

## How good is what?

### Agile Scientific

Geology is a descriptive science, which is to say, geologists are label-makers. We record observations by assigning labels to data. Labels can either be numbers or they can be words. As such, of the numerous tasks that machine learning is fit for attacking, **supervised classification** problems are perhaps the most accessible – the most intuitive – for geoscientists. Take data that already has labels. Build a model that learns the relationships between the data and labels. Use that model to make labels for new data. The concept is the same whether a geologist or an algorithm is doing it, and in both cases we want to test how well our classifier is at doing its label-making.

Say we have a classifier that will tell us whether a given combination of rock properties is either a dolomite (purple) or a sandstone (orange). Our classifier could be a person named Sally, who has seen a lot of rocks, or it could be a statistical model trained on a lot of rocks (e.g. this one on the right). For the sake of illustration, say we only have two tools to measure our rocks – that will make visualizing things easier. Maybe we have the gamma-ray tool that measures natural radioactivity, and the density tool that measures bulk density. Give these two measurements to our classifier, and they return to you a label.

## How good is my classifier?

Once you've trained your classifier – you've done the machine learning and all that – you've got yourself an automatic label maker. But that's not even the best part. The best part is that we get to analyze our system and get a handle on how good we can expect our predictions to be. We do this by seeing if the classifier returns the correct labels for samples *that it has never seen before*, using a dataset for which we know the labels. This dataset is called validation data.

Using the validation data, we can generate a suite of statistical scores to tell us unambiguously how this particular classifier is performing. In scikit-learn, this information compiled into a so-called classification report, and it's available to you with a few simple lines of code. It's a window into the behaviour of the classifier that warrants deeper inquiry.

To describe various elements in a classification report, it will be helpful to refer to some validation data:

Our Two-class Classifier (left) has not seen the Validation Data (middle). We can calculate a classification report by Analyzing the intersection of the two (right).

## Accuracy is not enough

When people straight up ask about a model's accuracy, it could be that they aren't thinking deeply enough about the performance of the classifier. Accuracy is a measure of the entire classifier. It tells us nothing about how well we are doing with one class compared to another, but there are other metrics that tell us this:

**Support** — how many instances there were of that label in the validation set.

**Precision** — the fraction of correct predictions for a given label. Also known as *positive predictive value*.

**Recall** — the proportion of the class that we correctly predicted. Also known as *sensitivity*.

**F1 score** — the harmonic mean of precision and recall. It's a combined metric for each class.

**Accuracy** – the total fraction of correct predictions for all classes. You can calculate this for each class, but it will be the same value for each of the class.

## DIY classification report

If you're like me and you find the grammar of true positives and false negatives confusing, it might help to treat each class within the classifier as its own mini diagnostic test, and build up data for the classification report row by row. Then it's as simple as counting hits and misses from the validation data and computing some fractions. Inspired by this diagram on the Wikipedia page for the F1 score, I've given both text and pictorial versions of the equations:

Have a go at filling in the scores for the two classes above. After that, fill in your answers into your own hand-drawn version of

the empty table below. Notice that there is only a single score for accuracy for the entire classifier, and that there may be a richer story between the various other scores in the table. Do you want to optimize accuracy overall? Or perhaps you care about maximizing recall in one class above all else? What matters most to you? Should you penalize some mistakes stronger than others?

When data sets get larger – by either increasing the number of samples, or increasing the dimensionality of the data – even though this scoring-by-hand technique becomes impractical, the implementation stays the same. In classification problems that have more than two classes we can add in a confusion matrix to our reporting, which is something that deserves a whole other post.

Upon finishing logging a slab of core, if you were to ask Sally the stratigrapher, "How accurate are your facies?", she may dismiss your inquiry outright, or maybe point to some samples she's not completely confident in. Or she might tell you that she was extra diligent in the transition zones, or point to regions where this is very sandy sand, or this is very hydrothermally altered. Sadly, we in geoscience – emphasis on the science – seldom take the extra steps to test and report our own performance. But we totally could. The ANSWERS. Upside Down. To two Decimal places.

---