On the Consistency of Ranking Algorithms

John Duchi   Lester Mackey   Michael I. Jordan

University of California, Berkeley

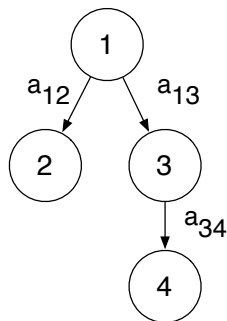International Conference on Machine Learning, 2010

Ranking

Goal: Order set of inputs/results to best match the preferences of an individual or a population

- ▶ Web search: Return most relevant results for user queries

- ▶ Recommendation systems:
  - ▶ Suggest movies to watch based on user's past ratings
  - ▶ Suggest news articles to read based on past browsing history

- ▶ Advertising placement: Maximize profit and click-through

Supervised ranking setup

Observe: Sequence of training examples

- Query q: e.g., search term
- Set of results x to rank
    - Items $\{1, 2, 3, 4\}$
- Weighted DAG G representing preferences over results
    - Item 1 preferred to $\{2, 3\}$ and item 3 to 4

Observe multiple preference graphs for the same query q and results x

Example G with
$x = \{1, 2, 3, 4\}$

Supervised ranking setup

Learn: Scoring function $f(x)$ to rank results x
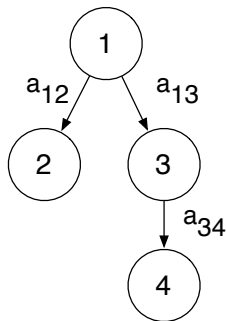
- Real-valued score for result i

$$s_i := f_i(x)$$

- Result i ranked above j iff $f_i(x) > f_j(x)$
- Loss suffered when scores s disagree with preference graph G:

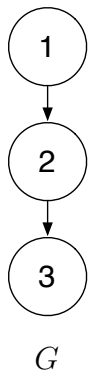$$L(s, G) = \sum_{i,j} a_{ij} 1_{(s_i < s_j)}$$

Example G with $x = \{1, 2, 3, 4\}$

Example:

$$L(s, G) = a_{12} 1_{(s_1 < s_2)} + a_{23} 1_{(s_1 < s_3)} + a_{34} 1_{(s_3 < s_4)}$$

Supervised ranking setup

Example: Scoring function f optimally ranks results in G
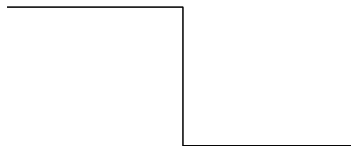

$$f_1(x) > f_2(x)$$

$$f_2(x) > f_3(x)$$

G

Detour to classification
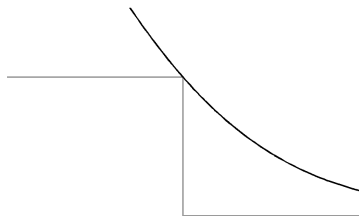
Consider the simpler problem of classification

- ▶ Given: Input x, label $y \in \{-1, 1\}$
- ▶ Learn: Classification function $f(x)$. Have *margin* $s = yf(x)$

Loss $L(s) = 1_{(s \leq 0)}$

Surrogate loss $\phi(s)$

Hard

Tractable

Classification and surrogate consistency

Question: Does minimizing expected ϕ-loss minimize expected L?

$$\text{Minimize } \sum_{i=1}^{n}\phi(y_i f(x_i)) \overset{n\to\infty}{\Rightarrow} \text{ Minimize } \mathbb{E}\phi(Yf(X))$$

$$\overset{?}{\Longleftrightarrow} \text{ Minimize } \mathbb{E}L(Yf(X))$$

Theorem: If ϕ is convex, procedure based on minimizing ϕ is consistent if and only if $\phi'(0) < 0$.[1]

What about ranking consistency?

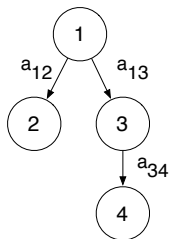Minimization of true ranking loss is hard

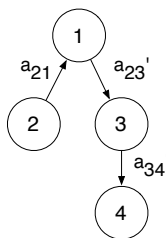▶ Replace ranking loss $L(s, G)$ with tractable surrogate $\varphi(s, G)$
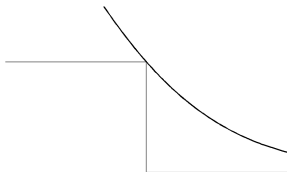
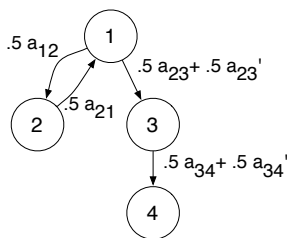**Question:** When is surrogate minimization consistent for ranking?

