Comparing apples and oranges

Reinterpreting common evaluation metrics in classification

Peter A. Flach

Intelligent Systems Laboratory, University of Bristol, United Kingdom

UCL, 17 October 2014

Talk outline

Acknowledgements I

This talk is based on joint work with José Hernández-Orallo and Cèsar Ferri:

- Peter A. Flach, José Hernández-Orallo, and Cèsar Ferri. A coherent interpretation of AUC as a measure of aggregated classification performance.
  In *Proceedings of the 28th International Conference on Machine Learning*, 2011
- José Hernández-Orallo, Peter A. Flach, and Cèsar Ferri. Brier curves: a new cost-based visualisation of classifier performance.
  In *Proceedings of the 28th International Conference on Machine Learning*, 2011
- José Hernández-Orallo, Peter A. Flach, and Cèsar Ferri. A unified view of performance metrics: translating threshold choice into expected classification loss.
  Journal of Machine Learning Research, 13:2813–2869, 2012
- José Hernández-Orallo, Peter A. Flach, and César Ferri. ROC curves in cost space.
  Machine Learning, 93(1):71–91, 2013.
  Special issue ECML-PKDD'13

Acknowledgements II

It was originally inspired by David Hand's critique of AUC as a measure of classification performance:

☞ D. J. Hand. Measuring classifier performance: a coherent alternative to the area under the ROC curve.
Machine Learning, 77(1):103–123, 2009

Many notational conventions are taken from that paper.

What's next?

Classifier

A function that maps instances x from an instance space X to classes y from an output space Y. We will assume binary classifiers, that is, $Y = \{0, 1\}$.

Model

A function $m : X \to \mathbb{R}$ that maps examples to real numbers (scores) on an unspecified scale. We use the convention that higher scores express a stronger belief that the instance is of class 1 (negative), which means that thresholds increase along ROC curves.

Probability estimator

A function $m : X \to [0, 1]$ that maps examples to estimates $\hat{p}(1|x)$ of the probability of example x to be of class 1. Given a predicted score $s = m(x)$ and a threshold t, the instance x is classified in class 1 if $s > t$, and in class 0 otherwise.

True/false positive rate at threshold t

$$F_k(t) = \int_{-\infty}^{t} f_k(s)\,ds = P(s \leq t|k)$$

where $f_k(s)$ is the (true) score density of class $k \in \{0,1\}$ points.

Accuracy, error rate and predicted positive rate

$$\begin{aligned} Acc(t) &= \pi F_0(t) + (1-\pi)(1-F_1(t)) \\ Err(t) &= \pi(1-F_0(t)) + (1-\pi)F_1(t) \\ R(t) &= \pi F_0(t) + (1-\pi)F_1(t) \end{aligned}$$

where π is the proportion of positives.

ROC curve for scoring classifiers

The ROC curve is defined as a plot of $F_1(t)$ on the x-axis against $F_0(t)$ on the y-axis, with both quantities monotonically non-decreasing with increasing t (remember that scores increase with $\hat{p}(1|x)$ and 1 stands for the negative class).

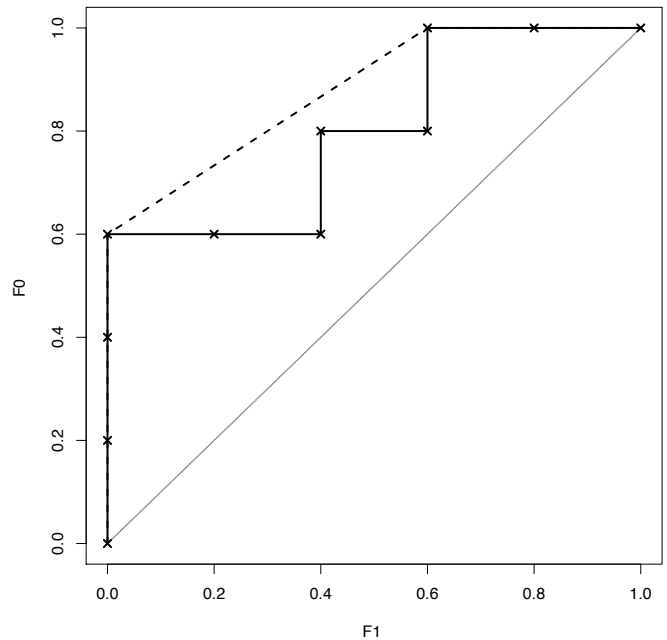


Figure: Empirical ROC curve arising from the test labels (0,0,0,1,1,0,1,0,1,1) with increasing scores (0.13,0.25,0.34,0.45,0.53,0.62,0.71,0.83,0.91,0.95). The dashed line indicates the *convex hull* which arises from considering optimal thresholds only.

AUC

The Area Under the ROC curve (AUC) is defined as:

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

AUC estimates the probability that a random class 1 example receives a higher score than a random class 0 example, and as such is a measure of ranking performance rather than classification performance.

Brier score for probability estimators

$$BS \triangleq \pi \int_0^1 s^2 f_0(s) ds + (1-\pi) \int_0^1 (1-s)^2 f_1(s) ds$$

The Brier score is a proper scoring rule (it is minimised by the true probabilities).

Misclassification costs and skew

$c_k \geq 0, k = 0, 1$ is the cost of misclassifying an example of true class k.
$b = c_0 + c_1$ is the *cost magnitude* ($b = 2$ ensures that loss is commensurate with error rate). $c = c_0 / b$ is the *cost proportion*.

Alternatively, c can be interpreted as a *change* in class distribution from π to $z = \frac{c\pi}{c\pi + (1-c)(1-\pi)}$. We will often refer to c simply as *skew*.

Loss at threshold t and skew c

$$Q(t; c) \triangleq 2\{c\pi(1 - F_0(t)) + (1 - c)(1 - \pi)F_1(t)\}.$$

Expected loss

Given a threshold choice method T and a probability density function over skews w, expected loss is defined as:

$$L_c \triangleq \int_0^1 Q(T(c); c)\, w(c)\, dc$$

Quiz

You have trained a classifier (say naive Bayes, or a decision tree) on a two-class data set with the default threshold (0.5 on the posterior).

You are now told that in the deployment context the skew is c relative to the training context. What do you do?

Quiz

You have trained a classifier (say naive Bayes, or a decision tree) on a two-class data set with the default threshold (0.5 on the posterior).

You are now told that in the deployment context the skew is c relative to the training context. What do you do?

☞ You keep the threshold at 0.5.

Quiz

You have trained a classifier (say naive Bayes, or a decision tree) on a two-class data set with the default threshold (0.5 on the posterior).

You are now told that in the deployment context the skew is c relative to the training context. What do you do?

- ☞ You keep the threshold at 0.5.
- ☞ You set the threshold at c.

Quiz

You have trained a classifier (say naive Bayes, or a decision tree) on a two-class data set with the default threshold (0.5 on the posterior).

You are now told that in the deployment context the skew is c relative to the training context. What do you do?

- ☞ You keep the threshold at 0.5.
- ☞ You set the threshold at c.
- ☞ You set the threshold such that the predicted positive rate is c.

Quiz

You have trained a classifier (say naive Bayes, or a decision tree) on a two-class data set with the default threshold (0.5 on the posterior).

You are now told that in the deployment context the skew is c relative to the training context. What do you do?

- ☞ You keep the threshold at 0.5.
- ☞ You set the threshold at c.
- ☞ You set the threshold such that the predicted positive rate is c.
- ☞ You estimate the optimal threshold for skew c.

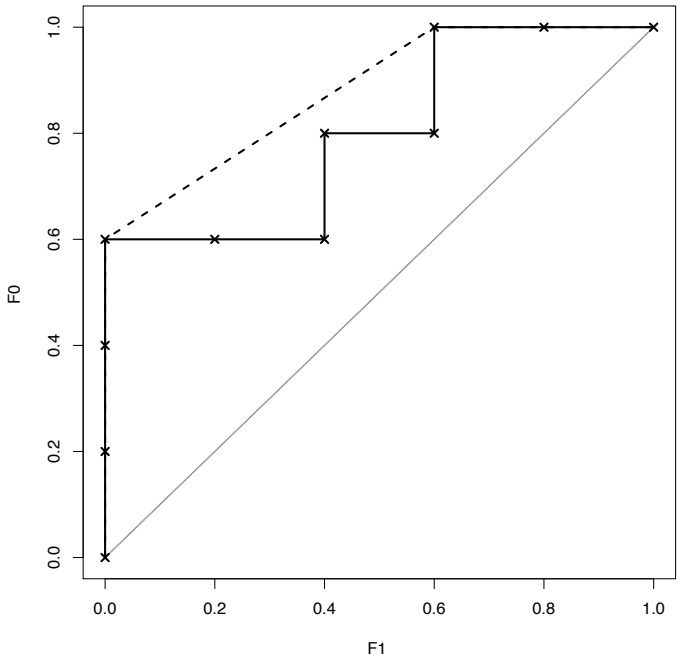
Quiz

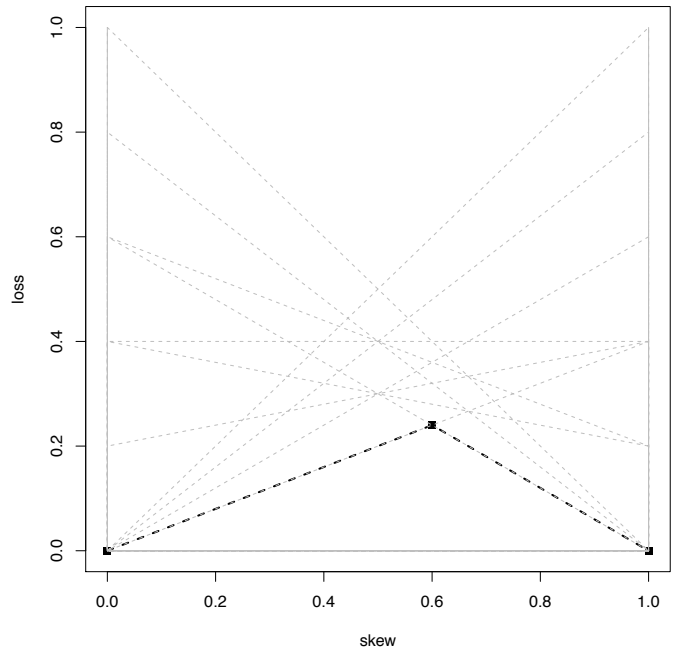
You have trained a classifier (say naive Bayes, or a decision tree) on a two-class data set with the default threshold (0.5 on the posterior).

