Brier Curves: a New Cost-Based Visualisation of Classifier Performance

J. Hernández-Orallo[1] and P. Flach[2] and C. Ferri[1]

[1]DSIC, UPV, València, Spain.
{jorallo,cferri}@dsic.upv.es

[2]Intelligent Systems Laboratory, University of Bristol, UK
Peter.flach@bristol.ac.uk

The 28th International Conference on Machine Learning.
ICML 2011

June 28 - July 2 , 2010
Seattle, USA

Outline

Introduction

Methods for evaluating classier performance

- Numerical
  - Usually represent the average or expected performance across a set of operating conditions
- Graphical
  - Especially useful when there is uncertainty about the misclassification costs or the class distribution
    - Can present a classifier's actual performance for a wide variety of different operating conditions

Graphical representations and tools for classifier evaluation

- ROC Curves and isometrics
- DET Curves
- Lift Charts
- Cost Curves

Some of these visualise two performance metrics as a function of an implicit operating condition while others have the operating condition explicitly on the x-axis, and a single performance metric on the y-axis.

ROC Space

- Draw the misclassification rate of one class (negative) on the x-axis and the accuracy of the other class (positive) on the y-axis
- Concentrate on ranking performance
- Ignores the magnitude of the scales

Cost Curves

- Represent the performance of the ROC convex hull of a classifier
  - This is a typically optimistic (and frequently unrealistic) assessment of a classifier
- Ignore the magnitude of the scores
- Draw loss on the y-axis against operating condition on the x-axis
- Visualise classification performance

ROC Space vs Cost Space

- Line segments in ROC space correspond to points in cost space and points in ROC space correspond to line segments in cost space
- The convex hull of a ROC curve corresponds to the lower envelope of the cost lines in cost space.

This paper introduces a new curve to graphically understand and assess classifiers.

- We assume that the classifier scores are posterior class probabilities
  - This provides a natural way of choosing the thresholds.
- This new curve depends on the quality of the probability estimates, and it shows the performance for the full range of operating conditions.
  - We can choose and discard classifiers depending on the operating conditions but we can also combine classifiers in order to obtain a lower overall loss.

Brier Score, ROC Curves and Cost Curves

Brier Score

- The Brier score is a well-known evaluation measure for probabilistic classifiers (Mean Squared Error or MSE loss) :

$$BS \triangleq \frac{1}{n} \sum_{i=1}^{n} (s_i - y_i)^2$$

- Where s_i is the score predicted for example i and y_i is the true class for example i.

$$BS = \pi_0 BS_0 + \pi_1 BS_1.$$

ROC Curves

- 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

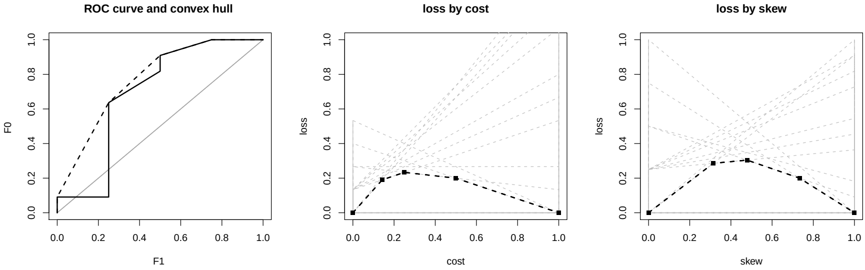
- A cost plot (Drummond & Holte) has loss

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

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:

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



Example

Scores	0.05	0.15	0.16	0.18	0.20	0.20	0.45	0.55
Classes	0	1	0	0	0	0	0	0
Scores	0.70	0.70	0.70	0.85	0.90	0.90	0.95	
Classes	0	1	0	0	1	0	1	

Brier Curves

Optimal Cost Curves

- Optimal cost curves assume that we set thresholds optimally
  - Thresholds that are optimal on a validation set may not carry over to a new test set.

Probabilistic threshold choice

- A natural way of setting the threshold for a probabilistic classifier.
  - Thresholds are set equal to the operating condition (cost proportion or skew).
- The probabilistic threshold choice method sets the threshold:

$$T_c^p(c) \triangleq c$$

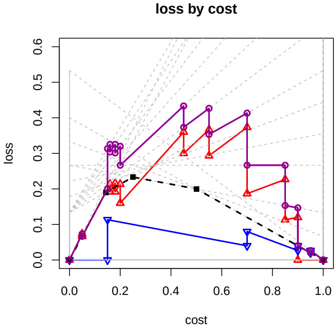
Brier Curves

- The *Brier curve* is defined as a plot of loss against operating condition using the probabilistic threshold choice method.
- If the operating condition is determined by cost proportion the Brier curve is defined by

$$\begin{aligned} BC_c(c) &\triangleq Q_c(T^p_c(c); c) = Q_c(c; c) \\ &= 2c\pi_0(1 - F_0(c)) + 2(1 - c)\pi_1 F_1(c) \end{aligned}$$

