Statistical Foundations of Learning

Debarghya Ghoshdastidar

School of Computation, Information and Technology

Technical University of Munich

Infinite hypothesis classes and uniform convergence

Recap

- Goal: Find bound on the generalisation error of ERM solution $\widehat{h}$

- If hypothesis class $\mathcal{H} \subset \{\pm 1\}^{\mathcal{X}}$ is finite:
  - Uniform convergence bound (for 0-1 loss)

$$\max_{h \in \mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \leq \sqrt{\frac{\ln(|\mathcal{H}|) + \ln(\frac{2}{\delta})}{2m}} \quad \text{with probability } 1 - \delta$$

  - Generalisation error bound

$$L_{\mathcal{D}}(\widehat{h}) \leq L_{\mathcal{D}}(\mathcal{H}) + \sqrt{\frac{2\ln(|\mathcal{H}|) + 2\ln(\frac{2}{\delta})}{m}} \qquad \text{with probability } 1 - \delta$$

From finite to infinite $\mathcal{H}$

- Uniform convergence bound when $\mathcal{H}$ is infinite
  - With uniform convergence, we can prove generalisation error bound for ERM as before
- Challenge: Previous bound depends on $|\mathcal{H}|$
  Which proof step led to $|\mathcal{H}|$ in bound?
  - Union bound over all $h \in \mathcal{H}$
- Do we need to consider all $h \in \mathcal{H}$?
  - No. For m training samples, there can be at most 2^m distinct predictions

Outline

- Growth function
  - How many distinct predictors can $\mathcal{H}$ provide on any m samples?
- Uniform convergence bound for infinite $\mathcal{H}$
  - Growth function replaces $|\mathcal{H}|$ in bound
- Proof of uniform convergence (main ideas; not needed for exam)

Growth function

- Consider sequence $C = (x_1, \ldots, x_m) \in \mathcal{X}^m$ $\qquad$ C only has features, not labels

- Restriction of hypothesis class $\mathcal{H} \subseteq \{-1, +1\}^{\mathcal{X}}$ to C

$$\mathcal{H}_{|C} = \big\{ (h(x_1), \ldots, h(x_m)) \ : \ h \in \mathcal{H} \big\}$$

  - Set of all possible labelling of the m data points in C using $\mathcal{H}$

- Growth function of $\mathcal{H}$

$$\tau_{\mathcal{H}}(m) = \max_{C \subseteq \mathcal{X} : |C| = m} \big| \mathcal{H}_{|C} \big|$$

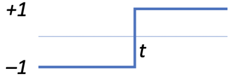
  - Maximum number of possible binary labelling for any m instances in $\mathcal{X}$ using $\mathcal{H}$

- Verify $\tau_{\mathcal{H}}(m) \leq \min \{ |\mathcal{H}|, 2^m \}$

Example: Threshold functions

- A threshold function $h_t : \mathbb{R} \to \{\pm 1\}$ has one parameter $t \in \mathcal{X}$

$$h_t(x) = \begin{cases} -1 & \text{if } x \leq t \\ +1 & \text{if } x > t \end{cases}$$



- Let $\mathcal{H}_{thr} = \{h_t(\cdot) \ : \ t \in \mathbb{R}\} \subset \{\pm 1\}^{\mathbb{R}}$
- Compute $\tau_{\mathcal{H}_{thr}}(1)$
  - Let $C = \{x_1\}$
  - We either have $h_t(x_1) = +1$ if $t \geq x_1$ or $h_t(x_1) = -1$ if $t < x_1$
  - $\mathcal{H}_{thr|C} = \{(+1), (-1)\}$ for every C of size 1 $\implies \tau_{\mathcal{H}_{thr}}(1) = 2$

