Multitask Spectral Learning of Weighted Finite Automata

Guillaume Rabusseau[1] Borja Balle[2] Joelle Pineau[1,3]

[1]McGill University [2]Amazon Research Cambridge [3]CIFAR

Overview

- Sequence data is ubiquitous in computer science and machine learning.

- Weighted Finite Automata (WFA) can model functions on sequences.
- We propose a spectral multitask learning algorithm for WFAs:
  - extends the spectral learning algorithm for WFAs [1]
  - relies on the novel model of vector-valued WFA.

Multitask Learning

- Common task in machine learning: estimate a function $f : \mathcal{X} \to \mathcal{Y}$ from a training sample $\{(x_i, y_i)\}_{i=1}^N$ where each $y_i \simeq f(x_i)$.
- In multitask learning, the learner is given several learning tasks $f_1, \cdots, f_m$.
- Jointly learning related tasks $f_1, \cdots, f_m$ can lead to better performances.
- This work: multitask learning when $\mathcal{X}$ consists of *sequence data*.

Weighted Finite Automata

- A weighted finite automaton (WFA) is a tuple $A = (\boldsymbol{\alpha}, \{\mathbf{A}^\sigma\}_{\sigma\in\Sigma}, \boldsymbol{\omega})$ and computes a function $f_A : \Sigma^* \to \mathbb{R}$ defined for each word $x = x_1 x_2 \cdots x_k \in \Sigma^*$ by

$$f_A(x_1 x_2 \cdots x_k) = \boldsymbol{\alpha}^\top \mathbf{A}^{x_1}\mathbf{A}^{x_2}\cdots\mathbf{A}^{x_k}\boldsymbol{\omega} = \boldsymbol{\alpha}^\top \mathbf{A}^x \boldsymbol{\omega}.$$

- The number of states of A is the size n of the matrices $\mathbf{A}^\sigma$ and A is minimal if any WFA B such that $f_A = f_B$ has at least n states, in which case n is the rank of the function f.

Spectral Learning of WFAs

- Hankel matrix $\mathbf{H}_f \in \mathbb{R}^{\Sigma^*\times\Sigma^*}$ associated with a function $f : \Sigma^* \to \mathbb{R}$

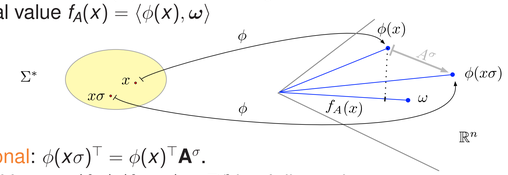
$$(\mathbf{H}_f)_{u,v} = f(uv) \quad \text{for all} \quad u, v \in \Sigma^*.$$

Theorem [3, 4] *For any function* $f : \Sigma^* \to \mathbb{R}$, $\operatorname{rank}(f) = \operatorname{rank}(\mathbf{H}_f)$.

- Spectral learning of WFAs (in a nutshell) [1, Lemma 4.1].
  1. Let $\mathbf{H}_f = \mathbf{PS}$ with $\mathbf{P}, \mathbf{S}^\top \in \mathbb{R}^{\Sigma^*\times n}$ where $n = \operatorname{rank}(f)$
  2. For each $\sigma \in \Sigma$, let $\mathbf{H}_f^\sigma \in \mathbb{R}^{\Sigma^*\times\Sigma^*}$ be defined by $(\mathbf{H}_f^\sigma)_{u,v} = f(u\sigma v)$ for all $u, v \in \Sigma^*$.
  3. WFA $(\boldsymbol{\alpha}, \{\mathbf{A}^\sigma\}_{\sigma\in\Sigma}, \boldsymbol{\omega})$ with $\boldsymbol{\alpha}^\top = \mathbf{P}_{\lambda,:}$, $\boldsymbol{\omega} = \mathbf{S}_{:,\lambda}$, and $\mathbf{A}^\sigma = \mathbf{P}^\dagger\mathbf{H}^\sigma\mathbf{S}^\dagger$ is a minimal WFA for f.

