Ranking Distributions based on Noisy Sorting

Adil El Mesaoudi-Paul, Eyke Hüllermeier & Róbert Busa-Fekete

ICML debrief - Robin Vogel - 04/10/18

Outline

Probability distributions on rankings

Ranking distributions based on sorting

Parameter estimation

Appendix

Introduction

Analysis of ranking data:
$\rightarrow$ longstanding tradition in statistics: [Mallows, 1957], [Plackett, 1975], [Luce, 1959], [Babington-Smith, 1950], ...
$\rightarrow$ various fields of application: psychology, social sciences, ...
$\rightarrow$ renewed interest with ML and IR.

Consider a fixed set $O = \{o_1, \ldots, o_K\}$ of K items.
A **ranking** over O is identified by $\pi \in \mathfrak{S}_K$, permutation over $[\![1, K]\!]$.
Ex: $\pi = (3, 1, 2)$ means first is o_3, then o_1, then o_2. And $\pi^{-1}(1) = 3$.

Contribution of the paper:

New class of **probability distributions** on rankings.

Usual probability distributions on rankings

The **Mallows model** (MM) [Mallows, 1957] param. $\tau \in \mathfrak{S}_k, \phi > 0$,

$$\mathbb{P}_{\tau,\phi}(\pi) = \frac{1}{C(\phi)} \exp\left(-\phi D(\pi,\tau)\right),$$

where D is the Kendall distance, i.e. the # of pairwise inversions between π and τ.

The **Plackett-Luce** (PL) model [Plackett, 1975] param. $\theta \in \mathbb{R}_+^K$,

$$\mathbb{P}_\theta(\pi) = \prod_{i=1}^{K} \frac{\theta_{\pi^{-1}(i)}}{\theta_{\pi^{-1}(i)} + \theta_{\pi^{-1}(i+1)} + \cdots + \theta_{\pi^{-1}(K)}}.$$

The **Babington-Smith** (BS) model [Babington-Smith, 1950] param. $P = (p_{i,j})_{i<j}$ s.t. $\forall 1 \leq i < j \leq K, p_{i,j} = 1 - p_{j,i}$,

$$\mathbb{P}_P(\pi) = \frac{1}{C(P)} \prod_{1 \leq i < j \leq K} p_{\pi^{-1}(i),\pi^{-1}(j)}.$$

Outline

Probability distributions on rankings

Ranking distributions based on sorting

Parameter estimation

Appendix

Ranking distribution based on sorting

See ranking as noisy sorting, given a ground truth $\tau \in \mathfrak{S}_K$:
$\rightarrow$ Run a sorting algorithm $\mathcal{A}$ on an initial ordering σ,
$\rightarrow$ When one compares o_i and o_j, right outcome with prob. p.

Introduce $P, D^{\pi,\sigma} \in \mathbb{R}^{K\times K}$, $P = (p_{i,j})_{i,j}$ and $D^{\pi,\sigma} = (d^{\pi,\sigma}_{i,j})_{i,j}$ with

$$p_{i,j} = \begin{cases} p & \text{if } \tau(o_i) < \tau(o_j), \\ 1-p & \text{if } \tau(o_i) > \tau(o_j). \end{cases} \quad \text{and} \quad d^{\pi,\sigma}_{i,j} = \begin{cases} 1 & \text{if } \mathcal{A} \text{ saw } o_i \succ o_j, \\ 0 & \text{otherwise.} \end{cases}$$

Insertion Sort Rank (ISR) model [Biernacki and Jacques, 2013]
param. $\tau \in \mathfrak{S}_K, p \in [0.5, 1]$,

$$\mathbb{P}_P(\pi) = \frac{1}{C'(P)} \sum_{\sigma\in\mathfrak{S}_k} \prod_{i=1}^{K} \prod_{j\neq i} p_{i,j}^{d^{\pi,\sigma}_{i,j}} .$$

