A Coherent Interpretation of AUC as a Measure of Aggregated Classification Performance

Peter Flach[1] José Hernández-Orallo[2] Cèsar Ferri[2]

[1]Intelligent Systems Laboratory
University of Bristol, United Kingdom

[2]Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València, Spain

The 28th International Conference on Machine Learning,
June 2011

Outline

Introduction and motivation

Preliminaries

Why *AUC* is perceived to be incoherent

AUC is coherent when including non-optimal thresholds

Expected minimum loss is measured by the area under the optimal cost curve

Conclusions

Motivation

- The area under the ROC curve (*AUC*) is a well-known measure of ranking performance, estimating the probability that a random positive is ranked before a random negative, without committing to a particular decision threshold.
- It is also often used as a measure of aggregated classification performance, on the grounds that *AUC* in some sense averages over all possible decision thresholds and operating conditions.
- Q: How exactly?
- A: in a model-dependent way! (David Hand, MLj 2009)

A summary of Hand's argument

- *AUC* can be interpreted as the expected true positive rate, averaged over all false positive rates.
- For any given classifier we don't have direct access to the false positive rate, and so we average over possible decision thresholds.
- The relationship between decision thresholds and operating conditions under which this threshold is optimal is model-specific, and so the way *AUC* aggregates performance over possible operating conditions is model-specific.
- Expectations over the operating condition are task-specific and not dependent on the model, and so *AUC* may make a model's classification performance look better or worse than it actually is.

Contributions of this paper

- We offer a novel, model-independent interpretation of *AUC* as an aggregation of macro-accuracy over all possible decision thresholds and operating conditions.
- In doing so we provide a unifying framework for classifier performance evaluation.
- We offer a natural interpretation of Hand's alternative to *AUC* (the *H* measure) as the area under the cost curve.

Expected Loss

Let c_0 and c_1 be the cost of misclassifying class 0 or 1 examples, and $b = c_0 + c_1$ and $c = c_0/b$. We set $b = 2$ to ensure loss is commensurate with error rate.

The loss produced at a decision threshold t and a cost proportion c is given by

$$\begin{aligned} Q_c(t;c) &\triangleq c_0\pi_0(1 - F_0(t)) + c_1\pi_1 F_1(t) \\ &= 2\{c\pi_0(1 - F_0(t)) + (1 - c)\pi_1 F_1(t)\} \end{aligned}$$

Expected loss is defined as

$$L_c \triangleq \int_0^1 Q_c(T_c(c);c)w_c(c)dc$$

T_c is a threshold choice method which maps cost proportions to decision thresholds, and $w_c(c)$ is a distribution for cost proportions over $[0, 1]$.

ROC curve and *AUC*

For a given, unspecified classifier and population from which data are drawn, we denote the score density for class k by f_k and the cumulative distribution function by F_k.

The ROC curve is defined as a plot of $F_1(t)$ (i.e., false positive rate at decision threshold t) on the x-axis against $F_0(t)$ (true positive rate at t) on the y-axis.

$$AUC = \int_0^1 F_0(s) dF_1(s) = \int_{-\infty}^{+\infty} F_0(s) f_1(s) ds$$

The convex hull of a ROC curve (ROCCH) includes only those points on the ROCCH with minimum loss for some c, using the *optimal* threshold choice method $T_c^o(c) \triangleq \arg\min_t \{Q_c(t; c)\}$

