Models of Language Acquisition

Matilde Marcolli
CS101: Mathematical and Computational Linguistics

Winter 2015

Language Acquisition Problem

- Target Grammar $\mathcal{G}^{(t)}$
- Example sentences $s_k \in \mathcal{L}_{\mathcal{G}^{(t)}}$
- Hypothesis Grammars $\mathfrak{h} \in \mathcal{H}$
- Learning Algorithm $\mathcal{A}$

• Learners construct from data s_k a model grammar $\mathfrak{h}$ used to generate new test sentences...

• the process converges to the target grammar $\mathcal{G}^{(t)}$

• with a selection procedure (learning algorithm $\mathcal{A}$) for the model grammars $\mathfrak{h} \in \mathcal{H}$

- main difference between child and adult language learning: child only exposed to s_k not to $\mathcal{G}^{(t)}$
- key aspect is passage from passive reception of sample sentences s_k to active forming of new test sentences
- after n sentences $s_1, \ldots, s_n \in \mathcal{L}_{\mathcal{G}^{(t)}}$: grammatical hypothesis $\mathfrak{h}_n \in \mathcal{H}$
- successful language learning requires $\mathfrak{h}_n \to \mathcal{G}^{(t)}$ as $n \to \infty$
- a notion of convergence requires a notion of distance between grammars

$$\lim_{n\to\infty} d(\mathfrak{h}_n, \mathcal{G}^{(t)}) = 0$$

Set of Grammars $\mathcal{H}$

- Context-free Grammars
- Tree-adjoining Grammars
- Probabilistic CFGs; probabilistic TAGs
- Head-driven Phrase Structure Grammars
- Lexical-Functional Grammars

• $\mathcal{H}$ is set of all grammars that can be hypothesized by learner

• in the case of Probabilistic CFG and TAGs: convergence statements should be made in the almost-everywhere sense with respect to the probability measure

Example

- suppose $\mathcal{H} = \{\mathfrak{h}_1, \mathfrak{h}_2\}$ two possibilities
- after N sample sentences $s_1, \ldots, s_N$ hypothesis $\mathfrak{h}_N \in \mathcal{H}$
- some part ϵ of the population will have $\mathfrak{h}_N = \mathfrak{h}_1$, and a part $1 - \epsilon$ will have $\mathfrak{h}_N = \mathfrak{h}_2$
- behavior of the next generation will depend on how similar $\mathfrak{h}_1$ and $\mathfrak{h}_2$ are, how large N, what the specific learning algorithm $\mathcal{A}$ is...
- want to construct a dynamical system that describes this type of learning process

Linguistics vs Biology

- long history of exchanging methods and ideas between Biology and Linguistics
  - Darwin's evolution and Historical Linguistics
  - Phylogenetic trees
  - Syntactic Parameters as Language DNA

- Evolutionary process: necessary ingredients
  - Variation across population
  - Heredity: offsprings resemble parents
  - Transmission with errors: mutation, change
  - Selection process (least effort)

Grammars and Languages

• Grammar $\mathcal{G}$ generates $\mathcal{L} = \mathcal{L}_{\mathcal{G}}$ language (all strings obtained from production rules of grammar)

• Given $\mathcal{L}$: not unique grammar $\mathcal{G}$ with $\mathcal{L} = \mathcal{L}_{\mathcal{G}}$

• Language $\mathcal{L}$ is in the class of recursively enumerable languages (Type 0): can enumerate grammars $\mathcal{G}_m$ with $\mathcal{L}_{\mathcal{G}_m} = \mathcal{L}$ (at most countable)

• Church thesis: partial recursive functions $\Leftrightarrow$ computable

• set $\mathcal{H}$ of hypothesis grammars is some enumerable set

• learning algorithm $\mathcal{A}$ is some partial recursive function from set of sample sentences to $\mathcal{H}$

Assumptions

- sample sentences s_k encountered one at a time:
learning independent of order

- learning algorithm $\mathcal{A}$ should drive convergence to a target grammar independently of order of the s_k

- also assume occurrences of sample sentences s_k as drawn according to independent identically distributed according to an underlying probability distribution

- probability distribution μ on $\mathfrak{A}^\star$, alphabet (lexicon) $\mathfrak{A}$

- only positive examples: μ supported on $\mathcal{L} \subset \mathfrak{A}^\star$

