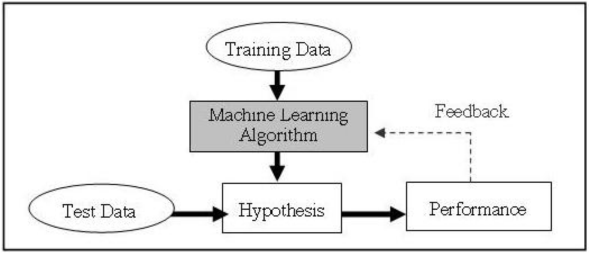
Mitigating the Risks of Thresholdless Metrics in Machine Learning Evaluation

Kendrick Boyd

Advisor: David Page

August 1, 2014

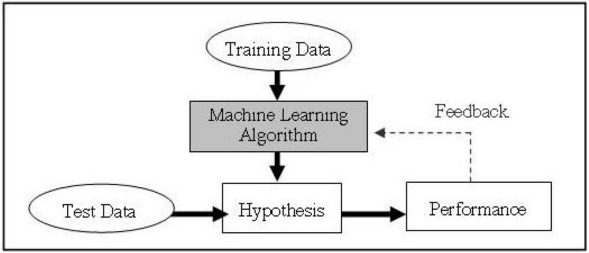
Supervised Machine Learning



- 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?

Image: Ala Fernandez, flickr.com

Thesis Statement

Not all methods of generating thresholdless metrics are created equal, and potential pitfalls and benefits accrue based on which methods are chosen.

Thesis Statement

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

Binary Classification

Dichotomous Classifiers

	Label	
Id	Actual	Predicted
1	Positive	Positive
2	Positive	Negative
3	Positive	Positive
4	Negative	Positive
5	Negative	Negative
6	Negative	Negative

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

Binary Classification

Dichotomous Classifiers

	Label	
Id	Actual	Predicted
1	Positive	Positive
2	Positive	Negative
3	Positive	Positive
4	Negative	Positive
5	Negative	Negative
6	Negative	Negative

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

Binary Classification

Dichotomous Classifiers

	Label	
Id	Actual	Predicted
1	Positive	Positive
2	Positive	Negative
3	Positive	Positive
4	Negative	Positive
5	Negative	Negative
6	Negative	Negative

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

Binary Classification

Dichotomous Classifiers

	Label	
Id	Actual	Predicted
1	Positive	Positive
2	Positive	Negative
3	Positive	Positive
4	Negative	Positive
5	Negative	Negative
6	Negative	Negative

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

Binary Classification

Dichotomous Classifiers

	Label	
Id	Actual	Predicted
1	Positive	Positive
2	Positive	Negative
3	Positive	Positive
4	Negative	Positive
5	Negative	Negative
6	Negative	Negative

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

Binary Classification

Dichotomous Classifiers

	Label	
Id	Actual	Predicted
1	Positive	Positive
2	Positive	Negative
3	Positive	Positive
4	Negative	Positive
5	Negative	Negative
6	Negative	Negative

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

Confusion Matrix

	Actual	
Predicted	Positive	Negative
Positive	tp	fp
Negative	fn	tn
Total	n	m

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: $\frac{tp}{tp+fp}$

Confusion Matrix

Predicted	Actual Positive	Actual Negative
Positive	tp	fp
Negative	fn	tn
Total	n	m

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

Id	Actual Label	Score
1	Positive	1.0
2	Positive	0.8
3	Negative	0.7
4	Positive	0.6
5	Negative	0.5
6	Negative	0.3
7	Positive	0.2
8	Negative	0.1

	Actual	
Predicted	Positive	Negative
Positive	2	0
Negative	2	4
Total	4	4

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

Binary Classification

Scoring Classifier

Id	Actual Label	Score
1	Positive	1.0
2	Positive	0.8
3	Negative	0.7
4	Positive	0.6
5	Negative	0.5
6	Negative	0.3
7	Positive	0.2
8	Negative	0.1

	Actual	
Predicted	Positive	Negative
Positive	4	3
Negative	0	1
Total	4	4

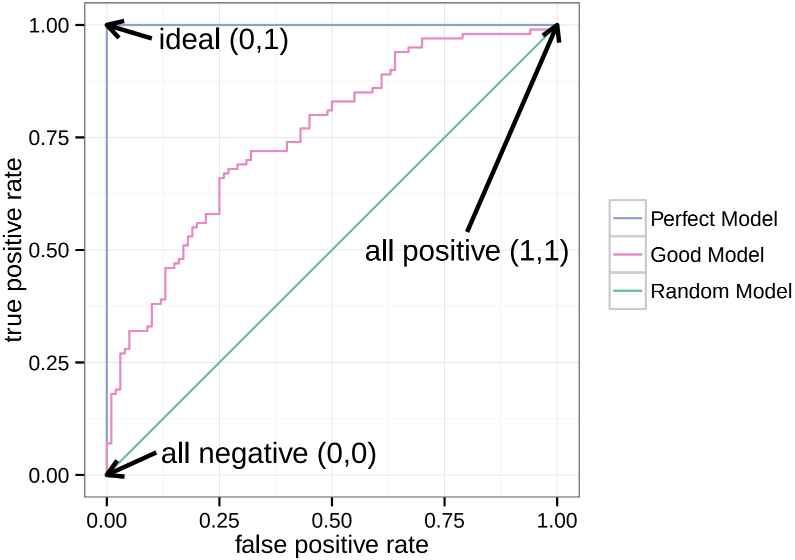
- Accuracy $= \frac{5}{8}$
- TPR $= \frac{4}{4}$
- FPR $= \frac{3}{4}$
- Precision $= \frac{4}{7}$

Thresholdless Metrics

Evaluate model without choosing a specific threshold

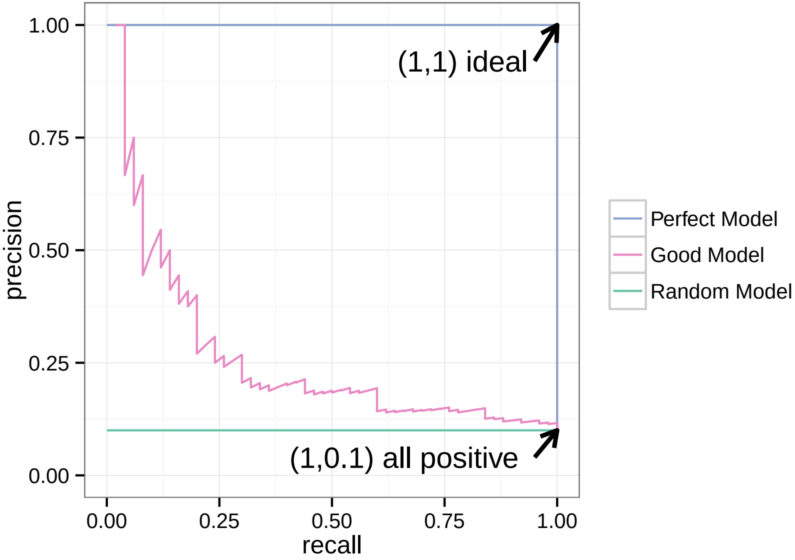
- Receiver operating characteristic (ROC) curves (Provost, Fawcett, et al., 1997)
  - Area under the ROC curve (AUCROC)
- Precision-recall (PR) curves (V Raghavan, Bollmann, and Jung, 1989)
  - Area under the PR curve (AUCPR)
- Lift curves (Piatetsky-Shapiro and Masand, 1999)
- Cost curves (Drummond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

ROC Curves



ideal (0,1)
all positive (1,1)
all negative (0,0)
true positive rate
false positive rate
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Perfect Model
Good Model
Random Model

PR Curves



1.00
0.75
0.50
0.25
0.00
precision
(1,1) ideal
(1,0.1) all positive
0.00
0.25
0.50
0.75
1.00
recall
Perfect Model
Good Model
Random Model

Outline

Empirical PR Points

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

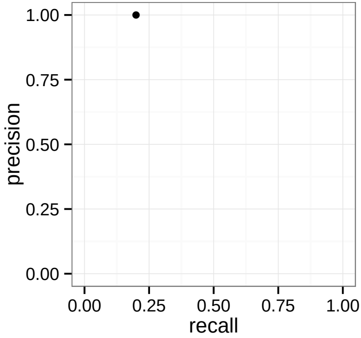
score	true label
1.00	Positive
0.95	Negative
0.90	Negative
0.85	Positive
0.80	Positive
0.75	Negative
0.70	Negative
0.65	Negative

$\ldots$

Empirical PR Points

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

score	true label
1.00	Positive
0.95	Negative
0.90	Negative
0.85	Positive
0.80	Positive
0.75	Negative
0.70	Negative
0.65	Negative
...	



