Statistical Foundations of Learning

Debarghya Ghoshdastidar

School of Computation, Information and Technology

Technical University of Munich

Algorithmic Stability

Context: Stability vs PAC learnability

- Previously: Learnable classes (finite VC classes) have low generalisation error
  - Generalisation bounds mainly depend on $\mathrm{VCdim}(\mathcal{H})$
  - We can have different algorithms to learn from same $\mathcal{H}$
  - $\mathcal{H}_{lin} = \{\mathrm{sign}(w^\top x + b) : w \in \mathbb{R}^p, b \in \mathbb{R}\}$ can be learned using ERM over $\mathcal{H}_{lin}$ / SVM / soft SVM / Linear Discriminant Analysis, $\ldots$
- This lecture: Stable learner generalises well (low generalisation error)
  - *Stability* is property of learner whereas *learnability* feature of hypothesis class
  - Stability based bounds take into account additional properties of data — margins for linear classification (will use for soft SVM)

Algorithmic stability

- Informal definition:
  Learning algorithm is stable if output does not change significantly if we change only one input (training example)

- Outline: Will cover few possible mathematical formulations
  - More types of stability: (don't have to read)
    O. Bousquet, A. Elisseeff. Stability and Generalization. *Journal of Machine Learning Research* 2, pp. 499-526, 2002

- Outline: Generalisation error bound for stable algorithms

On-average-replace-one stability

Notations

- S^i
  - Consider training sample $S \sim \mathcal{D}^m$
  - Replace (x_i, y_i) with an independent example $(x', y') \sim \mathcal{D}$
- $i \sim \text{Unif}(m)$
  - Pick one of the m examples in S uniformly at random
- β_m^{rep}
  - $\beta^{rep} : \mathbb{N} \to \mathbb{R}$, and $\beta_m^{rep} = \beta^{rep}(m)$

On-average-replace-one stability

- Given learner $\mathcal{A}$, loss function ℓ and $\beta^{rep} : \mathbb{N} \to \mathbb{R}$

- Learner $\mathcal{A}$ is on-average-replace-one stable

  - with rate β^{rep} with respect to ℓ

  - if for every sample size m and every distribution $\mathcal{D}$

$$\mathbb{E}_{S\sim\mathcal{D}^m,(x',y')\sim\mathcal{D},i\sim\mathrm{Unif}(m)}\left[\ell\big(\mathcal{A}_{S^i}(x_i),y_i\big) - \ell\big(\mathcal{A}_S(x_i),y_i\big)\right] \leq \beta^{rep}_m$$

- Difference of two losses:

  - 2^{nd} term: Training set contains (x_i, y_i) and tested also using (x_i, y_i)

  - 1^{st} term: (x_i, y_i) not used for training, but for testing

Generalisation from on-average-replace-one stability

Lemma Stab.1 (Expected generalisation error)

For any learner $\mathcal{A}$ and loss ℓ,

$$\mathbb{E}_{S\sim\mathcal{D}^m}\big[L_{\mathcal{D}}(\mathcal{A}_S) - L_S(\mathcal{A}_S)\big] = \mathbb{E}_{S\sim\mathcal{D}^m,(x',y')\sim\mathcal{D},i\sim\mathrm{Unif}(m)}\Big[\ell\big(\mathcal{A}_{S^i}(x_i),y_i\big) - \ell\big(\mathcal{A}_S(x_i),y_i\big)\Big]$$

If $\mathcal{A}$ is an on-average-replace-one stable learner with rate β, then

$$\mathbb{E}_{S\sim\mathcal{D}^m}\big[L_{\mathcal{D}}(\mathcal{A}_S)\big] \quad\leq\quad \mathbb{E}_{S\sim\mathcal{D}^m}\big[L_S(\mathcal{A}_S)\big] + \beta_m$$

Proof

- $S, (x', y') \sim \mathcal{D}^{m+1}$ is an iid sequence of $m+1$ examples

- Use any m of them for training, and the other one for testing

- No effect on the expected true risk

$$\begin{aligned}
\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(\mathcal{A}_S)\right] &= \mathbb{E}_{S,(x',y')\sim\mathcal{D}^{m+1}}\left[\ell\big(\mathcal{A}_S(x'), y'\big)\right] && \ldots\text{definition} \\
&= \mathbb{E}_{S\sim\mathcal{D}^m,(x',y')\sim\mathcal{D},i\sim\mathrm{Unif}(m)}\left[\ell\big(\mathcal{A}_{S^i}(x_i), y_i\big)\right] && \ldots\text{above equivalence}
\end{aligned}$$

