

Você ainda tem 2 histórias gratuitas exclusivas para membros neste mês. [Inscreva-se](#) no Medium e ganhe um extra.

◆ História exclusiva para membros

Micro, Macro e Médias Ponderadas da Pontuação F1, Claramente Explicadas

Compreender os conceitos por trás da média micro, média macro e média ponderada da pontuação F1 na classificação multiclasse com ilustrações simples



Kenneth Leung · Seguir

Publicado em Rumo à ciência de dados

7 minutos de leitura · 4 de janeiro de 2022

Listen

Share

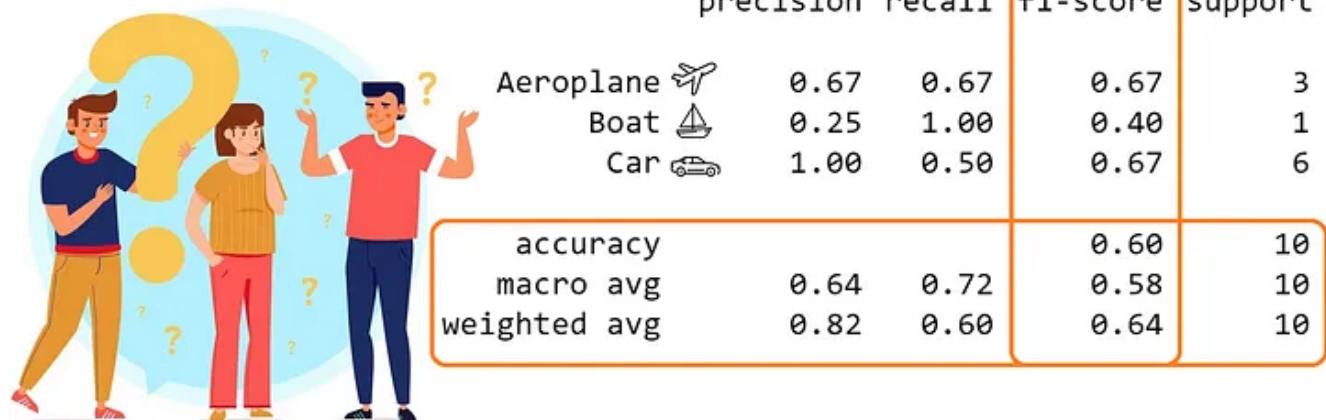


Imagen do autor e [Freepik](#)

A pontuação F1 (também conhecida como F-measure) é uma métrica popular para avaliar o desempenho de um modelo de classificação.

In the case of multi-class classification, we adopt averaging methods for F1 score calculation, resulting in a set of different average scores (macro, weighted, micro) in the classification report.

This article looks at the meaning of these averages, how to calculate them, and which one to choose for reporting.

Contents

-
- (1) [Recap of the Basics \(Optional\)](#)
 - (2) [Setting the Motivating Example](#)
 - (3) [Macro Average](#)
 - (4) [Weighted Average](#)
 - (5) [Micro Average](#)
 - (6) [Which average should I choose?](#)
-

(1) Recap of the Basics (Optional)

Note: Skip this section if you are already familiar with the concepts of precision, recall, and F1 score.

Precision

Layman definition: Of all the positive predictions I made, how many of them are truly positive?

Calculation: Number of True Positives (TP) divided by the Total Number of True Positives (TP) and False Positives (FP).

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

The equation for precision | Image by author

Recall

Layman definition: Of all the actual positive examples out there, how many of them did I correctly predict to be positive?

Calculation: Number of True Positives (TP) divided by the Total Number of True Positives (TP) and False Negatives (FN).

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

The equation for Recall | Image by author

If you compare the formula for precision and recall, you will notice both look similar. The only difference is the second term of the denominator, where it is False Positive for precision but False Negative for recall.

F1 Score

To evaluate model performance comprehensively, we should examine **both** precision and recall. The F1 score serves as a helpful metric that considers both of them.

Definition: Harmonic mean of precision and recall for a more balanced summarization of model performance.

Calculation:

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

The equation for F1 score | Image by author

If we express it in terms of True Positive (TP), False Positive (FP), and False Negative (FN), we get this equation:

$$\text{F1 Score} = \frac{\text{TP}}{\text{TP} + \frac{1}{2} (\text{FP} + \text{FN})}$$

The alternative equation for F1 score | Image by author

(2) Setting the Motivating Example

To illustrate the concepts of averaging F1 scores, we will use the following example in the context of this tutorial.

