Ranking with Ordered Weighted Pairwise Classification

N. Usunier, D. Buffoni, P. Gallinari

Laboratoire d'Informatique de Paris 6
Université Pierre et Marie Curie

ICML – June 15th 2009

Outline

Learning to Rank

- An example: $(z, \mathbf{y})$:
  - $z \Rightarrow (X_1(z), \ldots, X_{[z]}(z))$, the set of candidates,
  - $\mathbf{y}$, the set of indexes of *relevant* candidates.
- $X_j(z) \underset{\text{score}}{\xrightarrow{f}} f_j(z)$:

$\mathbf{y}$ | $\bar{\mathbf{y}}$	z	score	f
$+$	$X_{\sigma(1)}(z)$	$\uparrow$	$f_{\sigma(1)}(z)$
$-$	$X_{\sigma(2)}(z)$		$f_{\sigma(2)}(z)$
$\vdots$	$\vdots$		$\vdots$
$+$	$X_{\sigma([z])}(z)$		$f_{\sigma([z])}(z)$

where $f_{\sigma(j)}(z) \geq f_{\sigma(j+1)}(z)$

- Learn f with low *ranking error* using a training set $S = (z_i, \mathbf{y}_i)_{i=1}^m$.
- Applications: **Information Retrieval**, Multiclass classification, Recommender Systems, ...

Learning to Rank

- An example: $(z, \mathbf{y})$:
  - $z \Rightarrow (X_1(z), \dots, X_{[z]}(z))$, the set of candidates,
  - $\mathbf{y}$, the set of indexes of *relevant* candidates.
- $X_j(z) \underset{\text{score}}{\xrightarrow{f}} f_j(z)$:

$\mathbf{y} \mid \bar{\mathbf{y}}$	z	score	f
$+$	$X_{\sigma(1)}(z)$	$\uparrow$	$f_{\sigma(1)}(z)$
$-$	$X_{\sigma(2)}(z)$		$f_{\sigma(2)}(z)$
$\vdots$	$\vdots$		$\vdots$
$+$	$X_{\sigma([z])}(z)$		$f_{\sigma([z])}(z)$

where
$$f_{\sigma(j)}(z) \geq f_{\sigma(j+1)}(z)$$

- Learn f with low *ranking error* using a training set $S = (z_i, \mathbf{y}_i)_{i=1}^m$.

- Applications: **Information Retrieval**, Multiclass classification, Recommender Systems, ...

Learning to Rank

- An example: $(z, \mathbf{y})$:
  - $z \Rightarrow (X_1(z), \ldots, X_{[z]}(z))$, the set of candidates,
  - $\mathbf{y}$, the set of indexes of *relevant* candidates.
- $X_j(z) \underset{\text{score}}{\xrightarrow{f}} f_j(z)$:

$$\begin{array}{cccc} \mathbf{y} \mid \bar{\mathbf{y}} & \text{z} & \text{score} & f \\ + & X_{\sigma(1)}(z) & \uparrow & f_{\sigma(1)}(z) \\ - & X_{\sigma(2)}(z) & \mid & f_{\sigma(2)}(z) \\ \vdots & \vdots & \mid & \vdots \\ + & X_{\sigma([z])}(z) & \mid & f_{\sigma([z])}(z) \end{array} \qquad \text{where } f_{\sigma(j)}(z) \geq f_{\sigma(j+1)}(z)$$

- Learn f with low *ranking error* using a training set $S = (z_i, \mathbf{y}_i)_{i=1}^{m}$.
- Applications: **Information Retrieval**, Multiclass classification, Recommender Systems, ...

Background: the Pairwise Approach

- Based on pairwise comparisons: $\mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z)\right)$,
  for a relevant element $y \in \mathbf{y}$ and an irrelevant element $\bar{y} \in \bar{\mathbf{y}}$
- Aggregation of the pairwise errors:
  1. mean operator $\longrightarrow$ mean rank of the relevant elements,
  2. max operator $\longrightarrow$ check whether a relevant element is top-ranked.
- Convex losses for ranking: $\mathrm{I}(t \leq 0) \rightarrow \ell(t)$ with ℓ convex
  - hinge loss: Ranking SVM (mean), multiclass SVM (max)
  - exponential loss: RankBoost (mean)
- Information Retrieval: need high precision on the top of the list.
  The quality measures used in IR are not correlated to the errors defined in the pairwise approach.

Background: the Pairwise Approach

- Based on pairwise comparisons: $\mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z)\right)$, for a relevant element $y \in \mathbf{y}$ and an irrelevant element $\bar{y} \in \bar{\mathbf{y}}$
- Aggregation of the pairwise errors:
  1. mean operator $\longrightarrow$ mean rank of the relevant elements,
  2. max operator $\longrightarrow$ check whether a relevant element is top-ranked.
- Convex losses for ranking: $\mathrm{I}(t \leq 0) \rightarrow \ell(t)$ with ℓ convex
  - hinge loss: Ranking SVM (mean), multiclass SVM (max)
  - exponential loss: RankBoost (mean)
- Information Retrieval: need high precision on the top of the list. *The quality measures used in IR are not correlated to the errors defined in the pairwise approach.*

Background: the Pairwise Approach

- Based on pairwise comparisons: $\mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z)\right)$, for a relevant element $y \in \mathbf{y}$ and an irrelevant element $\bar{y} \in \bar{\mathbf{y}}$
- Aggregation of the pairwise errors:
  1. mean operator $\longrightarrow$ mean rank of the relevant elements,
  2. $\max$ operator $\longrightarrow$ check whether a relevant element is top-ranked.
- Convex losses for ranking: $\mathrm{I}(t \leq 0) \rightarrow \ell(t)$ with ℓ convex
  - hinge loss: Ranking SVM (mean), multiclass SVM ($\max$)
  - exponential loss: RankBoost (mean)
- Information Retrieval: need high precision on the top of the list. *The quality measures used in IR are not correlated to the errors defined in the pairwise approach.*

Learning to Rank for Information Retrieval

- Previous work:
  - no pairwise comparisons,
  - listwise loss functions, smooth approximations or convex upper bounds of IR metrics

ListNet, SoftRank, AdaRank, SVM^{map}, ...

