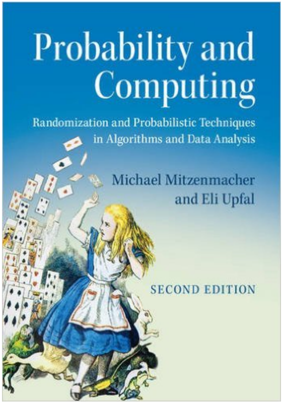
CS155/254: Probabilistic Methods in Computer Science

Chapter 14.2: Uniform Convergence - VC - Dimension



Learning a Binary Classifier

- An unknown probability distribution $\mathcal{D}$ on a domain $\mathcal{U}$
- An unknown correct classification – a partition c of $\mathcal{U}$ to *In* and *Out* sets
- Input:
  - Concept class $\mathcal{C}$ – a collection of possible classification rules (partitions of U).
  - A training set $\{(x_i, c(x_i)) \mid i = 1, \dots, m\}$, where $x_1, \dots, x_m$ are sampled from $\mathcal{D}$.
- Goal: With probability $1 - \delta$ the algorithm generates a *good* classifier.
- A classifier is *good* if the probability that it errs on an item generated from $\mathcal{D}$ is $\leq opt(\mathcal{C}) + \epsilon$, where $opt(\mathcal{C})$ is the error probability of the best classifier in $\mathcal{C}$.
- Realizable case: $c \in \mathcal{C}$, $Opt(C) = 0$.
- Unrealizable case: $c \notin \mathcal{C}$, $Opt(C) > 0$.

Probably Approximately Correct Learning (PAC Learning)

- The goal is to learn a concept (hypothesis) from a pre-defined concept class. (An interval, a rectangle, a k-CNF boolean formula, etc.)
- There is an unknown distribution D on input instances.
- Correctness of the algorithm is measured with respect to the distribution D.
- The goal: a polynomial time (and number of samples) algorithm that with probability $1 - \delta$ computes an hypothesis of the target concept that is correct (on each instance) with probability $1 - \epsilon$.

Formal Definition

- We have a unit cost function $Oracle(c, D)$ that produces a pair $(x, c(x))$, where x is distributed according to D, and $c(x)$ is the value of the concept c at x. Successive calls are independent.
- A concept class $\mathcal{C}$ over input set X is PAC learnable if there is an algorithm L with the following properties: For every concept $c \in \mathcal{C}$, every distribution D on X, and every $0 \leq \epsilon, \delta \leq 1/2$,
  - Given a function $Oracle(c, D)$, ϵ and δ, with probability $1 - \delta$ the algorithm output an hypothesis $h \in \mathcal{C}$ such that $Pr_D(h(x) \neq c(x)) \leq \epsilon$.
  - The concept class $\mathcal{C}$ is efficiently PAC learnable if the algorithm runs in time polynomial in the size of the problem,$1/\epsilon$ and $1/\delta$.

So far we showed that the concept class "intervals on the line" is efficiently PAC learnable.

The fundamental learning questions:

- What concept classes are PAC-learnable? How large training set is needed?
- What concept class are efficiently learnable (in polynomial time)?

A complete (and beautiful) characterization for the first question, not very satisfying answer for the second one.

Some Examples:

- Efficiently PAC learnable: Interval in R, rectangular in R^2, disjunction of up to n variables, 3-CNF formula,...
- PAC learnable, but not in polynomial time (unless $P = NP$): DNF formula, finite automata, ...
- Not PAC learnable: Convex body in R^2, $\{\sin(hx) \mid 0 \leq h \leq \pi\}$,...

The Weakness of Union Bound

Theorem

In the realizable case, any concept class $\mathcal{C}$ can be learned with $m = \frac{1}{\epsilon}(\ln|\mathcal{C}| + \ln\frac{1}{\delta})$ samples.

Learning an Interval:

- The true classification rule is defined by a sub-interval $[a, b] \subseteq [A, B]$. The concept class $\mathcal{C}$ is the collection of all intervals, $\mathcal{C} = \{[c, d] \mid [c, d] \subseteq [A, B]\}$

Theorem

There is a learning algorithm that given a sample from $\mathcal{D}$ of size $m = \frac{2}{\epsilon}\ln\frac{2}{\delta}$, with probability $1 - \delta$, returns a classification rule (interval) $[x, y]$ that is correct with probability $1 - \epsilon$.

This sample size bound is independent of the size of the concept class $|\mathcal{C}|$, which is infinite.

Uniform Convergence for Learning Binary Classifcation

- Given a concept class $\mathcal{C}$, and a training set sampled from $\mathcal{D}$, $\{(x_i, c(x_i)) \mid i = 1, \dots, m\}$.
- For any $h \in \mathcal{C}$, let $\Delta(c, h)$ be the set of items on which the two classifiers differ: $\Delta(c, h) = \{x \in U \mid h(x) \neq c(x)\}$
- For the realizable case we need a training set (sample) that with probability $1 - \delta$ intersects every set in

$$\{\Delta(c, h) \mid Pr(\Delta(c, h)) \geq \epsilon\} \quad (\epsilon\text{-net})$$

- For the unrealizable case we need a training set that with probability $1 - \delta$ estimates, within additive error ϵ, every set in

$$\Delta(c, h) = \{x \in U \mid h(x) \neq c(x)\} \quad (\epsilon\text{-sample}).$$

-

Uniform Convergence Sets

Given a collection R of sets in a universe X, under what conditions a finite sample N from an arbitrary distribution $\mathcal{D}$ over X, satisfies with probability $1-\delta$,

1. $$\forall r \in R, \ \Pr_{\mathcal{D}}(r) \geq \epsilon \Rightarrow \ r \cap N \neq \emptyset \qquad (\epsilon\text{-net})$$

2. for any $r \in R$,
$$\left| \Pr_{\mathcal{D}}(r) - \frac{|N \cap r|}{|N|} \right| \leq \varepsilon \qquad (\epsilon\text{-sample})$$

- Under what conditions on R can a finite sample achieve these requirements?
- What sample size is needed?

