Mitigating the Risks of Thresholdless Metrics in Machine Learning Evaluation

Kendrick Boyd

Advisor: David Page

August 1, 2014
Supervised Machine Learning

- Data
- Algorithm
- Evaluation

Image: Jinapattanah via Wikimedia Commons
Supervised Machine Learning

- Data
- Algorithm
- Evaluation

Image: Jinapattanah via Wikimedia Commons
Choices

Question: Which model should I use?
- Learning algorithm
  - SVM or random forests?
- Parameters
  - # of trees in random forest?
- Algorithm internals
  - Keep this rule or discard?

Answer: Evaluation
- How well does the model predict new examples?
Not all methods of generating thresholdless metrics are created equal, and potential pitfalls and benefits accrue based on which methods are chosen.
Not all methods of generating thresholdless metrics are created equal, and potential pitfalls and benefits accrue based on which methods are chosen.

Specific contributions

- Unachievable region in precision-recall (PR) space, Chapter 3 (Boyd, Santos Costa, Davis, and Page, ICML 2012)
- Area under the PR curve estimation, Chapter 4 (Boyd, Eng, and Page, ECML 2013)
- Differentially private evaluation, Chapter 5 (Boyd, Lantz, and Page, under review)
Outline

1. Introduction
2. Evaluation Background
3. AUCPR Estimation
4. Unachievable Region
5. Differentially Private Evaluation
6. Conclusion
## Binary Classification

### Dichotomous Classifiers

<table>
<thead>
<tr>
<th>Id</th>
<th>Actual</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Positive</td>
<td>Positive</td>
</tr>
<tr>
<td>2</td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>3</td>
<td>Positive</td>
<td>Positive</td>
</tr>
<tr>
<td>4</td>
<td>Negative</td>
<td>Positive</td>
</tr>
<tr>
<td>5</td>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td>6</td>
<td>Negative</td>
<td>Negative</td>
</tr>
</tbody>
</table>

- True positive \( (tp) \)
- True negative \( (tn) \)
- False positive \( (fp) \)
- False negative \( (fn) \)

---

**Mitigating the Risks of Thresholdless Metrics**

**Evaluation Background**

August 1, 2014
Dichotomous Classifiers

<table>
<thead>
<tr>
<th>Id</th>
<th>Actual</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Positive</td>
<td>Positive</td>
</tr>
<tr>
<td>2</td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>3</td>
<td>Positive</td>
<td>Positive</td>
</tr>
<tr>
<td>4</td>
<td>Negative</td>
<td>Positive</td>
</tr>
<tr>
<td>5</td>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td>6</td>
<td>Negative</td>
<td>Negative</td>
</tr>
</tbody>
</table>

- True positive ($tp$)
- True negative ($tn$)
- False positive ($fp$)
- False negative ($fn$)
## Binary Classification

**Dichotomous Classifiers**

<table>
<thead>
<tr>
<th>Id</th>
<th>Actual</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Positive</td>
<td>Positive</td>
</tr>
<tr>
<td>2</td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>3</td>
<td>Positive</td>
<td>Positive</td>
</tr>
<tr>
<td>4</td>
<td>Negative</td>
<td>Positive</td>
</tr>
<tr>
<td>5</td>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td>6</td>
<td>Negative</td>
<td>Negative</td>
</tr>
</tbody>
</table>

- True positive ($tp$)
- True negative ($tn$)
- False positive ($fp$)
- False negative ($fn$)

### Mitigating the Risks of Thresholdless Metrics
# Binary Classification

## Dichotomous Classifiers

<table>
<thead>
<tr>
<th>Id</th>
<th>Actual</th>
<th>Predicted</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Positive</td>
<td>Positive</td>
<td>True positive ((tp))</td>
</tr>
<tr>
<td>2</td>
<td>Positive</td>
<td>Negative</td>
<td>True negative ((tn))</td>
</tr>
<tr>
<td>3</td>
<td>Positive</td>
<td>Positive</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Negative</td>
<td>Positive</td>
<td>False positive ((fp))</td>
</tr>
<tr>
<td>5</td>
<td>Negative</td>
<td>Negative</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Negative</td>
<td>Negative</td>
<td>False negative ((fn))</td>
</tr>
</tbody>
</table>

- True positive (\(tp\))
- True negative (\(tn\))
- False positive (\(fp\))
- False negative (\(fn\))
Binary Classification

Dichotomous Classifiers

<table>
<thead>
<tr>
<th>Id</th>
<th>Actual</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Positive</td>
<td>Positive</td>
</tr>
<tr>
<td>2</td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>3</td>
<td>Positive</td>
<td>Positive</td>
</tr>
<tr>
<td>4</td>
<td>Negative</td>
<td>Positive</td>
</tr>
<tr>
<td>5</td>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td>6</td>
<td>Negative</td>
<td>Negative</td>
</tr>
</tbody>
</table>

- True positive \((tp)\)
- True negative \((tn)\)
- False positive \((fp)\)
- False negative \((fn)\)
### Binary Classification

#### Dichotomous Classifiers

<table>
<thead>
<tr>
<th>Id</th>
<th>Actual</th>
<th>Predicted</th>
<th>True positive ($tp$)</th>
<th>True negative ($tn$)</th>
<th>False positive ($fp$)</th>
<th>False negative ($fn$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Positive</td>
<td>Positive</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Positive</td>
<td>Negative</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Positive</td>
<td>Positive</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Negative</td>
<td>Positive</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Negative</td>
<td>Negative</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Negative</td>
<td>Negative</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Confusion Matrix

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
</tr>
<tr>
<td>Positive</td>
<td>$tp$</td>
</tr>
<tr>
<td>Negative</td>
<td>$fn$</td>
</tr>
<tr>
<td>Total</td>
<td>$n$</td>
</tr>
</tbody>
</table>

