Performance Evaluation in Machine Learning: The Good, The Bad, The Ugly and The Way Forward

Peter Flach
Intelligent Systems Laboratory, University of Bristol, UK
The Alan Turing Institute, London, UK
Peter.Flach@bristol.ac.uk

Abstract

This paper gives an overview of some ways in which our understanding of performance evaluation measures for machine-learned classifiers has improved over the last twenty years. I also highlight a range of areas where this understanding is still lacking, leading to ill-advised practices in classifier evaluation. This suggests that in order to make further progress we need to develop a proper *measurement theory* of machine learning. I then demonstrate by example what such a measurement theory might look like and what kinds of new results it would entail. Finally, I argue that key properties such as classification ability and data set difficulty are unlikely to be directly observable, suggesting the need for latent-variable models and causal inference.

Introduction

Data-driven AI systems typically operate in a rich ecosystem involving many different components and actors, between which a multitude of signals and messages are passed. Important signals include the predicted target value for a particular data case as estimated by a model (e.g., a class label in classification, or a real number in regression); the variability or uncertainty in these estimates (e.g., confidence intervals, calibrated class probabilities); and performance measurements of a machine learning model on a test data set (e.g., classification accuracy or F-score). The latter kind of signals will be our main concern in this paper.

It may appear that performance measurements and related signals are well-understood, at least in supervised machine learning. For example, classification performance can be measured by a range of evaluation measures including accuracy, true and false positive rate, precision and recall, F-score, Area Under (ROC) Curve, and Brier score. Each of these evaluation measures has a clear technical interpretation that can be linked to particular use cases. There are furthermore well-defined relationships between many of these evaluation measures.

However, one only has to dig a little deeper in the machine learning literature for problematic issues to emerge, often stemming from a limited appreciation of the importance of the *scale* on which evaluation measures are expressed. Several examples will be given later in the paper, which aims to give a balanced view of where we are in machine learning performance evaluation, where we should aim to be, and how we might get there. This short paper is therefore in part a review of my own and others' work on evaluation measures, in part a critique of current practice in empirical machine learning, and in part a suggested way forward towards a well-founded measurement theory for machine learning.

Performance Evaluation in Machine Learning

I will start by reviewing and critiquing current practice in performance evaluation of machine learning algorithms. I will highlight some good things, some not so good things, and some things to be avoided. This is intended to demonstrate, by example, the need for a more careful treatment of performance evaluation and the development of a specific measurement framework for machine learning, but should not be taken as complete in any sense.

ML Evaluation: The Good

An important development in the last twenty years has been the realisation that, even in the simplest scenarios, a single aggregated measurement is insufficient to accurately reflect the performance of a machine learning algorithm. Provost, Fawcett, and Kohavi (1998) pointed out the limitations of using predictive accuracy as the gold standard in classification and were among the early proponents of ROC analysis in this context. Provost and Fawcett (2001) demonstrated the usefulness of ROC analysis for dealing with changing class and cost distributions, and introduced the ROC convex hull method, which was later shown to be equivalent to using isotonic regression to obtain calibrated probabilities from a classifier (Fawcett and Niculescu-Mizil 2007; Flach and Matsubara 2007). An example ROC curve with its convex hull can be seen in Figure 1 (left).

ROC analysis derives its main ethos from multi-objective optimisation, which is to delay choosing a trade-off between the different objectives to be optimised for as long as possible. By discarding all dominated points (ones that cannot be optimal under any trade-off) one is left with the set of non-dominated solutions (the Pareto front) from which the optimal operating point can be obtained once a trade-off is fixed. In classification the optimisation objectives are per-class accuracies, and the Pareto front is the