Vapnik–Chervonenkis (VC) Dimension 1968/1971

(X, R) is called a "range space":

- $X =$ finite or infinite set (the set of objects to learn)
- R is a family of subsets of X, $R \subseteq 2^X$.
  - In learning, $R = \{\Delta(c, h) \mid h \in \mathcal{C}\}$, where $\mathcal{C}$ is the concept class, and c is the correct classification.
- For a finite set $S \subseteq X$, $s = |S|$, define the projection of R on S,

$$\Pi_R(S) = \{r \cap S \mid r \in R\}.$$

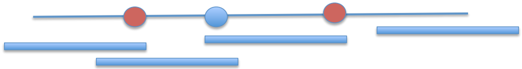
- If $|\Pi_R(S)| = 2^s$ we say that R shatters S.
- The VC-dimension of (X, R) is the maximum size of S that is shattered by R. If there is no maximum, the VC-dimension is ∞.

Theorem

A range space has a finite ϵ-net (ϵ-sample) iff its VC-dimension is finite.

The VC-Dimension of a Collection of Intervals

C = collections of intervals in [A,B] – can shatter 2 point but not 3. No interval includes only the two red points



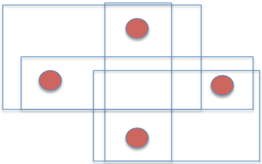
The VC-dimension of *C* is 2

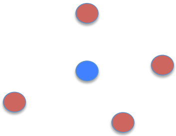
Collection of Half Spaces in the Plane

C – all half space partitions in the plane. Any 3 points can be shattered:

- Cannot partition the red from the blue points
- The VC-dimension of half spaces on the plane is 3
- The VC-dimension of half spaces in d-dimension space is d+1

Axis-parallel rectangles on the plane



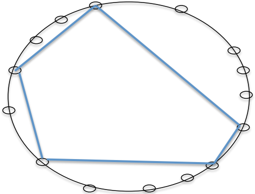


4 points that define a convex hull can be shattered.

No five points can be shattered since one of the points
must be in the convex hull of the other four.

Convex Bodies in the Plane

- *C* – all convex bodies on the plane



Any subset of the point can be included in a convex body.
The VC-dimension of *C* is ∞

A Few Examples

- $\mathcal{C} =$ set of intervals on the line. Any two points can be shattered, no three points can be shattered.
- $\mathcal{C} =$ set of linear half spaces in the plane. Any three points can be shattered but no set of 4 points. If the 4 points define a convex hull let one diagonal be 0 and the other diagonal be 1. If one point is in the convex hull of the other three, let the interior point be 1 and the remaining 3 points be 0.
- $\mathcal{C} =$ set of axis-parallel rectangles on the plane. 4 points that define a convex hull can be shattered. No five points can be shattered since one of the points must be in the convex hull of the other four.
- $\mathcal{C} =$ all convex sets in R^2. Let S be a set of n points on a boundary of a cycle. Any subset $Y \subset S$ defines a convex set that doesn't include $S \setminus Y$.

The Main Result

Theorem (A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth - 1989)

Let $\mathcal{C}$ be a concept class with VC-dimension d then

1. *$\mathcal{C}$ is PAC learnable in the realizable case with*

$$m = O\left(\frac{d}{\epsilon} \ln \frac{d}{\epsilon} + \frac{1}{\epsilon} \ln \frac{1}{\delta}\right) \qquad (\epsilon\text{-net})$$

   samples.

2. *$\mathcal{C}$ is PAC learnable in the unrealizable case with*

$$m = O\left(\frac{d}{\epsilon^2} \ln \frac{d}{\epsilon} + \frac{1}{\epsilon^2} \ln \frac{1}{\delta}\right) \qquad (\epsilon\text{-sample})$$

   samples.

The sample size is not a function of the number of concepts, or the size of the domain!

Sauer's Lemma

For a finite set $S \subseteq X$, $s = |S|$, define the projection of R on S,

$$\Pi_R(S) = \{r \cap S \mid r \in R\}.$$

Theorem

Let (X, R) be a range space with VC-dimension d, for any $S \subseteq X$, such that $|S| = n$,

$$|\Pi_R(S)| \leq \sum_{i=0}^{d} \binom{n}{i}.$$

For $n = d$, $|\Pi_R(S)| \leq 2^d$, and for $n > d \geq 2$, $|\Pi_R(S)| \leq n^d$.

The projection of R on $n > d$ elements grows polynomially in the VC-dimension and does not depend on $|R|$.

Proof

- By induction on d, and for a fixed d, by induction on n.
- True for $d = 0$ or $n = 0$, since $\Pi_R(S) = \{\emptyset\}$.
- Assume that the claim holds for $d' \leq d - 1$ and any n, and for d and all $|S'| \leq n - 1$.
- Fix $x \in S$ and let $S' = S - \{x\}$.

$$\begin{aligned}\Pi_R(S) &= \{r \cap S \mid r \in R\}\\ \Pi_R(S') &= \{r \cap S' \mid r \in R\}\\ \Pi_{R(x)}(S') &= \{r \cap S' \mid r \in R \text{ and } x \notin r \text{ and } r \cup \{x\} \in R\}\end{aligned}$$

- For $r_1 \cap S \neq r_2 \cap S$ we have $r_1 \cap S' = r_2 \cap S'$ iff $r_1 = r_2 \cup \{x\}$, or $r_2 = r_1 \cup \{x\}$. Thus,

$$|\Pi_R(S)| = |\Pi_R(S')| + |\Pi_{R(x)}(S')|$$

Fix $x \in S$ and let $S' = S - \{x\}$.

$$
\begin{aligned}
\Pi_R(S)| &= \{r \cap S \mid r \in R\}| \\
\Pi_R(S')| &= \{r \cap S' \mid r \in R\}| \\
\Pi_{R(x)}(S')| &= \{r \cap S' \mid r \in R \text{ and } x \notin r \text{ and } r \cup \{x\} \in R\}
\end{aligned}
$$

- The VC-dimension of $(S, \Pi_R(S))$ is no more than the VC-dimension of (X, R), which is d.
- The VC-dimension of the range space $(S', \Pi_R(S'))$ is no more than the VC-dimension of $(S, \Pi_R(S))$ and $|S'| = n - 1$, thus by the induction hypothesis

$$
|\Pi_R(S')| \leq \sum_{i=0}^{d} \binom{n-1}{i}.
$$

Fix $x \in S$ and let $S' = S - \{x\}$.

$$
\begin{aligned}
|\Pi_R(S)| &= |\{r \cap S \mid r \in R\}| \\
|\Pi_R(S')| &= |\{r \cap S' \mid r \in R\}| \\
|\Pi_{R(x)}(S')| &= |\{r \cap S' \mid r \in R \text{ and } x \notin r \text{ and } r \cup \{x\} \in R\}|
\end{aligned}
$$