- From definition of empirical risk

$$\begin{aligned}
\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_S(\mathcal{A}_S)\right] &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{S\sim\mathcal{D}^m}\left[\ell\big(\mathcal{A}_S(x_i), y_i\big)\right] && \ldots\text{from definition} \\
&= \mathbb{E}_{S\sim\mathcal{D}^m,i\sim\mathrm{Unif}(m)}\left[\ell\big(\mathcal{A}_S(x_i), y_i\big)\right] && \ldots\text{average} = \mathbb{E}_{i\sim\mathrm{Unif}(m)}[\cdot]
\end{aligned}$$

Generalisation from on-average-replace-one stability

Theorem Stab.2 (Expected generalisation error for stable ERM)

Assume

- $\mathcal{A} =$ *ERM for some hypothesis class* $\mathcal{H}$
- $\mathcal{A}$ *is on-average-replace-one stable with rate* β

Then

$$\mathbb{E}_{S\sim\mathcal{D}^m}\big[L_{\mathcal{D}}(\mathcal{A}_S)\big] \quad \leq \quad L_{\mathcal{D}}(\mathcal{H}) + \beta_m$$

Above implies PAC learnability if $\beta_m \to 0$ as $m \to \infty$

Proof

We know

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(\mathcal{A}_S)\right] \;\leq\; \mathbb{E}_{S\sim\mathcal{D}^m}\left[L_S(\mathcal{A}_S)\right] + \beta_m$$

Using fact that $\mathcal{A} = \text{ERM}$

$$\implies \quad L_S(\mathcal{A}_S) \;\leq\; L_S(h) \qquad \text{for all } h \in \mathcal{H}$$

$$\implies \quad \mathbb{E}_{S\sim\mathcal{D}^m}\left[L_S(\mathcal{A}_S)\right] \;\leq\; \underbrace{\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_S(h)\right]}_{=L_{\mathcal{D}}(h)} \qquad \text{for all } h \in \mathcal{H}$$

$$\implies \quad \mathbb{E}_{S\sim\mathcal{D}^m}\left[L_S(\mathcal{A}_S)\right] \;\leq\; \underbrace{\inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h)}_{=L_{\mathcal{D}}(\mathcal{H})}$$

Uniform stability: Yet another notion of stability

- Fix learner $\mathcal{A}$, loss function ℓ and $\beta^u : \mathbb{N} \to \mathbb{R}$
- Learner $\mathcal{A}$ is uniformly stable
  - with rate β^u with respect to ℓ
  - if for every sample size m

$$\sup_{S \in (\mathcal{X}\times\mathcal{Y})^m} \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \max_{i\in\{1,\ldots,m\}} \left|\ell\big(\mathcal{A}_{S_{-i}}(x), y\big) - \ell\big(\mathcal{A}_S(x), y\big)\right| \leq \beta^u_m$$

- On-average-replace-one stability can only give bound on $\mathbb{E}_S[L_\mathcal{D}(\mathcal{A}_S)]$

  In contrast, uniform stability can give high probability bound on $L_\mathcal{D}(\mathcal{A}_S)$ (will see later)

Validation: Estimating generalisation error

Context

- Generalisation bounds for ERM solution:

$$L_{\mathcal{D}}(\widehat{h}) \leq L_{\mathcal{D}}(\mathcal{H}) + O\left(\sqrt{\frac{\mathrm{VCdim}(\mathcal{H})}{m}}\right)$$

- Validation: Estimate the true risk of $\widehat{h}$
  - Get better estimates of $L_{\mathcal{D}}(\widehat{h})$ than obtained from uniform convergence bounds
- Validation vs empirical risk:
  - Both estimates of $L_{\mathcal{D}}(\cdot)$
  - $\widehat{h}$ depends on empirical risk, but not on validation error

Hold out set

- We have assumed access to m labelled examples
- Split labelled examples into two parts
  - $S =$ training set with m_s examples
  - $V =$ validation / hold out set with m_v examples $\qquad m = m_s + m_v$

$$\widehat{h} = \mathcal{A}(S) \qquad\qquad L_V(\widehat{h}) = \frac{1}{m_v} \sum_{(x,y)\in V} \ell(\widehat{h}(x), y)$$

- $\widehat{h}$ depends only on S, and $L_V(\widehat{h})$ is an “independent review” of $\widehat{h}$

Validation using hold out set

Theorem Valid.1 (Bound on generalisation error from validation error)

