Unachievable Region in Precision-Recall Space and Its Effect on Empirical Evaluation

Kendrick Boyd[1]
Vítor Santos Costa[2] Jesse Davis[3] David Page[1]

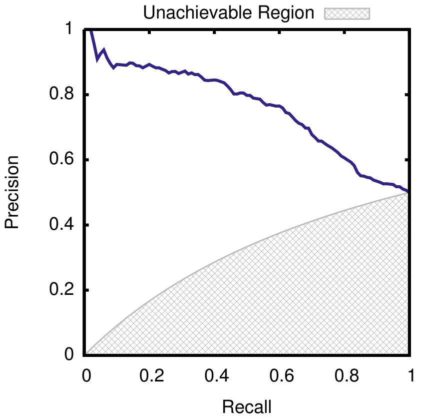
[1]University of Wisconsin – Madison

[2]University of Porto, Portugal

[3]KU Leuven, Belgium

June 28, 2012

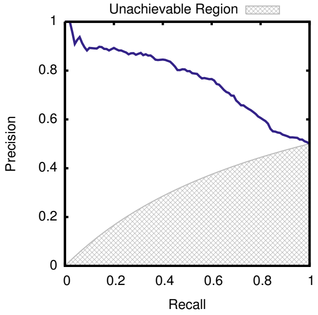
Introduction

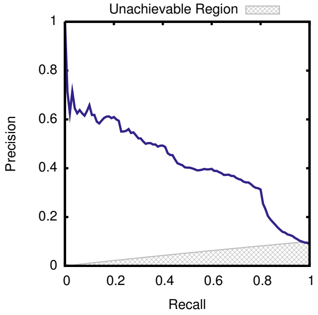


Unachievable Region

Precision-recall curves cannot go through the unachievable region.

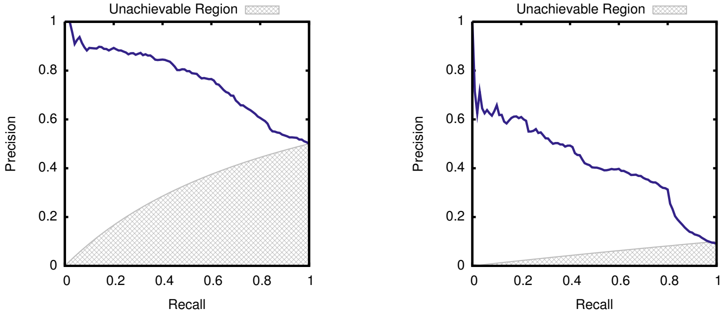
Introduction





Unachievable region varies with the data set.

Introduction



Unachievable region varies with the data set.

Our Goal

Understand and characterize the unachievable region.

Precision-Recall Analysis

$$\text{Precision} = \frac{TP}{TP + FP}$$
$$\text{Recall} = \frac{TP}{TP + FN}$$

	Actual P	Actual N
P	*TP*	*FP*
N	*FN*	*TN*
	pos	*neg*

Precision-Recall Analysis

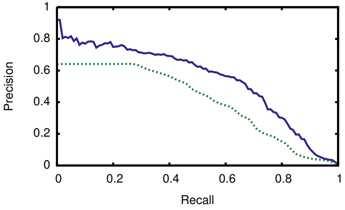
$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

	Actual	
	P	N
P	TP	FP
N	FN	TN
	pos	neg

Scores used in PR analysis

- Precision-recall (PR) curves
- Area under PR curve (AUCPR)
- F_β
- Mean average precision

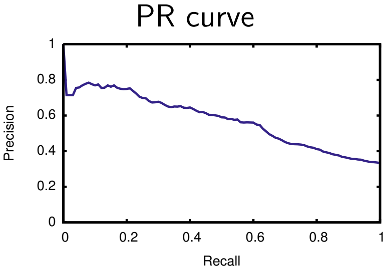


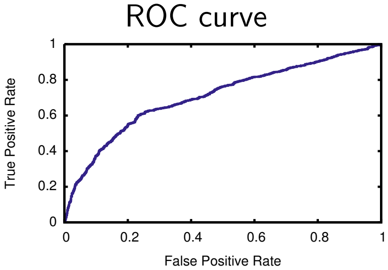
Known PR Space Properties

- Precision has high variance at low recall

Known PR Space Properties

- Precision has high variance at low recall
- Can translate between PR curves and ROC curves





Known PR Space Properties

- Precision has high variance at low recall
- Can translate between PR curves and ROC curves

Class Skew

Proportion of positive examples in a data set.