You are now told that in the deployment context the skew is c relative to the training context. What do you do?

- ☞ You keep the threshold at 0.5.
- ☞ You set the threshold at c.
- ☞ You set the threshold such that the predicted positive rate is c.
- ☞ You estimate the optimal threshold for skew c.
- ☞ None of the above.

What's next?





Cost plots[1] plot loss Q against operating condition c, here shown for fixed thresholds (yielding a *cost line* for each operating point) or optimal thresholds (yielding the *lower envelope* of cost lines). The area under the lower envelope is the expected loss for optimal thresholds assuming uniform skews.

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

The Brier curve

If a probability estimator is trained on a particular class distribution and we want to adapt it to a different skew c it makes sense to set the threshold equal to c, as this would be optimal for the true posterior probability.[a]

Plotting loss against c results in the *Brier curve*, the area underneath which is the Brier score. What this means is that – if we "trust" the probability estimates – expected loss over uniform c is equal to the Brier score.

The Brier curve alternates between following a given cost line corresponding to the ROC operating point, and jumping to a different cost line once c exceeds the score of the next test instance.

---

[a] Charles Elkan. The foundations of cost-sensitive searning.
In *Proceedings of the Seventeenth International Conference on Artificial Intelligence (IJCAI-01)*, pages 973–978, San Francisco, CA, 2001

Example Brier curve

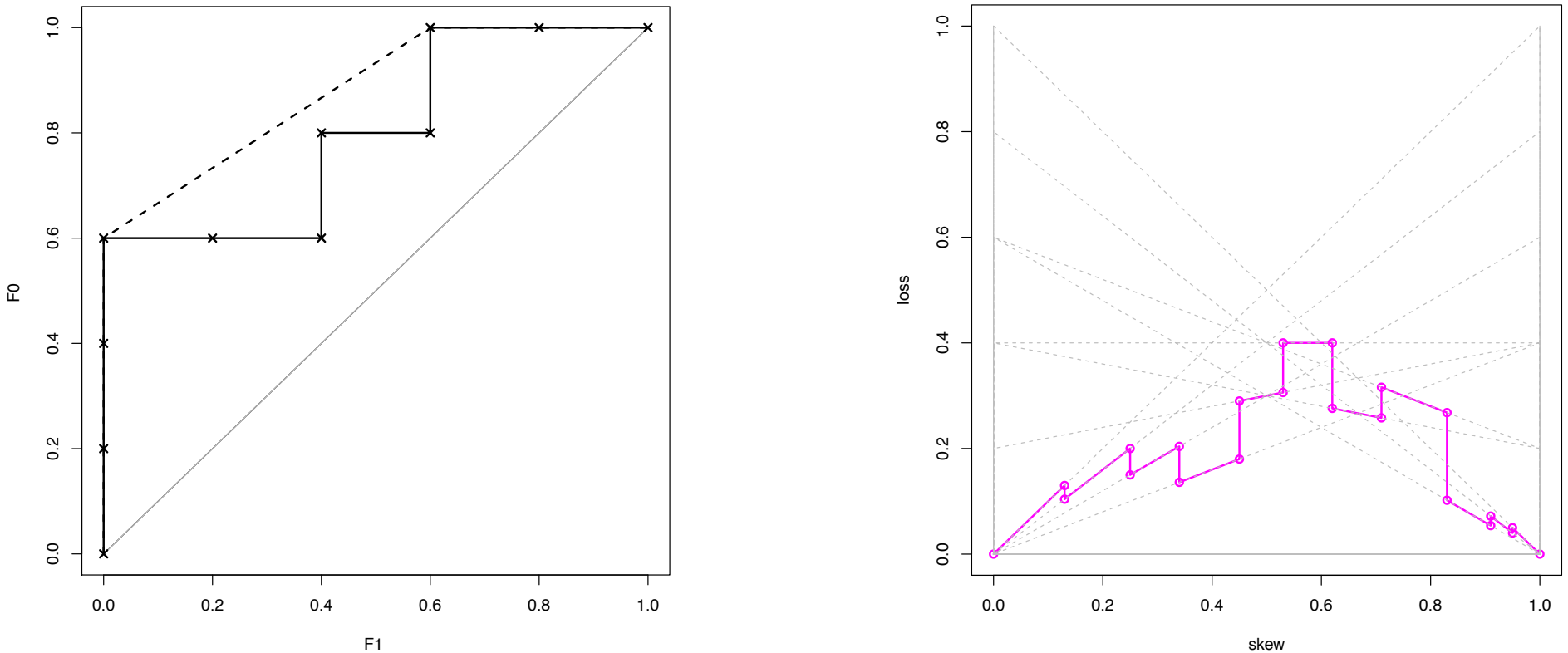
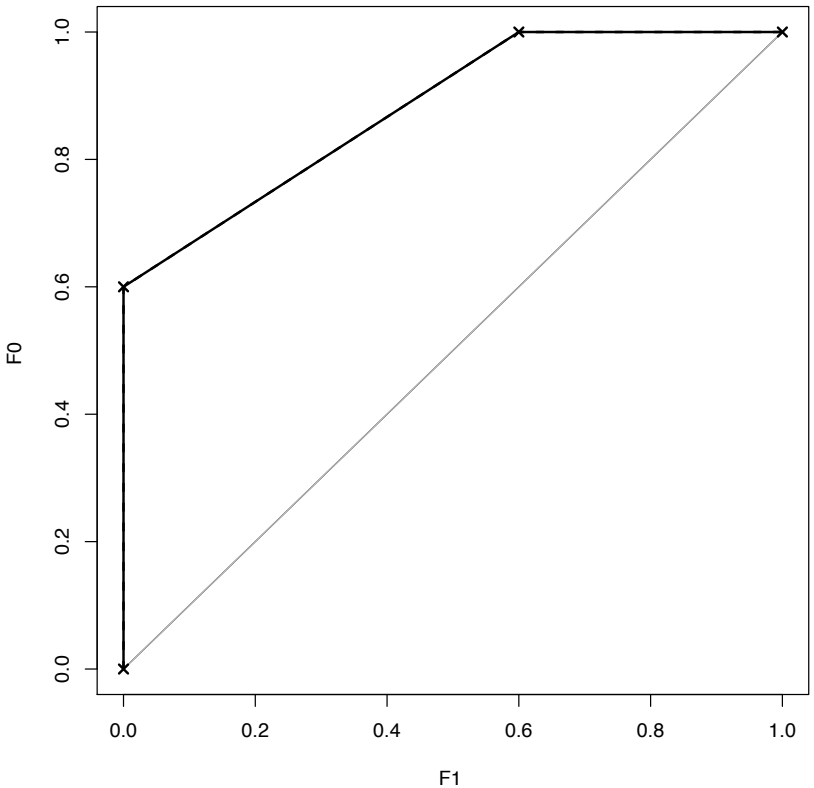


Figure: (Left) ROC curve arising from the test labels (0,0,0,1,1,0,1,0,1,1) with scores (0.13,0.25,0.34,0.45,0.53,0.62,0.71,0.83,0.91,0.95).
(Right) The Brier curve often jumps to non-optimal cost lines. The area under the Brier curve is the Brier score.

Using the convex hull



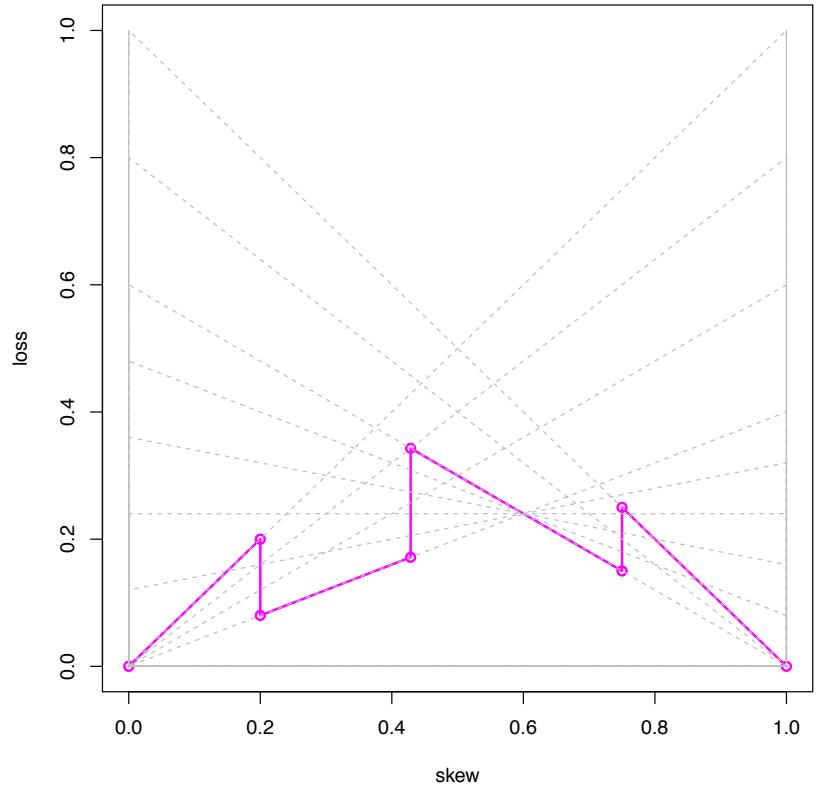
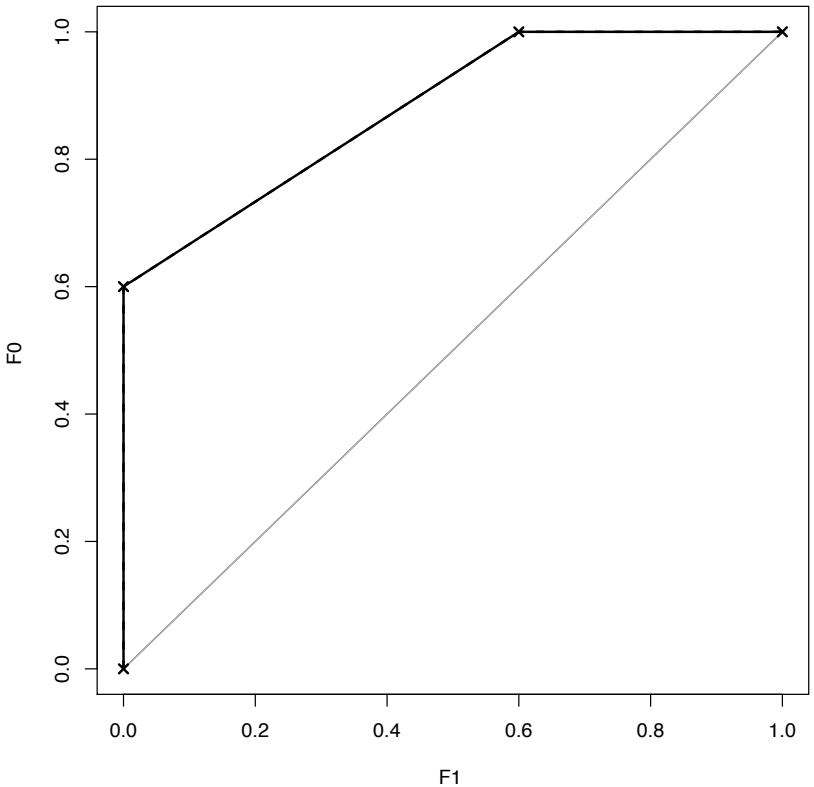


