The Geometry of ROC Space: Understanding Machine Learning Metrics through ROC Isometrics,
by Peter A. Flach
International Conference on
Machine Learning (ICML-2003)

http://www.cs.bris.ac.uk/Publications/Papers/1000704.pdf

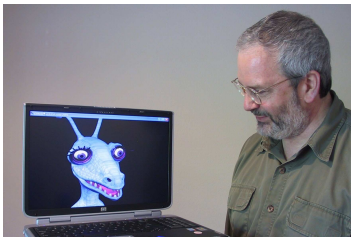
Robin Hewitt

Relevance for Computer Vision

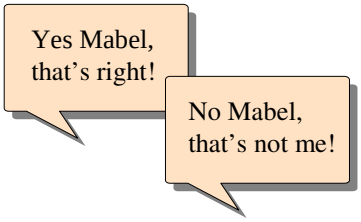
Robots and builders: Bruce with Leaf, Alex with Rocky, and Robin with Mabel.

The application context is visual object recognition in a social robot.

How can we translate user feedback into improved recognition behavior?



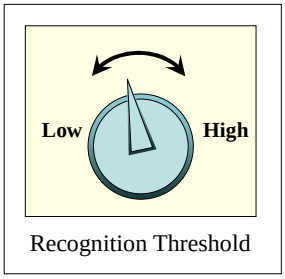
Relevance for Computer Vision



With more positive and negative examples, more sophisticated recognition algorithms can be deployed.

When should transitions occur?

Can we provide simple, intuitive tuning for a complex recognition system?

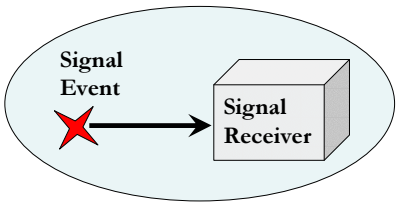


Relevance for Computer Vision

Methods based in ROC-curve analysis provide tools for

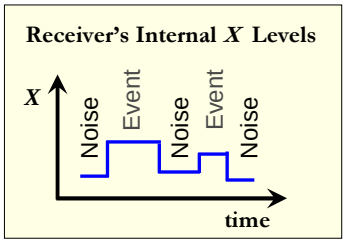
- Creating optimal hybrid classifiers
- Transitioning between classifiers
- Providing a simple, one-knob interface for "tuning" a complex, hybrid classifier.
- Separating use context – cost and class distribution – from classifier design.

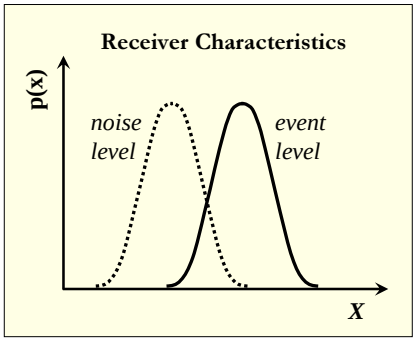
ROC Curves



ROC stands for Receiver Operating Characteristics.

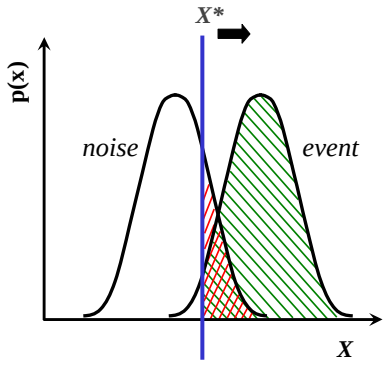
The terms and concepts come from signal-detection theory.