- Contribution of the paper:
  - extension of the pairwise approach to learning to rank for IR,
  - a family of ranking error functions:
    - focus on the top of the list,
    - easier to upper bound than IR evaluation measures.
  - convex upper bounds on the ranking error,
    $\rightarrow$ Ordered Weighted Averaging Aggregation Operators
  - margin-based error bounds.

Learning to Rank for Information Retrieval

- Previous work:
  - no pairwise comparisons,
  - listwise loss functions, smooth approximations or convex upper bounds of IR metrics

ListNet, SoftRank, AdaRank, SVM^{map}, ...

- Contribution of the paper:
  - extension of the pairwise approach to learning to rank for IR,
  - a family of ranking error functions:
    - focus on the top of the list,
    - easier to upper bound than IR evaluation measures.
  - convex upper bounds on the ranking error,
    $\rightarrow$ Ordered Weighted Averaging Aggregation Operators
  - margin-based error bounds.

Framework

Ranking Error Functions

The ranking error of f on the example $(z, \mathbf{y})$ is defined as:

$$\mathtt{err}(f, z, \mathbf{y}) \overset{def}{=} \frac{1}{[\mathbf{y}]} \sum_{y \in \mathbf{y}} \Phi_{[\bar{\mathbf{y}}]} \left(rank_y(f, z, \mathbf{y}) \right)$$

Where:

- $rank_y(f, z, \mathbf{y}) \overset{def}{=} \sum_{\bar{y} \in \bar{\mathbf{y}}} \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z)\right)$
- for all n, $\forall k \in \{0..n\}$ $\Phi_n(k) \overset{def}{=} \sum_{j=1}^{k} \alpha_j^n$

  with $\alpha_1^n \geq \alpha_2^n \geq ... \geq \alpha_n^n \geq 0$ and $\sum_{j=1}^{n} \alpha_j^n = 1$

- The α_js are *decreasing*

  $\Rightarrow$ lower error on functions with high precision on the top of the list

Special cases

Ranking error function

$$\mathtt{err}(f, z, \mathbf{y}) \stackrel{def}{=} \frac{1}{[\mathbf{y}]} \sum_{y \in \mathbf{y}} \Phi_{[\bar{\mathbf{y}}]} \left(rank_y(f, z, \mathbf{y}) \right)$$

with $\forall k \in \{0..n\}\ \Phi_n(k) \stackrel{def}{=} \sum_{j=1}^{k} \alpha_j^n$

- If $\alpha_1^n = 1$ and $\forall j > 1, \alpha_j^n = 0$: top rank error,
- if $\alpha_j^n = \frac{1}{n}$ for all j: mean rank,
- Other possibilities:
  - optimize the mean rank over the top p percent of the list,
  - ...

Ordered Weighted Pairwise Classification (1/2)

Ordered Weighted Averaging (OWA) Aggregation Operator

- Definition from (Yager, 88),
- Let α such that $\sum_{j=1}^{n} \alpha_j = 1$ and $\alpha_j \geq 0$,
- $\forall (t_1, ..., t_n) \in \mathbb{R}^n, \ \underset{j \in \{1..n\}}{\text{owa}} \ t_j = \sum_{j=1}^{n} \alpha_j t_{\sigma(j)}$

  where $\forall j, t_{\sigma(j)} \geq t_{\sigma(j+1)}$.

$$(t_1, ..., t_n) \underset{\text{sort}}{\longrightarrow} (t_{\sigma(1)}, ..., t_{\sigma(n)}) \underset{\text{weighted sum}}{\longrightarrow} \sum_{j=1}^{n} \alpha_j t_{\sigma(j)}$$

- $\left.\begin{array}{l} \alpha_1 = 1 \\ \alpha_j = 0 \text{ if } j > 1 \end{array}\right\} \Rightarrow \underset{j \in \{1..n\}}{\text{owa}} \ t_j = t_{\sigma(1)} = \max_j t_j$.
- $\alpha_j = \frac{1}{n}$ for all $j \Rightarrow \underset{j \in \{1..n\}}{\text{owa}} \ t_j = \text{mean}_j \ t_j$.

Ordered Weighted Pairwise Classification (1/2)

Ordered Weighted Averaging (OWA) Aggregation Operator

- Definition from (Yager, 88),
- Let α such that $\sum_{j=1}^{n} \alpha_j = 1$ and $\alpha_j \geq 0$,
- $$\forall (t_1, ..., t_n) \in \mathbb{R}^n, \ \underset{j \in \{1..n\}}{\mathrm{owa}} \ t_j = \sum_{j=1}^{n} \alpha_j t_{\sigma(j)}$$

  where $\forall j, t_{\sigma(j)} \geq t_{\sigma(j+1)}$.

$$(t_1, ..., t_n) \ \underset{\text{sort}}{\longrightarrow} (t_{\sigma(1)}, ..., t_{\sigma(n)}) \ \underset{\text{weighted sum}}{\longrightarrow} \sum_{j=1}^{n} \alpha_j t_{\sigma(j)}$$

- $$\left.\begin{array}{l} \alpha_1 = 1 \\ \alpha_j = 0 \text{ if } j > 1 \end{array}\right\} \Rightarrow \underset{j \in \{1..n\}}{\mathrm{owa}} \ t_j = t_{\sigma(1)} = \max_j t_j.$$
- $\alpha_j = \frac{1}{n}$ for all $j \Rightarrow \underset{j \in \{1..n\}}{\mathrm{owa}} \ t_j = \mathrm{mean}_j \, t_j.$