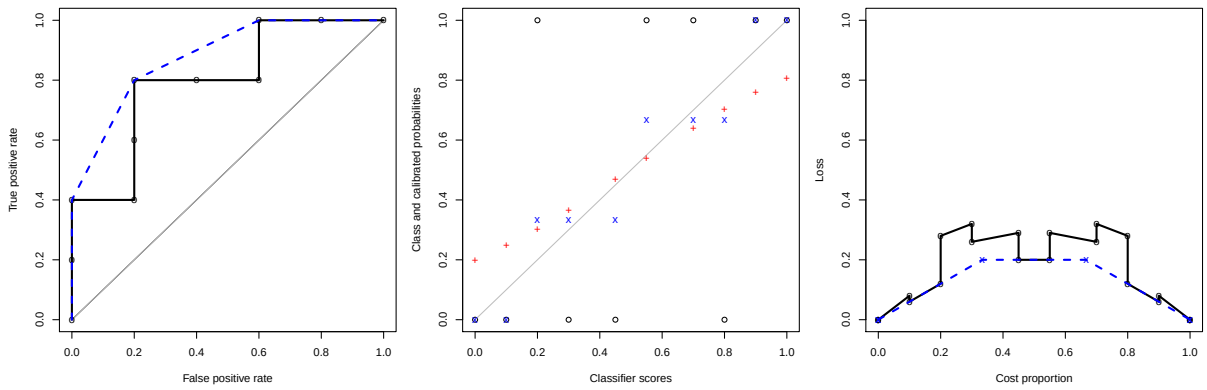

Figure 1: (left) Example ROC curve (black, solid line) and convex hull (blue, dashed line) on 10 instances with true labels $++-++--+--$, ranked on decreasing classifier scores. Starting in the origin, the curve steps up for a positive and along for a negative, until it reaches the NE corner; the ideal curve would go through the NW corner. This classifier mis-ranks 5 out of 25 $+/-$ pairs, corresponding to the five cells above the curve (AUC $= 20/25 = 0.8$). The convex hull introduces bins $[++][-++][--+][--]$, turning four ranking errors into half-errors (AUCH $= 22/25 = 0.88$). (centre) Uncalibrated scores against true classes (black circles), calibrated scores obtained by isotonic regression (blue crosses) and logistic calibration (red plusses). (right) Brier curves for the original scores (black, solid line) and isotonically calibrated scores (blue, dashed line); the difference between the two curves represents the decrease in Brier score achievable by calibration. Figure from (Flach 2016).

ROC convex hull. A trade-off between true and false positive rate manifests itself as an isometric: a straight line with a particular slope (e.g., an accuracy isometric has the ratio of negative to positive examples as its slope (Flach 2004; 2011)). The convex hull is constructed from isometrics with slopes that can be used to obtain calibrated probabilities.

Importantly, ROC analysis can also be used as a conceptual tool, and has been used to develop a new decision tree splitting criterion (Ferri, Flach, and Hernández-Orallo 2002) and to more generally understand performance evaluation metrics (Flach 2003; Fürnkranz and Flach 2005). Also very insightful are the cost curves introduced by Drummond and Holte (2006). Cost curves differ from ROC curves in that they explicitly represent trade-offs and and the loss incurred at specific operating points, but at the expense of losing isometrics as a visual representation of the trade-off under which that threshold is optimal (Figure 1 (right)).

What ROC curves and cost curves do have in common, though, is the fact that the *area under the curve* (AUC) is itself meaningful as an aggregated performance measure. The area under the ROC curve (the expected true positive rate when uniformly averaging over all false positive rates) can not just be interpreted as an estimate of the probability that a random positive is ranked higher than a random negative, but is also linearly related to the expected classification accuracy of a classifier that sets its rate of positive predictions in a certain way, a result first derived by Hernández-Orallo, Flach, and Ferri (2012). The area under the cost curve (i.e., the expected misclassification loss when uniformly averaging over all trade-offs) was shown by Hernández-Orallo, Flach, and Ramirez (2011) to be equal to the Brier score of a probabilistic classifier (the mean squared residuals compared to the 'ideal' probabilities 0 and 1). However, the practice of averaging performance measures in this way is not universally applicable – we will soon encounter an area-under-curve technique that should be avoided.

What links these techniques together is a move away from decision boundaries *per se*, considering instead the entire range of operating points: we don't want just $\hat{p}(Y|X) = 0.5$ to be in the right place, but also, say, $\hat{p}(Y|X) = 0.3$ or $\hat{p}(Y|X) = 0.8$. The advantage of calibrating over the whole range of predicted probabilities is that, if the class distribution changes from $50/50$ in training to $30/70$ or $80/20$ in testing, we can simply change the decision threshold accordingly.[1] There is a wealth of material on calibration and scoring of probabilistic forecasts that is now being exploited in machine learning, see for example Kull and Flach (2015). Figure 1 (centre) shows the results of two calibration methods, one using isotonic regression and the other using a logistic sigmoid.

ML Evaluation: The Bad

I now turn to some less laudable practices that can nevertheless be observed in experimental machine learning. An obvious tendency is to over-report evaluation measures: for example, in classification one often sees accuracy/error rate, F-score and AUC reported on the same experiments.[2] The key point is that *each of these measures assumes a different use case*: accuracy assumes that, within each class, the difference in cost of correctly classifying an instance and misclassifying it is the same, while F-score assumes addi-

[1] If the training distribution is imbalanced, calculation of the new threshold is marginally more involved (Flach 2016).

