Machine Learning Theory 2023 Lecture 5

Tim van Erven

Download these slides now from elo.mastermath.nl!

Focus on binary classification:

- ▶ Review
- ▶ Remaining proof:
  growth function controls uniform convergence

Uniform Convergence Upper Bound with VC-Dimension

> **Theorem**
>
> *Consider binary classification. Suppose* $\mathrm{VCdim}(\mathcal{H}) \leq v < \infty$. *Then there exists an absolute constant $C > 0$ such that*
>
> $$\sup_{h \in \mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon \qquad \text{with probability} \geq 1 - \delta,$$
>
> *whenever*
>
> $$m \geq C \frac{v \ln(1/\epsilon) + \ln(1/\delta) + 1}{\epsilon^2}.$$

Proof Approach

Growth function: $\tau_{\mathcal{H}}(m) = \max_{|\mathcal{C}|=m} |\mathcal{H}_{\mathcal{C}}|$

▶ Interpretation: How many truly different hypotheses are there when we only observe m inputs $\mathcal{C} = \{x_1, \ldots, x_m\}$?

Proof Approach

Growth function: $\tau_{\mathcal{H}}(m) = \max\limits_{|\mathcal{C}|=m} |\mathcal{H}_{\mathcal{C}}|$

▶ Interpretation: How many truly different hypotheses are there when we only observe m inputs $\mathcal{C} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$?

Part I: Growth function controls uniform convergence:

$$\sup_{h \in \mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \leq c\sqrt{\frac{\ln \tau_{\mathcal{H}}(m)}{m}} + c\sqrt{\frac{\ln(4/\delta)}{m}} \qquad \text{with probability} \geq 1 - \delta$$

Part II: VC-dimension controls growth function (Sauer's Lemma):

$$\ln \tau_{\mathcal{H}}(m) \leq v \ln\left(\frac{em}{v}\right) \qquad \text{for } m > v.$$

▶ Finish: combine Parts I and II, and find lower bound on m s.t. $\sup_{h \in \mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon$.

Proof Part I:
Growth Function Controls Uniform Convergence

Part I: Proof Outline

Lemma (Two-sided Bound)

Consider binary classification. Then there exists an absolute constant $c > 0$ such that, for any $\delta \in (0, 1]$,

$$\sup_{h \in \mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \leq c\sqrt{\frac{\ln \tau_{\mathcal{H}}(m)}{m}} + c\sqrt{\frac{\ln(4/\delta)}{m}} \qquad \text{w.p.} \geq 1 - \delta.$$

Part I: Proof Outline

Lemma (Two-sided Bound)

Consider binary classification. Then there exists an absolute constant $c > 0$ such that, for any $\delta \in (0, 1]$,

$$\sup_{h \in \mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \leq c\sqrt{\frac{\ln \tau_{\mathcal{H}}(m)}{m}} + c\sqrt{\frac{\ln(4/\delta)}{m}} \qquad w.p. \geq 1 - \delta.$$

Note: could measure loss in binary classification differently.
Sufficient to show:

Lemma (One-sided Bound)

For any loss function $\ell(h, \boldsymbol{X}, Y) = \tilde{\ell}(h(\boldsymbol{X}), Y)$ with range $[0, 1]$:

$$\sup_{h \in \mathcal{H}} \{L_S(h) - L_{\mathcal{D}}(h)\} \leq c\sqrt{\frac{\ln \tau_{\mathcal{H}}(m)}{m}} + c\sqrt{\frac{\ln(2/\delta)}{m}} \qquad w.p. \geq 1 - \delta.$$

Let $z = \sup\limits_{h \in \mathcal{H}} \{L_S(h) - L_D(h)\}$

$z' = \sup\limits_{h \in \mathcal{H}} \{L_D(h) - L_S(h)\}$

Applying one-sided bound with $\ell' = 1 - \ell$
controls z', because

$$L'_S(h) - L'_D(h) = \frac{1}{m} \sum_{i=1}^{m} (1 - \ell(h, x_i, y_i))$$
$$- \mathbb{E}[1 - \ell(h, x, y)]$$
$$= \mathbb{E}[\ell(h, x, y)] - \frac{1}{m} \sum_{i=1}^{m} \ell(h, x_i, y_i)$$
$$= L_D(h) - L_S(h)$$

Then

$$\sup_{h \in \mathcal{H}} |L_S(h) - L_D(h)| = \max \{z, z'\}$$

$$\leq c \sqrt{\frac{\ln \tau_{\mathcal{H}}(m)}{m}} + c \sqrt{\frac{\ln(4/\delta)}{m}} \quad w.p. \geq 1-\delta$$

by one-sided bounds for z and z' with $\delta' = \frac{\delta}{2}$

+ union bound.