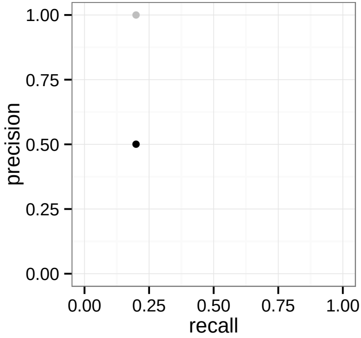
$$\text{Recall} = \frac{1}{5}$$
$$\text{Precision} = \frac{1}{1}$$

Empirical PR Points

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

score	true label
1.00	Positive
0.95	Negative
0.90	Negative
0.85	Positive
0.80	Positive
0.75	Negative
0.70	Negative
0.65	Negative

...

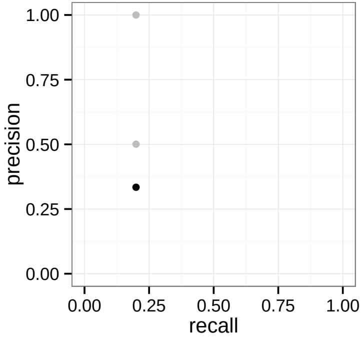


$$\text{Recall} = \frac{1}{5}$$
$$\text{Precision} = \frac{1}{2}$$

Empirical PR Points

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

score	true label
1.00	Positive
0.95	Negative
0.90	Negative
0.85	Positive
0.80	Positive
0.75	Negative
0.70	Negative
0.65	Negative
...	

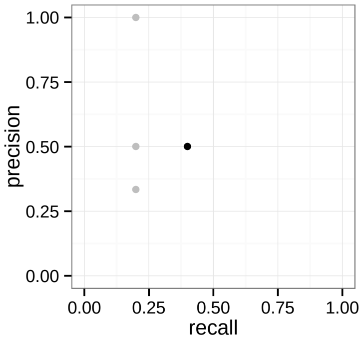


$\text{Recall} = \frac{1}{5}$
$\text{Precision} = \frac{1}{3}$

Empirical PR Points

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

score	true label
1.00	Positive
0.95	Negative
0.90	Negative
0.85	Positive
0.80	Positive
0.75	Negative
0.70	Negative
0.65	Negative
...	



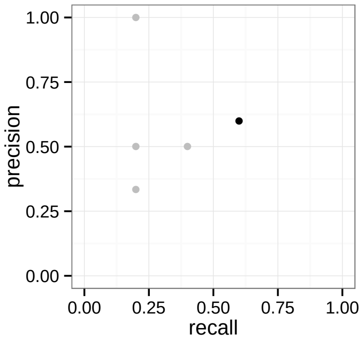
$$\text{Recall} = \frac{2}{5}$$
$$\text{Precision} = \frac{2}{4}$$

Empirical PR Points

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

score	true label
1.00	Positive
0.95	Negative
0.90	Negative
0.85	Positive
0.80	Positive
0.75	Negative
0.70	Negative
0.65	Negative

$\cdots$

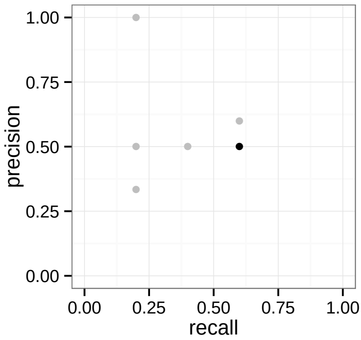


$$\text{Recall} = \frac{3}{5}$$
$$\text{Precision} = \frac{3}{5}$$

Empirical PR Points

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

score	true label
1.00	Positive
0.95	Negative
0.90	Negative
0.85	Positive
0.80	Positive
0.75	Negative
0.70	Negative
0.65	Negative
...	



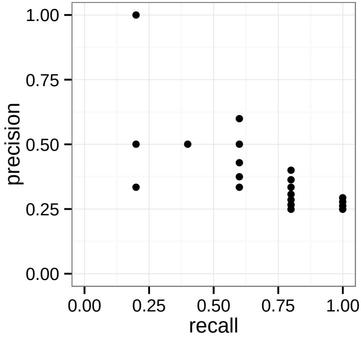
$$\text{Recall} = \frac{3}{5}$$
$$\text{Precision} = \frac{3}{6}$$

Empirical PR Points

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

score	true label
1.00	Positive
0.95	Negative
0.90	Negative
0.85	Positive
0.80	Positive
0.75	Negative
0.70	Negative
0.65	Negative

...

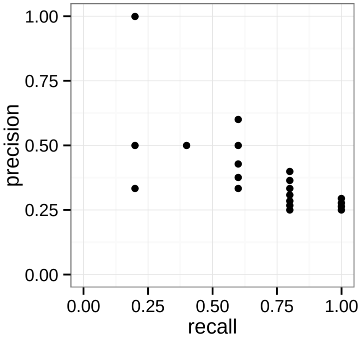


$\text{Recall} = \frac{3}{5}$

$\text{Precision} = \frac{3}{6}$

The Challenge

Given



Do

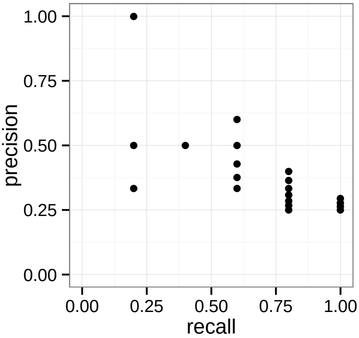
- Estimate area: AUCPR $= 0.5$
- Calculate confidence interval: $[0.4, 0.6]$

Our Goal

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

The Challenge

Given



Do

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

Our Goal

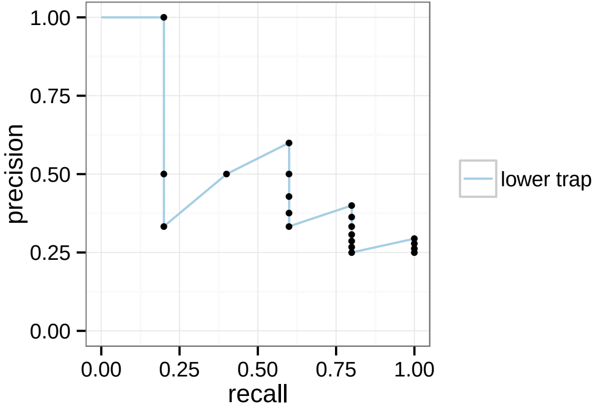
Empirically evaluate point estimates and confidence intervals of AUCPR to identify their differences and recommend best practices.

