

Confusion Matrix, Accuracy, Precision, Recall, F1 Score

 medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd

Harikrishnan N B

1 June 2020

Top highlight

Binary Classification Metric



Harikrishnan N B

How to evaluate the performance of a machine learning model?

Let us consider a task to classify whether a person is **pregnant** or **not pregnant**. If the test for pregnancy is positive (+ve), then the person is pregnant. On the other hand, if the test for pregnancy is negative (-ve) then the person is not pregnant.

Now consider the above classification (pregnant or not pregnant) carried out by a machine learning algorithm. The output of the machine learning algorithm can be mapped to one of the following categories.

1. A person who is actually pregnant (positive) and classified as pregnant (positive).
This is called



Figure 1: True Positive.

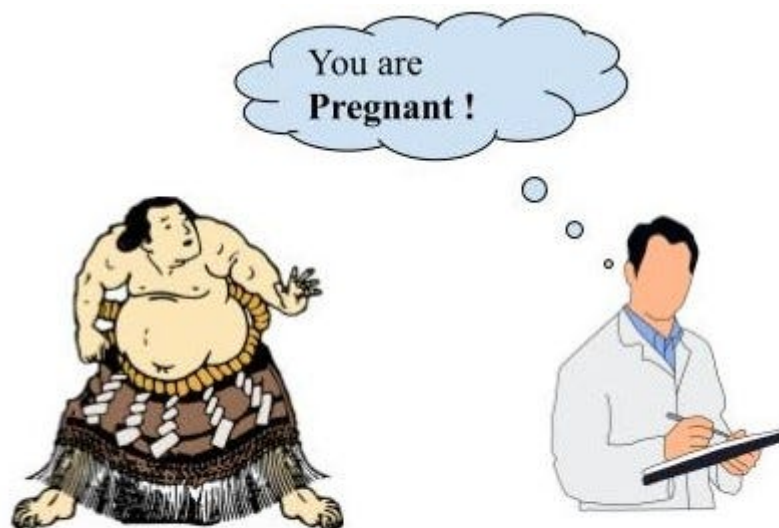
2. A person who is actually not pregnant (negative) and classified as not pregnant (negative). This is called **TRUE NEGATIVE ()**.



TRUE NEGATIVE

Figure 2: True Negative.

3. A person who is actually not pregnant (negative) and classified as pregnant (positive). This is called **FALSE POSITIVE** ().



FALSE POSITIVE

Figure 3: False Positive.

4. A person who is actually pregnant (positive) and classified as not pregnant (negative). This is called **FALSE NEGATIVE** ().



Figure 4. False Negative.

What we desire is **TRUE POSITIVE** and **TRUE NEGATIVE** but due to the misclassifications, we may also end up in **FALSE POSITIVE** and **FALSE NEGATIVE**. So there is a confusion in classifying whether a person is pregnant or not. This is because no machine learning algorithm is perfect. Soon we will describe this confusion in classifying the data in a matrix called confusion matrix.

Now, we select 100 people which includes pregnant women, not pregnant women and men with fat belly. Let us assume out of this 100 people 40 are pregnant and the remaining 60 people include not pregnant women and men with fat belly. We now use a machine learning algorithm to predict the outcome. The predicted outcome (pregnancy +ve or -ve) using a machine learning algorithm is termed as the and the true outcome (in this case which we know from doctor's/expert's record) is termed as the .

Now we will introduce the which is required to compute the of the machine learning algorithm in classifying the data into its corresponding labels.

Confusion matrix C is a square matrix where C_{ij} represents the number of data instances which are known to be in group i (true label) and predicted to be in group j (predicted label).

If we consider a binary classification problem,
 C_{00} represents the count of true negative
 C_{01} represents the count of false positive
 C_{10} represents the count of false negative and
 C_{11} represents the count of true positive.

The following diagram illustrates the confusion matrix for a binary classification problem.

		PREDICTED LABEL	
		NEGATIVE	POSITIVE
TRUE LABEL	NEGATIVE	TRUE NEGATIVE	FALSE POSITIVE
	POSITIVE	FALSE NEGATIVE	TRUE POSITIVE


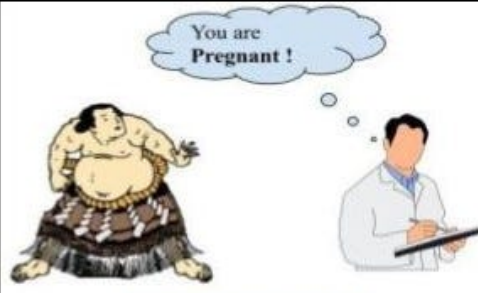
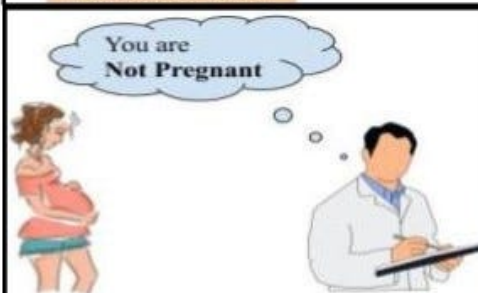