- For each $r \in \Pi_{R(x)}(S')$ the range set $\Pi_S(R)$ has two sets: r and $r \cup \{x\}$. If B is shattered by $(S', \Pi_{R(x)}(S'))$ then $B \cup \{x\}$ is shattered by (X, R), thus $(S', \Pi_{R(x)}(S'))$ has VC-dimension bounded by $d - 1$, and

$$
|\Pi_{R(x)}(S')| \leq \sum_{i=0}^{d-1} \binom{n-1}{i}.
$$

$$|\Pi_R(S)| = |\Pi_R(S')| + |\Pi_{R(x)}(S')|$$

$$\begin{aligned} |\Pi_R(S)| &\le \sum_{i=0}^{d} \binom{n-1}{i} + \sum_{i=0}^{d-1} \binom{n-1}{i} \\ &= 1 + \sum_{i=1}^{d} \left(\binom{n-1}{i} + \binom{n-1}{i-1}\right) \\ &= \sum_{i=0}^{d} \binom{n}{i} \le \sum_{i=0}^{d} \frac{n^i}{i!} \le n^d \end{aligned}$$

[We use $\binom{n-1}{i-1} + \binom{n-1}{i} = \frac{(n-1)!}{(i-1)!(n-i-1)!}(\frac{1}{n-i} + \frac{1}{i}) = \binom{n}{i}$]

The number of distinct concepts on n elements grows polynomially in the VC-dimension!

ϵ-net

Definition

Let (X, R) be a range space, with a probability distribution D on X. A set $N \subseteq X$ is an *ϵ-net* for X with respect to D if

$$\forall r \in R, \;\; \Pr_{\mathcal{D}}(r) \geq \epsilon \Rightarrow \;\; r \cap N \neq \emptyset.$$

Theorem

Let (X, R) be a range space with VC-dimension bounded by d. With probability $1 - \delta$, a random sample of size

$$m \geq \frac{8d}{\epsilon} \ln \frac{16d}{\epsilon} + \frac{4}{\epsilon} \ln \frac{4}{\delta}$$

is an ϵ-net for (X, R).

When is a Random Sample an ϵ-net?

- Let (X, R) be a range space with VC-dimension d. Let M be m independent samples from X.
- Let $E_1 = \{\exists r \in R \mid Pr(r) \geq \epsilon \text{ and } |r \cap M| = 0\}$. We want to show that $Pr(E_1) \leq \delta$.
- Choose a second sample T of m independent samples.
- Let $E_2 = \{\exists r \in R \mid Pr(r) \geq \epsilon \text{ and } |r \cap M| = 0 \text{ and } |r \cap T| \geq \epsilon m/2\}$

Lemma

$$Pr(E_2) \leq Pr(E_1) \leq 2Pr(E_2)$$

Lemma

$$Pr(E_2) \leq Pr(E_1) \leq 2Pr(E_2)$$

$E_1 = \{\exists r \in R \mid Pr(r) \geq \epsilon \text{ and } |r \cap M| = 0\}$

$E_2 = \{\exists r \in R \mid Pr(r) \geq \epsilon \text{ and } |r \cap M| = 0 \text{ and } |r \cap T| \geq \epsilon m/2\}$

$\frac{Pr(E_2)}{Pr(E_1)} = Pr(E_2 \mid E_1) \geq Pr(|T \cap r| \geq \epsilon m/2) \geq 1/2$

[The probability that $\exists r \in R$.... is at least the probability for a given $r \in R$.]

Since $|T \cap r|$ has a Binomial distribution $B(m, \epsilon)$,
$Pr(|T \cap r| < \epsilon m/2) \leq e^{-\epsilon m/8} < 1/2$ for $m \geq 8/\epsilon$.

$E_2 = \{\exists r \in R \mid Pr(r) \geq \epsilon \text{ and } |r \cap M| = 0 \text{ and } |r \cap T| \geq \epsilon m/2\}$

$E_2' = \{\exists r \in R \mid |r \cap M| = 0 \text{ and } |r \cap T| \geq \epsilon m/2\}$

Lemma

$$Pr(E_1) \leq 2Pr(E_2) \leq 2Pr(E_2') \leq 2(2m)^d 2^{-\epsilon m/2}.$$

For a fixed $r \in R$ and $k = \epsilon m/2$, let

$E_r = \{|r \cap M| = 0 \text{ and } |r \cap T| \geq k\} = \{|M \cap r| = 0 \text{ and } |r \cap (M \cup T)| \geq k\}$

$E_r = \{|r \cap M| = 0 \text{ and } |r \cap T| \geq k\}$

$E_2' = \cup_{r \in R} E_r.$

$E_2' = \{\exists r \in R \mid |r \cap M| = 0 \text{ and } |r \cap T| \geq \epsilon m/2\}$

For a fixed $r \in R$ and $k = \epsilon m/2$ let
$E_r = \{|r \cap M| = 0 \text{ and } |r \cap T| \geq k\}$

$E_2' = \cup_{r \in R} E_r$.

Choose an arbitrary set Z of size $2m$ and divide it randomly to M and T.

$$\begin{aligned} Pr(E_r) &= Pr(|M \cap r| = 0 \mid |r \cap (M \cup T)| \geq k) Pr(|r \cap (M \cup T)| \geq k) \\ &\leq Pr(|M \cap r| = 0 \mid |r \cap (M \cup T)| \geq k) \leq \frac{\binom{2m-k}{m}}{\binom{2m}{m}} \\ &= \frac{m(m-1)....(m-k+1)}{2m(2m-1)....(2m-k+1)} \leq 2^{-\epsilon m/2} \end{aligned}$$

The Main Idea: Switching Sample Space

We start with events defined on the distributions of samples from D that can intersect any set $r \in R$.

$E_1 = \{\exists r \in R \mid Pr(r) \geq \epsilon \text{ and } |r \cap M| = 0\}$

$E_2 = \{\exists r \in R \mid Pr(r) \geq \epsilon \text{ and } |r \cap M| = 0 \text{ and } |r \cap T| \geq \epsilon m/2\}$

$E_2' = \{\exists r \in R \mid |r \cap M| = 0 \text{ and } |r \cap T| \geq \epsilon m/2\}$

$E_r = \{|r \cap M| = 0 \text{ and } |r \cap T| \geq k\} = \{|M \cap r| = 0 \text{ and } |r \cap (M \cup T)| \geq k\}$

$E_2' = \cup_{r \in R} E_r$

Choosing a sample of $2n$ elements, $Z = M \cup T$, and partition it randomly

$$\begin{aligned} Pr(E_r) &= Pr(|M \cap r| = 0 \mid |r \cap (M \cup T)| \geq k) Pr(|r \cap (M \cup T)| \geq k) \\ &\leq Pr(|M \cap r| = 0 \mid |r \cap (M \cup T)| \geq k) \end{aligned}$$

$(|M \cap r| = 0 \mid |r \cap (M \cup T)| \geq k)$ is an event in the distribution of all partitions of Z to M and T. Therefore,

$Pr(E_2') \leq \sum_{r \in \Pi_R(Z)} Pr(|M \cap r| = 0 \mid |r \cap (M \cup T)| \geq k)$

We only need to consider sets in the projection of R on Z.

Since $|\Pi_R(Z)| \leq (2m)^d$,

$$Pr(E_2') \leq (2m)^d 2^{-\epsilon m/2}.$$

$$Pr(E_1) \leq 2Pr(E_2') \leq 2(2m)^d 2^{-\epsilon m/2}.$$

Theorem

Let (X, R) be a range space with VC-dimension bounded by d. With probability $1 - \delta$, a random sample of size

$$m \geq \frac{8d}{\epsilon} \ln \frac{16d}{\epsilon} + \frac{4}{\epsilon} \ln \frac{4}{\delta}$$

is an ϵ-net for (X, R).

We need to show that $(2m)^d 2^{-\epsilon m/2} \leq \delta$. for $m \geq \frac{8d}{\epsilon} \ln \frac{16d}{\epsilon} + \frac{4}{\epsilon} \ln \frac{1}{\delta}$.

Arithmetic

We show that $(2m)^d 2^{-\epsilon m/2} \leq \delta$. for $m \geq \frac{8d}{\epsilon} \ln \frac{16d}{\epsilon} + \frac{4}{\epsilon} \ln \frac{1}{\delta}$.
Equivalently, we require

$$\epsilon m/2 \geq \ln(1/\delta) + d \ln(2m).$$

Clearly $\epsilon m/4 \geq \ln(1/\delta)$, since $m > \frac{4}{\epsilon} \ln \frac{1}{\delta}$.

We need to show that $\epsilon m/4 \geq d \ln(2m)$.

Lemma

If $y \geq x \ln x > e$, *then* $\frac{2y}{\ln y} \geq x$.

Proof.

For $y = x \ln x$ we have $\ln y = \ln x + \ln \ln x \leq 2 \ln x$. Thus

$$\frac{2y}{\ln y} \geq \frac{2x \ln x}{2 \ln x} = x.$$

Differentiating $f(y) = \frac{\ln y}{2y}$ we find that $f(y)$ is monotonically decreasing when $y \geq x \ln x \geq e$, and hence $\frac{2y}{\ln y}$ is monotonically increasing on the same interval, proving the lemma. □

Let $y = 2m \geq \frac{16d}{\epsilon} \ln \frac{16d}{\epsilon}$ and $x = \frac{16d}{\epsilon}$, we have

$$\frac{4m}{\ln(2m)} \geq \frac{16d}{\epsilon},$$

so

$$\frac{\epsilon m}{4} \geq d \ln(2m)$$

as required.

Lower Bound on Sample Size

Theorem

A random sample of a range space with VC dimension d that with probability at least $1-\delta$ is an ϵ-net must have size $\Omega(\frac{d}{\epsilon})$.

Consider a range space (X, R), with $X = \{x_1, \ldots, x_d\}$, and $R = 2^X$.

Define a probability distribution D:

$$
\begin{aligned}
Pr(x_1) &= 1 - 4\epsilon \\
Pr(x_2) &= Pr(x_3) = \cdots = Pr(x_d) = \frac{4\epsilon}{d-1}
\end{aligned}
$$

Let $X' = \{x_2, \ldots, x_d\}$.

Let $X' = \{x_2, \ldots, x_d\}$.
$Pr(x_2) = Pr(x_3) = \cdots = Pr(x_d) = \frac{4\epsilon}{d-1}$

Let S be a sample of $m = \frac{(d-1)}{16\epsilon}$ examples from the distribution D.

Let B be the event $|S \cap X'| \leq (d-1)/2$, then $Pr(B) \geq 1/2.$

With probability $\geq 1/2$, the sample does not hit a set of probability

$$\frac{d-1}{2}\frac{4\epsilon}{d-1} = 2\epsilon$$

Corollary

A range space has a finite ϵ-net iff its VC-dimension is finite.

Back to Learning

- Let X be a set of items, $\mathcal{D}$ a distribution on X, and $\mathcal{C}$ a set of concepts on X.
- $\Delta(c, c') = \{c \setminus c' \cup c' \setminus c \mid c' \in \mathcal{C}\}$
- We take m samples and choose a concept c', while the correct concept is c.
- If $Pr_D(\{x \in X \mid c'(x) \neq c(x)\}) > \epsilon$ then, $Pr(\Delta(c, c')) \geq \epsilon$, and no sample was chosen in $\Delta(c, c')$
- How many samples are needed so that with probability $1 - \delta$ all sets $\Delta(c, c')$, $c' \in \mathcal{C}$, with $Pr(\Delta(c, c')) \geq \epsilon$, are hit by the sample?

Theorem

The VC-dimension of $(X, \{\Delta(c, c') \mid c' \in \mathcal{C}\})$ *is the same as* $(X, \mathcal{C})$.

Proof.

We show that
$\{c' \cap S \mid c' \in \mathcal{C}\} \to \{((c' \setminus c) \cup (c \setminus c')) \cap S \mid c' \in \mathcal{C}\}$ is a bijection.
Assume that $c_1 \cap S \neq c_2 \cap S$, then w.o.l.g. $x \in (c_1 \setminus c_2) \cap S$.

$x \notin c$ iff $x \in ((c_1 \setminus c) \cup (c \setminus c_1)) \cap S$ and
$x \notin ((c_2 \setminus c) \cup (c \setminus c_2)) \cap S$.

$x \in c$ iff $x \notin ((c_1 \setminus c) \cup (c \setminus c_1)) \cap S$ and $x \in ((c_2 \setminus c) \cup (c \setminus c_2)) \cap S$

Thus, $c_1 \cap S \neq c_2 \cap S$ iff
$((c_1 \setminus c) \cup (c \setminus c_1)) \cap S \neq ((c_2 \setminus c) \cup (c \setminus c_2)) \cap S$. The projection on S in both range spaces has equal size. □

PAC Learning

Theorem

In the realizable case, a concept class $\mathcal{C}$ is PAC-learnable iff the VC-dimension of the range space defined by $\mathcal{C}$ is finite.