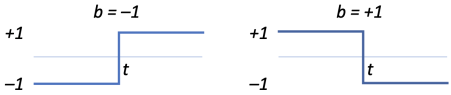
Example: Threshold functions

- Compute $\tau_{\mathcal{H}_{thr}}(2)$
  - Let $C = \{x_1, x_2\}$ with $x_1 < x_2$
  - $\mathcal{H}_{thr|C} = \Big\{ \underbrace{(-1,-1)}_{\text{if } t \geq x_2}, \underbrace{(-1,+1)}_{\text{if } x_1 \leq t < x_2}, \underbrace{(+1,+1)}_{\text{if } t < x_1} \Big\}$ $(+1,-1)$ cannot happen since $x_1 < x_2$
  - So $\tau_{\mathcal{H}_{thr}}(2) = 3$
- Does above imply that $|\mathcal{H}_{thr|C}| = 3$ for all C of size 2?
  - No. We could have $C = \{x_1, x_1\}$
- Use previous arguments to verify that $\tau_{\mathcal{H}_{thr}}(m) = m + 1$

Example: Decision stumps

- A one-dimensional decision stump has two parameters $t \in \mathcal{X}$ and $b \in \{\pm 1\}$

$$h_{t,b}(x) = \begin{cases} b & \text{if } x \leq t \\ -b & \text{if } x > t \end{cases}$$



- Let $\mathcal{H}_{ds\text{-}1} = \left\{ h_{t,b}(\cdot) \ : \ t \in \mathbb{R}, b \in \{\pm 1\} \right\} \subset \{\pm 1\}^{\mathbb{R}}$

- Compute $\tau_{\mathcal{H}_{ds\text{-}1}}(m)$

Example: Decision stumps

- Answer $\tau_{\mathcal{H}_{ds\text{-}1}}(m) = 2m$
- Take $C = \{x_1, x_2, \ldots, x_m\}$ with $x_1 < x_2 < \ldots < x_m$
- For $b = -1$, $m+1$ possible labellings $(-1, \ldots, -1), (-1, \ldots, -1, +1), \ldots, (+1, \ldots, +1)$
- For $b = 1$, signs reverse $(+1, \ldots, +1), (+1, \ldots, +1, -1), \ldots, (-1, \ldots, -1)$
- We have $(+1, \ldots, +1)$ and $(-1, \ldots, -1)$ in both cases (need to count only once)
- Hence $2m$ possible functions

Uniform convergence for infinite $\mathcal{H}$

Theorem UC.1 (Uniform convergence of $L_S(\cdot)$ for infinite $\mathcal{H}$)

Let $\epsilon \in (0,1)$ and $m > \frac{2\ln 4}{\epsilon^2}$. Let $\mathcal{H} \subset \{\pm 1\}^{\mathcal{X}}$ and we measure risk with respect to 0-1 loss.

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left(\sup_{h\in\mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\right) \leq \tau_{\mathcal{H}}(2m)\cdot 4e^{-m\epsilon^2/8}$$

Equivalent statement: *Let $\delta \in (0,1)$. With probability $\geq 1-\delta$,*

$$\sup_{h\in\mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)| \leq \sqrt{\frac{8\ln\left(\tau_{\mathcal{H}}(2m)\right) + 8\ln\left(\frac{4}{\delta}\right)}{m}}$$

Generalisation error for ERM

- Use previous result to verify that for ERM solution $\widehat{h}$

$$L_{\mathcal{D}}(\widehat{h}) \leq L_{\mathcal{D}}(\mathcal{H}) + 2\sqrt{\frac{8\ln(\tau_{\mathcal{H}}(2m)) + 8\ln(\frac{4}{\delta})}{m}} \qquad \text{with probability } 1-\delta$$

- Consider ERM over $\mathcal{H}_{ds\text{-}1}$. Use above result to derive generalisation error bound

$$L_{\mathcal{D}}(\widehat{h}) \leq L_{\mathcal{D}}(\mathcal{H}) + 2\sqrt{\frac{8\ln(4m) + 8\ln(\frac{4}{\delta})}{m}} \qquad \text{with probability } 1-\delta$$

  - Set $\delta = 0.01$ and large $m = 10^7$
  - There is 99% chance of having $L_{\mathcal{D}}(\widehat{h}) < L_{\mathcal{D}}(\mathcal{H}) + 0.01$ ... ERM finds nearly best solution