Figure: (Left) Convex ROC curve arising from the test labels (0,0,0,1,1,0,1,0,1,1) with scores (0.2,0.2,0.2,0.43,0.43,0.43,0.43,0.43,0.75,0.75).
(Right) The Brier curve gets closer to the lower envelope, so this model has lower expected loss. It can be further reduced through calibration.

Calibrated scores



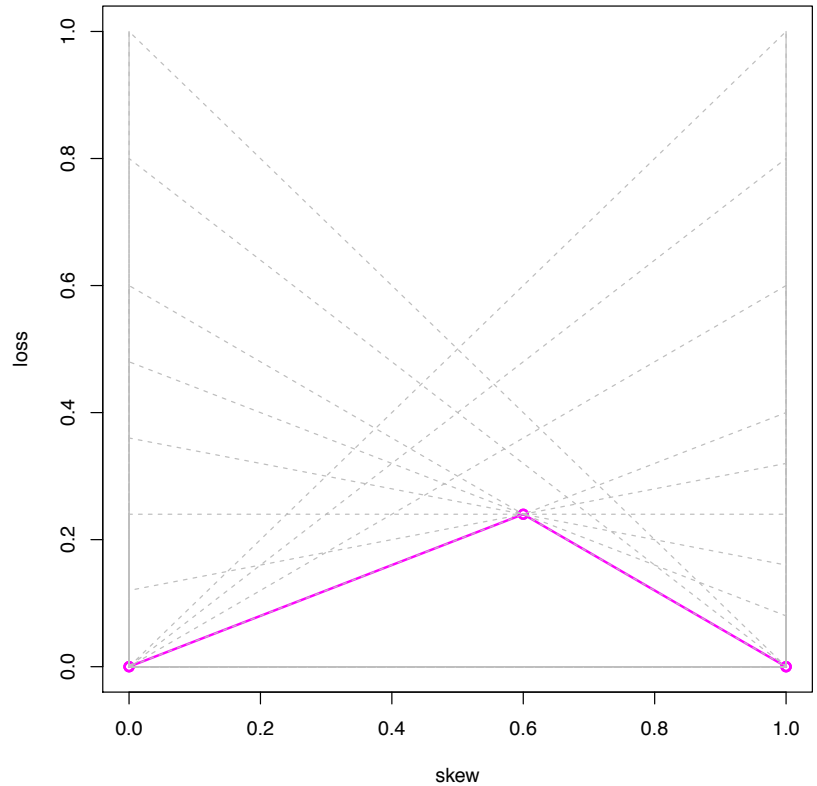


Figure: (Left) Convex ROC curve arising from the test labels (0,0,0,1,1,0,1,0,1,1) with calibrated scores (0.0,0.0,0.0,0.6,0.6,0.6,0.6,0.6,1.0,1.0).
(Right) The Brier curve coincides with the lower envelope, indicating that this model's loss cannot be further reduced.

A realistic example

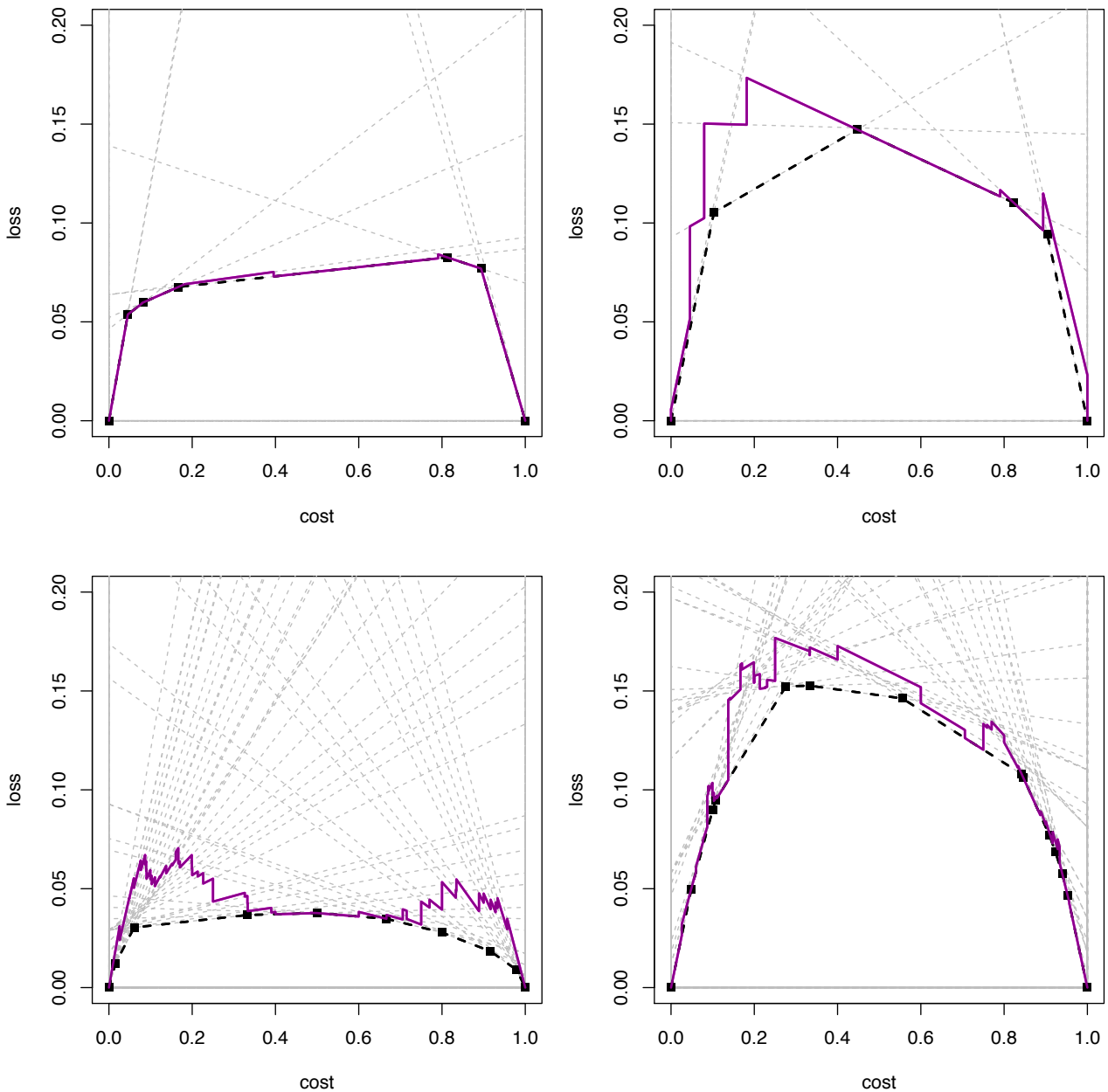
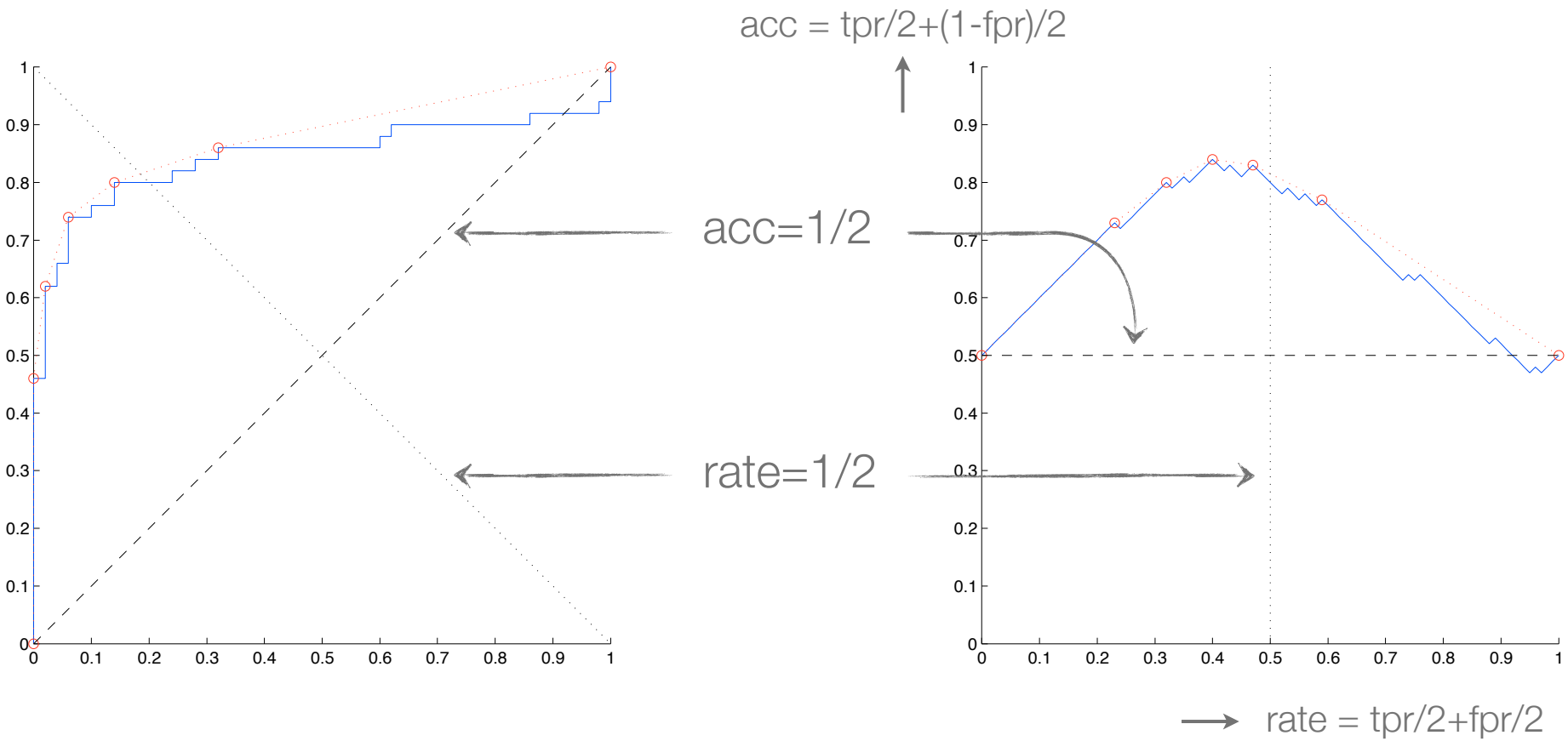