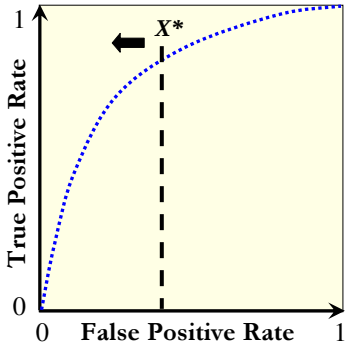




ROC Curves

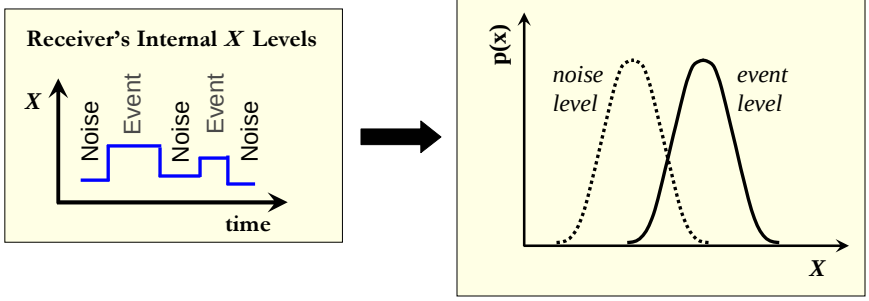
Each point on a receiver's ROC curve represents one cutoff threshold, ***X****. The red hatched area corresponds to the False Positive Rate. The green hatched area corresponds to the True Positive Rate. The ROC curve portrays (FPR, TPR) values at all possible threshold levels for a classifier.





ROC Curves

Discriminability, d′, is the distance between the two distributions – the distribution for the value of ***X*** when there's no signal and the distribution for the value of ***X*** when there is a signal.



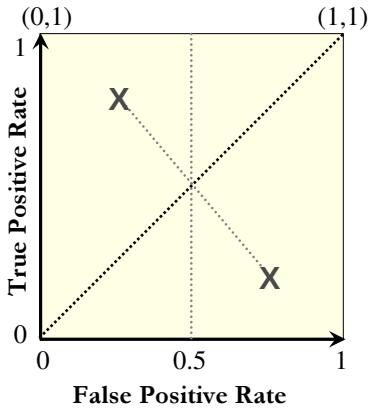
ROC Curves

The (0,1) point represents the Perfect Classifier. It's never wrong.

The ascending diagonal (0,0) to (1,1) represents d′ = 0. The internal levels for ***X*** in such a classifier carry no information for distinguishing between signal and noise.

Curves for real classifiers fall somewhere in between.

Any point that falls below the ascending diagonal can be mirrored across the point (0.5, 0.5). The information is there, but the class labels are reversed.

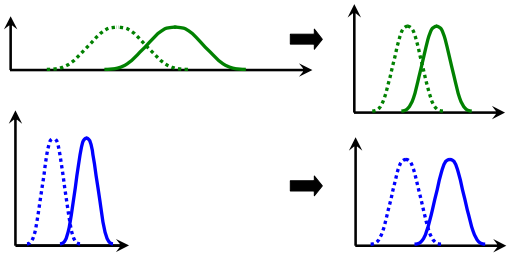


ROC Curves

Traditional practice has been to assume every classifier's internal signal and noise distributions are Gaussians with a shared standard deviation.

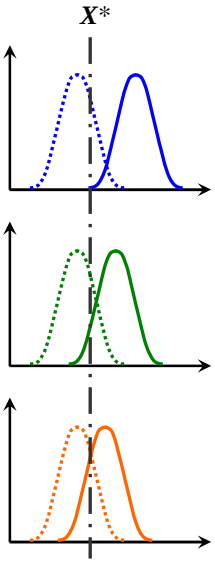
In that case, $d' = \frac{|\mu_1 - \mu_2|}{\sigma}$, so all distribution pairs could be rescaled to a standard height, σ_0, and discriminibility viewed as the absolute difference between rescaled means:

$$d' = \frac{\frac{\sigma_0}{\sigma}|\mu_1 - \mu_2|}{\sigma_0}$$

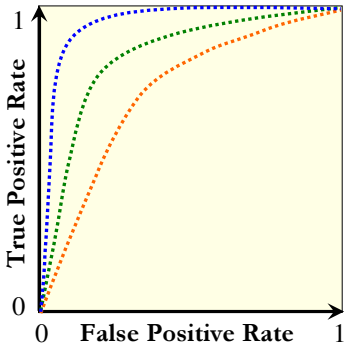


The difference between classifiers would then reduce to the linear distance between the rescaled means.

ROC Curves



If this assumption about classifiers were true, for every pair of classifiers, the one with the larger distance between the rescaled means would have a higher TPR value at every FPR value.



In this case, all possible ROC curves would nest.

