A Structured Prediction Approach for Label Ranking

Anna Korba, Alexandre Garcia, Florence d'Alché-Buc
LTCI, Télécom ParisTech, Université Paris-Saclay

Label ranking

Consider:

- A set of K items/labels : $\{1, \ldots, K\}$ (Ex: $\{1, 2, 3, 4\}$).
- A individual expresses her preferences as a *full ranking* (strict order $\succ$) over K :

$$a_1 \succ a_2 \succ \cdots \succ a_K \quad (\text{Ex: } 2 \succ 1 \succ 3 \succ 4)$$

- Also seen as the *permutation* σ that maps an item to its rank:

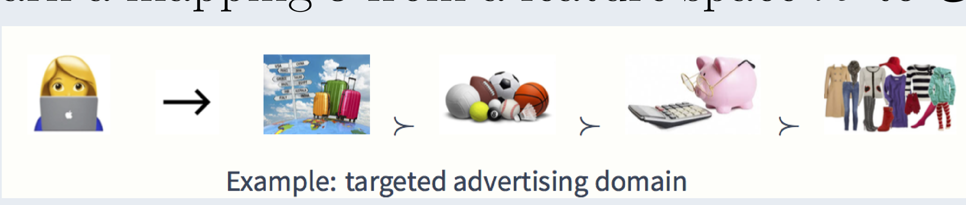
$$a_1 \succ \cdots \succ a_K \quad \Leftrightarrow \quad \sigma \in \mathfrak{S}_K \text{ s.t. } \sigma(a_i) = i$$

(Ex: $\sigma(2) = 1, \sigma(1) = 2, \cdots \Rightarrow \sigma = 2134$)

- $\mathfrak{S}_K$: set of permutations of $\{1, \ldots, K\}$.

Label ranking

Learn a mapping s from a feature space $\mathcal{X}$ to $\mathfrak{S}_K$:



Loss function: $\Delta : \mathfrak{S}_K \times \mathfrak{S}_K \to \mathbb{R}$

Learning problem

Goal: learn a function $s : \mathcal{X} \to \mathfrak{S}_K$ that minimizes the expected risk:

$$\min_{s:\mathcal{X}\to\mathfrak{S}_K} \mathcal{E}(s) = \int_{\mathcal{X}\times\mathfrak{S}_K} \Delta(s(x), \sigma) dP(x, \sigma). \tag{1}$$

with Δ some loss function, e.g.:

- Kendall's τ:
  $\Delta_\tau(\sigma, \sigma') = \sum_{i<j} \mathbb{I}[(\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0]$
- Hamming :
  $\Delta_H(\sigma, \sigma') = \sum_{i=1}^K \mathbb{I}[\sigma(i) \neq \sigma'(i)]$.

Idea: Consider a family of Δ loss functions:

$$\Delta(\sigma, \sigma') = \|\phi(\sigma) - \phi(\sigma')\|^2_{\mathcal{F}}. \tag{2}$$

with $\phi : \mathfrak{S}_K \to \mathcal{F}$ some ranking embedding, i.e. that maps the permutations $\sigma \in \mathfrak{S}_K$ into a Hilbert space $\mathcal{F}$ (e.g. $\mathbb{R}^d$ for $d \in \mathbb{N}$).

Motivation: There exist ϕ_τ, ϕ_H such that Δ_τ and Δ_H write as (2).

Structured prediction approach

Pb: (1) is hard to optimize.
Idea: Introduce a surrogate problem:

$$\min_{g\,:\,\mathcal{X}\to\mathcal{F}} \mathcal{R}(g), \quad \text{with} \quad \mathcal{R}(g) = \mathbb{E}\left[\|g(x) - \phi(\sigma)\|^2_{\mathcal{F}}\right]$$

$\Rightarrow$ **easier to optimize since g has values in $\mathcal{F}$.**

We can thus approach structured prediction in **two steps** (see [1, 2]) :

- **Step 1 (Regression)**: Learn g from $\mathcal{D}_N = (x_1, \phi(\sigma_1)), \ldots, (x_N, \phi(\sigma_N))$ with any regression method (kNN, RF, Ridge regression...)
  $\Longrightarrow$ Output $\widehat{g} : \mathcal{X} \to \mathcal{F}$
- **Step 2 (Pre-image)**: for any $x \in \mathcal{X}$:

$$\widehat{s}(x) = d \circ \widehat{g}(x) = \underset{\sigma \in \mathfrak{S}_K}{\operatorname{argmin}} \|\phi(\sigma) - \widehat{g}(x)\|^2_{\mathcal{F}}$$

$$\begin{tikzcd} \mathcal{X} \arrow[r, "\widehat{s}"] \arrow[d, "\widehat{g}"'] & \mathfrak{S}_n \\ \mathcal{F} \arrow[ur, "d"'] & \end{tikzcd}$$

$\Longrightarrow$ Choice of ϕ and regression method matter

Embeddings proposed

- **Kemeny**

$$\phi_\tau \colon \mathfrak{S}_K \to \mathbb{R}^{K(K-1)/2}$$
$$\sigma \mapsto (\text{sign}(\sigma(j) - \sigma(i)))_{1\leq i<j\leq K}$$

Ex: $\sigma = 132$
$\phi_\tau(\sigma) = (1, 1, -1)$

Properties : $\|\phi_\tau(\sigma) - \phi_\tau(\sigma')\|^2 = 4\Delta_\tau(\sigma, \sigma')$ and $\|\phi_\tau(\sigma)\| = \sqrt{K(K-1)/2}$.
Pre-image problem: NP-Hard