WFAs as Linear Models in a Feature Space

- Computation of a WFA A on $x \in \Sigma^*$:
  1. map x to feature vector $\phi(x) = \boldsymbol{\alpha}^\top\mathbf{A}^x$ through a compositional feature map $\phi : \Sigma^* \to \mathbb{R}^n$
  2. compute final value $f_A(x) = \langle\phi(x), \boldsymbol{\omega}\rangle$



- ϕ is compositional: $\phi(x\sigma)^\top = \phi(x)^\top\mathbf{A}^\sigma$.
- ϕ is minimal if $V = \operatorname{span}(\{\phi(x)\}_{x\in\Sigma^*}) \subset \mathbb{R}^n$ is of dimension n.

$\Rightarrow$ $\phi : x \mapsto \boldsymbol{\alpha}^\top\mathbf{A}^x$ is minimal if and only if $(\boldsymbol{\alpha}, \{\mathbf{A}^\sigma\}_{\sigma\in\Sigma}, \boldsymbol{\omega})$ is minimal.

A Notion of Relatedness between Functions on Sequences

Relatedness between WFAs: to which extent two WFAs can share a joint feature map ϕ.

- Let $f_1, f_2 : \Sigma^* \to \mathbb{R}$ of rank n_1 and n_2. with feature maps $\phi_1 : \Sigma^* \to \mathbb{R}^{n_1}$ and $\phi_2 : \Sigma^* \to \mathbb{R}^{n_2}$.
- $\phi = \phi_1 \oplus \phi_2 : \Sigma^* \to \mathbb{R}^{n_1+n_2}$ is a joint feature map for f_1 and f_2:

$$f_1(x) = \langle\phi(x), \boldsymbol{\omega}_1 \oplus \mathbf{0}\rangle \quad \text{and} \quad f_2(x) = \langle\phi(x), \mathbf{0} \oplus \boldsymbol{\omega}_2\rangle$$

but it may not be minimal.

$\rightarrow$ there may exist another feature map of dimension $n < n_1 + n_2$.

- The smaller n is, the more related f_1 and f_2 are.

Vector-Valued WFA

- A d-dimensional vector-valued weighted finite automaton (vv-WFA) with n states is a tuple $A = (\boldsymbol{\alpha}, \{\mathbf{A}^\sigma\}_{\sigma\in\Sigma}, \boldsymbol{\Omega})$ where
  - $\boldsymbol{\alpha} \in \mathbb{R}^n$ is the initial weights vector
  - $\boldsymbol{\Omega} \in \mathbb{R}^{n\times d}$ is the matrix of final weights
  - $\mathbf{A}^\sigma \in \mathbb{R}^{n\times n}$ is the transition matrix for each $\sigma \in \Sigma$.
- A vv-WFA computes a function $\vec{f}_A : \Sigma^* \to \mathbb{R}^d$ defined for each word $x = x_1 x_2 \cdots x_k \in \Sigma^*$ by

$$\vec{f}_A(x_1 x_2 \cdots x_k) = \boldsymbol{\alpha}^\top\mathbf{A}^{x_1}\mathbf{A}^{x_2}\cdots\mathbf{A}^{x_k}\boldsymbol{\Omega} = \boldsymbol{\alpha}^\top\mathbf{A}^x\boldsymbol{\Omega}.$$

$\Rightarrow$ Rank of $\vec{f} = [f_1, f_2] : \Sigma^* \to \mathbb{R}^2$ equal dimension of a minimal joint feature map for f_1 and f_2.

$\Rightarrow$ $\max\{\operatorname{rank}(f_1), \operatorname{rank}(f_2)\} \le \operatorname{rank}([f_1, f_2]) \le \operatorname{rank}(f_1) + \operatorname{rank}(f_2)$.