Cost curves

A cost plot (Drummond & Holte) has loss

$$Q_z(t;z) = z(1 - F_0(t)) + (1 - z)F_1(t)$$

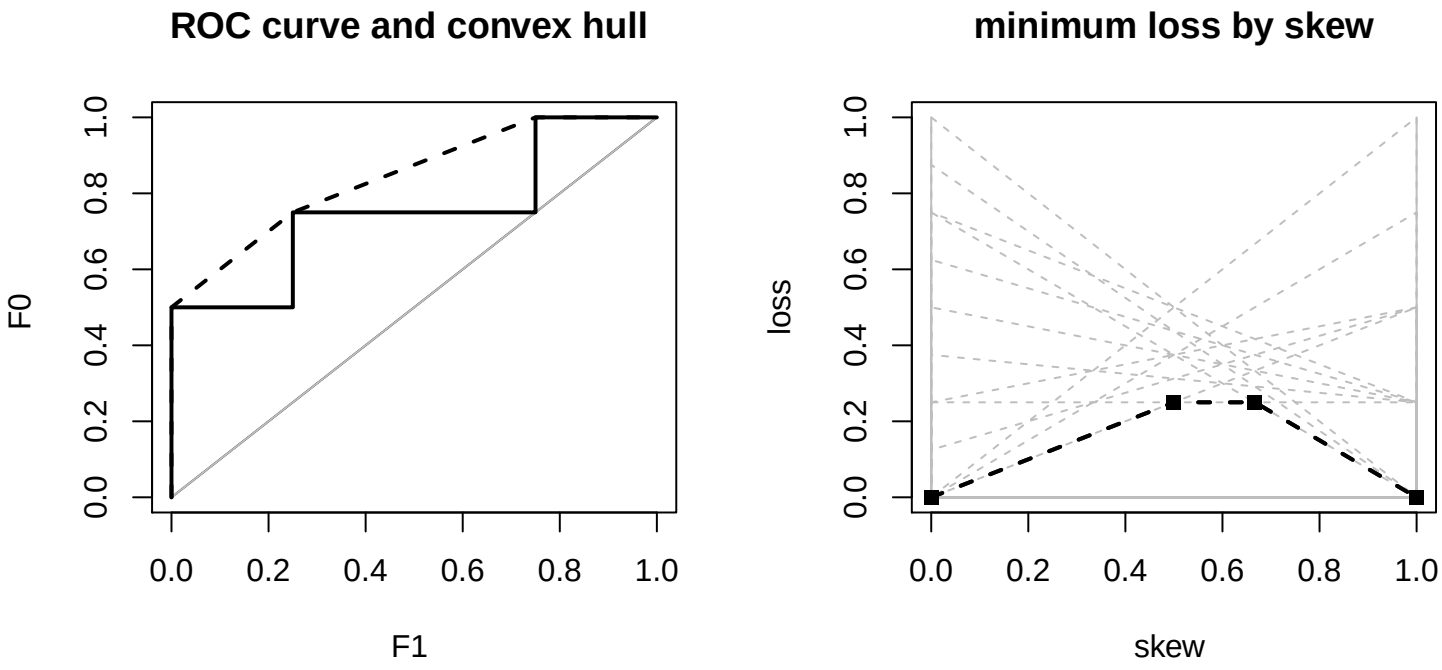
on the y-axis against skew $z = \frac{c_0\pi_0}{c_0\pi_0 + c_1\pi_1}$ on the x-axis.

Cost lines for a given decision threshold t are straight lines with intercept $F_1(t)$ and slope $1 - F_0(t) - F_1(t)$.

The optimal cost curve is the lower envelope of all the cost lines, and only considers the optimal threshold for each skew:

$$CC(z) \triangleq Q_z(T_z^o(z);z)$$

ROC curve, ROCCH and optimal cost curve



ROC curve and convex hull (left), and cost lines and optimal cost curve (right) for a classifier with scores (0.05,0.1,0.2,0.3,0.4,0.5,0.6,0.65,0.7,0.8,0.9,0.95) and corresponding true classes (0,0,1,1,0,1,1,1,1,0,1,1).

Hand's argument in more detail

Hand assumes that thresholds are chosen optimally using $T_c^o(c)$. Under the assumption of a convex and continuously differentiable ROC curve, the mapping from c to $T_c^o(c)$ is one-to-one, with inverse

$$c(T) = \pi_1 f_1(T) / \{\pi_0 f_0(T) + \pi_1 f_1(T)\}$$

By changing the variable of integration we obtain

$$L_c^o = \int_{-\infty}^{\infty} 2\{c(T)\pi_0(1 - F_0(T)) + (1 - c(T))\pi_1 F_1(T)\} W(T) dT$$

Assuming a particular threshold distribution $W_G(T) \triangleq \pi_0 f_0(T) + \pi_1 f_1(T)$ it is possible to derive

$$L_{c,G}^o = 4\pi_0\pi_1(1 - AUC)$$

Hand's argument in more detail (2)

In other words, optimising *AUC* means minimising expected minimum loss under threshold distribution W_G. As there is a one-to-one mapping from optimal thresholds to costs, this can be traced back to a cost distribution

$$w_G(c) = \{\pi_0 f_0(T_c^o(c)) + \pi_1 f_1(T_c^o(c))\} \left| \frac{dT_c^o(c)}{dc} \right|$$

which depends on the score densities and hence on the classifier.

So, two classifiers may have the same *AUC*, but that doesn't imply that they have equal expected minimum loss if a different distribution over cost proportions was used that was the same for both classifiers.

An alternative view

In our view, basing performance metrics on optimal thresholds is overly optimistic. This means that we need to consider thresholds and costs separately:

$$L_c^t \triangleq \int_0^1 \int_{-\infty}^{\infty} Q_c(t;c) W(t) dt \; w_c(c) dc$$

If we assume $w_c(c)$ uniform, this reduces to

$$L_{U(c)}^t = \int_{-\infty}^{\infty} \{\pi_0(1 - F_0(t)) + \pi_1 F_1(t)\} W(t) dt$$

Thresholds are sampled by the mixture distribution as before.