**Notation**
- $n$: # of positive examples
- $m$: # of negative examples
- $\pi = \frac{n}{n+m}$: proportion of positives (skew)

**Metrics**
- **Accuracy**: $\frac{tp+tn}{n+m}$
- **True positive rate**: $\frac{tp}{n}$
- **False positive rate**: $\frac{fp}{m}$
- **Precision**: $\frac{tp}{tp+fp}$

---

**Mitigating the Risks of Thresholdless Metrics**

**Evaluation Background**

August 1, 2014
## Confusion Matrix

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
</tr>
<tr>
<td>Positive</td>
<td>$tp$</td>
</tr>
<tr>
<td>Negative</td>
<td>$fn$</td>
</tr>
<tr>
<td>Total</td>
<td>$n$</td>
</tr>
</tbody>
</table>

### Notation
- $n$: # of positive examples
- $m$: # of negative examples
- $\pi = \frac{n}{n+m}$: proportion of positives (skew)

### Metrics
- **Accuracy**: $\frac{tp+tn}{n+m}$
- **True positive rate**: $\frac{tp}{n}$
- **False positive rate**: $\frac{fp}{m}$
- **Precision**: $\frac{tp}{tp+fp}$
## Binary Classification

### Scoring Classifier

<table>
<thead>
<tr>
<th>Id</th>
<th>Label</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Positive</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>Positive</td>
<td>0.8</td>
</tr>
<tr>
<td>3</td>
<td>Negative</td>
<td>0.7</td>
</tr>
<tr>
<td>4</td>
<td>Positive</td>
<td>0.6</td>
</tr>
<tr>
<td>5</td>
<td>Negative</td>
<td>0.5</td>
</tr>
<tr>
<td>6</td>
<td>Negative</td>
<td>0.3</td>
</tr>
<tr>
<td>7</td>
<td>Positive</td>
<td>0.2</td>
</tr>
<tr>
<td>8</td>
<td>Negative</td>
<td>0.1</td>
</tr>
</tbody>
</table>

**Actual**

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Negative</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

- **Accuracy** $= \frac{6}{8}$
- **TPR** $= \frac{2}{4}$
- **FPR** $= \frac{0}{4}$
- **Precision** $= \frac{2}{2}$
### Scoring Classifier

<table>
<thead>
<tr>
<th>Id</th>
<th>Label</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Positive</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>Positive</td>
<td>0.8</td>
</tr>
<tr>
<td>3</td>
<td>Negative</td>
<td>0.7</td>
</tr>
<tr>
<td>4</td>
<td>Positive</td>
<td>0.6</td>
</tr>
<tr>
<td>5</td>
<td>Negative</td>
<td>0.5</td>
</tr>
<tr>
<td>6</td>
<td>Negative</td>
<td>0.3</td>
</tr>
<tr>
<td>7</td>
<td>Positive</td>
<td>0.2</td>
</tr>
<tr>
<td>8</td>
<td>Negative</td>
<td>0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Positive</td>
</tr>
<tr>
<td>Negative</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
</tr>
</tbody>
</table>

- **Accuracy** = \( \frac{5}{8} \)
- **TPR** = \( \frac{4}{4} \)
- **FPR** = \( \frac{3}{4} \)
- **Precision** = \( \frac{4}{7} \)
Thresholdless Metrics

Evaluate model without choosing a specific threshold

- Receiver operating characteristic (ROC) curves (Provost, Fawcett, et al., 1997)
  - Area under the ROC curve (AUCROC)
- Precision-recall (PR) curves (V Raghavan, Bollmann, and Jung, 1989)
  - Area under the PR curve (AUCPR)
- Lift curves (Piatetsky-Shapiro and Masand, 1999)
- Cost curves (Drummond and Holte, 2006)
- Brier curves (Ferri, Hernández-Orallo, and Flach, 2011)
ROC Curves

- Ideal (0,1)
- All negative (0,0)
- All positive (1,1)

Perfect Model
Good Model
Random Model

Mitigating the Risks of Thresholdless Metrics

Evaluation Background
August 1, 2014
PR Curves

- Perfect Model
  - (1,1) ideal
- Good Model
  - (1,0.1) all positive
- Random Model

Evaluation Background

Mitigating the Risks of Thresholdless Metrics
Outline

1. Introduction
2. Evaluation Background
3. AUCPR Estimation
4. Unachievable Region
5. Differentially Private Evaluation
6. Conclusion
Empirical PR Points

- 5 positives \((n)\)
- 15 negatives \((m)\)
- \(\pi = 0.25\)

<table>
<thead>
<tr>
<th>score</th>
<th>true label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>Positive</td>
</tr>
<tr>
<td>0.95</td>
<td>Negative</td>
</tr>
<tr>
<td>0.90</td>
<td>Negative</td>
</tr>
<tr>
<td>0.85</td>
<td>Positive</td>
</tr>
<tr>
<td>0.80</td>
<td>Positive</td>
</tr>
<tr>
<td>0.75</td>
<td>Negative</td>
</tr>
<tr>
<td>0.70</td>
<td>Negative</td>
</tr>
<tr>
<td>0.65</td>
<td>Negative</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
Empirical PR Points

- 5 positives ($n$)
- 15 negatives ($m$)
- $\pi = 0.25$

<table>
<thead>
<tr>
<th>score</th>
<th>true label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>Positive</td>
</tr>
<tr>
<td>0.95</td>
<td>Negative</td>
</tr>
<tr>
<td>0.90</td>
<td>Negative</td>
</tr>
<tr>
<td>0.85</td>
<td>Positive</td>
</tr>
<tr>
<td>0.80</td>
<td>Positive</td>
</tr>
<tr>
<td>0.75</td>
<td>Negative</td>
</tr>
<tr>
<td>0.70</td>
<td>Negative</td>
</tr>
<tr>
<td>0.65</td>
<td>Negative</td>
</tr>
</tbody>
</table>

