Statistical Foundations of Learning

Debarghya Ghoshdastidar

School of Computation, Information and Technology
Technical University of Munich

Probably Approximately Correct (PAC) learning

Recap

- ERM over $\mathcal{H} \subset \{\pm 1\}^{\mathcal{X}}$: $\quad \widehat{h} = \arg\min_{h \in \mathcal{H}} L_S(h)$
- Goal: Bound generalisation error $L_{\mathcal{D}}(\widehat{h})$
- $\mathrm{VCdim}(\mathcal{H}) =$ size of largest set that $\mathcal{H}$ can shatter
  - If $\mathcal{H}$ has more complex functions, then $\mathrm{VCdim}(\mathcal{H})$ is larger
- Generalisation error bound when $\mathrm{VCdim}(\mathcal{H}) = d$:

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

Outline

- Probably Approximately Correct (PAC) learning:
  - Another view of results derived so far ... for which $\mathcal{H}$ is ERM good?
- No free lunch theorem
  - Shows that ERM does not have low generalisation error if $\mathcal{H} = \{\pm 1\}^{\mathcal{X}}$
- Fundamental theorem of statistical learning
  - Finite VC dimension necessary and sufficient for (agnostic) PAC learnability
  - Bounds on sample complexity
- Proof of no free lunch theorem

Realisable setting

- $\mathcal{D}$ is a distribution on $\mathcal{X} \times \{\pm 1\}$
- $\mathcal{D}$ is **realisable** with respect to hypothesis class $\mathcal{H}$ and loss function ℓ if

$$L_{\mathcal{D}}(\mathcal{H}) = 0$$

- What conditions do $\mathcal{D}$ satisfy if it is realisable?
  - Bayes risk $L^*_{\mathcal{D}} = 0$
  - There is $h^* \in \mathcal{H}$ such that $L_{\mathcal{D}}(h^*) = 0$ ... assuming $L_{\mathcal{D}}(\mathcal{H})$ is a min, not infimum
  - $\mathcal{D}$ can be decomposed as:

$$(x, y) \sim \mathcal{D} \qquad \Longrightarrow \qquad x \sim \mathcal{D}_{\mathcal{X}} \ \text{ and } \ \mathbb{P}_{\mathcal{Y}|\mathcal{X}}\big(y = h^*(x) \mid x\big) = 1$$

Understanding the meaning of PAC

- Probably Approximately Correct (PAC) Learning
  - Originally formulated in the realisable setting
  - There is $h^* \in \mathcal{H}$ such that $L_{\mathcal{D}}(h^*) = 0$
  - Note: Our notation differs from Ben David's book since we assume $\mathcal{D}$ is distribution over $\mathcal{X} \times \mathcal{Y}$
- Correct learning:
  - Find $h^* \in \mathcal{H}$ given training sample $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \sim \mathcal{D}^m$
  - More generally, design learning algorithm $\mathcal{A}$ such that

$$\widehat{h} = \mathcal{A}(S) \in \mathcal{H} \quad \text{satisfies} \quad L_{\mathcal{D}}(\widehat{h}) = 0$$

Understanding the meaning of PAC

- Approximately correct learning:
  - Given $S \sim \mathcal{D}^m$ and allowable approximation $\epsilon \in (0, 1)$
  - Design $\mathcal{A}$ such that $\widehat{h} = \mathcal{A}(S) \in \mathcal{H}$ satisfies $L_{\mathcal{D}}(\widehat{h}) \leq \epsilon$
- Probably approximately correct learning:
  - Find $\widehat{h} = \mathcal{A}(S)$ that is approximately correct with specified probability $1 - \delta, \quad \delta \in (0, 1)$

$$L_{\mathcal{D}}(\widehat{h}) \leq \epsilon \qquad \text{with probability } 1 - \delta$$

  - Equivalently, $\mathbb{P}_{S \sim \mathcal{D}^m} \left(L_{\mathcal{D}}(\mathcal{A}(S)) > \epsilon \right) \leq \delta$

