On the importance of using precision and recall

It may often seem unclear why we need both precision and recall as metrics, and the F1 score that computes the harmonic mean of both.

To explain the importance of using both these metrics, it's best to use examples where our classification model makes extreme predictions.

Precision and recall

In multiclass problems, we will compute precision and recall for each class. So we'll take one class at a time.

Working on the evaluation of a specific class A:

- Precision measures, out of the items that we predicted as A, how many of them were correct (i.e. actually A).
- Recall measures the coverage instead. Out of all those items whose actual label is A, how many of them have we predicted as A, i.e. how many of the A items have we identified as such.

Working example

For our example, we will assume we have a dataset where the test set contains 400 items. These items are equally distributed across 4 classes:

- class A: 100 items
- class B: 100 items
- class C: 100 items
- class D: 100 items

Example 1: Model with extremely good precision

Let's say that we have a model that, for the 400 items in test set above, predicts 1 item as belonging to class A. The remainder 399 items are predicted as B, C or D (we don't mind the specific predictions for this example).

The 1 item that our model predicted as being class A is actually correct. Hence:

- Precision (class A) = 1.0 (i.e. of the items predicted as A, they were all correct)
- Recall (class A) = 1 / 100 = 0.01 (i.e. of all the items that should have been predicted as A, we only managed to identify 1)

This extreme example shows how you can get perfect precision, but terrible recall. This is not a good model, because we missed 99 items that should have been predicted as A. However, if we only used precision as the metric, we would have concluded that our model was perfect.

Example 2: Model with extremely good recall

In this case we have a model that predicts all 400 items in the test set as class A. No predictions whatsoever for classes B, C and D.

When we are evaluating performance on class A:

- Precision (class A) = 100 / 400 = 0.25 (i.e. of the items predicted as A, we only got 25% of them right)
- Recall (class A) = 1.0 (i.e. of the 100 items whose label is actually A, we managed to identify all of them)

This is another extreme, yet opposite, example, where we get perfect recall, but poor precision. While the recall suggests that this model is great, we have actually classified 300 items incorrectly as A. If we only used recall as the metric here, our evaluation would be misleading, suggesting that we have a perfect outcome, when we don't.

Conclusion

These two examples show us that precision and recall provide complementary information. Ideally we want to optimise performance on both metrics, as a good score on one metric but bad score on the other suggests it's not a good model.

Because we often want to have a single value to easily compare models between them, the F1 score as the harmonic mean provides a convenient way of combining both precision and recall into a single metric, hence allowing us to optimise for both of them.