Example

$$\begin{cases} f_1(x) = 0.5|x|_a + 0.5|x|_b \\ f_2(x) = 0.3|x|_b - 0.6|x|_c \\ f_3(x) = |x|_c \end{cases}$$

$\operatorname{rank} f_2 = 4 = \operatorname{rank}([f_2, f_3])$

$\operatorname{rank}([f_1, f_3]) = 6 = \operatorname{rank}(f_1) + \operatorname{rank}(f_3)$

$\operatorname{rank}(f_1) = \operatorname{rank}(f_2) < \operatorname{rank}([f_1, f_2]) < \operatorname{rank}(f_1) + \operatorname{rank}(f_2)$

Spectral Learning of vv-WFAs

- Hankel tensor $\mathcal{H} \in \mathbb{R}^{\Sigma^*\times d\times\Sigma^*}$ associated with a function $\vec{f} : \Sigma^* \to \mathbb{R}^d$

$$\mathcal{H}_{u,:,v} = \vec{f}(uv) \quad \text{for all} \quad u, v \in \Sigma^*.$$

Theorem [Vector-Valued Fliess Theorem] *For any* $\vec{f} : \Sigma^* \to \mathbb{R}^d$, $\operatorname{rank}(\vec{f}) = \operatorname{rank}(\mathcal{H}_{(1)})$, *where* $\mathcal{H}_{(1)} = [\mathcal{H}_{:,1,:} \;\; \mathcal{H}_{:,2,:} \;\; \cdots \;\; \mathcal{H}_{:,d,:}]$ *is the flattening of the Hankel tensor.*

- Spectral learning of vv-WFAs. A vv-WFA computing $\vec{f}$ can be recovered from any rank n factorization of $\mathcal{H}_{(1)}$:
  1. Let $\mathcal{H}_{(1)} = \mathbf{P}\mathcal{S}_{(1)}$ with $\mathbf{P} \in \mathbb{R}^{\Sigma^*\times n}$ and $\mathcal{S} \in \mathbb{R}^{n\times d\times\Sigma^*}$.
  2. For each $\sigma \in \Sigma$, let $\mathcal{H}^\sigma \in \mathbb{R}^{\Sigma^*\times d\times\Sigma^*}$ be defined by $\mathcal{H}^\sigma_{u,:,v} = \vec{f}(u\sigma v)$ for all $u, v \in \Sigma^*$.
  3. The vv-WFA $A = (\boldsymbol{\alpha}, \{\mathbf{A}^\sigma\}_{\sigma\in\Sigma}, \boldsymbol{\Omega})$ where $\boldsymbol{\alpha}^\top = \mathbf{P}_{\lambda,:}$, $\boldsymbol{\Omega} = \mathcal{S}_{:,:,\lambda}$, and $\mathbf{A}^\sigma = \mathbf{P}^\dagger\mathcal{H}^\sigma_{(1)}(\mathcal{S}_{(1)})^\dagger$ is a minimal vv-WFA for $\vec{f}$.

Multitask Learning of WFAs

- Let $f_1, \cdots, f_m$ be related functions defined on Σ^*.
- Learning $\vec{f} = [f_1, \cdots, f_m]$ as a vv-WFA enforces discovering a shared feature map between tasks.

Algorithm 1

1: Compute the rank R truncated SVD $\hat{\mathcal{H}}_{(1)} \simeq \mathbf{UDV}^\top$.
2: Let $A = (\boldsymbol{\alpha}, \{\mathbf{A}^\sigma\}_{\sigma\in\Sigma}, \boldsymbol{\Omega})$ be the vv-WFA defined by

$$\boldsymbol{\alpha}^\top = \mathbf{U}_{\lambda,:}, \quad \boldsymbol{\Omega} = \mathbf{U}^\top(\hat{\mathcal{H}}_{:,:,\lambda}) \quad \text{and} \quad \mathbf{A}^\sigma = \mathbf{U}^\top\hat{\mathcal{H}}^\sigma_{(1)}(\hat{\mathcal{H}}_{(1)})^\dagger\mathbf{U} \;\text{ for each } \sigma \in \Sigma.$$