Imagine we have trained an **image classification model** on a **multi-class dataset** containing images of **three classes**: Airplane, Boat, and Car.



Image by [macrovector — freepik.com](#)

We use this model to **predict** the classes of **ten** test set images. Here are the **raw predictions**:

No	Actual	Predicted	Match
1	Airplane	Airplane	✓
2	Car	Boat	✗
3	Car	Car	✓
4	Car	Car	✓
5	Car	Boat	✗
6	Airplane	Boat	✗
7	Boat	Boat	✓
8	Car	Airplane	✗
9	Airplane	Airplane	✓
10	Car	Car	✓

Sample predictions of our demo classifier | Image by author

Upon running `sklearn.metrics.classification_report`, we get the following classification report:

	precision	recall	f1-score	support
Aeroplane	0.67	0.67	0.67	3
Boat	0.25	1.00	0.40	1
Car	1.00	0.50	0.67	6
accuracy			0.60	10
macro avg	0.64	0.72	0.58	10
weighted avg	0.82	0.60	0.64	10

Per-Class F1 scores

Average F1 scores

[Relatório de classificação do pacote scikit-learn | imagem do autor](#)

As colunas (em laranja) com as pontuações **por turma** (isto é, a pontuação de cada turma) e as pontuações **médias** são o foco de nossa discussão.

Podemos ver acima que o conjunto de dados está **desequilibrado** (apenas uma das dez instâncias do conjunto de teste é 'Boat'). Assim, a **proporção de correspondências corretas** (também conhecidas como precisão) seria ineficaz na avaliação do desempenho do modelo.

Em vez disso, vamos olhar para a **matriz de confusão** para uma compreensão holística das previsões do modelo.

		Predicted		
		Airplane	Boat	Car
Actual	Airplane	2	1	0
	Boat	0	1	0
	Car	1	2	3

Matriz de confusão | imagem do autor

A matriz de confusão acima nos permite calcular os valores críticos de Verdadeiro Positivo (TP), Falso Positivo (FP) e Falso Negativo (FN), conforme mostrado abaixo.

Label	True Positive (TP)	False Positive (FP)	False Negative (FN)
Airplane	2	1	1
Boat	1	3	0
Car	3	0	3

Valores calculados de TP, FP e FN da matriz de confusão | imagem do autor

A tabela acima nos prepara bem para calcular os valores **por classe de precisão**, recall e pontuação F1 para cada uma das três classes.

É importante lembrar que na **classificação multiclasse**, calculamos a pontuação F1 para cada classe em uma abordagem One-vs-Rest (OvR) em vez de uma única pontuação F1 geral, como visto na classificação binária.

Nesta abordagem **OvR**, determinamos as métricas para cada classe separadamente, como se houvesse um classificador diferente para cada classe. Aqui estão as métricas por classe (com o cálculo da pontuação F1 exibido):

Label	True Positive (TP)	False Positive (FP)	False Negative (FN)	Precision	Recall	F1 Score
 Airplane	2	1	1	0.67	0.67	$2 * (0.67 * 0.67) / (0.67 + 0.67) = \mathbf{0.67}$
 Boat	1	3	0	0.25	1.00	$2 * (0.25 * 1.00) / (0.25 + 1.00) = \mathbf{0.40}$
 Car	3	0	3	1.00	0.50	$2 * (1.00 * 0.50) / (1.00 + 0.50) = \mathbf{0.67}$

No entanto, em vez de ter várias pontuações F1 por classe, seria melhor tirar a média delas para obter um único número para descrever o desempenho geral.

Agora, vamos discutir os métodos de média que levaram às três pontuações médias diferentes da F1 do relatório de classificação .

(3) Média Macro

A **média macro** é talvez o mais direto entre os vários métodos de média.

A pontuação F1 macro-média (ou pontuação F1 macro) é calculada usando a média aritmética (também conhecida como média **não ponderada**) de todas as pontuações F1 por classe.

Esse método trata todas as classes igualmente, independentemente de seus valores de suporte .

