Models of Language Acquisition: Part II

Matilde Marcolli

MAT1509HS: Mathematical and Computational Linguistics

University of Toronto, Winter 2019, T 4-6 and W 4, BA6180

Probably Approximately Correct Model of Language Learning

• General setting of Statistical Learning Theory: objects of learning are functions

• concept class: set $\mathcal{F}$ of possible target functions

• hypothesis class set $\mathcal{H}$ of functions $f : X \to Y$
typically assume $\mathcal{F} = \mathcal{H}$

• for language case: $X =$ set $\mathfrak{A}^\star$ of all possible strings on an alphabet, $Y = \{0, 1\}$

• given a language $\mathcal{L} \subset \mathfrak{A}^\star$ consider the associated indicator function (characteristic function) $\chi_{\mathcal{L}} : \mathfrak{A}^\star \to \{0, 1\}$

$$\chi_{\mathcal{L}}(x) = \begin{cases} 1 & x \in \mathcal{L} \\ 0 & x \notin \mathcal{L} \end{cases}$$

- Language $\mathcal{L}$ can be seen as
  1. recursively enumerable subset $\mathcal{L} \subset \mathfrak{A}^\star$
  2. Turing machine (program) that recognizes $\mathcal{L}$
  3. indicator function $\chi_{\mathcal{L}}$

- distance function on the space of languages (something better than the discrete 0/1 metric): $L^1(\mathbb{P})$-distance
- use a probability measure $\mathbb{P}$ on $\mathfrak{A}^\star$ (Bernoulli, Markov,...)
- define $L^1(\mathbb{P})$-distance as

$$d_{\mathbb{P}}(\mathcal{L}, \mathcal{L}') = \sum_{s \in \mathfrak{A}^\star} |\chi_{\mathcal{L}}(s) - \chi_{\mathcal{L}'}(s)| \ \mathbb{P}(s)$$

- ϵ-neighborhood of a language $\mathcal{L}$

$$\mathcal{N}_\epsilon(\mathcal{L}) = \{\mathcal{L}' \,|\, d_{\mathbb{P}}(\mathcal{L}, \mathcal{L}') < \epsilon\}$$

- Examples are randomly presented to a learner according to the probability distribution $\mathbb{P}$
- Note: both positive and negative examples
- view examples as pairs (x, y) with $x \in \mathfrak{A}^\star$ and $y = \chi_{\mathcal{L}}(x)$
- Data set: $\mathcal{D} = \cup_k \mathcal{D}_k$

$$\mathcal{D}_k = \{(z_1, \ldots, z_k) \mid z_i = (x_i, y_i),\ x_i \in \mathfrak{A}^\star,\ y_i \in \{0, 1\}\}$$

- learning algorithm

$$\mathcal{A} : \mathcal{D} \longrightarrow \mathcal{H}$$

after k data points learner conjectures a function $\hat{h}_k \in \mathcal{H}$

- procedure that minimizes empirical risk:

$$\hat{h}_k(x_j) = \arg\min_{h \in \mathcal{H}} \frac{1}{k} \sum_{j=1}^{k} |y_j - h(x_j)|$$

- $\hat{h}_k = \mathcal{A}(\delta_k)$ with $\delta_k \in \mathcal{D}_k$ a random element
- successful learning: hypothesis $\hat{h}_k$ converges to target $\chi_{\mathcal{L}}$ as $k \to \infty$
- hypothesis $\hat{h}_k$ is a random function (because $\delta_k \in \mathcal{D}_k$ random) so need convergence in probabilistic sense

$$\lim_{k\to\infty} \mathbb{P}\left(d_{\mathbb{P}}(\hat{h}_k, \chi_{\mathcal{L}}) > \epsilon\right) = 0$$

- this means *weak convergence* of random variables

$$\hat{h}_k = \mathcal{A}(\delta_k) \xrightarrow{w} \chi_{\mathcal{L}}$$

$$\mathbb{E}_{\mathbb{P}}\left(\left|\hat{h}_k(s) - \chi_{\mathcal{L}}(s)\right|\right) = \sum_{s\in\mathfrak{A}^\star} \left|\hat{h}_k(s) - \chi_{\mathcal{L}}(s)\right| \mathbb{P}(s) \to 0$$

- Note double role of probability $\mathbb{P}$: in defining $L^1(\mathbb{P})$-distance for convergence; and also in drawing random data $\delta_k \in \mathcal{D}_k$

- target function is

$$\mathcal{L}^{(t)} = \arg\min_{\mathcal{L}} \mathbb{E}_{\mathbb{P}} \left(|\chi_{\mathcal{L}^{(t)}} - \chi_{\mathcal{L}}| \right)$$

- a set of elements $S = \{s_1, \ldots, s_n\}$ in $\mathfrak{A}^\star$ is shattered by the set of functions $\mathcal{H}$ if, for every set of binary vectors $b = (b_1, \ldots, b_n)$ there is a function $h_b \in \mathcal{H}$ such that $h_b(x_i) = 1$ iff $b_i = 1$

- this means that for every way of partitioning the set S into two parts, there is a function in $\mathcal{H}$ that implements the partition ($\mathcal{H}$ must have at least 2^n elements)