3: **for** $i = 1$ **to** m
4: Compute the rank R_i truncated SVD $\hat{\mathcal{H}}_{:,i,:} \simeq \mathbf{U}_i\mathbf{D}_i\mathbf{V}_i^\top$.
5: Let $A_i = (\mathbf{U}_i^\top\mathbf{U}\boldsymbol{\alpha}, \{\mathbf{U}_i^\top\mathbf{U}\mathbf{A}^\sigma\mathbf{U}^\top\mathbf{U}_i\}_{\sigma\in\Sigma}, \mathbf{U}_i^\top\mathbf{U}\boldsymbol{\Omega}_{:,i})$

- Additional step to the spectral learning algorithm (lines 3-5):
  - vv-WFA $A = (\boldsymbol{\alpha}, \{\mathbf{A}^\sigma\}_{\sigma\in\Sigma}, \boldsymbol{\Omega})$ is minimal $\not\Rightarrow$ WFA $A_i = (\boldsymbol{\alpha}, \{\mathbf{A}^\sigma\}_{\sigma\in\Sigma}, \boldsymbol{\Omega}_{:,i})$ is minimal.
  - Need to project down each A_i to its true dimension.

Theoretical Insight

Theorem Let $\mathbf{M} \in \mathbb{R}^{d_1\times d_2}$ of rank R, $\hat{\mathbf{M}} = \mathbf{M} + \mathbf{E}$, $\boldsymbol{\Pi}_U, \boldsymbol{\Pi}_{\hat{U}} \in \mathbb{R}^{d_1\times d_1}$ matrices of orthogonal projections on the top R left sing. vectors of $\mathbf{M}$ and $\hat{\mathbf{M}}$. Then, for any $\delta > 0$, with probability $\ge 1 - \delta$,

$$\|\boldsymbol{\Pi}_U - \boldsymbol{\Pi}_{\hat{U}}\|_F \le 4\left(\sqrt{\frac{(d_1 - R)R + 2\log(1/\delta)}{d_1 d_2}}\frac{\|\mathbf{E}\|_F}{s_R(\mathbf{M})} + \frac{\|\mathbf{E}\|_F^2}{s_R(\mathbf{M})^2}\right). \tag{1}$$

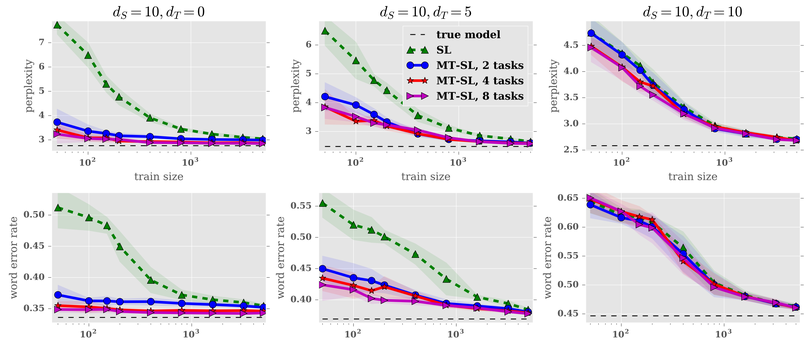
- Consider m tasks $f_1, \cdots, f_m$ with empirical Hankel matrices $\hat{\mathbf{H}}_1, \cdots, \hat{\mathbf{H}}_m \in \mathbb{R}^{P\times S}$, then

$$\hat{\mathcal{H}}_{(1)} = \left[\hat{\mathbf{H}}_1 \;\; \hat{\mathbf{H}}_2 \;\cdots\; \hat{\mathbf{H}}_m\right] \in \mathbb{R}^{P\times mS}.$$

- If the tasks are maximally related (i.e. $R = \operatorname{rank}(\vec{f}) = \operatorname{rank}(f_1) = \cdots = \operatorname{rank}(f_m)$) then as the number of tasks grows, the first term in Eq. (1) tends to 0 and the estimation error of the singular subspace decays quadratically instead of linearly.