Approach for One-Sided Bound

Lemma (One-sided Bound)

For any loss function $\ell(h, \boldsymbol{X}, Y) = \tilde{\ell}(h(\boldsymbol{X}), Y)$ with range $[0, 1]$:

$$\sup_{h \in \mathcal{H}} \{L_S(h) - L_{\mathcal{D}}(h)\} \leq c\sqrt{\frac{\ln \tau_{\mathcal{H}}(m)}{m}} + c\sqrt{\frac{\ln(2/\delta)}{m}} \qquad w.p. \geq 1 - \delta.$$

Remark:
▶ Book first derives suboptimal dependence on δ in Chapter 6
▶ I am taking a shortcut through Chapters 6, 26 and 28

Approach for One-Sided Bound

Lemma (One-sided Bound)

For any loss function $\ell(h, \boldsymbol{X}, Y) = \tilde{\ell}(h(\boldsymbol{X}), Y)$ with range $[0,1]$:

$$\sup_{h \in \mathcal{H}} \{L_S(h) - L_{\mathcal{D}}(h)\} \leq c\sqrt{\frac{\ln \tau_{\mathcal{H}}(m)}{m}} + c\sqrt{\frac{\ln(2/\delta)}{m}} \qquad w.p. \geq 1 - \delta.$$

Proof consists of 3 steps:

1. **Concentration:** Abbreviate $Z = \sup_{h \in \mathcal{H}} \{L_S(h) - L_{\mathcal{D}}(h)\}$. Then, for any loss function $\ell(h, \boldsymbol{X}, Y)$ with range $[0,1]$,

$$Z \leq \underset{S}{\mathbb{E}}[Z] + c\sqrt{\frac{\ln(2/\delta)}{m}} \qquad \text{w.p.} \geq 1 - \delta.$$

Approach for One-Sided Bound

Lemma (One-sided Bound)

For any loss function $\ell(h, \boldsymbol{X}, Y) = \tilde{\ell}(h(\boldsymbol{X}), Y)$ with range $[0,1]$:

$$\sup_{h \in \mathcal{H}} \{L_S(h) - L_{\mathcal{D}}(h)\} \leq c\sqrt{\frac{\ln \tau_{\mathcal{H}}(m)}{m}} + c\sqrt{\frac{\ln(2/\delta)}{m}} \qquad \textit{w.p.} \geq 1 - \delta.$$

Proof consists of 3 steps:

1. **Concentration:** Abbreviate $Z = \sup_{h \in \mathcal{H}} \{L_S(h) - L_{\mathcal{D}}(h)\}$. Then, for any loss function $\ell(h, \boldsymbol{X}, Y)$ with range $[0, 1]$,

$$Z \leq \mathop{\mathbb{E}}_{S}[Z] + c\sqrt{\frac{\ln(2/\delta)}{m}} \qquad \text{w.p.} \geq 1 - \delta.$$

2. **Symmetrization:** For any loss function:
$$\mathbb{E}[Z] \leq 2\,\mathbb{E}[\mathcal{R}(\ell, \mathcal{H}, S)].$$

Approach for One-Sided Bound

Lemma (One-sided Bound)

For any loss function $\ell(h, \boldsymbol{X}, Y) = \tilde{\ell}(h(\boldsymbol{X}), Y)$ with range $[0, 1]$:

$$\sup_{h \in \mathcal{H}} \{L_S(h) - L_{\mathcal{D}}(h)\} \leq c\sqrt{\frac{\ln \tau_{\mathcal{H}}(m)}{m}} + c\sqrt{\frac{\ln(2/\delta)}{m}} \qquad w.p. \geq 1 - \delta.$$

Proof consists of 3 steps:

1. **Concentration:** Abbreviate $Z = \sup_{h \in \mathcal{H}} \{L_S(h) - L_{\mathcal{D}}(h)\}$. Then, for any loss function $\ell(h, \boldsymbol{X}, Y)$ with range $[0, 1]$,

$$Z \leq \mathbb{E}_S[Z] + c\sqrt{\frac{\ln(2/\delta)}{m}} \qquad \text{w.p.} \geq 1 - \delta.$$

2. **Symmetrization:** For any loss function:

$$\mathbb{E}[Z] \leq 2\,\mathbb{E}[\mathcal{R}(\ell, \mathcal{H}, S)].$$

3. For any loss $\ell(h, \boldsymbol{X}, Y) = \tilde{\ell}(h(\boldsymbol{X}), Y)$ with range $[0, 1]$:

$$\mathcal{R}(\ell, \mathcal{H}, S) \leq \sqrt{\frac{2\ln|\mathcal{H}_S|}{m}} \leq \sqrt{\frac{2\ln \tau_{\mathcal{H}}(m)}{m}} \qquad \text{for all } S.$$

To show: loss $\ell(h, x, y) \in [0, 1]$

$$Z = \sup_{h \in H} L_S(h) - L_D(h)$$

$$Z \le \mathbb{E}[Z] + c\sqrt{\frac{\ln(2/\delta)}{m}} \qquad w.p. \ge 1 - \delta$$

Proof:

$\qquad\qquad\qquad$ ↗ some complicated function

$Z = f(A_1, \ldots, A_m)$ for $A_i = (x_i, y_i)$

Bounded differences property:

* If change $A_i \to A_i'$, then

$\qquad Z$ changes by at most $\frac{1}{m}$

(because $L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, x_i, y_i)$

$\qquad\qquad$ changes by at most $1/m$)

McDiarmid's Inequality

Suppose $A_1, \ldots, A_m$ are independent random variables, and $f : A^m \to \mathbb{R}$ satisfies for all i

$$\sup_{\substack{a_1, \ldots, a_m \\ a_i'}} |f(a_1, \ldots, a_m) - f(a_1, \ldots, a_{i-1}, a_i', a_{i+1}, \ldots, a_m)| \leq b.$$

Then, with probability $\geq 1 - \delta$,

$$|f(A_1, \ldots, A_m) - \mathbb{E}[f(A_1, \ldots, A_m)]| \leq b\sqrt{\frac{m}{2}\ln\left(\frac{2}{\delta}\right)}$$

z satisfies this with $b = 1/m$:

$$|z - \mathbb{E}[z]| \leq \sqrt{\frac{\ln\left(2/\delta\right)}{2m}} \quad w.p. \geq 1 - \delta$$

$$\leq c\sqrt{\frac{\ln\left(2/\delta\right)}{m}} \quad \text{for } c \geq \frac{1}{\sqrt{2}} \qquad \square$$

Rademacher Complexity

How much can the losses of $h \in \mathcal{H}$ on S
correlate with random errors?

Rademacher random variables: Let $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_m) \in \{-1, +1\}^m$ be i.i.d. with $\Pr(\sigma_i = -1) = \Pr(\sigma_i = +1) = 1/2$.

Rademacher complexity:

$$\mathcal{R}(\ell, \mathcal{H}, S) = \frac{1}{m} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^{m} \sigma_i \ell(h, \boldsymbol{X}_i, Y_i) \right]$$

▶ Interpret $\sum_{i=1}^{m} \sigma_i \ell(h, \boldsymbol{X}_i, Y_i)$ as correlation of losses with random errors

Step 2: Symmetrization

$\mathcal{R}(\ell, \mathcal{H}, S) = \frac{1}{m} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^{m} \sigma_i \ell(h, \boldsymbol{X}_i, Y_i) \right]$

Lemma

$$\mathbb{E}_S \left[\sup_{h \in \mathcal{H}} \{ L_S(h) - L_{\mathcal{D}}(h) \} \right] \ \leq \ 2 \, \mathbb{E}_S [\mathcal{R}(\ell, \mathcal{H}, S)]$$

Step 2: Symmetrization

$$\mathcal{R}(\ell, \mathcal{H}, S) = \frac{1}{m} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^{m} \sigma_i \ell(h, \boldsymbol{X}_i, Y_i) \right]$$

Lemma

$$\mathbb{E}_{S} [\sup_{h \in \mathcal{H}} \{L_S(h) - L_{\mathcal{D}}(h)\}] \ \leq \ 2 \mathbb{E}_{S} [\mathcal{R}(\ell, \mathcal{H}, S)]$$

Amazing because:

▶ $\sup_{h \in \mathcal{H}} \{L_S(h) - L_{\mathcal{D}}(h)\}$ may be large for very unlikely S

▶ But Rademacher complexity $\mathcal{R}(\ell, \mathcal{H}, S)$ is small for all S!