AUCPR Estimators

Many existing methods to estimate AUCPR:

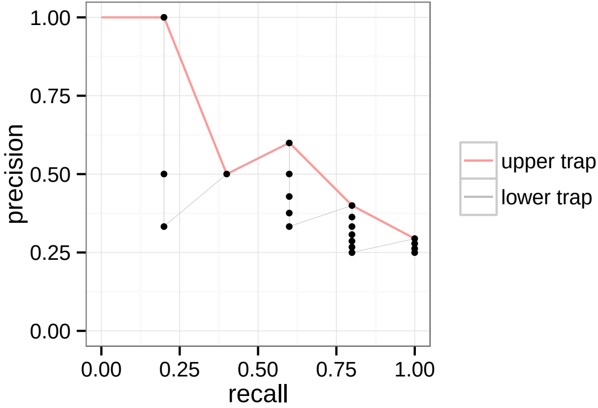
AUCPR Estimators

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



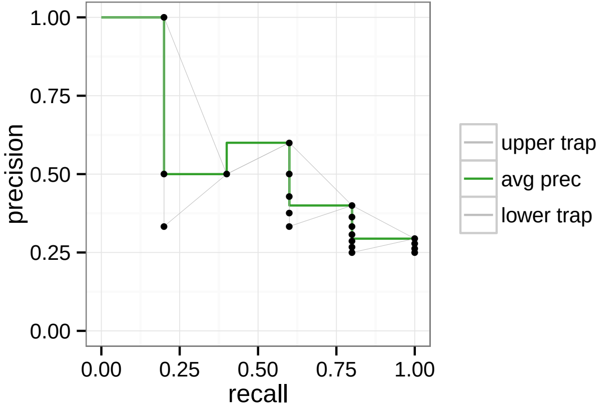
AUCPR Estimators

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



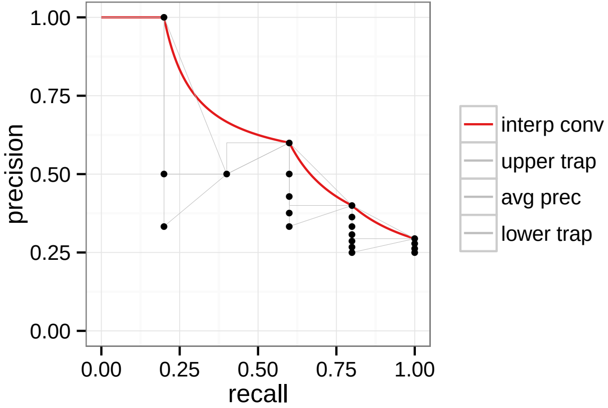
AUCPR Estimators

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



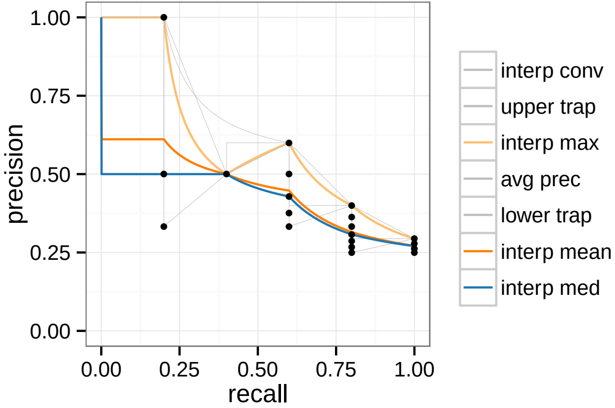
AUCPR Estimators

Many existing methods to estimate AUCPR:
(Davis and Goadrich, 2006)



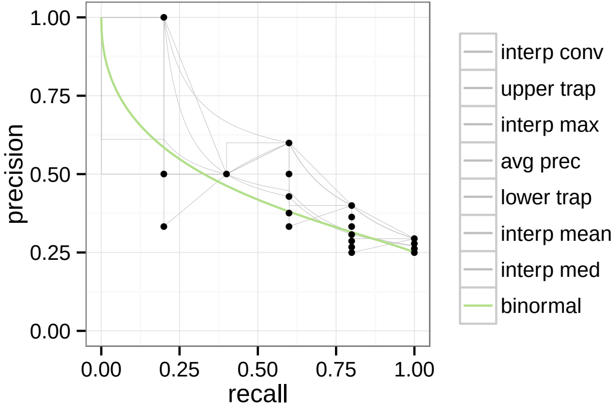
AUCPR Estimators

Many existing methods to estimate AUCPR:



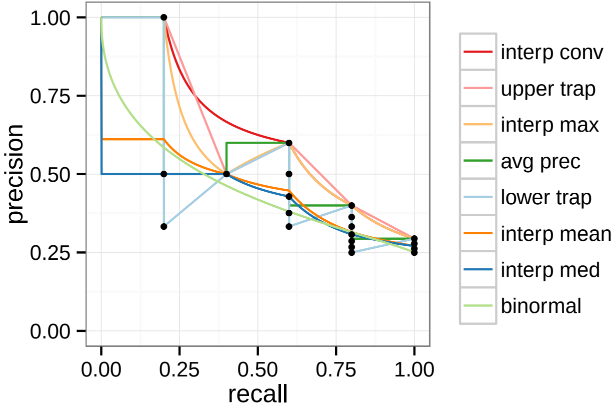
AUCPR Estimators

Many existing methods to estimate AUCPR:
(Brodersen et al., 2010)



AUCPR Estimators

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 (π) and data set sizes ($n + m$)

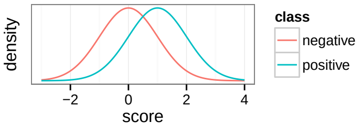
Simulation Setup

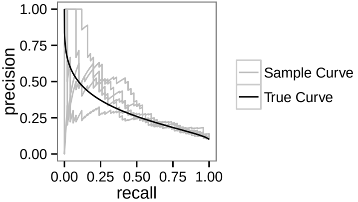
- Assume example scores are drawn from known distributions
  - Similar to binormal analysis on ROC curves (Pepe, 2004; Bamber, 1975)
  - Allows calculation of true PR curve and AUCPR
- Analyzed distributions
  - Binormal
  - Bibeta
  - Offset uniform
- Additional parameters
  - # of examples ($n + m$)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



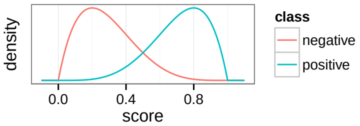


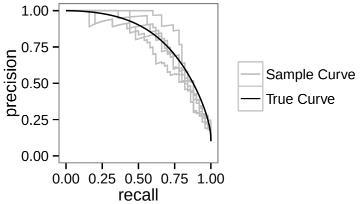
negative $\sim \mathrm{Normal}(0, 1)$
positive $\sim \mathrm{Normal}(1, 1)$

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



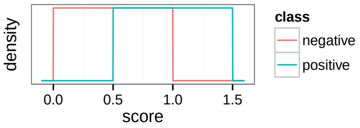


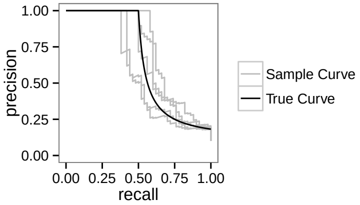
$$\text{negative} \sim \mathrm{Beta}(2, 5)$$
$$\text{positive} \sim \mathrm{Beta}(5, 2)$$

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 ($\pi = 0.1$)

Offset Uniform



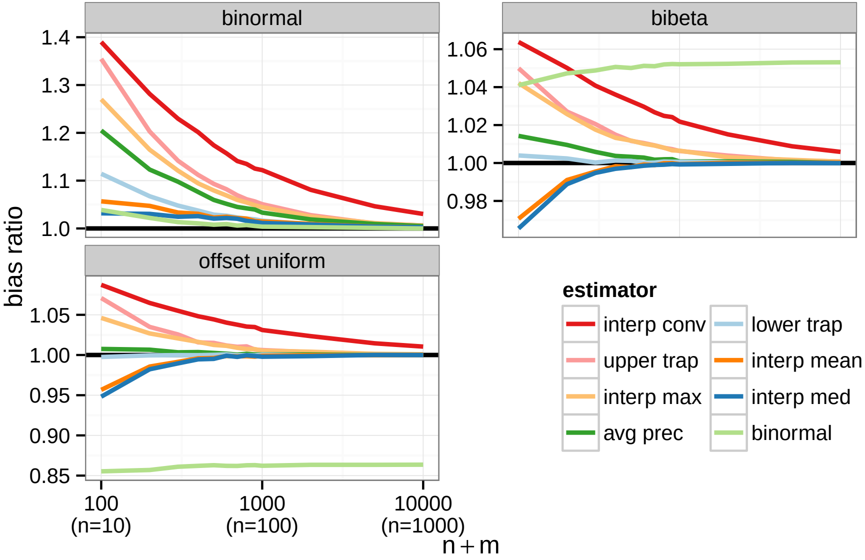


negative $\sim \mathrm{Uniform}(0, 1)$
positive $\sim \mathrm{Uniform}(0.5, 1.5)$