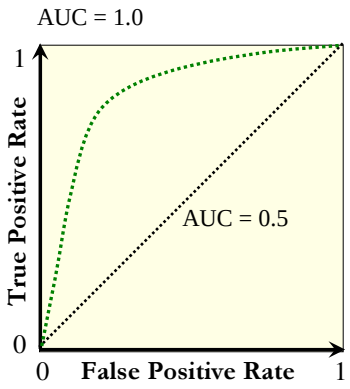
ROC Metrics

Under this assumption

- Classifier accuracy has only one degree of freedom (1DOF)
- Area Under the Curve (AUC) is a sufficient metric for classifiers.

In fact, this is exactly how ROC curves have traditionally been used. AUC is used to compare classifiers. It ranges from 0.5 along the ascending diagonal to 1.0 for the Perfect Classifier.

The Gini coefficient is similar:
GINI = 2AUC – 1.

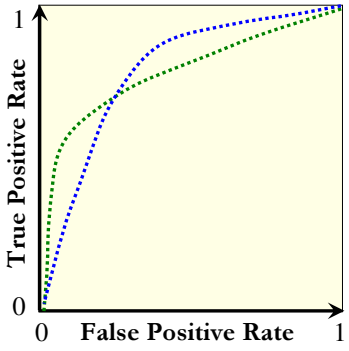


ROC Metrics

What if the normal-distribution-equal-sigma assumption for a classifier *isn't* valid?

Here are two classifiers with ROC curves that don't nest. Which is better?

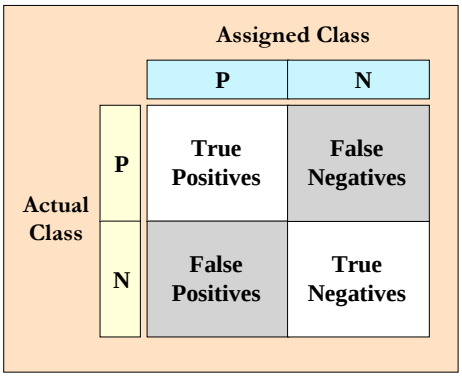
AUC value can't answer that. In this case, they have equal AUC. But even if their AUC values differed, the fact that the curves cross one another means that a 1DOF metric isn't sufficient.



Inherent Dimensionality

The overall space for quality metrics is three dimensional.

Confusion Matrix



Definitions

POS = Number of positive examples.

NEG = Number of negative examples.

TPR = True Positive Rate = TP/POS.

FPR = False Positive Rate = FP/NEG.

3DOF Total

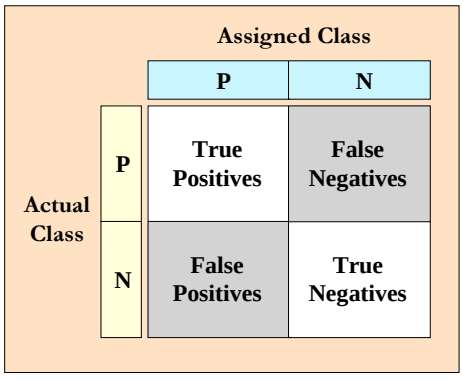
TPR, FPR, and NEG/POS.

(Flach uses pos = POS/(POS+NEG).)

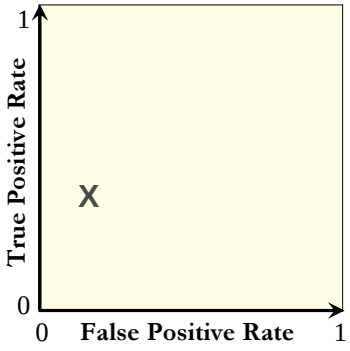
Inherent Dimensionality

ROC curves are *two* dimensional: TPR v. FPR.
A confusion matrix plots as one point on an ROC curve.

Confusion Matrix



TPR, FPR Space



ROC Metrics – Cost

Flach merges cost and class distribution as skew. Total estimated cost is, however, itself a potential classifier metric. It's also a convenient starting point for analyzing the geometry of 3D ROC metrics.

Dimensionality analysis indicates cost is a 3D function. To understand its shape, we find the equation for iso-cost curves in ROC-curve (FPR, TPR) space. This is a level-curve problem.

Let

C_{tot} = total expected cost, c_{fp} = false positive cost,

$L_c(FPR)$ = level curve for cost, c_{fn} = false negative cost.

