Machine Learning Theory 2024
Lecture 5

Tim van Erven

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*

$$\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon \qquad \textit{with probability} \geq 1 - \delta,$$

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 $\mathcal{C} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$?

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 $\mathcal{C} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$?

Part I: Growth function controls uniform convergence:

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

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. $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon$.

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 $\delta \in (0, 1]$,

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

Part I: Proof Outline

Lemma (Two-sided Bound)

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

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

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 $[0, 1]$:

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

One-sided Bound $\Rightarrow$ Two-sided Bound

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

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

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

Then

$$\sup_{h \in \mathcal{H}} |L_D(h) - L_S(h)| = \max\{Z, Z'\}$$
$$\leq c\sqrt{\frac{\ln \tau_{\mathcal{H}}(m)}{m}} + c\sqrt{\frac{\ln(4/\delta)}{m}} \quad \text{w.p.} \geq 1-\delta$$

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* $[0, 1]$*:*

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

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* $[0,1]$*:*

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

Proof consists of 3 steps:

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

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

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* $[0, 1]$*:*

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

Proof consists of 3 steps:

1. **Concentration:** Abbreviate $Z = \sup_{h \in \mathcal{H}} \{L_{\mathcal{D}}(h) - L_S(h)\}$. 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* $[0, 1]$*:*

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

Proof consists of 3 steps:

1. **Concentration:** Abbreviate $Z = \sup_{h \in \mathcal{H}} \{L_{\mathcal{D}}(h) - L_S(h)\}$. 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 $[0, 1]$:
$$\mathcal{R}(\ell, \mathcal{H}, S) \leq \sqrt{\frac{2 \ln |\mathcal{H}_S|}{m}} \leq \sqrt{\frac{2 \ln \tau_{\mathcal{H}}(m)}{m}} \qquad \text{for all } S.$$

Step 1: Concentration

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

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

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

Proof: some complicated function

$$Z = f(A_1, \dots, A_m) \quad \text{for } A_i = (x_i, y_i)$$

Bounded differences property:

* If change $A_i \to A_i'$, then Z changes by at most $\frac{1}{m}$

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

McDiarmid's Inequality

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

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

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

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

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

$$|Z - \mathbb{E}[Z]| \le \sqrt{\frac{\ln(2/\delta)}{2m}} \quad \text{w.p.} \geq 1 - \delta$$

$$\le c \sqrt{\frac{\ln(2/\delta)}{m}} \quad \text{for } c \geq \frac{1}{\sqrt{2}}$$

$\square$

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:

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

- 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

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

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

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

Amazing because:

- $\sup_{h \in \mathcal{H}} \{L_{\mathcal{D}}(h) - L_S(h)\}$ 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!

Step 2: "Symmetrization"

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

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

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

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

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

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

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 their distribution does not change.

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

$$\frac{1}{m} \mathop{\mathbb{E}}_{S,S'} \Big[\sup_{h\in\mathcal{H}} \sum_{i=1}^{m} \{ \ell(h,x_i',y_i') - \ell(h,x_i,y_i) \} \Big]$$

$$= \frac{1}{m} \mathop{\mathbb{E}}_{S,S'} \Big[\sup_{h\in\mathcal{H}} \sum_{i=1}^{m} \sigma_i \{ \ell(h,x_i',y_i') - \ell(h,x_i,y_i) \} \Big]$$

$$= \frac{1}{m} \mathop{\mathbb{E}}_{\sigma,S,S'} \Big[\sup_{h\in\mathcal{H}} \sum_{i=1}^{m} \sigma_i \{ \ell(h,x_i',y_i') - \ell(h,x_i,y_i) \} \Big]$$

$$\le \frac{1}{m} \mathop{\mathbb{E}}_{\sigma,S,S'} \Big[\sup_{h\in\mathcal{H}} \sum_{i=1}^{m} \sigma_i \ell(h,x_i',y_i') + \sup_{h\in\mathcal{H}} \sum_{i=1}^{m} -\sigma_i \ell(h,x_i,y_i) \Big]$$

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

$$= \frac{2}{m} \mathop{\mathbb{E}}_{\sigma,S} \Big[\sup_{h\in\mathcal{H}} \sum_{i=1}^{m} \sigma_i \ell(h,x_i,y_i) \Big] = 2 \mathop{\mathbb{E}}_{S} \big[R(\ell,\mathcal{H},S) \big] \quad \square$$

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*:*

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

Step 3

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

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

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

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

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

Hoeffding's Lemma (B.7 in Shai²):

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

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

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

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

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

(Jensen's inequality)

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

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

(Hoeffding's Lemma)

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

Take $\lambda = \sqrt{\frac{\ln |\mathcal{H}_S|}{m/2}}$: $\quad = \sqrt{2m \ln |\mathcal{H}_S|}$

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

$\square$

Back to the Big Picture

Part I: Growth function controls uniform convergence:

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

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:

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

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 $m > v$:

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

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