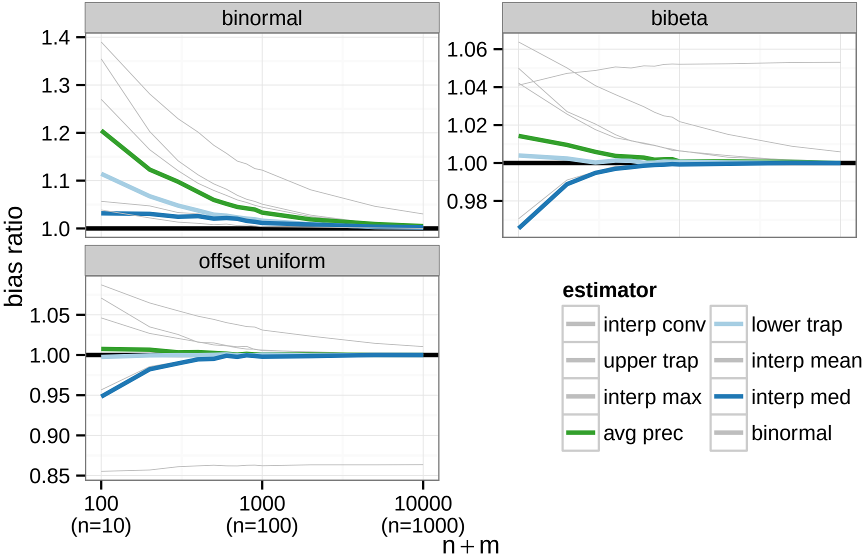
Simulation Setup

- 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



AUCPR Estimator Results



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 (π) 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 (π) 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: $\left[\frac{e^{\hat{\eta}-\Phi(1-\alpha/2)\hat{\tau}}}{1+e^{\hat{\eta}-\Phi(1-\alpha/2)\hat{\tau}}}, \frac{e^{\hat{\eta}+\Phi(1-\alpha/2)\hat{\tau}}}{1+e^{\hat{\eta}+\Phi(1-\alpha/2)\hat{\tau}}}\right]$
    $\hat{\eta} = \log\frac{\hat{\theta}}{1-\hat{\theta}}, \hat{\tau} = \left(n\hat{\theta}(1-\hat{\theta})\right)^{-1/2}$

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: $\left[\frac{e^{\hat{\eta}-\Phi(1-\alpha/2)\hat{\tau}}}{1+e^{\hat{\eta}-\Phi(1-\alpha/2)\hat{\tau}}}, \frac{e^{\hat{\eta}+\Phi(1-\alpha/2)\hat{\tau}}}{1+e^{\hat{\eta}+\Phi(1-\alpha/2)\hat{\tau}}}\right]$
    $\hat{\eta} = \log\frac{\hat{\theta}}{1-\hat{\theta}}, \hat{\tau} = \left(n\hat{\theta}(1-\hat{\theta})\right)^{-1/2}$

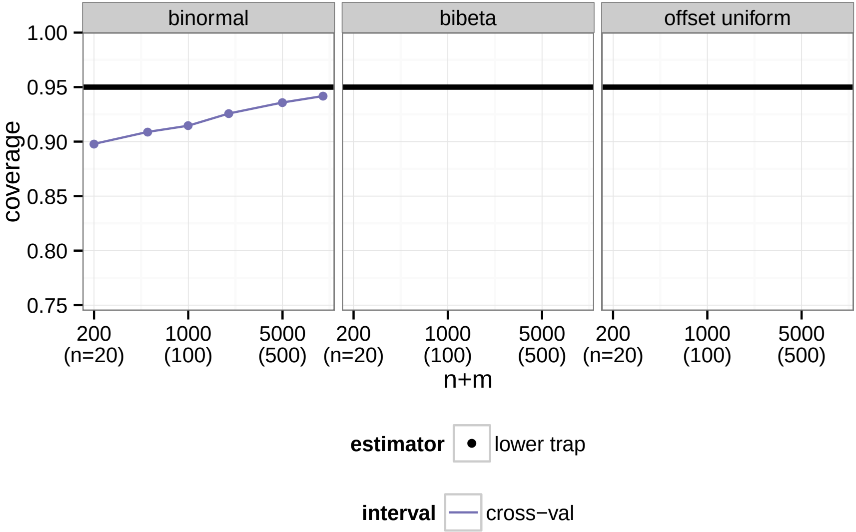
AUCPR Confidence Intervals

- 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: $\left[\frac{e^{\hat{\eta}-\Phi(1-\alpha/2)\hat{\tau}}}{1+e^{\hat{\eta}-\Phi(1-\alpha/2)\hat{\tau}}}, \frac{e^{\hat{\eta}+\Phi(1-\alpha/2)\hat{\tau}}}{1+e^{\hat{\eta}+\Phi(1-\alpha/2)\hat{\tau}}}\right]$
    $\hat{\eta} = \log\frac{\hat{\theta}}{1-\hat{\theta}}, \hat{\tau} = (n\hat{\theta}(1-\hat{\theta}))^{-1/2}$

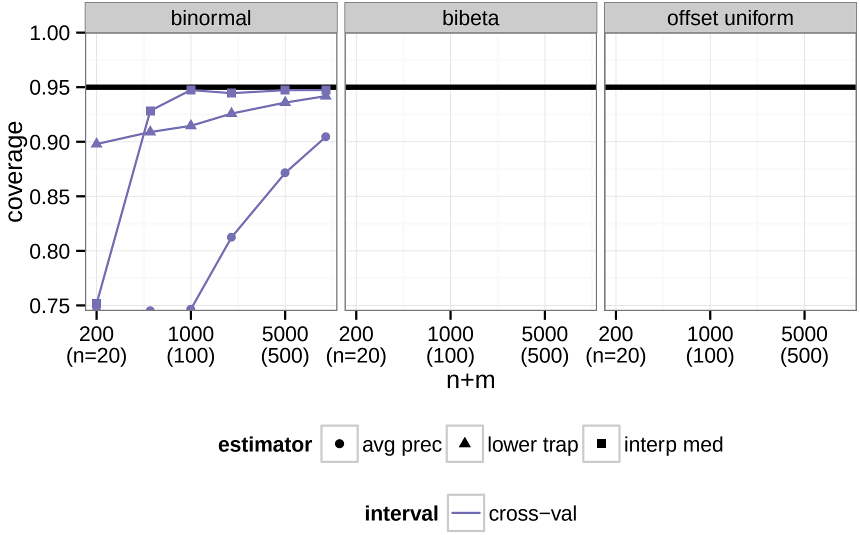
AUCPR Confidence Intervals

- 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: $\left[\frac{e^{\hat{\eta}-\Phi(1-\alpha/2)\hat{\tau}}}{1+e^{\hat{\eta}-\Phi(1-\alpha/2)\hat{\tau}}}, \frac{e^{\hat{\eta}+\Phi(1-\alpha/2)\hat{\tau}}}{1+e^{\hat{\eta}+\Phi(1-\alpha/2)\hat{\tau}}}\right]$
    $\hat{\eta} = \log\frac{\hat{\theta}}{1-\hat{\theta}}, \hat{\tau} = (n\hat{\theta}(1-\hat{\theta}))^{-1/2}$

