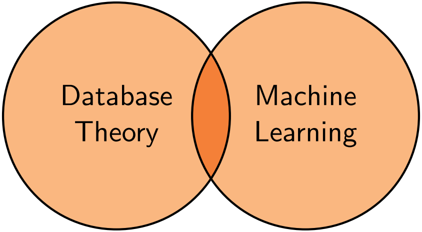
Certifiable Robustness for Nearest Neighbor Classifiers

Austen Z. Fan Paraschos Koutris

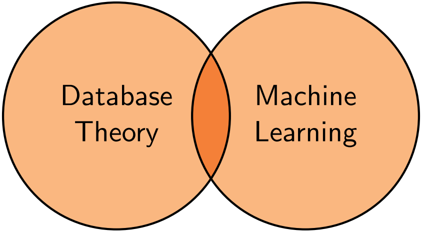
University of Wisconsin-Madison

DB Affiliates Workshop
Sep 22, 2022

Motivation

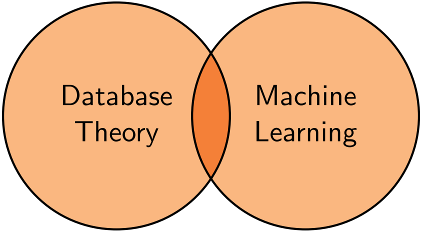


Motivation



How much should we trust a model prediction when the training data is inconsistent?

Motivation



How much should we trust a model prediction when the training data is inconsistent?

Certifiable robustness as a measure of such confidence.

Inconsistent Dataset and Repair, I

Inconsistent Dataset and Repair, I

Definition (Inconsistent dataset)

A dataset is called inconsistent if it violates its functional dependency (FD) constraints.

Inconsistent Dataset and Repair, I

Definition (Inconsistent dataset)

A dataset is called inconsistent if it violates its functional dependency (FD) constraints.

Definition (Repair)

Given an inconsistent dataset D, a repair is a subset of D such that it is consistent and maximal (not maximum).

Inconsistent Dataset and Repair, II

Inconsistent Dataset and Repair, II

FD: A $\rightarrow$ B

	A	B	C	label
t_1	1	0	a	0
t_2	1	2	b	0
t_3	2	0	a	2
t_4	2	5	c	1
t_5	3	1	a	0
t_6	4	2	d	2

Inconsistent Dataset and Repair, II

FD: A $\rightarrow$ B

	A	B	C	label
t_1	1	0	a	0
t_3	2	0	a	2
t_5	3	1	a	0
t_6	4	2	d	2

Inconsistent Dataset and Repair, II

FD: A $\rightarrow$ B

	A	B	C	label
t_1	1	0	a	0
t_4	2	5	c	1
t_5	3	1	a	0
t_6	4	2	d	2

Inconsistent Dataset and Repair, II

FD: A $\rightarrow$ B

	A	B	C	label
t_2	1	2	b	0
t_3	2	0	a	2
t_5	3	1	a	0
t_6	4	2	d	2

Inconsistent Dataset and Repair, II

FD: A $\rightarrow$ B

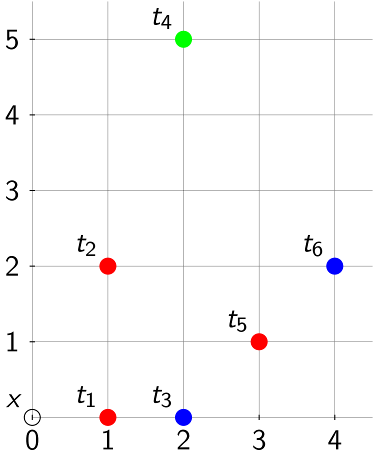
	A	B	C	label
t_2	1	2	b	0
t_4	2	5	c	1
t_5	3	1	a	0
t_6	4	2	d	2