Label	Per-Class F1 Score	Macro-Averaged F1 Score
 Airplane	0.67	$\frac{0.67 + 0.40 + 0.67}{3} = 0.58$
 Boat	0.40	
 Car	0.67	

Cálculo da pontuação F1 macro | imagem do autor

O valor de 0,58 que calculamos acima corresponde à pontuação F1 média macro em nosso relatório de classificação.

	precision	recall	f1-score	support
Aeroplane	0.67	0.67	0.67	3
Boat	0.25	1.00	0.40	1
Car	1.00	0.50	0.67	6
accuracy			0.60	10
macro avg	0.64	0.72	0.58	10
weighted avg	0.82	0.60	0.64	10

(4) Média ponderada

A pontuação F1 média ponderada é calculada tomando a média de todas as pontuações F1 por classe , considerando o suporte de cada classe .

O *suporte* refere-se ao número de ocorrências reais da classe no conjunto de dados. Por exemplo, o valor de suporte de 1 em *Boat* significa que há apenas uma observação com um rótulo real de *Boat*.

O 'peso' refere-se essencialmente à proporção do suporte de cada classe em relação à soma de todos os valores de suporte.

Label	Per-Class F1 Score	Support	Support Proportion	Weighted Average F1 Score
Airplane	0.67	3	0.3	
Boat	0.40	1	0.1	
Car	0.67	6	0.6	
Total	-	10	1.0	$(0.67 * 0.3) + (0.40 * 0.1) + (0.67 * 0.6) = 0.64$

Cálculo da pontuação F1 ponderada | imagem do autor

Com a média ponderada, a média de saída teria contabilizado a contribuição de cada classe ponderada pelo número de exemplos daquela determinada classe.

O valor calculado de 0,64 coincide com a pontuação F1 média ponderada em nosso relatório de classificação.

	precision	recall	f1-score	support
Aeroplane	0.67	0.67	0.67	3
Boat	0.25	1.00	0.40	1
Car	1.00	0.50	0.67	6
accuracy			0.60	10
macro avg	0.64	0.72	0.58	10
weighted avg	0.82	0.60	0.64	10

(5) Micromédia

A micromédia calcula uma pontuação F1 média global contando as somas dos Verdadeiros Positivos (TP), Falsos Negativos (FN) e Falsos Positivos (FP).

Primeiro somamos os respectivos valores de TP, FP e FN em todas as classes e, em seguida, os inserimos na equação F1 para obter nossa pontuação micro F1.

Label	True Positive (TP)	False Positive (FP)	False Negative (FN)	Micro-Averaged F1 Score
Airplane	2	1	1	$\frac{TP}{TP + \frac{1}{2}(FP+FN)} = \frac{6}{6 + \frac{1}{2}(4+4)} = 0.60$
Boat	1	3	0	
Car	3	0	3	
TOTAL	6	4	4	

Cálculo da pontuação micro F1 | imagem do autor

No relatório de classificação, você pode estar se perguntando por que nossa pontuação micro F1 de 0,60 é exibida como 'precisão' e por que NÃO há linha informando 'micro média'.

	precision	recall	f1-score	support
accuracy			0.60	10
macro avg	0.64	0.72	0.58	10
weighted avg	0.82	0.60	0.64	10

Isso ocorre porque a micromédia essencialmente calcula a proporção de observações classificadas corretamente de todas as observações. Se pensarmos sobre isso, essa definição é o que usamos para calcular a precisão geral .

Além disso, se fizermos uma micromédia para precisão e recuperação, obteremos o mesmo valor de 0,60 .

Label	True Positive (TP)	False Positive (FP)	False Negative (FN)	Micro-Averaged Values
Airplane	2	1	1	$\text{Precision} = \frac{6}{6+4} = 0.60$ $\text{Recall} = \frac{6}{6+4} = 0.60$ $\text{F1 Score} = \frac{6}{6 + \frac{1}{2}(4+4)} = 0.60$
Boat	1	3	0	
Car	3	0	3	
TOTAL	6	4	4	

Esses resultados significam que em casos de classificação multiclasse em que cada observação tem um único rótulo , micro-F1 , microprecisão , microrecall e exatidão compartilham o mesmo valor (ou seja, 0,60 neste exemplo).

E isso explica por que o relatório de classificação precisa exibir apenas um único valor de precisão, já que micro-F1, microprecisão e microrecall também têm o mesmo valor.

micro-F1 = precisão = microprecisão =
microrecuperação

Junte-se ao Medium com meu link de referência - Kenneth Leung

Acesse todo o meu conteúdo (e todos os artigos do Medium) pelo preço de apenas um café!

kennethleungty.medium.com

(6) Qual média devo escolher?