Ordered Weighted Pairwise Classification (1/2)

Ordered Weighted Averaging (OWA) Aggregation Operator

- Definition from (Yager, 88),
- Let α such that $\sum_{j=1}^{n} \alpha_j = 1$ and $\alpha_j \geq 0$,
- $\forall (t_1, ..., t_n) \in \mathbb{R}^n, \ \underset{j \in \{1..n\}}{\text{owa}} \ t_j = \sum_{j=1}^{n} \alpha_j t_{\sigma(j)}$

  where $\forall j, t_{\sigma(j)} \geq t_{\sigma(j+1)}$.

$$(t_1, ..., t_n) \ \underset{\text{sort}}{\longrightarrow} (t_{\sigma(1)}, ..., t_{\sigma(n)}) \ \underset{\text{weighted sum}}{\longrightarrow} \sum_{j=1}^{n} \alpha_j t_{\sigma(j)}$$

- $\left.\begin{array}{l} \alpha_1 = 1 \\ \alpha_j = 0 \text{ if } j > 1 \end{array}\right\} \Rightarrow \underset{j \in \{1..n\}}{\text{owa}} \ t_j = t_{\sigma(1)} = \max_j t_j.$
- $\alpha_j = \frac{1}{n}$ for all $j \Rightarrow \underset{j \in \{1..n\}}{\text{owa}} \ t_j = \underset{j}{\text{mean}} \ t_j.$

Ordered Weighted Pairwise Classification (1/2)

Ordered Weighted Averaging (OWA) Aggregation Operator

- Definition from (Yager, 88),
- Let α such that $\sum_{j=1}^{n} \alpha_j = 1$ and $\alpha_j \geq 0$,
- $\forall (t_1, ..., t_n) \in \mathbb{R}^n, \underset{j \in \{1..n\}}{\text{owa}} t_j = \sum_{j=1}^{n} \alpha_j t_{\sigma(j)}$

  where $\forall j, t_{\sigma(j)} \geq t_{\sigma(j+1)}$.

$$(t_1, ..., t_n) \quad \xrightarrow[\text{sort}]{} (t_{\sigma(1)}, ..., t_{\sigma(n)}) \xrightarrow[\text{weighted sum}]{} \sum_{j=1}^{n} \alpha_j t_{\sigma(j)}$$

- $\left.\begin{array}{l} \alpha_1 = 1 \\ \alpha_j = 0 \text{ if } j > 1 \end{array}\right\} \Rightarrow \underset{j \in \{1..n\}}{\text{owa}} t_j = t_{\sigma(1)} = \max_j t_j$.
- $\alpha_j = \frac{1}{n}$ for all $j \Rightarrow \underset{j \in \{1..n\}}{\text{owa}} t_j = \text{mean}_j\, t_j$.

Ordered Weighted Pairwise Classification (1/2)

Ordered Weighted Averaging (OWA) Aggregation Operator

- Definition from (Yager, 88),
- Let α such that $\sum_{j=1}^{n} \alpha_j = 1$ and $\alpha_j \geq 0$,
- $\forall (t_1, ..., t_n) \in \mathbb{R}^n, \ \underset{j \in \{1..n\}}{\text{owa}} \ t_j = \sum_{j=1}^{n} \alpha_j t_{\sigma(j)}$

  where $\forall j, t_{\sigma(j)} \geq t_{\sigma(j+1)}$.

$$(t_1, ..., t_n) \ \underset{\text{sort}}{\longrightarrow} (t_{\sigma(1)}, ..., t_{\sigma(n)}) \ \underset{\text{weighted sum}}{\longrightarrow} \sum_{j=1}^{n} \alpha_j t_{\sigma(j)}$$

- $\left.\begin{array}{l} \alpha_1 = 1 \\ \alpha_j = 0 \text{ if } j > 1 \end{array}\right\} \Rightarrow \underset{j \in \{1..n\}}{\text{owa}} \ t_j = t_{\sigma(1)} = \max_j t_j$.
- $\alpha_j = \frac{1}{n}$ for all $j \Rightarrow \underset{j \in \{1..n\}}{\text{owa}} \ t_j = \text{mean}_j \ t_j$.

Ordered Weighted Pairwise Classification (1/2)

Ordered Weighted Averaging (OWA) Aggregation Operator

- Definition from (Yager, 88),
- Let α such that $\sum_{j=1}^{n} \alpha_j = 1$ and $\alpha_j \geq 0$,
- $\forall (t_1, ..., t_n) \in \mathbb{R}^n, \ \underset{j \in \{1..n\}}{\mathrm{owa}} \ t_j = \sum_{j=1}^{n} \alpha_j t_{\sigma(j)}$

  where $\forall j, t_{\sigma(j)} \geq t_{\sigma(j+1)}$.

$$(t_1, ..., t_n) \ \underset{\text{sort}}{\longrightarrow} (t_{\sigma(1)}, ..., t_{\sigma(n)}) \ \underset{\text{weighted sum}}{\longrightarrow} \sum_{j=1}^{n} \alpha_j t_{\sigma(j)}$$

- $\left. \begin{array}{l} \alpha_1 = 1 \\ \alpha_j = 0 \text{ if } j > 1 \end{array} \right\} \Rightarrow \underset{j \in \{1..n\}}{\mathrm{owa}} \ t_j = t_{\sigma(1)} = \max_j t_j$.
- $\alpha_j = \frac{1}{n}$ for all $j \Rightarrow \underset{j \in \{1..n\}}{\mathrm{owa}} \ t_j = \underset{j}{\mathrm{mean}} \ t_j$.