k-NN

$k = 3$

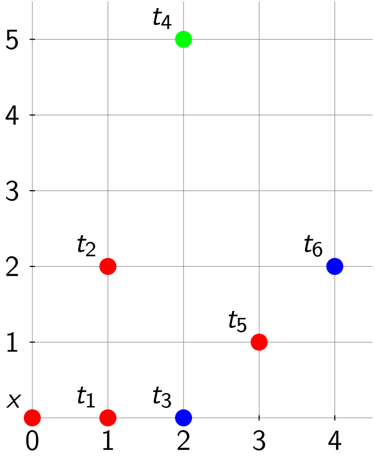
k-NN

$k = 3$



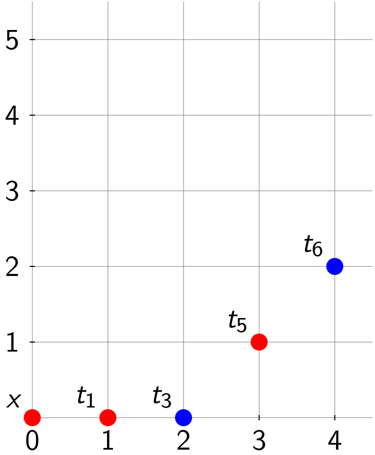
k-NN

$k = 3$



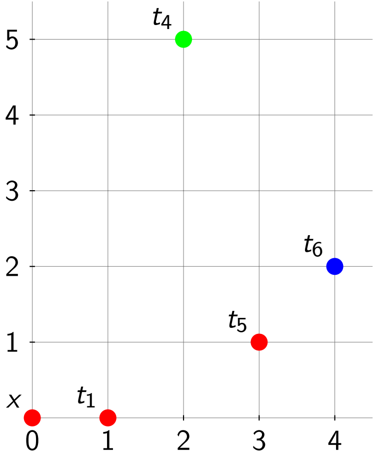
k-NN

$k = 3$



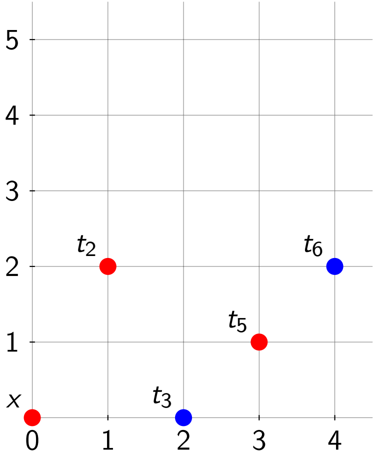
k-NN

$k = 3$



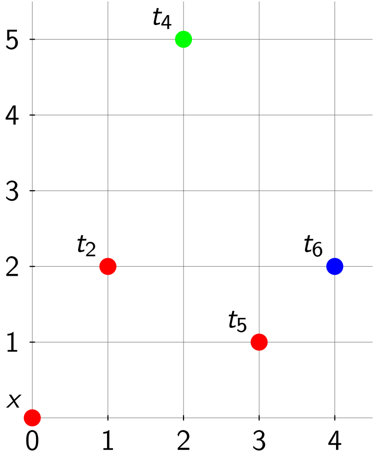
k-NN

$k = 3$



k-NN

$k = 3$



Certifiable Robustness, I

Certifiable Robustness, I

Definition (Certifiable Robustness for k-NN Classifier)

Given an inconsistent database D with labels and a test point x, then x is said to be certifiably robust for k-NN classifier if the prediction of k-NN about x on any repair of D is consistent.

Certifiable Robustness, II

Certifiable Robustness, II

A classification learning algorithm $\mathcal{L}$ with labels in $\mathcal{Y}$ takes a labeled instance I over the schema $R(A_1, \ldots, A_d)$ as training set, and returns a classifier which is a total function $\mathcal{L}_I : \mathbb{D}^d \to \mathcal{Y}$.