Em geral, se você estiver trabalhando com um conjunto de dados desequilibrado em que todas as classes são igualmente importantes, usar a média **macro** seria uma boa escolha, pois trata todas as classes igualmente.

Isso significa que, para nosso exemplo envolvendo a classificação de aviões, barcos e carros, usariamos a pontuação macro-F1.

Se você tiver um conjunto de dados desequilibrado, mas quiser atribuir maior contribuição a classes com mais exemplos no conjunto de dados, a média ponderada é preferida.

Isso ocorre porque, na média ponderada, a contribuição de cada classe para a média da F1 é ponderada pelo seu tamanho.

Suponha que você tenha um conjunto de dados equilibrado e queira uma métrica facilmente compreensível para desempenho geral, independentemente da classe. Nesse caso, você pode ir com precisão, que é essencialmente nossa pontuação **micro F1**.

Antes de você ir

Convido você a se juntar a mim em uma jornada de aprendizado de ciência de dados! Siga minha página [do Medium](#) e confira meu [GitHub](#) para ficar por dentro de conteúdos mais empolgantes sobre ciência de dados. Enquanto isso, divirta-se interpretando as pontuações da F1!

Suposições de Regressão Logística, Claramente Explicadas

Entenda e implemente verificações de suposições para uma das técnicas de modelagem mais essenciais

[Open in app](#) ↗

[Sign up](#)

[Sign In](#)



Meio de pesquisa



▼

Por que o Bootstrapping realmente funciona

Uma explicação simples de leigo sobre por que essa técnica popular faz sentido

em direção a [datascience.com](#)

O Problema ReLU Morrer, Claramente Explicado

Mantenha sua rede neural viva ao entender as desvantagens do ReLU

em direção a [datascience.com](#)

Junte-se ao Medium com meu link de referência - Kenneth Leung

Como membro do Medium, uma parte de sua taxa de assinatura vai para os escritores que você lê e você obtém acesso total a todas as...

[kennethleungty.medium.com](#)