[2] A regex search on the ICML 2018 PDFs suggests that of the 32 papers reporting AUC, nearly two-third (21) also report accuracy.

tionally that true negatives do not add value; both assume that the class distribution in the test set is meaningful. Furthermore, these two measures assume that the classifier has a fixed operating point, whereas AUC aggregates over operating points in response to changing costs or class distributions – or alternatively, deals with ranking performance rather than classification performance.

It is possible that each of these use cases are relevant for a particular study, but then this should be clearly stated: 'the goal of this experiment is to test the algorithm's performance in such-and-such use case, which is measured adequately by performance measure so-and-so'. Experimental set-ups are often inherited from previous studies, which quite possibly encourages this 'everything and the kitchen sink' approach. But it would be highly desirable, in my opinion, for machine learning experimenters to be more explicit about the objective of their experiments and to justify the reported evaluation measures from that perspective, so that it becomes more straightforward to translate the experimental measurements back to conclusions about the experimental objective.[3]

A concrete example where two related measures assess quite different use cases concerns mean absolute error (MAE) and mean squared error (MSE) of probabilistic classifiers. Mean squared error or Brier score is a useful measure as it decomposes into calibration loss and refinement loss, and can also be interpreted as the expected loss when a classifier sets its decision threshold equal to the cost parameter c (the cost of a false positive in proportion to the sum of the costs of false positive and a false negative) (Hernández-Orallo, Flach, and Ferri 2012). Mean absolute error can be interpreted as the expected loss if the classifier uses its predicted probability $\hat{p}$ to make a stochastic prediction: positive with probability $\hat{p}$, negative with probability $1-\hat{p}$. Note that this ignores the operating condition c, and also gives a loss that is always higher than MSE (except in edge cases), so is unlikely to be a practically useful scenario. This suggests that MAE is almost never worth reporting, and authors should prefer MSE.

The treatment of probabilistic classifier scores is an area where current practice can fall short more generally. This is perhaps best illustrated using the case of the naive Bayes classifier, which makes the simplifying – and simplistic – assumption that within each class features are mutually independent, and hence estimates the likelihood ratio jointly over all features as a product of the per-feature likelihood ratios. As a result, naive Bayes' estimates of these likelihoods are almost always woefully poor – its usefulness as a classifier stems solely from the fact that often it does a good job as a ranker (Domingos and Pazzani 1997). In terms of performance measures, one would expect a poor Brier score but a decent AUC. Because of this, the decision threshold cannot be fixed in advance or derived from Bayes' rule, but should be estimated from the ROC curve. In other words, despite having the appearance of a probabilistic classifier, naive Bayes is best treated as a scoring classifier whose scores happen to fall in the interval $[0,1]$; these scores must be calibrated in post-processing in order to be meaningful as estimates of the posterior class probability. That naive Bayes is neither probabilistic nor Bayesian is not widely acknowledged in the machine learning literature, and textbooks continue to invoke Bayes' rule as the decision rule.

In a related vein, logistic regression is often said to yield well-calibrated probability estimates, but this will only be true insofar the parametric assumptions of the model – logits deriving from normal distributions within each class – are satisfied, and can give arbitrarily bad results if they are not. Similarly, scores from a support vector machine are often calibrated by fitting a logistic sigmoid following Platt (1999), but this again assumes that the scores within each class are normally distributed, and there is nothing in the SVM model that guarantees it – indeed, the whole point of an SVM seems to be to avoid modelling distributions and instead identify support vectors. This makes the SVM + logistic calibration hybrid a somewhat curious mongrel. Logistic sigmoids are also widely used in deep neural networks, in the form of the parameter-free 'softmax' function or the more recently proposed 'temperature scaling' which learns a single shape parameter across all classes (Guo et al. 2017), but again the assumption that the preceding network produces logits is rarely justified.

Finally – and this is already looking to a way forward – there appears to be a widespread belief that properties of interest in experimental machine learning are directly measurable, e.g. by inspecting a confusion matrix. Here I suggest we take a leaf out of the psychometrician's book, who will be very familiar with the idea that many variables of interest – such as the difficulty of a test or the ability of a student – are *latent variables* that manifest themselves only indirectly through test results. Latent variable models are of course an important tool in the machine learner's toolbox and it is hence somewhat embarrassing that machine learners haven't yet caught on to using psychometrics-like tools in their own experimental practices, but this is fortunately starting to change (Bachrach et al. 2012; Martínez-Plumed et al. 2016).