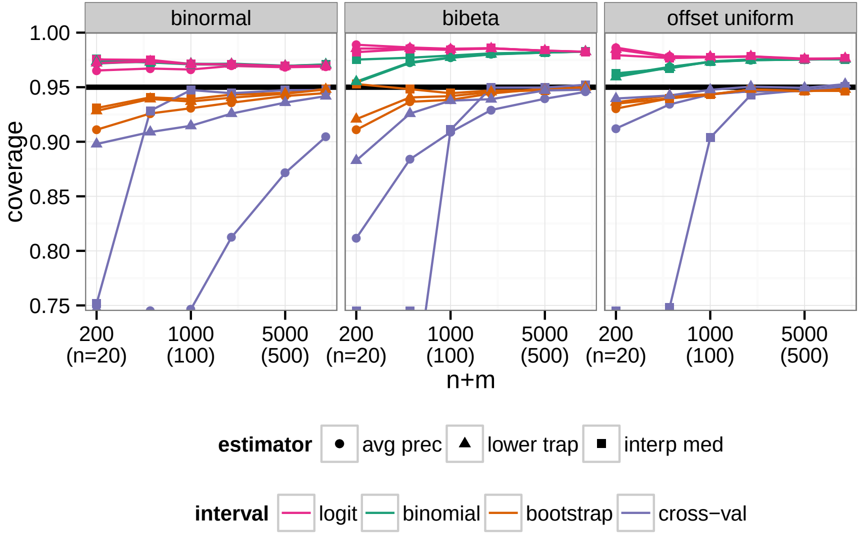
AUCPR Confidence Interval Results



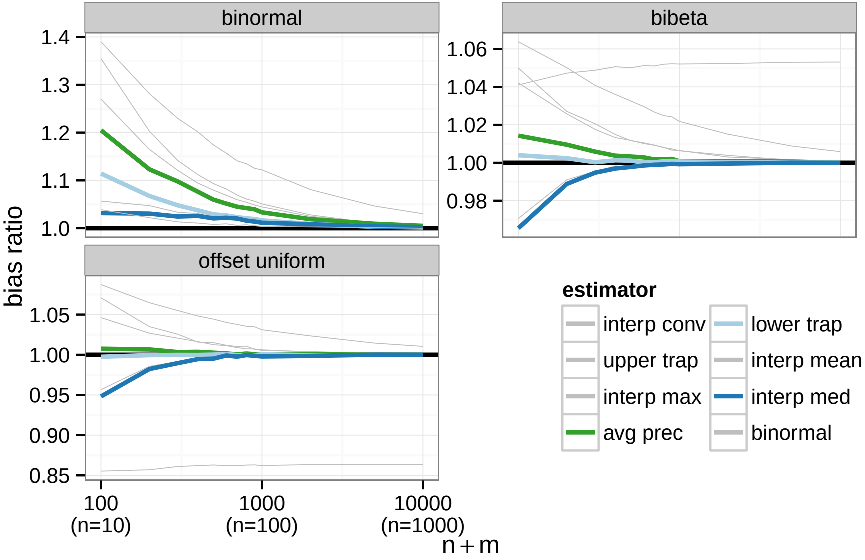
AUCPR Confidence Interval Results



AUCPR Confidence Interval Results



AUCPR Estimator Results



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

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

AUCPR Summary

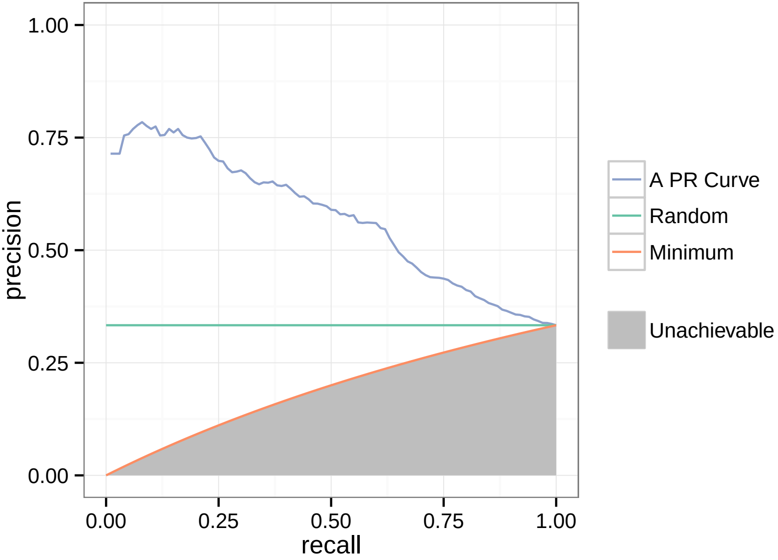
- Choice of AUCPR estimator and confidence interval is important
    - Particularly for small data sets
- Recommended estimators
    - Lower trapezoid
    - Average precision
    - Interpolated median
- Recommended confidence intervals
    - Binomial
    - Logit
    - What about cross-validation and bootstrap?
        - Converge to proper coverage, but from below
        - Problematic for small data sets and low numbers of positive examples

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

Outline

Unachievable Region in PR Space



Outline

Attacks on Evaluation Metrics

Can evaluation metrics disclose private information?

Attacks on Evaluation Metrics

Can evaluation metrics disclose private information?

Yes!

Attacks on Evaluation Metrics

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)

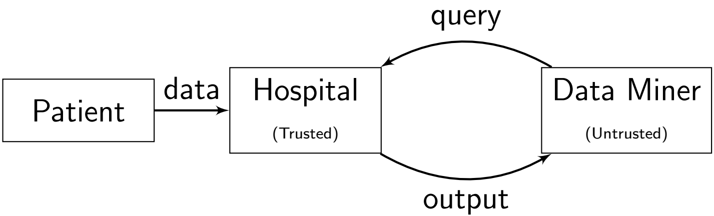
Image: www.lchcia.com

Need for Privacy

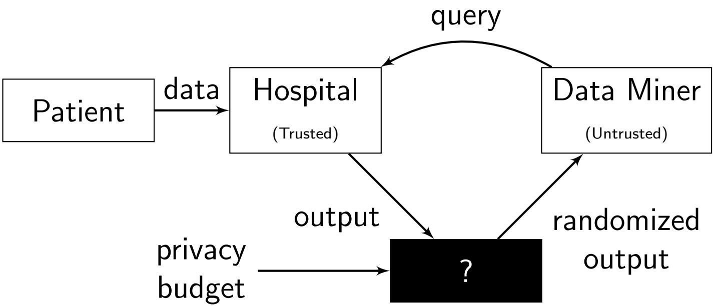
- Large databases of patient information
  - Regulations and expectations of privacy
  - Enormous potential gains from data mining
  - How to allow useful interaction with a database while preserving privacy?
- Privacy frameworks
  - k-anonymity (Sweeney, 2002)
  - Differential privacy (Dwork, 2006)

Image: www.lchcia.com

Privacy Blueprint



Privacy Blueprint



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.

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.

Differential Privacy: Formally

Definition (Dwork, 2006)

For any input database D, a randomized algorithm $f' : \mathbb{D} \to Range(f')$ is (ϵ, δ)-differentially private iff for any $\mathcal{S} \subset Range(f)$ and any database $D' \in \mathbb{D}$ where $d(D, D') = 1$,

$$\Pr(f'(D) \in \mathcal{S}) \leq e^{\epsilon} \Pr(f'(D') \in \mathcal{S}) + \delta$$

- $d(D, D')$ - number of rows that differ between D and D'
- ϵ and δ are the privacy budget
  - Smaller means more private
  - If $\delta = 0$, known as ϵ-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)
  - β-smooth sensitivity (Nissim, Raskhodnikova, and Smith, 2007)

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)
  - β-smooth sensitivity (Nissim, Raskhodnikova, and Smith, 2007)

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

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

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

guarantees ϵ-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 $f : \mathbb{D} \to \mathbb{R}$, 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 $f : \mathbb{D} \to \mathbb{R}$, the local sensitivity of f at $D \in \mathbb{D}$ is

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

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 β-smooth upper bound on the local sensitivity of f iff it satisfies:

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

- A β-smooth upper bound ensures neighboring databases will use a similar scale of noise
  - β-smooth sensitivity is the smallest such function
- Modified perturbation algorithms can use β-smooth sensitivity
  - Laplace noise provides (ϵ, δ)-differential privacy
  - Cauchy noise provides ϵ-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 β-smooth upper bound on the local sensitivity of f iff it satisfies:

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

- A β-smooth upper bound ensures neighboring databases will use a similar scale of noise
  - β-smooth sensitivity is the smallest such function
- Modified perturbation algorithms can use β-smooth sensitivity
  - Laplace noise provides (ϵ, δ)-differential privacy
  - Cauchy noise provides ϵ-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 β-smooth upper bound on the local sensitivity of f iff it satisfies:

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

- A β-smooth upper bound ensures neighboring databases will use a similar scale of noise
  - β-smooth sensitivity is the smallest such function
- Modified perturbation algorithms can use β-smooth sensitivity
  - Laplace noise provides (ϵ, δ)-differential privacy
  - Cauchy noise provides ϵ-differential privacy

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.