Figure: Top: pruned decision tree, bottom: unpruned tree. Left: training set, right: test set.

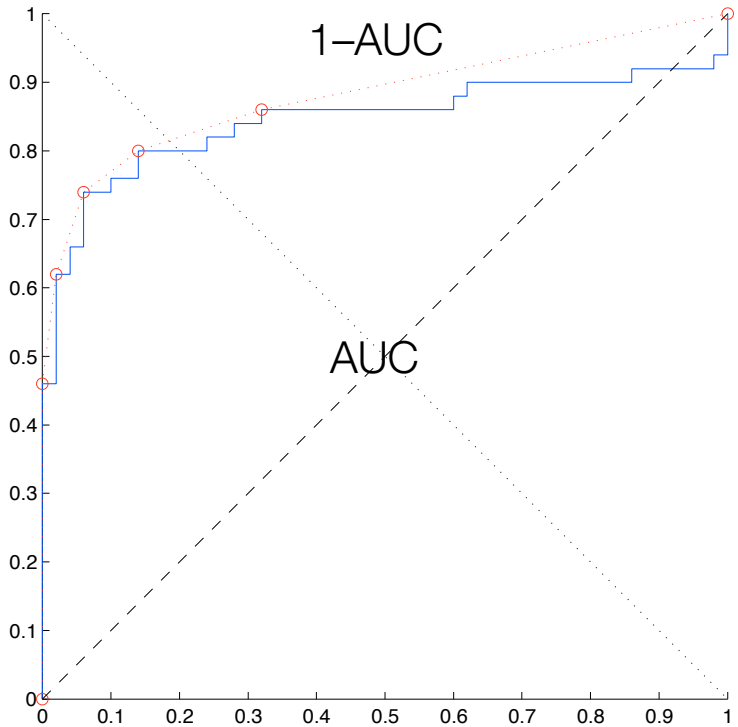
From ROC curve to ROL curve I

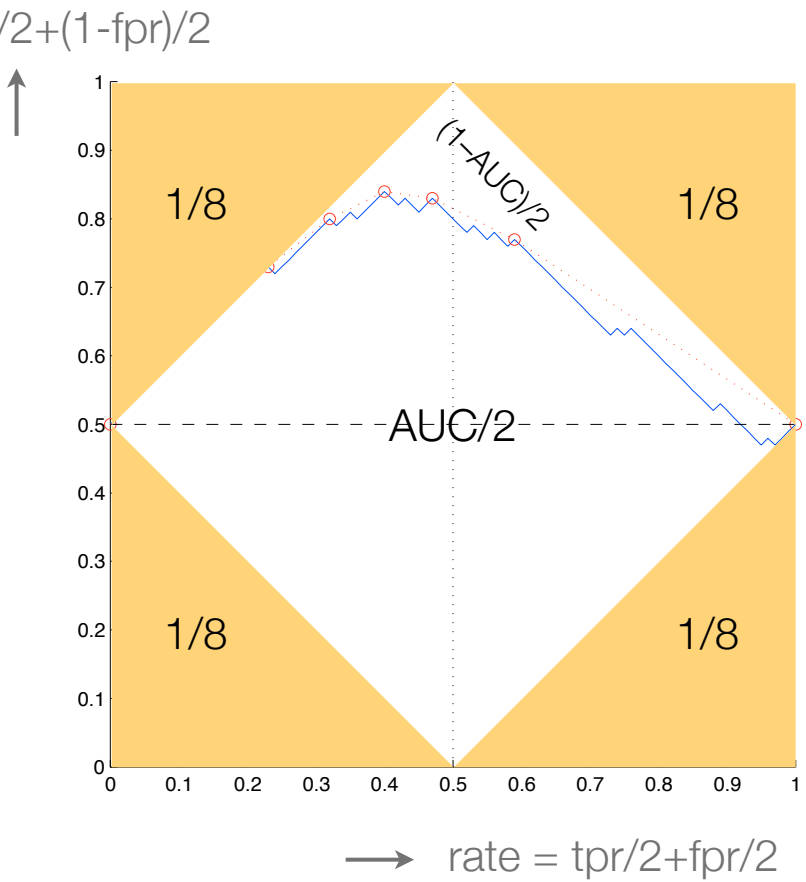
Plotting accuracy against rate for balanced classes ($\pi = 1/2$)



From ROC curve to ROL curve II

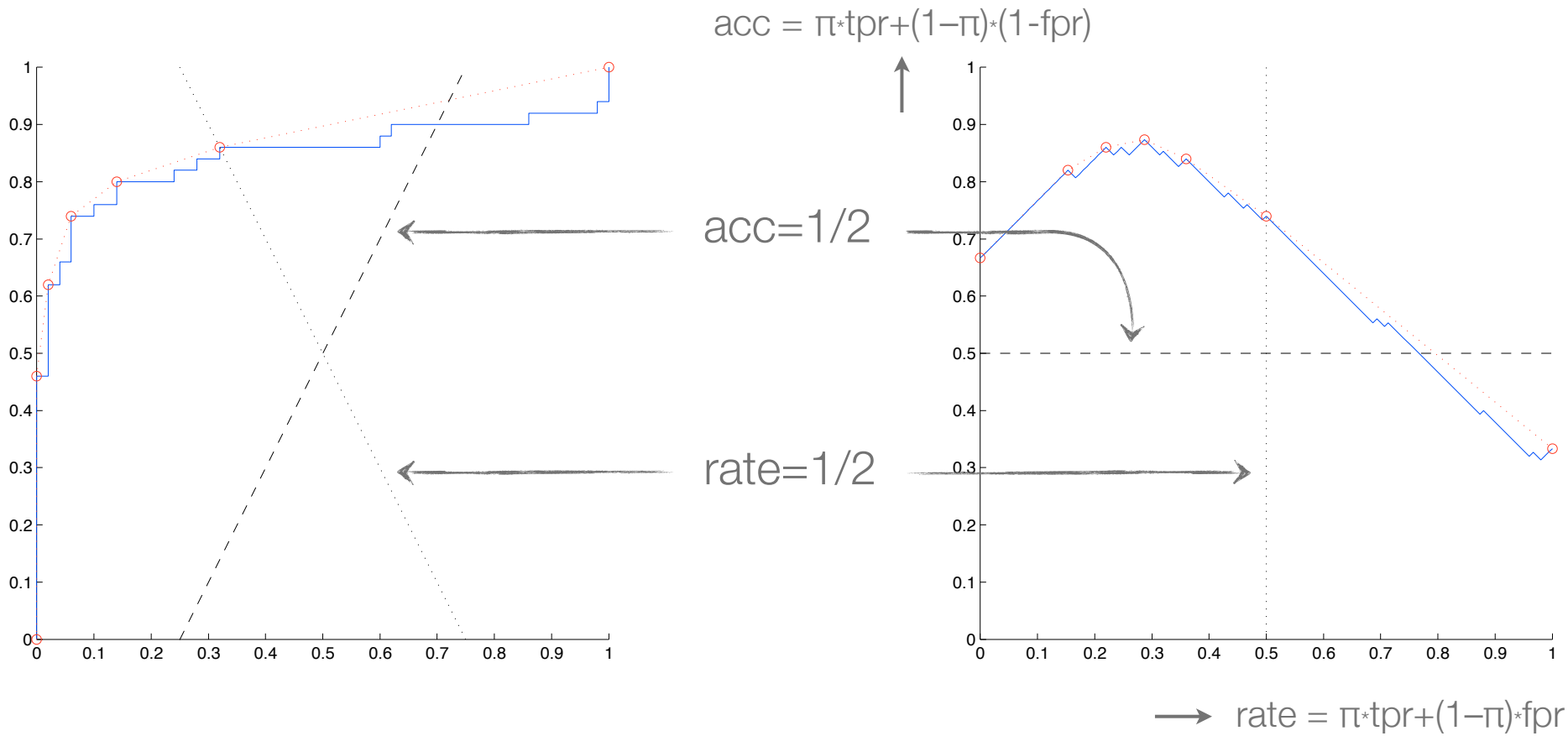
Expected loss for uniform rate and balanced classes is
$(1-AUC)/2+1/4=(1-2AUC)/4+1/2$