ML Evaluation: The Ugly

I now turn to some practices that are downright wrong yet not widely recognised as such. As two main culprits I mention the tendency to use the arithmetic mean as the sole way to obtain averages, and a related tendency to use linear interpolation without checking that it is indeed coherent to do so. These issues will be further discussed in the next section so I will only give one example here, as both mistakes are often made in the context of so-called precision-recall curves. These plot precision against recall while varying the decision threshold of the classifier, in much the same way as ROC curves are produced. Authors also often report the area under the precision-recall curve (AUPR), which was for example used in a well-known object classification challenge (Everingham et al. 2015). But this practice makes exactly the two mistakes just mentioned:

- while linear interpolation is correct in ROC space, in the sense that any operating point on a straight line between

[3] Berrar and Flach (2011) discuss further cases of misapplication and mis-understanding of AUC, such as the fallacy that $AUC = 1/2$ means random performance.

two classifiers can be achieved by random choice between those classifiers, it doesn't carry over to precision-recall space as pointed out by Davis and Goadrich (2006);

- even with the correct hyperbolic interpolation method the area under the precision-recall curve is meaningless as linear expectations are incoherent here (Flach and Kull 2015).

Finally, I mention the dangers of using parametric models in situations where the distributions involved are inappropriate. One example of this arises if one applies logistic calibration, which assumes scores on an unbounded scale as, e.g., output by a support vector machine, to a classifier which scores on a bounded scale, such as naive Bayes.[4] The solution proposed by (Kull, Silva Filho, and Flach 2017a; 2017b) is to replace the Gaussian distributions underlying the logistic sigmoid with Beta distributions as they have finite support.

The Way Forward: A Measurement Theory for Machine Learning

Having looked at practices good and bad in performance evaluation of machine learning algorithms, I postulate that many of the issues discussed relate to notions of *scale*. We therefore turn our attention to measurement theory, which is the study of concepts of measurement and scale. After a brief introduction I will discuss how insights from measurement theory can be brought to bear on machine learning evaluation.

Concatenation, Scales and Transformations

Representational measurement is one of the most developed formal systems for measurement (Krantz et al. 1971–1990). Representational measurement studies homomorphisms between an *empirical relational system* (ERS), describing the relationships between measured objects in the real world, and a *numerical relational system* (NRS) which aims to capture those relationships numerically (Hand 2004). The fundamental empirical relationship is *concatenation*: e.g., placing two rigid rods a and b end to end in a straight line would be denoted $a \circ b$. If M denotes the mapping of rods into some numerical scale representing their length, we would want this scale to be such that it allows an operation $\oplus$ in the NRS that corresponds to concatenation in the ERS, i.e., $M(a \circ b) = M(a) \oplus M(b)$, which represents the combined length of the concatenated rods. Furthermore, we would have an equivalence relation $\sim$ on rods and their concatenations, indicating that they are of the same length; this equivalence relation would map to equality in the NRS, i.e.: if $a \circ b \sim c$ then $M(a \circ b) = M(c)$.

There may be multiple concatenation relationships and hence multiple numerical operations: e.g., if we are concerned with electrical resistance then connecting two resistors in series would give one type of concatenation, say $M(a \circ_s b) = M(a) \oplus_s M(b)$, while putting them in parallel would give another, say $M(a \circ_p b) = M(a) \oplus_p M(b)$. If M measures resistance then we would have $r_1 \oplus_s r_2 = r_1 + r_2$ and $r_1 \oplus_p r_2 = \left(r_1^{-1} + r_2^{-1}\right)^{-1}$; whereas if M measures conductance (the reciprocal of resistance) then we would have $c_1 \oplus_s c_2 = \left(c_1^{-1} + c_2^{-1}\right)^{-1}$ and $c_1 \oplus_p c_2 = c_1 + c_2$. So, not only do we have multiple concatenation relationships in this case, but also a choice of measurement scales with associated transformations between them ($c = 1/r$ and $r = 1/c$).

