Statistical Machine Learning

Christoph Lampert

IST AUSTRIA

Institute of Science and Technology

Spring Semester 2015/2016 // Lecture 8

The Holy Grail of Statistical Machine Learning

Image: http://typemoon.wikia.com/

The Holy Grail of Statistical Machine Learning

Inferring the test error
from the training error

Generalization Bound

For every $f \in \mathcal{H}$ it holds:

$$\underbrace{\mathop{\mathbb{E}}_{(x,y)} \ell(y, f(x))}_{\text{generalization loss}} \quad \leq \quad \underbrace{\frac{1}{n} \sum_i \ell(y_i, f(x_i))}_{\text{training loss}} \quad + \quad ?$$

Image: http://typemoon.wikia.com/

$\mathcal{X}$: input set, $\mathcal{Y} = \{0, 1\}$, $\ell(y, y') = [\![y \neq y']\!]$, $p(x, y)$: data distribution
$\mathcal{H}$: hypothesis class of finite VC-dimension, $\mathrm{VC}(\mathcal{H})$

Theorem (VC bound – realizable case)

If labels are deterministic with a labeling function $f \in \mathcal{H}$, then the following inequality holds with probability at least $1 - \delta$ (over $\mathcal{D}_m \overset{i.i.d.}{\sim} p$) for all $h \in \mathcal{H}$ with $\hat{\mathcal{R}}_{\mathcal{D}_m}(h) = 0$:

$$\mathcal{R}_p(h) \leq \frac{1}{m}\left(\log VC(\mathcal{H}) + \log \frac{1}{\delta} \right)$$

Theorem (VC bound – general case)

For arbitrary $p(x, y)$ the following inequality holds with probability at least $1 - \delta$ (over $\mathcal{D}_m \overset{i.i.d.}{\sim} p$) for all $h \in \mathcal{H}$:

$$\mathcal{R}_p(h) \leq \hat{\mathcal{R}}_m(h) + \sqrt{\frac{8\, VC(\mathcal{H}) \log \frac{2em}{VC(\mathcal{H})} + 8 \log \frac{4}{\delta}}{m}}$$

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

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

Theorem (SVM radius/margin bound)

Let $\ell(x, y; w) := \mathbf{max}\{0, 1 - y\langle w, x\rangle\}$ be the hinge loss. Let p be a distribution on $\mathcal{X} \times \mathcal{Y}$ 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 $w \in \mathcal{H}$:

$$\Pr\{\text{sign}\langle w, x\rangle \neq y\} \leq \frac{1}{m}\sum_{i=1}^{m} \ell(x_i, y_i, w) + \frac{2BR}{\sqrt{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

Almost perfect justification of SVMs:

- uniform in w, i.e. holds even for minimizer of r.h.s.
- B is a upper bound on $\|w\| \rightarrow$ small $\|w\|$ are most promising
- dimensionality of x does not show up! also holds for kernels

Reminder: hard-margin SVM:

$$\mathbf{min}_w \|w\|^2 \text{ subject to } y_i\langle w, x_i\rangle \geq 1 \text{ for } i = 1, \ldots, n.$$

Theorem (SVM hard margin bound)

Let p be a distribution on $\mathcal{X} \times \mathcal{Y}$ with $\Pr\{\, \|x\| \leq R \,\} = 1$ for which there exists w^* with $\Pr\{\, y\langle w^*, x\rangle \leq 1\} = 1$ (linearly separable with a margin). Let w_S be the solution to the hard-margin SVM problem. Then, with prob. at least $1 - \delta$ over $\mathcal{D}_m \overset{i.i.d.}{\sim} p(x, y)$ the following inequality holds:

$$\Pr\{\text{sign}\langle w_S, x\rangle \neq y\} \;\leq\; \frac{2R\|w^*\|}{\sqrt{m}} + \sqrt{\frac{\log(1/\delta)}{2m}}$$

Also, with prob. $1 - \delta$, the following inequality holds

$$\Pr\{\text{sign}\langle w_S, x\rangle \neq y\} \;\leq\; \frac{4R\|w_S\|}{\sqrt{m}} + \sqrt{\frac{\log(\frac{4\log_2(\|w_S\|)}{\delta})}{2m}}$$

(stronger versions are possible where r.h.s scales like $\frac{1}{m}$ instead of $\frac{1}{\sqrt{m}}$)

Towards Modern Generalization Bounds

We need more modern measure of complexity than VC dimension:

- $\mathcal{Z}$: 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 $\mathcal{D}_m = \{z_1, \ldots, z_m\}$ 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 $\sigma_1, \ldots, \sigma_m$ are independent binary random variables with $p(+1) = p(-1) = \frac{1}{2}$ (called **Rademacher variables**).

