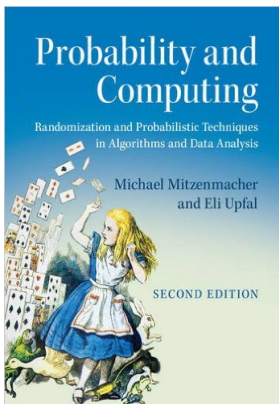






CS155/254: Probabilistic Methods in Computer Science

Chapter 14.3: Rademacher Complexity



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	DOG	SENSOR ERROR

Unrealizable (Agnostic) Learning

- We are given a training set $\{(x_1, c(x_1)), \dots, (x_m, c(x_m))\}$, 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:

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

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 .
- So far we used the 0-1 loss function:

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

- Alternatives that give higher/lower loss to false negative:

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

- 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 f_h , the loss of h .

Uniform Convergence

So far we only discussed binary classification with $0 - 1$ loss function.

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, \dots, z_m of size $m = m_{\mathcal{F}}(\epsilon, \delta)$ satisfies

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

The general supervised learning scheme:

- f_h is the loss (error) for hypothesis h . $\mathcal{F}_C = \{f_h \mid h \in C\}$.
- \mathcal{F}_C has the uniform convergence property \Rightarrow for any distribution D and hypothesis $h \in C$ we have a good estimate of the error of h
- An ERM (Empirical Risk Minimization) algorithm is ϵ -optimal

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, \dots, z_m of size $m = m_{\mathcal{F}}(\epsilon, \delta)$ satisfies

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

- 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 $S = \{z_1, \dots, z_m\}$ and split it randomly.
- Let $\sigma = \sigma_1, \dots, \sigma_m$ 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

Let $\sigma = \sigma_1, \dots, \sigma_m$ i.i.d $Pr(\sigma_i = -1) = Pr(\sigma_i = 1) = 1/2$.

The *Empirical Rademacher Average* of \mathcal{F} with respect to a sample $S = \{z_1, \dots, z_m\}$, 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

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

Intuition

Definition

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

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

Assume that $\forall f \in \mathcal{F}, f : Z \rightarrow \{-1, 1\}$.

If $|\mathcal{F}| = 1$, then $R_m(\mathcal{F}) = 0$.

If $|\mathcal{F}| = 2^n$, then $R_m(\mathcal{F}) = 1$. (For any assignment $\sigma_1, \dots, \sigma_m$, and z_1, \dots, z_m there is a function $f \in \mathcal{F}$ such that $\frac{1}{m} \sum_{i=1}^m \sigma_i f(z_i) = 1$.)

The *Rademacher Average* $0 \leq R_m(\mathcal{F}) \leq 1$ is another measure of the complexity or expressiveness of \mathcal{F} .

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 $S = \{z_1, \dots, z_m\}$ 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 $f : \mathbb{R}^m \rightarrow \mathbb{R}$ is said to be *convex* if, for any x_1, x_2 and $0 \leq \lambda \leq 1$,

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

Theorem (Jensen's Inequality)

If f is a convex function, then

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

In particular, $\mathbf{E}[\sup_{f \in \mathcal{F}} f] \geq \sup_{f \in \mathcal{F}} \mathbf{E}[f]$

Symmetrization Inequality: Proof

Function family: $\mathcal{F} \subseteq \mathcal{Z} \rightarrow \mathbb{R}$; samples: $z, z' \sim \mathcal{D}^m$;
 $\sigma \sim \text{Rademacher}^m = \{-1, 1\}^m$. $\Pr(-1) = \Pr(1) = 1/2$.

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

Start with the *supremum deviation* (which we want to bound)

Symmetrization Inequality: Proof

Function family: $\mathcal{F} \subseteq Z \rightarrow \mathbb{R}$; samples: $\mathbf{z}, \mathbf{z}' \sim \mathcal{D}^m$;
 $\sigma \sim \text{Rademacher}^m = \{-1, 1\}^m$. $\Pr(-1) = \Pr(1) = 1/2$.

$$\mathbf{E}_{\mathbf{z}} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \mathbf{E}_{\mathcal{D}}[f] \right) \right] = \mathbf{E}_{\mathbf{z}} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \mathbf{E}_{\mathbf{z}'} \left[\frac{1}{m} \sum_{i=1}^m f(z'_i) \right] \right) \right] \quad \text{Linearity of Expectation}$$

Symmetrization Inequality: Proof

Function family: $\mathcal{F} \subseteq \mathcal{Z} \rightarrow \mathbb{R}$; samples: $z, z' \sim \mathcal{D}^m$;
 $\sigma \sim \text{Rademacher}^m = \{-1, 1\}^m$. $\Pr(-1) = \Pr(1) = 1/2$.

$$\begin{aligned} \mathbb{E}_z \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \mathbb{E}_{\mathcal{D}}[f] \right) \right] &= \mathbb{E}_z \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \mathbb{E}_{z'} \left[\frac{1}{m} \sum_{i=1}^m f(z'_i) \right] \right) \right] && \text{Linearity of Expectation} \\ &\leq \mathbb{E}_z \left[\mathbb{E}_{z'} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \frac{1}{m} \sum_{i=1}^m f(z'_i) \right) \middle| z \right] \right] && \text{Jensen's Inequality} \end{aligned}$$

Jensen's Inequality: Easier to pick an $f \in \mathcal{F}$ to fit (maximize) each draw of z' individually than to pick an $f \in \mathcal{F}$ that has to work in expectation over z'

Symmetrization Inequality: Proof

Function family: $\mathcal{F} \subseteq Z \rightarrow \mathbb{R}$; samples: $z, z' \sim \mathcal{D}^m$;
 $\sigma \sim \text{Rademacher}^m = \{-1, 1\}^m$. $Pr(-1) = Pr(1) = 1/2$.

$$\begin{aligned} \mathbf{E}_z \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \mathbf{E}_{\mathcal{D}}[f] \right) \right] &= \mathbf{E}_z \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \mathbf{E}_{z'} \left[\frac{1}{m} \sum_{i=1}^m f(z'_i) \right] \right) \right] && \text{Linearity of Expectation} \\ &\leq \mathbf{E}_z \left[\mathbf{E}_{z'} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \frac{1}{m} \sum_{i=1}^m f(z'_i) \right) \middle| z \right] \right] && \text{Jensen's Inequality} \\ &= \mathbf{E}_{z, z'} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \frac{1}{m} \sum_{i=1}^m f(z'_i) \right) \right] && \text{Conditional Expectation} \end{aligned}$$

Symmetrization Inequality: Proof

Function family: $\mathcal{F} \subseteq \mathcal{Z} \rightarrow \mathbb{R}$; samples: $z, z' \sim \mathcal{D}^m$;
 $\sigma \sim \text{Rademacher}^m = \{-1, 1\}^m$. $Pr(-1) = Pr(1) = 1/2$.

$$\begin{aligned} \mathbb{E}_z \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \mathbb{E}_{\mathcal{D}}[f] \right) \right] &= \mathbb{E}_z \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \mathbb{E}_{z'} \left[\frac{1}{m} \sum_{i=1}^m f(z'_i) \right] \right) \right] && \text{Linearity of Expectation} \\ &\leq \mathbb{E}_z \left[\mathbb{E}_{z'} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \frac{1}{m} \sum_{i=1}^m f(z'_i) \right) \middle| z \right] \right] && \text{Jensen's Inequality} \\ &= \mathbb{E}_{z, z'} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \frac{1}{m} \sum_{i=1}^m f(z'_i) \right) \right] && \text{Conditional Expectation} \\ &= \mathbb{E}_{z, z'} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m (f(z_i) - f(z'_i)) \right] && \text{Algebra} \end{aligned}$$

Symmetrization Inequality: Proof

Function family: $\mathcal{F} \subseteq Z \rightarrow \mathbb{R}$; samples: $z, z' \sim \mathcal{D}^m$;
 $\sigma \sim \text{Rademacher}^m = \{-1, 1\}^m$. $Pr(-1) = Pr(1) = 1/2$.

$$\begin{aligned} \mathbf{E}_z \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \mathbf{E}_{\mathcal{D}}[f] \right) \right] &= \mathbf{E}_z \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \mathbf{E}_{z'} \left[\frac{1}{m} \sum_{i=1}^m f(z'_i) \right] \right) \right] && \text{Linearity of Expectation} \\ &\leq \mathbf{E}_{z, z'} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m (f(z_i) - f(z'_i)) \right] && \text{Jensen's Inequality} \end{aligned}$$

Consolidate the previous few steps

Symmetrization Inequality: Proof

Function family: $\mathcal{F} \subseteq \mathcal{Z} \rightarrow \mathbb{R}$; samples: $\mathbf{z}, \mathbf{z}' \sim \mathcal{D}^m$;
 $\sigma \sim \text{Rademacher}^m = \{-1, 1\}^m$. $\Pr(-1) = \Pr(1) = 1/2$.

$$\begin{aligned} \mathbb{E}_{\mathbf{z}} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \mathbb{E}_{\mathcal{D}}[f] \right) \right] &= \mathbb{E}_{\mathbf{z}} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \mathbb{E}_{\mathbf{z}'} \left[\frac{1}{m} \sum_{i=1}^m f(z'_i) \right] \right) \right] && \text{Linearity of Expectation} \\ &\leq \mathbb{E}_{\mathbf{z}, \mathbf{z}'} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m (f(z_i) - f(z'_i)) \right] && \text{Jensen's Inequality} \\ &= \mathbb{E}_{\mathbf{z}, \mathbf{z}', \sigma} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i (f(z_i) - f(z'_i)) \right] && \text{Symmetry} \end{aligned}$$

Symmetry: Since \mathbf{z}, \mathbf{z}' are i.i.d., swapping z_i, z'_i is *equally likely*.
 $\sigma_i = 1$: z_i, z'_i not swapped. $\sigma_i = -1$: z_i, z'_i swapped.