		PREDICTED LABEL	
		NEGATIVE	POSITIVE
TRUE LABEL	NEGATIVE	 TRUE NEGATIVE	 FALSE POSITIVE
	POSITIVE	 FALSE NEGATIVE	 TRUE POSITIVE

Figure 5: Confusion Matrix.

We will now go back to the earlier example of classifying 100 people (which includes 40 pregnant women and the remaining 60 are not pregnant women and men with a fat belly) as pregnant or not pregnant. Out of 40 pregnant women 30 pregnant women are classified correctly and the remaining 10 pregnant women are classified as not pregnant

by the machine learning algorithm. On the other hand, out of 60 people in the not pregnant category, 55 are classified as not pregnant and the remaining 5 are classified as pregnant.

In this case, $TP = 30$, $FP = 5$, $TN = 55$, $FN = 10$. The confusion matrix is as follows.

		PREDICTED LABEL	
		NEGATIVE	POSITIVE
TRUE LABEL	NEGATIVE	55 TRUE NEGATIVE	5 FALSE POSITIVE
	POSITIVE	10 FALSE NEGATIVE	30 TRUE POSITIVE

Figure 6: Confusion matrix for the pregnant vs not pregnant classification.

What is the accuracy of the machine learning model for this classification task?

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

Accuracy represents the number of correctly classified data instances over the total number of data instances.

In this example, $Accuracy = (55 + 30)/(55 + 5 + 30 + 10) = 0.85$ and in percentage the accuracy will be 85%.

Is accuracy the best measure?

Accuracy may not be a good measure if the dataset is not balanced (both negative and positive classes have different number of data instances). We will explain this with an example.

Consider the following scenario: There are 90 people who are healthy (negative) and 10 people who have some disease (positive). Now let's say our machine learning model perfectly classified the 90 people as healthy but it also classified the unhealthy people as healthy. What will happen in this scenario? Let us see the confusion matrix and find out the accuracy?

In this example, $TP = 90$, $FP = 0$, $FN = 10$ and $TP = 0$. The confusion matrix is as follows.

		PREDICTED LABEL	
		NEGATIVE	POSITIVE
TRUE LABEL	NEGATIVE	90 TRUE NEGATIVE	0 FALSE POSITIVE
	POSITIVE	10 FALSE NEGATIVE	0 TRUE POSITIVE

Figure 7: Confusion matrix for healthy vs unhealthy people classification task.

Accuracy in this case will be $(90 + 0)/(100) = 0.9$ and in percentage the accuracy is 90 %.

Is there anything fishy?

The accuracy, in this case, is 90 % but this model is very poor because all the 10 people who are unhealthy are classified as healthy. By this example what we are trying to say is that **accuracy is not a good metric when the data set is unbalanced**. Using accuracy in such scenarios can result in misleading interpretation of results.

So now we move further to find out another metric for classification. Again we go back to the pregnancy classification example.

Now we will find the **precision (positive predictive value)** in classifying the data instances. Precision is defined as follows:

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

What does precision mean?

should ideally be 1 (high) for a good classifier. becomes 1 only when the numerator and denominator are equal i.e , this also means is zero. As increases the value of denominator becomes greater than the numerator and value decreases (which we don't want).

So in the pregnancy example, $= 30/(30+ 5) = 0.857$

Now we will introduce another important metric called . is also known as or and is defined as follows:

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

should ideally be 1 (high) for a good classifier. becomes 1 only when the numerator and denominator are equal i.e , this also means is zero. As increases the value of denominator becomes greater than the numerator and value decreases (which we don't want).

So in the pregnancy example let us see what will be the recall.

$= 30/(30+ 10) = 0.75$

So ideally in a good classifier, we want both and to be one which also means and are zero. Therefore we need a metric that takes into account both and . is a metric which takes into account both and and is defined as follows:

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

becomes 1 only when and are both 1. becomes high only when both and are high. is the harmonic mean of and and is a better measure than .

In the pregnancy example, $= 2 * (0.857 * 0.75) / (0.857 + 0.75) = 0.799$.

Reading List

The following is an interesting article on the common binary classification metric by neptune.ai. The link to the article is available here: <https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc>