Statistical Machine Learning

https://cvml.ist.ac.at/courses/SML_W18

Christoph Lampert

IST AUSTRIA

Institute of Science and Technology

Spring Semester 2018/2019

Lecture 7

Overview (tentative)

Date		no.	Topic
Oct 08	Mon	1	A Hands-On Introduction
Oct 10	Wed	–	self-study (Christoph traveling)
Oct 15	Mon	2	Bayesian Decision Theory
			Generative Probabilistic Models
Oct 17	Wed	3	Discriminative Probabilistic Models
			Maximum Margin Classifiers
Oct 22	Mon	4	Generalized Linear Classifiers, Optimization
Oct 24	Wed	5	Evaluating Predictors; Model Selection
Oct 29	Mon	–	self-study (Christoph traveling)
Oct 31	Wed	6	Overfitting/Underfitting, Regularization
Nov 05	Mon	7	Learning Theory I: classical/Rademacher bounds
Nov 07	Wed	8	Learning Theory II: miscellaneous
Nov 12	Mon	9	Probabilistic Graphical Models I
Nov 14	Wed	10	Probabilistic Graphical Models II
Nov 19	Mon	11	Probabilistic Graphical Models III
Nov 21	Wed	12	Probabilistic Graphical Models IV
until Nov 25			final project

Classical generalization bounds

Reminder: Finite Hypothesis Set

Setup:

- $\ell(y, \bar{y}) = [\![y \neq \bar{y}]\!]$ (0-1 loss)
- finite number of possible classifiers $\mathcal{H} = \{f_1, \dots, f_T\} \subset \mathcal{Y}^{\mathcal{X}}$

For any $\delta > 0$, the following statement holds with probability at least $1 - \delta$ over the training set $\mathcal{D} = \{(x^1, y^1) \dots, (x^n, y^n)\} \overset{i.i.d.}{\sim} p(x, y)$:

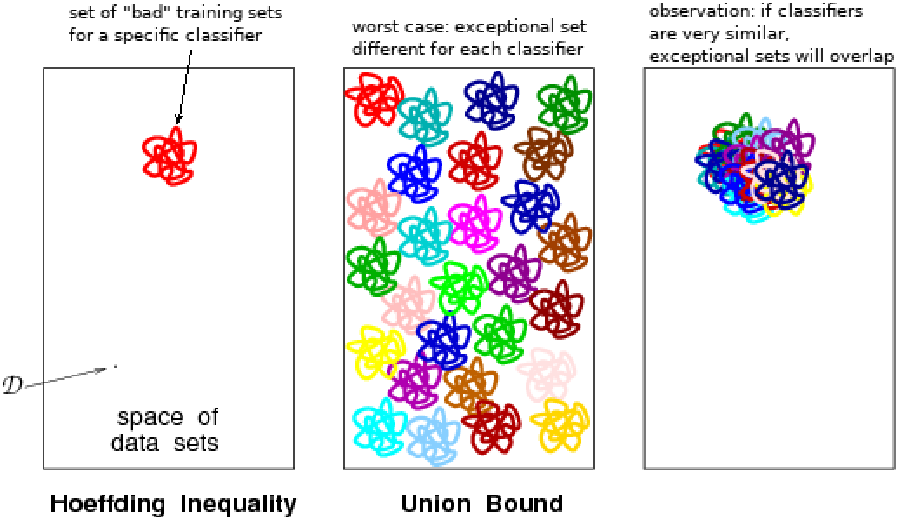
For all $f \in \mathcal{H}$:

$$\mathcal{R}(f) \leq \hat{\mathcal{R}}(f) + \sqrt{\frac{\log |\mathcal{H}| + \log 1/\delta}{2n}}$$

Proof steps:

- Bound prob. of $\mathcal{R}(f) - \hat{\mathcal{R}}(f) > \epsilon$ separately for each classifier f
- Combine by union bound $\to \log |\mathcal{H}|$ term

Discussion: union bound



Union bound is "worst case": usually overly pessimistic

Image: https://work.caltech.edu/library/

Can we find a better way to characterize hypothesis classes than simply the number of elements?

Suggested complexity measures:

- covering numbers
- growth function
- VC dimension
- Rademacher complexity

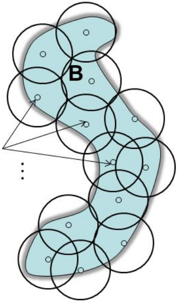
In particular, these work also for infinitely large hypothesis sets.