Known PR Space Properties

- Precision has high variance at low recall
- Can translate between PR curves and ROC curves
- ROC curves - independent of class skew
- PR curves - sensitive to class skew

Class Skew

Proportion of positive examples in a data set.

Outline

1. Introduction
2. **Unachievable Points**
3. Unachievable Region
4. Discussion
   - Downsampling
   - Aggregation
   - F1 Score
5. Conclusion

Illustration of an Unachievable Point

- 100 positive examples
- 200 negative examples
- Hold precision constant at 0.2

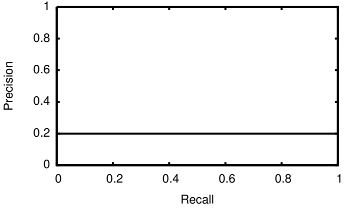
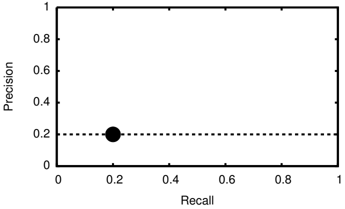


Illustration of an Unachievable Point

- 100 positive examples
- 200 negative examples
- Hold precision constant at 0.2

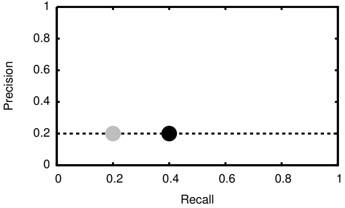


	Actual	
	P	N
P	20	80
N	80	120
	100	200

0.2 recall

Illustration of an Unachievable Point

- 100 positive examples
- 200 negative examples
- Hold precision constant at 0.2



	Actual	
	P	N
P	20	80
N	80	120
	100	200

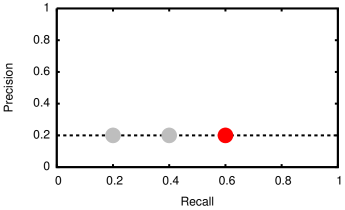
0.2 recall

	Actual	
	P	N
P	40	160
N	60	40
	100	200

0.4 recall

Illustration of an Unachievable Point

- 100 positive examples
- 200 negative examples
- Hold precision constant at 0.2



	Actual P	Actual N
P	20	80
N	80	120
	100	200

0.2 recall

	Actual P	Actual N
P	40	160
N	60	40
	100	200

0.4 recall

	Actual P	Actual N
P	60	240
N	40	**-40**
	100	200

0.6 recall

Unachievable Points in PR Space

Theorem

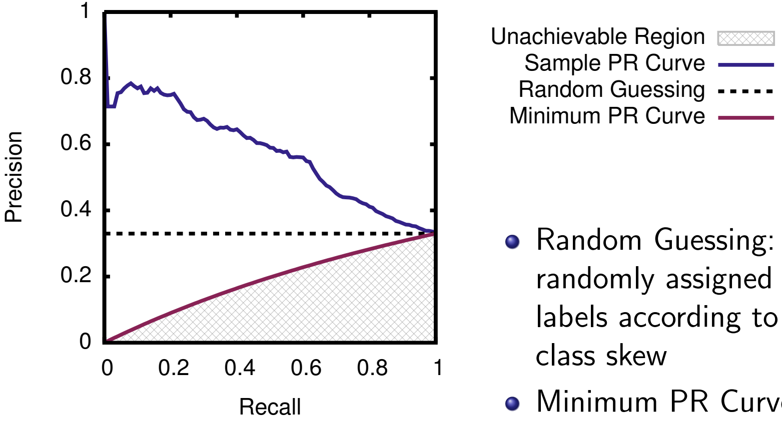
Precision (p) and recall (r) must satisfy

$$p \geq \frac{\pi r}{1 - \pi + \pi r}$$

where $\pi = \frac{pos}{pos+neg}$ *is the class skew.*

Outline

Unachievable Region



Example PR curves with $\pi = 0.33$.

- Random Guessing: randomly assigned labels according to class skew
- Minimum PR Curve: worst possible ranking

Minimum AUCPR

Theorem

The area of the unachievable region in PR space and the minimum AUCPR*, for class skew* π*, is*

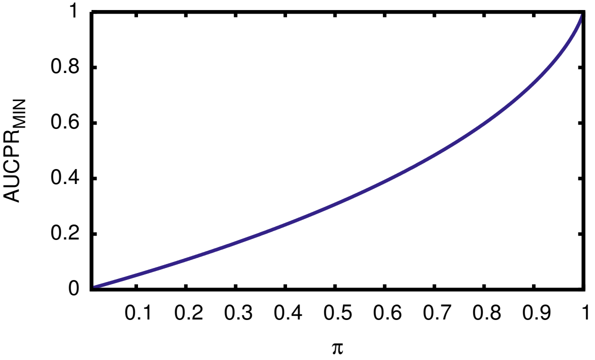
$$\mathrm{AUCPR}_{\mathrm{MIN}} = 1 + \frac{(1-\pi)\ln(1-\pi)}{\pi}$$

Minimum AUCPR

Theorem

The area of the unachievable region in PR space and the minimum AUCPR*, for class skew* π*, is*

$$\mathrm{AUCPR}_{\mathrm{MIN}} = 1 + \frac{(1-\pi)\ln(1-\pi)}{\pi}$$



Outline

AUCPR Comparisons

- Unachievable region always included “for free” in AUCPR

AUCPR Comparisons

- Unachievable region always included “for free” in AUCPR
- Range of AUCPR scores is $[\mathrm{AUCPR}_{\mathrm{MIN}}, 1]$ and thus depends on skew

AUCPR Comparisons

- Unachievable region always included "for free" in AUCPR
- Range of AUCPR scores is $[\mathrm{AUCPR}_{\mathrm{MIN}}, 1]$ and thus depends on skew
- AUCPR not comparable from different skews

AUCNPR

Normalized area under precision-recall curve

$$\mathrm{AUCNPR} = \frac{\mathrm{AUCPR} - \mathrm{AUCPR}_{\mathrm{MIN}}}{1 - \mathrm{AUCPR}_{\mathrm{MIN}}}$$

- Pros
  - Range of AUCNPR is $[0, 1]$ regardless of skew
  - With same skew, preserves ordering of AUCPR
- Cons
  - No good interpretation as an area in PR space
  - AUCNPR of random guessing is not simple
  - Still sensitive to skew

Downsampling

- Often downsample positives and negatives differently

Downsampling

- Often downsample positives and negatives differently
- Changes unachievable region for PR analysis

Downsampling

- Often downsample positives and negatives differently
- Changes unachievable region for PR analysis

Cohort Study

Should preserve the true skew

Downsampling

- Often downsample positives and negatives differently
- Changes unachievable region for PR analysis

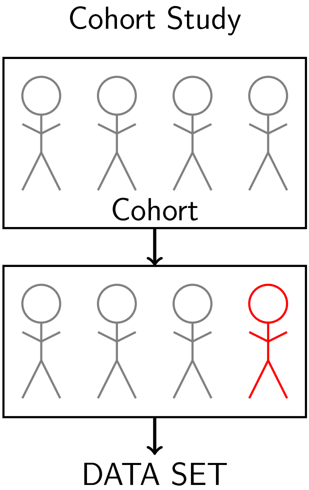
Cohort Study

Should preserve the true skew

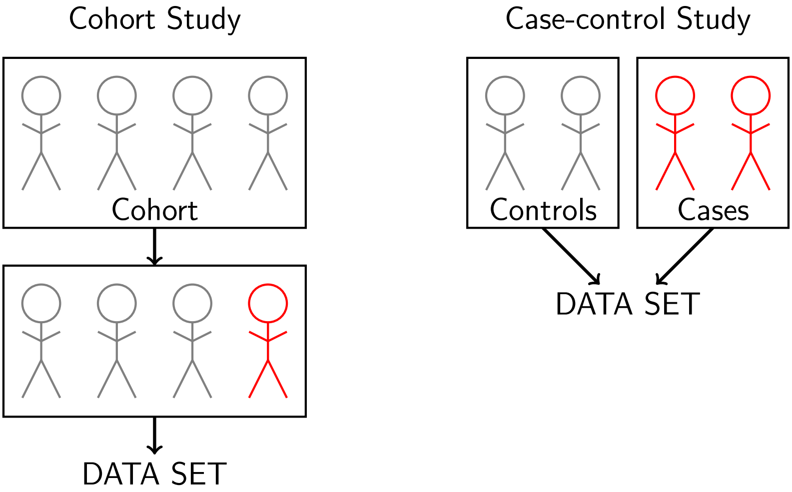
Case-control Study

Artificially makes the ratio of negatives to positives 1:1, 2:1, etc.

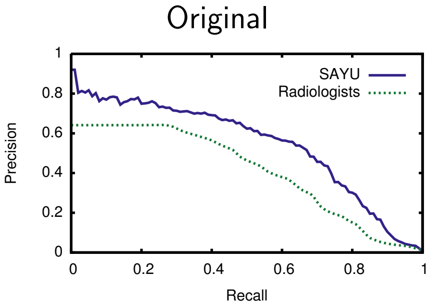
Downsampling on Mammography

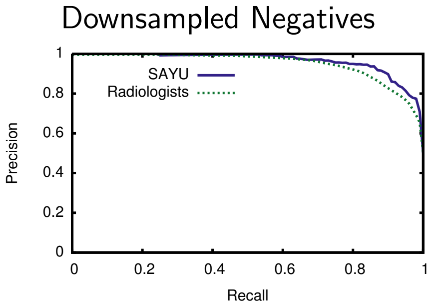


Downsampling on Mammography



Downsampling on Mammography





120:1 ratio ($\pi = 0.008$)

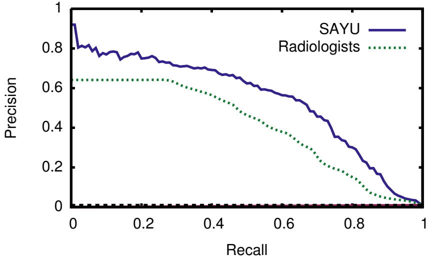
$\mathrm{AUCPR} \quad = 0.545$

1:1 ratio ($\pi = 0.5$)

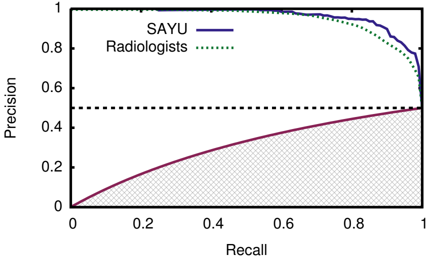
$\mathrm{AUCPR} \quad = 0.965$

Downsampling on Mammography

Original



Downsampled Negatives



120:1 ratio ($\pi = 0.008$)

$$\mathrm{AUCPR} = 0.545$$
$$\mathrm{AUCPR}_{\mathrm{MIN}} = 0.004$$
$$\mathrm{AUCNPR} = 0.543$$

1:1 ratio ($\pi = 0.5$)

$$\mathrm{AUCPR} = 0.965$$
$$\mathrm{AUCPR}_{\mathrm{MIN}} = 0.307$$
$$\mathrm{AUCNPR} = 0.950$$

Aggregating Results

Sometimes want to combine results from problems with different skews

- Cross-validation folds
- Multiple tasks

Aggregating Results

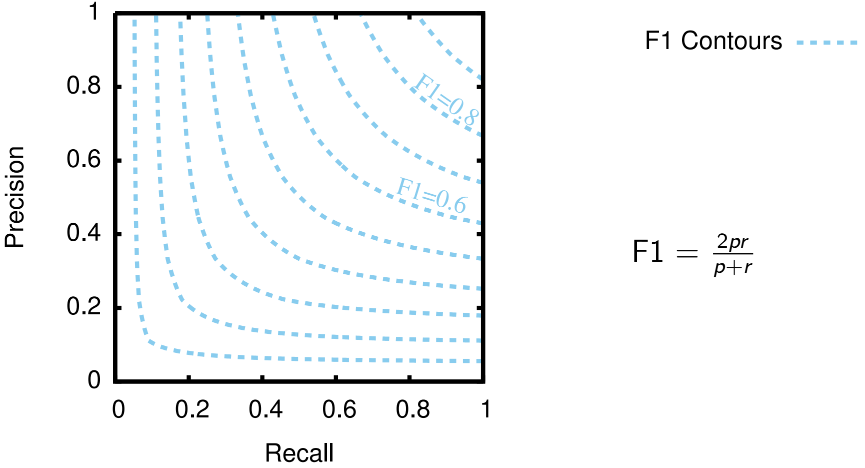
Sometimes want to combine results from problems with different skews

- Cross-validation folds
- Multiple tasks

AUCNPR is a step in the right direction

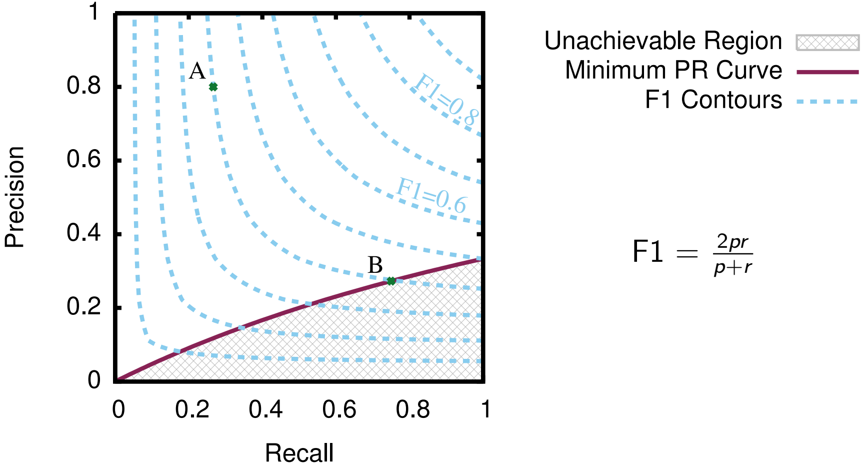
- AUCNPR range is [0,1] for each fold/task
- Mean AUCNPR gives mean fraction of achievable area obtained
- But more work is needed!

F1 Score and Unachievable Region



$$\text{F1} = \frac{2pr}{p+r}$$

F1 Score and Unachievable Region



$$F1 = \frac{2pr}{p+r}$$

Outline

Conclusions

- Awareness of the unachievable region is critical for precision-recall analysis

Conclusions

- Awareness of the unachievable region is critical for precision-recall analysis
- Shown how to compute this region from just the class skew

Conclusions

- Awareness of the unachievable region is critical for precision-recall analysis
- Shown how to compute this region from just the class skew
- Recommendation: Always show unachievable region in figures

Conclusions

- Awareness of the unachievable region is critical for precision-recall analysis
- Shown how to compute this region from just the class skew
- Recommendation: Always show unachievable region in figures
- AUCNPR: a first step towards scores that account for the unachievable region

Acknowledgments

- Elizabeth S. Burnside
- Jude Shavlik

- CIBM Training Program (NIH 5T15LM007359)
- NIGMS grant R01GM097618-01
- NLM grant R01LM011028-01
- NIEHS grant 4R01ES017400-03
- ERDF through the Programme COMPETE
- Portuguese Government through FCT
- Project HORUS ref. PTDC/EIA-EIA/100897/2008
- European Union Seventh Framework Programme FP7/2007-2013
- Research Fund K.U. Leuven (CREA/11/015 and OT/11/051)
- EU FP7 Marie Curie Farerr Integration Grant (#294068)
- FWO-Vlaanderen (G.0356.12)
- UW Carbone Cancer Center

Thank you

Questions?

Modifying F1 Score

A few desirable properties for a modified F1 score, f',

$$f'(r, p) = 0 \text{ if } p = \frac{r\pi}{1 - \pi + r\pi}$$
$$f'(r_1, p) < f'(r_2, p) \text{ iff } r_1 < r_2$$
$$f'(r, p_1) < f'(r, p_2) \text{ iff } p_1 < p_2$$

Modifying F1 Score

A few desirable properties for a modified F1 score, f',

$$f'(r,p) = 0 \text{ if } p = \frac{r\pi}{1-\pi+r\pi}$$
$$f'(r_1,p) < f'(r_2,p) \text{ iff } r_1 < r_2$$
$$f'(r,p_1) < f'(r,p_2) \text{ iff } p_1 < p_2$$

Impossible!

$$0 = f'(0,0) < f'(0,\pi) < f'(1,\pi) = 0$$

Modifying F1 Score

Relaxed properties for a modified F1 score, f',

$$f'(r, p) = 0 \text{ if } p = \frac{r\pi}{1 - \pi + r\pi}$$

$$f'(r_1, p) \leq f'(r_2, p) \text{ iff } r_1 \leq r_2$$

$$f'(r, p_1) \leq f'(r, p_2) \text{ iff } p_1 \leq p_2$$

Modifying F1 Score

Relaxed properties for a modified F1 score, f',

$$f'(r,p) = 0 \text{ if } p = \frac{r\pi}{1-\pi+r\pi}$$
$$f'(r_1,p) \le f'(r_2,p) \text{ iff } r_1 \le r_2$$
$$f'(r,p_1) \le f'(r,p_2) \text{ iff } p_1 \le p_2$$

One possible f'

$$f'(r,p) = \begin{cases} 0 & \text{if } p \le \pi \\ \frac{2(p-\pi)r}{p-\pi+(1-\pi)r} & \text{if } p > \pi \end{cases}$$

Proof of Unachievable Points Theorem

Theorem

Precision (p) and recall (r) must satisfy

$$p \geq \frac{\pi r}{1 - \pi + \pi r}$$

where $\pi = \frac{pos}{pos+neg}$ *is the proportion of positive examples.*

$$\begin{aligned} p &= \frac{TP}{TP + FP} \\ &\geq \frac{TP}{TP + (1-\pi)n} \\ &= \frac{r\pi n}{r\pi n + (1-\pi)n} \\ &= \frac{r\pi}{r\pi + (1-\pi)} \end{aligned}$$

Proof of Minimum AUCPR Theorem

Theorem

The area of the unachievable region in PR space and the minimum AUCPR, for proportion of positives π, is

$$\mathrm{AUCPR_{MIN}} = 1 + \frac{(1-\pi)\ln(1-\pi)}{\pi}$$

$$\begin{aligned}
\mathrm{AUCPR_{MIN}} &= \int_0^1 \frac{r\pi}{1-\pi+r\pi}\,\mathrm{d}r \\
&= \left.\frac{r\pi + (\pi-1)\ln(\pi(r-1)+1)}{\pi}\right|_{r=0}^{r=1} \\
&= \frac{1}{\pi}(\pi + (\pi-1)(\ln(1) - \ln(1-\pi))) \\
&= 1 + \frac{(1-\pi)\ln(1-\pi)}{\pi}
\end{aligned}$$

Proof of Minimum AP Theorem

Theorem

The minimum average precision, for pos and neg positive and negative examples, respectively, is

$$\mathrm{AP_{MIN}} = \frac{1}{pos} \sum_{i=1}^{pos} \frac{i}{i + neg}$$

$$\begin{aligned}
\mathrm{AP_{MIN}} &= \frac{1}{pos} \sum_{i=1}^{pos} \frac{\frac{\pi i}{pos}}{1 - \pi + \frac{\pi i}{pos}} \\
&= \frac{1}{pos} \sum_{i=1}^{pos} \frac{\frac{pos\, i}{(pos+neg)pos}}{1 + \frac{pos}{pos+neg}(\frac{i}{pos} - 1)} \\
&= \frac{1}{pos} \sum_{i=1}^{pos} \frac{\frac{i}{pos+neg}}{\frac{i+neg}{pos+neg}} = \frac{1}{pos} \sum_{i=1}^{pos} \frac{i}{i + neg}
\end{aligned}$$