- **Hamming**

$$\phi_H \colon \mathfrak{S}_K \to \mathbb{R}^{K\times K}$$
$$\sigma \mapsto (\mathbb{I}\{\sigma(i) = j\})_{1\leq i,j\leq K}$$

Ex: $\sigma = 132$
$\phi_H(\sigma) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$

Properties: $\|\phi_H(\sigma) - \phi_H(\sigma')\|^2 = \Delta_H(\sigma, \sigma')$ and $\|\phi_H(\sigma)\| = \sqrt{K}$.
Pre-image problem: computed by the Hungarian algorithm in $\mathcal{O}(K^3)$

- **Lehmer**

$$\phi_L \colon \mathfrak{S}_K \to \mathbb{R}^K$$
$$\sigma \mapsto (\#\{i \in [\![K]\!] : i < j, \sigma(i) > \sigma(j)\})_{j=1,\ldots,K}$$

Ex: $\sigma = 132$
$\phi_L(\sigma) = (0, 1, 0)$

Properties: Related to Δ_τ: $\frac{1}{K-1}\Delta_\tau(\sigma, \sigma') \leq |\phi_L(\sigma) - \phi_L(\sigma')| \leq \Delta_\tau(\sigma, \sigma')$.
Pre-image: by rounding towards closest integer and decoding Lehmer in $\mathcal{O}(K)$

Computational properties

Embedding	Embedding σ in $\phi(\sigma)$ **Step 1 a**	Pre-image in $\mathcal{F}$ **Step 2 b**
ϕ_τ	$\mathcal{O}(K^2N)$	NP-hard
ϕ_H	$\mathcal{O}(KN)$	$\mathcal{O}(K^3N)$
ϕ_L	$\mathcal{O}(KN)$	$\mathcal{O}(KN)$

Regressor	Learning $\hat{g}$ **Step 1 b**	Prediction in $\mathcal{F}$ $(dim(\mathcal{F}) = m)$ **Step 2 a**
kNN	$\mathcal{O}(1)$	$\mathcal{O}(Nm)$
Ridge	$\mathcal{O}(N^3)$	$\mathcal{O}(Nm)$

Table 1: Embeddings and regressors complexities.

$\Longrightarrow$ Fastest: kNN+Lehmer in $\mathcal{O}(KN)$

Numerical results

Table 2: Mean Kendall's τ coefficient on benchmark datasets

	authorship	glass	iris	vehicle	vowel	wine
kNN Hamming	0.01±0.02	0.08±0.04	-0.15±0.13	-0.21±0.04	0.24±0.04	-0.36±0.04
kNN Kemeny	**0.94**±0.02	0.85±0.06	0.95±0.05	0.85±0.03	0.85±0.02	0.94±0.05
kNN Lehmer	0.93±0.02	0.85±0.05	0.95±0.04	0.84±0.03	0.78±0.03	0.94±0.06
ridge Hamming	-0.00±0.02	0.08±0.05	-0.10±0.13	-0.21±0.03	0.26±0.04	-0.36±0.03
ridge Lehmer	0.92±0.02	0.83±0.05	**0.97**±0.03	0.85±0.02	0.86±0.01	0.84±0.08
ridge Kemeny	**0.94**±0.02	0.86±0.06	**0.97**±0.05	**0.89**±0.03	**0.92**±0.01	0.94±0.05
Cheng PL	**0.94**±0.02	0.84±0.07	0.96±0.04	0.86±0.03	0.85±0.02	**0.95**±0.05
Cheng LWD	0.93±0.02	0.84±0.08	0.96±0.04	0.85±0.03	0.88±0.02	0.94±0.05
Zhou RF	0.91	**0.89**	**0.97**	0.86	0.87	**0.95**

Cheng PL [3], Cheng LWD [4], Zhou RF [5]

Kendall's τ coefficient corresponds to a rescaling of Kendall's tau distance d_τ between [-1,1] (so the closer from 1 is the better)

Theory vs computation

For **Kemeny** and **Hamming** embedding:

- **consistency holds** ([1]):

$$\mathcal{E}(d \circ \widehat{g}) - \mathcal{E}(s^*) \leq c_\phi \sqrt{\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)}$$

with $c_{\phi_\tau} = \sqrt{\frac{K(K-1)}{2}}$ and $c_{\phi_H} = \sqrt{K}$ (constants with K the number of labels)

- but the **pre-image step is hard**

In contrast, for the **Lehmer** embedding:

- we **lose consistency**:

$$\mathcal{E}(d \circ \widehat{g}) - \mathcal{E}(s^*) \leq \sqrt{\frac{K(K-1)}{2}} \sqrt{\mathcal{R}(\widehat{g}) - \mathcal{R}(g^*)} + \mathcal{E}(d \circ g^*) - \mathcal{E}(s^*) + \mathcal{O}(K\sqrt{K})$$

- but the **pre-image step is fast**

References

[1] Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi.
A consistent regularization approach for structured prediction.
In *Advances in Neural Information Processing Systems*, pages 4412–4420, 2016.

[2] Céline Brouard, Marie Szafranski, and Florence d'Alché Buc.
Input output kernel regression: supervised and semi-supervised structured output prediction with operator-valued kernels.
Journal of Machine Learning Research, 17(176):1–48, 2016.

[3] Weiwei Cheng, Eyke Hüllermeier, and Krzysztof J Dembczynski.
Label ranking methods based on the plackett-luce model.
In *Proceedings of the 27th Annual International Conference on Machine Learning (ICML-10)*, pages 215–222, 2010.

[4] W. Cheng and E. Hüllermeier.
A nearest neighbor approach to label ranking based on generalized labelwise loss minimization, 2013.

[5] Yangming Zhou and Guoping Qiu.
Random forest for label ranking.
Expert Systems with Applications, 2018.