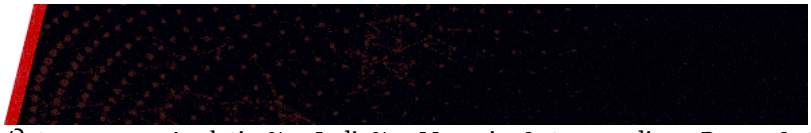

(https://praxis.ac.in/data-science-program/?utm_source=Analytics%20India%20Magazine&utm_medium=Banner&utm_campaign=DS-Mar2020)

DEVELOPERS CORNER (HTTPS://ANALYTICSINDIAMAG.COM/CATEGORY/LEARNING-CORNER/)

Beginners Guide To Understanding ROC Curve: How To Find The Perfect Probability Threshold

BY AMAL NAIR (HTTPS://ANALYTICSINDIAMAG.COM/AUTHOR/AMAL-NAIRANALYTICSINDIAMAG-COM/)
26/09/2019

In the Machine Learning spectrum, Binary Classification is one of the simplest and most popular problems that has been widely used across domains. From identifying fraudulent bank transactions to classifying or diagnosing diseases, Binary Classifiers have been in use since the inception of Machine Learning.

Many classification algorithms like Logistic Regressor uses probability to distribute samples into classes and in most cases the probability threshold defaults to 0.5. Which means that the algorithm classifies a sample as positive if the probability of that sample being positive is above 0.5(50%) and classifies a sample as negative if the probability of that sample being positive is less 0.5(50%)

This default threshold may not be enough when we consider a variety of problems that demand otherwise, say for example while diagnosing a disease it may be wise to choose a lower probability threshold to prevent any chance of the disease going misclassified. Thus the classification of critical data demands a more custom threshold which meets certain requirements. This is where the Receiver operating characteristic Curve shortly known as the ROC curve comes into play. It illustrates the diagnostic ability of a binary classifier.

(https://www.analytixlabs.co.in/)

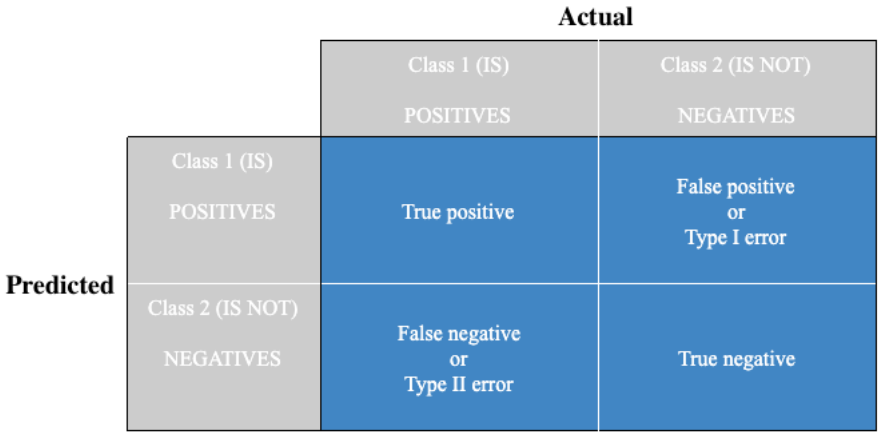
In layman's terms, the ROC curve visualises the effect of a chosen probability threshold on the classification efficiency. It helps analyse how the efficiency of Binary Classification changes with the values of Probability threshold.

The Confusion Matrix

The simplest way to analyse classification efficiency is by creating a Confusion Matrix with the predicted outcomes. The Confusion Matrix is also known as the Error Matrix as it determines the error rate in a classification prediction.

(https://www.jigsawacademy.com/online-analytics-training/?query=online&utm_source=AIM&utm_medium=in_article_banner&utm_campaign=low_arpu)

The following image represents a Confusion Matrix.



Consider a simple classification problem where the objective is to classify fraudulent and non-fraudulent transactions. Class 1 will be the positives which are the Fraudulent transactions and Class 2 will be the negatives or the non-fraudulent transactions.

Then,

- True Positive(TP) is the number of predicted fraudulent transactions that are actually fraudulent transactions.

