Illustrating Agnostic Learning

We want a classifier to distinguish between *cats* and *dogs*

	Image 1	Image 2	Image 3	Image 4
x				
$c(x)$	Cat	Dog	Ostrich?	Sensor Error

Unrealizable (Agnostic) Learning

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

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

- We estimate $\Pr_{\mathcal{D}}(h(x) \neq c(x))$ by

$$\tilde{Pr}(h(x) \neq c(x)) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{h(x_i) \neq c(x_i)}$$

Unrealizable (Agnostic) Learning

- We estimate $\Pr_{\mathcal{D}}(h(x) \neq c(x))$ by

$$\tilde{Pr}(h(x) \neq c(x)) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{h(x_i) \neq c(x_i)}$$

- If for all h we have:

$$\left| \widehat{\Pr}(h(x) \neq c(x)) - \Pr_{x \sim \mathcal{D}}(h(x) \neq c(x)) \right| \leq \frac{\epsilon}{2} \ ,$$

  then the ERM (Empirical Risk Minimization) algorithm

$$\hat{h} = \arg \min_{h \in \mathcal{C}} \hat{\Pr}(h(x) \neq c(x))$$

  is ϵ-optimal.

More General Formalization

- Let f_h be the loss (error) function for hypothesis h (often denoted by $\ell(h, x)$).
- Here we use the 0-1 loss function:

$$f_h(x) = \begin{cases} 0 & \text{if } h(x) = c(x) \\ 1 & \text{if } h(x) \neq c(x) \end{cases}$$

- Alternative that gives higher lose to false negative.

$$f_h(x) = \begin{cases} 0 & \text{if } h(x) = c(x) \\ 1 + c(x) & \text{if } h(x) \neq c(x) \end{cases}$$

- Let $\mathcal{F}_C = \{f_h \mid h \in C\}$.
- $\mathcal{F}_C$ has the uniform convergence property $\Rightarrow$ if for any distribution $\mathcal{D}$ and hypothesis $h \in C$ we have a good estimate for the loss function of h

Uniform Convergence

Definition

A range space $(X, \mathcal{R})$ has the *uniform convergence property if for every* $\epsilon, \delta > 0$ *there is a sample size* $m = m(\epsilon, \delta)$ *such that for every distribution* $\mathcal{D}$ *over* X*, if* S *is a random sample from* $\mathcal{D}$ *of size* m *then, with probability at least* $1 - \delta$*,* S *is an* ϵ*-sample for* X *with respect to* $\mathcal{D}$*.*

Theorem

The following three conditions are equivalent:

1. *A concept class* $\mathcal{C}$ *over a domain* X *is agnostic PAC learnable.*
2. *The range space* $(X, \mathcal{C})$ *has the uniform convergence property.*
3. *The range space* $(X, \mathcal{C})$ *has a finite VC dimension.*

Is Uniform Convergence Necessary?

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

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

The general supervised learning scheme:

- Let $\mathcal{F}_{\mathcal{H}} = \{f_h \mid h \in H\}$.
- $\mathcal{F}_H$ has the uniform convergence property $\Rightarrow$ for any distribution $\mathcal{D}$ and hypothesis $h\{\mathcal{C}\}$ we have a good estimate of the error of h
- An ERM (Empirical Risk Minimization) algorithm correctly identify an almost best hypothesis in $\mathcal{H}$.

Is Uniform Convergence Necessary?

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

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

- We don't need uniform convergence for any distribution $\mathcal{D}$, just for the input (training set) distribution– Rademacher average.
- We don't need tight estimate for all functions, only for functions in neighborhood of the optimal function – local Rademacher average.

Rademacher Complexity

Limitations of the VC-Dimension Approach:

- Hard to compute
- Combinatorial bound - ignores the distribution over the data.