Certifiable Robustness, II

A classification learning algorithm $\mathcal{L}$ with labels in $\mathcal{Y}$ takes a labeled instance I over the schema $R(A_1, \ldots, A_d)$ as training set, and returns a classifier which is a total function $\mathcal{L}_I : \mathbb{D}^d \to \mathcal{Y}$.

Definition (Certifiable Robustness)

Let $\mathcal{I}$ be a labeled uncertain instance over $R(A_1, \ldots, A_d)$ and $\mathcal{L}$ be a classification learning algorithm with labels in $\mathcal{Y}$. We say that a (test) point $x \in \mathbb{D}^d$ is certifiably robust in $\mathcal{I}$ if there exists a label $\ell \in \mathcal{Y}$ such that for every repair $I \in \mathcal{I}$, $\mathcal{L}_I(x) = \ell$. The label ℓ is called a certain label for x.

Consistent Query Answering

Consistent Query Answering

Definition (Repair)

Given an inconsistent instance D with FD constraint, a repair is a subset of D such that it is consistent and maximal (not maximum).

Consistent Query Answering

Definition (Repair)

Given an inconsistent instance D with FD constraint, a repair is a subset of D such that it is consistent and maximal (not maximum).

Definition (CQA)

Given a Boolean query Q, we say that D is certain for Q, denoted as $D \models Q$, if for every repair r of D, $Q(r)$ is true.

Consistent Query Answering

Definition (Repair)

Given an inconsistent instance D with FD constraint, a repair is a subset of D such that it is consistent and maximal (not maximum).

Definition (CQA)

Given a Boolean query Q, we say that D is certain for Q, denoted as $D \models Q$, if for every repair r of D, $Q(r)$ is true.

CERTAINTY(Q): Given an inconsistent instance D and Boolean query Q, does $D \models Q$?

Consistent Query Answering

Definition (Repair)

Given an inconsistent instance D with FD constraint, a repair is a subset of D such that it is consistent and maximal (not maximum).

Definition (CQA)

Given a Boolean query Q, we say that D is certain for Q, denoted as $D \models Q$, if for every repair r of D, $Q(r)$ is true.

CERTAINTY(Q): Given an inconsistent instance D and Boolean query Q, does $D \models Q$?

There has been a long line of reserach in establishing dichotomies in CQA [Wij09, Ber11, KP12, KS14, KW18, KOW21].

Our Question; Informal

Our Question; Informal

Definition (Certifiable Robustness for k-NN Classifier)

Given an inconsistent database D with labels and a test point x, then x is said to be certifiably robust if the prediction of k-NN about x on any repair of D is consistent.

Our Question; Informal

Definition (Certifiable Robustness for k-NN Classifier)

Given an inconsistent database D with labels and a test point x, then x is said to be certifiably robust if the prediction of k-NN about x on any repair of D is consistent.

Question (Informal)

Given an inconsistent database D with labels and a test point x, is x certifiably robust?

Our Question; Formal

Our Question; Formal

Definition (CR-$\text{NN}_p\langle \mathbf{R}, k\rangle$)

Given an inconsistent labeled instance D over an FD schema $\mathbf{R}$ and a test point x, is x certifiably robust for k-NN classification w.r.t. the p-norm?

Our Question; Formal

Definition ($\text{CR-NN}_p\langle \mathbf{R}, k \rangle$)

Given an inconsistent labeled instance D over an FD schema $\mathbf{R}$ and a test point x, is x certifiably robust for k-NN classification w.r.t. the p-norm?

Definition ($\text{CR-NN}_<\langle \mathbf{R}, k \rangle$)

Given an inconsistent labeled instance D over an FD schema $\mathbf{R}$ and a strict ordering of the points in D w.r.t. their distance from a test point x, is x certifiably robust for k-NN classification?

Our Question; Formal

Definition ($\text{CR-NN}_p\langle \mathbf{R}, k\rangle$)

Given an inconsistent labeled instance D over an FD schema $\mathbf{R}$ and a test point x, is x certifiably robust for k-NN classification w.r.t. the p-norm?