Symmetrization Inequality: Proof

Function family: $\mathcal{F} \subseteq Z \rightarrow \mathbb{R}$; samples: $z, z' \sim \mathcal{D}^m$;
 $\sigma \sim \text{Rademacher}^m = \{-1, 1\}^m$. $Pr(-1) = Pr(1) = 1/2$.

$$\begin{aligned} \mathbb{E}_z \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \mathbb{E}_{\mathcal{D}}[f] \right) \right] &= \mathbb{E}_z \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \mathbb{E}_{z'} \left[\frac{1}{m} \sum_{i=1}^m f(z'_i) \right] \right) \right] && \text{Linearity of Expectation} \\ &\leq \mathbb{E}_{z, z'} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m (f(z_i) - f(z'_i)) \right] && \text{Jensen's Inequality} \\ &= \mathbb{E}_{z, z', \sigma} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i (f(z_i) - f(z'_i)) \right] && \text{Symmetry} \\ &\leq \mathbb{E}_{z, z', \sigma} \left[\left(\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i f(z_i) \right) + \left(\sup_{f \in \mathcal{F}} -\frac{1}{m} \sum_{i=1}^m \sigma_i f(z'_i) \right) \right] && \text{Subadditivity} \end{aligned}$$

Symmetrization Inequality: Proof

Function family: $\mathcal{F} \subseteq \mathcal{Z} \rightarrow \mathbb{R}$; samples: $z, z' \sim \mathcal{D}^m$;
 $\sigma \sim \text{Rademacher}^m = \{-1, 1\}^m$. $Pr(-1) = Pr(1) = 1/2$.

$$\begin{aligned} \mathbb{E}_z \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \mathbb{E}_{\mathcal{D}}[f] \right) \right] &= \mathbb{E}_z \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \mathbb{E}_{z'} \left[\frac{1}{m} \sum_{i=1}^m f(z'_i) \right] \right) \right] && \text{Linearity of Expectation} \\ &\leq \mathbb{E}_{z, z'} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m (f(z_i) - f(z'_i)) \right] && \text{Jensen's Inequality} \\ &= \mathbb{E}_{z, z', \sigma} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i (f(z_i) - f(z'_i)) \right] && \text{Symmetry} \\ &\leq \mathbb{E}_{z, z', \sigma} \left[\left(\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i f(z_i) \right) + \left(\sup_{f \in \mathcal{F}} -\frac{1}{m} \sum_{i=1}^m \sigma_i f(z'_i) \right) \right] && \text{Subadditivity} \\ &= \mathbb{E}_{z, \sigma} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i f(z_i) \right] + \mathbb{E}_{z', \sigma} \left[\sup_{f \in \mathcal{F}} -\frac{1}{m} \sum_{i=1}^m \sigma_i f(z'_i) \right] && \text{Linearity of Expectation} \end{aligned}$$

Symmetrization Inequality: Proof

Function family: $\mathcal{F} \subseteq Z \rightarrow \mathbb{R}$; samples: $z, z' \sim \mathcal{D}^m$;
 $\sigma \sim \text{Rademacher}^m = \{-1, 1\}^m$. $Pr(-1) = Pr(1) = 1/2$.

$$\begin{aligned} \mathbf{E}_z \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \mathbf{E}_{\mathcal{D}}[f] \right) \right] &= \mathbf{E}_z \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \mathbf{E}_{z'} \left[\frac{1}{m} \sum_{i=1}^m f(z'_i) \right] \right) \right] && \text{Linearity of Expectation} \\ &\leq \mathbf{E}_{z, z'} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m (f(z_i) - f(z'_i)) \right] && \text{Jensen's Inequality} \\ &= \mathbf{E}_{z, z', \sigma} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i (f(z_i) - f(z'_i)) \right] && \text{Symmetry} \\ &\leq \mathbf{E}_{z, \sigma} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i f(z_i) \right] + \mathbf{E}_{z', \sigma} \left[\sup_{f \in \mathcal{F}} -\frac{1}{m} \sum_{i=1}^m \sigma_i f(z'_i) \right] && \text{Subadditivity} \end{aligned}$$

Consolidate the previous few steps

Symmetrization Inequality: Proof

Function family: $\mathcal{F} \subseteq \mathcal{Z} \rightarrow \mathbb{R}$; samples: $\mathbf{z}, \mathbf{z}' \sim \mathcal{D}^m$;
 $\sigma \sim \text{Rademacher}^m = \{-1, 1\}^m$. $\Pr(-1) = \Pr(1) = 1/2$.

$$\begin{aligned} \mathbf{E}_{\mathbf{z}} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \mathbf{E}_{\mathcal{D}}[f] \right) \right] &= \mathbf{E}_{\mathbf{z}} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \mathbf{E}_{\mathbf{z}'} \left[\frac{1}{m} \sum_{i=1}^m f(z'_i) \right] \right) \right] && \text{Linearity of Expectation} \\ &\leq \mathbf{E}_{\mathbf{z}, \mathbf{z}'} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m (f(z_i) - f(z'_i)) \right] && \text{Jensen's Inequality} \\ &= \mathbf{E}_{\mathbf{z}, \mathbf{z}', \sigma} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i (f(z_i) - f(z'_i)) \right] && \text{Symmetry} \\ &\leq \mathbf{E}_{\mathbf{z}, \sigma} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i f(z_i) \right] + \mathbf{E}_{\mathbf{z}', \sigma} \left[\sup_{f \in \mathcal{F}} -\frac{1}{m} \sum_{i=1}^m \sigma_i f(z'_i) \right] && \text{Subadditivity} \\ &= \mathbf{E}_{\mathbf{z}, \sigma} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i f(z_i) \right] + \mathbf{E}_{\mathbf{z}', \sigma} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i f(z'_i) \right] && \Pr(\sigma) = \Pr(-\sigma) \end{aligned}$$