PAC Learnability

- Let $\mathcal{D}$ be realisable w.r.t. $\mathcal{H}$ and loss function ℓ
- When can we learn $\mathcal{D}$?
  - Given $\epsilon, \delta \in (0, 1)$
  - There is a sample size m_0 and learner $\mathcal{A} : \bigcup_{m \geq 1} (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$ such that

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left(L_{\mathcal{D}}(\mathcal{A}(S)) > \epsilon \right) \leq \delta \qquad \text{for all } m \geq m_0$$

  - Equivalently, $L_{\mathcal{D}}(\mathcal{A}(S)) \leq \epsilon$ with probability $1 - \delta$

PAC Learnability

- $\mathcal{H}$ is PAC learnable w.r.t. loss ℓ if *there exists*
  - function $m_{\mathcal{H}} : (0,1) \times (0,1) \to \mathbb{N}$
  - learning algorithm $\mathcal{A} : \bigcup_{m \geq 1} (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$

  such that for every
  - $\epsilon, \delta \in (0,1)$
  - realisable distribution $\mathcal{D}$
  - sample size $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ and $S \sim \mathcal{D}^m$

$$L_{\mathcal{D}}(\mathcal{A}(S)) \leq \epsilon \qquad \text{with probability } 1 - \delta$$

Finite VC classes are PAC learnable

- We call $\mathcal{H}$ finite VC class if $\text{VCdim}(\mathcal{H}) = d < \infty$
- Show that finite VC classes are PAC learnable
- Let $\mathcal{A} = \text{ERM}$

$$L_{\mathcal{D}}(\mathcal{A}(S)) < \underbrace{L_{\mathcal{D}}(\mathcal{H})}_{0 \text{ for realisable } \mathcal{D}} + \underbrace{2\sqrt{\frac{8\left(d \ln\left(\frac{2em}{d}\right) + \ln\left(\frac{4}{\delta}\right)\right)}{m}}}_{\leq \epsilon \text{ for large } m} \quad \text{with probability } 1 - \delta$$

- Define $m_{\mathcal{H}}(\epsilon, \delta) = C\frac{d\ln\left(\frac{d}{\epsilon}\right) + \ln\left(\frac{1}{\delta}\right)}{\epsilon^2}$
  - For a large constant C, above bound $\leq \epsilon$ for all $m \geq m_{\mathcal{H}}(\epsilon, \delta)$

Agnostic setting

- Limitations of realisable setting:
  - Assumes the true function h^* lies in $\mathcal{H}$
    — If we use SVM, this is same as assuming data is always linearly separable
  - Assumes there is a true h^*
    — Bayes error $L^*_{\mathcal{D}} = 0$ (not always practical)
- Agnostic setting:
  - Does not make either assumption
  - Main implication: $L_{\mathcal{D}}(\mathcal{H}) > 0$ (already considered in previous chapter)
  - How do we define learning / learnability in agnostic setting?

Agnostic PAC Learning

- Given $\epsilon, \delta \in (0, 1)$, find $\mathcal{A}$ such that

$$L_{\mathcal{D}}(\mathcal{A}(S)) \leq L_{\mathcal{D}}(\mathcal{H}) + \epsilon \qquad \text{with probability } 1 - \delta$$

- We do not require $\mathcal{A}(S)$ to be ϵ-approximately correct
  - This may not be possible if $h^* \notin \mathcal{H}$
- We require $\mathcal{A}(S)$ to be ϵ-close to best predictor in $\mathcal{H}$

Agnostic PAC Learnability

- $\mathcal{H}$ is agnostic PAC learnable w.r.t. loss ℓ if *there exists*
  - function $m_{\mathcal{H}} : (0,1) \times (0,1) \to \mathbb{N}$
  - learning algorithm $\mathcal{A} : \bigcup_{m \geq 1} (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$

  such that for every
  - $\epsilon, \delta \in (0,1)$
  - distribution $\mathcal{D}$
  - sample size $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ and $S \sim \mathcal{D}^m$

$$L_{\mathcal{D}}(\mathcal{A}(S)) \leq L_{\mathcal{D}}(\mathcal{H}) + \epsilon \qquad \text{with probability } 1 - \delta$$