Ordered Weighted Pairwise Classification (2/2)

Previous definitions

- $\underset{j\in\{1..n\}}{\text{owa}}\ t_j = \sum_{j=1}^{n} \alpha_j t_{\sigma(j)}$ with $\forall j, t_{\sigma(j)} \geq t_{\sigma(j+1)}$,
- for non-increasing α_j^ns:
$$\mathtt{err}(f, z, \mathbf{y}) = \frac{1}{[\mathbf{y}]} \sum_{y\in\mathbf{y}} \Phi_{[\bar{\mathbf{y}}]}\left(rank_y(f, z, \mathbf{y})\right) \text{ with } \Phi_n(k) \overset{def}{=} \sum_{j=1}^{k} \alpha_j^n \, .$$

1. Fix an example $(z, \mathbf{y})$, a relevant element $y \in \mathbf{y}$, set the weights of the OWA operator to those of $\Phi_{[\bar{\mathbf{y}}]}$,
2. $$\left(\mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z)\right)\right)_{\bar{y}\in\bar{\mathbf{y}}} \xrightarrow[\text{sort}]{} (\underbrace{1, \dots, 1}_{k=rank_y(f,z,\mathbf{y})}, 0, \dots, 0) \xrightarrow[\text{weighted sum}]{} \sum_{j=1}^{k} \alpha_j$$

$$\Rightarrow \underset{\bar{y}\in\bar{\mathbf{y}}}{\text{owa}}\, \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z)\right) = \Phi_{[\bar{\mathbf{y}}]}\left(rank_y(f, z, \mathbf{y})\right)$$

Ordered Weighted Pairwise Classification (2/2)

Previous definitions

- $\operatorname*{owa}_{j \in \{1..n\}} t_j = \sum_{j=1}^{n} \alpha_j t_{\sigma(j)}$ with $\forall j, t_{\sigma(j)} \geq t_{\sigma(j+1)}$,
- for non-increasing α_j^ns:
  $\mathtt{err}(f, z, \mathbf{y}) = \frac{1}{[\mathbf{y}]} \sum_{y \in \mathbf{y}} \Phi_{[\bar{\mathbf{y}}]} \left(rank_y(f, z, \mathbf{y}) \right)$ with $\Phi_n(k) \stackrel{def}{=} \sum_{j=1}^{k} \alpha_j^n$.

1. Fix an example $(z, \mathbf{y})$, a relevant element $y \in \mathbf{y}$,
   set the weights of the OWA operator to those of $\Phi_{[\bar{\mathbf{y}}]}$,
2. $\left(\mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z) \right) \right)_{\bar{y} \in \bar{\mathbf{y}}} \xrightarrow{\text{sort}} (\underbrace{1, \ldots, 1}_{k = rank_y(f, z, \mathbf{y})}, 0, \ldots, 0) \xrightarrow{\text{weighted sum}} \sum_{j=1}^{k} \alpha_j$

$$\Rightarrow \operatorname*{owa}_{\bar{y} \in \bar{\mathbf{y}}} \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z) \right) = \Phi_{[\bar{\mathbf{y}}]}\left(rank_y(f, z, \mathbf{y}) \right)$$

Ordered Weighted Pairwise Classification (2/2)

Previous definitions

- $\underset{j \in \{1..n\}}{\mathrm{owa}}\, t_j = \sum_{j=1}^{n} \alpha_j t_{\sigma(j)}$ with $\forall j, t_{\sigma(j)} \geq t_{\sigma(j+1)}$,
- for non-increasing α_j^ns:
  $\mathtt{err}(f, z, \mathbf{y}) = \frac{1}{[\mathbf{y}]} \sum_{y \in \mathbf{y}} \Phi_{[\bar{\mathbf{y}}]} \left(rank_y(f, z, \mathbf{y}) \right)$ with $\Phi_n(k) \overset{def}{=} \sum_{j=1}^{k} \alpha_j^n$.

1. Fix an example $(z, \mathbf{y})$, a relevant element $y \in \mathbf{y}$,
   set the weights of the OWA operator to those of $\Phi_{[\bar{\mathbf{y}}]}$,
2. $$\Big(\mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z) \right) \Big)_{\bar{y} \in \bar{\mathbf{y}}} \xrightarrow[\text{sort}]{} \big(\underbrace{1, \ldots, 1}_{k = rank_y(f, z, \mathbf{y})}, 0, \ldots, 0 \big) \xrightarrow[\text{weighted sum}]{} \sum_{j=1}^{k} \alpha_j$$

$$\Rightarrow \underset{\bar{y} \in \bar{\mathbf{y}}}{\mathrm{owa}}\, \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z) \right) = \Phi_{[\bar{\mathbf{y}}]}\left(rank_y(f, z, \mathbf{y}) \right)$$

Ordered Weighted Pairwise Classification (2/2)

Previous definitions

- $\underset{j\in\{1..n\}}{\mathrm{owa}}\, t_j = \sum_{j=1}^{n} \alpha_j t_{\sigma(j)}$ with $\forall j, t_{\sigma(j)} \geq t_{\sigma(j+1)}$,
- for non-increasing α_j^ns:

  $\mathtt{err}(f, z, \mathbf{y}) = \frac{1}{[\mathbf{y}]} \sum_{y\in\mathbf{y}} \Phi_{[\bar{\mathbf{y}}]}\left(rank_y(f, z, \mathbf{y})\right)$ with $\Phi_n(k) \overset{def}{=} \sum_{j=1}^{k} \alpha_j^n$.

