06 Logistic Regression

Created Date: 2022-09-27

> Metadata 🧰
>
> - Title: Logistic Regression
> - Author: Andrew Jones
> - Reference: Data Science Infinity

> Links & Tags 🖇️
>
> - Index: Course Note Index
> - Atomic Tag: #datascience
> - Subatomic Tags: #machinelearning #logiticregression #classi

---

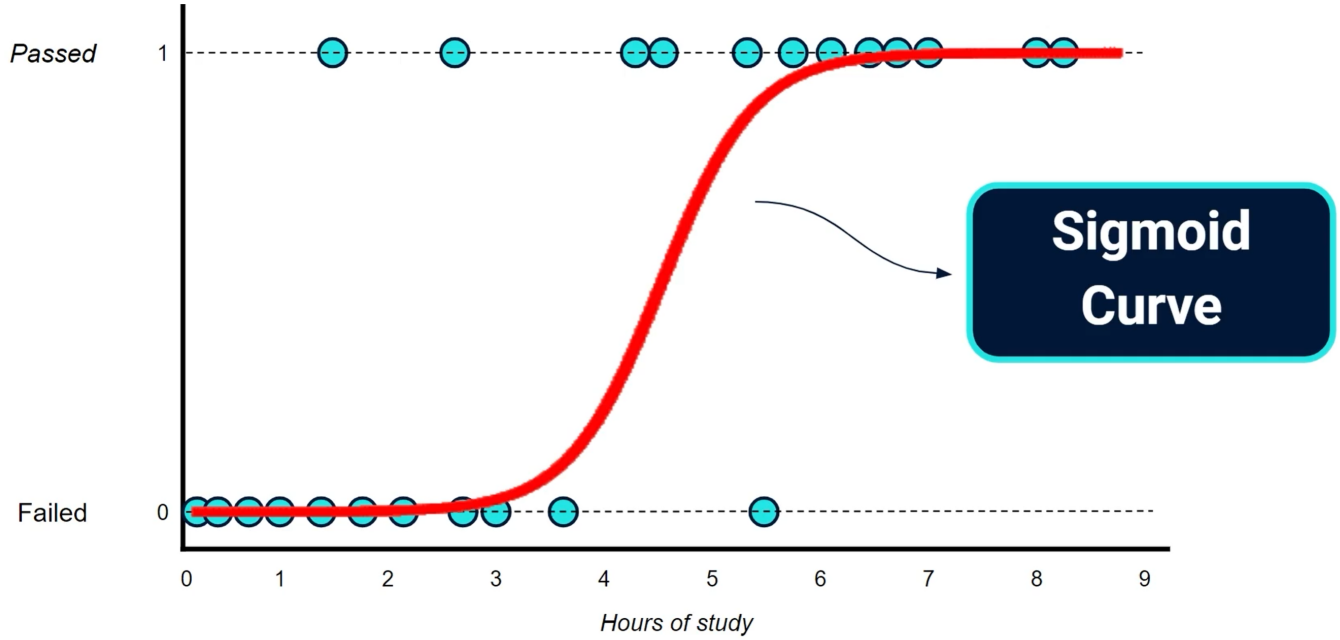
High-Level Overview

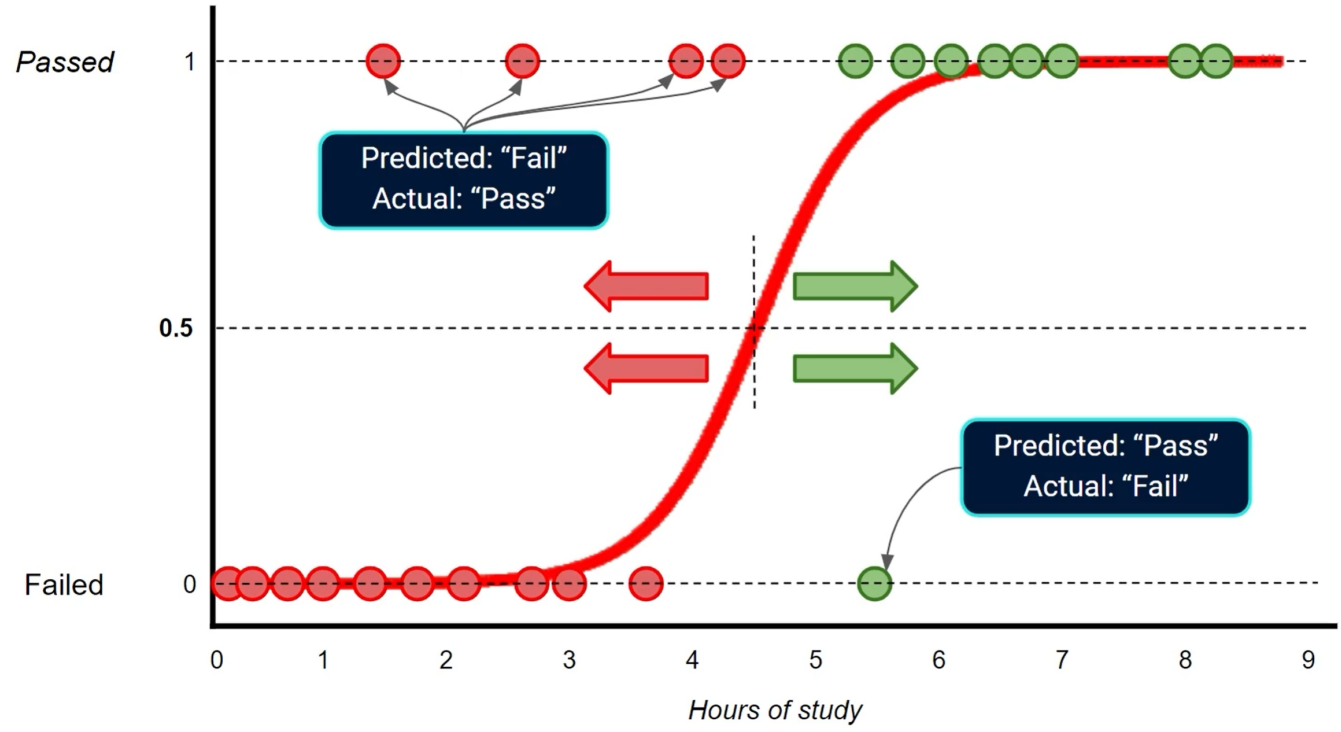
Logistic Regression in 3 Minutes Jupyter Notebook: Basic Logistic Regression Template

> Logistic Regression is a model used to predict the probability of a certain class or event based on one or more input variables. It transforms linear relationships to a probabilistic output through the Logistic Function.