Sample complexity

- Exercise: Show that finite VC classes are agnostic PAC learnable
- Sample complexity:
  - Function $m_{\mathcal{H}}(\cdot, \cdot)$ is called sample complexity for learning $\mathcal{H}$
  - Fix ϵ (allowable excess error) and δ (probability of not satisfying bound)
  - $m_{\mathcal{H}}(\epsilon, \delta) =$ minimum number of training samples needed to agnostically PAC learn best predictor in $\mathcal{H}$ for any $\mathcal{D}$

No Free Lunch Theorem: Significance

- Previous results:
  - If $\text{VCdim}(\mathcal{H}) < \infty$, then $\mathcal{H}$ is (agnostic) PAC learnable
  - For $m > m_{\mathcal{H}}(\epsilon, \delta)$, can PAC learn any $\mathcal{D}$
- No Free Lunch Theorem:
  - $|\mathcal{X}| = \infty$ and $\mathcal{H} = \{\pm 1\}^{\mathcal{X}}$
  - There is no $m_{\mathcal{H}}(\epsilon, \delta) < \infty$ that is sufficient for learning all $\mathcal{D}$
  - Extension: For any $\mathcal{H}$ with $\text{VCdim}(\mathcal{H}) = \infty$, there is no $m_{\mathcal{H}}(\epsilon, \delta) < \infty$ that is sufficient for learning all $\mathcal{D}$

No Free Lunch Theorem

Theorem PAC.1 (No Free Lunch Theorem)

Assume

- $\mathcal{A}$ *is a learner for binary classification over* $\mathcal{X}$*, that is,* $\mathcal{A}(\cdot) \in \{\pm 1\}^{\mathcal{X}}$
- *training sample size* $m < \frac{|\mathcal{X}|}{2}$

One can construct a distribution $\mathcal{D}$ *of the form* $\mathcal{D} = \mathcal{D}_{\mathcal{X}} \times \mathbb{P}_{\mathcal{Y}|\mathcal{X}}$ *such that:*

- $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y = h^*(x) \mid x) = 1$ *for some* h^* $\quad \ldots \mathcal{D}$ *realisable w.r.t.* $\{\pm 1\}^{\mathcal{X}}$*, or* $L_{\mathcal{D}}(h^*) = 0$
- *For* $S \sim \mathcal{D}^m$*,*

$$L_{\mathcal{D}}(\mathcal{A}(S)) \geq \frac{1}{8} \qquad \textit{with probability } > \frac{1}{7}$$

Interpreting No Free Lunch Theorem

- Fix any sample size $m < \frac{1}{2}|\mathcal{X}|$ and any algorithm $\mathcal{A}$
- One can find $\mathcal{D}$ such that $\mathcal{A}$ cannot PAC-learn $\mathcal{D}$ for $\epsilon < \frac{1}{8}$, $\delta < \frac{1}{7}$
- If $|\mathcal{X}| = \infty$, above is true for any finite m and any learner $\mathcal{A}$
- Why does NFL not contradict previous learnability results?
  - NFL: No learner $\mathcal{A}$ can PAC learn $\{\pm 1\}^{\mathcal{X}}$ if $|\mathcal{X}| = \infty$
  - PAC learnability: ERM over $\mathcal{H}$ can PAC learn $\mathcal{H}$ if it has finite VC-dim
  - Agnostic PAC: ERM over $\mathcal{H}$ satisfies $L_{\mathcal{D}}(\mathcal{A}(S)) \leq L_{\mathcal{D}}(\mathcal{H}) + \epsilon$ $\quad \ldots L_{\mathcal{D}}(\mathcal{A}(S) > \epsilon$ possible