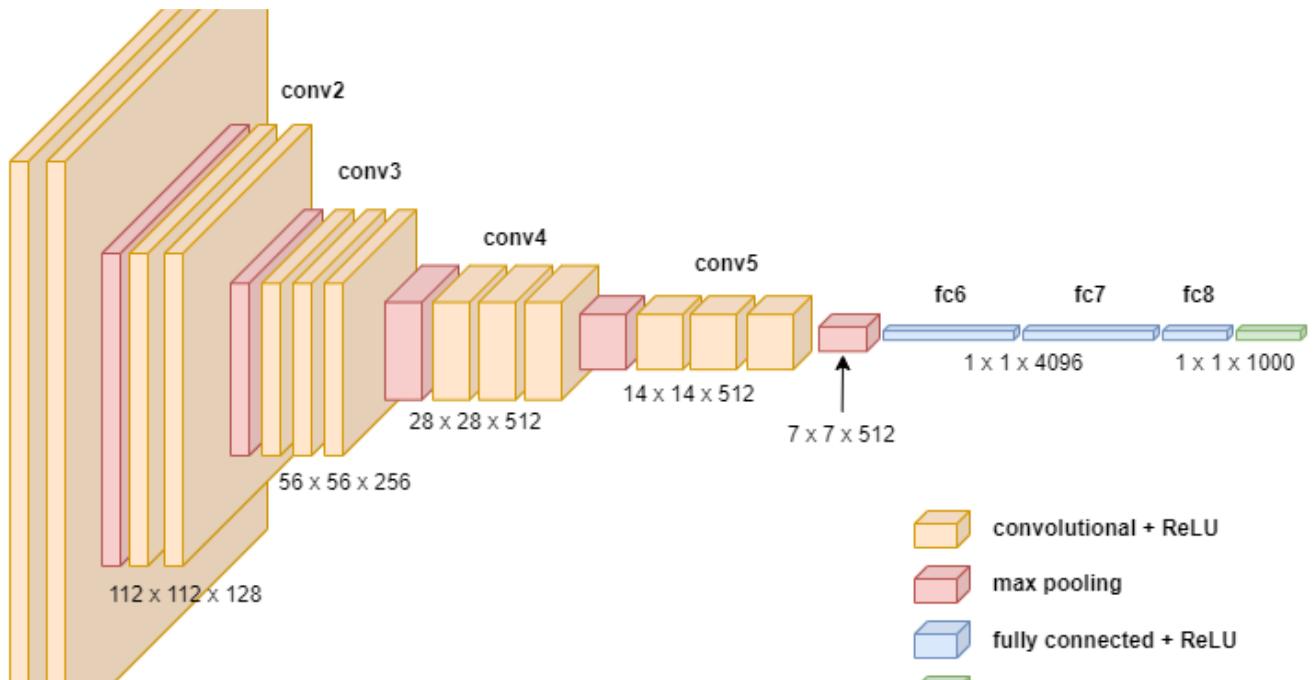

[Follow](#)


Written by Kenneth Leung

2.7K Followers · Writer for Towards Data Science

Data Scientist at BCG | ML Engineer | 1M+ views | linkedin.com/in/kennethleungty | Join me on Medium: kennethleungty.medium.com/membership

More from Kenneth Leung and Towards Data Science



Kenneth Leung in Towards Data Science

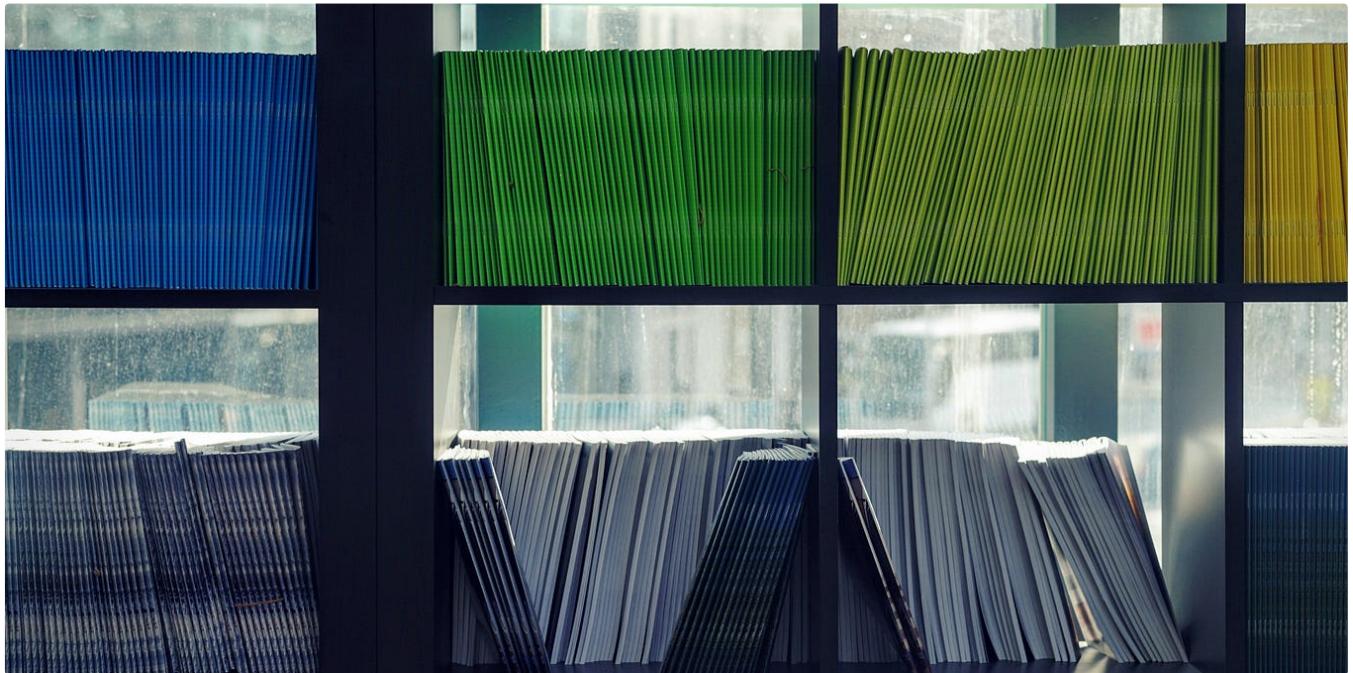
How to Easily Draw Neural Network Architecture Diagrams

Using the no-code diagrams.net tool to showcase your deep learning models with diagram visualizations