Generalisation error for ERM over other $\mathcal{H}$

- For arbitrary infinite $\mathcal{H}$, recall that $\tau_{\mathcal{H}}(2m) \leq 2^{2m}$
- Using this bound for growth function

$$L_{\mathcal{D}}(\widehat{h}) \leq L_{\mathcal{D}}(\mathcal{H}) + \underbrace{2\sqrt{\frac{16m + 8\ln(\frac{4}{\delta})}{m}}}_{\text{larger than } 1} \qquad \text{with probability } 1 - \delta$$

  - Bound is meaningless since $L_{\mathcal{D}}(\widehat{h}) \leq 1$ trivially
  - Next topic: We will derive non-trivial bound on $\tau_{\mathcal{H}}$ in terms of VC dimension

Proof Step 1: Symmetrisation – idea

- Need to show $\sup_{h \in \mathcal{H}} |L_S(h) - L_{\mathcal{D}}(h)|$ is not large
- Recall: Main challenge in the proof is union bound over all $\mathcal{H}$ (due to sup)
  - Cannot avoid this, but use a trick to reduce number of terms
- How many possible values of $|L_S(h) - L_{\mathcal{D}}(h)|$ can we have?
  - $L_{\mathcal{D}}(\cdot)$ can take at most $|\mathcal{H}|$ values (unique value for every $h \in \mathcal{H}$)
  - $L_S(\cdot)$ can take only $m+1$ values in set $\left\{0, \frac{1}{m}, \frac{2}{m}, \ldots, 1\right\}$
- Idea: "Replace" $L_{\mathcal{D}}(\cdot)$ by empirical risk $L_{S'}(h)$ over an independent set S' of size m

Proof Step 1: Symmetrisation – result

Lemma UC.2 (Symmetrisation by introducing independent copy of S)

Let $S, S' \sim \mathcal{D}^m$ be two independent training sets, each of size m. For $m\epsilon^2 > 2\ln 4$,

$$\mathbb{P}_S\left(\sup_{h\in\mathcal{H}} |L_S(h) - L_\mathcal{D}(h)| > \epsilon\right) \leq 2\mathbb{P}_{S,S'}\left(\sup_{h\in\mathcal{H}} |L_S(h) - L_{S'}(h)| > \frac{\epsilon}{2}\right)$$

- Intuition: If $L_S(h)$ is close to $L_\mathcal{D}(h)$, then
  - $L_{S'}(h)$ is also likely to be close to $L_\mathcal{D}(h)$ (since S' has same distribution as S)
  - $L_S(\cdot)$ and $L_{S'}(h)$ are likely to be close to each other (both close to $L_\mathcal{D}(h)$)
- Advantage of this step: $|L_S(\cdot) - L_{S'}(\cdot)|$ takes only $m+1$ distinct values for all $h \in \mathcal{H}$

Proof Step 2: Swapping permutations – idea

- Need to show $\sup_{h \in \mathcal{H}} |L_S(h) - L_{S'}(h)|$ is not large

- Naive idea (does not work, but informative):
    - $\sup_{h \in \mathcal{H}} |L_S(h) - L_{S'}(h)| = \max_{\mathbf{h} \in \mathcal{H}_{|S \cup S'}} |L_S(\mathbf{h}) - L_{S'}(\mathbf{h})|$
    - Can bound probability for every $\mathbf{h}$, and apply union bound over $\mathcal{H}_{|S \cup S'}$
    - Union bound leads to multiplicative factor of $|\mathcal{H}_{|S \cup S'}| \leq \tau_{\mathcal{H}}(2m)$

- Why doesn't this work?
    - $\mathcal{H}_{|S \cup S'}$ is random, depends on S, S' (can apply union bound only when union is fixed)

Proof Step 2: Swapping permutations – idea

- Idea that works: Can apply above if we condition on S, S' … makes $\mathcal{H}_{|S \cup S'}$ fixed
  - Introduce another source of randomness (Rademacher symmetrisation)