Concatenation also gives us averaging: e.g., a rod d with the average length of a and b would be such that $a \circ b \sim d \circ d$, hence $M(a \circ b) = M(d \circ d)$, hence $M(a) \oplus M(b) = M(d) \oplus M(d)$. If we use an additive scale to measure length we would have $M(a) + M(b) = M(d) + M(d)$, hence $M(d) = (M(a) + M(b))/2$ (arithmetic mean); if we use a multiplicative scale we would have $M'(a)M'(b) = M'(d)M'(d)$, hence $M'(d) = \sqrt{M'(a)M'(b)}$ (geometric mean); notice that such a scale can be construed as arising from an additive scale by the transformation $M'(a) = \exp M(a)$ and hence transformed back into an additive scale by means of $M(a) = \ln M'(a)$. The harmonic mean arises, e.g., when we put two resistors in parallel (series) and we want to achieve the same resistance (conductance) with two resistors with equal resistance (conductance).

There are also transformations that don't change the numerical operations in an essential way, such as transforming pounds into euros (today $p = 1.12e$), or degrees Celsius into Fahrenheit ($c = (f - 32) \cdot 5/9$). Notice that a change of currency only involves scaling, since zero pounds always corresponds to zero euros (or zero in any other currency). This means that statements such as 'this costs twice as much as that' are meaningful regardless of the currency – currency scales are said to be of *ratio* scale type. In contrast, if two object's temperatures are 50° and 100° Celsius then it doesn't make sense to say that the second object is twice as hot as the first. Only ratios of differences are meaningful here: e.g., one could say 'in the previous hour the patient's body temperature increased twice as much as in the hour before' which does not depend on the temperature scale used – temperature scales are said to be of *interval* scale type. So ratio scales are invariant under scaling $x \mapsto ax$ only whereas interval scales admit affine transformations $x \mapsto ax + b$. Other scale types distinguished in the literature include *log-interval*, which are scales invariant under transformations $x \mapsto cx^d$ (the name derives from the fact that a logarithmic transformation changes the scale into an interval scale – or a ratio scale if $c = 1$); *ordinal*, which are scales invariant under any monotonic transformation; and *nominal*, which allows any one-to-one transformation of scale values (Stevens 1946).

Measurements on Confusion Matrices

Thinking about measurement in machine learning, the first things that probably come to mind are absolute frequencies (counts), relative frequencies (proportions), (conditional) probabilities, etc. As we can meaningfully say things like 'classifier A correctly classifies twice as many examples as classifier B' or 'classifier C predicts a positive outcome as twice as likely as classifier D' these measurements seem to be expressed on a ratio scale. However, such frequen-

[4] See, for example, `http://scikit-learn.org/stable/modules/calibration.html` where logistic calibration is applied to naive Bayes scores.

cies and probabilities are also bounded from above, which is something not recognised by Stevens' levels of measurement. That is, we can also meaningfully say 'classifier B misclassifies twice as many examples as classifier A' – the complement of such measurements is also expressed on a ratio scale. Chrisman (1998) calls such scales *absolute*.

What would be sensible notions of concatenation in machine learning? Concentrating on performance evaluation as we do in this paper, let us consider what happens in one of the most common evaluation scenarios in machine learning: cross-validation of classifiers. In this scenario we partition a data set into k 'folds', train a model on $k-1$ of those and test it on the remaining fold. This is repeated so that each fold is used for testing exactly once. We calculate a performance measure of choice on each test fold and aggregate those k measures to arrive at a single measure of performance. For example, we could calculate the arithmetic mean of the accuracies obtained in each test fold.

Let us furthermore assume that our chosen performance measure can be solely evaluated in terms of the counts in a confusion matrix, and consider binary classification for simplicity. A two-class confusion matrix with marginals looks as follows:

	Predicted +	*Predicted −*	
Actual +	*TP*	*FN*	*Pos*
Actual −	*FP*	*TN*	*Neg*
	PPos	*PNeg*	*N*

Consider now an aggregated matrix which contains in each cell the sum of all corresponding values from the per-fold confusion matrices. This aggregated matrix is itself a confusion matrix over the entire data set, recording for each instance how it was classified when it was part of a test fold. Taking this summing of confusion matrices as our concatenation operation we can then obtain

$$acc(C_i) = \frac{TP_i + TN_i}{N_i}, i = 1,2 \tag{1}$$

$$acc(C_1 \circ C_2) = \frac{\sum_i TP_i + \sum_i TN_i}{\sum_i N_i} = \frac{N_1}{N} acc(C_1) + \frac{N_2}{N} acc(C_2) \tag{2}$$

In particular, if the two test sets have the same number of test instances (which is natural in cross-validation) we see that this form of concatenation of confusion matrices corresponds to arithmetic averaging of accuracies. This establishes a first component of an NRS for confusion matrices.