Theorem

Let $\mathcal{C}$ be a concept class that defines a range space with VC dimension d. For any $0 < \delta, \epsilon \leq 1/2$, there is an

$$m = O\left(\frac{d}{\epsilon} \ln \frac{d}{\epsilon} + \frac{1}{\epsilon} \ln \frac{1}{\delta}\right)$$

such that $\mathcal{C}$ is PAC learnable with m samples.

Unrealizable (Agnostic) Learning

- We are given a training set $\{(x_1, c(x_1)), \ldots, (x_m, c(x_m))\}$, and a concept class $\mathcal{C}$
- No hypothesis in the concept class $\mathcal{C}$ is consistent with all the training set ($c \notin \mathcal{C}$).
- Relaxed goal: Let c be the correct concept. Find $c' \in \mathcal{C}$ such that

$$\Pr_{\mathcal{D}}(c'(x) \neq c(x)) \leq \inf_{h \in \mathcal{C}} \Pr_{\mathcal{D}}(h(x) \neq c(x)) + \epsilon.$$

- An $\epsilon/2$-sample of the range space $(X, \Delta(c, c'))$ gives enough information to identify an hypothesis that is within ϵ of the best hypothesis in the concept class.

When does the sample identify the correct rule? The unrealizable (agnostic) case

- The unrealizable case - c may not be in $\mathcal{C}$.
- For any $h \in \mathcal{C}$, let $\Delta(c, h)$ be the set of items on which the two classifiers differ: $\Delta(c, h) = \{x \in U \mid h(x) \neq c(x)\}$
- For the training set $\{(x_i, c(x_i)) \mid i = 1, \ldots, m\}$, let

$$\tilde{P}r(\Delta(c, h)) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{h(x_i) \neq c(x_i)}$$

- Algorithm: choose $h^* = \arg\min_{h \in \mathcal{C}} \tilde{P}r(\Delta(c, h))$.
- If for every set $\Delta(c, h)$,

$$|Pr(\Delta(c, h)) - \tilde{P}r(\Delta(c, h))| \leq \epsilon,$$

then

$$Pr(\Delta(c, h^*)) \leq opt(\mathcal{C}) + 2\epsilon.$$

where $opt(\mathcal{C})$ is the error probability of the best classifier in $\mathcal{C}$.

If for every set $\Delta(c, h)$,

$$|Pr(\Delta(c, h)) - \tilde{P}r(\Delta(c, h))| \leq \epsilon,$$

then

$$Pr(\Delta(c, h^*)) \leq opt(\mathcal{C}) + 2\epsilon.$$

where $opt(\mathcal{C})$ is the error probability of the best classifier in $\mathcal{C}$.
Let $\bar{h}$ be the best classifier in $\mathcal{C}$. Since the algorithm chose h^*,

$$\tilde{P}r(\Delta(c, h^*)) \leq \tilde{P}r(\Delta(c, \bar{h})).$$

Thus,

$$\begin{aligned} Pr(\Delta(c, h^*)) - opt(\mathcal{C}) &\leq \tilde{P}r(\Delta(c, h^*)) - opt(\mathcal{C}) + \epsilon \\ &\leq \tilde{P}r(\Delta(c, \bar{h})) - opt(\mathcal{C}) + \epsilon \leq 2\epsilon \end{aligned}$$

ε-sample

Definition

An ε-sample for a range space (X, R), with respect to a probability distribution $\mathcal{D}$ defined on X, is a subset $N \subseteq X$ such that, for any $r \in R$,

$$\left| \Pr_{\mathcal{D}}(r) - \frac{|N \cap r|}{|N|} \right| \leq \varepsilon \ .$$

Theorem

Let $(X, \mathcal{R})$ be a range space with VC dimension d and let $\mathcal{D}$ be a probability distribution on X. For any $0 < \epsilon, \delta < 1/2$, there is an

$$m = O\left(\frac{d}{\epsilon^2} \ln \frac{d}{\epsilon} + \frac{1}{\epsilon^2} \ln \frac{1}{\delta} \right)$$

such that a random sample from $\mathcal{D}$ of size greater than or equal to m is an ϵ-sample for X with with probability at least $1 - \delta$.

An ϵ-sample for a finite R

Definition

An ε-sample for a range space (X, R), with respect to a probability distribution $\mathcal{D}$ defined on X, is a subset $N \subseteq X$ such that, for any $r \in R$,

$$\left| \Pr_{\mathcal{D}}(r) - \frac{|N \cap r|}{|N|} \right| \leq \varepsilon \ .$$

Theorem (Hoeffding's Inequality)

Let $X_1, \ldots, X_n$ *be independent random variables such that for all* $1 \leq i \leq n$, $E[X_i] = \mu$ *and* $Pr(a \leq X_i \leq b) = 1$. *Then*

$$Pr(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| \geq \epsilon) \leq 2e^{-2n\epsilon^2/(b-a)^2}$$

Sample m elements with distribution $\mathcal{D}$. For a given $r \in R$, let $X_i^r = 1$ if the i-th sample is in r, otherwise $X_i^r = 0$. $\quad Pr(X_i^r = 1) = \Pr_{\mathcal{D}}(r)$.

$$Pr\left(\left|\Pr_{\mathcal{D}}(r) - \frac{|N \cap r|}{|N|}\right| \geq \varepsilon\right) = Pr(|\frac{1}{m}\sum_{i=1}^{m} X_i^r - \Pr_{\mathcal{D}}(r)| \geq \epsilon) \leq 2e^{-2m\epsilon^2}$$

We need $|R|2e^{-2m\epsilon^2} \leq \delta$, which requires $m \geq \frac{\ln |R|}{2\epsilon^2} + \frac{\ln(2/\delta)}{2\epsilon^2}$

Proof of the ε-sample Bound:

Let N be a set of m independent samples from X according to $\mathcal{D}$.
Let

$$E_1 = \left\{ \exists r \in R \text{ s.t. } \left| \frac{|N \cap r|}{m} - \Pr(r) \right| > \varepsilon \right\}.$$

We want to show that $\Pr(E_1) \leq \delta$.