1. Fix an example $(z, \mathbf{y})$, a relevant element $y \in \mathbf{y}$,
   set the weights of the OWA operator to those of $\Phi_{[\bar{\mathbf{y}}]}$,
2. $\Big(\mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z)\right)\Big)_{\bar{y}\in\bar{\mathbf{y}}} \underset{\text{sort}}{\rightarrow} (\underbrace{1, ..., 1}_{k=rank_y(f,z,\mathbf{y})}, 0, ..., 0) \underset{\text{weighted sum}}{\rightarrow} \sum_{j=1}^{k} \alpha_j$

$$\Rightarrow \underset{\bar{y}\in\bar{\mathbf{y}}}{\mathrm{owa}}\, \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z)\right) = \Phi_{[\bar{\mathbf{y}}]}(rank_y(f, z, \mathbf{y}))$$

Ordered Weighted Pairwise Classification (2/2)

Previous definitions

- $\underset{j\in\{1..n\}}{\mathrm{owa}}\ t_j = \sum_{j=1}^{n} \alpha_j t_{\sigma(j)}$ with $\forall j, t_{\sigma(j)} \geq t_{\sigma(j+1)}$,
- for non-increasing α_j^ns:

  $\mathtt{err}(f,z,\mathbf{y}) = \frac{1}{[\mathbf{y}]} \sum_{y\in\mathbf{y}} \Phi_{[\bar{\mathbf{y}}]}\left(rank_y(f,z,\mathbf{y})\right)$ with $\Phi_n(k) \overset{def}{=} \sum_{j=1}^{k} \alpha_j^n$.

1. Fix an example $(z,\mathbf{y})$, a relevant element $y \in \mathbf{y}$,
   set the weights of the OWA operator to those of $\Phi_{[\bar{\mathbf{y}}]}$,
2. $$\left(\mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z)\right)\right)_{\bar{y}\in\bar{\mathbf{y}}} \xrightarrow[\text{sort}]{} (\underbrace{1,...,1}_{k=rank_y(f,z,\mathbf{y})},0,...,0) \xrightarrow[\text{weighted sum}]{} \sum_{j=1}^{k} \alpha_j$$

$$\Rightarrow \underset{\bar{y}\in\bar{\mathbf{y}}}{\mathrm{owa}}\ \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z)\right) = \Phi_{[\bar{\mathbf{y}}]}\left(rank_y(f,z,\mathbf{y})\right)$$

Ordered Weighted Pairwise Classification (2/2)

Previous definitions

- $\underset{j\in\{1..n\}}{\text{owa}} t_j = \sum_{j=1}^{n} \alpha_j t_{\sigma(j)}$ with $\forall j, t_{\sigma(j)} \geq t_{\sigma(j+1)}$,
- for non-increasing α_j^ns:
  $\mathtt{err}(f, z, \mathbf{y}) = \frac{1}{[\mathbf{y}]} \sum_{y\in\mathbf{y}} \Phi_{[\bar{\mathbf{y}}]} \left(rank_y(f, z, \mathbf{y}) \right)$ with $\Phi_n(k) \stackrel{def}{=} \sum_{j=1}^{k} \alpha_j^n$.

1. Fix an example $(z, \mathbf{y})$, a relevant element $y \in \mathbf{y}$,
   set the weights of the OWA operator to those of $\Phi_{[\bar{\mathbf{y}}]}$,
2. $\left(\mathrm{I} \left(f_y(z) \leq f_{\bar{y}}(z) \right) \right)_{\bar{y}\in\bar{\mathbf{y}}} \underset{\text{sort}}{\rightarrow} (\underbrace{1, \ldots, 1}_{k=rank_y(f,z,\mathbf{y})} , 0, \ldots, 0) \underset{\text{weighted sum}}{\rightarrow} \sum_{j=1}^{k} \alpha_j$

$$\Rightarrow \underset{\bar{y}\in\bar{\mathbf{y}}}{\text{owa}} \mathrm{I} \left(f_y(z) \leq f_{\bar{y}}(z) \right) = \Phi_{[\bar{\mathbf{y}}]} \left(rank_y(f, z, \mathbf{y}) \right)$$

Convex Losses with OWPC

- We have: $\mathtt{err}(f, z, \mathbf{y}) = \frac{1}{[\mathbf{y}]} \sum_{y \in \mathbf{y}} \operatorname*{owa}_{\bar{y} \in \bar{\mathbf{y}}} \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z)\right),$
- Additional properties:
  1. an OWA operator with non-increasing weights is convex,
  2. if moreover $t \mapsto \ell(t)$ is a convex upper bound on $\mathrm{I}(t \leq 0)$, then:

$$\operatorname*{owa}_{j \in \{1..n\}} \ell(t_j) \text{ is a convex upper bound on } \operatorname*{owa}_{j \in \{1..n\}} \mathrm{I}(t_j \leq 0)$$

Example

Consider:

- Linear score functions: $f_p(z) = \langle w, X_p(z) \rangle$
- hinge loss: $\ell(t) = [1 - t]_+$

$$L(w, z, \mathbf{y}) \stackrel{def}{=} \frac{1}{[\mathbf{y}]} \sum_{y \in \mathbf{y}} \operatorname*{owa}_{\bar{y} \in \bar{\mathbf{y}}} [1 - \langle w, X_y(z) - X_{\bar{y}}(z) \rangle]_+$$

is a convex upper bound on $\mathtt{err}(f, z, \mathbf{y})$

Convex Losses with OWPC

- We have: $\mathtt{err}(f, z, \mathbf{y}) = \frac{1}{[\mathbf{y}]} \sum_{y \in \mathbf{y}} \underset{\bar{y} \in \bar{\mathbf{y}}}{\mathrm{owa}} \, \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z)\right)$,
- Additional properties:
  1. an OWA operator with non-increasing weights is convex,
  2. if moreover $t \mapsto \ell(t)$ is a convex upper bound on $\mathrm{I}(t \leq 0)$, then:

$$\underset{j \in \{1..n\}}{\mathrm{owa}} \ \ell(t_j) \text{ is a convex upper bound on } \underset{j \in \{1..n\}}{\mathrm{owa}} \ \mathrm{I}(t_j \leq 0)$$