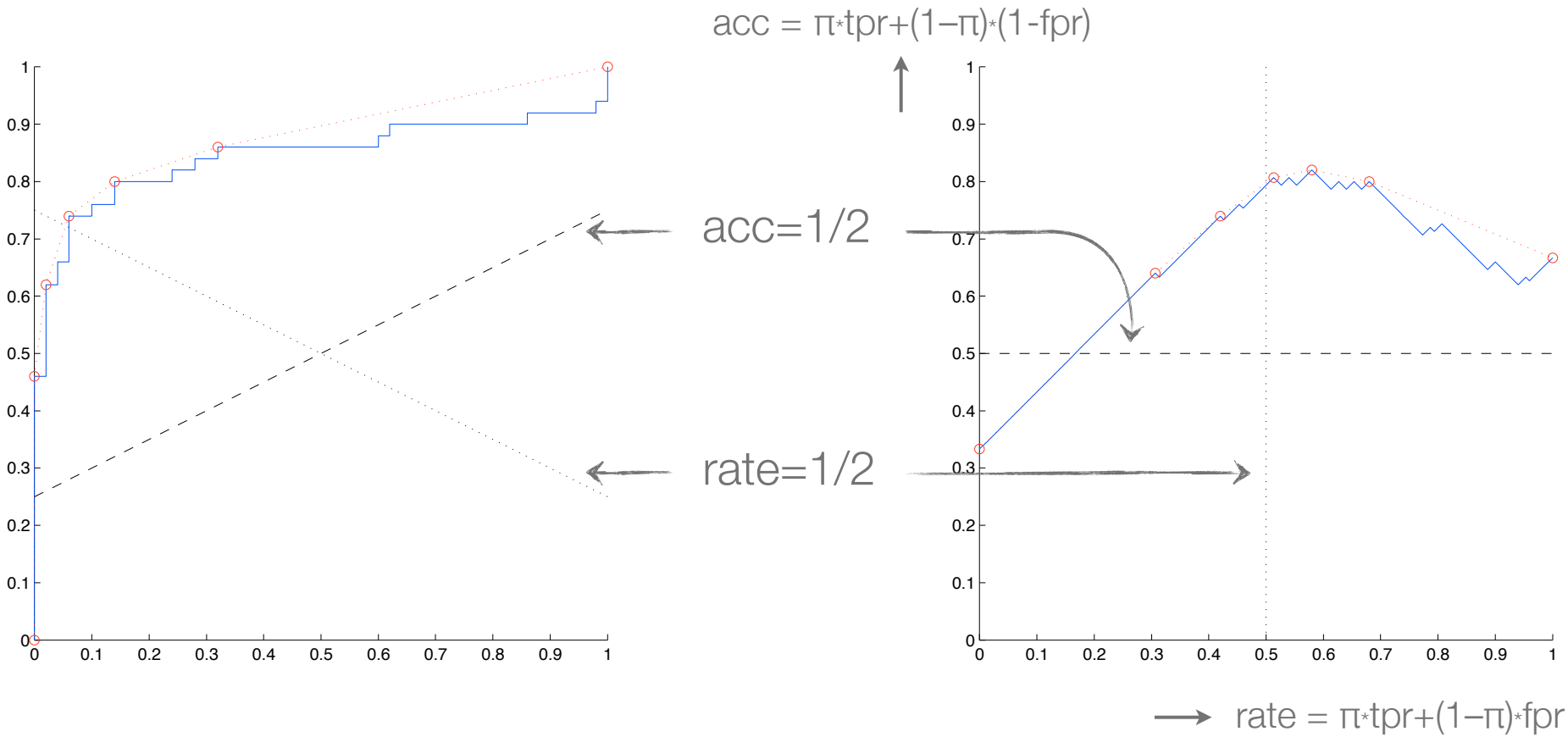
From ROC curve to ROL curve III

More negatives than positives ($\pi < 1/2$)



From ROC curve to ROL curve IV

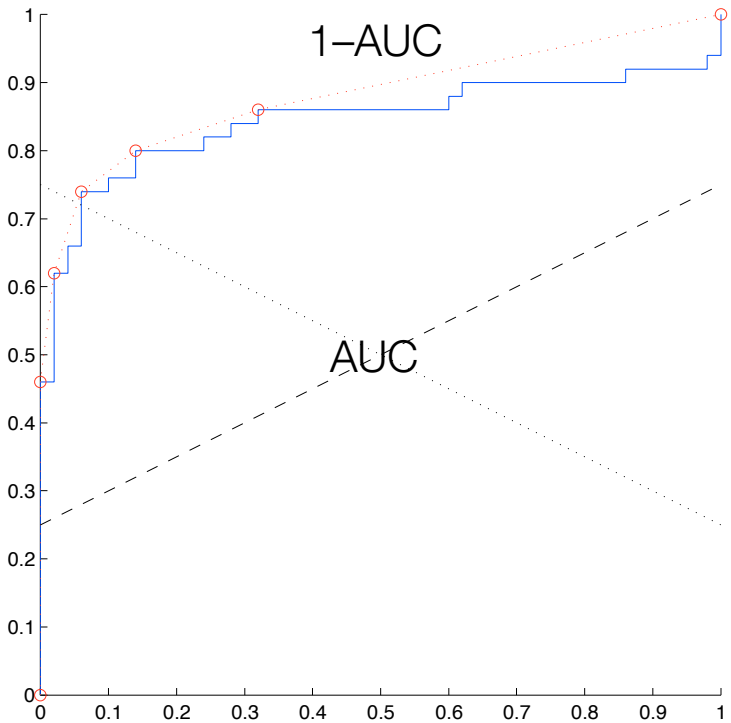
More positives than negatives ($\pi > 1/2$)

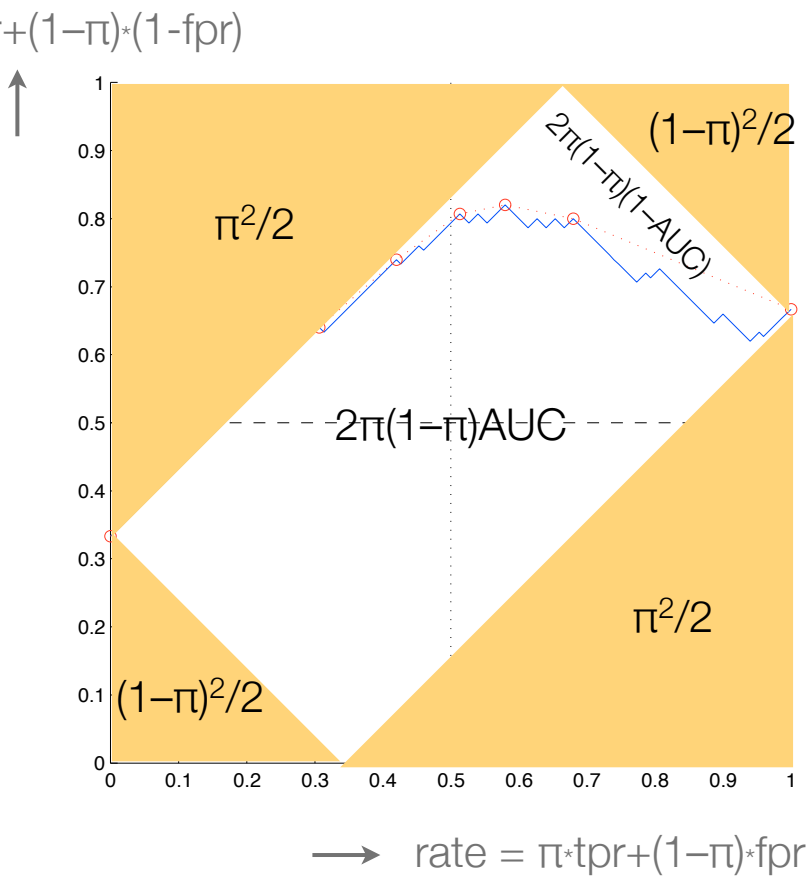


From ROC curve to ROL curve V

Expected loss for uniform rate is

$$2\pi(1-\pi)(1-AUC)+\pi^2/2+(1-\pi)^2/2=\pi(1-\pi)(1-2AUC)+1/2$$





Rate-driven thresholds

Expected loss for uniform skews when setting the rate equal to c is linearly related to AUC as follows:

$$L^{rd}_{U(c)} = \pi(1-\pi)(1-2AUC) + 1/3$$

This vindicates the use of AUC as a performance measure in classification and provides an answer to David Hand's critique.

The rate-driven threshold choice method can probably still be improved as for uniform c it leads to an expected rate of $1/2$, whereas it makes more sense to set thresholds such that the expected rate is π.

Rate-driven cost curve

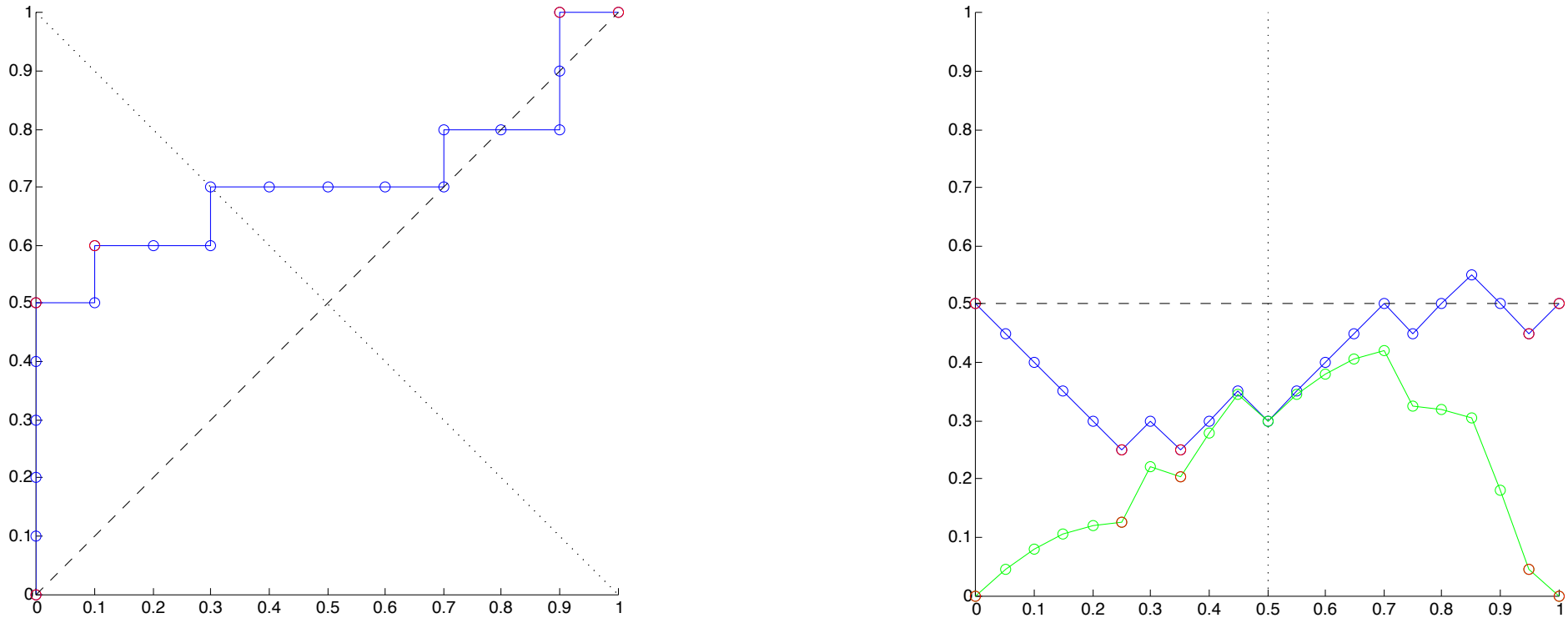


Figure: (Left) Empirical ROC curve. (Right) Rate-uniform cost curve in blue and rate-driven cost curve in green. The rate-driven threshold choice method is able to take advantage of knowing the operating condition, leading to a lower expected loss; the area between the two curves is 1/6.