Rademacher Averages:

- Incorporates the input distribution
- Applies to general functions not just classification
- Always at least as good bound as the VC-dimension
- Can be computed from a sample
- Still hard to compute

Rademacher Averages - Motivation

- Assume that S_1 and S_2 are sufficiently large samples for estimating the expectations of any function in $\mathcal{F}$. Then, for any $f \in \mathcal{F}$,

$$\frac{1}{|S_1|}\sum_{x\in S_1} f(x) \approx \frac{1}{|S_2|}\sum_{y\in S_2} f(y) \approx E[f(x)],$$

  or

$$E_{S_1,S_2\sim\mathcal{D}}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{|S_1|}\sum_{x\in S_1} f(x) - \frac{1}{|S_2|}\sum_{y\in S_2} f(y)\right)\right] \leq \epsilon$$

- *Rademacher Variables:* Instead of two samples, we can take one sample $S = \{z_1, \ldots, z_m\}$ and split it randomly.
- Let $\sigma = \sigma_1, \ldots, \sigma_m$ i.i.d $Pr(\sigma_i = -1) = Pr(\sigma_i = 1) = 1/2$. The *Empirical Rademacher Average* of $\mathcal{F}$ is defined as

$$\tilde{R}_m(\mathcal{F}, S) = E_\sigma\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i)\right]$$

Rademacher Averages (Complexity)

Definition

The *Empirical Rademacher Average* of $\mathcal{F}$ with respect to a sample $S = \{z_1, \ldots, z_m\}$ and $\sigma = \sigma_1, \ldots, \sigma_m$, is defined as

$$\tilde{R}_m(\mathcal{F}, S) = E_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i) \right]$$

Taking an expectation over the distribution $\mathcal{D}$ of the samples:

Definition

The *Rademacher Average* of $\mathcal{F}$ is defined as

$$R_m(\mathcal{F}) = E_{S \sim \mathcal{D}}[\tilde{R}_m(\mathcal{F}, S)] = E_{S \sim \mathcal{D}} E_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i) \right]$$

The Major Results

We first show that the Rademacher Average indeed captures the expected error in estimating the expectation of any function in a set of functions $\mathcal{F}$ (The Generalization Error).

- Let $E_{\mathcal{D}}[f(z)]$ be the true expectation of a function f with distribution $\mathcal{D}$.
- For a sample $S = \{z_1, \ldots, z_m\}$ the empirical estimate of $E_{\mathcal{D}}[f(z)]$ using the sample S is $\frac{1}{m}\sum_{i=1}^{m} f(z_i)$.

Theorem

$$E_{S\sim\mathcal{D}}\left[\sup_{f\in\mathcal{F}}\left(E_{\mathcal{D}}[f(z)] - \frac{1}{m}\sum_{i=1}^{m} f(z_i)\right)\right] \leq 2R_m(\mathcal{F}).$$

Jensen's Inequality

Definition

A function $f : \mathbb{R}^m \to \mathbb{R}$ is said to be *convex* if, for any x_1, x_2 and $0 \leq \lambda \leq 1$,

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2).$$

Theorem (Jenssen's Inequality)

If f is a convex function, then

$$f(\mathbf{E}[X]) \leq E[f(X)].$$

In particular

$$\sup_{f \in \mathcal{F}} E[f] \leq E[\sup_{f \in \mathcal{F}} f]$$

Proof

Pick a second sample $S' = \{z'_1, \ldots, z'_m\}$.

$$
\begin{aligned}
& E_{S\sim\mathcal{D}}\left[\sup_{f\in\mathcal{F}}\left(E_{\mathcal{D}}[f(z)] - \frac{1}{m}\sum_{i=1}^{m} f(z_i)\right)\right] \\
= \; & E_{S\sim\mathcal{D}}\left[\sup_{f\in\mathcal{F}}\left(E_{S'\sim\mathcal{D}}\frac{1}{m}\sum_{i=1}^{m} f(z'_i) - \frac{1}{m}\sum_{i=1}^{m} f(z_i)\right)\right] \\
\leq \; & E_{S,S'\sim\mathcal{D}}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m} f(z'_i) - \frac{1}{m}\sum_{i=1}^{m} f(z_i)\right)\right] \qquad \text{Jensen's Inequlity} \\
= \; & E_{S,S',\sigma}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m} \sigma_i(f(z_i) - f(z'_i))\right)\right] \\
\leq \; & E_{S,\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m} \sigma_i(f(z_i)\right] + E_{S',\sigma}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m} \sigma_i(f(z'_i)\right] \\
= \; & 2R_m(\mathcal{F})
\end{aligned}
$$