- A Brier curve for skew is defined by

$$\begin{aligned} BC_z(z) &\triangleq Q_z(T^p_z(z); z) = Q_z(z; z) \\ &= z(1 - F_0(z)) + (1 - z)F_1(z) \end{aligned}$$

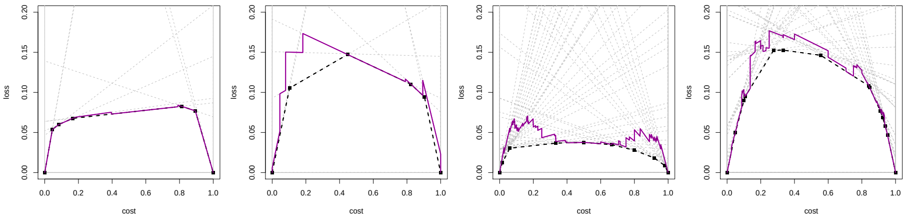


Scores	0.05	0.15	0.16	0.18	0.20	0.20	0.45	0.55
Classes	0	1	0	0	0	0	0	0

Scores	0.70	0.70	0.70	0.85	0.90	0.90	0.95	
Classes	0	1	0	0	1	0	1	

Brier Curves

- Top curve is the Brier Curve
- BC_0 blue line and BC_1 red line.
- Cost lines in thin dashed lines.



Brier curves of a real example

- Brier and optimal cost curves for two J48 classifiers evaluated on training and test sets both sampled from the credit rating UCI dataset.
    - TL: Pruned tree on training set (*AUC*: 0.937, *AUCH*: 0.937, *BS*: 0.068).
    - TR: Pruned tree on test set (*AUC*: 0.887, *AUCH*: 0.894, *BS*: 0.126).
    - BL: Unpruned tree on training set (*AUC*: 0.985, *AUCH*: 0.988, *BS*: 0.042).
    - BR: Unpruned tree on test set (*AUC*: 0.893, *AUCH*: 0.904, *BS*: 0.126).

Area under the Brier Score

The Area under the Brier Curve is the Brier Score

- The area under the Brier curve represents the expected loss averaged over the whole operating range.

$$L_c \triangleq \int_0^1 BC_c(c)dc = \int_0^1 Q_c(c;c)dc = \int_0^1 2\{c\pi_0(1-F_0(c)) + (1-c)\pi_1 F_1(c)\}dc$$

Theorem

- *The area under the Brier curve for cost proportions is equal to the Brier score.*

$$L_c \triangleq \int_0^1 BC_c(c)dc = BS$$

The Area under the Brier Curve is the Brier Score

- We state the corresponding result for skews.

$$L_z \triangleq \int_0^1 BC_z(z)dz = \int_0^1 Q_z(z;z)dz$$
$$= \int_0^1 \{z(1 - F_0(z)) + (1 - z)F_1(z)\}dz$$

Corollary

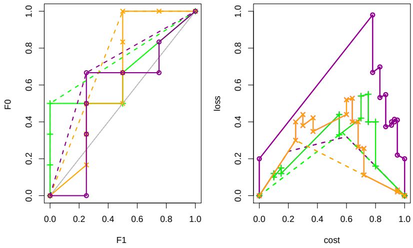
$L_z = (BS_0 + BS_1)/2.$

Properties of Brier Curves

- The BS equivalence of the area lends further credibility to Brier curves
  - The interpretation of *AUC* as the Wilcoxon-Mann-Whitney sum of ranks statistic lends credibility to ROC curves
- Offer a generalisation of the Brier score in the sense that we can investigate 'partial Brier scores' as expected loss over a more restricted range of operating conditions

Brier Curves for Comparing Classifiers

- With ROC analysis we can compare classifiers and identify regions where one classifier dominates other classifiers
- With optimal curves, we can do similarly, assuming optimal choices.
- In the same way, with Brier Curves, given an operating condition on the x-axis we can simply read off on the y-axis which classifier will have lowest loss.
- Given two classifiers A and B we say that A *dominates* B at a cost proportion c iff $Q_c^A(c; c) < Q_c^B(c; c)$.



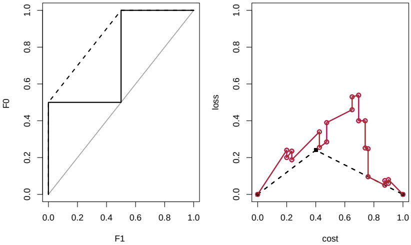
	Class	*A*	*B*	*C*	*D*
e_1	1	0.70 (4..5)	0.60 (5)	0.00 (1)	0.65 (5)
e_2	1	0.80 (7..10)	1.00 (10)	1.00 (9..10)	0.90 (10)
e_3	1	0.80 (7..10)	0.95 (9)	0.93 (7)	0.88 (9)
e_4	1	0.70 (4..5)	0.25 (1..2)	0.91 (6)	0.48 (4)
e_5	0	0.80 (7..10)	0.68 (7)	0.78 (2..3)	0.74 (7)
e_6	0	0.75 (6)	0.64 (6)	0.83 (4)	0.70 (6)
e_7	0	0.10 (1)	0.37 (4)	0.78 (2..3)	0.24 (2)
e_8	0	0.55 (3)	0.30 (3)	0.95 (8)	0.43 (3)
e_9	0	0.80 (7..10)	0.72 (8)	1.00 (9..10)	0.76 (8)
e_{10}	0	0.15 (2)	0.25 (1..2)	0.87 (5)	0.20 (1)