Other Assumptions

- Consistent learner: after N samples $\mathfrak{h}_N$ is consistent with all the s_k, for $k = 1, \ldots, N$
- Empirical risk minimizing learner:

$$\mathfrak{h}_N = \arg\min_{\mathfrak{h} \in \mathcal{H}} \mathcal{R}(\mathfrak{h} \,|\, (s_1, \ldots, s_N))$$

with $\mathcal{R}$ some risk function measuring the fit of $\mathfrak{h}$ to the data $(s_1, \ldots, s_N)$ (the argmin need not be unique)

- Memoryless learner: $\mathfrak{h}_{n+1}$ depends only on s_{n+1} and $\mathfrak{h}_n$ but not on $s_1, \ldots, s_n$

- Enumerative learner:

- first choose an enumeration of $\mathfrak{h} \in \mathcal{H}$

$$\mathcal{H} = \{\mathfrak{h}^{(1)}, \mathfrak{h}^{(2)}, \ldots, \mathfrak{h}^{(m)}, \ldots\}$$

- then start with $\mathfrak{h}^{(1)}$ and compare with datum s_1, stop if consistent
- if not continue down the list, stop at first $\mathfrak{h}^{(m)}$ consistent with s_1
- set first hypothesis $\mathfrak{h}_1 = \mathfrak{h}^{(m)}$
- compare this with s_2, if compatible stop and take as $\mathfrak{h}_2$
- if not continue down the list until find one compatible with s_1 and s_2, etc.

- Learnability: a set $\mathcal{H}$ of grammars is learnable if for all $\mathcal{G}$ in the set $d(\mathfrak{h}_n, \mathcal{G}) \to 0$ for $n \to \infty$

- generalization error: $d(\mathfrak{h}_n, \mathcal{G})$ distance between learner's hypothesis and target

Learning Algorithm

- $\mathcal{D}^k = \{(s_1, \dots, s_k) \,|\, s_i \in \mathfrak{A}^\star\} = \mathfrak{A}^k$ set of all possible sequences of k sample sentences
- under the hypothesis of only positive examples all $s_i \in \mathcal{L}$
- $\mathcal{D} = \cup_{k \geq 1} \mathcal{D}^k$ set of all finite data sequences
- $\mathcal{A} : \mathcal{D} \to \mathcal{H}$ partial recursive function

$$\mathcal{A} : t \in \mathcal{D} \mapsto \mathcal{A}(t) = \mathfrak{h}_t \in \mathcal{H}$$

the learner's hypothesis

Distance functions on the space of grammars

- different notions of convergence on the space $\mathcal{H}$ of grammars

- i-language vs e-language

- purely extensional form: $d(\mathfrak{h}, \mathfrak{h}')$ only depends on $\mathcal{L}_{\mathfrak{h}}$ and $\mathcal{L}_{\mathfrak{h}'}$ (so all grammars producing the same language have distance zero: metric on a quotient space of equivalence classes)

- purely intensional form: fix enumeration $\mathfrak{h}^{(k)}$ of the enumerable set $\mathcal{H}$ and set $d(\mathfrak{h}^{(k)}, \mathfrak{h}^{(\ell)}) = |k - \ell|$

- or distance in terms of grammar complexity (Kolmogorov ordering)

- distance by Hamming metric on the set of syntactic parameters (if think of identifying a grammar as setting parameters correctly)

Inductive Inference Approach

- text τ for a language $\mathcal{L}$: infinite sequence $s_1, \ldots, s_N, \ldots$ of examples, $s_k \in \mathcal{L}$
- assume every element of $\mathcal{L}$ appears at least once in τ
- $\tau_k \in \mathcal{D}^k$ subset of first k elements $(s_1, \ldots, s_k)$ of τ
- given a distance function d on $\mathcal{H}$, a target grammar $\mathcal{G}$ and a text τ for $\mathcal{L}_{\mathcal{G}}$, a learning algorithm $\mathcal{A}$ identifies $\mathcal{G}$ if

$$\lim_{k\to\infty} d(\mathcal{A}(\tau_k), \mathcal{G}) = 0$$

- given sequence $s = (s_1, \ldots, s_k)$ length $\ell(s) = k$; concatenation $x \circ y = (x_1, \ldots, x_k, y_1, \ldots, y_m)$