- Logistic Regression is commonly used for classification modeling
- The output variables are binary (true or false)
    - Did the event take place or not
- The logistic regression curve is called the Sigmoid curve, and falls between 0 and 1 (probability of an event occurring)

- A classification threshold is set and determines if the probability of an event is classified as true or false
- The accuracy score assesses how many classifications are predicted accurately when run on the test set





Advanced Theory

Jupyter Notebook: Advanced Logistic Regression Template

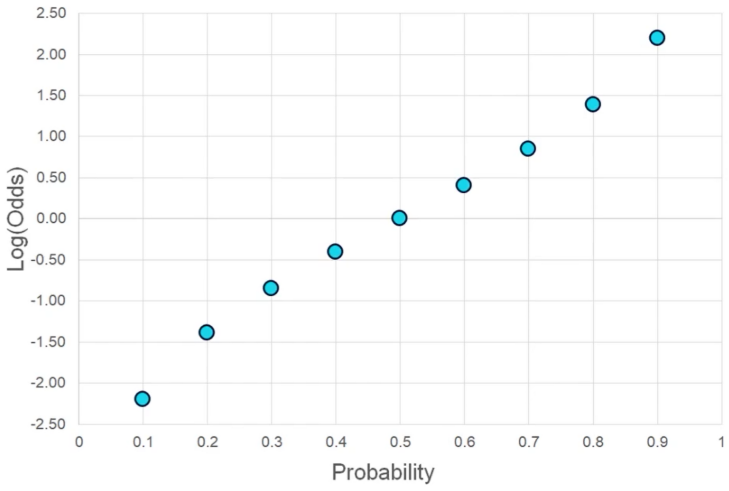
Probability, Odds, & Log(Odds)