Deviation Bounds

Theorem

Let $S = \{z_1, \ldots, z_n\}$ *be a sample from* $\mathcal{D}$ *and let* $\delta \in (0, 1)$. *If all* $f \in \mathcal{F}$ *satisfy* $A_f \leq f(z) \leq A_f + c$, *then*

1. *Bounding the estimate error using the Rademacher complexity:*

$$Pr(\sup_{f \in \mathcal{F}}(E_{\mathcal{D}}[f(z)] - \frac{1}{m}\sum_{i=1}^{m} f(z_i)) \geq 2R_m(\mathcal{F}) + \epsilon) \leq e^{-2m\epsilon^2/c^2}$$

2. *Bounding the estimate error using the empirical Rademacher complexity:*

$$Pr(\sup_{f \in \mathcal{F}}(E_{\mathcal{D}}[f(z)] - \frac{1}{m}\sum_{i=1}^{m} f(z_i)) \geq 2\tilde{R}_m(\mathcal{F}) + 2\epsilon) \leq 2e^{-2m\epsilon^2/c^2}$$

McDiarmid's Inequality

Applying Azuma inequality to Doob's martingale:

Theorem

Let $X_1, \ldots, X_n$ be independent random variables and let $h(x_1, \ldots, x_n)$ be a function such that a change in variable x_i can change the value of the function by no more than c_i,

$$\sup_{x_1,\ldots,x_n,x_i'} |h(x_1, \ldots, x_i, \ldots, x_n) - h(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i.$$

For any $\epsilon > 0$

$$Pr(h(X_1, \ldots, X_n) - E[h(X_1, \ldots, X_n)]| \geq \epsilon) \leq e^{-2\epsilon^2 / \sum_{i=1}^{n} c_i^2}.$$

Proof

- The generalization error

$$\sup_{f \in \mathcal{F}} (\mathbf{E}_{\mathcal{D}}[f(z)] - \frac{1}{m} \sum_{i=1}^{m} f(z_i))$$

  is a function of $z_1, \ldots, z_m$, and a change in one of the z_i changes the value of that function by no more than c/m.

- The *Empirical Rademacher Average*

$$\tilde{R}_m(\mathcal{F}, S) = E_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i) \right]$$

  is a function of m random variables, $z_1, \ldots, z_m$, and any change in one of these variables can change the value of $\tilde{R}_m(\mathcal{F}, S)$ by no more than c/m.

Estimating the Rademacher Complexity

Theorem (Massart's theorem)

Assume that $|\mathcal{F}|$ *is finite. Let* $S = \{z_1, \ldots, z_m\}$ *be a sample, and let*

$$B = \max_{f \in \mathcal{F}} \left(\sum_{i=1}^{m} f^2(z_i) \right)^{\frac{1}{2}}$$

then

$$\tilde{R}_m(\mathcal{F}, S) \leq \frac{B\sqrt{2 \ln |\mathcal{F}|}}{m}.$$

Hoeffding's Inequality

Large deviation bound for more general random variables:

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

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

Lemma

(Hoeffding's Lemma) Let X be a random variable such that $Pr(X \in [a, b]) = 1$ and $E[X] = 0$. Then for every $\lambda > 0$,

$$\mathbf{E}[e^{\lambda X}] \leq e^{\lambda^2(a-b)^2/8}.$$

Proof

For any $s > 0$,

$$
\begin{aligned}
e^{sm\tilde{R}_m(\mathcal{F},S)} &= e^{s\mathbf{E}_\sigma[\sup_{f\in\mathcal{F}}\sum_{i=1}^m \sigma_i f(z_i)]} \\
&\leq \mathbf{E}_\sigma\left[e^{s\sup_{f\in\mathcal{F}}\sum_{i=1}^m \sigma_i f(z_i)}\right] && \text{Jensen's Inequlity} \\
&= \mathbf{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left(e^{\sum_{i=1}^m s\sigma_i f(z_i)}\right)\right] \\
&\leq \sum_{f\in\mathcal{F}}\mathbf{E}_\sigma\left[\left(e^{\sum_{i=1}^m s\sigma_i f(z_i)}\right)\right] \\
&= \sum_{f\in\mathcal{F}}\mathbf{E}_\sigma\left[\prod_{i=1}^m e^{s\sigma_i f(z_i)}\right] \\
&= \sum_{f\in\mathcal{F}}\prod_{i=1}^m \mathbf{E}_\sigma\left[e^{s\sigma_i f(z_i)}\right]
\end{aligned}
$$

$$e^{sm\tilde{R}_m(\mathcal{F},S)} \leq \sum_{f\in\mathcal{F}} \prod_{i=1}^{m} \mathbf{E}_\sigma \left[e^{s\sigma_i f(z_i)} \right]$$