• Fact: if $\mathcal{A}$ identifies $\mathcal{G}$ then for all $\epsilon > 0$ there is a locking data set $\ell_\epsilon \subset \mathcal{D}$ with $\ell_\epsilon \subset \mathcal{L}_\mathcal{G}$ and $d(\mathcal{A}(\ell_\epsilon), \mathcal{G}) < \epsilon$ and

$$d(\mathcal{A}(\ell_\epsilon \circ x), \mathcal{G}) < \epsilon, \quad \forall x \in \mathcal{D} \cap \mathcal{L}_\mathcal{G}$$

• meaning: after encountering locking data, learner will remain ϵ-close to target with any additional input data

• argument: if no locking data set exists, for any ℓ there will be some x with $d(\mathcal{A}(\ell \circ x), \mathcal{G}) \geq \epsilon$... this can be used to construct a text τ for $\mathcal{L}$ on which $\mathcal{A}$ does not identify target $\mathcal{G}$:
- start with a given text $\rho = s_1, s_2, \ldots s_N, \ldots$... construct new one τ: set $\tau_1 = s_1$
- if $d(\mathcal{A}(\tau_1), \mathcal{G}) \geq \epsilon$ take $\tau_2 = \tau_1 \circ s_2$
- if $d(\mathcal{A}(\tau_1), \mathcal{G}) < \epsilon$ take the x such that $d(\mathcal{A}(\tau_1 \circ x), \mathcal{G}) \geq \epsilon$ and set $\tau_2 = \tau_1 \circ x \circ s_2$

- continue: $\tau_{k+1} = \tau_k \circ x_k \circ s_{k+1}$ if $d(\mathcal{A}(\tau_k), \mathcal{G}) < \epsilon$ and $\tau_{k+1} = \tau_k \circ s_k$ if $d(\mathcal{A}(\tau_k), \mathcal{G}) \geq \epsilon$

- valid text because s_i added at each stage

- but ... $\mathcal{A}(\tau_k)$ cannot converge to $\mathcal{G}$ because if at some stage τ_k hypothesis $\mathfrak{h}_k$ is in an ϵ-neighborhood of $\mathcal{G}$, at stage $\tau_k \circ x_k$ hypothesis is outside of ϵ-neighborhood (infinitely often)

- conclusion: if a grammar is learnable, then there is a locking data set that constraints the learner's hypothesis to an ϵ-neighborhood of the target... seems nice, but... it has some undesirable consequences

Unlearnability of Grammars

- take $d(\mathfrak{h},\mathfrak{h}')=0$ if $\mathcal{L}_{\mathfrak{h}}=\mathcal{L}_{\mathfrak{h}'}$ and $d(\mathfrak{h},\mathfrak{h}')=1$ otherwise
take $\epsilon=1/2$

- by previous if $\mathcal{A}$ identifies target grammar $\mathcal{G}$ there is a locking data set $\ell\subset\mathcal{L}_{\mathcal{G}}$ with $d(\mathcal{A}(\ell),\mathcal{G})=0$ and $d(\mathcal{A}(\ell\circ x),\mathcal{G})=0$ for all additional data x in $\mathcal{L}_{\mathcal{G}}$

- Consequence: if $\mathcal{H}$ contains all finite languages and at least one infinite language then $\mathcal{H}$ is not learnable

- argument: use metric as above, suppose learnable with algorithm $\mathcal{A}$, then can identify the infinite language $\mathcal{L}_\infty$ among other, using the (finite) locking set data $\ell_{\mathcal{L}_\infty}$ of length k... consider language made only of $\ell_{\mathcal{L}_\infty}$ (finite language in $\mathcal{H}$), construct text τ for this language with $\tau_k=\ell_{\mathcal{L}_\infty}$... on this text $\mathcal{A}$ converges to $\mathcal{L}_\infty$ hence it does not recognize the finite language from its text

Consequences

- the set of Regular Grammars is unlearnable
- the set of Context-free Grammars is unlearnable
- what if changing the metric? convergence in the 0/1 discrete metric = eventually constant
- this convergence "behaviorally plausible" (right extensional set) but "cognitively implausible" (no intensional model of grammar involved)
- but previous unlearnability result can be extended to other metrics
- criteria for learnability?