Conditional losses

$$p(G_1) = .5 \qquad p(G_2) = .5 \qquad \text{Aggregate}$$

- $\ell(p, s) = \sum_G p(G|x, q) L(s, G)$
- $\ell(p, s) = .5a_{21}1_{(s_2 < s_1)} + .5(a_{12} + a'_{12})1_{(s_1 < s_2)}$
  $\qquad\qquad + .5(a_{23} + a'_{23})1_{(s_1 < s_3)} + .5(a_{34} + a'_{34})1_{(s_3 < s_4)}$

- Optimal score vectors

$$A(p) = \operatorname*{argmin}_s \ell(p, s)$$

Consistency theorem

Theorem: Procedure minimizing φ is asymptotically consistent if and only if

$$\inf_s \left\{ \sum_G p(G)\varphi(s, G) \mid s \notin A(p) \right\} > \inf_s \left\{ \sum_G p(G)\varphi(s, G) \right\}$$

In other words, φ is consistent if and only if minimization gives correct order to the results

Goal: Find tractable φ so s that minimizes

$$\sum_G p(G)\varphi(s, G)$$

minimizes $\ell(p, s)$.

Consistent and Tractable?

Hard to get consistent and tractable φ

- In general, it is NP-hard even to *find* s minimizing
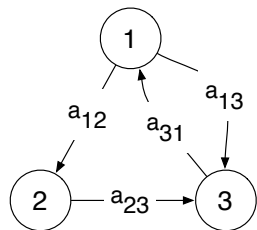
$$\sum_G p(G) L(s, G).$$

  (reduction from feedback arc-set problem)

Some restrictions on the problem space necessary...

Low noise setting

Definition: Low noise if $a_{ij} - a_{ji} > 0$ and $a_{jk} - a_{kj} > 0$

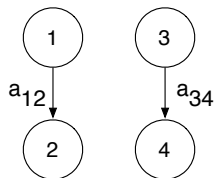implies $a_{ik} - a_{ki} \geq (a_{ij} - a_{ji}) + (a_{jk} - a_{kj})$

▶ Intuition: weight on path reinforces local weights, local weights reinforce paths.

▶ Reverse triangle inequality

▶ True when DAG derived from user ratings

$a_{13} - a_{31} \geq a_{12} + a_{23}$

Trying to achieve consistency

Try ideas from classification: ϕ is convex, bounded below, $\phi'(0) < 0$.
Common in ranking literature.[2]

$$\varphi(s, G) = \sum_{ij} a_{ij}\phi(s_i - s_j)$$

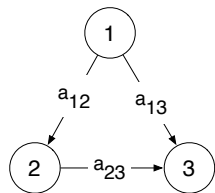$$\varphi(s, G) = a_{12}\phi(s_1 - s_2) + a_{34}\phi(s_3 - s_4)$$

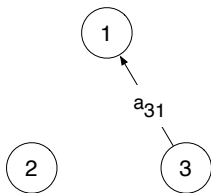**Theorem:** φ is not consistent, even in low noise settings.

---

[2]Herbrich et al., 2000; Freund et al., 2003; Dekel et al., 2004, etc.

What is the problem?

Surrogate loss $\varphi(s, G) = \sum_{ij} a_{ij} \phi(s_i - s_j)$
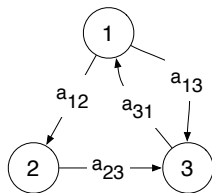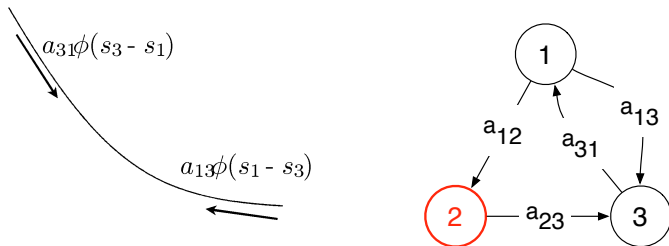

$p(G_1) = .5$ $\qquad\qquad$ $p(G_2) = .5$ $\qquad\qquad$ Aggregate

$$\sum_G p(G)\varphi(s, G) = \frac{1}{2}\varphi(s, G_1) + \frac{1}{2}\varphi(s, G_2)$$
$$\propto a_{12}\phi(s_1 - s_2) + a_{13}\phi(s_1 - s_3) + a_{23}\phi(s_2 - s_3) + a_{31}\phi(s_3 - s_1)$$

What is the problem?