Brier curve and rate-driven cost curve

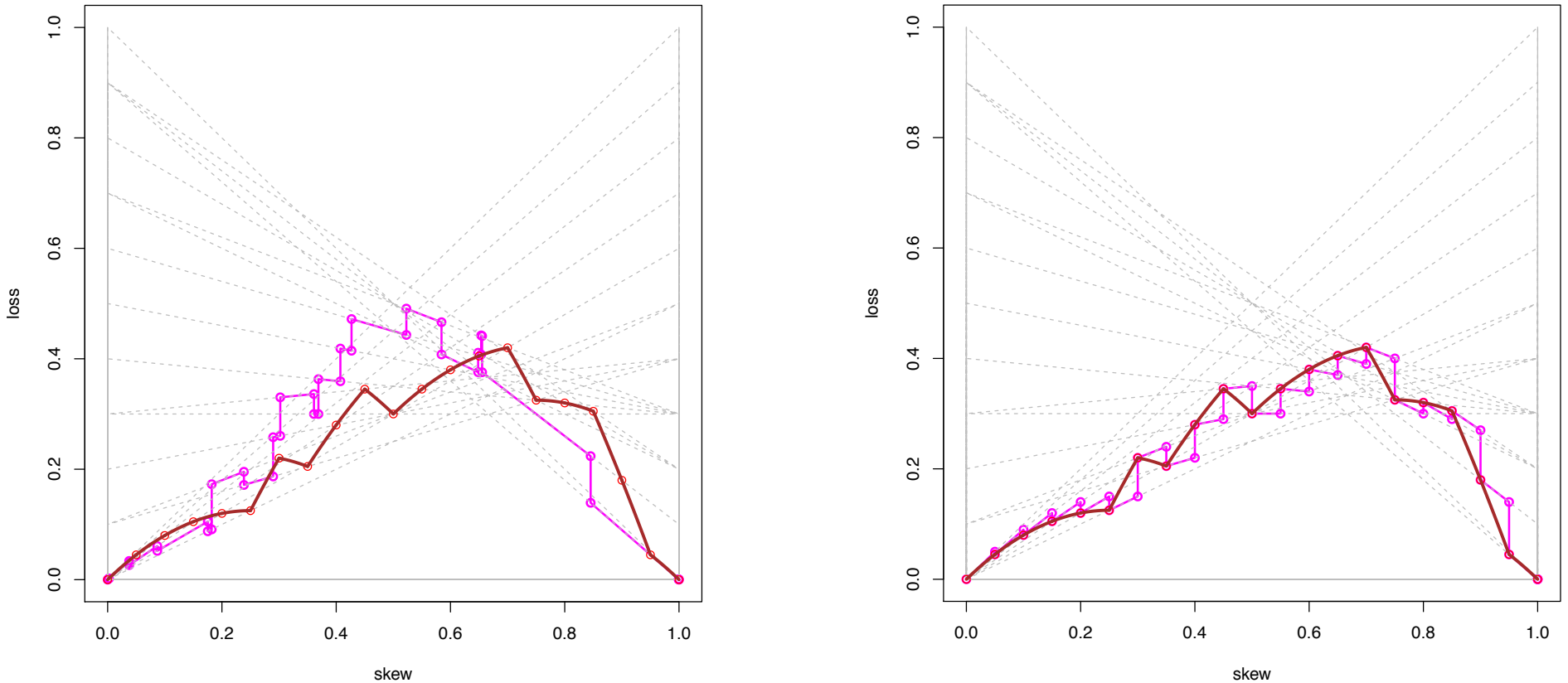


Figure: (Left) Brier curve in **violet** versus rate-driven curve in **red**. (Right) If scores are evenly spaced the two methods select the same thresholds – in the limit the area under the two curves is the same, which establishes the first known connection between Brier score and AUC.

What's next?

A general view of classifier calibration

Definition

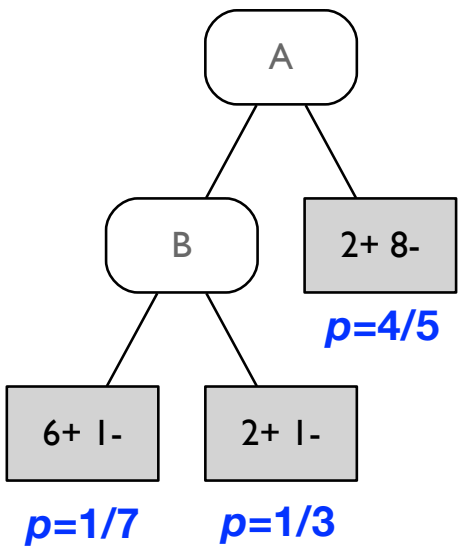
Classifier scores are well-calibrated for performance measure Q and context C if

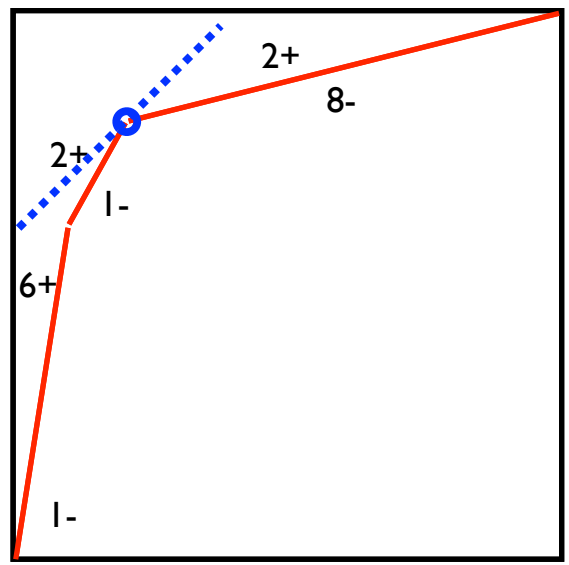
- ☞ the threshold $1/2$ is Q-optimal in context C;
- ☞ more generally, the threshold c is Q-optimal in context C', where c is the context change from C to C'.

Example

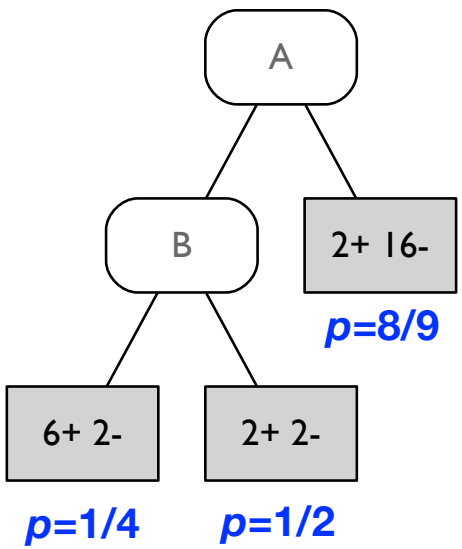
True posterior probabilities are well-calibrated for accuracy, given the class prior as context.

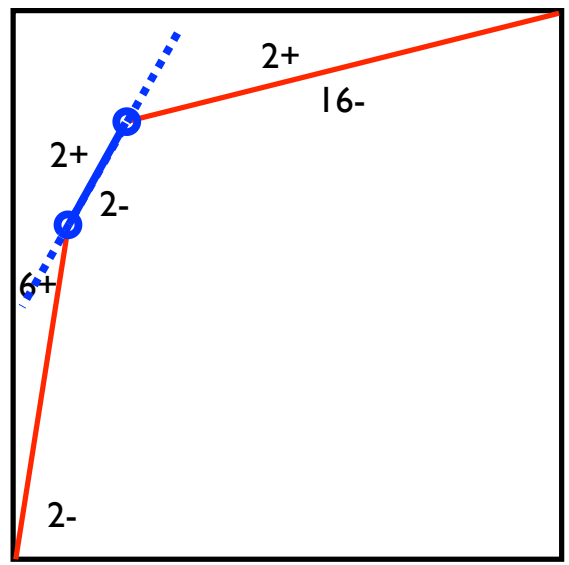
Example ($\pi = 1/2$)





Example ($\pi = 1/3$)





Trading off true and false positive rate

The results so far are derived using an accuracy-based loss function:

$$Q(t;\pi,b,c) = b\{c\pi(1-F_0(t)) + (1-c)(1-\pi)F_1(t)\}$$

Setting this equal to some constant q and solving for F_0 gives the equation for an accuracy-based *loss isometric* (line of constant loss in ROC space):

$$F_0(q;\pi,b,c) = \frac{1-\pi}{\pi}\frac{1-c}{c}F_1(q;\pi,b,c) + 1 - \frac{q/b}{c\pi}$$

These isometrics have constant slope $\frac{1-\pi}{\pi}\frac{1-c}{c}$. The ROC curve has slope $f_0(t)/f_1(t)$ which at an optimal point is equal to the slope of the isometric, from which we derive

$$c = \frac{(1-\pi)f_1(t)}{\pi f_0(t) + (1-\pi)f_1(t)}$$

This is therefore the accuracy-calibrated score of the classifier.

