The Relationship Between Precision-Recall and ROC Curves

Jesse Davis and Mark Goadrich, University of Wisconsin – Madison, USA

{jdavis,richm}@cs.wisc.edu

Abstract

Receiver Operator Characteristic (ROC) curves are commonly used to present results for binary decision problems in machine learning. However, when dealing with highly skewed datasets, Precision-Recall (PR) curves give a more informative picture of an algorithm's performance. We show that a deep connection exists between ROC space and PR space, such that a curve dominates in ROC space if and only if it dominates in PR space. A corollary is the notion of an achievable PR curve, which has properties much like the convex hull in ROC space; we show an efficient algorithm for computing this curve. Finally, we also note differences in the two types of curves are significant for algorithm design. For example, in PR space it is incorrect to linearly interpolate between points. Furthermore, algorithms that optimize the area under the ROC curve are not guaranteed to optimize the area under the PR curve.

	Actual Positive	Actual Negative
Predicted Positive	*TP*	*FP*
Predicted Negative	*FN*	*TN*

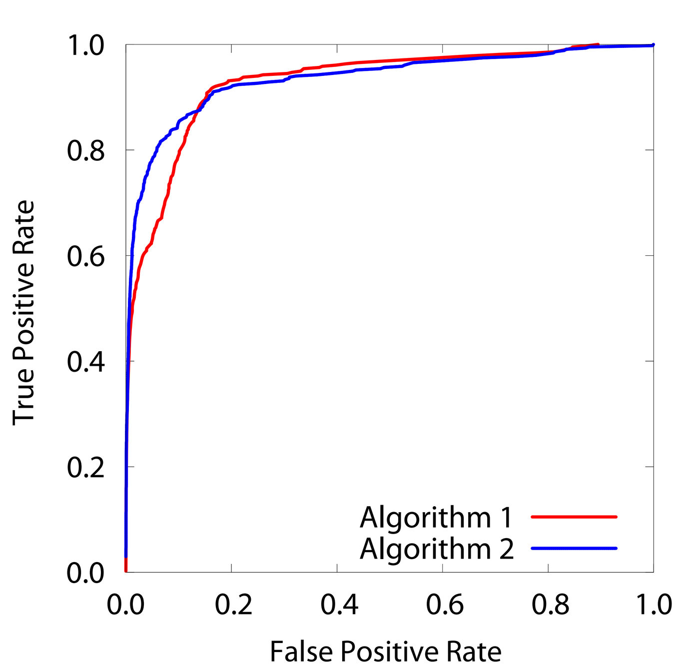
Confusion Matrix

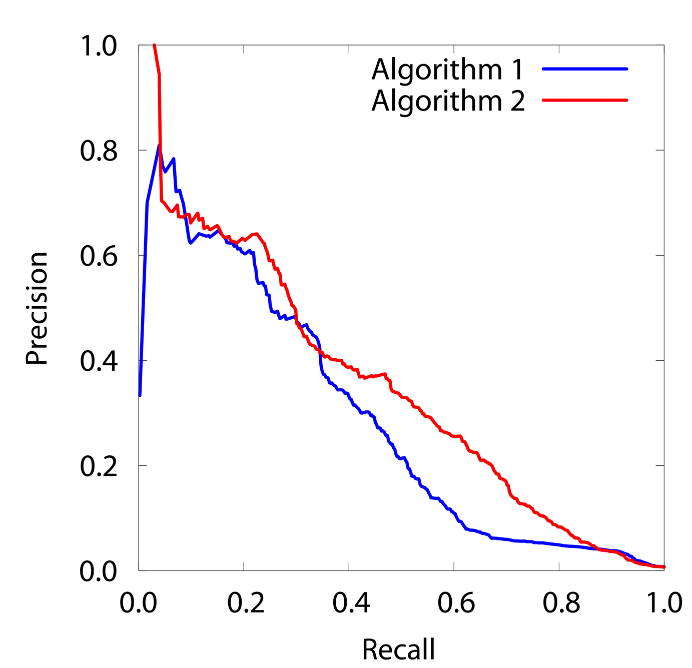
$$\text{Recall} = \frac{TP}{TP+FN}$$
$$\text{Precision} = \frac{TP}{TP+FP}$$
$$\text{True Positive Rate} = \frac{TP}{TP+FN}$$
$$\text{False Positive Rate} = \frac{TN}{TN+FP}$$

Evaluation Metrics

What are Precision-Recall and ROC Curves?