- True Negative(TN) is the number of predicted non-fraudulent transactions that are actually non-fraudulent transactions.
- False Positive(FP) is the number of non-fraudulent transactions which were incorrectly classified as fraudulent transactions.
- False Negative(FN) is the number of fraudulent transactions that were incorrectly predicted or classified as non-fraudulent transactions.

Plotting The ROC Curve

The ROC curve can be plotted using the Confusion matrix as follows.

The ROC curve is plotted against the False Positive Rates and False Negative Rates of a classification problem for various probability thresholds.

Lets us understand each of these terms.

False Positive Rate

False Positive Rate (FPR) also called fall out is the ratio of negative samples which are incorrectly classified.

For example, if we consider the bank transaction example stated above, the False Positive Rate is the ratio of non-fraudulent transactions that were incorrectly classified as fraudulent transactions.

It can be calculated as follows:

SEE ALSO

(https://analyticsindiamag.com/8-ways-companies-can-incentivise-data-scientists-to-be-more-productive/)

8 Ways Companies Can Incentivise Data Scientists To Be More Productive (https://analyticsindiamag.com/8-ways-companies-can-incentivise-data-scientists-to-be-more-productive/)

$$FPR = \frac{FP}{Number\ Of\ Negatives} = \frac{FP}{FP + TN}$$

True Positive Rate

True Positive Rate(TPR) also called as sensitivity, recall and hit rate is the ratio of Positive samples which were correctly classified. It can be calculated in the following ways.

True Positive Rate gives the ratio of fraudulent transactions which were correctly classified as fraudulent by the classifier.

$$TPR = \frac{TP}{Number\ Of\ Positives} = \frac{TP}{TP + FN}$$

AMAL NAIR (HTTPS://ANALYTICSINDIAMAG.COM/AUTHOR/AMAL-NAIRANALYTICSINDIAMAG-COM/)

in(https://www.linkedin.com/in/amal-v-nair-66a23512a)
A Computer Science Engineer turned Data Scientist who is passionate about AI and all related technologies. Contact: amal.nair@analyticsindiamag.com

SHARE

TWEET (https://twitter.com/share?text=Beginners%20Guide%20To%20Understanding%20ROC%20Curve%3A%20How%20To

P(https://pinterest.com/pin/create/bookmarklet/?url=https://analyticsindiamag.com/beginners-guide-to-understanding-roc-curve-how-t

in(https://www.linkedin.com/cws/share?url=https://analyticsindiamag.com/beginners-guide-to-understanding-roc-curve-how-to-find-th

(https://wa.me/?text=Beginners%20Guide%20To%20Understanding%20ROC%20Curve%3A%20How%20To%20Find%20The%20Perfect%

(mailto:?
subject=Beginners%20Guide%20To%20Understanding%20ROC%20Curve%3A%20How%20To%20Find%20The%20Perfect%20Probabilit
guide-to-understanding-roc-curve-how-to-find-the-perfect-probability-threshold/)

(https://reddit.com/submit?url=https://analyticsindiamag.com/beginners-guide-to-understanding-roc-curve-how-to-find-the-perfect-

(https://t.me/share/url?&text=Beginners%20Guide%20To%20Understanding%20ROC%20Curve%3A%20How%20To%20Find%20The%2