Definition (Covering)

Let $\mathcal{F}$ be a set of functions. We say $\mathcal{F}$ is ϵ-**covered** by $\mathcal{F}'$ with respect to a norm $\|\cdot\|$:

$$\forall f \in \mathcal{F} \quad \exists f' \in \mathcal{F}' \quad \|f - f'\| < \epsilon$$

$\mathcal{F}'$ is called an ϵ-**cover** of $\mathcal{F}$.



Definition (Covering Number)

Let $\mathcal{F}$ be a set of functions. The ϵ-**covering number**, $\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|)$, is the size of the smallest ϵ-cover of $\mathcal{F}$ with respect to $\|\cdot\|$.

Main idea: $\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|)$ can be small (finite), even if $\mathcal{F}$ is large (infinite). We can use the cover $\mathcal{F}'$ for everything, yet still only make a small error.

Image: Lee Wee Sun. https://slideplayer.com/slide/7277867/

Definition (Growth function)

Let $\mathcal{H} \subset \{f : \mathcal{X} \to \{\pm 1\}\}$ be a set of binary-valued hypotheses. The **growth function** $\Pi_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ of $\mathcal{H}$ is defined as:

$$\Pi_{\mathcal{H}}(n) = \max_{x_1,\ldots,x_n \in \mathcal{X}} \left| \left\{ \left(h(x_1), \ldots, h(x_n) \right) : h \in \mathcal{H} \right\} \right|$$

For any $n \in \mathbb{N}$, $\Pi_{\mathcal{H}}(n)$ is the largest number of different labelings that can be produced with functions in $\mathcal{H}$.

Growth function:

$$\Pi_{\mathcal{H}}(n) = \max_{x_1,\ldots,x_n \in \mathcal{X}} \left| \{ (\, h(x_1), \ldots, h(x_n) \,) : h \in \mathcal{H} \} \right|$$

Examples: growth function

- $\mathcal{H} = \{f_+, f_-\}$, where $f_+(x) = +1$ and $f_-(x) = -1$ (for all $x \in \mathcal{X}$)
  $\rightarrow$ $\Pi_{\mathcal{H}}(n) = 2$ for all $n \geq 1$

Growth function:

$$\Pi_{\mathcal{H}}(n) = \max_{x_1,\ldots,x_n \in \mathcal{X}} \left| \left\{ \left(h(x_1), \ldots, h(x_n) \right) : h \in \mathcal{H} \right\} \right|$$

Examples: growth function

- $\mathcal{H} = \{f_+, f_-\}$, where $f_+(x) = +1$ and $f_-(x) = -1$ (for all $x \in \mathcal{X}$)
  $\rightarrow$ $\Pi_{\mathcal{H}}(n) = 2$ for all $n \geq 1$
- $\mathcal{H} = \{f_1, \ldots, f_T\}$ $\quad\rightarrow\quad$ $\Pi_{\mathcal{H}}(n) \leq \mathbf{min}\{2^n, |\mathcal{H}|\}$

Growth function:

$$\Pi_{\mathcal{H}}(n) = \max_{x_1,\dots,x_n \in \mathcal{X}} \left| \{ (\, h(x_1), \dots, h(x_n) \,) : h \in \mathcal{H} \} \right|$$

Examples: growth function

- $\mathcal{H} = \{f_+, f_-\}$, where $f_+(x) = +1$ and $f_-(x) = -1$ (for all $x \in \mathcal{X}$)
  $\rightarrow$ $\Pi_{\mathcal{H}}(n) = 2$ for all $n \geq 1$
- $\mathcal{H} = \{f_1, \dots, f_T\}$ $\quad\rightarrow\quad$ $\Pi_{\mathcal{H}}(n) \leq \mathbf{min}\{2^n, |\mathcal{H}|\}$
- $\mathcal{H} = \{f : \mathcal{X} \rightarrow \{\pm 1\}\,\}$ (all binary values functions) and $|\mathcal{X}| = \infty$
  $\rightarrow$ $\Pi_{\mathcal{H}}(n) = 2^n$

Growth function:

$$\Pi_{\mathcal{H}}(n) = \max_{x_1,\ldots,x_n\in\mathcal{X}} \left|\{\,(\,h(x_1),\ldots,h(x_n)\,)\,:\,h\in\mathcal{H}\}\right|$$