Privacy Applications

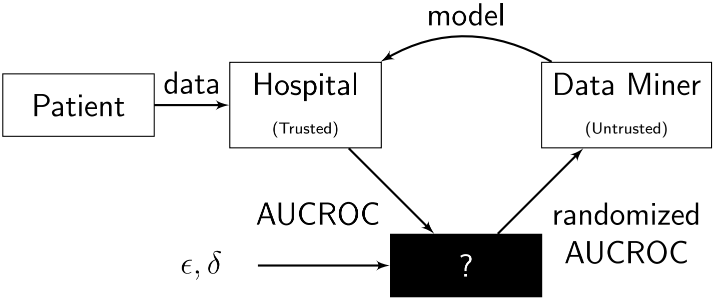
Existing applications of differential privacy

- 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)

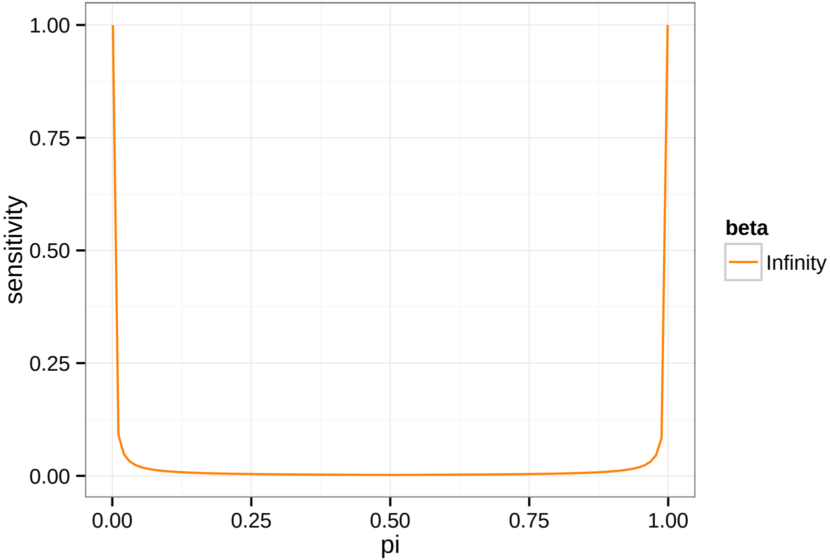
Local Sensitivity of AUCROC

Theorem

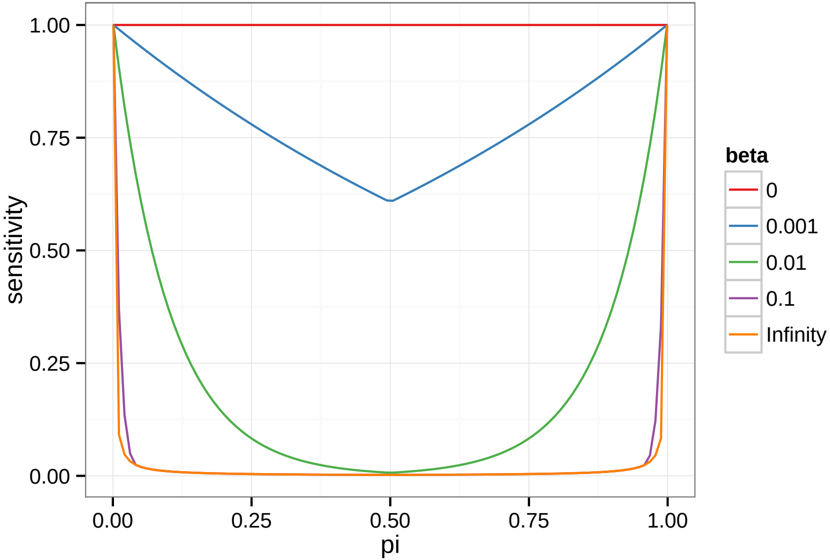
$$LS_{AUCROC}(n, m) = \begin{cases} \frac{1}{\min(n,m)} & \text{if } n > 1 \text{ and } m > 1 \\ 1 & \text{otherwise} \end{cases}$$

- n - number of positive examples in test set
- m - number of negative examples in test set

β-smooth Sensitivity of AUCROC



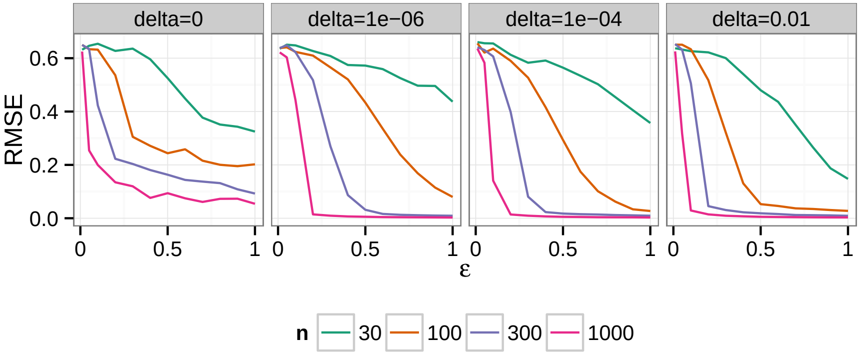
β-smooth Sensitivity of AUCROC



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



Private Evaluation Summary

- 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

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

Thesis Restatement

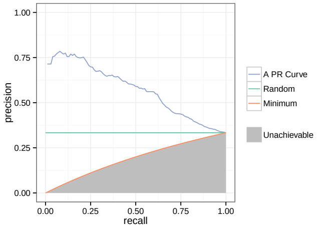
Not all methods of generating thresholdless metrics are created equal, and potential pitfalls and benefits accrue based on which methods are chosen.

Specific Contributions

Unachievable region in PR space

Recommendations

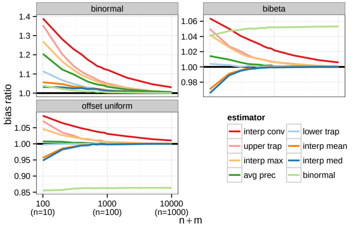
- Show unachievable region in PR curve plots
- Report skew with PR metrics (PR curve, AUCPR, F_1)
- Be aware of changing skew and aggregating from different skews

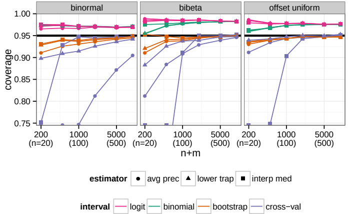


AUCPR estimators and confidence intervals

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



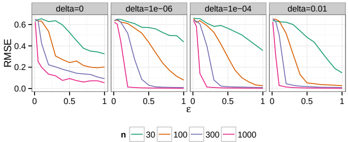


Specific Contributions

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



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

References I

Abeel, Thomas, Yves Van de Peer, and Yvan Saeys (2009). "Toward a Gold Standard for Promoter Prediction Evaluation". In: *Bioinformatics* 25.12, pp. i313–i320.

Bache, K. and M. Lichman (2013). *UCI Machine Learning Repository*. URL: `http://archive.ics.uci.edu/ml`.

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.

Barak, Boaz et al. (2007). "Privacy, accuracy, and consistency too: a holistic solution to contingency table release". In: *Proceedings of the 26th ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, pp. 273–282.

Blum, Avrim et al. (2005). "Practical privacy: the SuLQ framework". In: *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, pp. 128–138.

References II

Boyd, Kendrick, Vítor Santos Costa, et al. (July 2012). "Unachievable Region in Precision-Recall Space and Its Effect on Empirical Evaluation". In: *Proceedings of the 29th International Conference on Machine Learning*. Ed. by John Langford and Joelle Pineau. ICML '12. Edinburgh, Scotland, GB: Omnipress, pp. 639–646. ISBN: 978-1-4503-1285-1.

Boyd, Kendrick, Kevin H. Eng, and C. David Page (2013). "Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals". In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by Hendrik Blockeel et al. Vol. 8190. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 451–466.

Brodersen, Kay Henning et al. (Aug. 2010). "The Binormal Assumption on Precision-Recall Curves". In: *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, pp. 4263–4266.

References III

Davis, Jesse and Mark Goadrich (2006). "The Relationship between Precision-Recall and ROC curves". In: *Proceedings of the 23rd International Conference on Machine learning*. ICML '06. Pittsburgh, Pennsylvania: ACM, pp. 233–240. ISBN: 1-59593-383-2. DOI: 10.1145/1143844.1143874. URL: http://doi.acm.org/10.1145/1143844.1143874.

Drummond, Chris and Robert C. Holte (2006). "Cost curves: An improved method for visualizing classifier performance". English. In: *Machine Learning* 65.1, pp. 95–130. DOI: 10.1007/s10994-006-8199-5.