$$Probability = \frac{Possible\ Outcomes}{Total\ Outcomes}$$

$$Odds = \frac{Probability\ Of\ Event\ Occuring}{Probability\ Of\ Event\ Not\ Occuring} = \frac{p}{1-p}$$

- Probability $\neq$ Odds
- If the probability is against an event happening, the odds can only be between 0 and 1
- If the probability is in favor of an event happening, the odds can be from 1 to infinity
- Making it hard to compare the odds for similar probabilities on either side of an even chart
- The log(Odds) solves this problem by distributing the odds evenly

Probability	Odds	log(Odds)
0.0	0.0	NEG INFINITY
0.1	0.11	-2.2
0.2	0.25	-1.39
0.3	0.43	-0.85
0.4	0.66	-0.41
0.5	1	0
0.6	1.5	0.41
0.7	2.33	0.85
0.8	4	1.39
0.9	9	2.2
1.0	INFINITY	INFINITY



Logistic Regression Formula

- $y_{\ log(odds)} = bx + c$

Sigmoid Curve

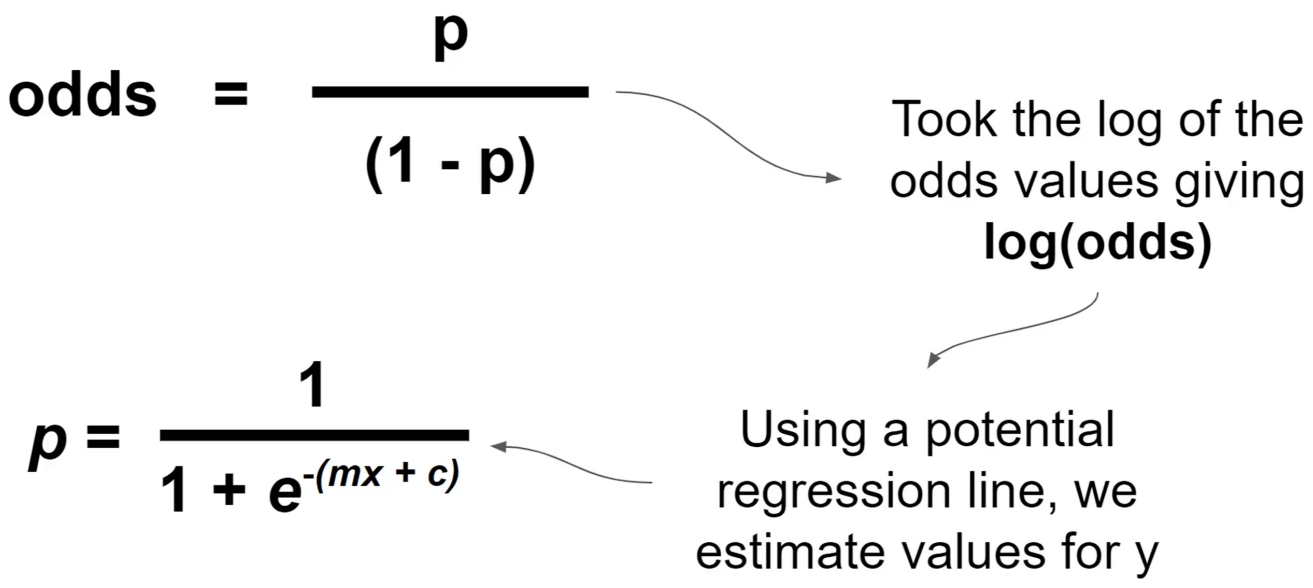
- Computes a probability prediction based on the log(Odds) values
    - Converts the straight line of best fit to a Sigmoid curve