Examples: growth function

- $\mathcal{H} = \{f_+, f_-\}$, where $f_+(x) = +1$ and $f_-(x) = -1$ (for all $x \in \mathcal{X}$)
  $\rightarrow$ $\Pi_{\mathcal{H}}(n) = 2$ for all $n \geq 1$
- $\mathcal{H} = \{f_1, \ldots, f_T\}$ $\quad\rightarrow\quad$ $\Pi_{\mathcal{H}}(n) \leq \mathbf{min}\{2^n, |\mathcal{H}|\}$
- $\mathcal{H} = \{f : \mathcal{X} \rightarrow \{\pm 1\}\,\}$ (all binary values functions) and $|\mathcal{X}| = \infty$
  $\rightarrow$ $\Pi_{\mathcal{H}}(n) = 2^n$
- $\mathcal{X} = \mathbb{R}^d$, $\mathcal{H} = \{\mathrm{sign}(\langle w, x\rangle + b) : w \in \mathbb{R}^d, b \in \mathbb{R}\,\}$ all linear classifiers
  $\rightarrow$ $\Pi_{\mathcal{H}}(n) = 2^n$ for $n \leq d+1$, but $\Pi_{\mathcal{H}}(n) < 2^n$ for $n > d+1$.

Growth function:

$$\Pi_{\mathcal{H}}(n) = \max_{x_1,\ldots,x_n \in \mathcal{X}} \Big| \{ (\, h(x_1), \ldots, h(x_n) \,) : h \in \mathcal{H} \} \Big|$$

Examples: growth function

- $\mathcal{H} = \{f_+, f_-\}$, where $f_+(x) = +1$ and $f_-(x) = -1$ (for all $x \in \mathcal{X}$)
  $\rightarrow$ $\Pi_{\mathcal{H}}(n) = 2$ for all $n \geq 1$
- $\mathcal{H} = \{f_1, \ldots, f_T\}$ $\quad\rightarrow\quad$ $\Pi_{\mathcal{H}}(n) \leq \mathbf{min}\{2^n, |\mathcal{H}|\}$
- $\mathcal{H} = \{f : \mathcal{X} \rightarrow \{\pm 1\}\,\}$ (all binary values functions) and $|\mathcal{X}| = \infty$
  $\rightarrow$ $\Pi_{\mathcal{H}}(n) = 2^n$
- $\mathcal{X} = \mathbb{R}^d$, $\mathcal{H} = \{\mathrm{sign}(\langle w, x\rangle + b) : w \in \mathbb{R}^d, b \in \mathbb{R}\,\}$ all linear classifiers
  $\rightarrow$ $\Pi_{\mathcal{H}}(n) = 2^n$ for $n \leq d+1$, but $\Pi_{\mathcal{H}}(n) < 2^n$ for $n > d+1$.
- $\mathcal{X} = \mathbb{R}$, $\mathcal{H} = \{\mathrm{sign}(\sin(\omega x)), \quad \omega \in \mathbb{R}\,\}$
  $\rightarrow$ $\Pi_{\mathcal{H}}(n) = 2^n$

Growth Function Generalization Bound

Setup:

- $\ell(y, \bar{y}) = [\![y \neq \bar{y}]\!]$ (0-1 loss)
- $\mathcal{H} \subset \{f : \mathcal{X} \to \{\pm 1\}\}$

For any $\delta > 0$, the following statement holds with probability at least $1 - \delta$ over the training set $\mathcal{D} = \{(x^1, y^1) \ldots, (x^n, y^n)\} \overset{i.i.d.}{\sim} p(x, y)$:

For all $f \in \mathcal{H}$:

$$\mathcal{R}(f) \leq \hat{\mathcal{R}}(f) + \sqrt{\frac{2 \log \Pi_{\mathcal{H}}}{n}} + \sqrt{\frac{\log 1/\delta}{2n}}$$

Growth Function Generalization Bound

Setup:

- $\ell(y, \bar{y}) = [\![y \neq \bar{y}]\!]$ (0-1 loss)
- $\mathcal{H} \subset \{f : \mathcal{X} \to \{\pm 1\}\}$

For any $\delta > 0$, the following statement holds with probability at least $1 - \delta$ over the training set $\mathcal{D} = \{(x^1, y^1) \ldots, (x^n, y^n)\} \overset{i.i.d.}{\sim} p(x, y)$:

For all $f \in \mathcal{H}$:

$$\mathcal{R}(f) \leq \hat{\mathcal{R}}(f) + \sqrt{\frac{2 \log \Pi_{\mathcal{H}}}{n}} + \sqrt{\frac{\log 1/\delta}{2n}}$$

- for $|\mathcal{H}| < \infty$, we (almost) recover the bound for finite hypothesis sets
- bound is vacuous for $\Pi_{\mathcal{H}}(n) = 2^n$, but interesting for $\Pi_{\mathcal{H}}(n) \ll 2^n$

Problem: growth function (for all $n \in \mathbb{N}$) is hard to determine
Easier: at what value does it change from $\Pi_{\mathcal{H}}(n) = 2^n$ to $\Pi_{\mathcal{H}}(n) < 2^n$?