Assume

- $S \sim \mathcal{D}^m$ *and* $V \sim \mathcal{D}^{m_v}$ *independent*
- $\widehat{h} = \mathcal{A}(S)$
- *Loss function* $\ell(\cdot,\cdot)$ *lies in* $[0,1]$

For every $\delta \in (0,1)$,

$$\mathbb{P}_{V\sim\mathcal{D}^{m_v}}\left(\left|L_{\mathcal{D}}(\widehat{h}) - L_V(\widehat{h})\right| > \sqrt{\frac{\ln(\frac{2}{\delta})}{2m_v}}\right) \leq \delta$$

Proof hints

- Probability only over V
- Can treat $\widehat{h} = \mathcal{A}(S)$ as a fixed function
- Apply Hoeffding's inequality to derive the bound
- Why did not we need union bound over $\mathcal{H}$?
  - We derive bound only for a fixed $\widehat{h}$
  - Contrast to uniform convergence:
    - Probability over $S \implies \widehat{h} = \mathcal{A}(S)$ is not fixed, but random
    - Needed to show $|L_S(\cdot) - L_{\mathcal{D}}(\cdot)|$ small for all $h \in \mathcal{H}$

Validation vs uniform convergence

- Consequence of uniform convergence: With probability $1-\delta$

$$L_S(\widehat{h}) - C\sqrt{\frac{\mathrm{VCdim}(\mathcal{H})\ln m_s + \ln(\frac{1}{\delta})}{m_s}} \leq L_{\mathcal{D}}(\widehat{h}) \leq L_S(\widehat{h}) + C\sqrt{\frac{\mathrm{VCdim}(\mathcal{H})\ln m_s + \ln(\frac{1}{\delta})}{m_s}}$$

- Validation using hold out set: With probability $1-\delta$,

$$L_V(\widehat{h}) - \sqrt{\frac{\ln(\frac{2}{\delta})}{2m_v}} \quad \leq \quad L_{\mathcal{D}}(\widehat{h}) \quad \leq \quad L_V(\widehat{h}) + \sqrt{\frac{\ln(\frac{2}{\delta})}{2m_v}}$$

No dependence on $\mathcal{H}$

- Validation provides tighter bounds for large m_v (practical choice: $m_v =$ 10-30% of m)

k-Fold cross validation

- Hold out set significantly reduces the number of training samples
- k-fold cross validation: another estimator for generalisation error
  - Split labelled data S into k partitions $S_1, \ldots, S_k$
  - Let $S_{-i} = S \backslash S_i$
  - For every $i = 1, \ldots, k$: train on S_{-i} and validate using S_i
  - Average k validation errors

$$L_{k-cv}(\mathcal{A}_S) = \frac{1}{k} \sum_{i=1}^{k} L_{S_i}(\mathcal{A}_{S_{-i}}) \qquad \text{notation: } \mathcal{A}_S = \mathcal{A}(S)$$

Leave one out

- k-fold cross validation with $k = m$
  - $S_i = (x_i, y_i)$ and $S_{-i} = S \backslash (x_i, y_i)$

$$L_{loo}(\mathcal{A}_S) = \frac{1}{m} \sum_{i=1}^{m} \ell\left(\mathcal{A}_{S_{-i}}(x_i), y_i\right)$$

Generalisation error from cross validation / loo error

- How can we bound $L_{\mathcal{D}}(\mathcal{A}_S)$ based on $L_{k-cv}(\mathcal{A}_S)$ or $L_{loo}(\mathcal{A}_S)$?
- $L_{k-cv}(\mathcal{A}_S)$ or $L_{loo}(\mathcal{A}_S)$ likely to over-estimate $L_{\mathcal{D}}(\mathcal{A}_S)$. Why?
  - Hint: Compute expectation of $L_{k-cv}(\mathcal{A}_S)$ or $L_{loo}(\mathcal{A}_S)$

Lemma Valid.2

In expectation, loo error with m samples equals true risk with $m-1$ samples

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{loo}(\mathcal{A}_S)\right] = \mathbb{E}_{S'\sim\mathcal{D}^{m-1}}\left[L_{\mathcal{D}}(\mathcal{A}_{S'})\right]$$

- $L_{\mathcal{D}}(\mathcal{A}_{S'}) =$ *generalisation error of predictor trained using $m-1$ iid samples*