Example

Consider:

- Linear score functions: $f_p(z) = \langle w, X_p(z) \rangle$
- hinge loss: $\ell(t) = [1 - t]_+$

$$L(w, z, \mathbf{y}) \overset{def}{=} \frac{1}{[\mathbf{y}]} \sum_{y \in \mathbf{y}} \underset{\bar{y} \in \bar{\mathbf{y}}}{\mathrm{owa}} \left[1 - \langle w, X_y(z) - X_{\bar{y}}(z) \rangle\right]_+$$

is a convex upper bound on $\mathtt{err}(f, z, \mathbf{y})$

Convex Losses with OWPC

- We have: $\mathtt{err}(f, z, \mathbf{y}) = \frac{1}{[\mathbf{y}]} \sum_{y \in \mathbf{y}} \underset{\bar{y} \in \bar{\mathbf{y}}}{\mathrm{owa}}\, \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z)\right)$,
- Additional properties:
  1. an OWA operator with non-increasing weights is convex,
  2. if moreover $t \mapsto \ell(t)$ is a convex upper bound on $\mathrm{I}(t \leq 0)$, then:

$$\underset{j \in \{1..n\}}{\mathrm{owa}} \ell(t_j) \text{ is a convex upper bound on } \underset{j \in \{1..n\}}{\mathrm{owa}} \mathrm{I}(t_j \leq 0)$$

Example

Consider:

- Linear score functions: $f_p(z) = \langle w, X_p(z) \rangle$
- hinge loss: $\ell(t) = [1 - t]_+$

$$L(w, z, \mathbf{y}) \stackrel{def}{=} \frac{1}{[\mathbf{y}]} \sum_{y \in \mathbf{y}} \underset{\bar{y} \in \bar{\mathbf{y}}}{\mathrm{owa}} [1 - \langle w, X_y(z) - X_{\bar{y}}(z) \rangle]_+$$

is a convex upper bound on $\mathtt{err}(f, z, \mathbf{y})$

Convex Losses with OWPC

- We have: $\mathtt{err}(f, z, \mathbf{y}) = \frac{1}{[\mathbf{y}]} \sum_{y \in \mathbf{y}} \underset{\bar{y} \in \bar{\mathbf{y}}}{\mathrm{owa}}\, \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z)\right),$
- Additional properties:
  1. an OWA operator with non-increasing weights is convex,
  2. if moreover $t \mapsto \ell(t)$ is a convex upper bound on $\mathrm{I}(t \leq 0)$, then:

$$\underset{j \in \{1..n\}}{\mathrm{owa}} \ell(t_j) \text{ is a convex upper bound on } \underset{j \in \{1..n\}}{\mathrm{owa}} \mathrm{I}(t_j \leq 0)$$

Example

Consider:

- Linear score functions: $f_p(z) = \langle w, X_p(z) \rangle$
- hinge loss: $\ell(t) = [1 - t]_+$

$$L(w, z, \mathbf{y}) \stackrel{def}{=} \frac{1}{[\mathbf{y}]} \sum_{y \in \mathbf{y}} \underset{\bar{y} \in \bar{\mathbf{y}}}{\mathrm{owa}} [1 - \langle w, X_y(z) - X_{\bar{y}}(z) \rangle]_+$$

is a convex upper bound on $\mathtt{err}(f, z, \mathbf{y})$

Convex Losses with OWPC

- We have: $\mathtt{err}(f, z, \mathbf{y}) = \frac{1}{[\mathbf{y}]} \sum_{y \in \mathbf{y}} \underset{\bar{y} \in \bar{\mathbf{y}}}{\mathrm{owa}} \, \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z)\right)$,
- Additional properties:
  1. an OWA operator with non-increasing weights is convex,
  2. if moreover $t \mapsto \ell(t)$ is a convex upper bound on $\mathrm{I}(t \leq 0)$, then:

$$\underset{j \in \{1..n\}}{\mathrm{owa}} \ \ell(t_j) \text{ is a convex upper bound on } \underset{j \in \{1..n\}}{\mathrm{owa}} \ \mathrm{I}(t_j \leq 0)$$

Example

Consider:

- Linear score functions: $f_p(z) = \langle w, X_p(z) \rangle$
- hinge loss: $\ell(t) = [1 - t]_+$

$$L(w, z, \mathbf{y}) \ \stackrel{def}{=} \ \frac{1}{[\mathbf{y}]} \sum_{y \in \mathbf{y}} \underset{\bar{y} \in \bar{\mathbf{y}}}{\mathrm{owa}} \left[1 - \langle w, X_y(z) - X_{\bar{y}}(z) \rangle\right]_+$$

is a convex upper bound on $\mathtt{err}(f, z, \mathbf{y})$

Special Cases and Optimization

Regularized empirical risk

$$\min_w \frac{1}{2}||w||^2 + C \sum_{(z,\mathbf{y}) \in S} L(w, z, \mathbf{y})$$

where $L(w, z, \mathbf{y}) \stackrel{def}{=} \frac{1}{[\mathbf{y}]} \sum_{y \in \mathbf{y}} \underset{\bar{y} \in \bar{\mathbf{y}}}{\text{owa}} \left[1 - \langle w, X_y(z) - X_{\bar{y}}(z) \rangle\right]_+$

- Special cases:
  1. when $[\mathbf{y}] = 1$ and owa is the $\max$
     $\longrightarrow$ SVM for multiclass classification (Crammer & Singer, 2001),
  2. when owa is the mean $\longrightarrow \sim$ Ranking SVM (e.g. Joachims, 2002),
  3. new possibilities, for instance:
     optimize the mean rank of the relevant elements on the top p% of the list.
- Training:
  - can be written as a structural SVM (e.g. (Tsochantaridis et al., 2005)),
  - optimization with existing methods (e.g. LaRank (Bordes et al., 2007)).