Since $\mathbf{E}[\sigma_i f(z_i)] = 0$ and $-f(z_i) \leq \sigma_i f(z_i) \leq f(z_i)$, we can apply Hoeffding's Lemma to obtain

$$\mathbf{E}\left[e^{s\sigma_i f(z_i)} \right] \leq e^{s^2(2f(z_i))^2/8} = e^{\frac{s^2}{2} f(z_i)^2}.$$

Thus,

$$\begin{aligned} e^{sm\tilde{R}_m(\mathcal{F},S)} &= e^{s\mathbf{E}[\sup_{f\in\mathcal{F}} \sum_{i=1}^{m} \sigma_i f(z_i)]} \\ &\leq \sum_{f\in\mathcal{F}} \prod_{i=1}^{m} e^{\frac{s^2}{2} f(z_i)^2} \\ &= \sum_{f\in\mathcal{F}} e^{\frac{s^2}{2} \sum_{i=1}^{m} f(z_i)^2} \\ &\leq |\mathcal{F}| e^{\frac{s^2 B^2}{2}}. \end{aligned}$$

$$e^{sm\tilde{R}_m(\mathcal{F},S)} \leq |\mathcal{F}| e^{\frac{s^2B^2}{2}}.$$

Hence, for any $s > 0$,

$$\tilde{R}_m(\mathcal{F}, S) \leq \frac{1}{m}\left(\frac{\ln|\mathcal{F}|}{s} + \frac{sB^2}{2}\right).$$

Setting $s = \frac{\sqrt{2\ln|\mathcal{F}|}}{B}$ yields

$$\tilde{R}_m(\mathcal{F}, S) \leq \frac{B\sqrt{2\ln|\mathcal{F}|}}{m}.$$

Application: Learning a Binary Classification

Let $\mathcal{C}$ be a binary concept class defined on a domain X, and let $\mathcal{D}$ be a probability distribution on X. For each $x \in X$ let $c(x)$ be the correct classification of x.
For each hypothesis $h \in C$ we define a function $f_h(x)$ by

$$f_h(x) = \begin{cases} 1 & \text{if } h(x) = c(x) \\ -1 & \text{otherwise} \end{cases}$$

Let $\mathcal{F} = \{f_h \mid h \in \mathcal{C}\}$. Our goal is to find $h' \in \mathcal{C}$ such that with probability at least $1 - \delta$

$$\mathbf{E}[f_{h'}] \geq \sup_{f_h \in \mathcal{F}} \mathbf{E}[f_h] - \epsilon.$$

We give an upper bound on the required size of the training set using Rademacher complexity.

For each hypothesis $h \in C$ we define a function $f_h(x)$ by

$$f_h(x) = \begin{cases} 1 & \text{if } h(x) = c(x) \\ -1 & \text{otherwise} \end{cases}$$

Let S be a sample of size m, then

$$B = \max_{f \in \mathcal{F}} \left(\sum_{i=1}^{m} f^2(z_i) \right)^{\frac{1}{2}} = \sqrt{m},$$

and

$$\tilde{R}_m(\mathcal{F}, S) \leq \sqrt{\frac{2 \ln |\mathcal{F}|}{m}}.$$

To use

$$Pr(\sup_{f \in \mathcal{F}} (E_{\mathcal{D}}[f(z)] - \frac{1}{m} \sum_{i=1}^{m} f(z_i)) \geq 2\tilde{R}_m(\mathcal{F}) + 2\epsilon) \leq 2e^{-2m\epsilon^2/c^2}$$

We need $\sqrt{\frac{2 \ln |\mathcal{F}|}{m}} \leq \frac{\epsilon}{4}$ and $2e^{-2m\epsilon^2/64} \leq \delta$.