Definition ($\text{CR-NN}_<\langle \mathbf{R}, k\rangle$)

Given an inconsistent labeled instance D over an FD schema $\mathbf{R}$ and a strict ordering of the points in D w.r.t. their distance from a test point x, is x certifiably robust for k-NN classification?

Definition ($\#\text{CR-NN}_<\langle \mathbf{R}\rangle$)

Given an inconsistent labeled instance D over an FD schema $\mathbf{R}$, a strict ordering of the points in D w.r.t. their distance from a test point x, and a label $\ell \in \mathcal{Y}$, output the number of repairs for which the k-NN classifier assigns label ℓ to x.

Our Contributions; I

Our Contributions; I

Definition (lhs Chain)

A set of FDs Σ has a left-hand-side chain (lhs chain for short) if for every two FDs $X_1 \to Y_1$ and $X_2 \to Y_2$ in Σ, either $X_1 \subseteq X_2$ or $X_2 \subseteq X_1$ [LKW21].

Our Contributions; I

Definition (lhs Chain)

A set of FDs Σ has a left-hand-side chain (lhs chain for short) if for every two FDs $X_1 \to Y_1$ and $X_2 \to Y_2$ in Σ, either $X_1 \subseteq X_2$ or $X_2 \subseteq X_1$ [LKW21].

Example

The FD set $\{A \to C, B \to C\}$ does not have an lhs chain, while the FD set $\{AB \to C, B \to D\}$ has an lhs chain.

Our Contributions; I

Definition (lhs Chain)

A set of FDs Σ has a left-hand-side chain (lhs chain for short) if for every two FDs $X_1 \rightarrow Y_1$ and $X_2 \rightarrow Y_2$ in Σ, either $X_1 \subseteq X_2$ or $X_2 \subseteq X_1$ [LKW21].

Example

The FD set $\{A \rightarrow C, B \rightarrow C\}$ does not have an lhs chain, while the FD set $\{AB \rightarrow C, B \rightarrow D\}$ has an lhs chain.

Theorem (Decision, F. & Koutris, 22')

Let $\mathbf{R}$ *be an FD schema. Then, the following hold:*

- *If* $\mathbf{R}$ *is equivalent to an FD schema with an lhs chain, then* $\mathrm{CR\text{-}NN}_<\langle\mathbf{R}\rangle$ *(and thus* $\mathrm{CR\text{-}NN}_p\langle\mathbf{R}\rangle$*) is in P.*
- *Otherwise, for any integer* $k \geq 1$, $\mathrm{CR\text{-}NN}_p\langle\mathbf{R}, k\rangle$ *(and thus* $\mathrm{CR\text{-}NN}_<\langle\mathbf{R}\rangle$*) is coNP-complete.*

Our Contributions; I

Definition (lhs Chain)

A set of FDs Σ has a left-hand-side chain (lhs chain for short) if for every two FDs $X_1 \rightarrow Y_1$ and $X_2 \rightarrow Y_2$ in Σ, either $X_1 \subseteq X_2$ or $X_2 \subseteq X_1$ [LKW21].

Example

The FD set $\{A \rightarrow C, B \rightarrow C\}$ does not have an lhs chain, while the FD set $\{AB \rightarrow C, B \rightarrow D\}$ has an lhs chain.