Dwork, Cynthia (2006). "Differential Privacy". In: *Automata, Languages and Programming*. Springer, pp. 1–12.

Dwork, Cynthia and Jing Lei (2009). "Differential privacy and robust statistics". In: *Proceedings of the 45st annual ACM symposium on Theory of computing*. ACM, pp. 371–380.

Ferri, Cèsar, José Hernández-Orallo, and Peter A Flach (2011). "Brier curves: a new cost-based visualisation of classifier performance". In: *Proceedings of the 28th International Conference on Machine Learning*. ICML '11, pp. 585–592.

References IV

Kasiviswanathan, Shiva Prasad et al. (2011). "What can we learn privately?" In: *SIAM Journal on Computing* 40.3, pp. 793–826.

Kifer, Daniel and Ashwin Machanavajjhala (2011). "No free lunch in data privacy". In: *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*. ACM, pp. 193–204.

Manning, Christopher D, Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

Matthews, Gregory J. and Ofer Harel (2013). "An Examination of Data Confidentiality and Disclosure Issues Related to Publication of Empirical ROC Curves". In: *Academic Radiology* 20.7, pp. 889–896. DOI: http://dx.doi.org/10.1016/j.acra.2013.04.011. URL: http://www.sciencedirect.com/science/article/pii/S1076633213002286.

McSherry, Frank and Kunal Talwar (2007). "Mechanism design via differential privacy". In: *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*. IEEE, pp. 94–103.

References V

Nissim, Kobbi, Sofya Raskhodnikova, and Adam Smith (June 2007). "Smooth sensitivity and sampling in private data analysis". In: *Proceedings of the 39th annual ACM symposium on Theory of computing*. STOC '07. New York, New York, USA: ACM Press, p. 75.

Pepe, Margaret Sullivan (2004). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, USA.

Piatetsky-Shapiro, Gregory and Brij Masand (1999). "Estimating campaign benefits and modeling lift". In: *Proceedings of the 5th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 185–193.

Provost, Foster J, Tom Fawcett, et al. (1997). "Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions." In: *In Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*. AAAI Press, pp. 43–48.

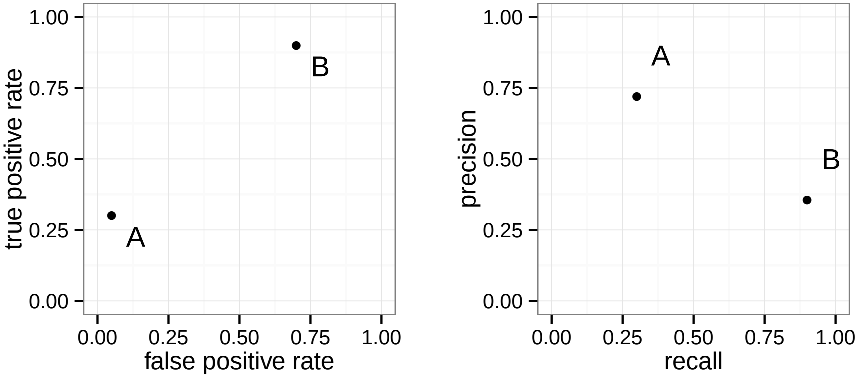
References VI

Raghavan, Vijay, Peter Bollmann, and Gwang S Jung (1989). "A critical investigation of recall and precision as measures of retrieval system performance". In: *ACM Transactions on Information Systems (TOIS)* 7.3, pp. 205–229.

Sweeney, Latanya (2002). "k-anonymity: A model for protecting privacy". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05, pp. 557–570.

Zhang, Jun et al. (2012). "Functional mechanism: regression analysis under differential privacy". In: *Proceedings of the VLDB Endowment* 5.11, pp. 1364–1375.

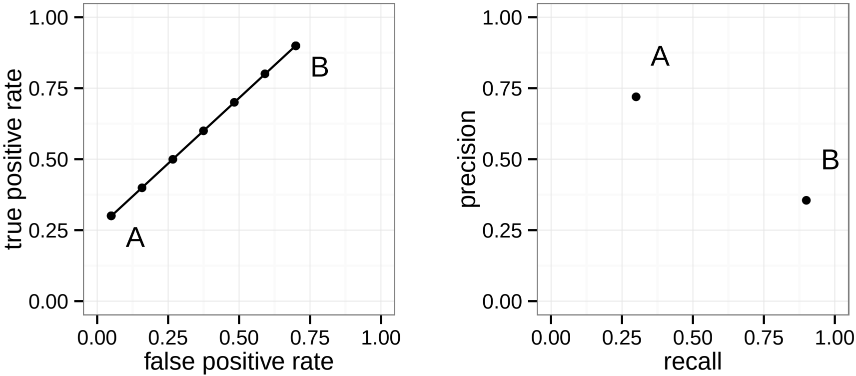
ROC/PR Space Interpolation



$$\pi = 0.3$$

(Davis and Goadrich, 2006)

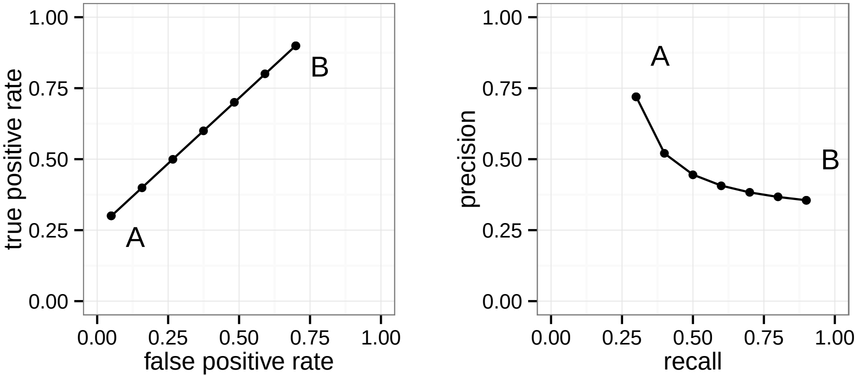
ROC/PR Space Interpolation



$$\pi = 0.3$$

(Davis and Goadrich, 2006)

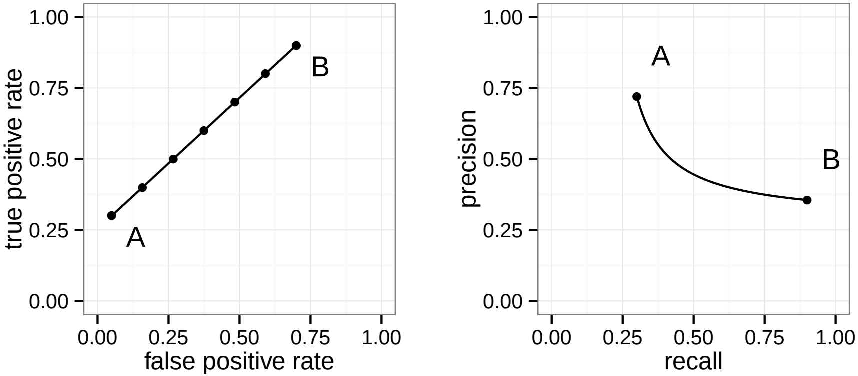
ROC/PR Space Interpolation



$$\pi = 0.3$$

(Davis and Goadrich, 2006)

ROC/PR Space Interpolation



$$\pi = 0.3$$

(Davis and Goadrich, 2006)