Find $\frac{dC_{tot}(FPR, L_c(FPR))}{d(FPR)} = 0$

ROC Metrics – Cost

Find $\frac{dC_{tot}(FPR, L_c(FPR))}{d(FPR)} = 0$

$C_{tot} = c_{fp} FP + c_{fn} FN$ Total cost

$FPR = \frac{FP}{NEG}, \quad FN = POS - TP, \quad TPR = \frac{TP}{POS}$ Replace FP and FN with FPR and TPR

so, $C_{tot}(FPR, TPR) = c_{fp} FPR \bullet NEG \; + \; c_{fn} POS(1 - TPR).$

By the chain rule,

$$\frac{dC_{tot}(FPR, L_c(FPR))}{d(FPR)} = \frac{\partial C_{tot}}{\partial(FPR)} \frac{d(FPR)}{d(FPR)} + \frac{\partial C_{tot}}{\partial(TPR)} \frac{dL_c}{d(FPR)} = 0$$

$$c_{fp} \bullet NEG - c_{fn} \bullet POS \frac{dL_c}{d(FPR)} = 0$$

ROC Metrics – Cost

$$c_{fp} \bullet NEG - c_{fn} \bullet POS \frac{dL_c}{d(FPR)} = 0 \quad \rightarrow \quad \frac{dL_c}{d(FPR)} = \frac{c_{fp} NEG}{c_{fn} POS}.$$

Use context determines the four values on the righthand side of this expression. For a given use context then, iso-cost curves in (FPR,TPR) space are lines with slope,

$$m = \frac{c_{fp} NEG}{c_{fn} POS}.$$

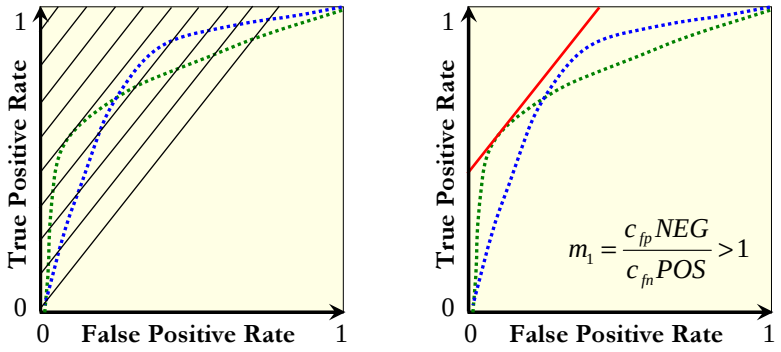
The value of C_{tot} for a level curve is specified by its TPR intercept as

$$C_{tot}(0, TPR) = c_{fn} POS(1 - TPR).$$

For a given slope, choosing the iso-cost line with the largest-possible TPR intercept minimizes C_{tot}.

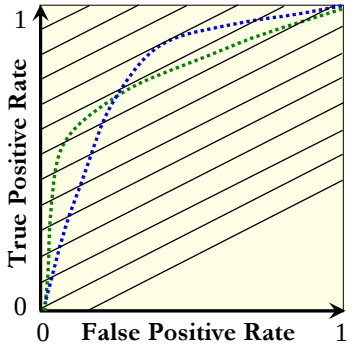
ROC Metrics – Cost

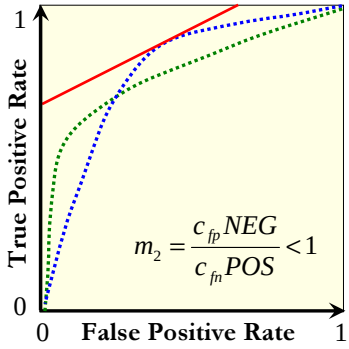
A family of iso-cost lines with identical slope, m_1. The line with highest TPR intercept that's tangent to an ROC curve minimizes total expected cost in this use context. In this situation, the classifier corresponding to the green ROC curve is better.



ROC Metrics – Cost

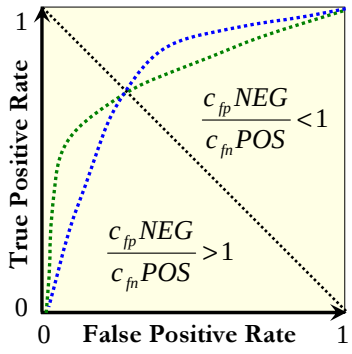
When slope = m_2, the classifier corresponding to the blue ROC curve is better.





ROC Metrics – Cost

The descending diagonal (0,1) to (1,0) separates the region where cost is mostly from false positives (iso-cost slope > 1) from the region where cost is mostly from false negatives (iso-cost slope < 1) .