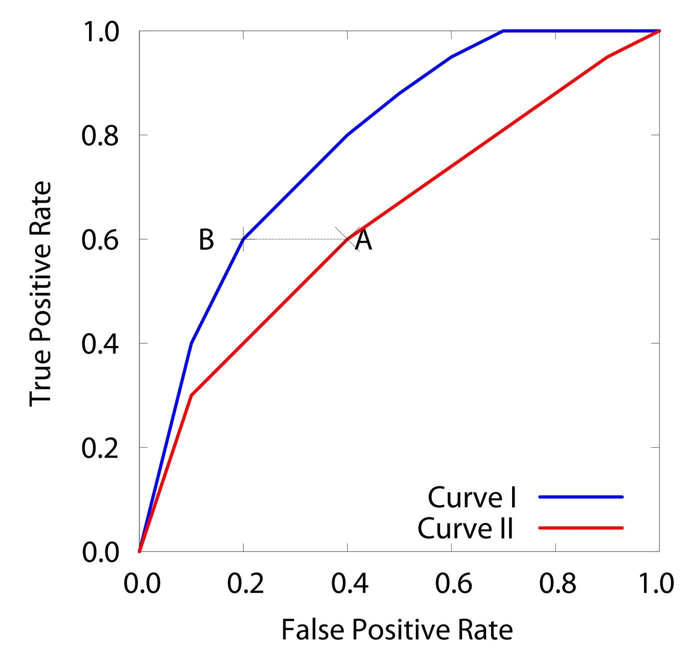
Receiver Operator Characteristic (ROC) curves show how the number of correctly classified positive examples varies with the number of incorrectly classified negative examples. However, ROC curves can present an overly optimistic view of an algorithm's performance if there is a large skew in the class distribution. Precision-Recall (PR) curves instead use metrics focused on the true positive examples.

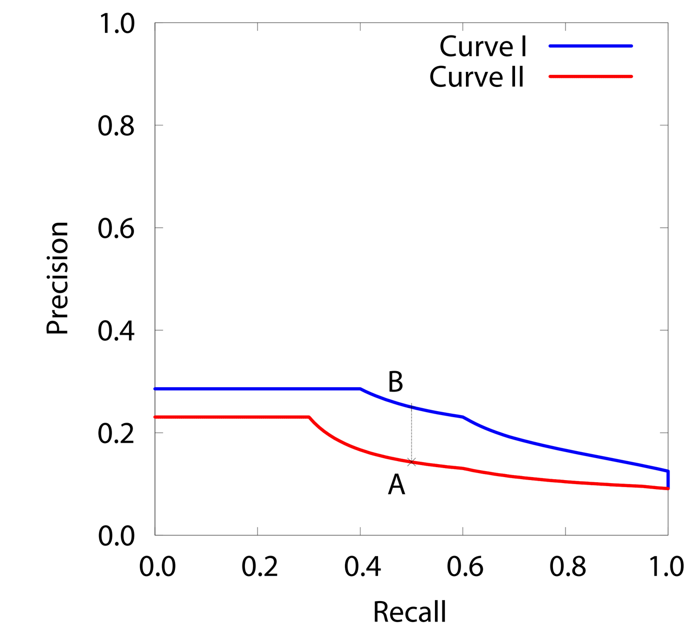




Acknowledgements

A Java program for calculating all of the discussed metrics can be found at http://www.cs.wisc.edu/~richm/programs/AUC/. We gratefully acknowledge the funding from USA NLM Grant 5T15LM007359-02 and USA Air Force Grant F30602-01-2-0571, Vitor Santos Costa, Louis Oliphant, our advisors David Page and Jude Shavlik and our anonymous reviewers for their helpful comments and suggestions.





