Agreement-Based Learning

P. Liang, D. Klein, M. I. Jordan

NIPS 2007

Discussion Notes written by
Xuejun Liao
08 August 2008
Assumptions

Suppose we have multiple sub-models \( p_m(x, z; \theta_m) \),
\( m = 1, 2, \ldots, M \)

- \( x \) are observed while \( z \) are unobserved
- the \( M \) sub-models have a common domain \( \mathcal{X} \times \mathcal{Z} \) (Notes: should also have the same definition for \( z \)).
- but they can be parameterized differently

Each sub-model is \textbf{tractable} and captures the data from a particular perspective
Independent versus Agreement-Based Learning

Independent learning, maximizing

\[ \ell_{\text{indep}}(\theta) = \log \prod_m \sum_z p_m(x, z; \theta_m) = \sum_m \log p_m(x; \theta_m) \]

Agreement-Based learning

\[ \ell_{\text{agree}}(\theta) = \log \sum_z \prod_m p_m(x, z; \theta_m) \]
\[ = \sum_m \log p_m(x; \theta_m) + \log \sum_z \prod_m p_m(z|x; \theta_m) \]
\[ = \ell_{\text{indep}}(\theta) + \text{agreement} \]

Notes: \( \sum_z \prod_m p_m(x, z; \theta_m) \) is an improper probability density because it is not normalized over \( x \). Since \( \sum_x \sum_z \prod_m p_m(x, z; \theta_m) \leq \prod_m \sum_x \sum_z p_m(x, z; \theta_m) = 1 \), \( \ell_{\text{agree}} \) is a lower bound to the normalized log-likelihood.
Suppose we have a joint model, \( p(x, z; \theta_1, \cdots, \theta_M) \), which cannot be (accurately) decomposed into sub-models. One learns by maximizing

\[
\ell_{\text{joint}}(\theta) = \log \sum_z p(x, z; \theta_1, \cdots, \theta_M) = \log p(x; \theta_1, \cdots, \theta_M)
\]

One may approximate

\[
p(x, z; \theta_1, \cdots, \theta_M) \approx \prod_m p_m(x, z; \theta_m)
\]

Independent learning completely ignores the correlations between the sub-models.

Agreement-based learning (partially) preserves the correlations through the agreement term.
Example I: Unsupervised Word Alignment (1/2)

Given \( x = (e, f) \), where \( e = (e_1, \cdots, e_{|e|}) \) is a sentence in the source language (e.g., English) and \( f = (f_1, \cdots, f_{|f|}) \) is a sentence in the target language (e.g., French)

The goal is to establish the correspondence between source words and target words.

(a) Submodel \( p_1 \)

(b) Submodel \( p_2 \)
Define two sub-models

- Let $z = a = (a_1, \cdots, a_{|f|})$, such that $f_j \leftarrow e_{a_j}$

$$p_1(x, z; \theta_1) = p_1(e, f, a; \theta_1) = p_1(e) \prod_{j=1}^{|f|} p_1(a_j)p(f_j|e_{a_j}; \theta_1)$$

- Let $z = a = (a_1, \cdots, a_{|e|})$, such that $f_{a_j} \rightarrow e_j$

$$p_2(x, z; \theta_2) = p_2(e, f, a; \theta_2) = p_2(f) \prod_{j=1}^{|e|} p_2(a_j)p(e_j|f_{a_j}; \theta_2)$$
Example II: Phylogenetic HMM Models (1/2)

Given a set of species $S$ in a fixed phylogeny, each $s \in S$ associated with a sequence of nucleotides $d_s = (d_{s1}, \cdots, d_{sL})$

Let $d = \{d_s : s \in S\}$ denote all such sequences, some of which are observed and other are unobserved

The goal is to predict the unobserved nucleotides

(a) Submodel $p_1$

(b) Submodel $p_2$
Example II: Phylogenetic HMM Models (2/2)

- The first sub-model

\[ p_1(x, z; \theta_1) = p_1(d; \theta_1) = \prod \prod \prod p_1(d_{s'j} | d_{sj}; \theta_1)p_1(d_{s',j+1} | d_{s'j}, d_{s,j+1}; \theta_1) \]

- The second sub-model

\[ p_2(x, z; \theta_2) = p_2(d; \theta_2) = \prod \prod \prod p_2(d_{s'j} | d_{sj}; \theta_2)p_2(d_{s',j+1} | d_{s'j}, d_{s,j+1}; \theta_2) \]
Product EM

- Lower bound: \( \ell_{\text{agree}}(\theta) \geq \mathbb{E}_{q(z)} \log \frac{\prod_{m} p_{m}(x, z; \theta_{m})}{q(z)} \)

- Joint E-step: 
  \[
  q(z) = \frac{\prod_{m} p_{m}(x, z; \theta_{m})}{\sum_{z} \prod_{m} p_{m}(x, z; \theta_{m})} \neq \prod_{m} p_{m}(z|x; \theta_{m})
  \]

- Independent M-steps:
  \[
  \theta_{\text{new}} = \arg \max_{\theta_{\text{new}}} \mathbb{E}_{q(z)} \log \prod_{m} p_{m}(x, z; \theta_{\text{new}}^{m})
  \]
  \[
  = \arg \max_{\theta_{\text{new}}} \sum_{m} \mathbb{E}_{q(z)} \log p_{m}(x, z; \theta_{\text{new}}^{m})
  \]
  equivalently, \( \theta_{m}^{\text{new}} = \arg \max_{\theta_{m}^{\text{new}}} \mathbb{E}_{q(z)} \log p_{m}(x, z; \theta_{m}^{\text{new}}), \ m = 1, \cdots, M \)
The experiments used 100K unaligned sentences for training and 137 for testing from the English-French Hansards data of the NAACL 2003 Shared Task. Alignments are evaluated using alignment error rate (AER). The results are as shown below.
Results on Nucleotides Prediction

The experiments used a multiple alignment consisting of $L = 20,000$ consecutive sites belonging to the L1 transposons in the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) gene (chromosome 7). Eight eutherian species were arranged in the phylogeny shown in the left below.

The results are as shown in the right below.