Conjunctive Noisy Sorting

Conjunctive Noisy Sorting (CNS) [Mesaoudi-Paul et al., 2018]
param. $\tau \in \mathfrak{S}_k, p \in [0.5, 1]$,

$$\mathbb{P}_P(\pi) = \frac{1}{C(P)} \prod_{\sigma \in \mathfrak{S}_k} \prod_{i=1}^{K} \prod_{j \neq i} p_{i,j}^{d_{i,j}^{\pi,\sigma}} . \tag{1}$$

$\rightarrow$ Difference with ISR model: $\sum_{\sigma \in \mathfrak{S}_k}$ replaced by $\prod_{\sigma \in \mathfrak{S}_k}$.

Generalized Conjunctive Noisy Sorting (GCNS) ,
param. $\tau \in \mathfrak{S}_k, p \in [0, 1]^{K \times K}$, different p for each (i, j), see eq. (1).

Introduce $D^{\pi} := \sum_{\sigma \in \mathfrak{S}_K} D^{\pi,\sigma} = (d_{i,j}^{\pi})_{i,j}$, eq. (1) rewrites

$$\mathbb{P}_P(\pi) = \frac{1}{C(P)} \prod_{i=1}^{K} \prod_{j \neq i} p_{i,j}^{d_{i,j}^{\pi}} .$$

Relations to other distributions

Compare (BS) and (GCNS), with $\gamma^\pi_{i,j} = \mathbb{I}\{\pi(i) < \pi(j)\}$:

$$\text{(BS)} \quad \mathbb{P}_P(\pi) = \frac{1}{C(P)} \prod_{1\le i<j\le K} p_{i,j}^{\gamma^\pi_{i,j}} (1 - p_{i,j})^{\gamma^\pi_{j,i}},$$

$$\text{(GCNS)} \quad \mathbb{P}_P(\pi) = \frac{1}{C(P)} \prod_{1\le i<j\le K} p_{i,j}^{d^\pi_{i,j}} (1 - p_{i,j})^{d^\pi_{j,i}}.$$

Difference: The $d^\pi_{i,j}$'s $\in \mathbb{N}$ and depend on $\mathcal{A}$.

With param. $\tau \in \mathfrak{S}_K, p \in [0.5, 1]$, setting $p_{i,j} = p^{\gamma^\tau_{i,j}} (1 - p)^{\gamma^\tau_{j,i}}$ implies:

- (BS) reduces to (MM),
- (GCNS) reduces to (CNS).

Outline

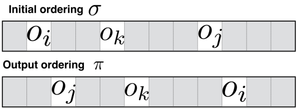
Instanciation of the model, i.e. calculus of D^π

The authors give an expression for D^π when:
→ $\mathcal{A}$ is the **insertion sort** algorithm $\mathcal{I}$,
→ $\mathcal{A}$ is the **quicksort** algorithm $\mathcal{Q}$.

Let B^π permutation matrix of π and σ_{Id} the identity permutation,
Values do not depend on the order of o_k's, hence $D^\pi = B^\pi D^{\sigma_{Id}} B^{\pi\top}$.

We now calculate $D^{\sigma_{Id}}$ for $\mathcal{I}$:

- **If** $i < j$, i.e. $o_i \prec o_j$ in final permutation σ_{Id},
  then i and j were compared once in the insertion phase,
  hence $d_{i,j}^{\sigma_{Id}} = \#\{\sigma \in \mathfrak{S}_k | \sigma(i) < \sigma(j)\} = K!/2$.
- **If** $i > j$, i.e. $o_i \succ o_j$ in final permutation σ_{Id},
  then i, j not compared if $\exists k$ s.t. o_k between o_i and o_j for (σ_{Id}, σ),
  Introduce $b_{i,j} = i - j - 1$, then $d_{i,j}^{\sigma_{Id}} = \binom{K}{b_{i,j}+2}(K - b_{i,j} - 2)!b_{i,j}!$.



In the case of $\mathcal{Q}$:
→ the pivot is deterministic,
→ $D^{\sigma_{Id}}$ is written by means of a recursive function.