Definition (VC Dimension)

The **VC dimension** of a hypothesis class $\mathcal{H}$, denoted VCdim($\mathcal{H}$), is the maximal value n, such that $\Pi_{\mathcal{H}}(n) = 2^n$. (i.e. $\Pi_{\mathcal{H}}(n+1) < 2^{n+1}$).
If no such value exists, we say that VCdim$(\mathcal{H}) = \infty$.

Problem: growth function (for all $n \in \mathbb{N}$) is hard to determine
Easier: at what value does it change from $\Pi_{\mathcal{H}}(n) = 2^n$ to $\Pi_{\mathcal{H}}(n) < 2^n$?

Definition (VC Dimension)

The **VC dimension** of a hypothesis class $\mathcal{H}$, denoted VCdim($\mathcal{H}$), is the maximal value n, such that $\Pi_{\mathcal{H}}(n) = 2^n$. (i.e. $\Pi_{\mathcal{H}}(n+1) < 2^{n+1}$).
If no such value exists, we say that VCdim($\mathcal{H}$) $= \infty$.

Examples:

- $\mathcal{H} = \{f_+, f_-\}$ for $f_+(x) = +1$ and $f_-(x) = -1$. $\rightarrow$ VCdim($\mathcal{H}$) $= 1$
- $\mathcal{H} = \{f_1, \dots, f_T\}$ $\rightarrow$ VCdim($\mathcal{H}$) $\leq \lfloor \log_2 |\mathcal{H}| \rfloor$
- $\mathcal{H} = \{f : \mathcal{X} \rightarrow \{\pm 1\}\}$ (all binary values functions) and $|\mathcal{X}| = \infty$
  $\rightarrow$ VCdim($\mathcal{H}$) $= \infty$
- $\mathcal{X} = \mathbb{R}^d$, $\mathcal{H} = \{\mathrm{sign}(\langle w, x\rangle + b) : w \in \mathbb{R}^d, b \in \mathbb{R}\}$ (linear classifiers)
  $\rightarrow$ VCdim($\mathcal{H}$) $= d + 1$
- $\mathcal{X} = \mathbb{R}$, $\mathcal{H} = \{\mathrm{sign}(\sin(\omega x)), \quad \omega \in \mathbb{R}\}$
  $\rightarrow$ VCdim($\mathcal{H}$) $= \infty$

Definition (VC Dimension)

The **VC dimension** of a hypothesis class $\mathcal{H}$, denoted VCdim($\mathcal{H}$), is the maximal value n, such that $\Pi_{\mathcal{H}}(n) = 2^n$, or ∞ if no such value exists.

Lemma (Sauer's Lemma)

For any $\mathcal{H}$ with VCdim($\mathcal{H}$) $< \infty$, for any m: $\Pi_{\mathcal{H}}(n) \leq \sum_{i=0}^{\mathit{VCdim}(\mathcal{H})} \binom{n}{i}$.

Consequence:

- up to $n =$ VCdim($\mathcal{H}$), growth function grows **exponentially**
- for $n \geq$ VCdim($\mathcal{H}$)$+1$, growth function grows only **polynomially**:

$$\Pi_{\mathcal{H}}(n) \leq (en/d)^d. \qquad \text{(proof: textbook)}$$

- complexity term $\sqrt{\frac{2\log\Pi_{\mathcal{H}}(n)}{n}}$ starts decreasing for $n >$ VCdim($\mathcal{H}$)

VC-Dimension Generalization Bound

Setup:

- $\ell(y, \bar{y}) = [\![y \neq \bar{y}]\!]$ (0-1 loss)
- $\mathcal{H} \subset \{f : \mathcal{X} \to \{\pm 1\}\}$

For any $\delta > 0$, the following statement holds with probability at least $1 - \delta$ over the training set $\mathcal{D} = \{(x^1, y^1) \dots, (x^n, y^n)\} \overset{i.i.d.}{\sim} p(x, y)$:

For all $f \in \mathcal{H}$:

$$\mathcal{R}(f) \leq \hat{\mathcal{R}}(f) + \sqrt{\frac{2d \log \frac{en}{d}}{n}} + \sqrt{\frac{\log 1/\delta}{2n}}$$

where $d = \text{VCdim}(\mathcal{H})$

Crucial quantity: $\frac{d}{n}$. Non-trivial bound only for $n > d$.