Note: $\hat{\mathfrak{R}}_{\mathcal{D}_m}$ is a data-dependent complexity measure (it depends on $\mathcal{D}_m$)

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

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 $\mathbf{sup}$ 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)

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 a Hilbert space (e.g. $\mathbb{R}^d$, or given by a kernel). Let $\mathcal{F} = \{f = \langle w, z\rangle : \mathcal{Z} \to \mathbb{R}\}$ be linear functions with $\|w\| \leq B$. 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 $\mathcal{D}_m = \{z_1, \ldots, z_m\}$

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

If $\langle \cdot, \cdot \rangle$ is given by a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, then

$$\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}) = \frac{B}{m}\sqrt{\mathrm{trace}(K)}.$$

where $K \in \mathbb{R}^{m \times m}$ is the kernel matrix, $k_{ij} = k(z_i, z_j) = \langle z_i, z_j \rangle$.

Proof: blackboard/notes

Definition

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

$$\mathfrak{R}_m(\mathcal{F}) = \mathop{\mathbb{E}}_{\mathcal{D}_m}[\ \hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F})\ ]$$

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

In some cases one can derive convenient upper bounds:

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: use that $\hat{\mathfrak{R}}_{\mathcal{D}_m}(\mathcal{F}) \leq \frac{B}{m}\sqrt{mR^2}$ with prob. 1.

Example: kernels of the form $e^{-d(x,x')}$ (e.g. Gaussian) fulfill $\|z\|^2 \leq 1$.

Slightly more general notation:

- loss function $\ell : \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \to \mathbb{R}$, e.g. $\ell(x, y, h) = [\![h(x) \neq y]\!]$, $\ell(x, y, h) = (h(x) - y)^2$, $\ell(x, y, h) = \mathbf{max}\{0, 1 - yh(x)\}$, etc.
- $\mathcal{R}_p^\ell(h) = \mathbb{E}_{(x,y)\sim p}\, \ell(x, y, h)$, $\hat{\mathcal{R}}_{\mathcal{D}_m}^\ell(h) = \frac{1}{m} \sum_{i=1}^m \ell(x_i, y_i, h)$

Theorem

Rademacher-based generalization bound Let $\ell(x, y, h) \leq c$ be a bounded loss function and for a hypothesis set $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ let $\mathcal{F} = \{\ell \circ h : h \in \mathcal{H}\}$. Then, with prob. at least $1 - \delta$, it holds for all $h \in \mathcal{H}$:

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

Proof. blackboard/notes □

Example (Hard-margin SVM)

- $\|x\| \leq R$ with probability 1
- $\mathcal{H} = \{h(x) = \langle w, x\rangle : \|w\| \leq B\}$ for B that we'll specify later
- ramp-loss: $\ell(x, y, h) = \mathbf{min}\{\ \mathbf{max}\{0, 1 - y\langle w, x\rangle\},\ 1\ \} \in [0, 1]$
- ℓ is an upper bounds to the 0/1 error

$$\Pr\{h(x) \neq y\} = \mathcal{R}_p^{0/1}(h) \leq \mathcal{R}_p^{\ell}(h)$$

- hard-margin h fulfills $y_i\langle w, x_i\rangle \geq 1$ for $i = 1, \ldots, m$: $\quad \hat{\mathcal{R}}^{\ell}_{\mathcal{D}_m}(h) = 0$
- ℓ is 1-Lipschitz, i.e. for $\mathcal{F} = \{\ell \circ h : h \in \mathcal{H}\}$:

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

- $B = \|w^*\|$ ensures that hard-margin SVM $h_S \in \mathcal{H}$.

With prob. $1 - \delta$: $\qquad \Pr\{h_S(x) \neq y\} \leq \dfrac{2R\|w^*\|}{\sqrt{m}} + \sqrt{\dfrac{\log(1/\delta)}{2m}}$

Example (Soft-margin SVM)

- $\|x\| \le R$ with probability 1
- $\mathcal{H} = \{h(x) = \langle w, x\rangle : \|w\| \le B\}$ for fixed B
- hinge loss: $\ell(x, y, h) = \mathbf{max}\{0, 1 - y\langle w, x\rangle\} \ \in [0, 1 + BR]$
- ℓ is 1-Lipschitz, i.e. for $\mathcal{F} = \{\ell \circ h : h \in \mathcal{H}\}$:

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

- ℓ is an upper bounds to the 0/1 error

$$\Pr\{h(x) \ne y\} = \mathcal{R}_p^{0/1}(h) \le \mathcal{R}_p^{\ell}(h)$$

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

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