Theorem (Counting, F. & Koutris, 22')

Let **R** *be an FD schema. Then, the following hold:*

- *If* **R** *is equivalent to an FD schema with an lhs chain, then* $\#\mathrm{CR\text{-}NN}_{<}\langle \mathbf{R} \rangle$ *is in FP.*
- *Otherwise,* $\#\mathrm{CR\text{-}NN}_{<}\langle \mathbf{R}, 1 \rangle$ *is #P-complete.*

Our Contributions; II

Our Contributions; II

- Design a linear-time algorithm when the only FD constraint is a primary key [KLW$^+$20];

Our Contributions; II

- Design a linear-time algorithm when the only FD constraint is a primary key [KLW+20];
- Investigate Min-Repair: to find the repair with the smallest total weight [LKR20];

Our Contributions; II

- Design a linear-time algorithm when the only FD constraint is a primary key [KLW+20];
- Investigate Min-Repair: to find the repair with the smallest total weight [LKR20];
- Investigate CR for three widely used uncertain models:

Our Contributions; II

- Design a linear-time algorithm when the only FD constraint is a primary key [KLW+20];
- Investigate Min-Repair: to find the repair with the smallest total weight [LKR20];
- Investigate CR for three widely used uncertain models:
  - ?-sets;
  - or-sets;
  - Codd-tables.

Algorithm for decision problem $\mathrm{CR\text{-}NN}_<\langle \mathbf{R}, k\rangle$

Algorithm for decision problem $\mathrm{CR\text{-}NN}_{<}\langle \mathbf{R}, k\rangle$

Similar to the recursive algorithm in [LKR20], divided into Base Case (FD is empty), Consensus FD (there exists an FD $\emptyset \to A$), and Common Attribute (attribute A in the lhs of all FDs).

Algorithm for decision problem $\mathrm{CR\text{-}NN}_{<}\langle \mathbf{R}, k \rangle$

Similar to the recursive algorithm in [LKR20], divided into Base Case (FD is empty), Consensus FD (there exists an FD $\emptyset \to A$), and Common Attribute (attribute A in the lhs of all FDs).

Idea: say we want to know whether label 1 certifiably robust, we will then try to falsify this by finding a repair which maximizes the difference between the numbers of top-k tuples with label i and label 1, where i loops over all other labels.

Algorithm for counting problem $\#\mathrm{CR}\text{-}\mathrm{NN}_{<}\langle \mathbf{R} \rangle$

Algorithm for counting problem $\#\mathrm{CR\text{-}NN}_<\langle \mathbf{R} \rangle$

Similar to the algorithm for decision problem, while building a matrix M to keep track of the number of repairs that predicts a given label.

Algorithm for counting problem $\#\mathrm{CR\text{-}NN}_<\langle \mathbf{R} \rangle$

Similar to the algorithm for decision problem, while building a matrix M to keep track of the number of repairs that predicts a given label.

Truth: the "matrix" M is multi-dimensional with size $O(|D| \cdot k^m)$, where m is the number of labels, and to compute each entry we need $O(k^m)$ time.

Algorithm for counting problem $\#\mathrm{CR\text{-}NN}_<\langle \mathbf{R} \rangle$

Similar to the algorithm for decision problem, while building a matrix M to keep track of the number of repairs that predicts a given label.

Truth: the "matrix" M is multi-dimensional with size $O(|D| \cdot k^m)$, where m is the number of labels, and to compute each entry we need $O(k^m)$ time.

Open question: can we do better?

Hardness for decision problem CR-NN$_p\langle \mathbf{R}, k \rangle$, I

Hardness for decision problem CR-NN$_p\langle \mathbf{R}, k\rangle$, I

Step 1: Reduce from SAT-3-RESTRICTED to MAXIMAL-MATCHING of a labelled bipartite graph G.

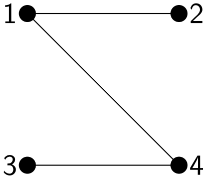
Hardness for decision problem CR-NN$_p\langle \mathbf{R}, k \rangle$, I

Step 1: Reduce from SAT-3-Restricted to Maximal-Matching of a labelled bipartite graph G.