◆ · 5 min read · Aug 23, 2021

👏 538

💬 6



 Jacob Marks, Ph.D. in Towards Data Science

How I Turned My Company's Docs into a Searchable Database with OpenAI

And how you can do the same with your docs

15 min read · Apr 25

👏 2.1K

💬 26





Leonie Monigatti in Towards Data Science

Getting Started with LangChain: A Beginner's Guide to Building LLM-Powered Applications

A LangChain tutorial to build anything with large language models in Python

◆ · 12 min read · Apr 25

998

12



Kenneth Leung in Towards Data Science

Assumptions of Logistic Regression, Clearly Explained

Understand and implement assumption checks (in Python) for one of the most important data science modeling techniques

◆ · 8 min read · Oct 4, 2021

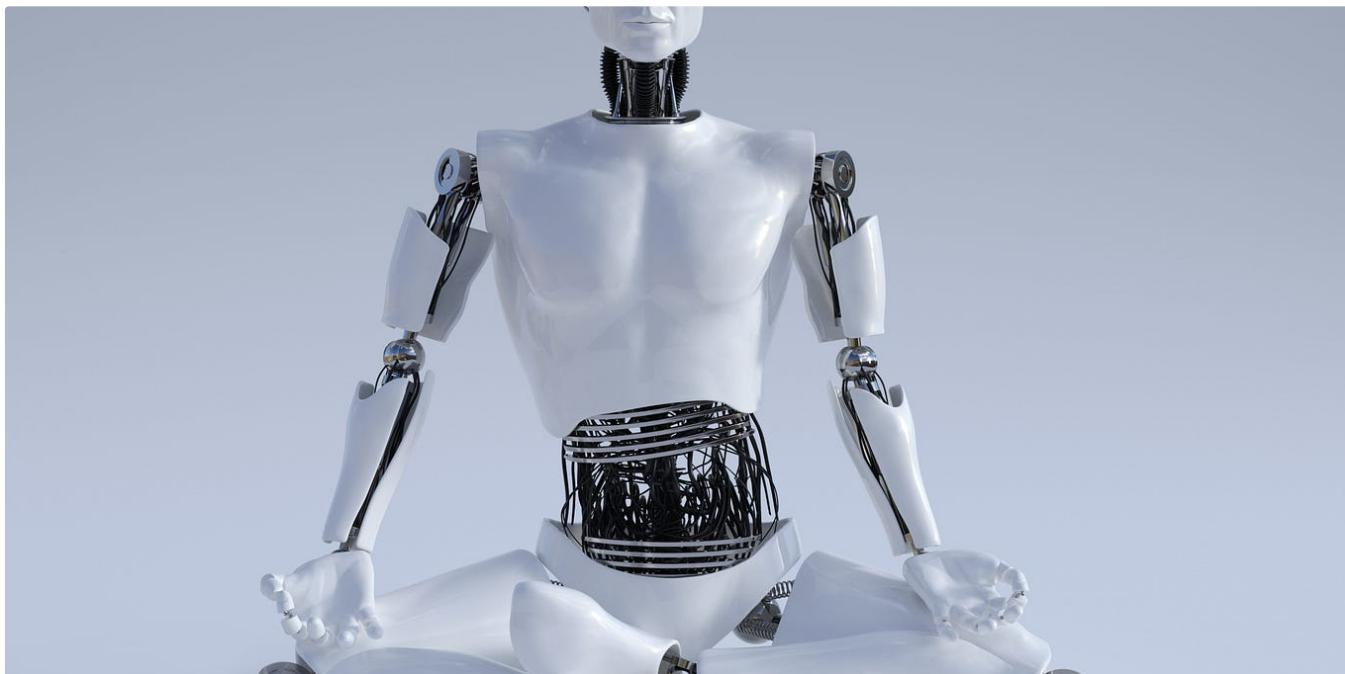
👏 316 🎧 5



See all from Kenneth Leung

See all from Towards Data Science

Recommended from Medium



 The PyCoach in Artificial Corner

You're Using ChatGPT Wrong! Here's How to Be Ahead of 99% of ChatGPT Users

Master ChatGPT by learning prompt engineering.

◆ · 7 min read · Mar 17

👏 17.6K 🎧 316





 Matt Chapman in Towards Data Science

The Portfolio that Got Me a Data Scientist Job

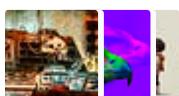
Spoiler alert: It was surprisingly easy (and free) to make

★ · 10 min read · Mar 24

 2.8K  44



Lists



What is ChatGPT?