Special Cases and Optimization

Regularized empirical risk

$$\min_w \frac{1}{2}||w||^2 + C \sum_{(z,\mathbf{y})\in S} L(w,z,\mathbf{y})$$

where $L(w,z,\mathbf{y}) \stackrel{def}{=} \frac{1}{[\mathbf{y}]} \sum_{y\in\mathbf{y}} \operatorname{owa}_{\bar{y}\in\bar{\mathbf{y}}} [1 - \langle w, X_y(z) - X_{\bar{y}}(z)\rangle]_+$

- Special cases:
  1. when $[\mathbf{y}] = 1$ and owa is the $\max$
     $\longrightarrow$ SVM for multiclass classification (Crammer & Singer, 2001),
  2. when owa is the mean $\longrightarrow \sim$ Ranking SVM (e.g. Joachims, 2002),
  3. new possibilities, for instance:
     optimize the mean rank of the relevant elements on the top p% of the list.
- Training:
  - can be written as a structural SVM (e.g. (Tsochantaridis et al., 2005)),
  - optimization with existing methods (e.g. LaRank (Bordes et al., 2007)).

Margin-Based Generalization Analysis

- We have: $\mathtt{err}(f, z, \mathbf{y}) = \frac{1}{[\mathbf{y}]} \sum_{y \in \mathbf{y}} \underset{\bar{y} \in \bar{\mathbf{y}}}{\mathrm{owa}} \, \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z)\right),$
- Assume for simplicity $[\mathbf{y}] = 1$ and $[\bar{\mathbf{y}}]$ is constant,
- assume S contains m examples drawn i.i.d. according to some distribution $\mathcal{D}$.
- Then, for any $\gamma > 0$, any $\delta \in (0, 1]$, we have:

$$\forall f \in \mathcal{F}, \underset{(z,\mathbf{y}) \sim \mathcal{D}}{\mathbb{E}} \underset{\bar{y} \in \bar{\mathbf{y}}}{\mathrm{owa}} \, \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z)\right) \leq \underset{(z,\mathbf{y}) \sim S}{\hat{\mathbb{E}}} \underset{\bar{y} \in \bar{\mathbf{y}}}{\mathrm{owa}} \, \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z) + \gamma\right) + \sqrt{\frac{\ln\left(\mathcal{N}(\mathcal{H}, \gamma/2)/\delta\right)}{2m}}$$

where $\mathcal{H} \stackrel{def}{=} \{(z, p), (z, q) \mapsto f_p(z) - f_q(z) | f \in \mathcal{F}\}$,
and $\mathcal{N}(\mathcal{H}, \gamma/2)$ are the L_∞ covering numbers of $\mathcal{H}$ at $\gamma/2$.

Margin-Based Generalization Analysis

- We have: $\mathtt{err}(f, z, \mathbf{y}) = \frac{1}{[\mathbf{y}]} \sum_{y \in \mathbf{y}} \operatorname{owa}_{\bar{y} \in \bar{\mathbf{y}}} \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z)\right)$,
- Assume for simplicity $[\mathbf{y}] = 1$ and $[\bar{\mathbf{y}}]$ is constant,
- assume S contains m examples drawn i.i.d. according to some distribution $\mathcal{D}$.
- Then, for any $\gamma > 0$, any $\delta \in (0, 1]$, we have:

$$\forall f \in \mathcal{F}, \quad \mathop{\mathbb{E}}_{(z,\mathbf{y}) \sim \mathcal{D}} \operatorname{owa}_{\bar{y} \in \bar{\mathbf{y}}} \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z)\right) \leq \mathop{\hat{\mathbb{E}}}_{(z,\mathbf{y}) \sim S} \operatorname{owa}_{\bar{y} \in \bar{\mathbf{y}}} \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z) + \gamma\right) + \sqrt{\frac{\ln\left(\mathcal{N}(\mathcal{H}, \gamma/2)/\delta\right)}{2m}}$$

where $\mathcal{H} \stackrel{def}{=} \{(z, p), (z, q) \mapsto f_p(z) - f_q(z) | f \in \mathcal{F}\}$,
and $\mathcal{N}(\mathcal{H}, \gamma/2)$ are the L_∞ covering numbers of $\mathcal{H}$ at $\gamma/2$.

Margin-Based Generalization Analysis

- We have: $\mathtt{err}(f, z, \mathbf{y}) = \frac{1}{[\mathbf{y}]} \sum_{y \in \mathbf{y}} \underset{\bar{y} \in \bar{\mathbf{y}}}{\mathrm{owa}}\, \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z)\right)$,
- Assume for simplicity $[\mathbf{y}] = 1$ and $[\bar{\mathbf{y}}]$ is constant,
- assume S contains m examples drawn i.i.d. according to some distribution $\mathcal{D}$.
- Then, for any $\gamma > 0$, any $\delta \in (0, 1]$, we have:

$$\forall f \in \mathcal{F}, \underset{(z,\mathbf{y}) \sim \mathcal{D}}{\mathbb{E}} \underset{\bar{y} \in \bar{\mathbf{y}}}{\mathrm{owa}}\, \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z)\right) \leq \underset{(z,\mathbf{y}) \sim S}{\hat{\mathbb{E}}} \underset{\bar{y} \in \bar{\mathbf{y}}}{\mathrm{owa}}\, \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z) + \gamma\right) + \sqrt{\frac{\ln\left(\mathcal{N}(\mathcal{H}, \gamma/2)/\delta\right)}{2m}}$$

where $\mathcal{H} \stackrel{def}{=} \left\{(z, p), (z, q) \mapsto f_p(z) - f_q(z) | f \in \mathcal{F}\right\}$,
and $\mathcal{N}(\mathcal{H}, \gamma/2)$ are the L_∞ covering numbers of $\mathcal{H}$ at $\gamma/2$.