Relation to VC-dimension

We express this bound in terms of the VC dimension of the concept class $\mathcal{C}$.
Each function $f_h \in \mathcal{F}$ corresponds to an hypothesis $h \in \mathcal{C}$.
Let d be the VC dimension of $\mathcal{C}$.
The projection of the range space $(X, \mathcal{C})$ on a sample of size m has no more than m^d different sets.
Thus, the set of different functions we need to consider is bounded by m^d, and

$$\tilde{R}_m(\mathcal{F}, S) \leq \sqrt{\frac{2d \ln m}{m}}.$$

Exercise: compare the the bounds obtained using the VC-dimension and the Rademacher complexity methods.

Back to Frequent Itemsets [Riondato and U. - KDD'15]

We define the task as an expectation estimation task:

- The domain is the dataset $\mathcal{D}$ (set of transactions)
- The family of functions is $\mathcal{F} = \{\mathbf{1}_A, A \subseteq 2^{\mathcal{I}}\}$, where $\mathcal{I}_A(\tau) = 1$ if $A \subseteq \tau$, else $\mathcal{I}_A(\tau) = 0$.
- The distribution π is uniform over $\mathcal{D}$: $\pi(\tau) = 1/|\mathcal{D}|$, for each $\tau \in \mathcal{D}$

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

Given a sample $z_1, \ldots, z_m$ of m transactions we need to bound the empirical Rademacher average

$$\tilde{R}_m(\mathcal{F}, S) = E_\sigma \left[\sup_{A \subseteq 2^{\mathcal{I}}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \mathbf{1}_A(z_i) \right]$$

How can we bound the Rademacher average? (high level picture)

Efficiency Constraint: use only information that can be obtained with a single scan of $\mathcal{S}$

How:

1. Prove a variant of Massart's Theorem.
2. Show that it's sufficient to consider only Closed Itemsets (CIs) in $\mathcal{S}$ (An itemset is closed iff none of its supersets has the same frequency)
3. We use the frequency of the single items and the lengths of the transactions to define a (conceptual) partitioning of the CIs into classes, and to compute upper bounds to the size of each class and to the frequencies of the CIs in the class
4. We use these bounds to compute an upper bound to $\mathrm{R}(\mathcal{S})$ by minimizing a convex function in $\mathbb{R}^+$ (no constraints)

Progressive Random Sampling

- Key question: How much to sample from $\mathcal{D}$ to obtain an (ε, δ)-approximation?
- The VC-dimension method gives a sufficient sample size, for a worst-case dataset with a given VC-dimension
- Instead, start sampling, and have the data tell us when to stop – we can get a better characterization of the data from the sample, and use it to reduce sample size

Progressive Random Sampling is an iterative sampling scheme

Algorithm for approximating $\mathsf{FI}(\mathcal{D}, \theta)$

At each iteration,

1. create sample $\mathcal{S}$ by drawing transactions from $\mathcal{D}$ uniformly and independently at random
2. Check a stopping condition on $\mathcal{S}$, by computing $\tilde{R}_m(\mathcal{F}, S)$ and checking if it gives an (ε, δ)-approximation
3. If stopping condition is satisfied, mine $\mathsf{FI}(\mathcal{S}, \gamma)$ for some $\gamma < \theta$ and output it
4. Else, iterate with a larger sample

Experimental Evaluation

Greatly improved runtime over exact algorithm, one-shot sampling (vc), and fixed geometric schedules. Better and better than exact as $\mathcal{D}$ grows

