Understanding Metrics to Assess the Quality of Classifiers

The quality of any classifier is measured in terms of True Positive, False Positive, True Negative and False Negative, Precision, Recall and F-Measure.

To understand this concept, let us first consider the story of boy who shouted 'Wolf' to fool the villagers. The story goes as follows.

The boy who cried wolf

Once a boy was getting bored and thought of making fools out of fellow villagers. To have some fun, he shouted out, 'Wolf!' even though no wolf was in sight. The villagers ran to rescue him, but then got angry when they realized that the boy was playing a joke on them. The boy repeated the same prank a number of times and each time the villagers rushed out. They got angrier when they found he was joking.

One night, the boy saw a real wolf approaching and shouted 'Wolf!'. This time villagers stayed in their houses and the hungry wolf turned the flock into lamb chops.

Let's make the following definitions: Here, 'Wolf' is a **positive class** and 'No wolf' is a **negative class**. We can summarize our 'wolf-prediction' model using a 2x2 confusion matrix that depicts all four possible outcomes as shown below.

True Positive (TP): Reality: A wolf threatened. Boy said: 'Wolf.' Outcome: Boy is a hero.	**False Positive (FP):** Reality: No wolf threatened. Boy said: 'Wolf.' Outcome: Villagers are angry at Boy for waking them up.
False Negative (FN): Reality: A wolf threatened. Boy said: 'No wolf.' Outcome: The wolf ate all the flock.	**True Negative (TN):** Reality: No wolf threatened. Boy said: 'No wolf.' Outcome: Everyone is fine.

Now consider a classifier, whose task is to predict whether the image is of a bird or not. In this case, let us assume, 'Bird' is a positive class and 'Not a Bird' is a negative class.

Let us suppose we have a dataset of 15,000 images having 6000 images of birds and 9000 images of anything that is not a bird. The matrix illustrating actual vs. predicted results also known as confusion matrix is given in Figure 5.46 below.

Results for 15,000 Validation Images
(6000 images are birds, 9000 images are not birds)

	Predicted 'bird'	Predicted 'not a bird'
Bird	5,450 *True Positives*	550 *False Negatives*
Not a Bird	162 *False Positives*	8,838 *True Negatives*

Figure 5.46   Confusion matrix for bird classifier

5.7.1 True positive

Those instances where predicted class is equal to the actual class are called as true positive or a true positive is an outcome where the model correctly predicts the positive class.

For example in the case of our bird classifier, the birds that are correctly identified as birds are called true positive.

5.7.2 True negative

Those instances where predicted class and actual class are both negative are called as true negative or a true negative is an outcome where the model correctly predicts the negative class.

For example, in the case of our bird classifier there are images that are not of birds which our classifier correctly identified as 'not a bird' are called true negatives.

5.7.3 False positive

Those instances where predicted class/answer is positive, but actually the instances are negative or a false positive is an outcome where the model incorrectly predicts the positive class.

For example, in case of our bird classifier there are some images that the classifier predicted as birds but they were something else. These are our false positives.

5.7.4 False negative

Those instances where predicted class is negative, but actually the instances are positive or a false negative is an outcome where the model incorrectly predicts the negative class.

For example, in case of our bird classifier there are some images of birds that the classifier did not correctly recognize as birds. These are our false negatives.

In simple words, predicted 'bird' column is considered as Positive and if the prediction is correct then cell is labeled as true positive, otherwise it is false positive. The column where prediction is 'not a bird' is considered as negative and if prediction is correct, the cell is labeled as true negative otherwise it is false negative as shown in Figure 5.46.

5.7.5 Confusion matrix

Confusion matrix is an $N \times N$ table that summarizes the accuracy of a classification model's predictions. Here, N represents the number of classes. In a binary classification problem, $N = 2$. In simple words, it is a correlation between the actual labels and the model's predicted labels. One axis of a confusion matrix is the label that the model predicted, and the other axis is the actual label.

For example, consider a sample confusion matrix for a binary classification problem to predict if a patient has a tumor or not.