Can we obtain similar results for other evaluation measures? Let's consider true positive rate, the proportion of actual positives correctly classified:

$$tpr(C_i) = \frac{TP_i}{Pos_i}, i = 1,2 \tag{3}$$

$$tpr(C_1 \circ C_2) = \frac{\sum_i TP_i}{\sum_i Pos_i} = \frac{Pos_1}{Pos} tpr(C_1) + \frac{Pos_2}{Pos} tpr(C_2) \tag{4}$$

We again obtain a weighted arithmetic mean, which will be unweighted if the test sets have the same class distribution (a practice called stratified cross-validation). We can obtain similar numerical relationships for false positive rate, and for true/false negative rate: for all these measures *it is sufficient to record their value on each test set and average them afterwards, without needing to inspect the individual confusion matrices or to form the concatenated matrix*.

What about precision, the proportion of positive predictions that are correct? Algebraically this is again straightforward:

$$prec(C_i) = \frac{TP_i}{PPos_i}, i = 1,2 \tag{5}$$

$$prec(C_1 \circ C_2) = \frac{\sum_i TP_i}{\sum_i PPos_i} = \frac{PPos_1}{PPos} prec(C_1) + \frac{PPos_2}{PPos} prec(C_2) \tag{6}$$

However, the number of positive predictions will not, in general, be constant across test folds, and *the 'weights' in this expression are themselves measurements* that require inspection of the individual confusion matrices. That is, if we want to use cross-validation to estimate the precision of a classifier, we need to either (i) record both the true positives and false positives in each test fold, or (ii) change the cross-validation protocol so that every classifier makes the same number of positive predictions.[5]

Let us now consider the F-score, which is customarily defined as the harmonic mean of precision and recall (the latter being another name for true positive rate). The use of the harmonic rather than the arithmetic mean here is often glossed over, but is itself of measure-theoretic interest. It could be justified as follows: the first row and the first column of the confusion matrix are vectors of the form $(G(ood), B(ad))$, and both precision and recall calculate the quantity $M_{PR}((G,B)) = G/(G+B)$ from their respective vector. We can concatenate the two vectors by entry-wise arithmetic averaging, yielding

$$\begin{aligned} M_{PR}((TP,FP) \circ (TP,FN)) &= M_{PR}((TP,(FP+FN)/2)) \\ &= 2TP/(2TP+FP+FN) \end{aligned} \tag{7}$$

which is easily seen to be equivalent to the harmonic mean of precision and recall. So the F-score arises as the counterpart of a very natural notion of concatenation of good/bad vectors; and the arithmetic mean of counts corresponds, *via a change of scale*, to the harmonic mean at the level of precision and recall.

Going back to the previous discussion of aggregating confusion matrices in cross-validation, we can derive the following expression for the F-score of two concatenated confusion matrices:

$$\begin{aligned} Fscore(C_1 \circ C_2) = &\frac{Pos_1 + PPos_1}{Pos + PPos} Fscore(C_1) \\ &+ \frac{Pos_2 + PPos_2}{Pos + PPos} Fscore(C_2) \end{aligned} \tag{8}$$

[5] Changing the concatenation operation is theoretically possible but does not really seem a viable option: for example, an aggregated confusion matrix with $(TP_1 \cdot PPos_2 + TP_2 \cdot PPos_1)/2$ in its true positive cell and $PPos1 \cdot PPos2$ in its predicted positive cell would have precision equal to the arithmetic mean of the two individual precision values but is hard if not impossible to interpret.

As with precision, we will need access to the number of positive predictions in each test set or force them to be equal. Flach and Kull (2015) present a way of avoiding the harmonic mean altogether by introducing precision/recall gain, but even then one would need to record the number of true positives in each test set in order to properly aggregate.

$$recGain(C_1 \circ C_2) = \frac{TP_1}{TP} recGain(C_1) + \frac{TP_2}{TP} recGain(C_2) \tag{9}$$

$$precGain(C_1 \circ C_2) = \frac{TP_1}{TP} precGain(C_1) + \frac{TP_2}{TP} precGain(C_2) \tag{10}$$

$$FGain(C_1 \circ C_2) = \frac{TP_1}{TP} FGain(C_1) + \frac{TP_2}{TP} FGain(C_2) \tag{11}$$

Some Initial Results

If we further develop a measurement theory along these lines, what kinds of result can we hope to obtain? Here are two conjectures that I believe would be easy to prove.