Proof (try by yourself before seeing this slide)

$$
\begin{aligned}
\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{loo}(\mathcal{A}_S)\right] &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{S\sim\mathcal{D}^m}\left[\ell\left(\mathcal{A}_{S_{-i}}(x_i), y_i\right)\right] \\
&= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{S_{-i}\sim\mathcal{D}^{m-1}}\left[\mathbb{E}_{(x_i,y_i)\sim\mathcal{D}}\left[\ell\left(\mathcal{A}_{S_{-i}}(x_i), y_i\right)\right]\right] \\
&= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{S_{-i}\sim\mathcal{D}^{m-1}}\left[L_{\mathcal{D}}(\mathcal{A}_{S_{-i}})\right] \\
&= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{S'\sim\mathcal{D}^{m-1}}\left[L_{\mathcal{D}}(\mathcal{A}_{S'})\right] \qquad \text{replace } S_{-i} \text{ by any } S' \sim \mathcal{D}^{m-1}
\end{aligned}
$$

Derive corresponding result for $L_{k-cv}(\mathcal{A}_S)$ assuming all k parts of equal size

Confidence interval for $L_{\mathcal{D}}(\mathcal{A}_S)$

- Previous statement is only about expectation
- We are interested in following types of bounds: With probability $1-\delta$

$$\left|L_{\mathcal{D}}(\mathcal{A}_S) - L_{loo}(\mathcal{A}_S)\right| \leq \epsilon \qquad \text{or} \qquad L_{\mathcal{D}}(\mathcal{A}_S) \leq L_{loo}(\mathcal{A}_S) + \epsilon$$

- Can we use Hoeffding's inequality to bound $\left|L_{loo}(\mathcal{A}_S) - \mathbb{E}_S[L_{loo}(\mathcal{A}_S)]\right|$?
  - No. L_{loo} is mean of dependent terms
- Need tools based on algorithmic stability

Uniform stability: Yet another notion of stability

- Fix learner $\mathcal{A}$, loss function ℓ and $\beta^u : \mathbb{N} \to \mathbb{R}$
- Learner $\mathcal{A}$ is uniformly stable
  - with rate β^u with respect to ℓ
  - if for every sample size m

$$\sup_{S \in (\mathcal{X} \times \mathcal{Y})^m} \; \sup_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \; \max_{i \in \{1,\ldots,m\}} \left| \ell\big(\mathcal{A}_{S_{-i}}(x), y\big) - \ell\big(\mathcal{A}_S(x), y\big) \right| \leq \beta^u_m$$

- On-average-replace-one stability can only give bound on $\mathbb{E}_S[L_{\mathcal{D}}(\mathcal{A}_S)]$

  In contrast, uniform stability can give high probability bound on $L_{\mathcal{D}}(\mathcal{A}_S)$

Generalisation from uniform stability

- $L_{\mathcal{D}}^{0-1}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbf{1}\{h(x) \neq y\}]$ ... risk with respect to 0-1 loss
- Ramp loss:
  - Allow $h(x)$ to be any real value

$$\ell^{ramp}(h(x), y) = \begin{cases} 0 & \text{if } y \cdot h(x) \geq 1 \\ 1 & \text{if } y \cdot h(x) \leq 0 \\ 1 - y \cdot h(x) & \text{if } 0 < y \cdot h(x) < 1 \end{cases}$$

- $L_{\mathcal{D}}^{ramp} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell^{ramp}(h(x), y)\right]$ ... risk with respect to ramp loss
- Later: $L_{\mathcal{D}}^{0-1}(h) \leq L_{\mathcal{D}}^{ramp}(h)$

Generalisation from uniform stability (Bousquet-Elisseeff, Thm 17)

Theorem Valid.3 (Generalisation error bound from loo and training error)

Assume $\mathcal{A}$ has uniform stability rate β with respect to ramp loss

For $\delta \in (0,1)$, training sample $S \sim \mathcal{D}^m$, with probability $1-\delta$,

$$L_{\mathcal{D}}^{0-1}(A_S) \;\leq\; L_{\mathcal{D}}^{ramp}(\mathcal{A}_S) \;<\; L_S^{ramp}(\mathcal{A}_S) + 2\beta + (4m\beta + 1)\sqrt{\frac{\ln(\frac{1}{\delta})}{2m}}$$

$$L_{\mathcal{D}}^{0-1}(A_S) \;\leq\; L_{\mathcal{D}}^{ramp}(\mathcal{A}_S) \;<\; L_{loo}^{ramp}(\mathcal{A}_S) + \beta + (4m\beta + 1)\sqrt{\frac{\ln(\frac{1}{\delta})}{2m}}$$

Proof skipped

Meaningful only if $\beta_m \ll \frac{1}{\sqrt{m}}$. For soft SVM, $\beta_m = O(\frac{1}{m})$