$$a_{12}\phi(s_1 - s_2) + a_{13}\phi(s_1 - s_3) + a_{23}\phi(s_2 - s_3) + a_{31}\phi(s_3 - s_1)$$

More bang for your \$\$ by increasing to 0 from left: $s_1 \downarrow$. Result:

$$s^* = \underset{s}{\operatorname{argmin}} \sum_{ij} a_{ij}\phi(s_i - s_j)$$
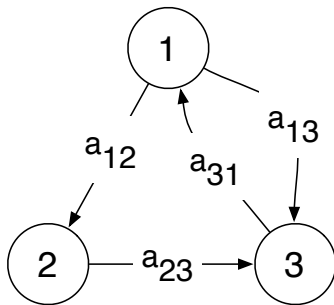
can have $s_2^* > s_1^*$, *even if* $a_{13} - a_{31} > a_{12} + a_{23}$.

Trying to achieve consistency, II

Idea: Use margin-based penalty[3]

$$\varphi(s, G) = \sum_{ij} \phi(s_i - s_j - a_{ij})$$

Inconsistent: Take $a_{ij} \equiv c$; can reduce to previous case

---

[3]Shashua and Levin 2002

Ranking is challenging

▶ Inconsistent in general
▶ Low noise settings
  ▶ Inconsistent for edge-based convex losses

  $$\varphi(s, G) = \sum_{ij} a_{ij} \phi(s_i - s_j)$$

  ▶ Inconsistent for margin-based convex losses

  $$\varphi(s, G) = \sum_{ij} \phi(s_i - s_j - a_{ij})$$

  ▶ Question: Do tractable consistent losses exist?

Yes.

A solution in the low noise setting

Recall reverse triangle inequality
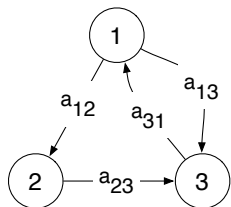
- ▶ Idea 1: make loss reduction proportional to weight difference $a_{ij} - a_{ji}$

- ▶ Idea 2: regularize to keep loss well-behaved

Theorem: For r strongly convex, following loss is consistent:

$$\varphi(s, G) = \sum_{ij} a_{ij}(s_j - s_i) + \sum_j r(s_j)$$

Consistency proof sketch

Write surrogate, take derivatives:

$$\sum_G p(G)\varphi(s, G) = \sum_{ij} a_{ij}(s_j - s_i) + \sum_j r(s_j)$$

$$\frac{\partial}{\partial s_i} = \sum_j (a_{ij} - a_{ji}) + r'(s_i) = 0$$

Simply note that r' is strictly increasing, see that

$$s_i > s_k \quad \Leftrightarrow \quad \sum_j a_{ij} - a_{ji} > \sum_j a_{kj} - a_{jk}$$

Last holds by low-noise assumption.

Experimental results

- MovieLens dataset:[4] 100,000 ratings for 1682 movies by 943 users

- Query is user u, results $X = \{1, \ldots, 1682\}$ are movies

- Scoring function: $f_i(x, u) = w^T \psi(x_i, u)$

- ψ maps from movie x_i and user u to features

- Per-user pair weight a_{ij}^u is difference of user's ratings for movies x_i, x_j

---

[4]GroupLens Lab, 2008

Surrogate risks

Losses based on pairwise comparisons

$$\text{Ours} \quad \sum_{i,j,u} a_{ij}^u w^T(\psi(x_j, u) - \psi(x_i, u)) + \theta \sum_{i,u} (w^T \psi(x_i, u))^2$$

$$\text{Hinge} \quad \sum_{i,j,u} a_{ij}^u \left[1 - w^T(\psi(x_j, u) - \psi(x_i, u)) \right]_+$$

$$\text{Logistic} \quad \sum_{i,j,u} a_{ij}^u \log \left(1 + e^{w^T(\psi(x_j,u) - \psi(x_i,u))} \right)$$

Experimental results

Test losses for each surrogate (standard error in parenthesis)

Num training pairs	Hinge	Logistic	Ours
20000	.478 (.008)	.479 (.010)	**.465** (.006)
40000	.477 (.008)	.478 (.010)	**.464** (.006)
80000	.480 (.007)	.478 (.009)	**.462** (.005)
120000	.477 (.008)	.477 (.009)	**.463** (.006)
160000	.474 (.007)	.474 (.007)	**.461** (.004)

Conclusions

- General theorem for consistency of ranking algorithms

- General inconsistency results as well as inconsistency results for several natural and commonly used losses, even in low noise settings

- Consistent loss for low noise settings

Open questions

▶ What are appropriate ranking losses? Click-based loss, ratings-based losses?

▶ Other consistent losses?

▶ Convergence rates?