- Vapnik–Chervonenkis dimension of $\mathcal{H}$ is D if there is at least one set of D elements that is shattered by $\mathcal{H}$ and no set of $D + 1$ elements is (if no such D then $\dim_{VC} \mathcal{H} = \infty$)

Learnability

• Fact: Set $\mathcal{H}$ of languages (identified with indicator functions $\chi_{\mathcal{L}}$): languages $\mathcal{L}$ of $\mathcal{H}$ are learnable iff Vapnik–Chervonenkis dimension $\dim_{VC} \mathcal{H} < \infty$

• here learnability as weak convergence $\hat{h}_k = \mathcal{A}(\delta_k) \xrightarrow{w} \chi_{\mathcal{L}}$

$$\hat{h}_k(x_j) = \chi_{\hat{\mathcal{L}}_k}(x_j) = \arg\min_{\mathcal{L}} \frac{1}{k} \sum_{j=1}^{k} |y_j - \chi_{\mathcal{L}}(x_j)|$$

empirical risk minimization

• why finite Vapnik–Chervonenkis dimension is needed? Lower bound on learnability...

Lower bound for learning

- suppose Vapnik–Chervonenkis dimension $\dim_{VC} \mathcal{H} = D$
- construct a probability distribution $\mathbb{P}$ on $\mathfrak{A}^\star$ with respect to which learner needs to draw at least

$$m \geq \frac{D}{4} \log_2(\frac{3}{2}) + \log_2(\frac{1}{8\delta})$$

in order to have

$$\mathbb{P}(d(\hat{h}_m, h) > \epsilon) < \delta$$

- so in particular if $D = \infty$ don't have learnability (for this $\mathbb{P}$)

construction of $\mathbb{P}$

- since $\dim_{VC} \mathcal{H} = D$ have a set $x_1, \ldots, x_D$ that is shattered by $\mathcal{H}$:
  - assign measure $\mathbb{P}(x_i) = \frac{1}{D}$ to these points
  - assign measure zero to all other points in $\mathfrak{A}^\star$
- in this measure two functions $h_1, h_2 \in \mathcal{H}$ have distance $d_{\mathbb{P}}(h_1, h_2) = 0$ iff they agree on all points x_i
- mod out $\mathcal{H}$ by equivalence relation $h_1 \sim h_2$ if $h_1(x_i) = h_2(x_i)$ for all $1 \leq i \leq D$: set of equivalence classes $\mathcal{H}_{/\sim}$ has 2^D points

Partitioning of $\mathcal{H}$

- draw a sequence $z = (z_1, \ldots, z_m)$ of random data according to the probability distribution $\mathbb{P}$
- suppose z contains ℓ distinct elements among the $X = \{x_1, \ldots, x_D\}$ (the remaining $D - \ell$ do not occur in z)
- there are then 2^ℓ possible ways in which can label $z = z_h$ by a potential candidate target function $h \in \mathcal{H}_{/\sim}$
- these choices determine a partitioning of $\mathcal{H}$ into disjoint subsets

$$\mathcal{H} = \mathcal{H}_1 \cup \cdots \cup \mathcal{H}_{2^\ell}$$

each $\mathcal{H}_i$ in this partition contains exactly $2^{D-\ell}$ different functions that agree on the ℓ distinct elements in z

Estimate of sum

$$\sum_{h\in\mathcal{H}} d(\mathcal{A}(z_h), h) = \sum_{i=1}^{2^\ell} \sum_{h\in\mathcal{H}_i} d(\mathcal{A}(z_h), h)$$

• the $2^{D-\ell}$ functions in $\mathcal{H}_i$ all agree on data set z_h while on remaining $D-\ell$ elements of X the functions h and $\mathcal{A}(z_h)$ in $\mathcal{H}_i$ disagree somewhere

• if $\mathcal{A}(z_h)$ and h disagree in j places then $d(\mathcal{A}(z_h), h) \geq j/D$ and this can happen in $\binom{D-\ell}{j}$ possible ways:

$$\sum_{h\in\mathcal{H}_i} d(\mathcal{A}(z_h), h) \geq \sum_{j=0}^{D-\ell} \binom{D-\ell}{j} \frac{j}{D} \geq \frac{2^{D-\ell}(D-\ell)}{2D}$$

$\Rightarrow \quad \sum_{h\in\mathcal{H}} d(\mathcal{A}(z_h), h) \geq \frac{2^D(D-\ell)}{2D}$

Candidate target function

- set $S_\ell = \{z \,|\, z \text{ has } \ell \text{ distinct elements}\}$

$$\sum_{z \in S_\ell} \mathbb{P}(z) \frac{1}{2^D} \sum_{h \in \mathcal{H}} d(\mathcal{A}(z), h) \geq \frac{D-\ell}{2D} \mathbb{P}(S_\ell)$$

- change order of sum: $2^{-D} \sum_h \sum_z \mathbb{P}(z) d(\mathcal{A}(z), h)$
- to have inequality there must be at least one $h = h_\star$ with

$$\sum_{z \in S_\ell} \mathbb{P}(z) d(\mathcal{A}(z), h_\star) \geq \frac{D-\ell}{2D} \mathbb{P}(S_\ell)$$