Margin-Based Generalization Analysis

- We have: $\texttt{err}(f,z,\mathbf{y}) = \frac{1}{[\mathbf{y}]} \sum_{y\in\mathbf{y}} \underset{\bar{y}\in\bar{\mathbf{y}}}{\mathrm{owa}}\, \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z)\right)$,
- Assume for simplicity $[\mathbf{y}] = 1$ and $[\bar{\mathbf{y}}]$ is constant,
- assume S contains m examples drawn i.i.d. according to some distribution $\mathcal{D}$.
- Then, for any $\gamma > 0$, any $\delta \in (0,1]$, we have:

$$\forall f \in \mathcal{F}, \underset{(z,\mathbf{y})\sim\mathcal{D}}{\mathbb{E}} \underset{\bar{y}\in\bar{\mathbf{y}}}{\mathrm{owa}}\, \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z)\right) \leq \underset{(z,\mathbf{y})\sim S}{\hat{\mathbb{E}}} \underset{\bar{y}\in\bar{\mathbf{y}}}{\mathrm{owa}}\, \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z) + \gamma\right) + \sqrt{\frac{\ln\left(\mathcal{N}(\mathcal{H},\gamma/2)/\delta\right)}{2m}}$$

where $\mathcal{H} \stackrel{def}{=} \{(z,p),(z,q) \mapsto f_p(z) - f_q(z) | f \in \mathcal{F}\}$,
and $\mathcal{N}(\mathcal{H},\gamma/2)$ are the L_∞ covering numbers of $\mathcal{H}$ at $\gamma/2$.

Margin-Based Generalization Analysis

- We have: $\texttt{err}(f, z, \mathbf{y}) = \frac{1}{[\mathbf{y}]} \sum_{y \in \mathbf{y}} \underset{\bar{y} \in \bar{\mathbf{y}}}{\mathrm{owa}} \, \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z)\right)$,
- Assume for simplicity $[\mathbf{y}] = 1$ and $[\bar{\mathbf{y}}]$ is constant,
- assume S contains m examples drawn i.i.d. according to some distribution $\mathcal{D}$.
- Then, for any $\gamma > 0$, any $\delta \in (0, 1]$, we have:

$$\forall f \in \mathcal{F}, \underset{(z,\mathbf{y}) \sim \mathcal{D}}{\mathbb{E}} \underset{\bar{y} \in \bar{\mathbf{y}}}{\mathrm{owa}} \, \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z)\right) \leq \underset{(z,\mathbf{y}) \sim S}{\hat{\mathbb{E}}} \underset{\bar{y} \in \bar{\mathbf{y}}}{\mathrm{owa}} \, \mathrm{I}\left(f_y(z) \leq f_{\bar{y}}(z) + \gamma\right) + \sqrt{\frac{\ln\left(\mathcal{N}(\mathcal{H}, \gamma/2)/\delta\right)}{2m}}$$

where $\mathcal{H} \overset{def}{=} \left\{(z, p), (z, q) \mapsto f_p(z) - f_q(z) | f \in \mathcal{F}\right\}$,
and $\mathcal{N}(\mathcal{H}, \gamma/2)$ are the L_∞ covering numbers of $\mathcal{H}$ at $\gamma/2$.

Experiments

Benchmark datasets (Liu et al., 2007)

- Letor 3.0: 6 datasets from TREC competitions (.GOV document collection),
- train/validation/test sets for each dataset,
- from 50 to 150 queries, ~ 1000 documents per query,
- 60 features for the joint (document, query) representation.

Experimental protocol

- Evaluation measures: MAP, NDCG, Precision
- hyperparameter selection: best MAP on the validation set,
- weights of the OWA operator fixed to $\alpha_j \propto 1/j$.

Test Performance (MAP)

Mean Average Precision (MAP)

Average Precision (for a given query):

$$\frac{1}{[\mathbf{y}]} \sum_{y \in \mathbf{y}} \frac{1 + \text{\# rel. docs before } y}{1 + \text{\# docs before } y}$$

MAP: Mean over the queries of the Average Precision

	TD03	TD04	HP03	HP04	NP03	NP04
RSVM	0.263	0.224	0.741	0.668	**0.696**	0.659
SVM^{map}	0.245	0.205	0.742	0.718	0.687	0.662
Adarank	0.228	0.219	**0.771**	0.722	0.678	0.622
ListNet	0.275	0.223	0.766	0.690	0.690	0.672
OWPC	**0.290**	**0.229**	0.757	**0.726**	0.685	**0.683**

Test Performance (Prec@1)

Precision at 1

percentage of queries for which the first document retrieved is relevant

	TD03	TD04	HP03	HP04	NP03	NP04
RSVM	0.320	0.413	0.693	0.573	**0.580**	0.507
SVM^{map}	0.320	0.293	0.713	**0.627**	0.560	0.520
Adarank	0.360	0.427	0.713	0.587	0.560	0.507
ListNet	0.400	0.360	**0.720**	0.600	0.567	0.533
OWPC	**0.440**	**0.453**	**0.720**	0.613	**0.580**	**0.560**

Conclusion

- The OWPC approach:
  - defines convex losses for ranking,
  - generalizes the classical pairwise approaches,
  - allows to learn score functions with high precision on the top of the list,
- margin-based generalization errors,
- state-of-the-art results on Letor 3.0.

- Perspectives:
  - Extension to real-valued relevance judgements,
  - Learning the weights of the OWA operator depending on the task.