A Spectral Learning Algorithm for Finite State Transducers

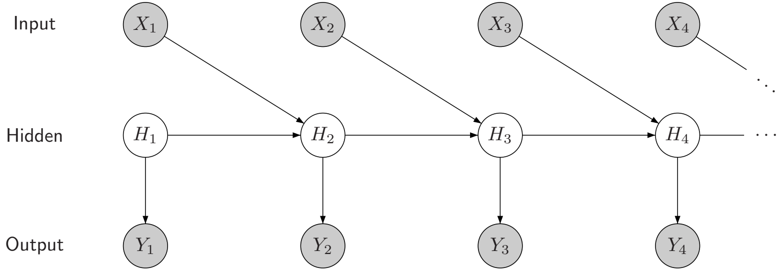
Borja Balle, Ariadna Quattoni, Xavier Carreras

LARCA. Laboratory for Relational Algorithmics, Complexity and Learning
UNIVERSITAT POLITÈCNICA DE CATALUNYA

ECML PKDD — September 7, 2011

Probabilistic Transducers

- Model input-output relations with hidden states
- As conditional distribution $\Pr[\, y \mid x \,]$ over strings
- With certain independence assumptions



- Used in many applications: NLP, biology, ...
- Hard to learn in general — usually EM algorithm is used

Spectral Learning Probabilistic Transducers

Our contribution:

- Fast learning algorithm for probabilistic FST
- With PAC-style theoretical guarantees
- Based on Observable Operator Model for FST
- Using spectral methods (Chang '96, Mossel-Roch '05, Hsu et al. '09, Siddiqi et al. '10)
- Performing better than EM in experiments with real data

Outline

Deriving Observable Operator Models

Given $(x, y) \in (\mathcal{X} \times \mathcal{Y})^t$ aligned sequences, model computes conditional probability (i.e. $|x| = |y|$)

$$
\begin{aligned}
\Pr[\, y \,|\, x \,] &= \textstyle\sum_{h \in \mathcal{H}^t} \Pr[\, y, h \,|\, x \,] && \text{(marginalize states)} \\
&= \textstyle\sum_{h_{t+1} \in \mathcal{H}} \Pr[\, y, h_{t+1} \,|\, x \,] && \text{(independence assumptions)} \\
&= 1^\top \alpha_{t+1} && \text{(vector form, } \alpha_{t+1} \in \mathbb{R}^m\text{)} \\
&= 1^\top A_{x_t}^{y_t} \alpha_t && \text{(forward-backward equations)} \\
&= 1^\top A_{x_t}^{y_t} \cdots A_{x_1}^{y_1} \alpha && \text{(induction on } t\text{)}
\end{aligned}
$$

The choice of an operator A_a^b depends only on observable symbols

Observable Operator Model Parameters

Given $\mathcal{X} = \{a_1, \dots, a_k\}$, $\mathcal{Y} = \{b_1, \dots, b_l\}$, $\mathcal{H} = \{c_1, \dots, c_m\}$, then

$\Pr[\, y \,|\, x \,] = 1^\top A_{x_t}^{y_t} \cdots A_{x_1}^{y_1} \alpha$ with parameters:

$$A_a^b = T_a \, D_b \in \mathbb{R}^{m \times m} \quad \text{(factorized operator)}$$

$$T_a(i,j) = \Pr[H_s = c_i | X_{s-1} = a, H_{s-1} = c_j] \in \mathbb{R}^{m \times m} \quad \text{(state transition)}$$

$$D_b(i,j) = \delta_{i,j} \Pr[Y_s = b | H_s = c_j] \in \mathbb{R}^{m \times m} \quad \text{(observation emission)}$$

$$O(i,j) = \Pr[Y_s = b_i | H_s = c_j] \in \mathbb{R}^{l \times m} \quad \text{(collected emissions)}$$

$$\alpha(i) = \Pr[H_1 = c_i] \in \mathbb{R}^m \quad \text{(initial probabilites)}$$

The choice of an operator A_a^b depends only on observable symbols ...

... but operator parameters are conditioned by hidden states

A Learnable Set of Observable Operators

Note that for any invertible $Q \in \mathbb{R}^{m \times m}$

$$\Pr[\, y \mid x \,] = 1^\top Q^{-1} \, (Q \, A_{x_t}^{y_t} \, Q^{-1}) \cdots (Q \, A_{x_1}^{y_1} \, Q^{-1}) \, Q \, \alpha$$

Idea

(*subspace identification methods for linear systems, '80s*)

Find a basis for the state space such that operators in the new basis are related to observable quantities

Following multiplicity automata and spectral HMM learning . . .

A Learnable Set of Observable Operators

Find a basis Q where operators can be expressed in terms of unigram, bigram and trigram probabilities

$$\rho(i) = \Pr[Y_1 = b_i] \in \mathbb{R}^I$$
$$P(i,j) = \Pr[Y_1 = b_j, Y_2 = b_i] \in \mathbb{R}^{I \times I}$$
$$P_a^b(i,j) = \Pr[Y_1 = b_j, Y_2 = b, Y_3 = b_i | X_2 = a] \in \mathbb{R}^{I \times I}$$

Theorem (ρ, P and P_a^b are sufficient statistics)

Let $P = U \Sigma V^*$ be a thin SVD decomposition, then $Q = U^\top O$ yields (under certain assumptions)

$$Q\,\alpha = U^\top \rho$$
$$1^\top Q^{-1} = \rho^\top (U^\top P)^+$$
$$Q\,A_a^b\,Q^{-1} = (U^\top P_a^b)(U^\top P)^+$$