The F-measure and its isometrics I

The F-measure has been introduced as an alternative to accuracy in order to deal with situations with many more negatives than positives but true negatives do not add value.

$$FM \triangleq \frac{TP}{TP + (FP + FN)/2} = \frac{2}{1/prec + 1/rec} = \frac{2\,prec \cdot rec}{prec + rec}$$

In terms of true and false positive rate the corresponding loss is then

$$FQ(t) = \frac{\pi(1 - F_0(t)) + (1 - \pi)F_1(t)}{2\pi F_0(t) + \pi(1 - F_0(t)) + (1 - \pi)F_1(t)}$$

F-measure isometrics in ROC space are straight lines with varying slope, rotating around the (virtual) point $(F_1 = -\pi/(1-\pi), F_0 = 0)$.

The F-measure and its isometrics II

Incorporating a skew parameter c as before, the F-measure loss can be expressed as

$$FQ(t;c) \triangleq \frac{c\pi(1-F_0(t)) + (1-c)(1-\pi)F_1(t)}{\pi F_0(t) + c\pi(1-F_0(t)) + (1-c)(1-\pi)F_1(t)}$$

Setting this equal to some value q and solving for F_0 gives the equation of an F-measure isometric:

$$F_0(q;c) = \frac{1-\pi}{\pi}\frac{(1-c)(1-q)}{c+(1-c)q}F_1(q;c) + \frac{c(1-q)}{c+(1-c)q}$$

We see that the slope of an F-measure isometric depends on the loss q: the added value of an increased true positive rate or a decreased false positive rate is not constant throughout ROC space, as with accuracy-based loss.

The F-measure and its isometrics III

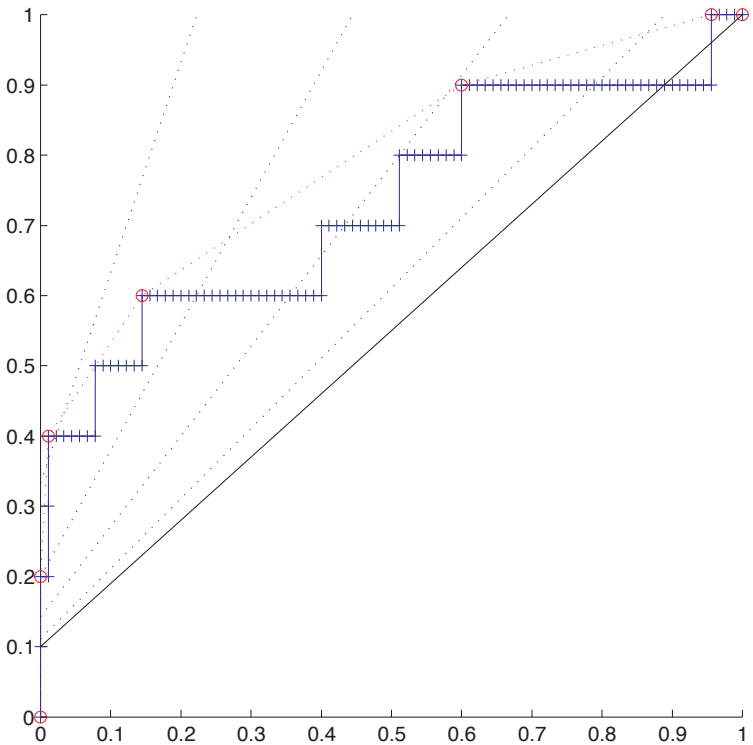


Figure: Example empirical ROC curve with F-measure isometrics. The ROC-convex hull consists of the seven points indicated in **red**. There are 10 positives and 90 negatives so the isometrics rotate around $F_1 = -1/9$. From top to bottom the F-measure values follow the harmonic series 1 (through ROC heaven), $1/2$, $1/3$, $1/4$ and $1/5$. The solid isometric corresponds to the default all-positive classifier.

FROC curves I

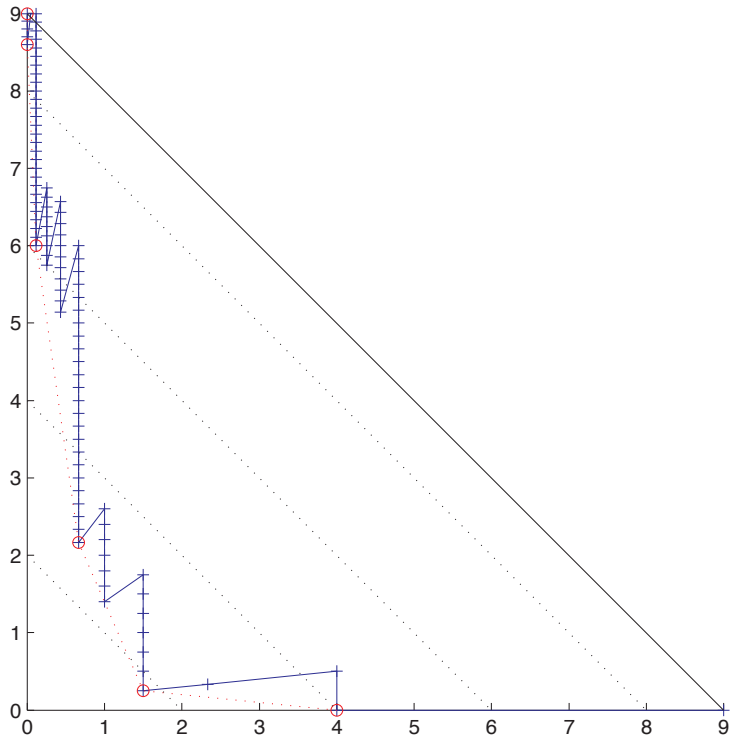


Figure: FROC plot with $G_y = 1/prec - 1 = FP/TP$ on the y-axis and $G_x = 1/rec - 1 = FN/TP$ on the x-axis. FROC heaven is in the origin and F-measure isometrics are parallel lines with slope -1.

FROC curves II

FROC isometrics are given by

$$G_x + G_y = \frac{1}{rec} - 1 + \frac{1}{prec} - 1 = \frac{2}{FM} - 2 = 2\frac{FQ}{1-FQ}$$

The loss on the right is a monotonic transformation of FQ, similar to a transformation of probabilities into odds.

We can incorporate skews c as follows:

$$2\frac{FQ}{1-FQ} = 2\frac{c\pi(1-F_0) + (1-c)(1-\pi)F_1}{\pi F_0} = 2cG_x + 2(1-c)G_y$$

from which we see that FQ isometrics have slope $-c/(1-c)$.