	Tumor (predicted)	*No-Tumor (predicted)*
Tumor (actual)	18 (TP)	1 (FN)
No-Tumor (actual)	*6 (FP)*	*452 (TN)*

Figure 5.47   Confusion matrix for tumor prediction

The confusion matrix shown in Figure 5.47 has 19 samples that actually had tumor, the model correctly classified 18 as having tumor (18 true positives), and incorrectly classified 1 as not having a tumor (1 false negative). Similarly, of 458 samples that actually did not have tumor, 452 were correctly classified (452 true negatives) and 6 were incorrectly classified (6 false positives).

Confusion matrices contain sufficient information to calculate a variety of performance metrics, including precision, recall and F-Measure. Instead of just looking at overall accuracy, we usually calculate precision, recall and F-Measure. These metrics give us a clear picture of how well the classifier performed.

5.7.6 Precision

Precision identifies the frequency with which a model was correct when predicting the positive class. It is defined as:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

Or we can say that Precision = True Positives/All Positive predicted. In other words, if we predict positive then how often was it really a positive instance.

For example, in case of our bird classifier, it refers to the situation where a bird was predicted, but how often was it really a bird.

For classifier shown in Figure 5.46, the Precision = 5450/(5450 + 162) =5450/5612 =0.9711. In terms of percentage it is 97.11%.

5.7.7 Recall

Recall identifies out of all the possible positive labels, how many did the model correctly identify? In simple word, it refers to what percentage of actual positive instances we are able to find.

For example, In case of our bird classifier, it refers what percentage of the actual birds did we find?

So, $$\text{Recall} = \frac{\text{True Positive}}{\text{All actual positive instances}} = \frac{\text{True Positive}}{\text{True Positive + False Negatives}}$$

For classifier shown in Figure 5.46, the Recall = 5450/(5450 + 550) = 5450/6000 = 0.9083 and in terms of percentage it is 90.83%.

The table below shows both precision and recall for the bird classifier discussed above.

Precision *If we predicted 'bird', how often was it really a bird?*	**97.11%** *(True positives ÷ All Positive Guesses)*
Recall *What percentage of the actual birds did we find?*	**90.83%** *(True positives ÷ Total Bird in Dataset)*

In our example of bird classifier, this tells us that 97.11% of the time the prediction was right! But it also tells us that we only found 90.83% of the actual birds in the dataset. In other words, we might not find every bird but we are pretty sure about it when we do find one!

Now, let us consider a multi class classifier, where a classifier has to predict one class out of three classes. Here, we have three classes A, B and C.

		Predicted Class		
		A	B	C
Actual Class	A	25	3	2
	B	2	34	5
	C	2	3	20

Let's calculate Precision and Recall for each predicted class.

For Class A

Precision = True Positives/All Positive Predicted

Precision = 25/29

Recall = True Positive/All actual positive instances = 25/30

For Class B

Precision = True Positives/All Positive Predicted

Precision = 34/40

Recall = True Positive/All actual positive instances =34/41

For Class C

Precision = True Positives/All Positive Predicted

Precision = 20/27

Recall = True Positive/All actual positive instances =20/25

5.7.8 F-Measure

In statistical analysis of binary classification, the F-Measure (also known as F-score or F-1 score) is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score. Here, p is the number of correct positive results divided by the number of all positive results returned by the classifier, and r is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive). The F1 score is the harmonic average of the precision and the recall, where, an F1 score reaches its best value at 1 (perfect precision and recall), and its worst at 0.

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Remind Me

◆ Classification and regression are two main techniques generally used for data analysis. Classification predicts categorical values, while regression models predict continuous values.

◆ Classification is further defined as two types that are posteriori classification and priori classification. Posterior classification is supervised machine learning approach, where classes are already given. On

the other hand, priori classification is unsupervised machine learning approach, where classes are notgiven.

- Information gain measures the quality of an attribute for predicting the class of each sample of thetraining data. The attribute with the highest information gain is selected as the next split
- attribute. Gini Index was developed by Italian scientist Corrado Gini in 1912. It is basically used to represent level of equality or inequality among objects and it can also be used to develop the decision tree. It alwaysranges between 0 and 1.
- Bayes theorem was postulated by Thomas Bayes and the classification system using Bayes theorem isbased on a hypothesis that the given data belongs to a particular class. In this theorem, probability iscalculated for the hypothesis to be true.