Spectral Learning Algorithm

Given

- Input $\mathcal{X}$ and output $\mathcal{Y}$ alphabet
- Number of hidden states m
- Training sample $S = \{(x^1, y^1), \ldots, (x^n, y^n)\}$

Do

- Compute unigram $\widehat{\rho}$, bigram $\widehat{P}$ and trigram $\widehat{P}_a^b$ relative frequencies in S
- Perform SVD on $\widehat{P}$ and take $\widehat{U}$ with top m left singular vectors
- Return operators computed using $\widehat{\rho}$, $\widehat{P}$, $\widehat{P}_a^b$ and $\widehat{U}$

In Time

- $O(n)$ to compute relative frequencies
- $O(|\mathcal{Y}|^3)$ to compute SVD

PAC-Style Result

- Input distribution D_X over $\mathcal{X}^*$ with $\lambda = \mathrm{E}[|X|]$, $\mu = \min_a \Pr[X_2 = a]$
- Conditional distributions $D_{Y|x}$ on $\mathcal{Y}^*$ given $x \in \mathcal{X}^*$ modeled by an FST with m states (satisfying certain rank assumptions)
- Sampling i.i.d. from joint distribution $D_X \otimes D_{Y|X}$

Theorem

For any $0 < \varepsilon, \delta < 1$, if the algorithm receives a sample of size

$$n \geq O\left(\frac{\lambda^2 m |\mathcal{Y}|}{\varepsilon^4 \mu \sigma_O^2 \sigma_P^4} \log \frac{|\mathcal{X}|}{\delta}\right) \quad ,$$

(σ_O and σ_P are mth singular values of O and P in target)

then with probability at least $1 - \delta$ the hypothesis $\widehat{D}_{Y|x}$ satisfies

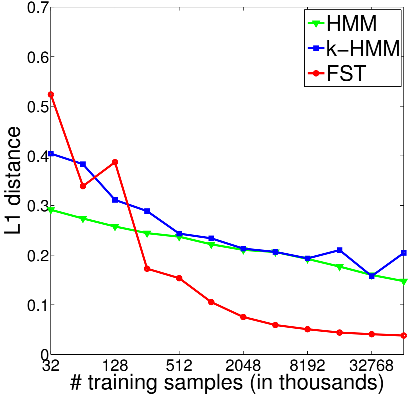
$$\mathrm{E}_X\left[\sum_{y \in \mathcal{Y}^*} \left| D_{Y|X}(y) - \widehat{D}_{Y|X}(y) \right| \right] \leq \varepsilon \quad .$$

(L_1 distance between joint distributions $D_X \otimes D_{Y|X}$ and $D_X \otimes \widehat{D}_{Y|X}$)

Synthetic Experiments

Goal: Compare against baselines when learning hypothesis hold

Target: Randomly generated with $|\mathcal{X}| = 3$, $|\mathcal{Y}| = 3$, $|\mathcal{H}| = 2$

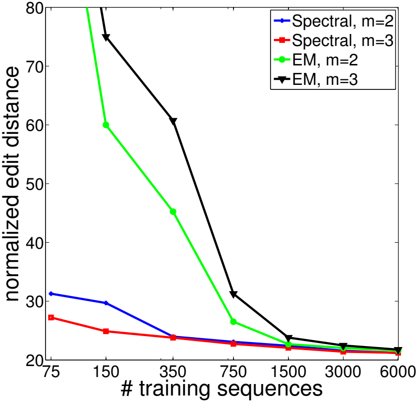


- HMM: model input-output jointly
- k-HMM: one model for each input symbol
- Results averaged over 5 runs

Transliteration Experiments

Goal: Compare against EM in a real task (where modeling assumptions fail)

Task: English to Russian transliteration (brooklyn $\rightarrow$ бруклин)



Training times	
Spectral	26 s
EM (iteration)	37 s
EM (best)	1133 s

- Sequence alignment done in preprocessing
- Standard techniques used for inference
- Test size: 943, $|\mathcal{X}| = 82$, $|\mathcal{Y}| = 34$

Summary of Contributions

- Fast spectral method for learning input-output OOM
- Strong theoretical guarantees with few assumptions on input distribution
- Outperforming previous spectral algorithms on FST
- Faster and better than EM in some real tasks

A Spectral Learning Algorithm for Finite State Transducers

Borja Balle, Ariadna Quattoni, Xavier Carreras

LARCA. Laboratory for Relational Algorithmics, Complexity and Learning
UNIVERSITAT POLITÈCNICA DE CATALUNYA

ECML PKDD — September 7, 2011

Technical Assumptions

$\mathcal{X} = \{a_1, \ldots, a_k\}, \mathcal{Y} = \{b_1, \ldots, b_l\}, \mathcal{H} = \{c_1, \ldots, c_m\}$

Parameters

$$T_a(i,j) = \Pr[H_s = c_i | X_{s-1} = a, H_{s-1} = c_j] \in \mathbb{R}^{m \times m} \quad \text{(state transition)}$$

$$T = \sum_a T_a \Pr[X_1 = a] \in \mathbb{R}^{m \times m} \quad \text{(``mean'' transition matrix)}$$

$$O(i,j) = \Pr[Y_s = b_i | H_s = c_j] \in \mathbb{R}^{l \times m} \quad \text{(collected emissions)}$$

$$\alpha(i) = \Pr[H_1 = c_i] \in \mathbb{R}^{m} \quad \text{(initial probabilites)}$$

Assumptions

1. $l \geq m$
2. $\alpha > 0$
3. $\mathrm{rank}(T) = \mathrm{rank}(O) = m$
4. $\min_a \Pr[X_2 = a] > 0$