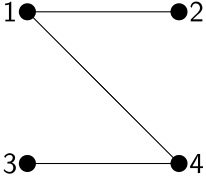
Step 2: View a maximal matching of G as a repair of a labeled instance D with FD schema $\{A \rightarrow B, B \rightarrow A\}$;

Hardness of decision problem CR-NN$_p\langle \mathbf{R}, k \rangle$, II

Hardness of decision problem CR-NN$_p\langle \mathbf{R}, k\rangle$, II



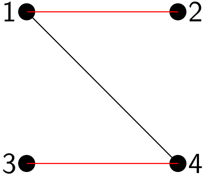
Hardness of decision problem CR-NN$_p\langle \mathbf{R}, k \rangle$, II



FD: A $\rightarrow$ B, B $\rightarrow$ A

A	B
1	2
1	4
3	4

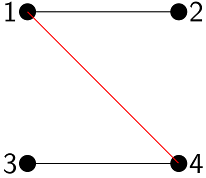
Hardness of decision problem CR-NN$_p\langle \mathbf{R}, k\rangle$, II



FD: A $\rightarrow$ B, B $\rightarrow$ A

A	B
1	2
3	4

Hardness of decision problem CR-NN$_p\langle \mathbf{R}, k\rangle$, II



FD: A $\rightarrow$ B, B $\rightarrow$ A

A	B
1	4

Hardness for decision problem, III

Step 1: Reduce from SAT-3-RESTRICTED to MAXIMAL-MATCHING of a labelled bipartite graph G.

Step 2: View a maximal matching of G as a repair of a labeled instance D with FD schema $\{A \rightarrow B, B \rightarrow A\}$;

Hardness for decision problem, III

Step 1: Reduce from SAT-3-RESTRICTED to MAXIMAL-MATCHING of a labelled bipartite graph G.

Step 2: View a maximal matching of G as a repair of a labeled instance D with FD schema $\{A \to B, B \to A\}$; use fact-wise reduction to transform D into FD schema $\mathbf{R}$.

Hardness for decision problem, III

Step 1: Reduce from SAT-3-RESTRICTED to MAXIMAL-MATCHING of a labelled bipartite graph G.

Step 2: View a maximal matching of G as a repair of a labeled instance D with FD schema $\{A \to B, B \to A\}$; use fact-wise reduction to transform D into FD schema $\mathbf{R}$.

Step 3: Encode the entries of D into numerics so that everything (p-norm values) goes through.

Hardness for counting problem

Hardness for counting problem

Theorem

If the FD schema $\mathbf{R}$ *is not equivalent to some FD schema with an lhs chain, then* $\#\mathrm{CR\text{-}NN}_{<}\langle\mathbf{R}\rangle$ *is #P-hard.*

Hardness for counting problem

Theorem

If the FD schema **R** *is not equivalent to some FD schema with an lhs chain, then* $\#\mathrm{CR\text{-}NN}_{<}\langle \mathbf{R} \rangle$ *is #P-hard.*

Proof.

Livshits, Kimelfeld, and Wijsen showed that it is #P-hard to count the number of repairs. Now, given any instance D, we can pick any ordering of the points and assign the same label ℓ to every tuple. Then, the number of repairs that predict label ℓ is the same as the total number of repairs. □

Algorithm when single primary key (Decision)

Algorithm when single primary key (Decision)

Karlas et al. gave two algorithms, SortScan (SS) and MinMax (MM), for “certain prediction” under the Codd table model.

Algorithm when single primary key (Decision)

Karlas et al. gave two algorithms, SortScan (SS) and MinMax (MM), for “certain prediction” under the Codd table model.

This is the certifiable robustness question under a further restriction that any two tuples in the same block have the same labels.

Algorithm when single primary key (Decision)

Karlas et al. gave two algorithms, SortScan (SS) and MinMax (MM), for "certain prediction" under the Codd table model.

This is the certifiable robustness question under a further restriction that any two tuples in the same block have the same labels.