PR Space Interpolation Theorem

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

$$p' = \frac{r'}{ar' + b}$$

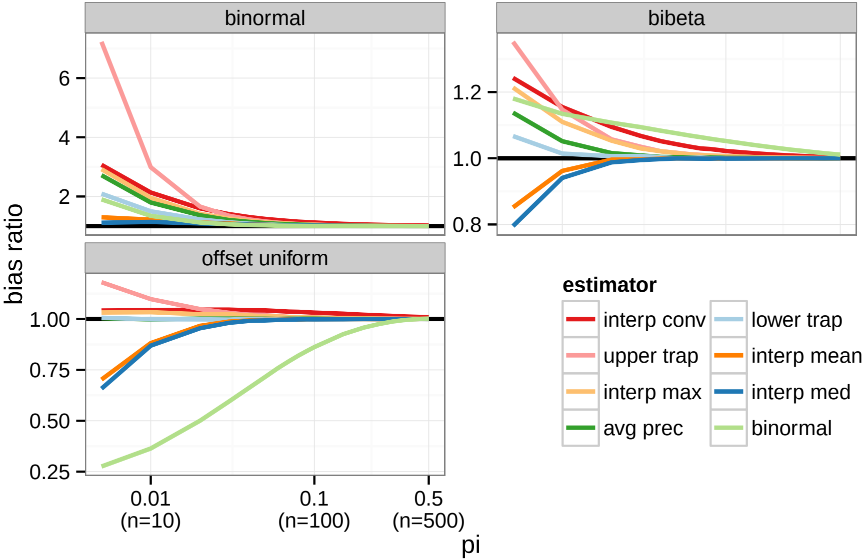
and the area under the interpolated curve between r_1 and r_2 is

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

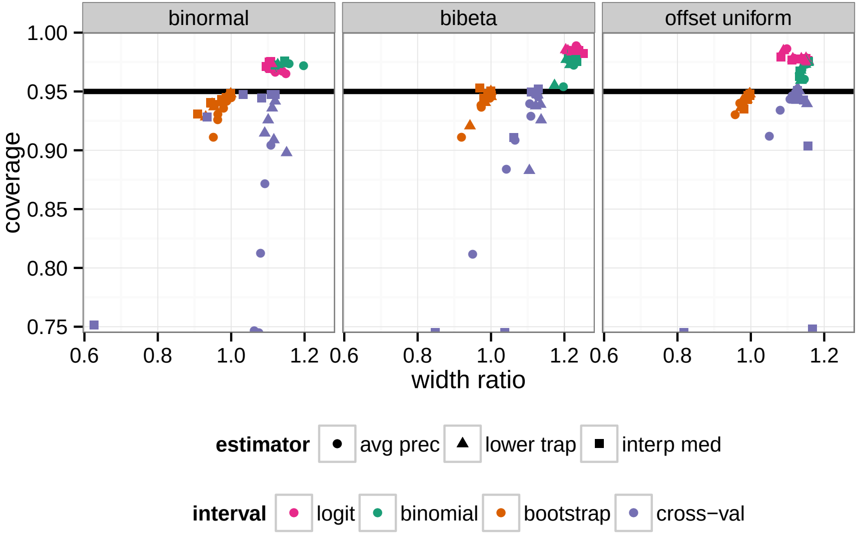
where

$$a = 1 + \frac{(1 - p_2)r_2}{p_2(r_2 - r_1)} - \frac{(1 - p_1)r_1}{p_1(r_2 - r_1)}$$
$$b = \frac{(1 - p_1)r_1}{p_1} - \frac{(1 - p_2)r_1 r_2}{p_2(r_2 - r_1)} + \frac{(1 - p_1)r_1^2}{p_1(r_2 - r_1)}$$

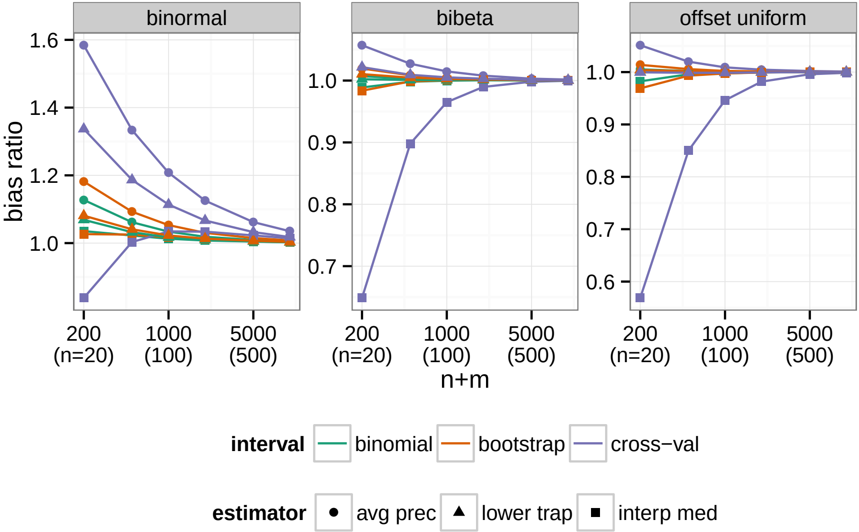
AUCPR Estimator Results by Skew



AUCPR Confidence Interval Widths



AUCPR Confidence Interval Locations



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)

β-smooth Sensitivity

Definition (Nissim, Raskhodnikova, and Smith, 2007)

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

$$S^*_{f,\beta}(D) = \max_{D' \in \mathbb{D}} LS_f(D') e^{-\beta d(D,D')}$$

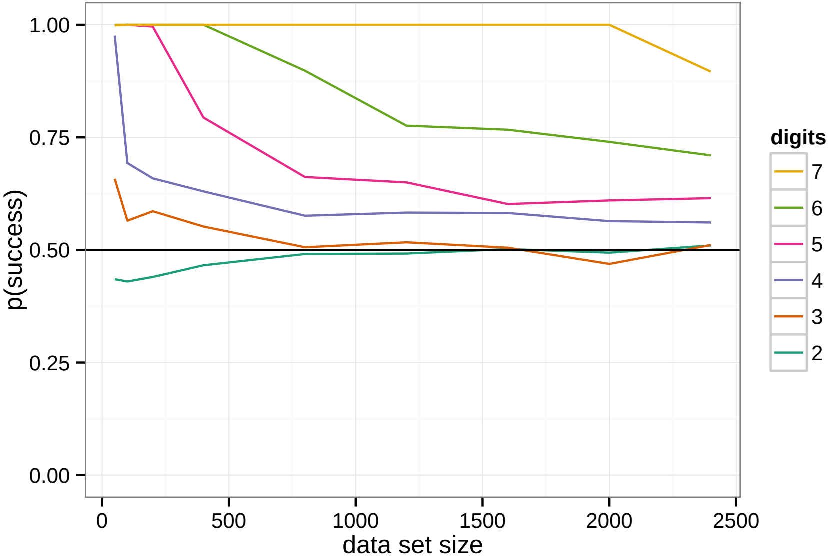
Differential Privacy using β-smooth Sensitivity

Theorem (Nissim, Raskhodnikova, and Smith, 2007)

Let $f : \mathbb{D} \to \mathbb{R}$ be any real-valued function and let $S : \mathbb{D} \to \mathbb{R}$ be the β-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 η is sampled from the distribution with density $h(z) \propto \frac{1}{1+|z|^\gamma}$, is ϵ-differentially private. Note that when $\gamma = 2$, η 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 $\eta \sim Laplace(1)$, is (ϵ, δ)-differentially private.*

AUCROC Attack Results



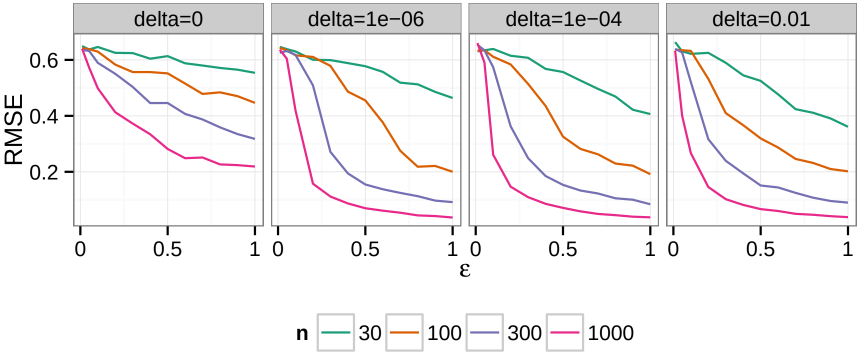
Local Sensitivity of Average Precision

Theorem

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

- n - number of positive examples in test set

AP Attack Results



Unachievable Region Theorems

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.*

Unachievable Region Theorems

Theorem (Minimum AUCPR)

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

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

Unachievable Region Theorems

Theorem (Minimum AP)

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

$$\mathrm{AP}_{\mathrm{MIN}} = \frac{1}{n}\sum_{i=1}^{n}\frac{i}{i+m}$$

Outline