- The y-axis needs to be converted to log(Odds) in order to apply a line of best fit

- We can then extrapolate predicted log(Odds) values and convert them back to probabilities (range between 0 and 1) using the Sigmoid function

- The logistic formula can now be interpreted as a one unit change with the input variable causes a log unit change in the output variable

-
    - $p = \frac{1}{1+e^{-y}} = \frac{1}{1+e^{-bx+c}}$
    - e Euler's Number (Inverse of Log)
    - $bx + c$ Logistic Regression Formula



Maximum Likelihood Estimation

- Used to find the best fitting line in logistic regression
- We can't use the least squares method because y is converted to log

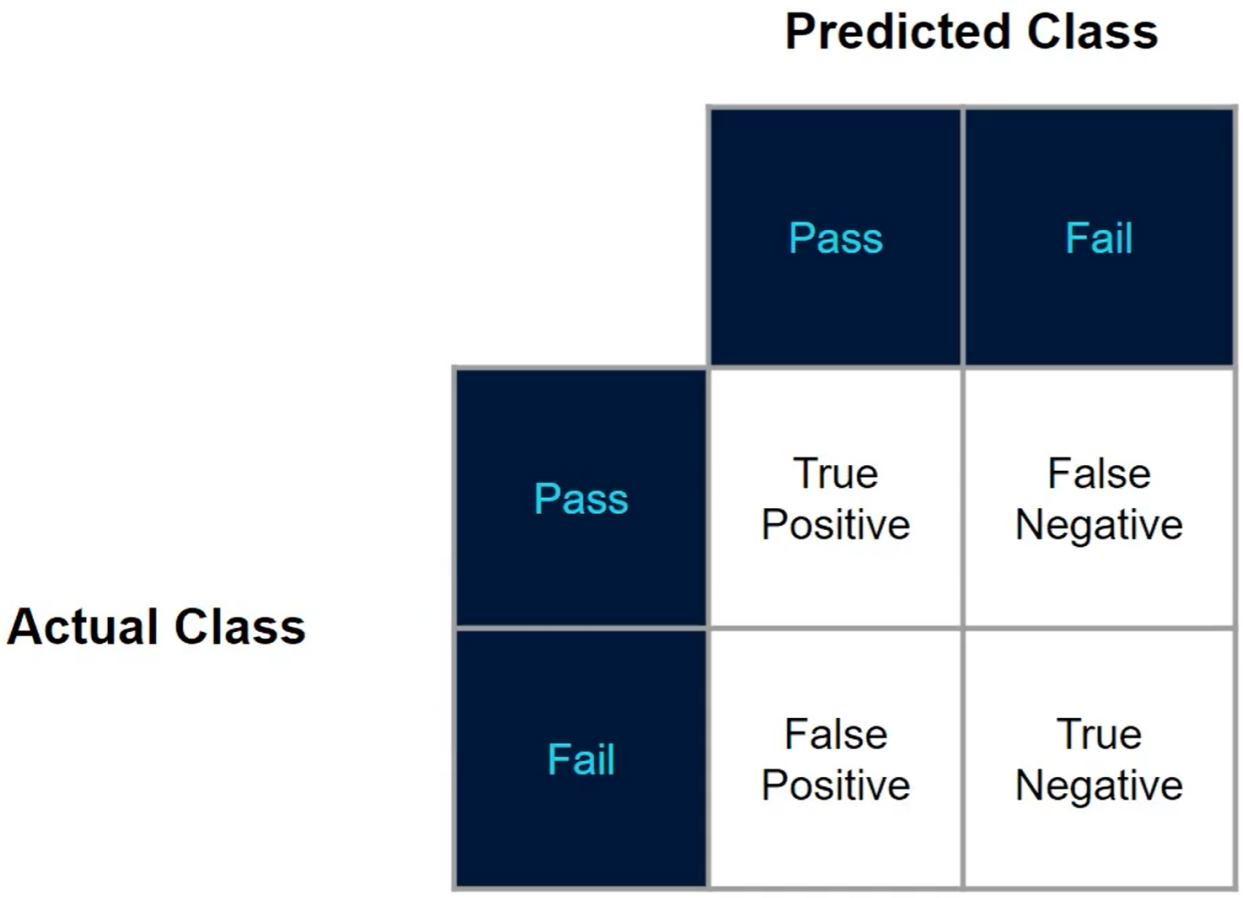
- We sum all the log (likelihoods) of events that passed and failed to find how well the line fits the data
    - This process is repeated until the line with the highest sum of log (likelihoods) is found, which is the line of best fit for logistic regression