Learnability of infinite VC classes

Theorem PAC.2 (Infinite VC classes are not PAC learnable)

Assume

- $\mathcal{X}$ *is infinite domain*
- $\mathcal{H} \subset \{\pm 1\}^{\mathcal{X}}$ *has infinite VC-dimension*
- $\mathcal{A}$ *is a learner that outputs predictors from* $\mathcal{H}$

For every $m < \infty$*, there is a distribution* $\mathcal{D}$*, realisable w.r.t.* $\mathcal{H}$*, such that:*

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left(L_{\mathcal{D}}(\mathcal{A}(S)) \geq \frac{1}{8} \right) > \frac{1}{7} \, .$$

Proof is exercise

Fundamental theorem of statistical learning

Theorem PAC.3 (Fundamental theorem for binary classification)

Let $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ and consider the 0-1 loss. The following are equivalent:

(1) *$\mathcal{H}$ has finite VC dimension*

(2) *Any ERM rule is a successful agnostic PAC learner for $\mathcal{H}$*

(3) *Any ERM rule is a successful PAC learner for $\mathcal{H}$*

(4) *$\mathcal{H}$ is agnostic PAC learnable*

(5) *$\mathcal{H}$ is PAC learnable*

Fundamental theorem for binary classification

- Significance:
  - Finite VC dimension necessary and sufficient for (agnostic) PAC learnability
  - When $\mathcal{H}$ is learnable, ERM finds nearly best predictor in $\mathcal{H}$
  - Learnability is an algorithm independent concept (not specific to ERM), yet ERM works
- Generalisation to other problems:
  - Equivalent results for other loss functions, and other problems (example: regression)
  - Based on alternatives for VC dimension (pseudo-dimension, fat-shattering dimension)

How to prove equivalence of statements?

- Approach 1: Prove $i \implies j$ and $j \implies i$ for all i, j
  - Too many proofs; Some may not be straightforward
- Approach 2: Prove any cycle (examples: $1 \implies 2 \implies 4 \implies 5 \implies 3 \implies 1$)
  - Proving all implications in a cycle may not be easy
- Approach 3: Prove multiple cycles that overlap
  - For theorem, show $1 \implies 2 \implies 4 \implies 5 \implies 1$ and $1 \implies 3 \implies 5 \implies 1$
  - Exercise: Use generalisation error bound for $1 \implies 2$ or $1 \implies 3$
    $2 \implies 4 \implies 5$ and $3 \implies 5$ straightforward
    $5 \implies 1$ proved by contradiction using non PAC learnability of infinite VC classes

Sample complexity for finite VC classes

Theorem PAC.4 (Sample complexity for agnostic learnability)

$\text{VCdim}(\mathcal{H}) = d < \infty$

$\mathcal{H}$ is agnostic PAC learnable with sample complexity

$$C_1 \frac{d + \log_2(\frac{1}{\delta})}{\epsilon^2} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d + \ln(\frac{1}{\delta})}{\epsilon^2}$$

for some constants $C_1, C_2 > 0$

Proof skipped. Weaker upper bound follows from generalisation error bounds

$$m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d \log_2(\frac{d}{\epsilon}) + \ln(\frac{1}{\delta})}{\epsilon^2}$$

Significance of sample complexity bounds

- Number of sample necessary and sufficient for nearly optimal prediction using $\mathcal{H}$

$$m = \text{ linear function of } \ \mathrm{VCdim}(\mathcal{H}) \ , \ \frac{1}{\epsilon^2} \ , \ \ln\left(\frac{1}{\delta}\right)$$

- Quantitative version of fundamental theorem
  - Finite VC classes can be learned with $\implies m < \infty$ sample
  - Learning with $m < \infty$ samples not possible for infinite VC classes