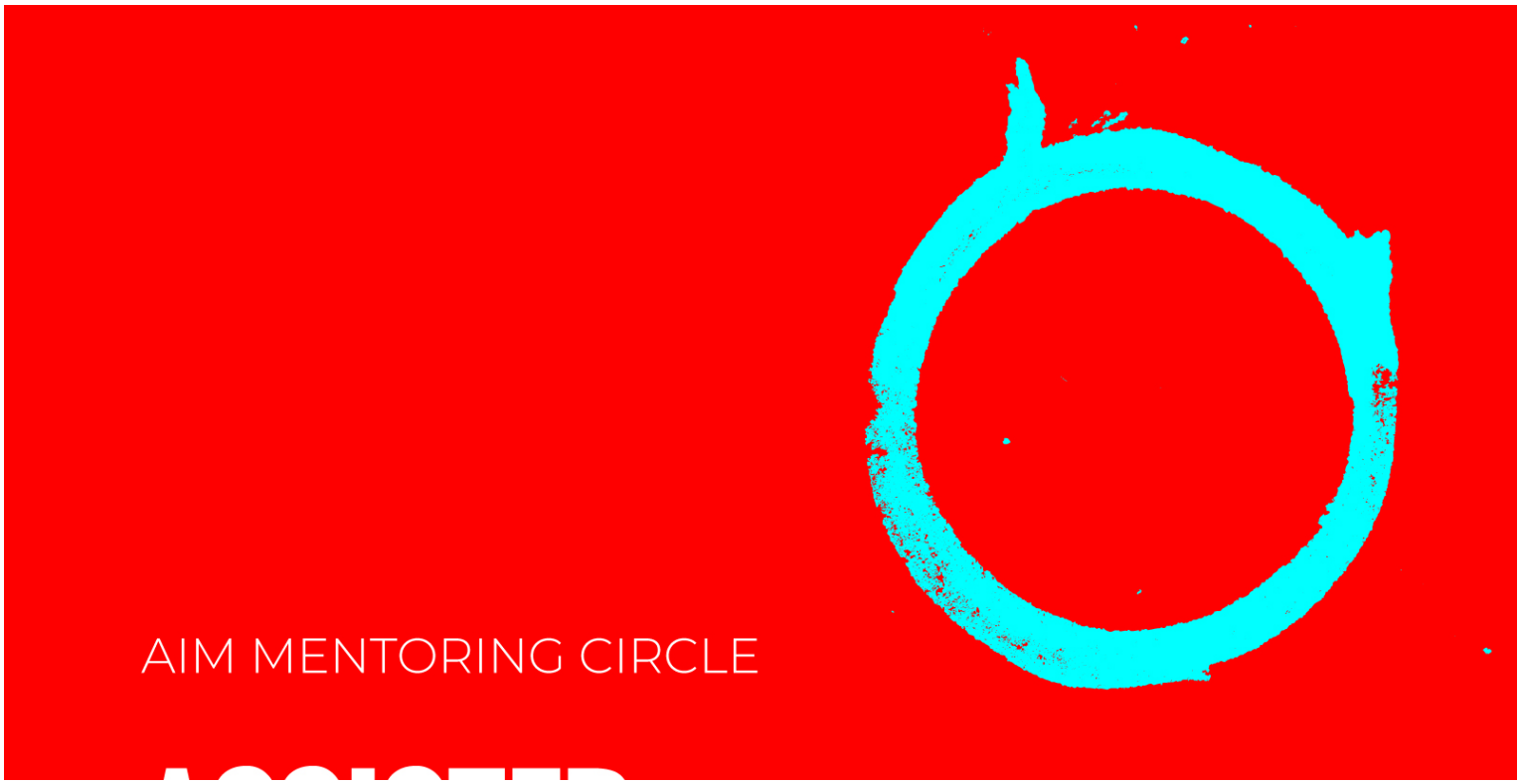
(https://www.adasci.org/cds/)

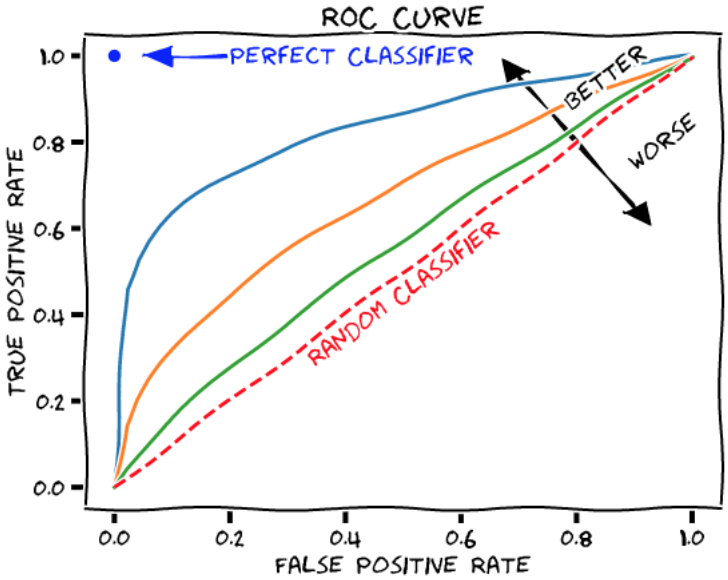
(https://business.louisville.edu/learnmore/msba-india/?utm_source=analyticsIndia&utm_medium=Banner&utm_term=India&utm_content=GetPaidIndia&utm_campaign=MSBAIndia)