Evaluating Classification Accuracy

- $Classification\ Accuracy = \frac{n\ Outcomes\ Classified\ Correctly}{Total\ Outcomes}$

- n Number of Outcomes Classified Correctly

- - Type I Error: False Positive - Type II Error: False Negative - Diagonal (True Positive & True Negative): Outcomes Correctly Classified



Advanced Evaluation Techniques

- When there is a large bias towards one of the classes we have an imbalanced data set
- Advanced techniques help evaluate models when we have imbalanced data
    - Precision evaluates how many observations were predicted as positive who were actually positive
    - Recall (Sensitivity) evaluates how many observations were predicted as positive who were actually positive (also referred to as the True Positive Rate)
    - False Positive Rate evaluates how many observations were predicted as positive who were actually negative
    - F1-Score evaluates the harmonic mean of Precision and Recall
        - A good F1-Score comes when there is a balance between Precision & Recall, rather than a disparity between them
- Precision & Recall can not be optimized together, sometimes it makes sense to adapt a model to optimize one of these metrics
    - As an example, in a disease diagnoses model this would evaluate observations that were not predicted to have a disease who actually have the disease
    - In this example, it may make sense to optimize Recall while still being cognizant that we don't want to misdiagnose people as positive when they are in fact negative

Precision	Recall	Meaning
High	High	The model is differentiating between classes well
High	Low	The model is struggling to detect the class, but when it does it is very trustworthy
Low	High	The model is identifying most of the class, but is also incorrectly including a high number of data points from another class
Low	Low	The model is struggling to differentiate between classes