Recall = $\frac{1}{5}$

Precision = $\frac{1}{1}$
Empirical PR Points

- 5 positives \((n)\)
- 15 negatives \((m)\)
- \(\pi = 0.25\)

<table>
<thead>
<tr>
<th>score</th>
<th>true label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>Positive</td>
</tr>
<tr>
<td>0.95</td>
<td>Negative</td>
</tr>
<tr>
<td>0.90</td>
<td>Negative</td>
</tr>
<tr>
<td>0.85</td>
<td>Positive</td>
</tr>
<tr>
<td>0.80</td>
<td>Positive</td>
</tr>
<tr>
<td>0.75</td>
<td>Negative</td>
</tr>
<tr>
<td>0.70</td>
<td>Negative</td>
</tr>
<tr>
<td>0.65</td>
<td>Negative</td>
</tr>
</tbody>
</table>

Recall = \(\frac{1}{5}\)

Precision = \(\frac{1}{2}\)
Empirical PR Points

- 5 positives ($n$)
- 15 negatives ($m$)
- $\pi = 0.25$

<table>
<thead>
<tr>
<th>score</th>
<th>true label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>Positive</td>
</tr>
<tr>
<td>0.95</td>
<td>Negative</td>
</tr>
<tr>
<td>0.90</td>
<td>Negative</td>
</tr>
<tr>
<td>0.85</td>
<td>Positive</td>
</tr>
<tr>
<td>0.80</td>
<td>Positive</td>
</tr>
<tr>
<td>0.75</td>
<td>Negative</td>
</tr>
<tr>
<td>0.70</td>
<td>Negative</td>
</tr>
<tr>
<td>0.65</td>
<td>Negative</td>
</tr>
</tbody>
</table>

Recall = $\frac{1}{5}$
Precision = $\frac{1}{3}$
Empirical PR Points

- 5 positives \((n)\)
- 15 negatives \((m)\)
- \(\pi = 0.25\)

<table>
<thead>
<tr>
<th>score</th>
<th>true label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>Positive</td>
</tr>
<tr>
<td>0.95</td>
<td>Negative</td>
</tr>
<tr>
<td>0.90</td>
<td>Negative</td>
</tr>
<tr>
<td>0.85</td>
<td>Positive</td>
</tr>
<tr>
<td>0.80</td>
<td>Positive</td>
</tr>
<tr>
<td>0.75</td>
<td>Negative</td>
</tr>
<tr>
<td>0.70</td>
<td>Negative</td>
</tr>
<tr>
<td>0.65</td>
<td>Negative</td>
</tr>
</tbody>
</table>

\[
\text{Recall} = \frac{2}{5} \\
\text{Precision} = \frac{2}{4}
\]
Empirical PR Points

- 5 positives \((n)\)
- 15 negatives \((m)\)
- \(\pi = 0.25\)

<table>
<thead>
<tr>
<th>score</th>
<th>true label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>Positive</td>
</tr>
<tr>
<td>0.95</td>
<td>Negative</td>
</tr>
<tr>
<td>0.90</td>
<td>Negative</td>
</tr>
<tr>
<td>0.85</td>
<td>Positive</td>
</tr>
<tr>
<td>0.80</td>
<td>Positive</td>
</tr>
<tr>
<td>0.75</td>
<td>Negative</td>
</tr>
<tr>
<td>0.70</td>
<td>Negative</td>
</tr>
<tr>
<td>0.65</td>
<td>Negative</td>
</tr>
</tbody>
</table>

\[
\text{Recall} = \frac{3}{5} \\
\text{Precision} = \frac{3}{5}
\]
Empirical PR Points

- 5 positives ($n$)
- 15 negatives ($m$)
- $\pi = 0.25$

<table>
<thead>
<tr>
<th>score</th>
<th>true label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>Positive</td>
</tr>
<tr>
<td>0.95</td>
<td>Negative</td>
</tr>
<tr>
<td>0.90</td>
<td>Negative</td>
</tr>
<tr>
<td>0.85</td>
<td>Positive</td>
</tr>
<tr>
<td>0.80</td>
<td>Positive</td>
</tr>
<tr>
<td>0.75</td>
<td>Negative</td>
</tr>
<tr>
<td>0.70</td>
<td>Negative</td>
</tr>
<tr>
<td>0.65</td>
<td>Negative</td>
</tr>
</tbody>
</table>

Recall $= \frac{3}{5}$

Precision $= \frac{3}{6}$
Empirical PR Points

- 5 positives ($n$)
- 15 negatives ($m$)
- $\pi = 0.25$

<table>
<thead>
<tr>
<th>score</th>
<th>true label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>Positive</td>
</tr>
<tr>
<td>0.95</td>
<td>Negative</td>
</tr>
<tr>
<td>0.90</td>
<td>Negative</td>
</tr>
<tr>
<td>0.85</td>
<td>Positive</td>
</tr>
<tr>
<td>0.80</td>
<td>Positive</td>
</tr>
<tr>
<td>0.75</td>
<td>Negative</td>
</tr>
<tr>
<td>0.70</td>
<td>Negative</td>
</tr>
<tr>
<td>0.65</td>
<td>Negative</td>
</tr>
</tbody>
</table>

Recall = $\frac{3}{5}$

Precision = $\frac{3}{6}$
The Challenge

Given

Do

- Estimate area: AUCPR = 0.5
- Calculate confidence interval: [0.4, 0.6]

Our Goal
Empirically evaluate point estimates and confidence intervals of AUCPR to identify their differences and recommend best practices.
The Challenge

Given

- Estimate area: AUCPR = 0.5
- Calculate confidence interval: [0.4, 0.6]

Our Goal

Empirically evaluate point estimates and confidence intervals of AUCPR to identify their differences and recommend best practices.
Many existing methods to estimate AUCPR:
AUCPR Estimators

Many existing methods to estimate AUCPR: (Abeel, Van de Peer, and Saeys, 2009)

Mitigating the Risks of Thresholdless Metrics
Many existing methods to estimate AUCPR: (Abeel, Van de Peer, and Saey, 2009)
AUCPR Estimators

Many existing methods to estimate AUCPR: (Manning, P Raghavan, and Schütze, 2008)
AUCPR Estimators

Many existing methods to estimate AUCPR: (Davis and Goadrich, 2006)
Many existing methods to estimate AUCPR:
Many existing methods to estimate AUCPR: (Brodersen et al., 2010)
Many existing methods to estimate AUCPR:
Estimator Desiderata

- Unbiased: expected estimate is equal to true AUCPR
- Robust to different output distributions
- Robust to various skews ($\pi$) and data set sizes ($n + m$)
Simulation Setup

- Assume example scores are drawn from known distributions
  - Similar to binormal analysis on ROC curves (Pepe, 2004; Bamber, 1975)
  - Allows calculation of true PR curve and AUCPR

- Analyzed distributions
  - Binormal
  - Bibeta
  - Offset uniform

- Additional parameters
  - # of examples $(n + m)$
  - skew ($\pi = 0.1$)
Simulation Setup

- Assume example scores are drawn from known distributions
  - Similar to binormal analysis on ROC curves (Pepe, 2004; Bamber, 1975)
  - Allows calculation of true PR curve and AUCPR
- Analyzed distributions
  - Binormal
  - Bibeta
  - Offset uniform
- Additional parameters
  - # of examples \((n + m)\)
  - skew \((\pi = 0.1)\)

**Binormal**

\[
\text{negative} \sim \text{Normal}(0, 1) \\
\text{positive} \sim \text{Normal}(1, 1)
\]
Simulation Setup

- Assume example scores are drawn from known distributions
  - Similar to binormal analysis on ROC curves (Pepe, 2004; Bamber, 1975)
  - Allows calculation of true PR curve and AUCPR
- Analyzed distributions
  - Binormal
  - Bibeta
  - Offset uniform
- Additional parameters
  - # of examples \((n + m)\)
  - skew \((\pi = 0.1)\)

![Bibeta Distribution](image)

**negative \sim Beta(2, 5)**
**positive \sim Beta(5, 2)**
Simulation Setup

- Assume example scores are drawn from known distributions
  - Similar to binormal analysis on ROC curves (Pepe, 2004; Bamber, 1975)
  - Allows calculation of true PR curve and AUCPR
- Analyzed distributions
  - Binormal
  - Bibeta
  - Offset uniform
- Additional parameters
  - # of examples \((n + m)\)
  - skew \((\pi = 0.1)\)

Negative \(\sim\) Uniform\((0, 1)\)
Positive \(\sim\) Uniform\((0.5, 1.5)\)
Simulation Setup

- Assume example scores are drawn from known distributions
  - Similar to binormal analysis on ROC curves (Pepe, 2004; Bamber, 1975)
  - Allows calculation of true PR curve and AUCPR
- Analyzed distributions
  - Binormal
  - Bibeta
  - Offset uniform
- Additional parameters
  - # of examples ($n + m$)
  - skew ($\pi = 0.1$)
AUCPR Estimator Results

Mitigating the Risks of Thresholdless Metrics

August 1, 2014
AUCPR Estimator Results

Mitigating the Risks of Thresholdless Metrics
Definition

A \((1 - \alpha)\)% confidence interval is an interval that contains the true value with probability at least \((1 - \alpha)\).

Desiderata

- Valid - at least \((1 - \alpha)\)% coverage
- Prefer narrower (but must still be valid)
- Robust to different output distributions
- Robust to various skews \((\pi)\) and data sizes \((n + m)\)
# Confidence Intervals

## Definition

A $(1 - \alpha)\%$ confidence interval is an interval that contains the true value with probability at least $(1 - \alpha)$.

## Desiderata

- Valid - at least $(1 - \alpha)\%$ coverage
- Prefer narrower (but must still be valid)
- Robust to different output distributions
- Robust to various skews ($\pi$) and data sizes ($n + m$)
AUCPR Confidence Intervals

- **Empirical**
  - Cross-validation: compute interval using mean and variance of $K$ estimates from $K$-fold cross-validation
  - Bootstrap: choose interval that contains $(1 - \alpha)\%$ of empirical distribution of AUCPR estimates

- **Parametric**
  - Binomial: $\hat{\theta} \pm \Phi_{1-\alpha/2} \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$
    - $\hat{\theta}$ is the estimated AUCPR
  - Logit: $\left[ \frac{e^{\hat{\eta}-\Phi(1-\alpha/2)\hat{\tau}}}{1+e^{\hat{\eta}-\Phi(1-\alpha/2)\hat{\tau}}}, \frac{e^{\hat{\eta}+\Phi(1-\alpha/2)\hat{\tau}}}{1+e^{\hat{\eta}+\Phi(1-\alpha/2)\hat{\tau}}} \right]$
    - $\hat{\eta} = \log \frac{\hat{\theta}}{1-\hat{\theta}}$, $\hat{\tau} = (n\hat{\theta}(1-\hat{\theta}))^{-1/2}$
AUCPR Confidence Intervals

- **Empirical**
  - Cross-validation: compute interval using mean and variance of K estimates from K-fold cross-validation
  - Bootstrap: choose interval that contains \((1 - \alpha)\)% of empirical distribution of AUCPR estimates