Fitting the (CNS) model

Aim: fit a sample $\mathcal{D} = \{\pi_1, \ldots, \pi_n\} \subset \mathfrak{S}_K$.

(CNS) reduces to:

$$\begin{aligned}
&\max_{P^\tau} \sum_{\ell=1}^{n} \sum_{i \neq j} d_{i,j}^{\pi_\ell} \log p_{i,j}^\tau - n \log C(P^\tau), \\
&\text{s.t. } p_{i,j}^\tau = \begin{cases} p & \text{if } \tau(i) < \tau(j), \\ 1-p & \text{if } \tau(i) > \tau(j), \end{cases} \quad \forall i,j \in [\![K]\!], i \neq j.
\end{aligned} \tag{2}$$

Optimize eq. (2) iteratively on τ and p:

$\rightarrow$ Use **hill climbing** for τ at fixed p,
Init: Borda ranking | Neighborhood: swap of adjacent items.

$\rightarrow$ Convex problem for p at fixed τ, use the **golden section method**.
i.e. use three functions evaluations to locate a minimum.

Fitting the (GCNS) reduces to model

(GCNS) reduces to reduces to:

$$\max_{P^\tau} \sum_{\ell=1}^{n} \sum_{i=1}^{K} \sum_{j \neq i} d_{i,j}^{\pi_\ell} \log p_{i,j} - n \log C(P), \tag{3}$$
$$\text{s.t. } p_{i,j} + p_{j,i} = 1, \quad \forall i, j \in [\![K]\!], i \neq j.$$

In this case, **no closed form** for $C(P)$!

They optimize eq. (3) using **generalized iterative scaling** (GIS).

Experiments compare the (MM), (ISR), (CNS) and (GCNS) with both $\mathcal{I}$ and $\mathcal{Q}$ on 213 real-world datasets.
$\rightarrow$ Comparison by splitting data and computing **KL divergences**.
$\rightarrow$ (GCNS) $\succ$ (CNS) $\succ$ others.
$\rightarrow$ $\mathcal{I}$ performs better than $\mathcal{Q}$.

Sampling from (GCNS)

Based on the **acceptance ratio**:

$$q(\pi, \pi') := \log \frac{\mathbb{P}_P(\pi)}{\mathbb{P}_P(\pi')} = \sum_{i=1}^{K} \sum_{j \neq i} \left(d_{i,j}^{\pi} - d_{i,j}^{\pi'} \right) \log p_{i,j},$$

we can use the **Metropolis-Hastings** algorithm with (MM) proposal.
The (MM) is symmetric, i.e. $\mathbb{P}_{\phi,\pi}(\pi') = \mathbb{P}_{\phi,\pi'}(\pi)$.

1. **Init:** Set $\mathcal{D} = \emptyset$, σ_0 initial ordering.
2. **For** $i = 1$ to T **do**
3. $\quad \pi_i \sim \mathbb{P}_{\phi,\sigma_{i-1}}$, (MM)
4. $\quad q_i \leftarrow q(\pi_i, \pi_{i-1})$,
5. $\quad$ With probability $\min(1, \exp(\pi_i))$, set $D = D \cup \{\sigma_i\}$.
6. **Return** $\mathcal{D}$.

References

Babington-Smith, B. (1950).
Discussion of professor ross' paper.
Journal of the Royal Statistical Society B, 12:153–162.

Biernacki, C. and Jacques, J. (2013).
A generative model for rank data based on an insertion sorting algorithm.
Computational Statistics and Data Analysis, 58:162–176.

Luce, R. D. (1959).
Individual Choice Behavior: A Theoretical analysis.
Wiley, New York, NY, USA.

Mallows, C. L. (1957).
Non-null ranking models.
Biometrika, 44(1-2):114–130.

Mesaoudi-Paul, A. E., Hüllermeier, E., and Busa-Fekete, R. (2018).
Ranking distributions based on noisy sorting.
In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 3472–3480.