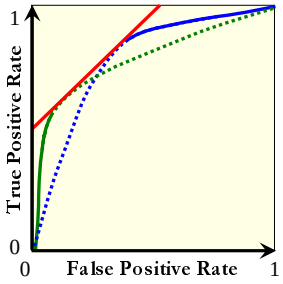


The classifier that dominates in the lower-left region minimizes cost due to false positives most successfully.

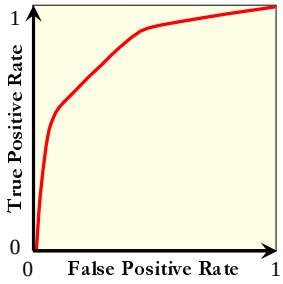
The classifier that dominates in the upper-right region minimizes cost due to false negatives most successfully.

ROC Curves – Convex Hull*

The slope shown in red is tangent to both ROC curves. For this value of *m*, the classifiers are equal. The dashed portions of these curves will never correspond to a minimum-cost tangent line, so they can be eliminated.



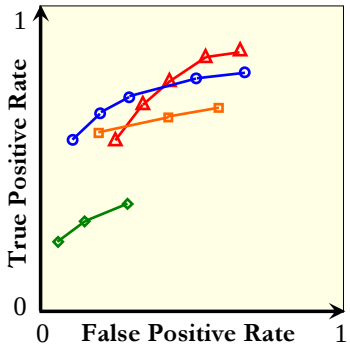
The remaining segments can be combined into an ROC curve for a hybrid classifier.

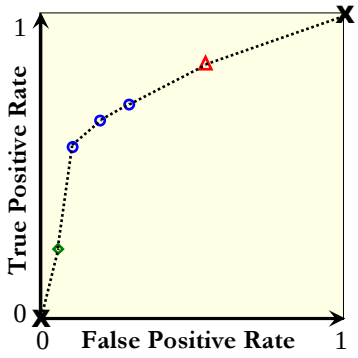


*The ROC Convex Hull method is from Provost and Fawcett, 2001.

ROC Curves – Convex Hull

This idea can be extended to N individual classifiers by taking the convex hull of the (FPR,TPR) points for all classifiers. In addition, the two trivial classifiers, AlwaysNeg, at (0,0), and AlwaysPos, at (1,1) are added to complete the convex hull. Any classifier entirely inside the convex hull can be eliminated from the hybrid classifier.





ROC Curves – Convex Hull

We don't have to compute *m* explicitly. Instead, we can give users a single threshold slider to adjust it. This leads to a simple, intuitive interface for adjusting classifier behavior:

1. Construct the ROC convex hull.

2. Set a value for *m*. (Use $m = 1$ if there's no information to guide the initial setting.)

3. When the user slides the threshold up, increase *m*. When the user moves the threshold down, decrease *m*.

4. Select the classifier and parameters to use for visual object recognition by finding the tangency point on the ROC convex hull for the current value of *m*.

ROC Metrics – Convex Hull

There's no "typical operating region" for cost and class-distribution.

$$m = \frac{c_{fp} NEG}{c_{fn} POS}$$

Example scenarios:

Identifying a fraudulent credit-card signature. Fradulent transactions are rare, and the penalty for a false positive is high (losing a customer, bad PR). $m >> 1$.

Equipment safety. Failure to detect human presence may lead to injury or death. *m* may be very small.

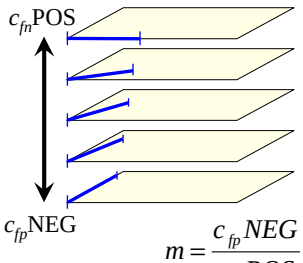
The 3D Cost Function and Skew

The 3D cost function is a family of twisted sheets. Vertical position specifies horizontal slope, *m*. The TPR intercept determines total expected cost.

Flach's "skew" is identical to *m*.

Changing the ratio of c_{fp} to c_{fn} has the same effect as changing the ratio of POS to NEG.

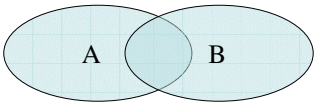
When cost and class distribution are merged into one ratio (skew), Minimum Cost = Maximum Accuracy. Flach's iso-accuracy lines are iso-cost lines with skew as *m*.



$$m = \frac{c_{fp}NEG}{c_{fn}POS}$$

Information-Retrieval Metrics

Precision, Recall, and F-measure come from the Information-Retrieval field.