	Our algorithm	Karlas et al.'s
Time Complexity	$O(\|D\| \cdot m)$	$O(\|D\| \cdot m)$ for MM $\Omega(\|D\| \cdot \binom{m+k-1}{k})$ for SS
Applicability	any k and m	MM only for $m = 2$ only under the restriction

Optimal Repairs Revisited, I

Optimal Repairs Revisited, I

OPT-REPAIR: to find the repair with the largest total weight where each tuple is associated with a positive weight [LKR20].

Optimal Repairs Revisited, I

OPT-REPAIR: to find the repair with the largest total weight where each tuple is associated with a positive weight [LKR20].

MIN-REPAIR: to find the subset repair that has tuples with the smallest total weight.

Optimal Repairs Revisited, I

OPT-REPAIR: to find the repair with the largest total weight where each tuple is associated with a positive weight [LKR20].

MIN-REPAIR: to find the subset repair that has tuples with the smallest total weight.

FORBIDDEN-REPAIR: given an inconsistent instance D and a subinstance $S \subseteq D$, does there exist a subset repair $I \subseteq D$ such that $I \cap S = \emptyset$?

Optimal Repairs Revisited, II

Optimal Repairs Revisited, II

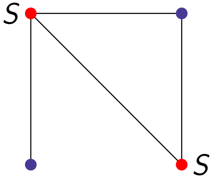
Lemma

There exists a many-one polynomial time reduction from $\textsc{Forbidden-Repair}\langle \mathbf{R} \rangle$ *to the complement of* $\mathrm{CR\text{-}NN}_{<}\langle \mathbf{R}, 1 \rangle$.

Optimal Repairs Revisited, II

Lemma

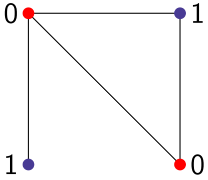
There exists a many-one polynomial time reduction from $\textsc{Forbidden-Repair}\langle \mathbf{R} \rangle$ *to the complement of* $\mathrm{CR\text{-}NN}_{<}\langle \mathbf{R}, 1 \rangle$.



Optimal Repairs Revisited, II

Lemma

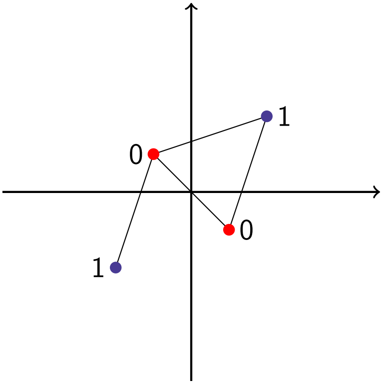
There exists a many-one polynomial time reduction from $\textsc{Forbidden-Repair}\langle \mathbf{R} \rangle$ *to the complement of* $\mathrm{CR\text{-}NN}_{<}\langle \mathbf{R}, 1 \rangle$.



Optimal Repairs Revisited, II

Lemma

There exists a many-one polynomial time reduction from $\text{Forbidden-Repair}\langle \mathbf{R} \rangle$ *to the complement of* $\text{CR-NN}_{<}\langle \mathbf{R}, 1 \rangle$.



Open Problems

Open Problems

Certifiable robustness

Open Problems

Certifiable robustness

- for other widely used classification algorithms, such as decision trees, Naive Bayes classifiers and linear classifiers?

Open Problems

Certifiable robustness

- for other widely used classification algorithms, such as decision trees, Naive Bayes classifiers and linear classifiers?
- when the instance D is not a single table but the join of several tables?

Open Problems

Certifiable robustness

- for other widely used classification algorithms, such as decision trees, Naive Bayes classifiers and linear classifiers?
- when the instance D is not a single table but the join of several tables?
- for other integrity constraints, e.g. inclusion dependencies?

Open Problems

Certifiable robustness

- for other widely used classification algorithms, such as decision trees, Naive Bayes classifiers and linear classifiers?
- when the instance D is not a single table but the join of several tables?
- for other integrity constraints, e.g. inclusion dependencies?