Conjecture 1 *Concatenation of confusion matrices by cell-wise summing corresponds to arithmetic averaging of evaluation measures with weights not requiring further measurements if and only the latter have parallel ROC isometrics.*

This conjecture applies to accuracy, true/false positive/negative rate, weighted relative accuracy (Lavrač, Flach, and Zupan 1999), and variants that manipulate the slope of the isometrics taking external factors into account (such as misclassification costs).

Conjecture 2 *Concatenation of confusion matrices by cell-wise summing corresponds to arithmetic averaging of evaluation measures with weights possibly requiring further measurements if and only if the latter have straight ROC isometrics.*

This extends the first conjecture to cover evaluation measures such as precision and F-score and their gain versions alluded to above, but excludes measures such as the arithmetic (or geometric) mean of precision and recall as these would give non-linear isometrics (as would most decision tree splitting criteria (Fürnkranz and Flach 2005)). The weights require further measurements if the slope of the isometrics varies across ROC space, which means Conjecture 1 doesn't apply.

Conclusions and Outlook

Measurement is of evident importance in machine learning in at least two ways:

- the features used by machine learning models are themselves measurements, which is a perspective explored in (Flach 2012, Chapter 10);
- performance evaluation of learned models, which was the topic of this paper.

The development of a measurement theory for machine learning is lagging behind the remarkable achievements of machine learning technology itself, a situation that is in urgent need of rectification if we want that technology to be accepted and trusted by users. It is hoped that this paper will provide an impetus in that direction.

On the positive side, I have demonstrated that much progress has been made in providing a multi-objective optimisation perspective on machine learning evaluation metrics, using tools such as ROC curves and cost plots. There are many opportunities to expand this line of work as only a few works go beyond two-class ROC analysis (Mossman 1999; Dreiseitl, Ohno-Machado, and Binder 2000; Ferri, Hernández-Orallo, and Salido 2003; Everson and Fieldsend 2006). There will be much value derived from techniques for determining and approximating the Pareto front developed in the multi-objective optimisation field.

I have outlined the contours of a measurement theory for evaluation measures based on confusion matrices. The key concept is concatenation, of which I have given two realistic examples: one within a matrix (to derive F-score) and the main one between confusion matrices as might be obtained from cross-validation. I have then shown that some evaluation measures commute in the sense that (measure then average) gives the same result as (concatenate then measure), but others don't – F-score in particular. This means that, in the absence of a sensible concatenation operator which commutes, arithmetic averages of F-scores are incoherent. I have also outlined the kind of formal results that might be expected from such a measurement theory, linking them to geometric properties in ROC space (Flach 2003).

Other highly promising opportunities involve the use of latent-variable models. Here we envisage a trained classifier to have an ability, and data sets (or data points) to have a difficulty, both unobserved. Using experimental results such as collected, e.g., by `openml.org`, we can estimate these latent variables using models similar to item-response theory (IRT) models in psychometrics (Embretson and Reise 2013). This would be useful, for example, to develop adaptive tests for machine learning algorithms (a kind of binary search in ability space using few data sets with varying difficulties). We have in fact started such work and will be reporting on it in due course.

Would a measurement theory endowed with latent variables be all we require? It seems to me that, ultimately, the kinds of conclusions we want to draw from our machine learning experiments are causal: 'this algorithm outperforms that algorithm because ...'. This would neatly tie in with the 'causal revolution' that has been declared (Pearl and Mackenzie 2018).

Acknowledgements

Many thanks to José Hernández-Orallo for fruitful discussions and to Kacper Sokol, Miquel Perello-Nieto and Simon Price for helping with the regex search on ICML proceedings. This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

References

Bachrach, Y.; Minka, T.; Guiver, J.; and Graepel, T. 2012. How to grade a test without knowing the answers: a

Bayesian graphical model for adaptive crowdsourcing and aptitude testing. In *Proc 29th Int Conf Machine Learning*, 819–826.

Berrar, D., and Flach, P. 2011. Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). *Briefings in Bioinformatics* 13(1):83–97.

Chrisman, N. R. 1998. Rethinking levels of measurement for cartography. *Cartography and Geographic Information Systems* 25(4):231–242.

Davis, J., and Goadrich, M. 2006. The relationship between precision-recall and ROC curves. In *23rd Int Conf on Machine Learning*, 233–240.

Domingos, P., and Pazzani, M. 1997. On the optimality of the simple bayesian classifier under zero-one loss. *Machine learning* 29(2-3):103–130.

Dreiseitl, S.; Ohno-Machado, L.; and Binder, M. 2000. Comparing three-class diagnostic tests by three-way ROC analysis. *Medical Decision Making* 20(3):323–331.