- **Parametric**
  - Binomial: \(\hat{\theta} \pm \Phi_{1-\alpha/2} \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\)
    - \(\hat{\theta}\) is the estimated AUCPR
  - Logit: \(\left[\frac{e^{\hat{\eta}} - \Phi(1-\alpha/2)\hat{\tau}}{1+e^{\hat{\eta}} - \Phi(1-\alpha/2)\hat{\tau}}, \frac{e^{\hat{\eta} + \Phi(1-\alpha/2)\hat{\tau}}}{1+e^{\hat{\eta} + \Phi(1-\alpha/2)\hat{\tau}}}\right]\)
    - \(\hat{\eta} = \log \frac{\hat{\theta}}{1-\hat{\theta}}, \hat{\tau} = \left(n\hat{\theta}(1-\hat{\theta})\right)^{-1/2}\)
AUCPR Confidence Intervals

- **Empirical**
  - Cross-validation: compute interval using mean and variance of K estimates from K-fold cross-validation
  - Bootstrap: choose interval that contains \((1 - \alpha)\)% of empirical distribution of AUCPR estimates

- **Parametric**
  - Binomial: \(\hat{\theta} \pm \Phi_{1-\alpha/2} \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\)
    \(\hat{\theta}\) is the estimated AUCPR
  - Logit:
    \[
    \left[ \frac{e^{\hat{\eta}-\Phi(1-\alpha/2)\hat{\tau}}}{1+e^{\hat{\eta}-\Phi(1-\alpha/2)\hat{\tau}}}, \frac{e^{\hat{\eta}+\Phi(1-\alpha/2)\hat{\tau}}}{1+e^{\hat{\eta}+\Phi(1-\alpha/2)\hat{\tau}}} \right]
    \]
    \(\hat{\eta} = \log \frac{\hat{\theta}}{1-\hat{\theta}}, \hat{\tau} = (n\hat{\theta}(1-\hat{\theta}))^{-1/2}\)
AUCPR Confidence Intervals

- Empirical
  - Cross-validation: compute interval using mean and variance of $K$ estimates from $K$-fold cross-validation
  - Bootstrap: choose interval that contains $(1 - \alpha)\%$ of empirical distribution of AUCPR estimates

- Parametric
  - Binomial: $\hat{\theta} \pm \Phi_{1-\alpha/2} \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$
    $\hat{\theta}$ is the estimated AUCPR
  - Logit: $\begin{bmatrix} e^{\hat{\eta} - \Phi(1-\alpha/2)\hat{\tau}} & e^{\hat{\eta} + \Phi(1-\alpha/2)\hat{\tau}} \\ 1 + e^{\hat{\eta} - \Phi(1-\alpha/2)\hat{\tau}} & 1 + e^{\hat{\eta} + \Phi(1-\alpha/2)\hat{\tau}} \end{bmatrix}$
    $\hat{\eta} = \log \frac{\hat{\theta}}{1-\hat{\theta}}, \hat{\tau} = (n\hat{\theta}(1 - \hat{\theta}))^{-1/2}$
AUCPR Confidence Interval Results

Mitigating the Risks of Thresholdless Metrics

AUCPR Estimation

August 1, 2014
Mitigating the Risks of Thresholdless Metrics

AUCPR Estimation

August 1, 2014
Mitigating the Risks of Thresholdless Metrics

AUCPR Estimation

August 1, 2014

21
AUCPR Estimator Results

Mitigating the Risks of Thresholdless Metrics

August 1, 2014
Choice of AUCPR estimator and confidence interval is important

- Particularly for small data sets

**Recommended estimators**
- Lower trapezoid
- Average precision
- Interpolated median

**Recommended confidence intervals**
- Binomial
- Logit

What about cross-validation and bootstrap?
- Converge to proper coverage, but from below
- Problematic for small data sets and low numbers of positive examples
AUCPR Summary

- Choice of AUCPR estimator and confidence interval is important
  - Particularly for small data sets

- Recommended estimators
  - Lower trapezoid
  - Average precision
  - Interpolated median

- Recommended confidence intervals
  - Binomial
  - Logit

- What about cross-validation and bootstrap?
  - Converge to proper coverage, but from below
  - Problematic for small data sets and low numbers of positive examples
Choice of AUCPR estimator and confidence interval is important
- Particularly for small data sets

Recommended estimators
- Lower trapezoid
- Average precision
- Interpolated median

Recommended confidence intervals
- Binomial
- Logit

What about cross-validation and bootstrap?
- Converge to proper coverage, but from below
- Problematic for small data sets and low numbers of positive examples
Choice of AUCPR estimator and confidence interval is important
  Particularly for small data sets

Recommended estimators
  Lower trapezoid
  Average precision
  Interpolated median

Recommended confidence intervals
  Binomial
  Logit
  What about cross-validation and bootstrap?
    Converge to proper coverage, but from below
    Problematic for small data sets and low numbers of positive examples
1. Introduction

2. Evaluation Background

3. AUCPR Estimation

4. Unachievable Region

5. Differentially Private Evaluation

6. Conclusion
Unachievable Region in PR Space

Mitigating the Risks of Thresholdless Metrics

August 1, 2014
Outline

1. Introduction
2. Evaluation Background
3. AUCPR Estimation
4. Unachievable Region
5. Differentially Private Evaluation
6. Conclusion
Attacks on Evaluation Metrics

Can evaluation metrics disclose private information?
Can evaluation metrics disclose private information?

Yes!
Can evaluation metrics disclose private information?

Yes!

Disclosive methods
- Empirical ROC curves (Matthews and Harel, 2013)
- AUCROC (Section 5.2)

Information leaked
- Class label
- Score range from model (e.g., risk of disease)
Need for Privacy

- Large databases of patient information
  - Regulations and expectations of privacy
  - Enormous potential gains from data mining
  - How to allow useful interaction with a database while preserving privacy?