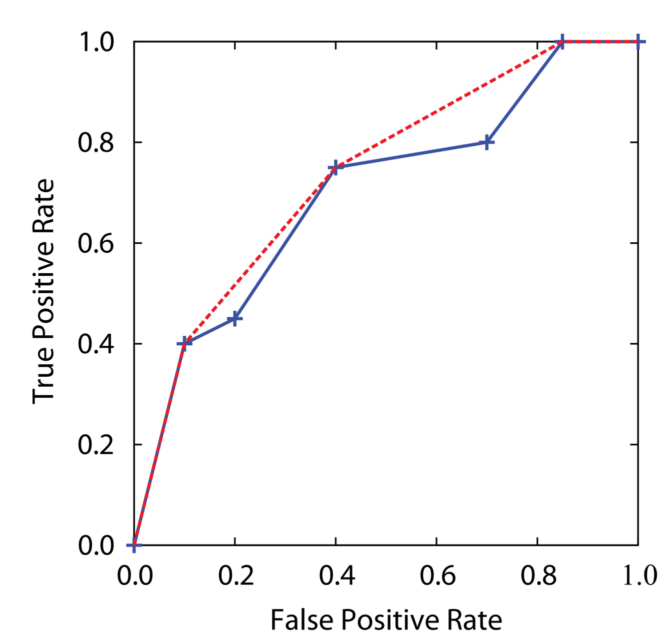
Dominance Relationship

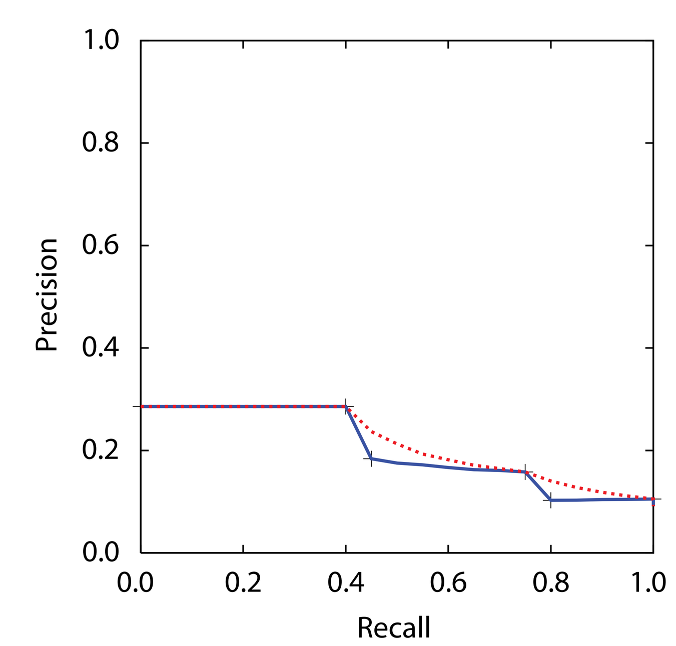
Our main theorem is as follows: *For a fixed number of positive and negative examples, one curve dominates a second curve in ROC space if and only if the first dominates the second in Precision-Recall space.*

Proof by contradiction of ROC to PR (PR to ROC proof is isomorphic): Suppose we have curve I and curve II such that curve I dominates in ROC space, yet curve I does not dominate in PR space. This means there exists some point A on curve II such that the point B on curve I with identical Recall has lower Precision.

(1) $PRECISION(A) > PRECISION(B)$	Given
(2) $RECALL(A) = RECALL(B)$	Given
(3) $TPR(A) = TPR(B)$	from 2
(4) $TP_A = TP_B = TP$	from 3
(5) $PRECISION(A) = \frac{TP}{TP+FP_A}$	from 4
(6) $PRECISION(B) = \frac{TP}{TP+FP_B}$	from 4
(7) $FPR(A) \geq FPR(B)$	Given
(8) $FP_A \geq FP_B$	from 7
(9) **$PRECISION(A) \leq PRECISION(B)$?**	from 5, 6, 8

Contradicts with 1, so curve I must dominate in PR space.