Drummond, C., and Holte, R. C. 2006. Cost curves: An improved method for visualizing classifier performance. *Machine Learning* 65(1):95–130.

Embretson, S. E., and Reise, S. P. 2013. *Item response theory*. Psychology Press.

Everingham, M.; Eslami, S. A.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2015. The Pascal visual object classes challenge: A retrospective. *Int J of Computer Vision* 111(1):98–136.

Everson, R. M., and Fieldsend, J. E. 2006. Multi-class ROC analysis from a multi-objective optimisation perspective. *Pattern Recognition Letters* 27(8):918–927.

Fawcett, T., and Niculescu-Mizil, A. 2007. PAV and the ROC convex hull. *Machine Learning* 68(1):97–106.

Ferri, C.; Flach, P.; and Hernández-Orallo, J. 2002. Learning decision trees using the area under the ROC curve. In *Int Conf on Machine Learning*, 139–146.

Ferri, C.; Hernández-Orallo, J.; and Salido, M. A. 2003. Volume under the ROC surface for multi-class problems. In *Eur Conf on Machine Learning*, 108–120. Springer.

Flach, P., and Kull, M. 2015. Precision-recall-gain curves: PR analysis done right. In *Advances in Neural Information Processing Systems*, 838–846.

Flach, P., and Matsubara, E. T. 2007. A simple lexicographic ranker and probability estimator. In *Eur Conf on Machine Learning*, 575–582. Springer.

Flach, P. 2003. The geometry of ROC space: understanding machine learning metrics through ROC isometrics. In *20th Int Conf on Machine Learning*, 194–201.

Flach, P. 2004. The many faces of ROC analysis in machine learning. ICML Tutorial, `http://people.cs.bris.ac.uk/~flach/ICML04tutorial/`.

Flach, P. 2011. ROC analysis. In *Encyclopedia of machine learning*. Springer. 869–875.

Flach, P. 2012. *Machine Learning: the art and science of algorithms that make sense of data*. Cambridge University Press.

Flach, P. 2016. Classifier calibration. In *Encyclopedia of Machine Learning and Data Mining*. Springer.

Fürnkranz, J., and Flach, P. 2005. Roc 'n' rule learning – towards a better understanding of covering algorithms. *Machine Learning* 58(1):39–77.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *Int Conf on Machine Learning*, 1321–1330.

Hand, D. J. 2004. *Measurement theory and practice*. Hodder Arnold.

Hernández-Orallo, J.; Flach, P.; and Ferri, C. 2012. A unified view of performance metrics: translating threshold choice into expected classification loss. *J Machine Learning Research* 13:2813–2869.

Hernández-Orallo, J.; Flach, P.; and Ramirez, C. F. 2011. Brier curves: a new cost-based visualisation of classifier performance. In *Int Conf on Machine Learning*, 585–592.

Krantz, D. H.; Luce, R. D.; Suppes, P.; and Tversky, A. 1971–1990. *Foundations of Measurement: Volumes I-III*. Academic Press.

Kull, M., and Flach, P. 2015. Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In *Joint Eur Conf on Machine Learning and Knowledge Discovery in Databases*, 68–85.

Kull, M.; Silva Filho, T.; and Flach, P. 2017a. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial Intelligence and Statistics*, 623–631.

Kull, M.; Silva Filho, T. M.; and Flach, P. 2017b. Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electr J of Statistics* 11(2):5052–5080.

Lavrač, N.; Flach, P.; and Zupan, B. 1999. Rule evaluation measures: A unifying view. In *Int Conf on Inductive Logic Programming*, 174–185. Springer.

Martínez-Plumed, F.; Prudêncio, R. B.; Martínez-Usó, A.; and Hernández-Orallo, J. 2016. Making sense of item response theory in machine learning. In *22nd Eur Conf on Artificial Intelligence*, 1140–1148.

Mossman, D. 1999. Three-way ROCs. *Medical Decision Making* 19(1):78–89.

Pearl, J., and Mackenzie, D. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Platt, J. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 10(3):61–74.

Provost, F., and Fawcett, T. 2001. Robust classification for imprecise environments. *Machine Learning* 42(3):203–231.

Provost, F. J.; Fawcett, T.; and Kohavi, R. 1998. The case against accuracy estimation for comparing induction algorithms. In *Int Conf on Machine Learning*, 445–453.

Stevens, S. S. 1946. On the theory of scales of measurement. *Science* 103(2684).