Symmetrization Inequality: Proof

Function family: $\mathcal{F} \subseteq \mathcal{Z} \rightarrow \mathbb{R}$; samples: $\mathbf{z}, \mathbf{z}' \sim \mathcal{D}^m$;
 $\sigma \sim \text{Rademacher}^m = \{-1, 1\}^m$. $\Pr(-1) = \Pr(1) = 1/2$.

$$\begin{aligned} \mathbb{E}_{\mathbf{z}} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \mathbb{E}_{\mathcal{D}}[f] \right) \right] &= \mathbb{E}_{\mathbf{z}} \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^m f(z_i) - \mathbb{E}_{\mathbf{z}'} \left[\frac{1}{m} \sum_{i=1}^m f(z'_i) \right] \right) \right] && \text{Linearity of Expectation} \\ &\leq \mathbb{E}_{\mathbf{z}, \mathbf{z}'} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m (f(z_i) - f(z'_i)) \right] && \text{Jensen's Inequality} \\ &= \mathbb{E}_{\mathbf{z}, \mathbf{z}', \sigma} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i (f(z_i) - f(z'_i)) \right] && \text{Symmetry} \\ &\leq \mathbb{E}_{\mathbf{z}, \sigma} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i f(z_i) \right] + \mathbb{E}_{\mathbf{z}', \sigma} \left[\sup_{f \in \mathcal{F}} -\frac{1}{m} \sum_{i=1}^m \sigma_i f(z'_i) \right] && \text{Subadditivity} \\ &= \mathbb{E}_{\mathbf{z}, \sigma} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i f(z_i) \right] + \mathbb{E}_{\mathbf{z}', \sigma} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i f(z'_i) \right] && \Pr(\sigma) = \Pr(-\sigma) \\ &= 2\mathbb{E}_{\mathbf{z}, \sigma} \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \sigma_i f(z_i) \right] = 2R_m(\mathcal{F}, \mathcal{D}) && \mathbf{z}, \mathbf{z}' \sim \mathcal{D}^m \end{aligned}$$

Deviation Bounds

Theorem

Let $S = \{z_1, \dots, z_n\}$ 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, \dots, X_n be independent random variables and let $h(x_1, \dots, x_n)$ be a function such that a change in variable x_i can change the value of the function by no more than c_i ,

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

For any $\epsilon > 0$

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

Proof

The generalization error:

Let $g(z_1, \dots, z_n) = \sup_{f \in \mathcal{F}} (\mathbf{E}_{\mathcal{D}}[f(z)] - \frac{1}{m} \sum_{i=1}^m f(z_i))$ We want to bound

$$g(z_1, \dots, z_n) - E[g(z_1, \dots, z_n)] \leq g(z_1, \dots, z_n) - 2R_m(\mathcal{F})$$

$g(z_1, \dots, z_n)$ is a function of independent z_1, \dots, z_m . Assume that we change z_i to y_i .

If the $\arg \sup_{f \in \mathcal{F}}$ doesn't change then the value of the function changes by no more than c/m .

Assume that $\arg \sup_{f \in \mathcal{F}}$ changes from h to h' .

$$h(z_1, \dots, z_n) \geq h'(z_1, \dots, z_n) \geq h'(z_1, \dots, y_i, \dots, z_n) - c/m$$

and a change is again no more than c/m .

The estimation error:

We want to bound

$$\tilde{R}_m(\mathcal{F}, S) - R_m(\mathcal{F}, S).$$

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, \dots, 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 .

Why Data Dependent Bounds?

- The Vapnik-Chervonenkis Dimension:
 - Applies only to binary classification
 - Complicated generalizations for regression or multi-class classification
 - Combinatorial bound
 - Ignores *data distribution* (worst-case over all distributions)
 - Can be hard to compute
- Rademacher Averages:
 - Handles general learning problems
 - Only need a *loss function*
 - Classification, regression, clustering, data mining
 - Sensitive to *data distribution* (distribution-dependent)
 - Approximated with *training sample* (data-dependent)
 - Always at least as good bound as the VC-dimension
 - Still hard to compute

Bounding Rademacher Averages: Massart's Inequality

Theorem (Massart's Finite Class Inequality)

Assume that $|\mathcal{F}|$ is finite. Let $S = \{z_1, \dots, z_m\}$ be a sample. Then

$$\hat{R}_m(\mathcal{F}, S) \leq \sqrt{\sup_{f \in \mathcal{F}} \sum_{i=1}^m f^2(z_i)} * \frac{\sqrt{2 \ln |\mathcal{F}|}}{m}.$$