- Swapping permutation:
  - Let (x_i, y_i) be the i^{th} instance in S, and (x'_i, y'_i) be i^{th} instance in S'
  - Define $Y_{(\sigma_1,\ldots,\sigma_m)} = \dfrac{1}{m}\displaystyle\sum_{i=1}^{m} \sigma_i \cdot \left(\mathbf{1}\left\{h(x_i) \neq y_i\right\} - \mathbf{1}\left\{h(x'_i) \neq y'_i\right\}\right)$ … for $\sigma_i \in \{\pm 1\}$
  - Note $Y_{(1,\ldots,1)} = L_S(h) - L_{S'}(h)$
  - $\sigma_i = -1$ means we swap i^{th} instances in S and S'

Proof Step 2: Swapping permutations – idea

- $Y_{(\sigma_1,\ldots,\sigma_m)}$ has same distribution as $Y_{(1,\ldots,1)}$

$$\begin{aligned}\mathbb{P}_{S,S'}\left(\sup_{h\in\mathcal{H}}|L_S(h)-L_{S'}(h)|>\frac{\epsilon}{2}\right) &= \mathbb{P}_{S,S'}\left(\sup_{h\in\mathcal{H}}|Y_{(1,\ldots,1)}|>\frac{\epsilon}{2}\right)\\ &= \frac{1}{2^m}\sum_{\sigma_1,\ldots,\sigma_m\in\{\pm1\}}\mathbb{P}_{S,S'}\left(\sup_{h\in\mathcal{H}}|Y_{(\sigma_1,\ldots,\sigma_m)}|>\frac{\epsilon}{2}\right)\end{aligned}$$

- Random swapping / Rademacher symmetrisation
  - Average can be viewed as an expectation
  - $\sigma_1,\ldots,\sigma_m$ i.i.d., each takes values ±1 with equal probability (Rademacher variables)

Proof Step 2: Swapping permutations – result

Lemma UC.3 (Symmetrisation by introducing Rademacher variables)

Let $\sigma = (\sigma_1, \ldots, \sigma_m)$ *where* $\sigma_1, \ldots, \sigma_m$ *i.i.d. Rademacher variable*

$$\mathbb{P}_{S,S'}\left(\sup_{h\in\mathcal{H}} |L_S(h) - L_{S'}(h)| > \frac{\epsilon}{2}\right) = \mathbb{P}_{S,S',\sigma}\left(\sup_{h\in\mathcal{H}} |Y_\sigma| > \frac{\epsilon}{2}\right)$$
$$= \mathbb{E}_{S,S'}\left[\mathbb{P}_{\sigma|S,S'}\left(\sup_{h\in\mathcal{H}} |Y_\sigma| > \frac{\epsilon}{2}\right)\right]$$

- Advantage of this step:
  Probability is conditioned over S, S'. Can apply the union bound over $\mathcal{H}_{|S\cup S'}$

Proof Step 3: Union bound

Can apply union bound since we condition over S, S' (that is, S, S' kept fixed)

$$\begin{aligned}
\mathbb{P}_{\sigma|S,S'}\left(\sup_{h\in\mathcal{H}} |Y_\sigma| > \frac{\epsilon}{2}\right) &= \mathbb{P}_{\sigma|S,S'}\left(\max_{\mathbf{h}\in\mathcal{H}_{|S\cup S'}} |Y_\sigma| > \frac{\epsilon}{2}\right) && Y_\sigma \text{ is function of } S, S', h \\
&\leq \sum_{\mathbf{h}\in\mathcal{H}_{|S\cup S'}} \mathbb{P}_{\sigma|S,S'}\left(|Y_\sigma(\mathbf{h})| > \frac{\epsilon}{2}\right) && \text{union bound} \\
&\leq |\mathcal{H}_{|S\cup S'}| \cdot 2e^{-m\epsilon^2/8} && \text{Hoeffding's inequality} \\
&\leq \tau_{\mathcal{H}}(2m) \cdot 2e^{-m\epsilon^2/8}
\end{aligned}$$

Bound does not depend on S, S'. Does not change after taking $\mathbb{E}_{S,S'}[\,\cdot\,]$