# An Introduction to the ROC-AUC in Classification Tasks

## What does the curve mean?

Yong Cui Sep 16, 2020·

Source: <https://towardsdatascience.com/an-introduction-to-the-roc-auc-in-classification-tasks-94c2a147dd04>

Desenho de uma pessoa

Descrição gerada automaticamente com confiança baixa

Photo by Francesco Ungaro on Unsplash

In machine learning, one essential learning objective is to classify data into groups. Although classification can include unsupervised learning (e.g., clustering), in most cases, our tasks involve known labels and thus we’re conducting supervised learning to classify our data. Your model will generate predicted labels, which will allow us to compare whether our prediction is accurate or not. When the predicted label and the true label match, we say the prediction is correct, and apparently when they don’t, we say the prediction is wrong.

To put our discussion into context, suppose that we have clinical data for some subjects whose diagnoses on diabetes are known. Just a quick disclaimer before we proceed — these data are not real data and they don’t constitute any medical advice.

Based on these data, we build a logistic regression model to predict whether people have diabetes or not based on their fasting glucose level. In the table below, the fasting glucose level is expressed as mg/dL. The diabetic\_clinical column shows you the clinical diagnosis while the diabetic\_predicted shows you the prediction result from the logistic regression model. Based on the clinical and predicted results, we can simply know whether a prediction is correct or not, as indicated by the last column.

Table 1. Fasting Glucose Level and Diabetes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| subject | fasting\_glucose\_level | diabetic\_clinical | diabetic\_predicted | prediction\_correct |
| 1 | 82 | N | N | correct |
| 2 | 93 | N | N | correct |
| 3 | 102 | N | N | correct |
| 4 | 123 | N | Y | wrong |
| 5 | 127 | N | Y | wrong |
| 6 | 118 | Y | N | wrong |
| 7 | 137 | Y | Y | correct |
| 8 | 144 | Y | Y | correct |
| 9 | 151 | Y | Y | correct |
| 10 | 162 | Y | Y | correct |

## Confusion Matrix

Based on this binary evaluation outcome (correct vs. wrong) in relation to the true labels, we can build the 2 x 2 confusion matrix, as shown in Figure 1.

Figure 1. Confusion Matrix.

Uma imagem contendo Diagrama

Descrição gerada automaticamente

by Author.

## Correct Predictions:

* **True Positive (TP)**: both predicted and true labels are positive. In the example shown above, those people have diabetes and they’re predicted to be diabetic using their fasting glucose level.
* **True Negative (TN)**: both predicted and true labels are negative. Those are non-diabetic subjects, who are predicted to be non-diabetic.

## Wrong Predictions:

* **False Positive (FP)**: the predicted label is positive, while the true label is negative. Those are predicted to be diabetic, but they’re actually not.
* **False Negative (FN)**: the predicted label is negative, while the true label is positive. Those are diabetic subjects who’re predicted to be non-diabetic.

We can derive many metrics essential for classification model evaluation. Some commonly used ones are listed below, and you can find a full list of generated metrics at the [Wikipédia](https://en.wikipedia.org/wiki/Confusion_matrix) page.

* **Accuracy**: the number of correct predictions divided by the total number of predictions: ***(TP + TN) / (TP + TN + FP + FN)***. In the example, the accuracy of our model is 0.7 (i.e., 7 / 10).
* **True Positive Rate (TPR, Sensitivity or Recall)**: the number of true positive labels divided by the number of positive labels: ***TP / (TP + FN)***. The TPR of our model is 0.8 (i.e., 4 / 5).
* **False Positive Rate (FPR, Fall-out)**: the number of false positive labels divided by the number of negative labels: ***FP / (FP + TN)***. The FPR of our model is 0.4 (i.e., 2 / 5).
* **True Negative Rate (Specificity)**: the number of true negative labels divided by the number of negative labels: ***TN / (FP + TN)****.* You can notice that ***specificity = 1 — FPR***. The specificity of our model is 0.6 (i.e., 3 / 5).

## Receiver Operating Characteristic (ROC)

The Receiver Operating Characteristic curve is a graph showing you how your classification model performs at all thresholds. The following graph is a hypothetical ROC curve.

Figure 2. ROC Example Graph.

Diagrama

Descrição gerada automaticamente

by Author.

* In an ROC curve graph, the x-axis is the FPR, while the y-axis is the TPR.
* For a perfect model, its FPR is 0 and its TPR is 1. By contrast, for the worst model, its FPR is 1 and its TPR is 0. The ROC curves for these two extreme scenarios are shown in the graph.
* In a typical model, we should see real curves. **Specifically, by varying the thresholds, our model will produce different TPR and FPR, and these points can be plotted on this graph. Connecting these points, we can generate an ROC curve**.

To put the ROC curve into the context of the diabetes diagnosis example, let’s suppose that we can have different thresholds for predicting the diabetes diagnosis. As you can expect, if we have an extreme low threshold, we will classify all subjects as diabetic. Although we can get a TPR of 1, but the FPR will become 1 too. Considering some moderate thresholds, we should be able to find many different combinations of TPR and FPR.

Because the data shown are small is size, if you have many more data points, by varying the threshold, you should create more points as a function of TPR and FPR. If we connect all of these points and smooth the curve, we’re getting the ROC curve for the particular model that we’re building.

## Area Under ROC curve (AUC)

How can we quantify the performance of our model? As discussed above, we say that the ROC is to show our model performs, but how can we exactly evaluate the performance with the ROC curves?

If you compare the two extreme scenarios (perfect vs. worst), you’ll probably notice that the areas under the curve seems to mean something. Your guess is exactly right. Suppose you consider a typical model, in most cases, your model’s TPR should be larger than the FPR at various thresholds. In this case, you’ll see your ROC curve above the diagonal line of the graph.

More importantly, if your model is better, you should see greater differences between the TPR and the FPR, which drives the curve towards the perfect model situation. The following graph shows you some possible scenarios that you may encounter with realistic models.

Figure 3. AUC for ROC curves.

Gráfico

Descrição gerada automaticamente com confiança média

by Author.

* The gray area depicts the area under the ROC curve.
* The AUC of Model 2 is greater than the AUC of Model 1. **We typically say that Model 2 outperforms Model 1.** In other words, we can roughly equate the performance of classifying models to their respective AUC.
* The diagonal line depicts the the random classifier for a binary classification task. Here’s [additional discussion](https://datascience.stackexchange.com/questions/31872/auc-roc-of-a-random-classifier) regarding why the AUC for such random guessing is 0.5 for both TPR and FPR.

## Before You Go

Although ROC and AUC are often used to evaluate classification task performance, they’re not always preferred. The major reason is that it doesn’t consider the actual question under examination. For instance, different models can have similar AUC, but they can have different shapes, which means that the models have different combinations of TPR and FPR.

Thus, we should consider other factors to pick the desired model. Questions to ask yourself can include. Do you care more about correctly identifying positive cases? Or do you care more about correctly identifying negative cases?

For your reference, here’s [a brief discussion](https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc) on the ROC and AUC, which I find very useful for beginners.