Corollary

Therefore, if $\mathcal{F} \subseteq \mathcal{X} \rightarrow [-1, 1]$, then

$$\sqrt{\sup_{f \in \mathcal{F}} \sum_{i=1}^m f^2(z_i)} \leq \sqrt{m}, \text{ thus } R_m(\mathcal{F}, \mathcal{D}) \leq \sqrt{\frac{2 \ln |\mathcal{F}|}{m}}.$$

Massart's Inequality: Proof Volume I

For any $\lambda > 0$:

$$\exp(\lambda m \hat{R}_m(\mathcal{F}, S)) = \exp\left(\lambda \mathbf{E}_\sigma \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^m \sigma_i f(z_i) \right]\right)$$

Definition

Massart's Inequality: Proof Volume I

For any $\lambda > 0$:

$$\begin{aligned}\exp(\lambda m \hat{R}_m(\mathcal{F}, S)) &= \exp\left(\lambda \mathbf{E}_\sigma \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^m \sigma_i f(z_i) \right]\right) \\ &\leq \mathbf{E}_\sigma \left[\exp\left(\lambda \sup_{f \in \mathcal{F}} \sum_{i=1}^m \sigma_i f(z_i)\right) \right]\end{aligned}$$

Definition

$$\exp(\mathbf{E}[\cdot]) \leq \mathbf{E}[\exp(\cdot)]$$

Jensen's Inequality

Massart's Inequality: Proof Volume I

For any $\lambda > 0$:

$$\begin{aligned}\exp(\lambda m \hat{R}_m(\mathcal{F}, S)) &= \exp\left(\lambda \mathbf{E}_\sigma \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^m \sigma_i f(z_i) \right]\right) \\ &\leq \mathbf{E}_\sigma \left[\exp\left(\lambda \sup_{f \in \mathcal{F}} \sum_{i=1}^m \sigma_i f(z_i)\right) \right] \\ &= \mathbf{E}_\sigma \left[\sup_{f \in \mathcal{F}} \exp\left(\sum_{i=1}^m \lambda \sigma_i f(z_i)\right) \right]\end{aligned}$$

Definition

$$\exp(\mathbf{E}[\cdot]) \leq \mathbf{E}[\exp(\cdot)]$$

Jensen's Inequality

$$\sup(\exp(\cdot)) = \exp(\sup(\cdot))$$

Monotonicity

Massart's Inequality: Proof Volume I

For any $\lambda > 0$:

$$\begin{aligned}\exp(\lambda m \hat{R}_m(\mathcal{F}, S)) &= \exp\left(\lambda \mathbf{E}_\sigma \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^m \sigma_i f(z_i) \right]\right) \\ &\leq \mathbf{E}_\sigma \left[\exp\left(\lambda \sup_{f \in \mathcal{F}} \sum_{i=1}^m \sigma_i f(z_i)\right) \right] \\ &= \mathbf{E}_\sigma \left[\sup_{f \in \mathcal{F}} \exp\left(\sum_{i=1}^m \lambda \sigma_i f(z_i)\right) \right] \\ &\leq \sum_{f \in \mathcal{F}} \mathbf{E}_\sigma \left[\exp\left(\sum_{i=1}^m \lambda \sigma_i f(z_i)\right) \right]\end{aligned}$$

Definition

$$\exp(\mathbf{E}[\cdot]) \leq \mathbf{E}[\exp(\cdot)]$$

Jensen's Inequality

$$\sup(\exp(\cdot)) = \exp(\sup(\cdot))$$

Monotonicity

$$\sup_{f \in \mathcal{F}} (\cdot) \leq \sum_{f \in \mathcal{F}} (\cdot) \text{ when positive}$$

Massart's Inequality: Proof Volume I

For any $\lambda > 0$:

$$\begin{aligned}\exp(\lambda m \hat{R}_m(\mathcal{F}, S)) &= \exp\left(\lambda \mathbf{E}_\sigma \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^m \sigma_i f(z_i) \right]\right) \\ &\leq \mathbf{E}_\sigma \left[\exp\left(\lambda \sup_{f \in \mathcal{F}} \sum_{i=1}^m \sigma_i f(z_i)\right) \right] \\ &= \mathbf{E}_\sigma \left[\sup_{f \in \mathcal{F}} \exp\left(\sum_{i=1}^m \lambda \sigma_i f(z_i)\right) \right] \\ &\leq \sum_{f \in \mathcal{F}} \mathbf{E}_\sigma \left[\exp\left(\sum_{i=1}^m \lambda \sigma_i f(z_i)\right) \right] \\ &= \sum_{f \in \mathcal{F}} \mathbf{E}_\sigma \left[\prod_{i=1}^m \exp(\lambda \sigma_i f(z_i)) \right]\end{aligned}$$

Definition

$$\exp(\mathbf{E}[\cdot]) \leq \mathbf{E}[\exp(\cdot)]$$

Jensen's Inequality

$$\sup(\exp(\cdot)) = \exp(\sup(\cdot))$$

Monotonicity

$$\sup_{f \in \mathcal{F}} (\cdot) \leq \sum_{f \in \mathcal{F}} (\cdot) \text{ when positive}$$