F-cost curves

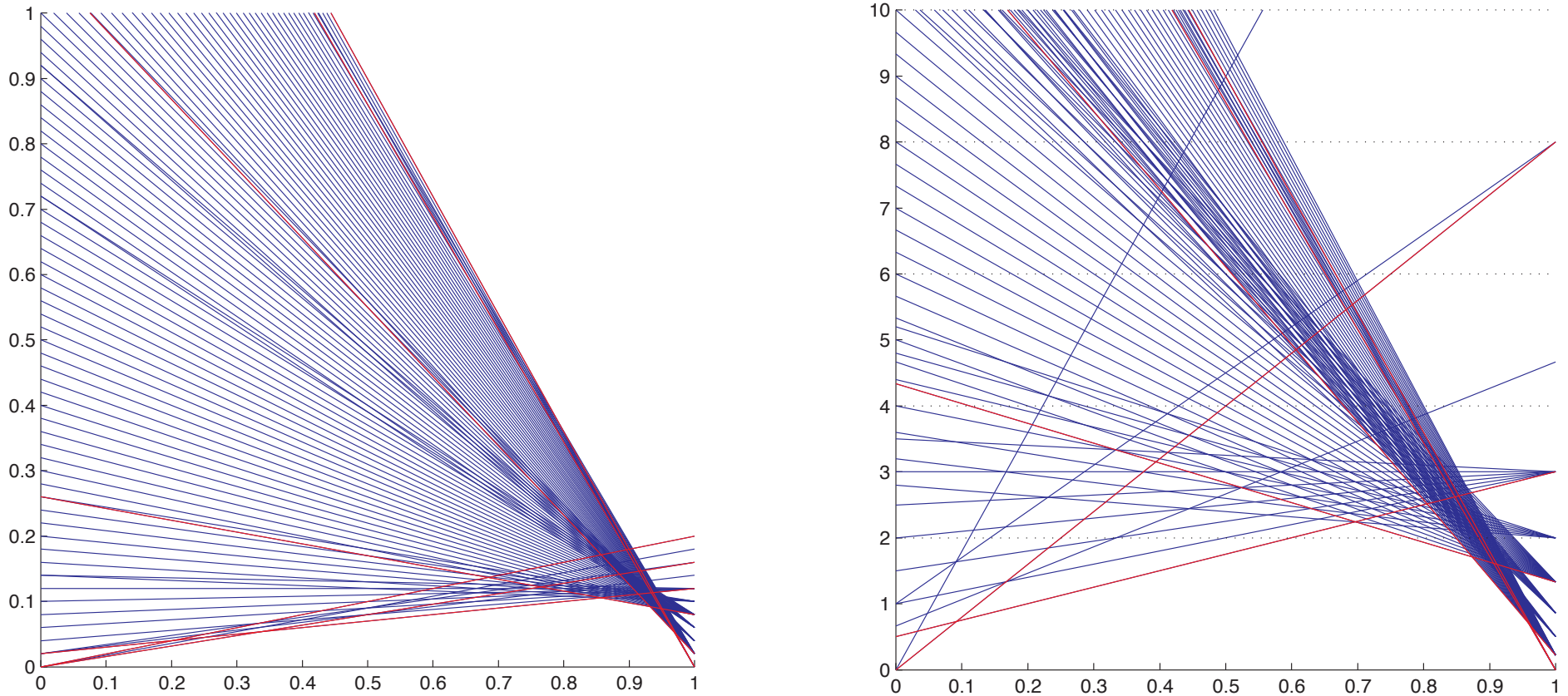
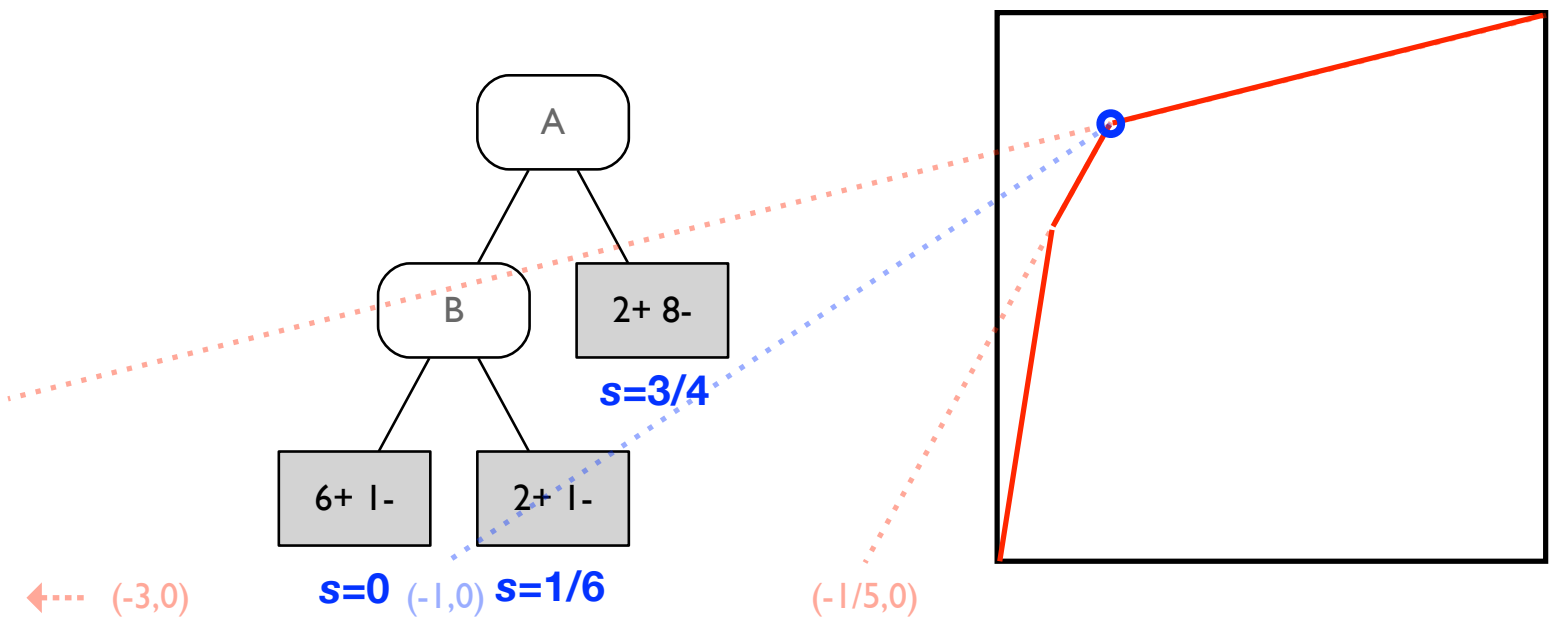
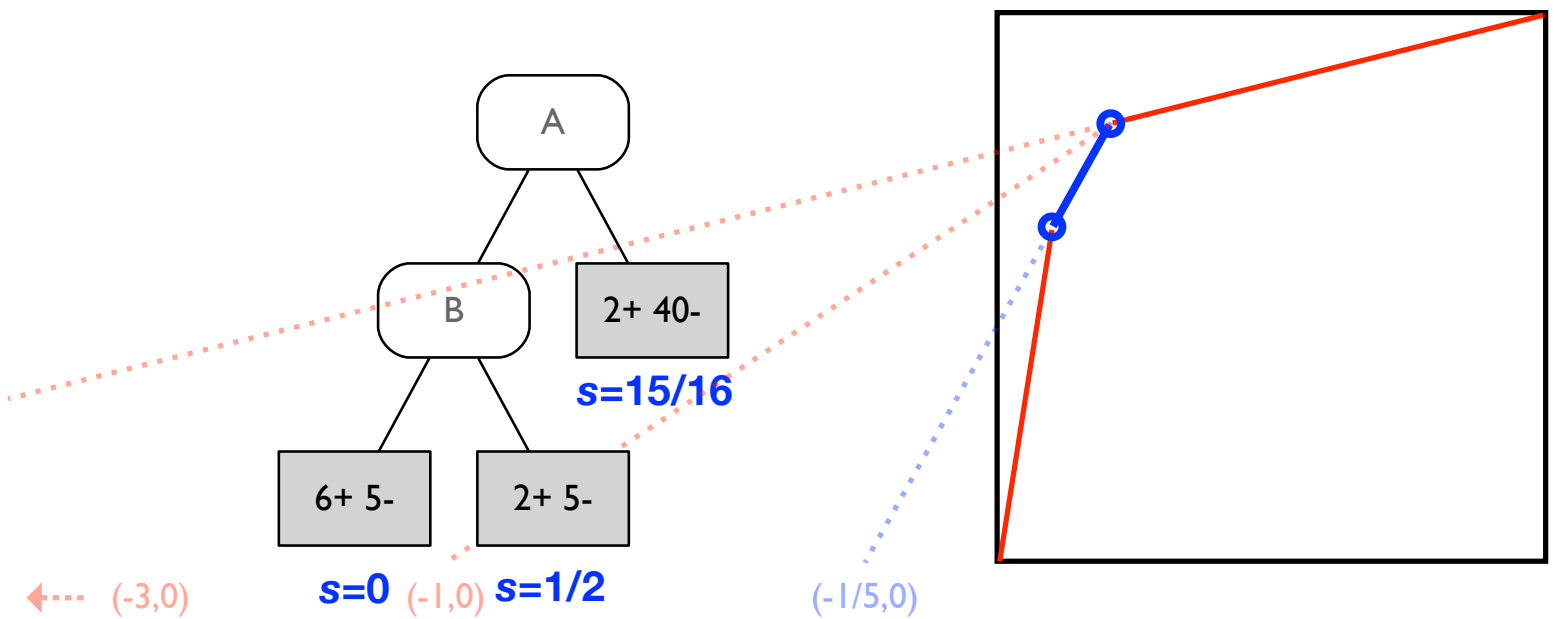


Figure: (Left) Accuracy-based cost lines. The x-axis shows c and the y-axis shows accuracy-based loss.
(Right) Cost lines for F-measure loss. The y-axis shows $2FQ/(1-FQ)$. We can see that the optimal operating points are chosen for lower values of c.

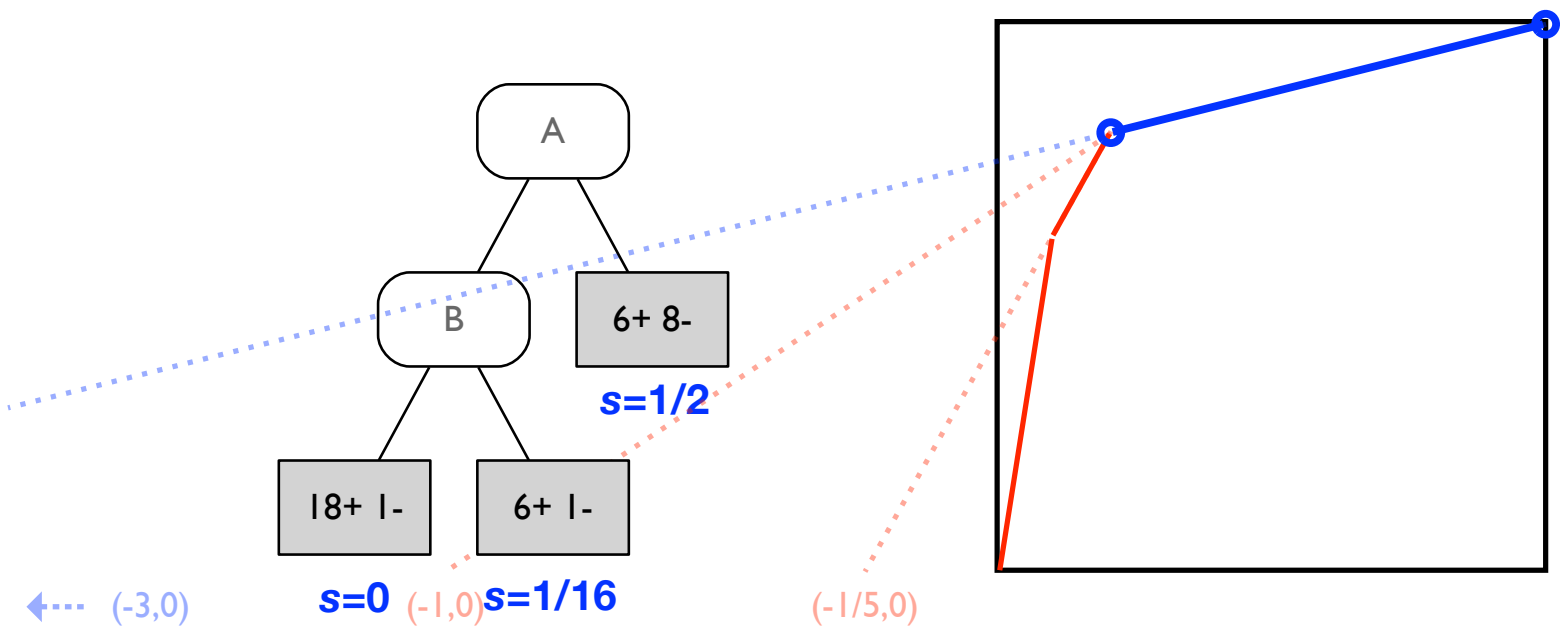
Scores calibrated for F-measure ($\pi = 1/2$)



Scores calibrated for F-measure ($\pi = 1/6$)



Scores calibrated for F-measure ($\pi = 3/4$)



What's next?

Concluding remarks

Apples and oranges

Expected classification loss is a convenient ‘common currency’ for model evaluation and selection.

Calibration

Classification models become more versatile when they are easily adapted to operating context changes.

Changing the loss function

Calibration is somewhat more involved but still possible for F-measure.

Concluding remarks

Apples and oranges

Expected classification loss is a convenient ‘common currency’ for model evaluation and selection.

Calibration

Classification models become more versatile when they are easily adapted to operating context changes.

Changing the loss function

Calibration is somewhat more involved but still possible for F-measure.

- ☞ Open problem: probabilistic interpretation of F-calibrated scores