A – Relevant Documents
B – Retrieved Documents

Precision, $P = \frac{|A \cap B|}{|B|}$ Recall, $R = \frac{|A \cap B|}{|A|}$ F-Measure, $F = 1 - \frac{|A \oplus B|}{|A| + |B|}$

Recall and Precision are competing objectives. Van Rijsbergen proposed using F-Measure as a way of balancing them. It turns out to be their harmonic mean.

ROC Metrics - Precision

Precision, $P = \frac{TP}{TP + FP} = \frac{POS \bullet TPR}{(POS \bullet TPR) + (NEG \bullet FPR)}$.

To find the level curve L_p, solve for $\frac{dL_P}{d(FPR)}$ in

$$\frac{dP(FPR, L_P(FPR))}{d(FPR)} = \frac{\partial P}{\partial(FPR)}\frac{d(FPR)}{d(FPR)} + \frac{\partial P}{\partial(TPR)}\frac{dL_P}{d(FPR)} = 0$$

➡ $\frac{dL_P}{d(FPR)} = \frac{TPR}{FPR}$. So along L_p, $\frac{d(TPR)}{TPR} = \frac{d(FPR)}{FPR}$

➡ $\int \frac{d(TPR)}{TPR} = \int \frac{d(FPR)}{FPR}$ ➡ $TPR = k \bullet FPR$ is the level curve for Precision.

ROC Metrics - Precision

What is the effect of changing k?

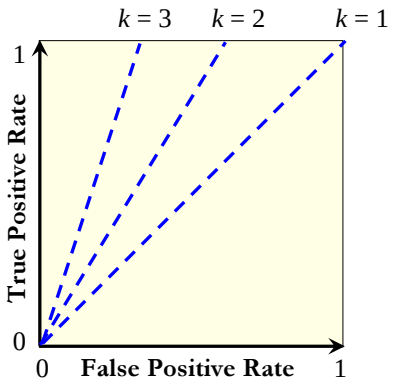
Level curve: $TPR = k \bullet FPR$

$$P = \frac{POS \bullet TPR}{(POS \bullet TPR) + (NEG \bullet FPR)}$$

Substituting for TPR in P:

$$P_L = \frac{POS \bullet k \bullet FPR}{(POS \bullet k \bullet FPR) + (NEG \bullet FPR)}$$

$$P_L = \frac{1}{1 + k \bullet NEG/POS}$$



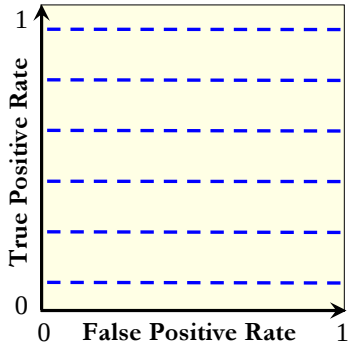
Lower values for *k* give higher Precision!

ROC Metrics - Recall

Recall, $R = \frac{TP}{POS} = TPR.$

The level curve for Recall is simply TPR = *k*.

Higher values for *k* give higher Recall.



ROC Metrics – F-Measure

F-Measure, $F = \frac{2POS \bullet TPR}{POS(1+TPR)+(NEG \bullet FPR)}.$

$$\frac{dL_F}{d(FPR)} = \frac{NEG \bullet TPR}{POS + NEG \bullet FPR} \quad \Rightarrow \quad \int \frac{d(TPR)}{TPR} = \int \frac{d(FPR)}{POS/NEG + FPR}$$

$TPR = k(FPR + POS/NEG)$
is the level curve for F-Measure.

(It appears that Flach didn't multiply skew by *k*, the constant of integration. This oversight makes it appear that offset and slope for *F* can vary independently.)