Properties of the Exponent

Massart's Inequality: Proof Volume I

For any $\lambda > 0$:

$$\begin{aligned}\exp(\lambda m \hat{R}_m(\mathcal{F}, S)) &= \exp\left(\lambda \mathbf{E}_\sigma \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^m \sigma_i f(z_i) \right]\right) && \text{Definition} \\ &\leq \mathbf{E}_\sigma \left[\exp\left(\lambda \sup_{f \in \mathcal{F}} \sum_{i=1}^m \sigma_i f(z_i)\right) \right] && \exp(\mathbf{E}[\cdot]) \leq \mathbf{E}[\exp(\cdot)] \\ &= \mathbf{E}_\sigma \left[\sup_{f \in \mathcal{F}} \exp\left(\sum_{i=1}^m \lambda \sigma_i f(z_i)\right) \right] && \text{Jensen's Inequality} \\ &\leq \sum_{f \in \mathcal{F}} \mathbf{E}_\sigma \left[\exp\left(\sum_{i=1}^m \lambda \sigma_i f(z_i)\right) \right] && \sup(\exp(\cdot)) = \exp(\sup(\cdot)) \\ &= \sum_{f \in \mathcal{F}} \mathbf{E}_\sigma \left[\prod_{i=1}^m \exp(\lambda \sigma_i f(z_i)) \right] && \text{Monotonicity} \\ &= \sum_{f \in \mathcal{F}} \prod_{i=1}^m \mathbf{E}_\sigma [\exp(\lambda \sigma_i f(z_i))] && \sup(\cdot) \leq \sum_{f \in \mathcal{F}} (\cdot) \text{ when positive} \\ & && \text{Properties of the Exponent} \\ & && \text{Independence}\end{aligned}$$

Massart's Inequality: Proof Volume II

Take $B^2 = \sup_{f \in \mathcal{F}} \sum_{i=1}^m f^2(z_i)$. For any $\lambda > 0$:

$$\exp(\lambda m \hat{R}_m(\mathcal{F}, S)) = \sum_{f \in \mathcal{F}} \prod_{i=1}^m \mathbb{E}_{\sigma} [\exp(\lambda \sigma_i f(z_i))]$$

[Previous Slide](#)

Massart's Inequality: Proof Volume II

Take $B^2 = \sup_{f \in \mathcal{F}} \sum_{i=1}^m f^2(z_i)$. For any $\lambda > 0$:

$$\begin{aligned} \exp(\lambda m \hat{R}_m(\mathcal{F}, S)) &= \sum_{f \in \mathcal{F}} \prod_{i=1}^m \mathbb{E}_{\sigma} [\exp(\lambda \sigma_i f(z_i))] \\ &= \sum_{f \in \mathcal{F}} \prod_{i=1}^m \frac{\exp(\lambda f(z_i)) + \exp(-\lambda f(z_i))}{2} \end{aligned}$$

Previous Slide

Definition of Expectation

Massart's Inequality: Proof Volume II

Take $B^2 = \sup_{f \in \mathcal{F}} \sum_{i=1}^m f^2(z_i)$. For any $\lambda > 0$:

$$\begin{aligned} \exp(\lambda m \hat{R}_m(\mathcal{F}, S)) &= \sum_{f \in \mathcal{F}} \prod_{i=1}^m \mathbb{E}_{\sigma} [\exp(\lambda \sigma_i f(z_i))] \\ &= \sum_{f \in \mathcal{F}} \prod_{i=1}^m \frac{\exp(\lambda f(z_i)) + \exp(-\lambda f(z_i))}{2} \\ &\leq \sum_{f \in \mathcal{F}} \prod_{i=1}^m \exp\left(\frac{\lambda^2}{2} f^2(z_i)\right) \end{aligned}$$

Previous Slide

Definition of Expectation

Hyperbolic Cosine Inequality

Massart's Inequality: Proof Volume II

Take $B^2 = \sup_{f \in \mathcal{F}} \sum_{i=1}^m f^2(z_i)$. For any $\lambda > 0$:

$$\exp(\lambda m \hat{R}_m(\mathcal{F}, S)) = \sum_{f \in \mathcal{F}} \prod_{i=1}^m \mathbb{E}_{\sigma} [\exp(\lambda \sigma_i f(z_i))]$$

Previous Slide

$$= \sum_{f \in \mathcal{F}} \prod_{i=1}^m \frac{\exp(\lambda f(z_i)) + \exp(-\lambda f(z_i))}{2}$$

Definition of Expectation

$$\leq \sum_{f \in \mathcal{F}} \prod_{i=1}^m \exp\left(\frac{\lambda^2}{2} f^2(z_i)\right)$$

