Probabilistic Deterministic Infinite Automata

David Pfau, Nicholas Bartlett, Frank Wood
{pfau@neurotheory, bartlett@stat, fwood@stat}.columbia.edu

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

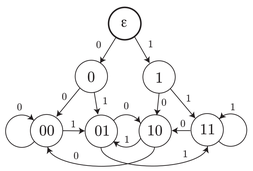
Overview

A Probabilistic Deterministic Infinite Automata (PDIA) is a Probabilistic Deterministic Finite Automata (PDFA)[1] with an unbounded number of states. We take a nonparametric Bayesian approach to PDIA inference.

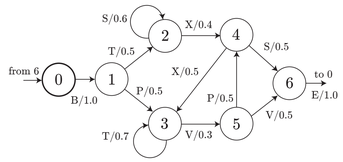
PDIA are an attractive compromise between the computational costs of Hidden Markov Models and the storage requirements of smoothed Markov models for predicting sequence data.

Finite Automata

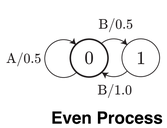
Probabilistic Deterministic Finite Automata



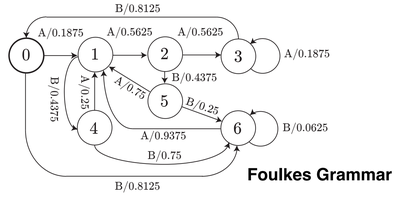
Trigram as DFA (without probability)



Reber Grammar

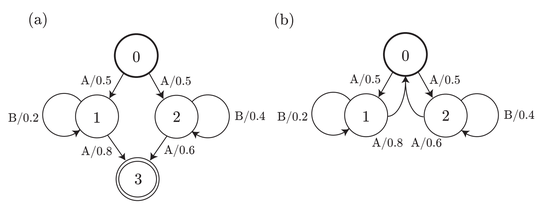


Even Process



Foulkes Grammar

Probabilistic Nondeterministic Finite Automata



(a) in mixtures of PDFA but not PDFA
(b) in PNFA but not mixtures of PDFA

Given a finite training sequence, the posterior distribution over PDIA parameters is an infinite mixture of PDFAs.

Samples for this distribution drawn via MCMC form a finite mixture approximating this posterior.

The class of models we deal with, from least to most general -

n*th*-order Markov ⊊ PDFA ⊊ finite mixture of PDFA ⊊ PNFA = HMM

(Technically, HMM = PNFA without final state)

Generative Model

A Prior over PDFA with a bounded number of states

$$\delta$$

	σ_0	σ_1	σ_2
q_0	q_3	q_1	q_{12}
q_1	•	q_3	•
q_2	q_6	q_4	•
q_3	q_3	•	q_5
q_4	q_3	q_{42}	q_5
q_5	•	q_1	q_2
⋮	⋮	⋮	⋮

$$\pi$$

	σ_0	σ_1	σ_2
q_0	(iid probability vector)		
q_1	(iid probability vector)		
⋮	⋮	⋮	⋮

$$\frac{q_0\ q_1\ q_3\ q_3\ q_3\ q_5\ q_2\ q_4\ q_5\ q_2\ \cdots}{1\ 1\ 0\ 0\ 2\ 2\ 1\ 2\ 1\ 0\ \ldots}$$

$$\mu \sim \mathrm{Dir}(\alpha_0/|Q|)$$
$$\phi_j \sim \mathrm{Dir}(\alpha\mu) \qquad j = 0 \ldots |\Sigma| - 1$$
$$\delta(q_i, \sigma_j) = \delta_{ij} \sim \phi_j \qquad i = 0 \ldots |Q| - 1$$
$$\pi_{q_i} \sim \mathrm{Dir}(\beta/|\Sigma|) \qquad i = 0 \ldots |Q| - 1$$
$$\xi_0 = q_0,\ \xi_t = \delta(\xi_{t-1}, x_{t-1})$$
$$x_t \sim \pi_{\xi_t}$$

Notation:

- Q - finite set of states
- Σ - finite alphabet
- $\delta : Q \times \Sigma \to Q$ - transition function
- $\pi : Q \times \Sigma \to [0, 1]$ - emission probabilities
- $q_0 \in Q$ - initial state
- $x_t \in \Sigma$ - data at time t
- $\xi_t \in Q$ - state at time t
- $\alpha, \alpha 0, \beta$ - hyperparameters

This prior biases towards state reuse in two ways: the global bias toward some states due to μ and the column-specific bias towards certain states when the same symbol is emitted due to ϕ_j.