More examples: VC dimension (from the literature)

1) **threshold functions**, $\mathcal{H} = \{h_\theta(x) = \operatorname{sign}(x - \theta), \text{for } \theta \in \mathbb{R}\}$.
   VCdim$(\mathcal{H}) = 1$

More examples: VC dimension (from the literature)

1) **threshold functions**, $\mathcal{H} = \{h_\theta(x) = \text{sign}(x - \theta), \text{for } \theta \in \mathbb{R}\}$.
VCdim$(\mathcal{H}) = 1$

- $n = 1$, $\mathcal{D} = \{x_1\}$
  - for $\theta < x_1 : h_\theta(x_1) = 1$,
  - for $\theta \geq x_1$, $h_\theta(x_1) = 0$.

$\Pi_{\mathcal{H}}(1) = 2 = 2^1$.

- $\mathcal{D} = \{x_1, x_2\}$, w.l.o.g. $x_1 < x_2$
  - for $\theta < x_1 :$ $(\, h_\theta(x_1), h_\theta(x_2)\,) = (1, 1)$
  - for $c_1 \leq \theta < c_2 :$ $h_\theta(c_1), h_\theta(c_2) = 0, 1$
  - for $\theta \geq c_2 :$ $h_\theta(c_1), h_\theta(c_2) = 0, 0$
  - there is no $h \in \mathcal{H}$ with $h_\theta(c_1), h_\theta(c_2) = 1, 0$.

$\Pi_{\mathcal{H}}(2) = 3 < 2^2$, no matter what c_1, c_2 are (except $c_1 = c_2$).

$\mathcal{H}$ can arbitrarily label a set of size 1, but no set of size 2
$\Rightarrow$ VCdim$(\mathcal{H}) = 1$

More examples: VC dimension (from the literature)

1) **threshold functions**, $\mathcal{H} = \{h_\theta(x) = \mathrm{sign}(x - \theta), \text{for } \theta \in \mathbb{R}\}$.
   $\mathrm{VCdim}(\mathcal{H}) = 1$
2) **polynomial classifiers**,

$$\mathcal{H} = \{h(x) = \mathrm{sign}\, f(x), \text{for } f \text{ any polynomial of degree } k \text{ in } \mathbb{R}^d\}.$$

$$\mathrm{VCdim}(\mathcal{H}) = \sum_{i=0}^{k} \binom{d+1}{i}$$

More examples: VC dimension (from the literature)

1) **threshold functions**, $\mathcal{H} = \{h_\theta(x) = \operatorname{sign}(x-\theta), \text{for } \theta \in \mathbb{R}\}$.
   VCdim$(\mathcal{H}) = 1$
2) **polynomial classifiers**,

   $$\mathcal{H} = \{h(x) = \operatorname{sign} f(x), \text{for } f \text{ any polynomial of degree } k \text{ in } \mathbb{R}^d\}.$$

   VCdim$(\mathcal{H}) = \sum_{i=0}^{k} \binom{d+1}{i}$
3) **boosting**: base set, $\mathcal{F}$, of weak classifiers with VCdim D.

   $$\mathcal{H} = \Big\{ f(x) = \sum_{t=1}^{T} \alpha_t g_t(x), \text{ for } g_1, \dots, g_T \in \mathcal{F} \text{ and } \alpha_1, \dots, \alpha_T \in \mathbb{R} \Big\}$$

   VCdim$(\mathcal{H}) \leq T(D+1) \cdot (3 \log(T(D+1)) + 2)$

More examples: VC dimension (from the literature)

1) **threshold functions**, $\mathcal{H} = \{h_\theta(x) = \operatorname{sign}(x-\theta), \text{for } \theta \in \mathbb{R}\}$.
   $\text{VCdim}(\mathcal{H}) = 1$
2) **polynomial classifiers**,

$$\mathcal{H} = \{h(x) = \operatorname{sign} f(x), \text{for } f \text{ any polynomial of degree } k \text{ in } \mathbb{R}^d\}.$$

$$\text{VCdim}(\mathcal{H}) = \sum_{i=0}^{k} \binom{d+1}{i}$$

3) **boosting**: base set, $\mathcal{F}$, of weak classifiers with VCdim D.