Learnability Criterion

• Result: a family $\mathcal{H}$ is learnable iff for all $\mathfrak{h} \in \mathcal{H}$ there is a subset $\mathcal{D}_{\mathfrak{h}} \subset \mathcal{L}_{\mathfrak{h}}$ such that if $\mathfrak{h}' \in \mathcal{H}$ has $\mathcal{D}_{\mathfrak{h}} \subset \mathcal{L}_{\mathfrak{h}'}$ then $\mathcal{L}_{\mathfrak{h}'} \not\subset \mathcal{L}_{\mathfrak{h}}$

• avoids previous problem where lock data set for one language determines another language

• argument:
(1) assume $\mathcal{H}$ learnable then have $\mathcal{A}$ and for $\mathfrak{h}$ a locking data set $\ell_{\mathfrak{h}}$ suppose this belongs to some other language $\ell_{\mathfrak{h}} \subset \mathcal{L}_{\mathfrak{h}'}$ with
$\mathcal{L}_{\mathfrak{h}'} \subsetneq \mathcal{L}_{\mathfrak{h}}$
then can construct a text τ for $\mathcal{L}_{\mathfrak{h}'}$ using $\ell_{\mathfrak{h}}$ with $d(\mathcal{A}(\tau_k), \mathfrak{h}') \not\to 0$
this contradicts learnability

(2) Conversely, assume property in the statement holds and show can construct $\mathcal{A}$ that makes $\mathcal{H}$ learnable
enumerate $\mathcal{H} = \{\mathfrak{h}^{(k)}\}_{k\in\mathbb{N}}$ and take $\mathcal{D}_k = \mathcal{D}_{\mathfrak{h}^{(k)}}$
define $\mathcal{A}$ by procedure:
- given τ_k search in list smallest $i \leq k$ with $\mathcal{D}_i \subset \tau_k \subset \mathcal{L}_{\mathfrak{h}^{(k)}}$
- if none take $\mathfrak{h}^{(1)}$
show this $\mathcal{A}$ identifies all $\mathcal{L}_k = \mathcal{L}_{\mathfrak{h}^{(k)}}$ correctly:
- at τ_k can hypothesize $\mathcal{L}_k$ (correct) or could have chosen some $\mathcal{L}_j$ with $j < k$, need to exclude this possibility
- it cannot hypothesizes $\mathfrak{h}^{(j)}$ with $j < k$ if $\mathcal{L}_j \subset \mathcal{L}_k$ because cannot have $\mathcal{D}_k \subset \mathcal{L}_j$
- if $\mathcal{L}_j \not\subset \mathcal{L}_k$ some sentence s in $\mathcal{L}_k \smallsetminus \mathcal{L}_j$ will appear in some τ_m and after than cannot hypothesize $\mathcal{L}_j$

Probabilistic Learnability

- $\mathcal{G}$ target grammar, measure $\mu = \mu_{\mathcal{G}}$ on $\mathfrak{A}^\star$ with support on $\mathcal{L}_{\mathcal{G}}$
- text τ for $\mathcal{G}$ produced as independent identically distributed random variables according to μ
- almost everywhere learning (with probability one): $\exists \mathcal{A}$ such that

$$\mu_\infty \left(\{ \tau \mid \lim_{n\to\infty} d(\mathcal{A}(\tau_k), \mathcal{G}) = 0 \} \right) = 1$$

where μ_∞ probability measure on the ω-language (Cantor set) determined by measure μ on the cylinder sets

- family $\mathcal{H}$ is probability-one-learnable if all $\mathcal{G}$ in $\mathcal{H}$ is almost everywhere learnable for $\mu = \mu_{\mathcal{G}}$

Recursively Enumerable languages are probabilistically learnable

• Result: with prior knowledge of the probability distributions $\mu_{\mathcal{L}}$ the set $\mathcal{H}$ of recursively enumerable languages is probability-one-learnable

• Comments: knowledge of the measure is needed in the argument (need to know the $d(n)$ = number of examples after which high probability of assigning correct membership)

• a better notion of probabilistic learnability, probability-one-learnable in a distribution-free sense: $\exists \mathcal{A}$ that learns target grammar with measure one for all measures

• ... but in distribution-free sense class of learnable languages same as in non-probabilistic sense, no improvement argument:

- enumeration $\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_m, \ldots$ of all recursively enumerable languages

- choose enumeration $s_1, s_2, \ldots, s_n, \ldots$ of all the finite strings in $\mathfrak{A}^\star$