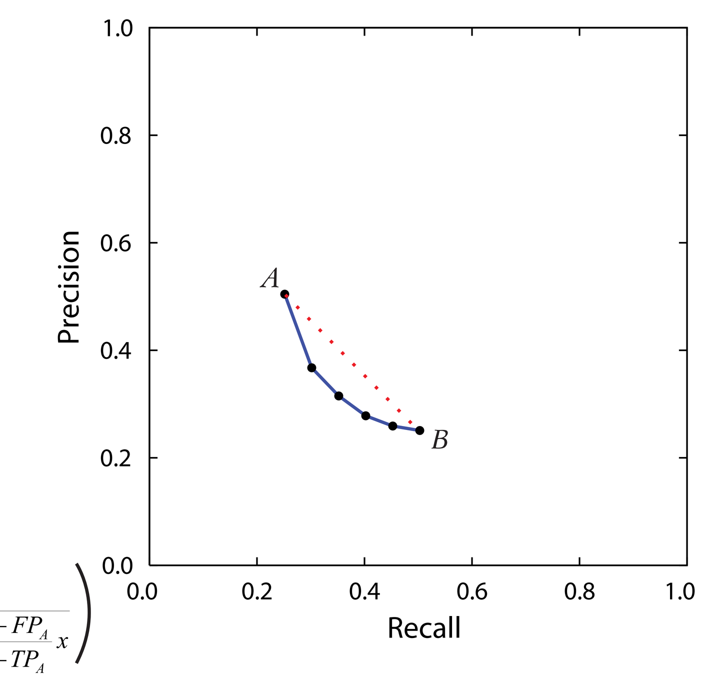




Convex Hull and Achievable Curve

Given a set of points in ROC space, the convex hull must meet the following three criteria: linear interpolation is used between adjacent points, no point lies above the final curve, and for any pair of points used to construct the curve, the line segment connecting them is equal to or below the curve. In our paper, we prove the following theorem: *Given a set of points in PR space, there exists an achievable PR curve that dominates the other valid curves that could be constructed with these points.* This is based on the dominance proof above.

	TP	*FP*	*REC*	*PREC*
A	5	5	0.25	0.5
.	6	10	0.3	0.375
.	7	15	0.35	0.318
.	8	20	0.4	0.286
.	9	25	0.45	0.265
B	10	30	0.5	0.25