Hyperbolic Cosine Inequality

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

Exponent Laws

Massart's Inequality: Proof Volume II

Take $B^2 = \sup_{f \in \mathcal{F}} \sum_{i=1}^m f^2(z_i)$. For any $\lambda > 0$:

$$\exp(\lambda m \hat{R}_m(\mathcal{F}, S)) = \sum_{f \in \mathcal{F}} \prod_{i=1}^m \mathbf{E}_{\sigma} [\exp(\lambda \sigma_i f(z_i))] \quad \text{Previous Slide}$$

$$= \sum_{f \in \mathcal{F}} \prod_{i=1}^m \frac{\exp(\lambda f(z_i)) + \exp(-\lambda f(z_i))}{2} \quad \text{Definition of Expectation}$$

$$\leq \sum_{f \in \mathcal{F}} \prod_{i=1}^m \exp\left(\frac{\lambda^2}{2} f^2(z_i)\right) \quad \text{Hyperbolic Cosine Inequality}$$

$$= \sum_{f \in \mathcal{F}} \exp\left(\frac{\lambda^2}{2} \sum_{i=1}^m f^2(z_i)\right) \quad \text{Exponent Laws}$$

$$\leq \sum_{f \in \mathcal{F}} \exp\left(\frac{\lambda^2 B^2}{2}\right) \quad \forall f_i \in \mathcal{F} : \sum_{i=1}^m f_i^2(z_i) \leq \sup_{f \in \mathcal{F}} \sum_{i=1}^m f^2(z_i) = B^2$$

Massart's Inequality: Proof Volume II

Take $B^2 = \sup_{f \in \mathcal{F}} \sum_{i=1}^m f^2(z_i)$. For any $\lambda > 0$:

$$\exp(\lambda m \hat{R}_m(\mathcal{F}, S)) = \sum_{f \in \mathcal{F}} \prod_{i=1}^m \mathbf{E}_{\sigma} [\exp(\lambda \sigma_i f(z_i))] \quad \text{Previous Slide}$$

$$= \sum_{f \in \mathcal{F}} \prod_{i=1}^m \frac{\exp(\lambda f(z_i)) + \exp(-\lambda f(z_i))}{2} \quad \text{Definition of Expectation}$$

$$\leq \sum_{f \in \mathcal{F}} \prod_{i=1}^m \exp\left(\frac{\lambda^2}{2} f^2(z_i)\right) \quad \text{Hyperbolic Cosine Inequality}$$

$$= \sum_{f \in \mathcal{F}} \exp\left(\frac{\lambda^2}{2} \sum_{i=1}^m f^2(z_i)\right) \quad \text{Exponent Laws}$$

$$\leq \sum_{f \in \mathcal{F}} \exp\left(\frac{\lambda^2 B^2}{2}\right) \quad \forall f_i \in \mathcal{F} : \sum_{i=1}^m f_i^2(z_i) \leq \sup_{f \in \mathcal{F}} \sum_{i=1}^m f^2(z_i) = B^2$$

$$\leq |\mathcal{F}| \exp\left(\frac{\lambda^2 B^2}{2}\right) \quad \text{Summation}$$

Massart's Inequality: Proof Volume III

Take $B^2 = \sup_{f \in \mathcal{F}} \sum_{i=1}^m f^2(z_i)$. For any $\lambda > 0$:

Previously: $\exp\left(\lambda m \hat{R}_m(\mathcal{F}, S)\right) \leq |\mathcal{F}| \exp\left(\frac{\lambda^2 B^2}{2}\right)$

Take logarithms, rearrange, and minimize with $\lambda = \frac{\sqrt{2 \ln |\mathcal{F}|}}{B}$

$$\hat{R}_m(\mathcal{F}, S) \leq \inf_{\lambda} \frac{1}{m} \left(\frac{\ln |\mathcal{F}|}{\lambda} + \frac{\lambda B^2}{2} \right) = \frac{B \sqrt{2 \ln |\mathcal{F}|}}{m} = \sqrt{\sup_{f \in \mathcal{F}} \sum_{i=1}^m f^2(z_i)} * \frac{\sqrt{2 \ln |\mathcal{F}|}}{m}$$

Takeaways:

- Lots of steps; all very simple
- Proof similar to Azuma-Hoeffding inequality
- Often loose, but lays foundation for better bounds

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 $h \in \mathcal{C}$ 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 $1 - \delta$

$$\mathbf{E}[f_{h'}] \leq \inf_{f_h \in \mathcal{F}} \mathbf{E}[f_h] + \epsilon.$$

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

For each hypothesis $h \in \mathcal{C}$ 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 $\epsilon' \leq \epsilon/4$, $\sqrt{\frac{2 \ln |\mathcal{F}|}{m}} \leq \frac{\epsilon}{4}$ 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

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

To have

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

We need $\tilde{R}_m(\mathcal{F}, S) \leq \sqrt{\frac{2d \ln m}{m}} \leq \frac{\epsilon}{4}$ and $2e^{-2m\epsilon^2/64} \leq \delta$, which requires

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

Using VC-dimension ϵ -sample we had

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

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

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 $\text{FI}(\mathcal{D}, \theta) = \{A \subseteq \mathcal{I} : f_{\mathcal{D}}(A) \geq \theta\}$