Privacy frameworks
- k-anonymity (Sweeney, 2002)
- Differential privacy (Dwork, 2006)
Need for Privacy

- Large databases of patient information
  - Regulations and expectations of privacy
  - Enormous potential gains from data mining
  - How to allow useful interaction with a database while preserving privacy?
- Privacy frameworks
  - k-anonymity (Sweeney, 2002)
  - Differential privacy (Dwork, 2006)
Mitigating the Risks of Thresholdless Metrics

Differentially Private Evaluation

August 1, 2014

29
Differential Privacy (Dwork, 2006)

**Goal**
Small added risk of adversary learning (private) information about an individual if his or her data is in the private database versus not in the database.

**Informal Definition**
Query output does not change much between neighboring databases.
Goal
Small added risk of adversary learning (private) information about an individual if his or her data is in the private database versus not in the database.

Informal Definition
Query output does not change much between neighboring databases.
Differential Privacy: Formally

Definition (Dwork, 2006)

For any input database $D$, a randomized algorithm $f' : \mathbb{D} \rightarrow \text{Range}(f')$ is $(\epsilon, \delta)$-differentially private iff for any $S \subset \text{Range}(f)$ and any database $D' \in \mathbb{D}$ where $d(D, D') = 1$,\[\Pr(f'(D) \in S) \leq e^\epsilon \Pr(f'(D') \in S) + \delta\]

- $d(D, D')$ - number of rows that differ between $D$ and $D'$
- $\epsilon$ and $\delta$ are the privacy budget
  - Smaller means more private
  - If $\delta = 0$, known as $\epsilon$-differential privacy
Obtaining Differential Privacy

- **Perturbation (Dwork, 2006)**
  - Calculate correct answer: $f(D)$
  - Add noise: $f(D) + \eta$

- **Soft-max (McSherry and Talwar, 2007)**
  - Quality function: $q(D, s)$
  - Exponential weighting: $\exp(\epsilon q(D, s))$

- **Extensions**
  - Propose-test-release (Dwork and Lei, 2009)
  - $\beta$-smooth sensitivity (Nissim, Raskhodnikova, and Smith, 2007)
Perturbation (Dwork, 2006)
- Calculate correct answer: $f(D)$
- Add noise: $f(D) + \eta$

Soft-max (McSherry and Talwar, 2007)
- Quality function: $q(D, s)$
- Exponential weighting: $\exp(\epsilon q(D, s))$

**Extensions**
- Propose-test-release (Dwork and Lei, 2009)
- $\beta$-smooth sensitivity (Nissim, Raskhodnikova, and Smith, 2007)
**Global Sensitivity**

**Definition (Dwork, 2006)**

Given a function \( f : D \rightarrow R \), the global sensitivity of \( f \) is,

\[
GS_f = \max_{D, D' \in D : d(D, D') = 1} |f(D) - f(D')|
\]

- Worst case
- Once \( D \) and \( f \) are chosen, global sensitivity is fixed
Global Sensitivity

Definition (Dwork, 2006)

Given a function \( f : \mathbb{D} \to \mathbb{R} \), the global sensitivity of \( f \) is,

\[
GS_f = \max_{D,D' \in \mathbb{D} : d(D,D') = 1} |f(D) - f(D')|
\]

- Worst case
- Once \( \mathbb{D} \) and \( f \) are chosen, global sensitivity is fixed
Laplace Noise

Theorem (Dwork, 2006)

Given a function $f : \mathbb{D} \rightarrow \mathbb{R}$, the computation

$$f'(D) = f(D) + \text{Laplace} \left( \frac{GS_f}{\epsilon} \right)$$

guarantees $\epsilon$-differential privacy.
Local Sensitivity

Example

Median

- For most databases, barely affected by changing a value
- But worst case change is large

Definition (Nissim, Raskhodnikova, and Smith, 2007)

Given a function $f : \mathbb{D} \rightarrow \mathbb{R}$, the local sensitivity of $f$ at $D \in \mathbb{D}$ is

$$LS_f(D) = \max_{D' \in \mathbb{D} : d(D, D') = 1} |f(D) - f(D')|.$$
Local Sensitivity

Example

Median
- For most databases, barely affected by changing a value
- But worst case change is large

Definition (Nissim, Raskhodnikova, and Smith, 2007)
Given a function \( f : \mathbb{D} \rightarrow \mathbb{R} \), the \textit{local sensitivity} of \( f \) at \( D \in \mathbb{D} \) is

\[
LS_f(D) = \max_{D' \in \mathbb{D} : d(D, D') = 1} |f(D) - f(D')|.
\]
Using Local Sensitivity

- Local sensitivity is not a direct replacement for global sensitivity

**Definition (Nissim, Raskhodnikova, and Smith, 2007)**

For $\beta > 0$, a function $S : \mathbb{D} \rightarrow \mathbb{R}^+$ is a $\beta$-smooth upper bound on the local sensitivity of $f$ iff it satisfies:

$$\forall D \in \mathbb{D} : S(D) \geq LS_f(D) \text{ and }$$

$$\forall D, D' \in \mathbb{D}, d(D, D') = 1 : S(D) \leq e^\beta S(D')$$

- A $\beta$-smooth upper bound ensures neighboring databases will use a similar scale of noise
- $\beta$-smooth sensitivity is the smallest such function
- Modified perturbation algorithms can use $\beta$-smooth sensitivity
  - Laplace noise provides $(\epsilon, \delta)$-differential privacy
  - Cauchy noise provides $\epsilon$-differential privacy
Using Local Sensitivity