$$\mathcal{H} = \Big\{ f(x) = \sum_{t=1}^{T} \alpha_t g_t(x), \text{ for } g_1, \dots, g_T \in \mathcal{F} \text{ and } \alpha_1, \dots, \alpha_T \in \mathbb{R} \Big\}$$

$$\text{VCdim}(\mathcal{H}) \leq T(D+1) \cdot (3\log(T(D+1)) + 2)$$

4) **neural networks** with binary activation functions,
   $\text{VCdim}(\mathcal{H}) \leq O(d \log d)$ where d is number of network weights
5) **neural networks** with binary and linear activation functions,
   $\text{VCdim}(\mathcal{H}) \leq O(d^2)$ where d is number of network weights

From classical to modern generalization bounds

Generalization bounds so far: with probability at least $1-\delta$:

$$\forall f \in \mathcal{H}: \quad \mathcal{R}(f) \leq \hat{\mathcal{R}}(f) + B(\mathcal{H}, n, \delta)$$

Observation:

- $B(\mathcal{H}, n, \delta)$ is data-independent
- data distribution does not show up anywhere
  $\rightarrow$ holds for "easy" as well as "hard" learning problems

Recently, more interest in distribution-dependent bounds.

- $\mathcal{Z}$: input set (later: $\mathcal{Z} = \mathcal{X}$ or $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$)
- $p(z)$: probability distribution over $\mathcal{Z}$
- $\mathcal{F} \subseteq \{f : \mathcal{Z} \to \mathbb{R}\}$: set of real-valued functions

Definition

Let $\mathcal{F} = \{f : \mathcal{Z} \to \mathbb{R}\}$ be a set of real-valued functions and $\mathcal{D}_m = \{z_1, \dots, z_m\}$ a finite set. The **empirical Rademacher complexity** of $\mathcal{F}$ with respect to $\mathcal{D}_m$ is defined as

$$\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}) = \mathop{\mathbb{E}}_{\sigma}\left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m}\sum_{i=1}^{m} \sigma_i f(z_i)\right)\right]$$

where $\sigma_1, \dots, \sigma_m$ are independent binary random variables with $p(+1) = p(-1) = \frac{1}{2}$ (called **Rademacher variables**).

Intuition: think of σ_i as random noise. The $\sup$ measures how well the function can correlate to arbitrary values (=memorize random noise).

Note: $\hat{\mathfrak{R}}_{\mathcal{D}_m}$ is data-dependent, it depends on $\mathcal{D}_m$.

Example

Let $\mathcal{F} = \{f\}$ (a single function). Then, for any m,

$$\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}) = \mathop{\mathbb{E}}_{\sigma} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i) \right) = \frac{1}{m} \sum_{i=1}^{m} \mathop{\mathbb{E}}_{\sigma}[\sigma_i] f(z_i) = 0$$

Example

Let $\mathcal{F} = \{f : \mathcal{Z} \to [-B, B]\}$ all bounded functions. Then, when there are no duplicates in $\mathcal{D}$,

$$\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}) = \mathop{\mathbb{E}}_{\sigma} \mathbf{sup}_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i) \right) \overset{f(z_i) = B\sigma_i}{=} \mathop{\mathbb{E}}_{\sigma} \frac{1}{m} \sum_{i=1}^{m} B = \mathop{\mathbb{E}}_{\sigma} B = B$$

(same argument would work, e.g., for piecewise linear functions)

Example

Let $\mathcal{F} = \{f_1, \ldots, f_K\}$ with $f_i : \mathcal{X} \to [-B, B]$ for $i = 1, \ldots, K$ (finitely many bounded function). Then

$$\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}) \leq B\sqrt{\frac{2\log K}{m}}$$

Proof: textbook

Example

Let $\mathcal{F} = \{f = w^\top z : \mathbb{R}^d \to \mathbb{R}\}$ with $\|w\| \leq B$ all *linear* functions with bounded slope. If $m > d$, then $z_1, \ldots, z_m$ are linearly dependent and $\sup$ can't fit all possible signs $\quad\to\quad$ $\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F})$ will decrease with m.

(we'll prove a more rigorous statement later)

Definition

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

$$\mathfrak{R}_m(\mathcal{F}) = \mathop{\mathbb{E}}_{\mathcal{D}_m \sim p^{\otimes m}} \left[\, \hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}) \, \right]$$

Note: $\mathfrak{R}_m$ is a distribution-dependent quantity (w.r.t. p).

In some cases, more convenient to compute than the empirical one.