A Prior over infinite models

The limit as $|Q| \to \infty$ is a PDIA prior. The two-level Dirichlet construction of the finite model becomes a Hierarchical Dirichlet Process (HDP)[2] which we give geometric base distribution. When μ and ϕ_j are marginalized out, δ_{ij} are exchangeable.

Inference

MCMC sampler for $\delta_{ij} | \delta_{-ij}, x_{0:T}, \alpha, \alpha_0, \beta$

Marginalizing out π in likelihood gives form that only depends on counts

$$p(x_{0:T} | \delta, c, \beta) = \prod_{i=0}^{|Q|-1} \frac{\Gamma(\beta)}{\Gamma(\frac{\beta}{|\Sigma|})^{|\Sigma|}} \frac{\prod_{j=1}^{|\Sigma|} \Gamma(\frac{\beta}{|\Sigma|} + c_{ij})}{\Gamma(\beta + \sum_{j=1}^{|\Sigma|} c_{ij})} \qquad c_{ij} = \sum_{t=0}^{T} \mathbf{1}_{ij}(\xi_t, x_t)$$

Propose δ^*_{ij} from $\delta^*_{ij} | \delta_{-ij}, \alpha, \alpha_0$, easy if $|Q| < \infty$, use *Chinese Restaurant Franchise Process*[2] if $|Q| = \infty$

Accept with probability $\min\left(1, \frac{p(x_{0:T} | \delta^*_{ij}, \delta_{-ij})}{p(x_{0:T} | \delta_{ij}, \delta_{-ij})}\right)$

If $c_{ij} = 0$, δ_{ij} can be ignored

Natural Language, DNA Prediction

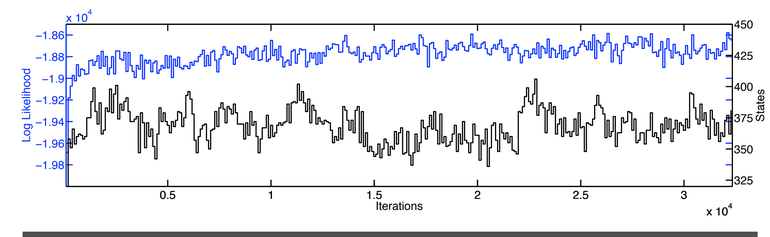
	PDIA	PDIA-MAP	HMM-EM	bigram	trigram	4-gram	5-gram	6-gram	SSM
AIW	5.13	5.46	7.89	9.71	6.45	5.13	4.80	4.69	4.78
	365.6	379	52	28	382	2,023	5,592	10,838	19,358
DNA	3.72	3.72	3.76	3.77	3.75	3.74	3.73	3.72	3.56
	64.7	54	19	5	21	85	341	1,365	314,166

Top rows: perplexity of held out data. Bottom rows: number of states

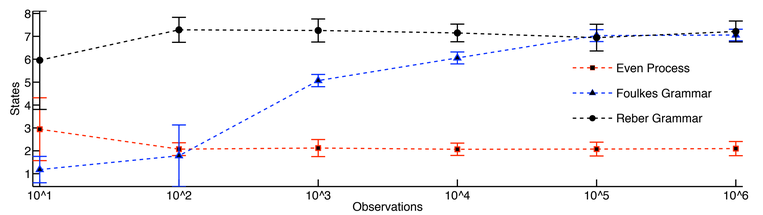
Predictions from average of many samples better than MAP sample

Low model complexity relative to other models with the same predictive performace

Alice in Wonderland:10k training characters, 4k test *"alice was beginning to..."*
Mouse DNA: 150k train, 50k test *"CGTATATGCGCC..."*
Controls: EM-trained HMM, HPYP smoothed n-gram[3], sequentially-trained sequence memoizer[4]



Synthetic Grammar Induction



Future Directions

Evaluation on larger data sets
Sampler that proposes "similar" PDFA - spilt-merge sampling?
How to smooth emission distributions?

References

[1] M.O. Rabin. Probabilistic Automata. *Information and Control*, 6(3):230-245, 1963
[2] Y. W. Teh, M. I. Jordan, M. J. Beal, D. M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*. 101(476):1566-1581, 2006
[3] Y. W. Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. *Proceedings of the ACL*. 985-992, 2006
[4] F. Wood, C. Archambeau, J. Gasthaus, L. James, Y. W. Teh. A Stochastic Memoizer for Sequence Data. *Proceedings of the 26th ICML*. 1129-1136, 2009

This research was supported by the NSF GRFP