OUR UPCOMING EVENTS

Webinar: AI-Powered Applications with AWS AI Service (https://register.gotowebinar.com/register/9205288776900109325) | 23rd Jul | Online

Computer Vision DevCon 2020 (https://cvdc.adasci.org/) | 13-14th Aug | Online





(https://commons.wikimedia.org/wiki/File:Roc-draft-xkcd-style.svg)

We plot the (TPR, FPR) for different values of the probability threshold to obtain the ROC curve.

- In the above figure, the red line indicates the points where the TPR and FPR are equal and any point on the line suggests that the ratio of correctly classified Positives is the same as the ratio of incorrectly classified Negatives which is only as good as a random classifier.
- Any point above the red line is better than a random classifier.
- Any point to the left of the red line indicates that the ratio of true positives are greater than that of false positives indicating a better probability threshold as the point moves towards left.
- The blue, orange and green lines represent ROC curves. Since the blue line is further away from the random classifier, it represents a better classifier in terms of efficiency.
- The blue dot represents a perfect classifier. At coordinate (0,1) the FPR is 0 and TPR is 1, meaning that all the positives are correctly classified and no negatives are incorrectly classified. Simply this means that there are no incorrect predictions at all.

Closing Note

The ROC Graph thus provides a summary of all the Confusion matrices generated with different probability thresholds and helps us choose the right probability based on the required/acceptable False Positives.

Provide your comments below

0 comments

0 Comments Sort by Oldest

Add a comment...

Facebook Comments Plugin

JOIN OUR TELEGRAM GROUP
Be part of an engaging community

(https://t.me/joinchat/NJLxnhZB7GkX3CPvjs9QGQ)

(https://analyticsindiamag.com/mentorship-circle/assisted-mentoring/)

CVDC 2020

VIRTUAL

COMPUTER VISION DEVCON 2020

13-14TH AUG 2020

ADaSci | THE ASSOCIATION OF DATA SCIENTISTS

(https://cvdc.adasci.org/)

More than 1,00,000 people are subscribed to our newsletter

Subscribe now to receive in-depth stories on AI & Machine Learning.

ENTER YOUR EMAIL

SUBSCRIBE NOW

ABOUT US(HTTPS://ANALYTICSINDIAMAG.COM/ABOUT/)

ADVERTISE(HTTPS://ANALYTICSINDIAMAG.COM/ADVERTISE-WITH-US/)

GET INVOLVED(HTTPS://ANALYTICSINDIAMAG.COM/GET-INVOLVED/)

COPYRIGHT(HTTPS://ANALYTICSINDIAMAG.COM/COPYRIGHT-TRADEMARKS/)

(https://analyticsindiamag.com/)

(HTTPS://FACEBOOK.COM/ANALYTICSINDIAMAGAZINE)

(HTTPS://TWITTER.COM/ANALYTICSINDIAM)

(HTTPS://INSTAGRAM.COM/ANALYTICSINDIAMAGAZINE)

(HTTPS://PINTEREST.COM/ANALYTICSINDIAM)

PRIVACY(HTTPS://ANALYTICSINDIAMAG.COM/PRIVACY-POLICY/)

TERMS OF USE(HTTPS://ANALYTICSINDIAMAG.COM/TERMS-USE/)

CONTACT US(HTTPS://ANALYTICSINDIAMAG.COM/CONTACT-US/)

(HTTPS://YOUTUBE.COM/CHANNEL/UCALWRSGEJAVG1VW9QSFOUMA)

(HTTPS://MEDIUM.COM/ANALYTICS-INDIA-MAGAZINE)

COPYRIGHT ANALYTICS INDIA MAGAZINE PVT LTD

/