Plackett, R. L. (1975).
The analysis of permutations.
Applied Statistics, 24:193–202.

Outline

Other models

The **Generalized Mallows Model** (GMM)

$$\mathbb{P}_{\tau,\phi}(\pi) = \frac{1}{C(\phi)} \exp\left(-\sum_{j=1}^{K-1} \phi_j V_j(\pi)\right),$$

where $V_j(\pi) = \sum_{i>j} \mathbb{I}\{\pi^{-1}(i) < \pi^{-1}(j)\}$ is the number of inversions for π w.r.t. $\sigma_{Id} \in \mathfrak{S}_K$.

Recursive Inversion Model (RIM)
Binary recursive decomposition τ with vertices $\mathcal{I}$ that weight inversions with $\theta_i, \forall i \in \mathcal{I}$. The probability of a ranking π is proportional to

$$\prod_{i \in \mathcal{I}} \exp\left(-\theta_i v_i(\pi, \pi_\tau)\right),$$

with $v_i(\pi, \pi_\tau)$ the number of inversions at vertex i of τ for π.

Reminders (1/2)

The **permutation matrix** of $\pi \in \mathfrak{S}_K$ is written:

$$P_\pi = \left(e_{\pi^{-1}(1)}, e_{\pi^{-1}(2)}, \ldots, e_{\pi^{-1}(K)}\right),$$

where e_i is the ith standard basis vector.

The **Kendall-distance** D between $\pi, \tau \in \mathfrak{S}_K$:

$$D(\pi, \tau) = \sum_{i<j} \mathbb{I}\{(\pi(i) - \pi(j))(\tau(i) - \tau(j)) < 0\}.$$

The **Borda ranking** is obtained by averaging the ranks of the items.
The **Borda score** s verifies:

$$s(i) = \frac{1}{n} \sum_{\ell=1}^{n} (n + 1 - \pi(i))$$

Reminders (2/2)

The **Lehmer code** associates to $\sigma \in \mathfrak{S}_k$ the application f, where:

$$f(i) = \#\{j \mid 1 \leq j \leq i \text{ and } \sigma(j) < \sigma(i)\},$$

which can be summarized by an element in $[\![1]\!] \times [\![2]\!] \times \cdots \times [\![K]\!]$.
The Lehmer code is bijective.

Kullback-Leibler divergence:

$$D_{KL}(P||Q) = \int_{\mathcal{X}} p \log\left(\frac{p}{q}\right) d\mu.$$

Proofs for the equivalence of the models

Justify the relationship between (BS) and (GCNS)

Since the permutation π is bijective $[\![N]\!] \to [\![N]\!]$,

$$\prod_{1\leq k,l\leq K} p_{k,l}^{\mathbb{I}\{\pi(k)<\pi(l)\}} = \prod_{1\leq i,j\leq K} p_{\pi^{-1}(i),\pi^{-1}(j)}^{\mathbb{I}\{i<j\}} = \prod_{1\leq i<j\leq K} p_{\pi^{-1}(i),\pi^{-1}(j)}.$$

Justify the reduction of (BS) to (MM)

Fixing $p_{i,j} = p^{\gamma^{\tau}_{i,j}}(1-p)^{\gamma^{\tau}_{j,i}}$, (then $1 - p_{i,j} = p^{\gamma^{\tau}_{j,i}}(1-p)^{\gamma^{\tau}_{i,j}}$) leads to

$$\mathbb{P}_P(\pi) = \frac{1}{C(P)} \prod_{1\leq i<j\leq K} p^{\gamma^{\tau}_{i,j}\gamma^{\pi}_{i,j}+\gamma^{\tau}_{j,i}\gamma^{\pi}_{j,i}} \cdot (1-p)^{\gamma^{\tau}_{i,j}\gamma^{\pi}_{j,i}+\gamma^{\pi}_{i,j}\gamma^{\tau}_{j,i}}.$$