Consequence:

▶ Can measure complexity of $\mathcal{H}$ conditional on S

▶ So only restriction of $\mathcal{H}$ to inputs $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ in S matters!

To show: $\mathcal{R}(\ell, \mathcal{H}, S) = \frac{1}{m} \mathbb{E}_{\sigma} \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^{m} \sigma_i \, \ell(h, x_i, y_i) \right]$

$\mathbb{E}_{S} \left\{ \sup_{h \in \mathcal{H}} L_S(h) - L_D(h) \right\} \leq 2 \, \mathbb{E}_{S} \left[\mathcal{R}(\ell, \mathcal{H}, S) \right]$

<u>Proof</u>: Let $S' = \binom{y_1'}{x_1'}, \ldots, \binom{y_m'}{x_m}$ be independent sample.

$\mathbb{E}_{S} \left[\sup_{h} L_S(h) - L_D(h) \right] = \mathbb{E}_{S} \left[\sup_{h} L_S(h) - \mathbb{E}_{S'} \left\{ L_{S'}(h) \right\} \right]$

$= \mathbb{E}_{S} \left[\sup_{h} \mathbb{E}_{S'} \left\{ L_S(h) - L_{S'}(h) \right\} \right] \leq \mathbb{E}_{S,S'} \left[\sup_{h} L_S(h) - L_{S'}(h) \right]$

$= \frac{1}{m} \mathbb{E}_{S,S'} \left[\sup_{h} \sum_{i=1}^{m} \left\{ \ell(h, x_i, y_i) - \ell(h, x_i', y_i') \right\} \right]$

Homogenize the two samples

N.B. If we swap any $\binom{y_i}{x_i}$ and $\binom{y_i'}{x_i'}$ between S and S', then distribution does not change.

Hence, for any $\sigma_i \in \{-1, +1\}$,

$$\frac{1}{m} \mathbb{E}_{S,S'} \left[\sup_h \sum_{i=1}^m \{ \ell(h, x_i, y_i) - \ell(h, x_i', y_i') \} \right]$$

$$= \frac{1}{m} \mathbb{E}_{S,S'} \left[\sup_h \sum_{i=1}^m \sigma_i \{ \ell(h, x_i, y_i) - \ell(h, x_i', y_i') \} \right]$$

$$= \frac{1}{m} \mathbb{E}_{\sigma, S, S'} \left[\sup_h \sum_{i=1}^m \sigma_i \{ \ell(h, x_i, y_i) - \ell(h, x_i', y_i') \} \right]$$

$$\leq \frac{1}{m} \mathbb{E}_{\sigma, S, S'} \left[\sup_h \sum_i \sigma_i \ell(h, x_i, y_i) + \sup_h \sum_i -\sigma_i \ell(h, x_i', y_i') \right]$$

(using that $-\sigma$ has same distribution as σ)

$$= \frac{2}{m} \mathbb{E}_{\sigma, S} \left[\sup_h \sum_i \sigma_i \ell(h, x_i, y_i) \right] = 2 \mathbb{E}_S \{ \mathcal{R}(\ell, \mathcal{H}, S) \} \quad \square$$

Step 3: Bound the Rademacher Complexity

Lemma

For any loss function $\ell(h, \boldsymbol{X}, Y) = \tilde{\ell}(h(\boldsymbol{X}), Y)$ with range $[0,1]$ and any sample S:

$$\mathcal{R}(\ell, \mathcal{H}, S) \leq \sqrt{\frac{2\ln |\mathcal{H}_S|}{m}} \leq \sqrt{\frac{2\ln \tau_{\mathcal{H}}(m)}{m}}.$$

$$\boxed{\text{Step 3}}$$

To show: $\ell(h, x, y) = \tilde{\ell}(h(x), y) \in [0, 1]$

For any S: $\mathcal{R}(\ell, \mathcal{H}, S) \leq \eta \sqrt{\dfrac{2 \ln |\mathcal{H}_S|}{m}}$