Theorem

$$L_{U(c)}^{U(i)} = \frac{L_{c,G}^o}{2} + \frac{\pi_0^2 + \pi_1^2}{2} = 2\pi_0\pi_1(1 - AUC) + \frac{\pi_0^2 + \pi_1^2}{2}$$

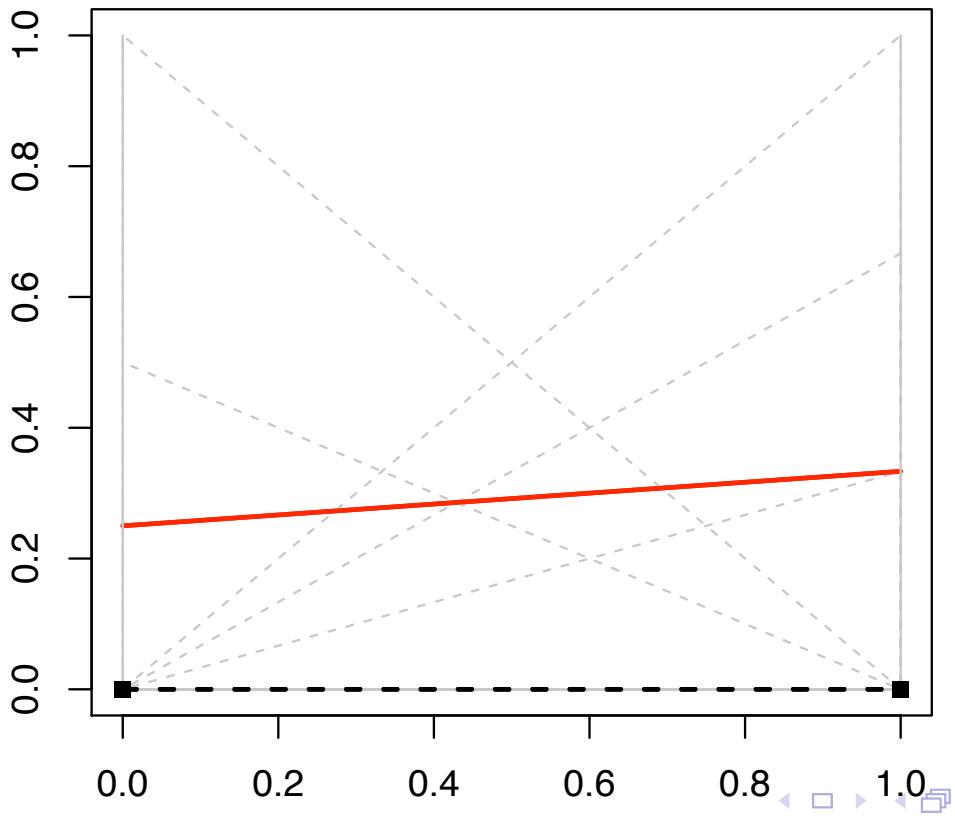
An alternative view (2)

- We again derive a linear relationship between expected loss and *AUC*, assuming a uniform distribution over operating conditions that is therefore the same for all classifiers.
- Expected loss ranges from $(\pi_0^2 + \pi_1^2)/2$ for a perfect ranker that is harmed by sub-optimal threshold choices, to $1 - (\pi_0^2 + \pi_1^2)/2$ for the worst possible ranker that gains some performance by putting the threshold close to one of the extremes.
- Sampling thresholds according to the mixture distribution corresponds to setting the threshold equal to the score of a uniformly selected instance.

The case of empirical ROC curves

Ranking $(0,0,0,1,1)$, scores $(0.1,0.2,0.7,0.8,0.9)$.

	t_1	t_2	t_3	t_4	t_5	t_6	Avg
$F_0(t_i)$	1	1	1	$2/3$	$1/3$	0	$2/3$
$F_1(t_i)$	1	$1/2$	0	0	0	0	$3/12$
$Q_z(t_i;z)$	$1-z$	$\frac{1-z}{2}$	0	$z/3$	$2z/3$	z	$\frac{3+z}{12}$
$\int_0^1 Q_z(t_i;z)dz$	$1/2$	$1/4$	0	$1/6$	$1/3$	$1/2$	$\mathbf{7/24}$



The case of empirical ROC curves (2)

Theorem

Let AUC be an empirical estimate obtained from a dataset with n examples, then the expected loss for uniform skew and (discrete) uniform instance selection is

$$L_{U(z)}^{\mathscr{U}(i)} = \left(\frac{n}{n+1}\right)\frac{1-AUC}{2} + \left(\frac{n+2}{n+1}\right)\frac{1}{4}$$