Choose another set T of m independent samples from X according to $\mathcal{D}$. Let

$$E_2 = \left\{ \exists r \in R \text{ s.t. } \left| \frac{|N \cap r|}{m} - \Pr(r) \right| > \varepsilon \ \wedge \ \left| \Pr(r) - \frac{|T \cap r|}{m} \right| \leq \varepsilon/2 \right\}$$

Lemma

$$\Pr(E_2) \leq \Pr(E_1) \leq 2 \Pr(E_2).$$

Lemma

$\Pr(E_2) \leq \Pr(E_1) \leq 2\Pr(E_2)$.

$$E_1 = \left\{\exists r \in R \text{ s.t. } \left|\frac{|N \cap r|}{m} - \Pr(r)\right| > \varepsilon\right\}$$

$$E_2 = \left\{\exists r \in R \text{ s.t. } \left|\frac{|N \cap r|}{m} - \Pr(r)\right| > \varepsilon \;\wedge\; \left|\frac{|T \cap r|}{m} - \Pr(r)\right| \leq \varepsilon/2\right\}$$

For $m \geq \frac{24}{\varepsilon^2}$,

$$\begin{aligned}\frac{\Pr(E_2)}{\Pr(E_1)} &= \frac{\Pr(E_1 \cap E_2)}{\Pr(E_1)} = \Pr(E_2|E_1) \geq \Pr(|\frac{|T \cap r|}{m} - \Pr(r)| \leq \varepsilon/2) \\ &\geq 1 - 2e^{-\varepsilon^2 m/12} \geq 1/2\end{aligned}$$

[In bounding $\Pr(E_2|E_1)$ we use the fact that the probability that $\exists r \in R$ is not smaller than the probability that the event holds for a fixed r]

Instead of bounding the probability of

$$E_2 = \left\{ \exists r \in R \text{ s.t. } \left| \frac{|N \cap r|}{m} - \Pr(r) \right| > \varepsilon \ \wedge \ \left| \frac{|T \cap r|}{m} - \Pr(r) \right| \leq \varepsilon/2 \right\}$$

we bound the probability of

$$E_2' = \{\exists r \in R \mid ||r \cap N| - |r \cap T|| \geq \frac{\epsilon}{2} m\}.$$

By the triangle inequality ($|A| + |B| \geq |A + B|$):

$$||r \cap N| - |r \cap T|| + ||r \cap T| - m \Pr_{\mathcal{D}}(r)| \geq ||r \cap N| - m \Pr_{\mathcal{D}}(r)|.$$

or

$$||r \cap N| - |r \cap T|| \geq ||r \cap N| - m \Pr_{\mathcal{D}}(r)| - ||r \cap T| - m \Pr_{\mathcal{D}}(r)| \geq \frac{\epsilon}{2} m.$$

Since N and T are random samples, we can first choose a random sample Z of $2m$ elements, and partition it randomly into two sets of size m each. The event E_2' is in the probability space of random partitions of Z.

Lemma

$$\Pr(E_1) \leq 2\Pr(E_2) \leq 2\Pr(E_2') \leq 2(2m)^d e^{-\epsilon^2 m/8}.$$

- Since N and T are random samples, we can first choose a random sample of $2m$ elements $Z = z_1, \ldots, z_{2m}$ and then partition it randomly into two sets of size m each.
- Since Z is a random sample, any partition that is independent of the actual values of the elements generates two random samples.
- We will use the following partition: for each pair of sampled items z_{2i-1} and z_{2i}, $i = 1, \ldots, m$, with probability $1/2$ (independent of other choices) we place z_{2i-1} in T and z_{2i} in N, otherwise we place z_{2i-1} in N and z_{2i} in T.

For $r \in R$, let E_r be the event

$$E_r = \left\{ \left| |r \cap N| - |r \cap T| \right| \geq \frac{\varepsilon}{2} m \right\}.$$

We have $E_2' = \{\exists r \in R \mid ||r \cap N| - |r \cap T|| \geq \frac{\epsilon}{2} m\} = \bigcup_{r \in R} E_r$.

- If $z_{2i-1}, z_{2i} \in r$ or $z_{2i-1}, z_{2i} \notin r$ they don't contribute to the value of $||r \cap N| - |r \cap T||$.
- If just one of the pair z_{2i-1} and z_{2i} is in r then their contribution is $+1$ or -1 with equal probabilities.
- There are no more than m pairs that contribute $+1$ or -1 with equal probabilities. Applying the Chernoff bound we have

$$Pr(E_r) \leq e^{-(\epsilon m/2)^2/2m} \leq e^{-\epsilon^2 m/8}.$$

- Since the projection of X on $T \cup N$ has no more than $(2m)^d$ distinct sets we have the bound.

To complete the proof we show that for

$$m \geq \frac{32d}{\epsilon^2} \ln \frac{64d}{\epsilon^2} + \frac{16}{\epsilon^2} \ln \frac{1}{\delta}$$

we have

$$(2m)^d e^{-\epsilon^2 m/8} \leq \delta.$$

Equivalently, we require

$$\epsilon^2 m/8 \geq \ln(1/\delta) + d \ln(2m).$$

Clearly $\epsilon^2 m/16 \geq \ln(1/\delta)$, since $m > \frac{16}{\epsilon^2} \ln \frac{1}{\delta}$.

To show that $\epsilon^2 m/16 \geq d \ln(2m)$ we use:

Lemma

If $y \geq x \ln x > e$, *then* $\frac{2y}{\ln y} \geq x$.

Proof.

For $y = x \ln x$ we have $\ln y = \ln x + \ln \ln x \leq 2 \ln x$. Thus

$$\frac{2y}{\ln y} \geq \frac{2x \ln x}{2 \ln x} = x.$$

Differentiating $f(y) = \frac{\ln y}{2y}$ we find that $f(y)$ is monotonically decreasing when $y \geq x \ln x \geq e$, and hence $\frac{2y}{\ln y}$ is monotonically increasing on the same interval, proving the lemma. □

Let $y = 2m \geq \frac{64d}{\epsilon^2} \ln \frac{64d}{\epsilon^2}$ and $x = \frac{64d}{\epsilon^2}$, we have $\frac{4m}{\ln(2m)} \geq \frac{64d}{\epsilon^2}$, so $\frac{\epsilon^2 m}{16} \geq d \ln(2m)$ as required.