- string (text) τ in $\mathfrak{A}^\omega$ and a language $\mathcal{L}_k$ agree on membership up to order n if for all $i \leq n$ have $s_i \in \tau$ iff $s_i \in \mathcal{L}_k$

- consider set of all texts for $\mathcal{L}_k$ for which one of the first n elements in $\mathfrak{A}^\star$ is in $\mathcal{L}_k$ but not in τ_m

$$A_{k,n,m} = \{\tau \text{ text for } \mathcal{L}_k \,|\, \exists i \leq n : \; s_i \in \mathcal{L}_k \smallsetminus \tau_m\}$$

- $A_{k,n,m} \supseteq A_{k,n,m+1}$ and $\cap_{m=1}^{\infty} A_{k,n,m} = \emptyset$ so

$$\lim_{m\to\infty} \mu_{\infty,k}(A_{k,n,m}) = 0$$

• number of examples after which high probability of assigning correct membership to s_i for $i \leq n$, if target is some $\mathcal{L}_k$ with $k \leq n$

$$d(n) = \min n \text{ such that } \mu_{\infty,i}(A_{i,n,m}) \leq 2^{-n}, \quad \forall i \leq n$$

monotonically increasing function: eventually identify target language with measure one

• how the learning algorithm $\mathcal{A}$ works:
- given input sequence of length m, find first $n \leq m$ with $d(n) \leq m$
- among languages $\mathcal{L}_1, \ldots, \mathcal{L}_n$ find least integer $k \leq n$ for which $\mathcal{L}_k$ agrees with test sequence up to n (if can't find one take $k = 1$)

• now need to show the set of texts on which $\mathcal{A}$ does not converge to $\mathcal{L}_k$ is of measure zero

$$\mathcal{B} = \{\tau \,|\, \mathcal{A}(\tau_n) \neq \mathcal{L}_k, \text{ for infinitely many } n\}$$

- if τ in B then $\mathcal{A}(\tau_m) \neq \mathcal{L}_k$ infinitely often: it can happen because τ_m and $\mathcal{L}_k$ do not agree through n or because there is some other $\mathcal{L}_j$ with $j < k$ that agrees with τ_m to order n
- can't be second case infinitely often because τ and $\mathcal{L}_j$ eventually disagree... so first case
- consider set

$$X_k = \cap_i \cup_{m>i} A_{k,n(m),m}$$

with $n = n(m)$ the first $n \leq m$ with $d(n) \leq m$

- previous observation implies $B \subset X_k$
- also can check that

$$\cap_i \cup_{m>i} A_{k,n(m),m} \subseteq \cap_i \cup_{n>i} A_{k,n,d(n)}$$

- by construction have finite sum of measures hence

$$\sum_m \mu_{\infty,k}(A_{k,n,d(n)}) < \infty \quad \Rightarrow \quad \mu_{\infty,k}(X_k) = 0$$

Borel–Cantelli lemma: $\sum_n \mathbb{P}(Y_n) < \infty \Rightarrow \mathbb{P}(\cap_n \cup_{k \geq n} Y_n) = 0$

Other notions of learnability

- active learner: learner can make queries about membership of arbitrary elements $s \in \mathfrak{A}^\star$; then regular languages are learnable (in polynomial time) but context-free remain unlearnable

- recursive texts: τ such that $\{\tau_n, n \in \mathbb{N}\}$ is a recursive set, algorithm should converge to target language on recursive set; then Phrase Structure Grammars are learnable, but $\mathcal{A}$ is not a computable function

- informant texts: text τ contains both positive and negative examples, all $s \in \mathfrak{A}^\star$ in the text with label for belonging to $\mathcal{L}$ or not; then all recursively enumerable languages are learnable

- observations on language learning in children suggests mostly positive examples though

- learning with mistakes: learning target language up to k mistakes; this gives a hierarchy of learnable languages increasing with k

References

- Partha Niyogi, *The Computational Nature of Language Learning and Evolution*, MIT Press, 2006
- M. Blum, L. Blum, *Towards a mathematical theory of inductive inference*, Information and Control, 28 (1975) 125–155
- E. Mark Gold, *Language identification in the limit*, Information and Control, 10 (1967) N.5, 447–474.
- Dana Angluin, *Inductive inference of formal languages from positive data*, Information and Control, 45 (1980) 117–135.