- Better sample complexity for PAC learnability (proof skipped)

$$C_1 \frac{d + \log_2(\frac{1}{\delta})}{\epsilon} \ \leq \ m_{\mathcal{H}}(\epsilon, \delta) \ \leq \ C_2 \frac{d \ln(\frac{1}{\epsilon}) + \ln(\frac{1}{\delta})}{\epsilon}$$

No Free Lunch Theorem

Theorem PAC.5 (No Free Lunch Theorem)

Assume

- $\mathcal{A}$ *is a learner for binary classification over* $\mathcal{X}$*, that is,* $\mathcal{A}(\cdot) \in \{\pm 1\}^{\mathcal{X}}$
- *training sample size* $m < \frac{|\mathcal{X}|}{2}$

One can construct a distribution $\mathcal{D}$ *of the form* $\mathcal{D} = \mathcal{D}_{\mathcal{X}} \times \mathbb{P}_{\mathcal{Y}|\mathcal{X}}$ *such that:*

- $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y = h^*(x) \mid x) = 1$ *for some* h^* $\quad \ldots \mathcal{D}$ *realisable w.r.t.* $\{\pm 1\}^{\mathcal{X}}$*, or* $L_{\mathcal{D}}(h^*) = 0$
- *For* $S \sim \mathcal{D}^m$*,*

$$L_{\mathcal{D}}(\mathcal{A}(S)) \geq \frac{1}{8} \qquad \textit{with probability } > \frac{1}{7}$$

Proof idea

- Simplifying the theorem statement
  - Given learner $\mathcal{A}$ and sample size $m < \frac{1}{2}|\mathcal{X}|$
  - Find $\mathcal{D} = \mathcal{D}_{\mathcal{X}} \times h^*$ $\quad (x \sim \mathcal{D}_{\mathcal{X}}, y = h^*(x))$ $\quad \ldots$ abusing notation
  - Such that $L_{\mathcal{D}}(\mathcal{A}(S))$ is large
- We do not construct single $\mathcal{D}_{\mathcal{X}}, h^*$, instead:
  - Fix $\mathcal{D}_{\mathcal{X}}$
  - Give a collection $\{h_1, \ldots, h_T\}$ such that $\mathcal{D}_{\mathcal{X}}, h_i$ satisfies above for at least one h_i

Proof idea (contd)

- Consider $C \subset \mathcal{X}$ of size $2m$
  - C is shattered by $\{\pm 1\}^{\mathcal{X}}$
  - $T = 2^{2m}$ functions from C to $\{\pm 1\}$. Call them as $h_1, \ldots, h_T$ — chosen collection of functions
  - For the case of infinite VC classes, take C that is shattered
- Choose $\mathcal{D}_{\mathcal{X}}$ = uniform over C
- Need to show $\mathcal{A}$ cannot learn at least one of above h_i

Proof idea (contd)

- Define $\mathcal{D}_i = \mathcal{D}_{\mathcal{X}} \times h_i$
- Formally, need to show — note: $[T] = \{1, 2, \ldots, T\}$

$$\mathbb{P}_{S \sim \mathcal{D}_i^m} \left(L_{\mathcal{D}_i}(\mathcal{A}(S)) \geq \frac{1}{8} \right) > \frac{1}{7} \qquad \text{for some } i \in [T]$$

- Suffices to show

$$\max_{i \in [T]} \mathbb{E}_{S \sim \mathcal{D}_i^m} \left[L_{\mathcal{D}_i}(\mathcal{A}(S)) \right] \geq \frac{1}{4}$$

- Exercise: Prove the following using Markov inequality, and verify that it leads to the above correspondence

$$\text{For r.v. } Z \in [0, 1], \qquad \mathbb{P}(Z \geq a) > \frac{\mathbb{E}[Z] - a}{1 - a} \; \forall a \in (0, 1)$$

Proof idea: Characterising all possible S

- For $S = \{(x_j, y_j)\}_{j=1}^m \sim \mathcal{D}_i^m$, let

$$X = (x_1, \ldots, x_m) \sim \mathcal{D}_{\mathcal{X}}^m \qquad \ldots \text{ randomly observed data / features}$$

