Research seminar week 8

Tamás Biró
Humanities Computing
University of Groningen
t.s.biro@rug.nl

This week:

- Background in learning theory
- Niyogi, chapter 2 (3 and 4).

Framework for Learning

- Finite $alphabet\ \Sigma$.
- $Language\ L \subset \Sigma^*$.
- Grammar g, generating language L_g.
- Family of grammars $\mathcal{L}$,
  family of languages $\mathcal{L} = \{L_g | g \in \mathcal{G}\}$.

NB: Language not an input-output mapping now.

Framework for Learning

- *Example sentences* s_i: from target L_t.
- Set of possible example sentence sequences $\mathcal{D} = (\Sigma^*)^*$.
- *Hypothesis languages* $h \in \mathcal{H}$.
- Learning algorithm $\mathcal{A}$: effective procedure $\mathcal{A} : \mathcal{D} \rightarrow \mathcal{H}$.

Framework for Learning

- $d(\cdot, \cdot)$: distance of grammars/languages.
- Criterion of success:

$$\lim_{n \to \infty} d(g_t, h_n) = 0$$

where $h_n = \mathcal{A}(s_1, ..., s_n)$.

Approaches to Learning

- Inductive inference / identification in the limit / Gold-learning.
- Probably Approximately Correct (PAC) learning

Identification in the limit

- *Text* t for language L: infinite sequence $s_1, ..., s_n, ...$, s.t. each $s_i \in L$, and all elements of L appears at least once in t.
- t_k: first k elements of text t.
- $t(k) = s_k$.

Identification in the limit

- Learning algorithm $\mathcal{A}$ identifies (learns) target g_t on text t *in the limit*, if $\lim_{k\to\infty} d(\mathcal{A}(t_k), g_t) = 0$.
- $calA$ identifies g_t in the limit, if it identifies g_t in the limit for all texts of L_{g_t}.
- Family $\mathcal{G}$ is *identifiable in the limit* if there is an algorithm $\mathcal{A}$ that identifies every $g \in \mathcal{G}$ in the limit.

Identification in the limit

- Gold's theorem (1967): family consisting of all finite languages and at least one infinite language is not learnable in the limit.
- Not learnable: regular languages; context free languages; infinite regular languages.
- Poverty of Stimulus; nativist arguments.

PAC Learning

Probably Approximately Correct Learning

(Vapnik and Chervonenkis 1971)

- "Probably": on "almost every" sequences of data.
- "Approximately correct": the algorithm gets close enough to target.

PAC Learning

Results:

- PAC unlearnable: all finite languages; regular languages; context free languages.

Complexity of learning

- Speed of convergence.