Slightly more general notation than before:

- hypothesis set $\mathcal{H} \subset \{\mathcal{X} \to \mathbb{R}\}$ (can be real-valued)
- loss $\ell : \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \to \mathbb{R}$,   e.g. $\ell(x, y, h) = \mathbf{max}\{0, 1 - yh(x)\}$,
- $\mathcal{R}^{\ell}(h) = \mathbb{E}_{(x,y)\sim p}\, \ell(x, y, h)$, $\hat{\mathcal{R}}^{\ell}(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(x_i, y_i, h)$

Slightly more general notation than before:

- hypothesis set $\mathcal{H} \subset \{\mathcal{X} \to \mathbb{R}\}$ (can be real-valued)
- loss $\ell : \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \to \mathbb{R}$, e.g. $\ell(x, y, h) = \mathbf{max}\{0, 1 - yh(x)\}$,
- $\mathcal{R}^\ell(h) = \mathbb{E}_{(x,y)\sim p}\, \ell(x, y, h)$, $\hat{\mathcal{R}}^\ell(h) = \frac{1}{m}\sum_{i=1}^m \ell(x_i, y_i, h)$

Theorem (Rademacher-based generalization bound)

Let $\ell(x, y, h) \le c$ be a bounded loss function and set
$\mathcal{F} = \{\ell \circ h : h \in \mathcal{H}\} \quad = \{\ell(x, y, h(x)) : h \in \mathcal{H}\} \subset \{f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}\}$

Then, with prob. at least $1 - \delta$ over $\mathcal{D}_m \overset{i.i.d.}{\sim} p$, it holds for all $h \in \mathcal{H}$:

$$\mathcal{R}^\ell(h) \le \hat{\mathcal{R}}^\ell(h) + 2\mathfrak{R}_m(\mathcal{F}) + c\sqrt{\frac{\log(1/\delta)}{2m}}.$$

Also, with prob. at least $1 - \delta$, it holds for all $h \in \mathcal{H}$:

$$\mathcal{R}^\ell(h) \le \hat{\mathcal{R}}^\ell(h) + 2\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}) + 3c\sqrt{\frac{2\log(4/\delta)}{m}}.$$

Proof. blackboard/notes

Useful properties:

Lemma

For $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ let $\mathcal{F}' := \{f + f_0 : f \in \mathcal{F}\}$ be a translated version for some $f_0 : \mathcal{X} \to \mathbb{R}$. Then, for any m,

$$\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}') = \hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F})$$

Lemma

For $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ let $\mathcal{F}' := \{\lambda f : f \in \mathcal{F}\}$ be scaled by a constant $\lambda \in \mathbb{R}$. Then, for any m,

$$\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}') = \lambda \hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F})$$

Lemma

For $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ and $\phi : \mathbb{R} \to \mathbb{R}$ let $\mathcal{F}' := \{\phi \circ f : f \in \mathcal{F}\}$. If ϕ is L-Lipschitz continuous, i.e. $|\phi(t) - \phi(t')| \leq L|t - t'|$, then for any m,

$$\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}') \leq L \cdot \hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F})$$

Lemma

Let $\mathcal{Z}$ be an inner-product space (e.g. $\mathbb{R}^d$ with $\langle \cdot, \cdot \rangle$). Let $\mathcal{F} = \{f = \langle w, z \rangle : \mathcal{X} \to \mathbb{R}\}$ be the set of linear functions with $\|w\| \leq B$. Then, for any $\mathcal{D}_m = \{z_1, \dots, z_m\}$,

$$\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}) \leq \frac{B}{m}\sqrt{\sum_i \|z_i\|^2}$$

Proof: blackboard/notes

Lemma

Let $\mathcal{F} = \{f = \langle w, z \rangle : \mathcal{X} \to \mathbb{R}\}$ be linear functions with $\|w\| \leq B$ and let p be such that $\Pr\{\|z\| < R\} = 1$ Then

$$\mathfrak{R}_m(\mathcal{F}) \leq BR\sqrt{\frac{1}{m}}$$

Proof: $\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}) \leq \frac{B}{m}\sqrt{mR^2}$ with prob. 1, so $\mathbb{E}_{\mathcal{D}}\,\hat{\mathfrak{R}} \leq \frac{B}{m}\sqrt{mR^2}$, too.

Reminder: (soft-margin) support vector machine (SVM):

$$\min_{w} \; \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_i \max\{0, 1 - y_i\langle w, x_i\rangle\}$$

Reminder: (soft-margin) support vector machine (SVM):

$$\min_w \ \frac{\lambda}{2}\|w\|^2 + \frac{1}{m}\sum_i \max\{0, 1 - y_i\langle w, x_i\rangle\}$$

Example: SVM "radius/margin" bound

Let $\ell(x,y;w) := \max\{0, 1 - y\langle w, x\rangle\}$ be the *hinge loss*. Let p be a distribution on $\mathbb{R}^d \times \mathcal{Y}$ such that $\Pr\{\, \|x\| \leq R\,\} = 1$ and let $\mathcal{H} = \{w : \|w\| \leq B\}$.