- green lines with '+' points: classifier *A* (*AUC*: 0.667, *AUCH*: 0.750, *BS*: 0.244);
- orange lines with 'x' points: classifier *B* (*AUC*: 0.646, *AUCH*: 0.750, *BS*: 0.240);
- magenta lines with 'o' points: classifier *C* (*AUC*: 0.563, *AUCH*: 0.708, *BS*: 0.558).

Brier Curves for Combining Classifiers

Dominance in performance graphics

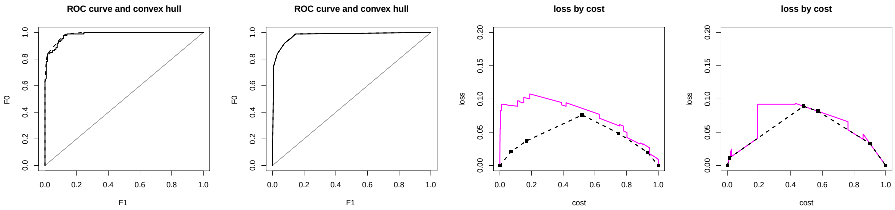
- With ROC analysis we can combine classifiers, or modify a classifier in a given operating range, in order to improve performance.
  - Concavities in the ROC curve of a scoring classifier can be repaired by randomising or inverting the ranking in the corresponding operating range
- Brier curves open up new ways of combining classifiers
  - Make a random choice between two probabilistic classifiers for each prediction
  - Average the predicted probabilities of the classifiers
  - Hybrid classifier: we can construct a hybrid classifier AB, which uses A's predictions if the cost proportion is in either interval $[0.1, 0.5]$ or $[0.55, 0.65]$ and B's predictions otherwise.



	Class	A	B	D
e_1	1	0.70 (4..5)	0.60 (5)	0.65 (5)
e_2	1	0.80 (7..10)	1.00 (10)	0.90 (10)
e_3	1	0.80 (7..10)	0.95 (9)	0.88 (9)
e_4	1	0.70 (4..5)	0.25 (1..2)	0.48 (4)
e_5	0	0.80 (7..10)	0.68 (7)	0.74 (7)
e_6	0	0.75 (6)	0.64 (6)	0.70 (6)
e_7	0	0.10 (1)	0.37 (4)	0.24 (2)
e_8	0	0.55 (3)	0.30 (3)	0.43 (3)
e_9	0	0.80 (7..10)	0.72 (8)	0.76 (8)
e_{10}	0	0.15 (2)	0.25 (1..2)	0.20 (1)

ROC curve and Brier curve of classifier D which predicts the average of the probabilities predicted by classifiers A and B (AUC: 0.750, $AUCH$: 0.875, BS: 0.231).

Brier Curves and Calibration



ROC curves and Brier curves for a Naive Bayes classifier on the vote UCI dataset before and after PAV calibration (50% train, 50% test).

- Top left: Non-calibrated ROC curve.
- Top right: PAV-calibrated ROC curve.
- Bottom left: Non-calibrated Brier curve.
- Bottom right: PAV-calibrated Brier curve.

Brier Curves and Calibration

- The Brier curve clearly locates the loss due to bad calibration between scores 0 and 0.5, although this has little effect on the ranking quality.
- Calibration improves both curves.
    - With ROC curves, calibration has the potential to fix the concavities of the curve.
    - With Brier curves it moves the curve closer to the optimal cost curve.
    - We can see where calibration fails.
        - Calibration has failed between 0.2 and 0.4, which corresponds to the strong discontinuity of the slope of the ROC curve.

Conclusions

- Brier Curves:
  - A new graphical tool to understand the performance of classifiers.
  - Are built setting the threshold equal to the operating condition, either cost proportion or skew.
  - Represent the performance of a probabilistic classifier for a range of operating conditions defined by cost proportion or skew.
- ROC curves are useful to represent and analyse rankers, Brier curves are useful to represent probabilistic classifiers.
- Optimal cost curves and Brier curves :
  - Summarise most of the information about the performance of a classifier
  - Allow us to consider different ways of choosing the thresholds, and their resulting performance.

Future Work

- The relationship between confidence intervals for the Brier curve and for the Brier score.
- Brier curves for the improvement of classifiers, (calibration).
- Brier score decomposition and the notion of calibration in the plots.
- Build hybrid classifiers using Brier curves.