<u>Proof</u>: $m \cdot \mathcal{R}(\ell, \mathcal{H}, S) = \underset{\sigma}{\mathbb{E}} \left[\underset{h \in \mathcal{H}}{\sup} \sum_{i=1}^{m} \sigma_i \, \tilde{\ell}(h(x_i), y_i) \right]$

$$= \underset{\sigma}{\mathbb{E}} \left[\underset{h \in \mathcal{H}_S}{\max} \sum_{i=1}^{m} \sigma_i \, \hat{\ell}(h(x_i), y_i) \right]$$

Let $z_i(h) = \sigma_i \, \hat{\ell}(h(x_i), y_i) \in [-1, +1]$. Then $\underset{\sigma}{\mathbb{E}} \{ z_i(h) \} = 0$

<u>Hoeffding's Lemma (B.4 in Shai²):</u>

Suppose z takes values in $[a, b]$ and $\mathbb{E}\{z\} = 0$. Then

$$\mathbb{E}\{ e^{\lambda z} \} \leq e^{\lambda^2 (b-a)^2 / 8} \qquad \text{for any } \lambda > 0.$$

$$m \cdot R(\ell, H, S) = \underset{\sigma}{\mathbb{E}} \left[\max_{h \in H_S} \sum_{i=1}^{m} z_i(h) \right]$$

$$= \frac{1}{\lambda} \underset{\sigma}{\mathbb{E}} \left[\ln \max_{h \in H_S} e^{\sum_{i=1}^{m} \lambda z_i(h)} \right] \quad \text{for any } \lambda > 0$$

$$\leq \frac{1}{\lambda} \mathbb{E} \left[\ln \sum_{h \in H_S} e^{\sum_{i=1}^{m} \lambda z_i(h)} \right]$$

(Jensen's inequality)

$$\leq \frac{1}{\lambda} \ln \left(\mathbb{E} \left[\sum_{h \in H_S} e^{\sum_{i=1}^{m} \lambda z_i(h)} \right] \right)$$

$$= \frac{1}{\lambda} \ln \left(\sum_{h \in H_S} \prod_{i=1}^{m} \mathbb{E} \left[e^{\lambda z_i(h)} \right] \right)$$

(Hoeffding's Lemma)

$$\leq \frac{1}{\lambda} \ln \left(\sum_{h \in H_S} \prod_{i=1}^{m} e^{\lambda^2/2} \right) = \frac{1}{\lambda} \ln |H_S| + \lambda \frac{m}{2}$$

Take $\lambda = \sqrt{\frac{\ln |H_S|}{m/2}}$: $= \sqrt{2m \ln |H_S|}$

$$R(\ell, H, S) \leq \sqrt{\frac{2 \ln |H_S|}{m}} \qquad \square$$

Back to the Big Picture

Part I: Growth function controls uniform convergence:

$$\sup_{h \in \mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \leq c\sqrt{\frac{\ln \tau_{\mathcal{H}}(m)}{m}} + c\sqrt{\frac{\ln(4/\delta)}{m}} \qquad \text{with probability} \geq 1 - \delta$$

Part II: VC-dimension controls growth function (Sauer's Lemma):

$$\ln \tau_{\mathcal{H}}(m) \leq v \ln\left(\frac{em}{v}\right) \qquad \text{for } m > v.$$

Back to the Big Picture

Part I: Growth function controls uniform convergence:

$$\sup_{h \in \mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \leq c\sqrt{\frac{\ln \tau_{\mathcal{H}}(m)}{m}} + c\sqrt{\frac{\ln(4/\delta)}{m}} \qquad \text{with probability} \geq 1 - \delta$$

Part II: VC-dimension controls growth function (Sauer's Lemma):

$$\ln \tau_{\mathcal{H}}(m) \leq v \ln\left(\frac{em}{v}\right) \qquad \text{for } m > v.$$

For $m > v$:

$$\sup_{h \in \mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \leq c\sqrt{\frac{v \ln\left(\frac{em}{v}\right)}{m}} + c\sqrt{\frac{\ln(4/\delta)}{m}} \qquad \text{with probability} \geq 1 - \delta$$

▶ Remaining: find lower bound on m s.t. bound is at most ϵ.