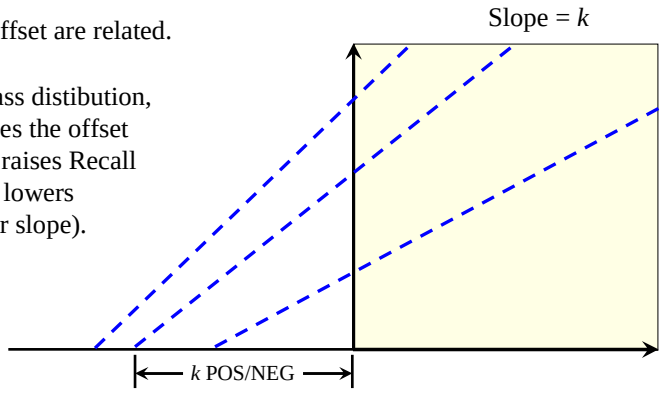
ROC Metrics – F-Measure

$TPR = k(FPR + POS/NEG)$

is the level curve for F-Measure.

Slope and TPR offset are related.

For any given class distibution, increasing k moves the offset further left. That raises Recall (TPR value), but lowers Precision (steeper slope).



ROC Metrics – G-Measure

Flach's proposed G-Measure has the same level curves as F-Measure.
Proof:

$$G = \frac{TP}{FP + POS} = \frac{POS \bullet TPR}{NEG \bullet FPR + POS}$$

$$\frac{dL_G}{d(FPR)} = \frac{NEG \bullet TPR}{NEG \bullet FPR + POS} \quad \Rightarrow \int \frac{d(TPR)}{TPR} = \int \frac{d(FPR)}{FPR + POS/NEG}$$

$TPR = k(FPR + POS/NEG)$, the level-curve eq. for F-Measure is also the level-curve eq. for G-Measure. □

Summary

- There's more to ROC curves than Area.
- The slope of the iso-cost curve in (FPR, TPR) space indicates which classifier is better for a particular use context.
- The ROC Convex Hull method is a principled way to create a hybrid classifier.
- ROC-based analysis provides a way to separate classifier design from costs and class distributions associated with use context.
- Level-curve analysis in (FPR,TPR) space provides new insights into a metric's utility and limitations.

Additional References

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

P.A. Flach, "The Many Faces of ROC Analysis in Machine Learning." ICML'04 Tutorial.
http://www.cs.bris.ac.uk/~flach/ICML04tutorial/index.html

C. Elkan, The foundations of cost-sensitive learning. *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-01),* pages 973-978, 2001.

C.J. Van Rijsbergen, *Information Retrieval*, London, 1979.
http://www.dcs.gla.ac.uk/~iain/keith/index.htm

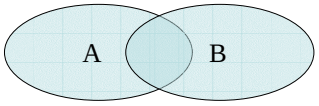
J.A. Hanley and B.J. McNeil, The meaning and use of the area under a receiver operating characteristic curve. *Radiology* 1982;143:29-36.

Extra Slides

F-Measure Proof

Following is a proof that the set-theoretic expression for F-Measure, F, is the same as the harmonic mean of Precision, P, and Recall, R.

Definitions:



A – Relevant Documents
B – Retrieved Documents
$|A \cap B| = TP$

$$P = \frac{TP}{|B|}, \qquad R = \frac{TP}{|A|}, \qquad F = 1 - \frac{|A \oplus B|}{|A| + |B|}, \qquad H(x, y) = \frac{1}{\frac{1}{2x} + \frac{1}{2y}}.$$

Proof:

$$\frac{|A \oplus B|}{|A| + |B|} = \frac{|A| - TP + |B| - TP}{|A| + |B|} = \frac{|A| + |B| - 2TP}{|A| + |B|} = \frac{|A| + |B|}{|A| + |B|} - \frac{2TP}{|A| + |B|}$$

F-Measure Proof, cont.

$$F = 1 - \frac{|A \oplus B|}{|A| + |B|} = \frac{2TP}{|A| + |B|} = \frac{1}{\left(\frac{|A| + |B|}{2TP}\right)} = \frac{1}{\frac{|A|}{2TP} + \frac{|B|}{2TP}} = \frac{1}{\frac{1}{2R} + \frac{1}{2P}}.$$ □

F(FPR,TPR)

Following are the steps for expressing F-Measure as F(FPR,TPR).

Definitions:

$$|A| = POS, \qquad |B| = TP + FP.$$

Starting with an expression for F from the harmonic-means proof,

$$F = \frac{2TP}{|A| + |B|} = \frac{2TP}{POS + TP + FP}$$

$$= \frac{2POS \bullet TPR}{POS + POS \bullet TPR + NEG \bullet FPR} = \frac{2POS \bullet TPR}{POS(1 + TPR) + (NEG \bullet FPR)}.$$