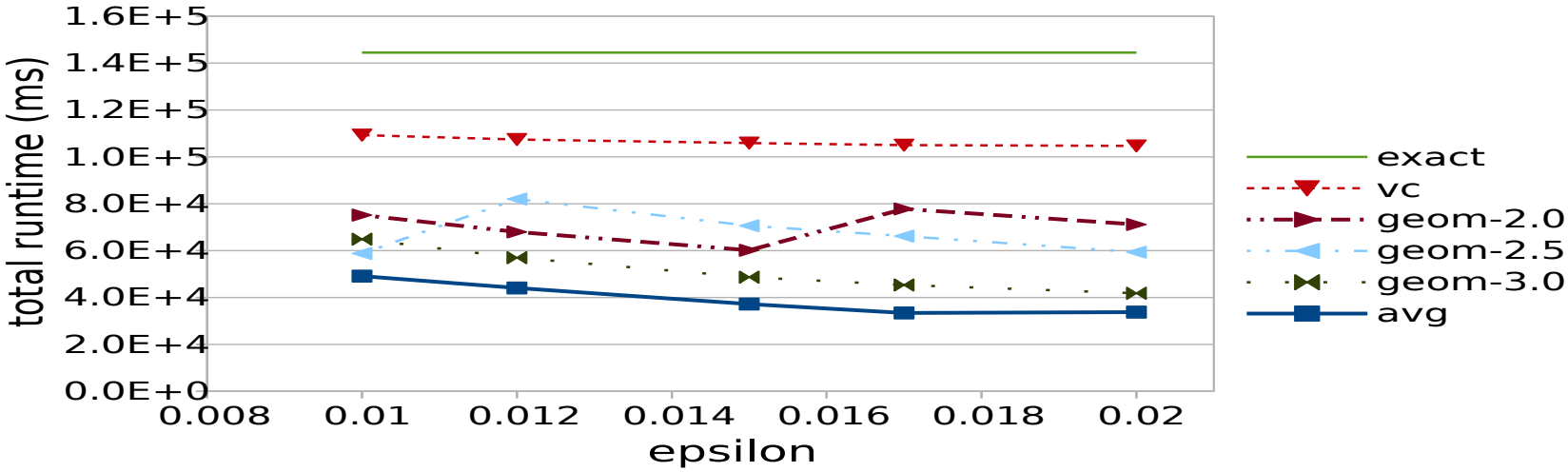


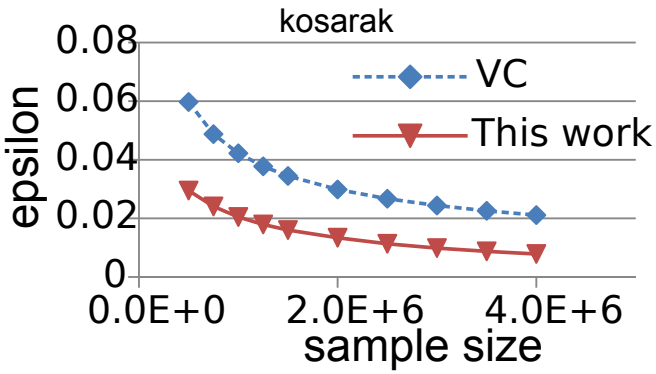
Figure: Running time for BMS-POS, $\theta = 0.015$.

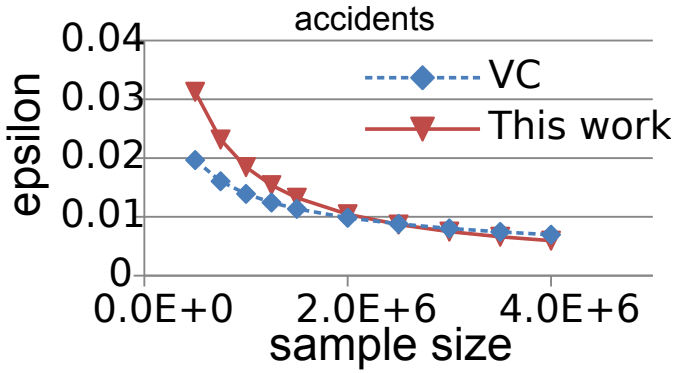
In 10K+ runs, the output was always an ε-approximation, not just with prob. $\geq 1 - \delta$

$\sup_{A \subseteq \mathcal{I}} |f_{\mathcal{D}}(A) - f_S(A)|$ is 10x smaller than ε (50x smaller on average)

How does it compare to the VC-dimension algorithm?

Given a sample $\mathcal{S}$ and some $\delta \in (0,1)$, what is the smallest ε such that $\mathsf{FI}(\mathcal{S}, \theta - \varepsilon/2)$ is a (ε, δ)-approximation?





Note that this comparison is unfavorable to our algorithm: as we are allowing the VC-dimension approach to compute the d-index of $\mathcal{D}$ (but we don't have access to $\mathcal{D}$!)

We strongly believe that this is because we haven't optimized all the aspects of the bound to the Rademacher average. Once we do it, the Rademacher avg approach will most probably always be better