- Local sensitivity is not a direct replacement for global sensitivity

**Definition (Nissim, Raskhodnikova, and Smith, 2007)**

For $\beta > 0$, a function $S : \mathbb{D} \to \mathbb{R}^+$ is a $\beta$-smooth upper bound on the local sensitivity of $f$ iff it satisfies:

$$
\forall D \in \mathbb{D} : S(D) \geq LS_f(D) \quad \text{and} \quad \forall D, D' \in \mathbb{D}, d(D, D') = 1 : S(D) \leq e^\beta S(D')
$$

- A $\beta$-smooth upper bound ensures neighboring databases will use a similar scale of noise
- $\beta$-smooth sensitivity is the smallest such function
- Modified perturbation algorithms can use $\beta$-smooth sensitivity
  - Laplace noise provides $(\epsilon, \delta)$-differential privacy
  - Cauchy noise provides $\epsilon$-differential privacy
Using Local Sensitivity

- Local sensitivity is not a direct replacement for global sensitivity

**Definition (Nissim, Raskhodnikova, and Smith, 2007)**

For \( \beta > 0 \), a function \( S : \mathbb{D} \rightarrow \mathbb{R}^+ \) is a \( \beta \)-smooth upper bound on the local sensitivity of \( f \) iff it satisfies:

\[
\forall D \in \mathbb{D} : \quad S(D) \geq LS_f(D) \quad \text{and} \\
\forall D, D' \in \mathbb{D}, d(D, D') = 1 : \quad S(D) \leq e^{\beta} S(D')
\]

- A \( \beta \)-smooth upper bound ensures neighboring databases will use a similar scale of noise
  - \( \beta \)-smooth sensitivity is the smallest such function
- Modified perturbation algorithms can use \( \beta \)-smooth sensitivity
  - Laplace noise provides \((\epsilon, \delta)\)-differential privacy
  - Cauchy noise provides \( \epsilon \)-differential privacy
Existing applications of differential privacy

- Consistent marginals (Barak et al., 2007)
- PAC learning (Kasiviswanathan et al., 2011)
- Learning algorithms (Blum et al., 2005; Nissim, Raskhodnikova, and Smith, 2007; Dwork and Lei, 2009; Zhang et al., 2012)
- Auctions (McSherry and Talwar, 2007)

Our Application: Evaluation

No previous usage of differential privacy specifically to the release of evaluation metrics after testing a model on a private database.
Privacy Applications

Existing applications of differential privacy

- Consistent marginals (Barak et al., 2007)
- PAC learning (Kasiviswanathan et al., 2011)
- Learning algorithms (Blum et al., 2005; Nissim, Raskhodnikova, and Smith, 2007; Dwork and Lei, 2009; Zhang et al., 2012)
- Auctions (McSherry and Talwar, 2007)

Our Application: Evaluation

No previous usage of differential privacy specifically to the release of evaluation metrics after testing a model on a private database.
Private Evaluation Setup

Private metrics
- Accuracy is a simple application of Laplace noise
- AUCROC (Section 5.4)
- Average precision (Section 5.5)
Local Sensitivity of AUCROC

**Theorem**

\[ LS_{AUCROC}(n, m) = \begin{cases} \frac{1}{\min(n,m)} & \text{if } n > 1 \text{ and } m > 1 \\ 1 & \text{otherwise} \end{cases} \]

- \( n \) - number of positive examples in test set
- \( m \) - number of negative examples in test set
$\beta$-smooth Sensitivity of AUCROC

Mitigating the Risks of Thresholdless Metrics

Differentially Private Evaluation

August 1, 2014
\( \beta \)-smooth Sensitivity of AUCROC
Private Evaluation Experiments

- Adult data set (Bache and Lichman, 2013)
  - Predict yearly income greater or less than $50,000
  - Features: capitol gain/loss, work status

- Procedure
  - Train logistic regression model on half of the data
  - Calculate private metric on subsets of other half
  - Compare with non-private metric (RMSE)
Private AUCROC Results

Mitigating the Risks of Thresholdless Metrics

Differentially Private Evaluation

August 1, 2014
Privacy of test sets
- Necessary due to demonstrated attacks on ROC curves
- Just as important as privacy of train sets

Private evaluation metrics
- Confusion matrix based metrics (accuracy, recall, etc.)
- AUCROC
- Average precision
Outline

1. Introduction
2. Evaluation Background
3. AUCPR Estimation
4. Unachievable Region
5. Differentially Private Evaluation
6. Conclusion
Future Work

- Unachievable region in PR space
  - PR curve and AUCPR aggregation across different skews
- AUCPR estimation
  - Less biased AUCPR estimators for small data sets
  - Tighter parametric AUCPR confidence intervals
- Differentially private evaluation
  - Private ROC and PR curves
  - Private cross-validation mechanisms
Not all methods of generating thresholdless metrics are created equal, and potential pitfalls and benefits accrue based on which methods are chosen.
Unachievable region in PR space

Recommendations

- Show unachievable region in PR curve plots
- Report skew with PR metrics (PR curve, AUCPR, $F_1$)
- Be aware of changing skew and aggregating from different skews
Specific Contributions

AUCPR estimators and confidence intervals

Recommendations

- Choose estimator and interval methods carefully based on task
- Default to average precision, lower trapezoid, or interpolated median estimators
- Default to binomial and logit confidence intervals
- Be aware of the tendencies of bootstrap and cross-validation
Specific Contributions

Differentially private evaluation

Recommendations

- Be aware that evaluation metrics can disclose private information
- Use private versions of evaluation algorithms
  - Accuracy, sensitivity, specificity, etc.
  - AUCROC
  - Average precision

<table>
<thead>
<tr>
<th>epsilon</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.5</td>
</tr>
<tr>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>1.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>n</th>
<th>30</th>
<th>100</th>
<th>300</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>delta=0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>delta=1e-6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>delta=1e-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>delta=0.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Thank You

Questions?