Application: Unrealizable (Agnostic) Learning

- We are given a training set $\{(x_1, c(x_1)), \ldots, (x_m, c(x_m))\}$, and a concept class $\mathcal{C}$
- No hypothesis in the concept class $\mathcal{C}$ is consistent with all the training set ($c \notin \mathcal{C}$).
- Relaxed goal: Let c be the correct concept. Find $c' \in \mathcal{C}$ such that

$$\Pr_{\mathcal{D}}(c'(x) \neq c(x)) \leq \inf_{h \in \mathcal{C}} \Pr_{\mathcal{D}}(h(x) \neq c(x)) + \epsilon.$$

- An $\epsilon/2$-sample of the range space $(X, \Delta(c, c'))$ gives enough information to identify an hypothesis that is within ϵ of the best hypothesis in the concept class.

ε-sample

Definition

An ε-sample for a range space (X, R), with respect to a probability distribution $\mathcal{D}$ defined on X, is a subset $N \subseteq X$ such that, for any $r \in R$,

$$\left| \Pr_{\mathcal{D}}(r) - \frac{|N \cap r|}{|N|} \right| \leq \varepsilon \ .$$

Theorem

Let $(X, \mathcal{R})$ be a range space with VC dimension d and let $\mathcal{D}$ be a probability distribution on X. For any $0 < \epsilon, \delta < 1/2$, there is an

$$m = O\left(\frac{d}{\epsilon^2} \ln \frac{d}{\epsilon} + \frac{1}{\epsilon^2} \ln \frac{1}{\delta} \right)$$

such that a random sample from $\mathcal{D}$ of size greater than or equal to m is an ϵ-sample for X with with probability at least $1 - \delta$.

Uniform Convergence [Vapnik – Chervonenkis 1971]

Definition

A set of functions $\mathcal{F}$ has the *uniform convergence* property with respect to a domain Z if there is a function $m_{\mathcal{F}}(\epsilon, \delta)$ such that

- for any $\epsilon, \delta > 0$, $m(\epsilon, \delta) < \infty$
- for any distribution D on Z, and a sample $z_1, \ldots, z_m$ of size $m = m_{\mathcal{F}}(\epsilon, \delta)$,

$$Pr(\sup_{f \in \mathcal{F}} |\frac{1}{m} \sum_{i=1}^{m} f(z_i) - E_{\mathcal{D}}[f]| \leq \epsilon) \geq 1 - \delta.$$

Let $f_E(z) = \mathbf{1}_{z \in E}$ then $\mathbf{E}[f_E(z)] = Pr(E)$.

Application: Frequent Itemsets Mining (FIM)?

Frequent Itemsets Mining: classic data mining problem with many applications

Settings:

Dataset $\mathcal{D}$
bread, milk
bread
milk, eggs
bread, milk, apple
bread, milk, eggs

Each line is a transaction, made of items from an alphabet $\mathcal{I}$

An itemset is a subset of $\mathcal{I}$. E.g., the itemset $\{\text{bread,milk}\}$

The frequency $f_{\mathcal{D}}(A)$ of $A \subseteq \mathcal{I}$ in $\mathcal{D}$ is the fraction of transactions of $\mathcal{D}$ that A is a subset of. E.g., $f_{\mathcal{D}}(\{\text{bread,milk}\}) = 3/5 = 0.6$

Problem: Frequent Itemsets Mining (FIM)

Given $\theta \in [0, 1]$ find (i.e., mine) all itemsets $A \subseteq \mathcal{I}$ with $f_{\mathcal{D}}(A) \geq \theta$

I.e., compute the set $\mathsf{FI}(\mathcal{D}, \theta) = \{A \subseteq \mathcal{I} \ : \ f_{\mathcal{D}}(A) \geq \theta\}$

There exist exact algorithms for FI mining (Apriori, FP-Growth, ...)

How to make FI mining faster?

Exact algorithms for FI mining do not scale with $|\mathcal{D}|$ (no. of transactions):

They scan $\mathcal{D}$ multiple times: painfully slow when accessing disk or network

How to get faster? We could develop faster exact algorithms (difficult) or...

... only mine random samples of $\mathcal{D}$ that fit in main memory

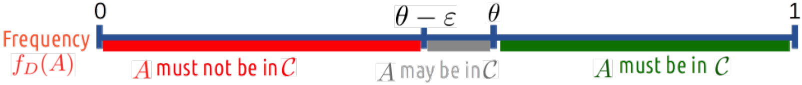
Trading off accuracy for speed: we get an approximation of $\mathsf{FI}(\mathcal{D}, \theta)$ but we get it fast

Approximation is OK: FI mining is an exploratory task (the choice of θ is also often quite arbitrary)

Key question: How much to sample to get an approximation of given quality?

How to define an approximation of the FIs?

For $\varepsilon, \delta \in (0,1)$, a (ε, δ)-approximation to $\mathsf{FI}(\mathcal{D}, \theta)$ is a collection $\mathcal{C}$ of itemsets s.t., with prob. $\geq 1 - \delta$:



"Close" False Positives are allowed, but no False Negatives
This is the price to pay to get faster results: we lose accuracy

Still, $\mathcal{C}$ can act as set of candidate FIs to prune with fast scan of $\mathcal{D}$

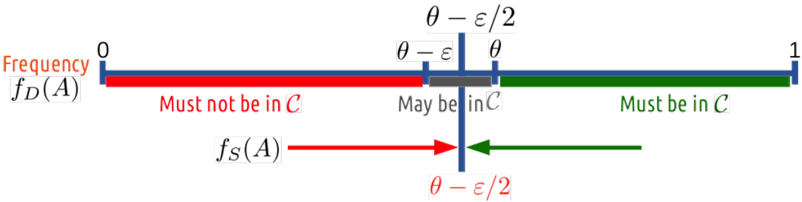
What do we really need?

We need a procedure that, given ε, δ, and $\mathcal{D}$, tells us how large should a sample $\mathcal{S}$ of $\mathcal{D}$ be so that

$$\Pr(\exists \text{ itemset } A \ : \ |f_{\mathcal{S}}(A) - f_{\mathcal{D}}(A)| > \varepsilon/2) < \delta$$

Theorem: When the above inequality holds, then $\mathsf{FI}(\mathcal{S}, \theta - \varepsilon/2)$ is an (ε, δ)-approximation

Proof (by picture):



What can we get with a Union Bound?

