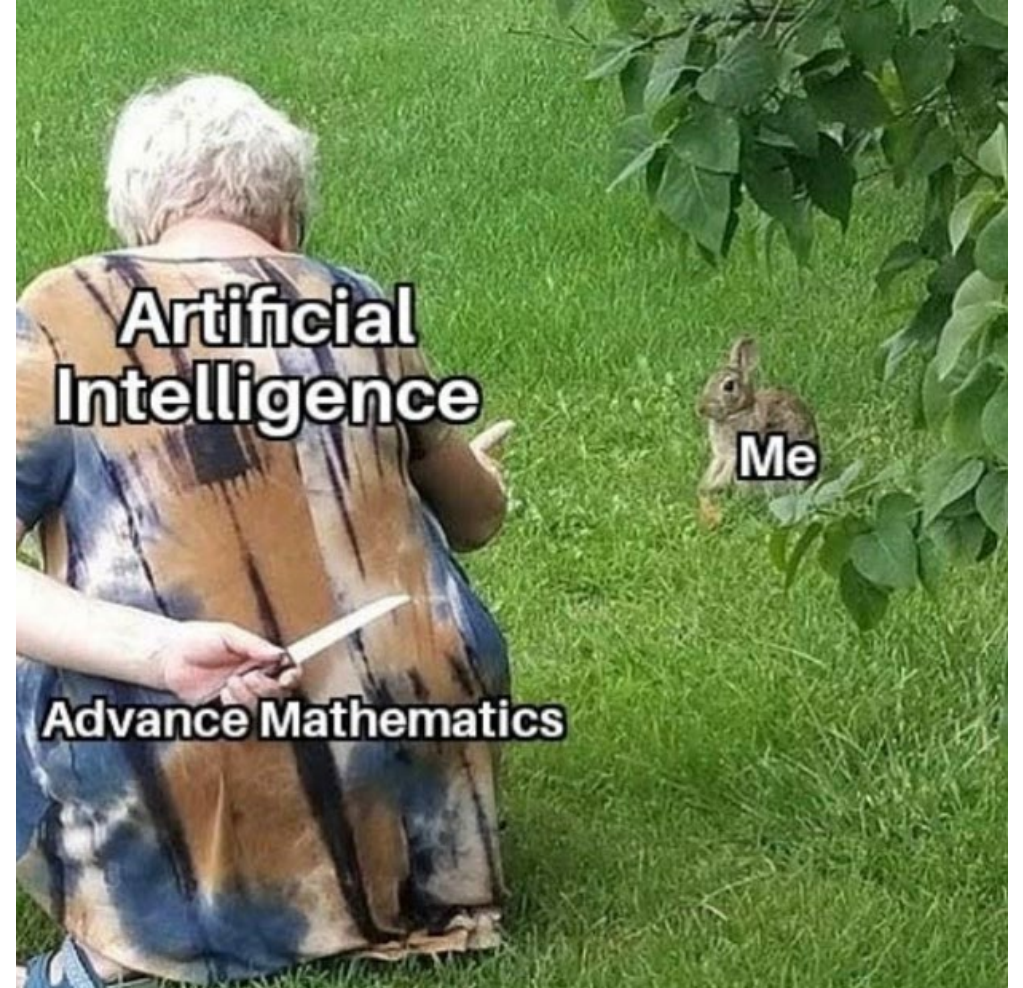






# CREDITS

1. [Understanding Confusion Matrix](#)
2. [Understanding ROC Curves](#)
3. [A Gentle Introduction to Maximum Likelihood Estimation](#)





# LICENSE

This material is originally made by [Hongjun Wu](#) for the course [CSE416: Introduction to Machine Learning](#) in the Summer 2020 quarter taught by [Vinitra Swamy](#), at University of Washington Paul G. Allen School of Computer Science and Engineering.

It was originally made for educational purpose, in a section taught by teaching assistants to help students explore material in more depth.

Any other materials used are cited in the Credits section.

This material is licensed under the [Creative Commons License](#).

Anyone, especially other educators and students, are welcomed and strongly encouraged to study and use this material.