Acknowledgments

- Advisor
  - David Page

- Committee
  - Mark Craven
  - Jeffrey Naughton
  - Jude Shavlik
  - Jerry Zhu

- Funding
  - NIH grant 5T15LM007359
  - NIGMS grant R01GM097618
  - NLM grant R01LM011028
  - NIH grant 8466993
  - NIH grant 8531347

- Coauthors
  - Jesse Davis
  - Kevin Eng
  - Eric Lantz
  - Vítor Santos Costa

- Colleagues
  - Debbie Chasman
  - Alex Cobian
  - Finn Kuusisto
  - Jie Liu
  - Deborah Muganda
  - Jeremy Weiss

- Family

Mitigating the Risks of Thresholdless Metrics

Conclusion

August 1, 2014


Bamber, Donald (1975). “The area above the ordinal dominance graph and the area below the receiver operating characteristic graph”. In: Journal of Mathematical Psychology 12.4, pp. 387–415.


Mitigating the Risks of Thresholdless Metrics
ROC/PR Space Interpolation

\[ \pi = 0.3 \]

(Davis and Goadrich, 2006)
ROC/PR Space Interpolation

\[ \pi = 0.3 \]

(Davis and Goadrich, 2006)
ROC/PR Space Interpolation

\[ \pi = 0.3 \]

(Davis and Goadrich, 2006)
Theorem (Boyd, Eng, and Page, 2013)

For two points, \((r_1, p_1)\) and \((r_2, p_2)\), in PR space, the interpolated curve and \(r'\) is

\[
p' = \frac{r'}{ar' + b}
\]

and the area under the interpolated curve between \(r_1\) and \(r_2\) is

\[
\frac{ar_2 - b \log(ar_2 + b) - ar_1 + b \log(ar_1 + b)}{a^2}
\]

where

\[
a = 1 + \frac{(1 - p_2)r_2}{p_2(r_2 - r_1)} - \frac{(1 - p_1)r_1}{p_1(r_2 - r_1)}
\]

\[
b = \frac{(1 - p_1)r_1}{p_1} - \frac{(1 - p_2)r_1r_2}{p_2(r_2 - r_1)} + \frac{(1 - p_1)r_1^2}{p_1(r_2 - r_1)}
\]
AUCPR Estimator Results by Skew

Mitigating the Risks of Thresholdless Metrics
AUCPR Confidence Interval Widths

Mitigating the Risks of Thresholdless Metrics

Backup Slides

August 1, 2014
AUCPR Confidence Interval Locations

- binormal
- bibeta
- offset uniform

- interval: binomial, bootstrap, cross-val
- estimator: avg prec, lower trap, interp med
Bounded versus Unbounded

Differential Privacy

Output does not change much between neighboring databases.

- Bounded: replace value of exactly one row
- Unbounded: add or remove exactly one row

(Kifer and Machanavajjhala, 2011)
Definition (Nissim, Raskhodnikova, and Smith, 2007)

For $\beta > 0$, the $\beta$-smooth sensitivity of $f$ is

$$S^*_{f,\beta}(D) = \max_{D' \in D} LS_f(D') e^{-\beta d(D, D')}$$
Differential Privacy using $\beta$-smooth Sensitivity

Theorem (Nissim, Raskhodnikova, and Smith, 2007)

Let $f : \mathcal{D} \rightarrow \mathbb{R}$ be any real-valued function and let $S : \mathcal{D} \rightarrow \mathbb{R}$ be the $\beta$-smooth sensitivity of $f$, then

1. If $\beta \leq \frac{\epsilon}{2(\gamma+1)}$ and $\gamma > 1$, the algorithm $f'(D) = f(D) + \frac{2(\gamma+1)S(D)}{\epsilon} \eta$, where $\eta$ is sampled from the distribution with density $h(z) \propto \frac{1}{1+|z|^\gamma}$, is $\epsilon$-differentially private. Note that when $\gamma = 2$, $\eta$ is drawn from a standard Cauchy distribution.

2. If $\beta \leq \frac{\epsilon}{2 \ln(\frac{2}{\delta})}$ and $\delta \in (0, 1)$, the algorithm $f'(D) = f(D) + \frac{2S(D)}{\epsilon} \eta$, where $\eta \sim \text{Laplace}(1)$, is $(\epsilon, \delta)$-differentially private.
Local Sensitivity of Average Precision

Theorem

\[ LS_{AP} = \begin{cases} \max \left( \frac{\log(n+1)}{n}, \frac{9+\log(n-1)}{4(n-1)} \right) + \max \left( \frac{\log(n+1)}{n}, \frac{9+\log n}{4n} \right) & \text{if } n > 1 \\ 1 & \text{if } n \leq 1 \end{cases} \]

- \( n \) - number of positive examples in test set
AP Attack Results

Mitigating the Risks of Thresholdless Metrics

Backup Slides

August 1, 2014
**Theorem (Achievable Points)**

An achievable point in PR space with precision $p$ and recall $r$ must satisfy

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

where $\pi = \frac{n}{n+m}$ is the skew.
Theorem (Minimum AUCPR)

The area of the unachievable region in PR space and the minimum AUCPR, for skew \( \pi \), is

\[
\text{AUCPR}_{\text{MIN}} = 1 + \frac{(1 - \pi) \ln(1 - \pi)}{\pi}
\]
Theorem (Minimum AP)

The minimum AP, for a data set with \( n \) positive and \( m \) negative examples is

\[
AP_{\text{MIN}} = \frac{1}{n} \sum_{i=1}^{n} \frac{i}{i + m}
\]
Outline

1. Introduction
2. Evaluation Background
3. AUCPR Estimation
4. Unachievable Region
5. Differentially Private Evaluation
6. Conclusion