Precision vs Recall: Understanding the Trade-off in Classification Models

When building a classification model, achieving high accuracy in distinguishing between classes is crucial. However, relying solely on accuracy as a performance metric can be misleading. In this blog post, we will explore the concepts of precision and recall, and why they are essential in evaluating model performance. We will also delve into the trade-off between precision and recall, and how it impacts decision-making in classification models.

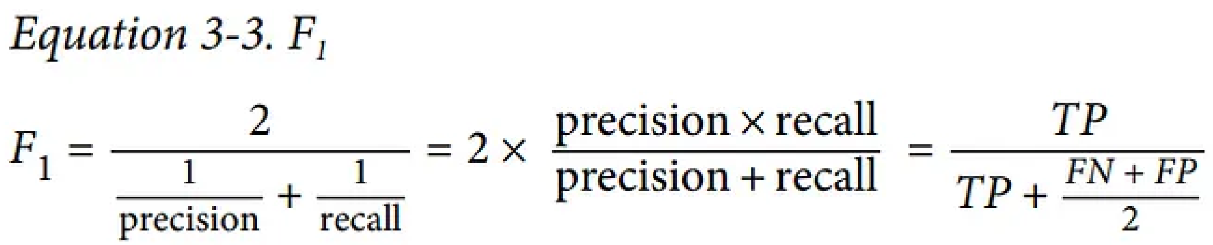
Precision and Recall

Consider the famous “hello world” of machine learning, the MNIST dataset, for simplicity. Suppose we have a binary classification task of differentiating between the digit 5 and non-5 digits. If we build a model using an algorithm like SGDClassifier and achieve an accuracy of 90%, it may seem impressive at first glance. However, upon closer inspection, we discover that approximately 90% of the images in the dataset are non-5 digits. This means that even a naive classifier that outputs “Not 5” for every instance would still achieve 90% accuracy. Hence, relying solely on accuracy is not reasonable, and alternative measures such as precision, recall, and F1 score come into play.

To assess the performance of our model, we calculate precision and recall. In our case, the model outputs a precision of 72.9% and a recall of 75.6%. This means that when the model claims an image represents a 5, it is correct only 72.9% of the time, and it detects only 75.6% of the actual 5 digits. These metrics provide a more comprehensive understanding of the model's behavior and its ability to classify correctly.

F1 Score

To simplify the comparison between classifiers, precision and recall can be combined into a single metric known as the F1 score. The F1 score represents the harmonic mean of precision and recall. Unlike the regular mean, which treats all values equally, the harmonic mean gives more weight to low values. As a result, classifier will achieve a high F1 score only if both precision and recall are high. F1 score favors classifiers that have balanced precision and recall, highlighting the need to find an optimal trade-off.



Equation 3-3. F_1

$$F_1 = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{TP}{TP + \frac{FN + FP}{2}}$$

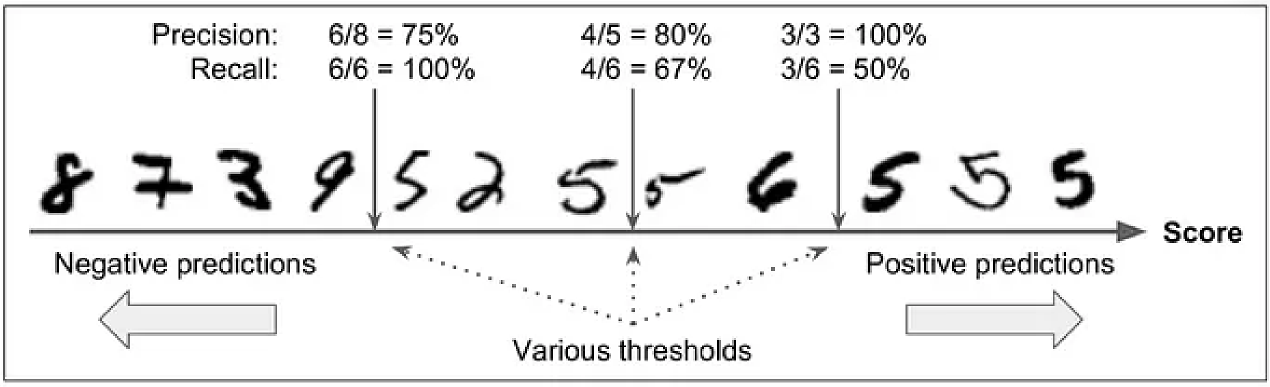
F1 Score.

Precision vs Recall: The Trade-off

Increasing precision often comes at the cost of reducing recall, and vice versa. This trade-off can be understood by examining how the SGDClassifier makes classification decisions. For each instance, the model computes a score based on a decision function. If the score exceeds a threshold, the instance is assigned to the positive class; otherwise, it is assigned to the negative class. Adjusting the threshold impacts precision and recall differently.

A Simple Illustration

To understand this trade-off, let's look at how the SGDClassifier makes its classification decisions. For each instance, it computes a score based on a decision function. If that score is greater than a threshold, it assigns the instance to the positive class; otherwise it assigns it to the negative class. Below figure shows a few digits positioned from the lowest score on the left to the highest score on the right. Suppose the decision threshold is positioned at the central arrow (between the two 5s): you will find 4 true positives (actual 5s) on the right of that threshold, and 1 false positive (actually a 6). Therefore, with that threshold, the precision is 80% (4 out of 5). But out of 6 actual 5s, the classifier only detects 4, so the recall is 67% (4 out of 6). If you raise the threshold (move it to the arrow on the right), the false positive (the 6) becomes a true negative, thereby increasing the precision (up to 100% in this case), but one true positive becomes a false negative, decreasing recall down to 50%. Conversely, lowering the threshold increases recall and reduces precision.



In this precision/recall trade-off, images are ranked by their classifier score, and those above the chosen decision threshold are considered positive; the higher the threshold, the lower the recall, but higher the precision. (Image credits: Hands on Machine Learning by Geron Aurelien, page 94).