Then, with prob. at least $1-\delta$ over $\mathcal{D}_m \overset{i.i.d.}{\sim} p$ the following inequality holds for all $w \in \mathcal{H}$:

$$\mathop{\mathbb{E}}_{(x,y)\sim p} [\![\operatorname{sign}\langle w, x\rangle \neq y]\!] \leq \frac{1}{m}\sum_{i=1}^m \max\{0, 1 - y^i\langle w, x^i\rangle\} + \frac{2BR}{\sqrt{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

Properties:

- complexity terms decrease with rate $O(\sqrt{\frac{1}{m}})$
- short $\|w\|$ is better than long $\|w\|$
- dimensionality of x does not show up, no curse of dimensionality!

Proof sketch:

- $\|x\| \leq R$ (with probability 1)
- "ramp loss": $\ell(x, y, h) = \mathbf{min}\{\ \mathbf{max}\{0, 1 - y\langle w, x\rangle\}, 1\ \}\ \in [0, 1]$
- $\mathcal{H} = \{h(x) = \langle w, x\rangle : \|w\| \leq B\}, \quad \mathcal{F} = \{\ell \circ h,\ h \in \mathcal{H}\}$

With prob. $1-\delta$: $\quad \forall h \in \mathcal{H} : \mathcal{R}^{\ell}(h) \leq \hat{\mathcal{R}}^{\ell}(h) + 2\mathfrak{R}_m(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2m}}$

- ℓ is 1-Lipschitz, i.e. for $\mathcal{F} = \{\ell \circ h : h \in \mathcal{H}\}$:

$$\mathfrak{R}_m(\mathcal{F}) \overset{\text{1-Lip.}}{\leq} \mathfrak{R}_m(\mathcal{H}) \overset{\text{Lemma}}{\leq} BR\sqrt{\frac{1}{m}}$$

- ℓ is upper bounds to 0/1 error and lower bound to hinge loss

$$\Pr\{h(x) \neq y\} \leq \mathcal{R}^{\ell}(h) \qquad \hat{\mathcal{R}}^{\ell}(h) \leq \frac{1}{m}\sum_{i=1}^{m} \mathbf{max}\{0, 1 - y_i\langle w, x_i\rangle\}$$

With prob. $1-\delta$ for every $w \in \mathcal{H}$:

$$\Pr\{\operatorname{sign}\langle w, x\rangle \neq y\} \leq \frac{1}{m}\sum_{i=1}^{m} \mathbf{max}\{0, 1 - y_i\langle w, x_i\rangle\} + \frac{2RB}{\sqrt{m}} + \sqrt{\frac{\log(1/\delta)}{2m}}$$

Theorem (Connections to other complexity measures)

Let $\mathcal{H} = \{h : \mathcal{X} \to \{\pm 1\}\}$ *be a hypothesis class. Then*

$$\hat{\mathfrak{R}}_m(\mathcal{H}) \leq \sqrt{\frac{2 \log |\mathcal{H}|}{m}} \qquad \text{if } |\mathcal{H}| \text{ is finite,}$$

$$\hat{\mathfrak{R}}_m(\mathcal{H}) \leq \sqrt{\frac{2 \log \Pi_{\mathcal{H}}(m)}{m}} \qquad \text{where } \Pi_{\mathcal{H}}(m) \text{ is the growth function,}$$

$$\hat{\mathfrak{R}}_m(\mathcal{H}) \leq \sqrt{\frac{2d \log m}{m}} \qquad \text{where } d = \text{VCdim}(\mathcal{H}).$$

Theorem (Connections to covering numbers)

Let $\mathcal{H} \subset \{\mathcal{X} \to [-1, 1]\}$ *and* $\mathcal{D} \overset{i.i.d.}{\sim} p(x, y)$ *with* $|\mathcal{D}| = m$. *Then*

$$\hat{\mathfrak{R}}_m(\mathcal{H}) \leq \mathbf{inf}_\alpha \left[\alpha + \sqrt{\frac{\mathcal{N}(\alpha, \mathcal{H}|_{\mathcal{D}}, \|\cdot\|_{L_1})}{m}}\ \right]$$

where $\mathcal{N}$ *are covering numbers of the set of values that* $\mathcal{H}$ *assigns to* $\mathcal{D}$.