For any itemset A, the number of transactions that include A is distributed

$$|\mathcal{S}|f_{\mathcal{S}}(A) \sim Binomial(|\mathcal{S}|, f_{\mathcal{D}}(A))$$

Applying Chernoff bound

$$\Pr(|f_{\mathcal{S}}(A) - f_{\mathcal{D}}(A)| > \varepsilon/2) \leq 2e^{-|\mathcal{S}|\varepsilon^2/12}$$

We then apply the union bound over all the itemsets to obtain uniform convergence

There are $2^{|\mathcal{I}|}$ itemsets, a priori. We need

$$2e^{-|\mathcal{S}|\varepsilon^2/12} \leq \delta/2^{|\mathcal{I}|}$$

Thus

$$|\mathcal{S}| \geq \frac{12}{\varepsilon^2}\left(|\mathcal{I}| + \ln 2 + \ln \frac{1}{\delta}\right)$$

Assume that we have a bound ℓ on the maximum transaction size.

There are $\sum_{i \leq \ell} \binom{|\mathcal{I}|}{i} \leq |\mathcal{I}|^\ell$ possible itemsets. We need

$$2e^{-|\mathcal{S}|\varepsilon^2/12} \leq \delta / |\mathcal{I}|^\ell$$

Thus,

$$|\mathcal{S}| \geq \frac{12}{\varepsilon^2}\left(\ell \log |\mathcal{I}| + \ln 2 + \ln \frac{1}{\delta}\right)$$

The sample size depends on $\log |\mathcal{I}|$ which can still be very large.
E.g., all the products sold by Amazon, all the pages on the Web, ...

Can have a smaller sample size that depends on some characteristic quantity of $\mathcal{D}$

How do we get a smaller sample size?

[R. and U. 2014, 2015]: Let's use VC-dimension!

We define the task as an expectation estimation task:

- The domain is the dataset $\mathcal{D}$ (set of transactions)
- The family is $\mathcal{F} = \{\mathcal{T}_A, A \subseteq 2^{\mathcal{I}}\}$, where $\mathcal{T}_A = \{\tau \in \mathcal{D} \ : \ A \subseteq \tau\}$ is the set of the transactions of $\mathcal{D}$ that contain A
- The distribution π is uniform over $\mathcal{D}$: $\pi(\tau) = 1/|\mathcal{D}|$, for each $\tau \in \mathcal{D}$

We sample transactions according to the uniform distribution, hence we have:

$$\mathbb{E}_\pi[\mathbb{1}_{\mathcal{T}_A}] = \sum_{\tau \in \mathcal{D}} \mathbb{1}_{\mathcal{T}_A}(\tau)\pi(\tau) = \sum_{\tau \in \mathcal{D}} \mathbb{1}_{\mathcal{T}_A}(\tau)\frac{1}{|\mathcal{D}|} = f_{\mathcal{D}}(A)$$

We then only need an efficient-to-compute upper bound to the VC-dimension

Bounding the VC-dimesion

Theorem: The VC-dimension is less or the maximum transaction size ℓ.

Proof:

- Let $t > \ell$ and assume it is possible to shatter a set $T \subseteq \mathcal{D}$ with $|T| = t$.
- Then any $\tau \in T$ appears in at least 2^{t-1} ranges $\mathcal{T}_A$ (there are 2^{t-1} subsets of T containing τ)
- Any τ only appears in the ranges $\mathcal{T}_A$ such that $A \subseteq \tau$. So it appears in $2^{\ell} - 1$ ranges
- But $2^{\ell} - 1 < 2^{t-1}$ so τ^* can not appear in 2^{t-1} ranges
- Then T can not be shattered. We reach a contradiction and the thesis is true

By the VC ε-sample theorem we need $|S| \geq O(\frac{1}{\varepsilon^2}\left(\ell \log \ell + \ln \frac{1}{\delta}\right))$

Better bound for the VC-dimension

Enters the d-index of a dataset $\mathcal{D}$!

The d-index d of a dataset $\mathcal{D}$ is the maximum integer such that $\mathcal{D}$ contains at least d different transactions of length at least d

Example: The following dataset has d-index 3

bread	beer	milk	coffee
chips	coke	pasta	
bread	coke	chips	
milk	coffee		
pasta	milk		

It is similar but not equal to the h-index for published authors

It can be computed easily with a single scan of the dataset

Theorem: The VC-dimension is less or equal to the d-index d of $\mathcal{D}$

How do we prove the bound?

Theorem: The VC-dimension is less or equal to the d-index d of $\mathcal{D}$

Proof:

- Let $\ell > d$ and assume it is possible to shatter a set $T \subseteq \mathcal{D}$ with $|T| = \ell$.
- Then any $\tau \in T$ appears in at least $2^{\ell-1}$ ranges $\mathcal{T}_A$ (there are $2^{\ell-1}$ subsets of T containing τ)
- But any τ only appears in the ranges $\mathcal{T}_A$ such that $A \subseteq \tau$. So it appears in $2^{|\tau|} - 1$ ranges
- From the definition of d, T must contain a transaction τ^* of length $|\tau^*| < \ell$
- This implies $2^{|\tau^*|} - 1 < 2^{\ell-1}$, so τ^* can not appear in $2^{\ell-1}$ ranges
- Then T can not be shattered. We reach a contradiction and the thesis is true

This theorem allows us to use the VC ε-sample theorem

What is the algorithm then?

$d \leftarrow$ d-index of $\mathcal{D}$
$r \leftarrow \frac{1}{\varepsilon^2}\left(d + \ln\frac{1}{\delta}\right)$
`sample size`
$\mathcal{S} \leftarrow \emptyset$
for $i \leftarrow 1, \ldots, r$ **do**
    $\tau_i \leftarrow$ random transaction from $\mathcal{D}$, chosen uniformly
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{\tau_i\}$
end
Compute $\mathsf{FI}(\mathcal{S}, \theta - \varepsilon/2)$ using exact algorithm `// Faster algos make our approach faster!`
Output $\mathsf{FI}(\mathcal{S}, \theta - \varepsilon/2)$

Theorem: The output of the algorithm is a (ε, δ)-approximation
We just proved it!

How does it perform in practice?

Very well!

Great speedup w.r.t. an exact algorithm mining the whole dataset

Gets better as $\mathcal{D}$ grows, because the sample size does not depend on $|\mathcal{D}|$

Sample is small: 10^5 transactions for $\varepsilon = 0.01$, $\delta = 0.1$

The output always had the desired properties, not just with prob. $1 - \delta$

Maximum error $|f_{\mathcal{S}}(A) - f_{\mathcal{D}}(A)|$ much smaller than ε