Experiments on Synthetic Data

- Randomly generated stochastic WFAs following the PAutomaC competition process [6].
- Related WFAs: joint feature space of dimension $d_S = 10$ and task specific space of dimension d_T (i.e. $\operatorname{rank}(f_i) = d_S + d_T$ and $\operatorname{rank}(\vec{f}) = \operatorname{rank}([f_1, \cdots, f_m]) = d_S + md_T$).
- Training sample drawn from target task f_1 and training samples of size $5,000$ for tasks $f_2, \cdots, f_m$.



Experiments on Real Data

- Universal Dependencies [5]: sentences from 33 languages labeled with 17 PoS tags.

$\Rightarrow$ Samples drawn from 33 distributions over strings on an alphabet of size 17.

- For each language, (80%, 10%, 10%)-split between training, validation and test sets.
- Two ways of selecting related tasks:
  1. use all other languages
  2. select the 4 closest languages w.r.t. the distance between the (top-50) left singular subspaces of the Hankel matrices.

Training size	100	500	1000	5000	all available data
	Related tasks: all other languages				
Perplexity	7.0744 (±7.76)	3.6666 (±5.22)	3.2879 (±5.17)	3.4187 (±5.57)	3.1574 (±5.48)
WER	1.4919 (±2.37)	1.3786 (±2.94)	1.2281 (±2.62)	1.4964 (±2.70)	1.4932 (±2.77)
	Related tasks: 4 closest languages				
Perplexity	6.0069 (±6.76)	4.3670 (±5.83)	4.4049 (±5.50)	2.9689 (±5.87)	2.8229 (±5.90)
WER	2.0883 (±3.26)	1.5175 (±2.87)	1.2961 (±2.57)	1.3080 (±2.55)	1.2160 (±2.31)

Table: Average relative improvement over all languages (in %) of MT-SL vs. SL on the UNIDEP dataset (e.g. for perplexity we report $100 \cdot (p_{\text{SL}} - p_{\text{MT-SL}})/p_{\text{SL}}$).

Target task	4 closest tasks w.r.t. subspace distance (closest first)			
Basque	Finnish	Polish	Czech	Indonesian
Croatian	Estonian	Slovenian	Czech	Finnish
French	Italian	Spanish	German	English
Hungarian	Danish	Ancient Greek	German	Portuguese
Gothic	Old Church Slavonic	Latin	Ancient Greek	Finnish
Italian	English	French	Spanish	Dutch
Japanese	Hindi	Persian	Arabic	Tamil
Latin	Old Church Slavonic	Ancient Greek	Gothic	Finnish
Swedish	Danish	Norwegian	Finnish	Estonian

Table: Some related tasks used in the UNIDEP experiment.

- Cherry picked example: on the Basque task with a training set of size 500, the WER was reduced from $\sim 76\%$ for SL to $\sim 70\%$ using all other languages as related tasks, and to $\sim 65\%$ using the 4 closest tasks (Finnish, Polish, Czech and Indonesian).

References

[1] Borja Balle, Xavier Carreras, Franco M Luque, and Ariadna Quattoni. Spectral learning of weighted automata. *Machine learning*, 96(1-2):33–63, 2014.

[2] T Tony Cai and Anru Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *arXiv preprint arXiv:1605.00353*, 2016.

[3] Jack W. Carlyle and Azaria Paz. Realizations by stochastic finite automata. *Journal of Computer and System Sciences*, 5(1):26–40, 1971.

[4] Michel Fliess. Matrices de hankel. *Journal de Mathématiques Pures et Appliquées*, 53(9):197–222, 1974.

[5] Joakim Nivre, Zeljko Agić, Lars Ahrenberg, et al. Universal dependencies 1.4, 2016. URL `http://hdl.handle.net/11234/1-1827`. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

Emails: guillaume.rabusseau@mail.mcgill.ca, pigem@amazon.co.uk, jpineau@cs.mcgill.ca