Spectral Learning of Sequence Taggers over Continuos Sequences

Adrià Recasens and Ariadna Quattoni
Universitat Politècnica de Catalunya

— Tagging Continuous Sequences —

- Examples: Gesture Recognition, Robot Navigation.
- Setting: We are given aligned sequences $< [x_1 \dots x_T], [y_1 \dots y_T] >$.
- $x_i \in \mathbb{R}^k$ and $y_i \in \Sigma$, for some discrete set of Tags.
- Goal: Learn a model of $\mathbb{P}(x, y)$ and use it to make predictions, i.e. compute $\mathrm{argmax}_y \mathbb{P}(x, y)$
- Contribution: A Spectral Algorithm for this task.

— Spectral Background —

HMMs [HKZ09]

- m states – $S_t \in \{1, \dots, m\}$
- k symbols – $x_t \in \{\sigma_1, \dots, \sigma_k\}$
- Forward-backward equations with $A_\sigma \in \mathbb{R}^{m \times m}$:

$$\mathbb{P}(x) = \alpha_1^\top A_{x_1} \cdots A_{x_t} \vec{1}$$

- Observable statistics:

$$H(i,j) = \mathbb{P}(x_{t-1} = \sigma_i,\ x_t = \sigma_j)$$
$$H_\sigma(i,j) = \mathbb{P}(x_{t-1} = \sigma_i,\ x_t = \sigma,\ x_{t+1} = \sigma_j)$$

- Algorithm: Compute SVD $H = UDV^\top$ and take top m right singular vectors V_m $A_\sigma = (HV_m)^+ H_\sigma V_m$

Finite State Taggers (FST)

- Input alphabet Δ and output alphabet Σ
- Operators: $A_\delta^\sigma \in \mathbb{R}^{m \times m}$, depend on input and output.
- $\mathbb{P}(x, y) = \alpha_1^\top A_{x_1}^{y_1} \cdots A_{x_T}^{y_T} \alpha_\infty$
- Algorithm: Balle et al, [ECML 2011].

— Continuous Sequence Taggers (CFST) —

- A *CFST* over $(\Phi(\mathcal{X}) \times \Sigma)^*$ with m states is a tuple: $A = \langle \Phi, \alpha_1, \alpha_\infty, O_l^\sigma \rangle$
- Φ is a set of k feature functions: $\phi_l(x) : \mathcal{X} \to \mathbb{R}$
- $O_l^\sigma \in \mathbb{R}^{m \times m}$ are the $k \times |\Sigma|$ operators and $A(\Phi(x_t), y_t) = \sum_{l=1}^k \phi_l(x_t) O_l^{y_t}$
- The function f_A realised by the CFST is defined by:

$$f_A(x, y) = \alpha_1^\top\ A(\Phi(x_1), y_1) \cdots A(\Phi(x_T), y_T)\ \alpha_\infty$$

— Example: Transitions as Mixture Models —

- $\mathbb{P}(x, y) = \sum_h \mathbb{P}(h_0) \prod_{t=1}^{T-1} \mathbb{P}(h_{t+1}, x_t, y_t \mid h_t)$
- $\mathbb{P}(h_{t+1}, x_t, y_t \mid h_t) = \sum_{l=1}^k \mathbb{P}_l(h_{t+1}, y_t \mid h_t) \mathbb{P}(z = l, x_t)$
- $\phi(x) = [\mathbb{P}(z = 1, x) \dots \mathbb{P}(z = k, x)]$
- $O_l^y = \mathbb{P}_l(h_{t+1}, y \mid h_t)$

Abstract

1. We address the problem of sequence tagging where the input is a continuous sequence and the output is discrete.
2. We generlizate the class of FSTs over discrete input-output sequences to a class where transitions are linear combinations of elementary transitions.

1. Intuitively, the atomic transition functions operate on a soft partition of the input space.
2. We derive a spectral learning algorithm for this model that is both simple and fast.

Observable Statistics:

- $H_1(i) = \mathbb{E}_\mathbb{P}[\phi_i(x_t)]$ — Input unigram expectations.
- $H_2(i,j) = \mathbb{E}_\mathbb{P}[\phi_i(x_t)\phi_j(x_{t+1})]$ — Input bigram expectations.
- $H_l^\sigma(i,j) = \mathbb{E}_{\mathbb{P}y_t=\sigma}(\phi_i(x_{t-1})\phi_l(x_t)\phi_j(x_{t+1}))$
  Input trigram expectations conditioned on y_t.
- $C(i,j) = \mathbb{E}_\mathbb{P}[\phi_i(x_t)\phi_j(x_t)]$ — Covariance.

— Duality: CFST and factorizations of H_2 —

Theorem: Minimal CFST A $\Longleftrightarrow$ Rank factorization of H_2.

Remarks $\Rightarrow$:

- Hypothesis: minimal CFST.
- H_2 can be wrote as $H_2 = FB$.
- $H_l^\sigma = F \sum_{i=1}^k O_i^\sigma C(l,i) B$ and $H_1 = F\alpha_\infty = \alpha_1^\top B$

Remarks $\Leftarrow$:

- Hypothesis: $H_2 = FB$, a rank factorization.
- $A = \langle \Phi, \alpha_1, \alpha_\infty, O_l^\sigma \rangle$ can be defined as:

$$\alpha_\infty = F^+ H_1 \quad \alpha_1^\top = H_1 B^+ \quad Q_l^\sigma = F^+ H_l^\sigma B^+$$
$$[O_1^\sigma(i,j), \dots, O_k^\sigma(i,j)]^\top = C^{-1}[Q_1^\sigma(i,j), \dots, Q_k^\sigma(i,j)]$$

— The Algorithm —

Algorithm `LearnCWFST`$(\mathcal{X}, \Phi, \Sigma, S, m)$

1. For every pair of sequences (x, y) in S and every index $1 < t < |x|$compute $\phi(x_t) = [\phi_1(x_t), \dots, \phi_k(x_t)]$
2. Use S to estimate matrix statistics $\widehat{H}_1 \in \mathbb{R}^k$, $\widehat{H}_2 \in \mathbb{R}^{k \times k}$, $\widehat{H}_l^\sigma \in \mathbb{R}^{k \times k}$ and covariance matrix $\widehat{C} \in \mathbb{R}^{k \times k}$.
3. Compute the m rank compact SVD of $\widehat{H}_2 = (U\Lambda)V^\top$.
4. Compute the inverse of $\widehat{C}$ and $Q_l^\sigma = (\widehat{H}_2 V)^+ \widehat{H}_l^\sigma V$.
5. Compute the start and ending parameters of the CWFST as: $\alpha_1^\top = \widehat{H}_1 V \quad \alpha_\infty = (\widehat{H}_2 V)^+ \widehat{H}_1$
6. Compute the transition matrices O_l^σ:

$$\begin{bmatrix} O_1^\sigma(i,j) \\ \vdots \\ O_k^\sigma(i,j) \end{bmatrix} = \widehat{C}^{-1} \begin{bmatrix} Q_1^\sigma(i,j) \\ \vdots \\ Q_k^\sigma(i,j) \end{bmatrix}$$

— Experimental Results —

Task Robot Navegation:

- Input: Sequence of Sensor Readings.
- Output: Sequence of Optimal Actions.

Features:

- Select $\{z_1 \dots z_k\}$ points in $\mathbb{R}^k$ (e.g. via kmeans)
- Define $\phi_l(x) = \exp^{\frac{-D(z_l, x)}{\tau}}$ for some distance function D.

Inference:

- Max marginals.

Compare:

- FST spectral learning (discretized inputs).
- Different Feature Functions.