9 stories · 13 saves



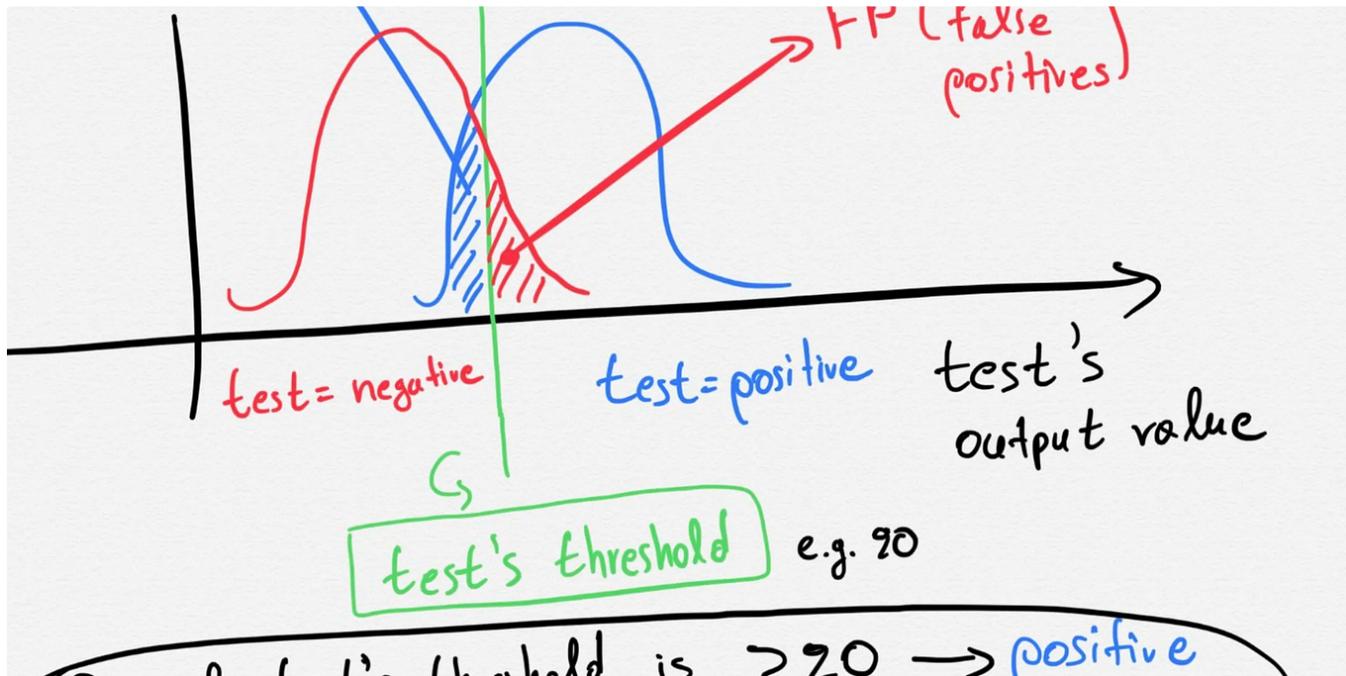
Stories to Help You Level-Up at Work

19 stories · 12 saves



Staff Picks

294 stories · 56 saves



Serafeim Loukas, PhD in Towards AI

How To Estimate FP, FN, TP, TN, TPR, TNR, FPR, FNR & Accuracy for Multi-Class Data in Python in 5...

In this post, I explain how someone can read a confusion matrix and how to extract several performance metrics for a multi-class...

★ · 6 min read · Feb 5

64



Samuel Flender in Towards Data Science

Class Imbalance in Machine Learning Problems: A Practical Guide

Five lessons from the trenches of applied data science

◆ · 8 min read · Oct 3, 2022

👏 300

💬 2



 Terence Shin

All Machine Learning Algorithms You Should Know for 2023

Intuitive explanations of the most popular machine learning models

◆ · 8 min read · Jan 6

👏 1.1K

💬 13





 Saupin Guillaume  in Towards Data Science

How Does XGBoost Handle Multiclass Classification?

É crucial entender o funcionamento subjacente da classificação usando esse tipo de modelo, pois isso afeta o desempenho.

◆ · 7 minutos de leitura · 7 de janeiro

 79



 +

Ver mais recomendações