- this $h = h_\star$ is a candidate target function with a certain estimate of inaccuracy of learning hypothesis

Inaccuracy estimate

• Set of draws of m data on which learner's hypothesis $\mathcal{A}(z)$ differs from candidate target $h_\star$ by more than a given size β:

$$\mathbb{S}_\beta = \{z \in S_\ell \,|\, d(\mathcal{A}(z), h_\star) > \beta\}$$

• lower bound on $\mathbb{P}(\mathbb{S}_\beta)$:

$$\frac{D-\ell}{2D}\mathbb{P}(S_\ell) \leq \sum_{z\in\mathbb{S}_\beta} \mathbb{P}(z)d(\mathcal{A}(z), h_\star) + \sum_{z\in S_\ell \smallsetminus \mathbb{S}_\beta} \mathbb{P}(z)d(\mathcal{A}(z), h_\star)$$

$$\leq \mathbb{P}(\mathbb{S}_\beta) + \beta(\mathbb{P}(S_\ell) - \mathbb{P}(\mathbb{S}_\beta))$$

gives $\mathbb{P}(\mathbb{S}_\beta) \geq (1-\beta)\mathbb{P}(\mathbb{S}_\beta) \geq (\frac{D-\ell}{2D} - \beta)\mathbb{P}(S_\ell)$

arrange for $\mathbb{P}(\mathbb{S}_\epsilon) > \delta$

- if target $h_\star$ then with probability $\mathbb{P}(\mathbb{S}_\epsilon)$ learner hypothesis more than ϵ away from target
- take arbitrary ℓ to be $\ell = D/2$ and $\epsilon < 1/8$, then

$$(\frac{D-\ell}{2D} - \beta)\mathbb{P}(S_\ell) > \frac{1}{8}\mathbb{P}(S_\ell)$$

- if have $\mathbb{P}(S_{D/2}) > 8\delta$ get also $\mathbb{P}(\mathbb{S}_\epsilon) > \delta$
- so can arrange that probability of learner hypothesis differing from target more than ϵ is greater than δ
- find conditions for $\mathbb{P}(S_{D/2}) > 8\delta$

arrange for $\mathbb{P}(S_{D/2}) > 8\delta$

- $\mathbb{P}(S_\ell) =$ probability of drawing ℓ distinct elements of $X = \{x_1, \ldots, x_D\}$ in m identically distributed trials
- $\binom{D}{\ell}$ ways of choosing ℓ elements; for each choice $\ell!$ ways in which items can appear in first ℓ positions
- $S_\ell^{(i)} \subset S_\ell$ set of all $z = (z_1, \ldots, z_m)$ with i-th choice of placing the ℓ distinct elements in first ℓ positions (remaining $m - \ell$ positions: same ℓ elements disposed in any way)

$$\mathbb{P}(S_\ell^{(i)}) = (\frac{1}{D})^\ell (\frac{\ell}{D})^{m-\ell}$$

- the $S_\ell^{(i)}$ disjoint so

$$\mathbb{P}(S_\ell) \geq \binom{D}{\ell} \ell! \; \mathbb{P}(S_\ell^{(i)}) = \binom{D}{\ell} \ell! \; (\frac{1}{D})^\ell (\frac{\ell}{D})^{m-\ell}$$

- for $D = 2\ell$ have

$$\binom{D}{\ell} \ell! \, (\frac{1}{D})^{\ell} (\frac{\ell}{D})^{m-\ell} = \frac{(2\ell)!}{\ell! \, \ell^{\ell}} 2^{-m}$$

- also have

$$\frac{(2\ell)!}{\ell! \, \ell^{\ell}} = \prod_{j=1}^{\ell} (1 + \frac{j}{\ell}) \geq (1 + \frac{1}{2})^{\ell/2}$$

$$\mathbb{P}(S_{D/2}) \geq 2^{-m} (\frac{3}{2})^{D/4}$$

- then have $\mathbb{P}(S_{D/2}) > 8\delta$ for

$$m < \frac{D}{4} \log_2(\frac{3}{2}) + \log_2(\frac{1}{8\delta})$$

- conclusion: in constructed probability $\mathbb{P}$ learner needs at least m larger than above to achieve $\mathbb{P}(d(\mathcal{A}(z), h_\star) > \epsilon) < \delta$

Unlearnability problem remains!

- The set of all finite languages is unlearnable
- The set of all regular languages is unlearnable
- The set of all context-free languages is unlearnable

- impose further constraints on learning
- limit the size of grammars... constraint on the number of production rules

Example

- $\mathcal{H}_{n,k}$ = class of Regular Grammars recognized by deterministic finite state automata
- with at most n states
- with fix number of letters $\#\mathfrak{A} = k$
- size of this family

$$\#\mathcal{H}_{n,k} \leq \binom{n}{k}^n$$

- then Vapnik–Chervonenkis dimension

$$\dim_{VC} \mathcal{H}_{n,k} \leq \log_2(\binom{n}{k}^n) \leq nk \log_2(n)$$