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

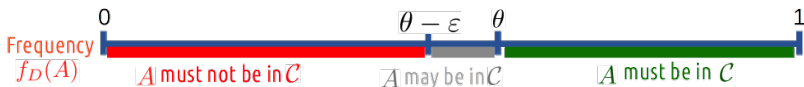
Trading off accuracy for speed: we get an approximation of $FI(\mathcal{D}, \theta)$ but we get it fast

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 $\text{FI}(\mathcal{D}, \theta)$ 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}

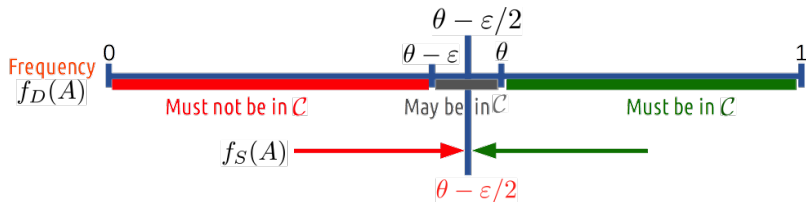
What do we really need?

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 $\text{FI}(\mathcal{S}, \theta - \varepsilon/2)$ 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

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

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

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

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

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

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

Thus,

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

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 we 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 of sets 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 of range space $(\mathcal{D}, \mathcal{T}_A)$

Bounding the VC-dimension

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 $2^\ell - 1 < 2^\ell \leq 2^{t-1}$ 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 $|S| \geq O(\frac{1}{\varepsilon^2} (\ell \log \ell + \ln \frac{1}{\delta}))$

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}

$r \leftarrow \frac{1}{\varepsilon^2} (d + \ln \frac{1}{\delta})$

sample size

$\mathcal{S} \leftarrow \emptyset$

for $i \leftarrow 1, \dots, r$ **do**

$\tau_i \leftarrow$ random transaction from \mathcal{D} , chosen uniformly

$\mathcal{S} \leftarrow \mathcal{S} \cup \{\tau_i\}$

end

Compute $\text{FI}(\mathcal{S}, \theta - \varepsilon/2)$ using exact algorithm // Faster
algorithms make our approach faster!

Output $\text{FI}(\mathcal{S}, \theta - \varepsilon/2)$

Theorem: The output of the algorithm is a (ε, δ) -approximation

We just proved it!

How does it perform in practice?

Very well!

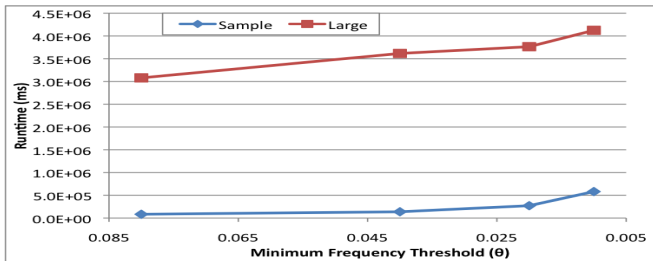
Great speedup w.r.t. an exact algorithm mining the whole dataset

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

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

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

Maximum error $|f_S(A) - f_D(A)|$ much smaller than ϵ



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 $\mathcal{F} = \{\mathbb{I}_A, A \subseteq 2^{\mathcal{I}}\}$, where $\mathbb{I}_A(\tau) = 1$ if $A \subseteq \tau$, else $\mathbb{I}_A(\tau) = 0$.
- The distribution π is uniform over \mathcal{D} : $\pi(\tau) = 1/|\mathcal{D}|$, for each $\tau \in \mathcal{D}$

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

Given a sample $\mathbf{z}_1, \dots, \mathbf{z}_m$ of m transactions we need to bound the empirical Rademacher average

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

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 $R(\mathcal{S})$ by minimizing a convex function in \mathbb{R}^+ (no constraints)

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

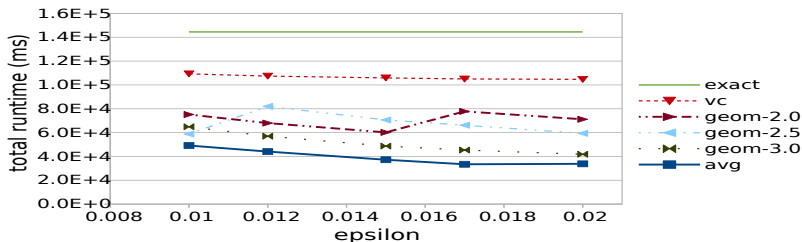


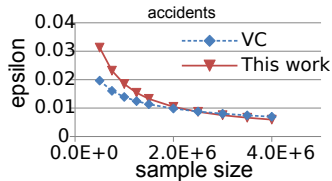
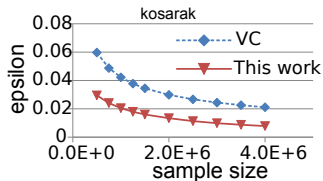
Figure: Running time for BMS-POS, $\theta = 0.015$.

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

$\sup_{A \subseteq \mathcal{I}} |f_{\mathcal{D}}(A) - f_{\mathcal{S}}(A)|$ 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 $\text{FI}(\mathcal{S}, \theta - \varepsilon/2)$ 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