Open Problems

Certifiable robustness

- for other widely used classification algorithms, such as decision trees, Naive Bayes classifiers and linear classifiers?
- when the instance D is not a single table but the join of several tables?
- for other integrity constraints, e.g. inclusion dependencies?

More on pratical side

Open Problems

Certifiable robustness

- for other widely used classification algorithms, such as decision trees, Naive Bayes classifiers and linear classifiers?
- when the instance D is not a single table but the join of several tables?
- for other integrity constraints, e.g. inclusion dependencies?

More on pratical side

- deriving fast heuristic or approximation algorithm for CR-NN$\langle \mathbf{R}, k \rangle$ or $\#$ CR-NN$\langle \mathbf{R}, k \rangle$ (under some assumption on simple FD schema structure).

Open Problems

Certifiable robustness

- for other widely used classification algorithms, such as decision trees, Naive Bayes classifiers and linear classifiers?
- when the instance D is not a single table but the join of several tables?
- for other integrity constraints, e.g. inclusion dependencies?

More on pratical side

- deriving fast heuristic or approximation algorithm for CR-NN$\langle \mathbf{R}, k \rangle$ or $\#$ CR-NN$\langle \mathbf{R}, k \rangle$ (under some assumption on simple FD schema structure).
- deriving fast heuristic or approximation algorithm for (almost uniformly) sampling a repair that predicts a certain label.

Open Problems

Certifiable robustness

- for other widely used classification algorithms, such as decision trees, Naive Bayes classifiers and linear classifiers?
- when the instance D is not a single table but the join of several tables?
- for other integrity constraints, e.g. inclusion dependencies?

More on pratical side

- deriving fast heuristic or approximation algorithm for CR-NN$\langle \mathbf{R}, k \rangle$ or $\#$ CR-NN$\langle \mathbf{R}, k \rangle$ (under some assumption on simple FD schema structure).
- deriving fast heuristic or approximation algorithm for (almost uniformly) sampling a repair that predicts a certain label.
- ...

Thank You!

References I

- Leopoldo E. Bertossi, *Database repairing and consistent query answering*, Synthesis Lectures on Data Management, Morgan & Claypool Publishers, 2011.
- Bojan Karlas, Peng Li, Renzhi Wu, Nezihe Merve Gürel, Xu Chu, Wentao Wu, and Ce Zhang, *Nearest neighbor classifiers over incomplete information: From certain answers to certain predictions*, Proc. VLDB Endow. **14** (2020), no. 3, 255–267.
- Paraschos Koutris, Xiating Ouyang, and Jef Wijsen, *Consistent query answering for primary keys on path queries*, PODS, ACM, 2021, pp. 215–232.
- Phokion G. Kolaitis and Enela Pema, *A dichotomy in the complexity of consistent query answering for queries with two atoms*, Inf. Process. Lett. **112** (2012), no. 3, 77–85.

References II

Paraschos Koutris and Dan Suciu, *A dichotomy on the complexity of consistent query answering for atoms with simple keys*, ICDT, OpenProceedings.org, 2014, pp. 165–176.

Paraschos Koutris and Jef Wijsen, *Consistent query answering for primary keys and conjunctive queries with negated atoms*, PODS, ACM, 2018, pp. 209–224.

Ester Livshits, Benny Kimelfeld, and Sudeepa Roy, *Computing optimal repairs for functional dependencies*, ACM Trans. Database Syst. **45** (2020), no. 1, 4:1–4:46.

Ester Livshits, Benny Kimelfeld, and Jef Wijsen, *Counting subset repairs with functional dependencies*, J. Comput. Syst. Sci. **117** (2021), 154–164.

References III

Jef Wijsen, *Consistent query answering under primary keys: a characterization of tractable queries*, ICDT, ACM International Conference Proceeding Series, vol. 361, ACM, 2009, pp. 42–52.