Advanced Evaluation Metrics

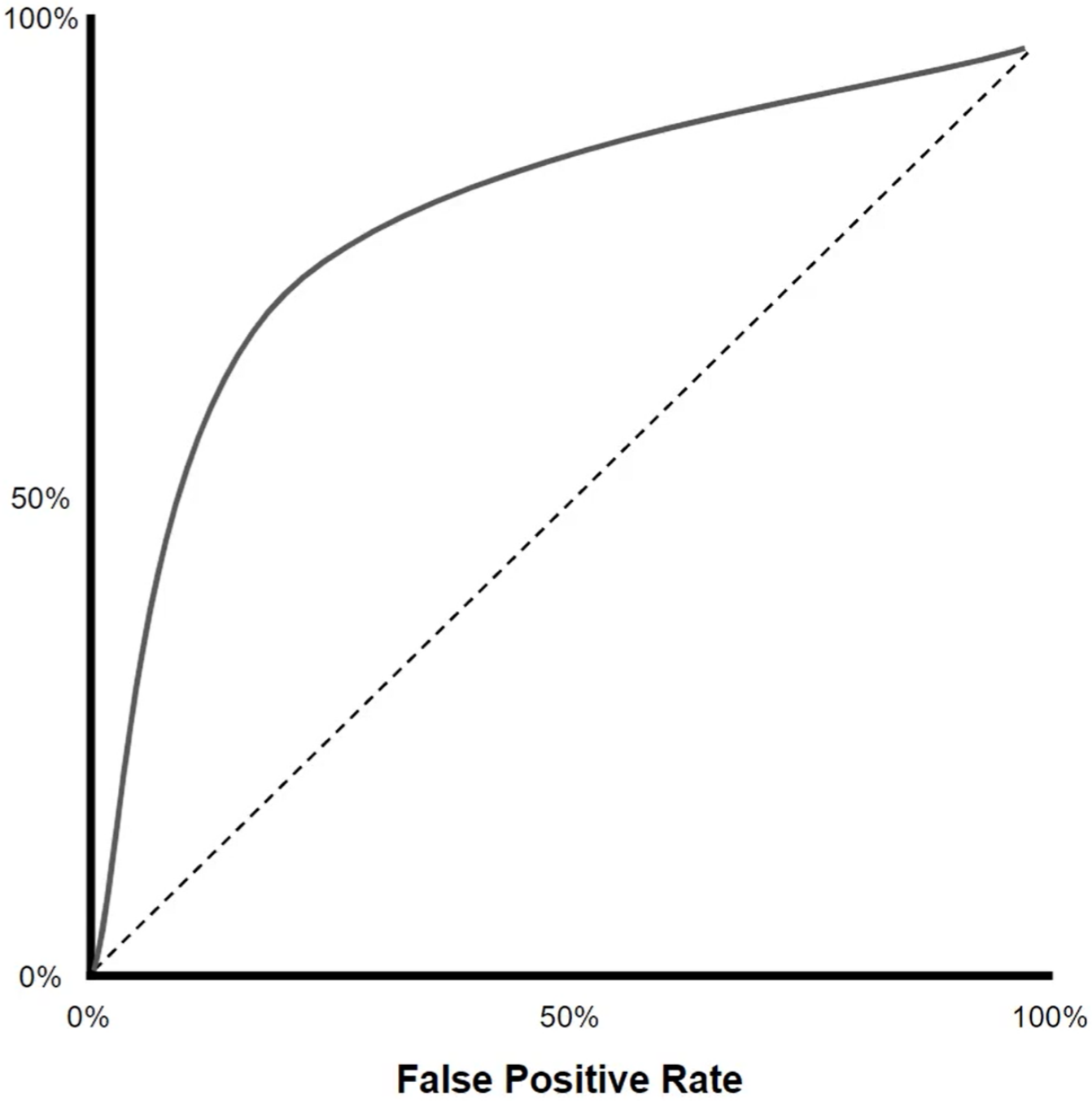
- $Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$
- $True\ Positive\ Rate\ (Recall\ Sensitivity) = \frac{True\ Positive}{True\ Positive + False\ Negative}$
- $False\ Positive\ Rate = \frac{False\ Positive}{False\ Positive + TrueNegative}$
- $F1\ Score = \frac{2 * (Recall * Precision)}{Recall + Precision}$

Changing the Classification Threshold

- The default classification threshold is 50%
- A low threshold will classify more observations as positive, while a high threshold will classify more as negative
- Changing the threshold will impact the Precision and Recall evaluation metrics
- We can visualize the impact of changing the threshold on the TPR and FPR metrics using an ROC curve

- The dashed lines would represent observations that had equal TPR and FPR results
- The solid line represents the actual results of the TPR and FPR metrics calculated for varying thresholds
- Observations to the left of the dashed line are good, as they infer the model has proportionately lower incorrect classifications (false positives)
- We can optimize the threshold by picking a threshold that results in the furthest point from the dashed line
- ROC curves can also be used to compare the accuracy of different classification models by calculating the area under the curve (AUC)
    - A larger AUC is considered to be a better performing model
- The ROC curve can be misleading when we have an imbalanced data set
    - In this case, we aim to optimize the F1 Score

> ROC (Receiver Operator Characteristic) Curve visualizes the trade-off between the *True Positive Rate* and the *False Positive Rate* across varying classification thresholds.


100%
50%
0%
0%
50%
100%
False Positive Rate