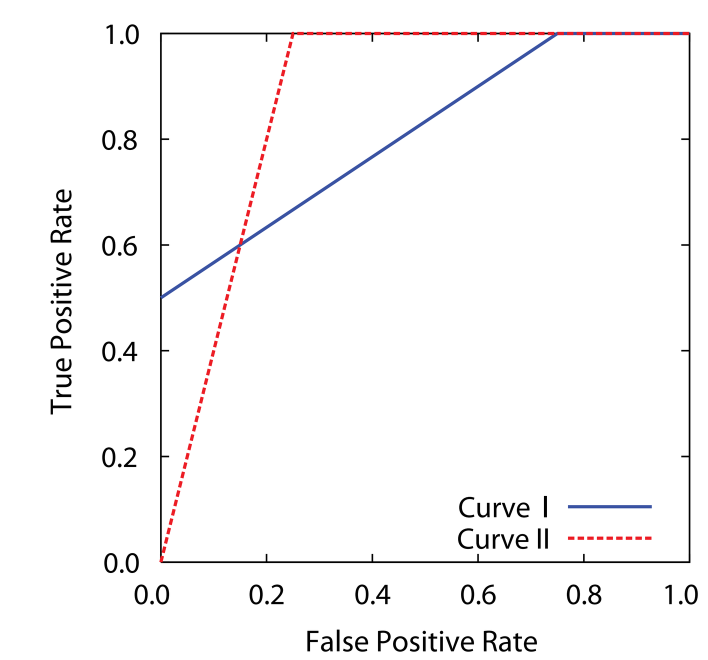


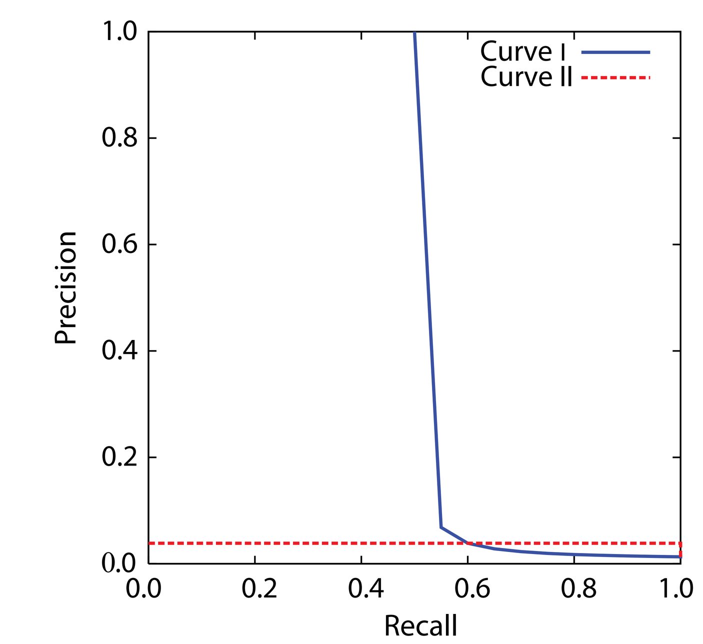
Interpolated points for a dataset with 20 positive and 2000 negative examples can be found with the following equation:

$$\left(\frac{TP_A+x}{TP+FN}, \frac{TP_A+x}{TP_A+x+FP_A+\frac{FP_B-FP_A}{TP_B-TP_A}x}\right)$$

Interpolation in PR Space

As the level of Recall varies, the Precision does not necessarily change linearly due to the fact that FP replaces FN in the denominator of the Precision metric. In these cases, linear interpolation is a mistake that yields an overly-optimistic estimate of performance.





Optimizing AUC in PR and ROC

Algorithms which optimize the AUC-ROC do not optimize the AUC-PR. We show two overlapping curves in ROC space for a domain with 20 positive examples and 2000 negative examples, where each curve individually is a convex hull. The AUC-ROC for curve I is 0.813 and the AUC-ROC for curve II is 0.875, so an algorithm optimizing the AUC-ROC and choosing between these two rankings would choose curve II. However, the same curves translated into PR space, and the difference here is drastic. The AUC-PR for curve I is now 0.514 due to the high ranking of over half of the positive examples, while the AUC-PR for curve II is far less at 0.038, so the direct opposite choice of curve I should be made to optimize the AUC-PR.

Selected References

Bradley, A. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30, 1145-1159.

Cortes, C., & Mohri, M. (2003). AUC optimization vs. error rate minimization. *Neural Information Processing Systems* 15 (*NIPS*). MIT Press.

Davis, J., Burnside, E., Dutra, I., Page, D., Ramakrishnan, R., Costa, V. S., & Shavlik, J. (2005). View learning for statistical relational learning: With an application to mammography. *Proceeding of the 19th International Joint Conference on Artifcial Intelligence*. Edinburgh, Scotland.

Goadrich, M., Oliphant, L., & Shavlik, J. (2004). Learning ensembles of first-order clauses for recall-precision curves: A case study in biomedical information extraction. *Proceedings of the 14th International Conference on Inductive Logic Programming (ILP).* Porto, Portugal.

Provost, F., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. *Proceeding of the 15th International Conference on Machine Learning* (pp. 445-453). Morgan Kaufmann, San Francisco, CA.