Theorem

Let CL_{t_i} denote the cost line corresponding to the i-th example with score t_i, then

$$\begin{aligned} L_{U(z)}^{\mathscr{U}(i)} &= \frac{1}{(n+1)}\sum_{i=1}^{n+1}\int_0^1 CL_{t_i}(s)ds \\ &= \frac{1}{(n+1)}\sum_{i=1}^{n+1}(1-MAcc(t_i)) \end{aligned}$$

Expected minimum loss

Hand's alternative to *AUC* is an explicit expected minimum loss measure with the cost distribution $w_c(c)$ equal to the beta distribution $B(c, \alpha, \beta)$ (Hand suggests to use $\alpha = \beta = 2$):

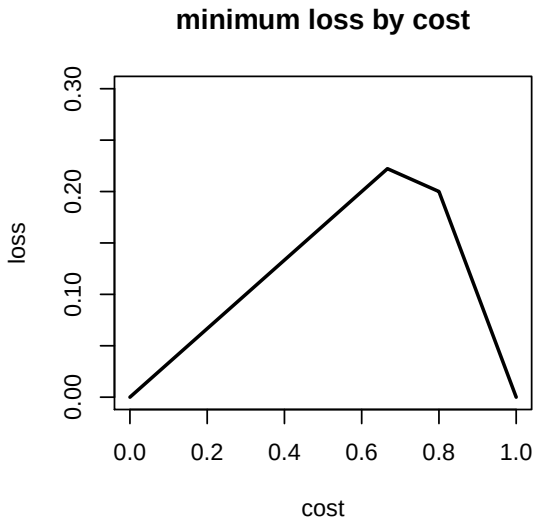
$$L_{\alpha,\beta} \triangleq \int_0^1 Q_c(T_c^o(c); c) B(c, \alpha, \beta) dc$$

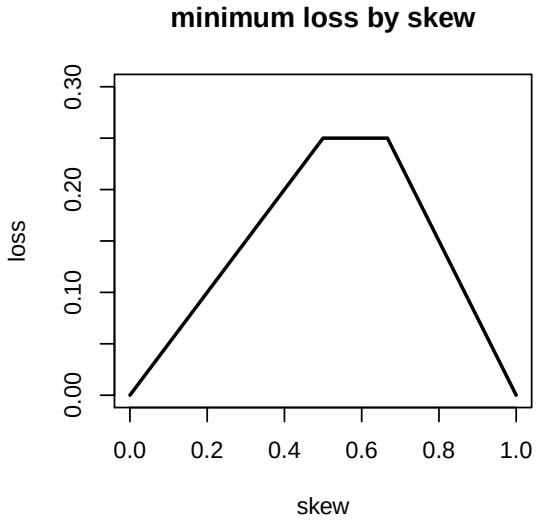
$$H \triangleq 1 - \frac{L_{\alpha,\beta}}{L_{Max}}$$

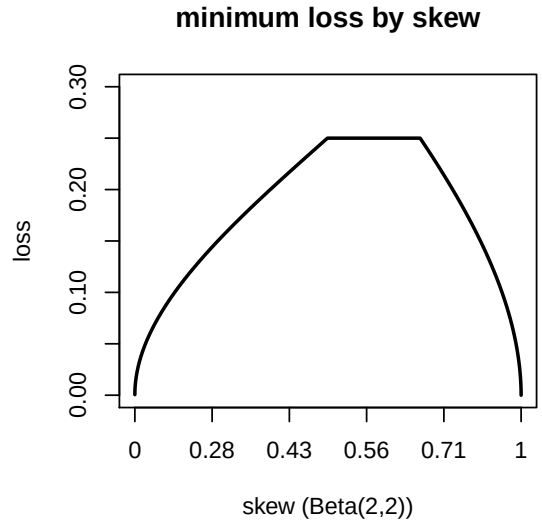
This proposal is very closely related to the area under the optimal cost curve:

$$L_z^o \triangleq \int_0^1 CC(z) dz = \int_0^1 Q_z(T_z^o(z); z) dz$$

Area under the optimal cost curve







Conclusions

- We have shown that *AUC* can be a coherent measure of aggregated classification performance when we consider all scores that have been assigned to data points as thresholds.
- We have also strengthened the connection between ROC plots and cost plots, by visualising *AUC* in cost space as an average of cost lines.
- Instance-uniform threshold selection is realistic in cases where the deployment operating condition is unknown and no validation data is available to set the threshold. Our current work is aimed at threshold selection methods that do take the operating condition into account when it is known.