- $Q = (2m)^m$ ways to sample m examples from C
  - Denote possible sequences by $X_1, X_2, \ldots, X_Q \in C^m$
  - All equally likely, since $\mathcal{D}_{\mathcal{X}}$ is uniform
- $Q \cdot T = (2m)^m 2^{2m}$ possibilities of S $\qquad \ldots$ if $\mathcal{D}$ can be any of $\mathcal{D}_1, \ldots, \mathcal{D}_T$

$$S_{i,j} = (X_j, h_i(X_j)) \qquad i \in [T], \quad j \in [Q]$$

  - m features in X_j observe, and labelled with h_i

Proof idea: Lower bound on expectation

- Write $\mathbb{E}_{S\sim\mathcal{D}_i^m}[\cdot]$ as average over $S_{i,1},\ldots,S_{i,Q}$
- Note that maximum $\geq$ average, and average $\geq$ minimum

$$\max_{i\in[T]} \mathbb{E}_{S\sim\mathcal{D}_i^m}\left[L_{\mathcal{D}_i}(\mathcal{A}(S))\right] \;\geq\; \frac{1}{T}\sum_{i=1}^{T}\frac{1}{Q}\sum_{j=1}^{Q} L_{\mathcal{D}_i}(\mathcal{A}(S_{i,j})) \;\geq\; \min_{j\in[Q]}\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(\mathcal{A}(S_{i,j}))$$

- Suffices to show: For every X_j,

$$\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(\mathcal{A}(S_{i,j})) \;\geq\; \frac{1}{4}$$

Proof idea: Dealing with single sequence X_j

- For every $X_j \in C^m$, there are $v_1, \ldots, v_p \in C$
  - $v_1, \ldots, v_p$ do not appear in X_j
  - $p \geq m$ (why?)
- For any predictor $h : C \to \{\pm 1\}$, can show

$$L_{\mathcal{D}_i}(h) \;=\; \frac{1}{2m} \sum_{x \in C} \mathbf{1}\left\{h(x) \neq h_i(x)\right\} \;\geq\; \frac{1}{2} \cdot \frac{1}{p} \sum_{k=1}^{p} \mathbf{1}\left\{h(v_k) \neq h_i(v_k)\right\}$$

  - We will use this lower bound for the case of $h = \mathcal{A}(S_{i,j})$

Proof idea: Dealing with single sequence X_j (contd)

- Averaging over all $i \in [T]$ gives

$$\begin{aligned}\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(\mathcal{A}(S_{i,j})) &\geq \frac{1}{2}\cdot\frac{1}{p}\sum_{k=1}^{p}\frac{1}{T}\sum_{i=1}^{T}\mathbf{1}\left\{\mathcal{A}(S_{i,j})(v_k) \neq h_i(v_k)\right\} \\ &\geq \frac{1}{2}\cdot\min_{k\in[p]}\frac{1}{T}\underbrace{\sum_{i=1}^{T}\mathbf{1}\left\{\mathcal{A}(S_{i,j})(v_k) \neq h_i(v_k)\right\}}_{=\frac{T}{2}\text{ (reason below)}} = \frac{1}{4}\end{aligned}$$

- For every h_i and $v_k \in C$, can find $h_{i'}$ such that

$$h_i(v_k) \neq h_{i'}(v_k) \text{ and } h_i(x) = h_{i'}(x)\ \forall x \in C\backslash\{v_k\} \qquad \text{(why?)}$$

- Group $h_1, \ldots, h_T$ into $\frac{T}{2}$ pairs of $(h_i, h_{i'})$ as above. Can show

$$\mathbf{1}\left\{\mathcal{A}(S_{i,j})(v_k) \neq h_i(v_k)\right\} + \mathbf{1}\left\{\mathcal{A}(S_{i',j})(v_k) \neq h_{i'}(v_k)\right\} = 1$$