Statistical Analysis of Bipartite and Multipartite Ranking by Convex Risk Minimization

Yoonkyung Lee*
Department of Statistics
The Ohio State University
*joint work with Kazuki Uematsu

August 26-27, 2010
Summer School at CIMAT, Guanajuato, Mexico
Ranking

- Aims to order a set of objects or instances reflecting their underlying utility, relevance or quality.
- Has gained increasing attention in machine learning and information retrieval for website search and document retrieval.

(Source: Google images of “ranking”)
Data for Ranking

```
  object_1 | positive
  object_2 | negative
  
  
  object_{n-1} | positive
  object_n    | negative
```

How do we order objects so that positive cases are ranked higher than negative cases?
Main Questions

- How to rank?
- How to formulate ranking problems?
- What loss criteria to use?
- What is the best ranking function given a criterion?
- How is it related to the underlying distribution?
Notation

- $X \in \mathcal{X}$: an instance to rank
- $Y \in \mathcal{Y} = \{1, \cdots, k\}$: an ordinal response in multipartite ranking (bipartite ranking when $k = 2$)
- $f: \mathcal{X} \rightarrow \mathbb{R}$: a ranking function

In bipartite ranking,

- $\mathcal{Y} = \{1, -1\}$
- $X$ and $X'$ denote a positive instance ($Y = 1$) and a negative instance ($Y = -1$) with pdf $g_+$ and $g_-$. 
Bipartite Ranking

- Training data: $n$ pairs of $(X, Y)$ from $\mathcal{X} \times \{1, -1\}$
  - $\{x_i, i = 1, \ldots, n_+\}$: $n_+$ positive instances
  - $\{x'_j, j = 1, \ldots, n_-\}$: $n_-$ negative instances

- Learn a ranking function $f$ which generally places a positive instance ahead of a negative instance.

- How to measure goodness of $f$?
Outline

- Loss criteria for bipartite ranking
- Optimal ranking function for consistency
- Convex risk minimization for ranking
- Numerical illustration
- Extension to multipartite ranking
Bipartite Ranking Loss

- For a pair of a positive $x$ and a negative $x'$,

$$\ell_0(f; x, x') = I(f(x) < f(x')) + \frac{1}{2} I(f(x) = f(x'))$$

- Note the invariance of the loss under order-preserving transformations.

- Find $f$ minimizing the ranking risk

$$R_{n_+, n_-}(f) = \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} \ell_0(f; x_i, x'_j)$$
Ranking Risk and AUC

ROC curve of $f$:
true positive rate $\text{TPR}(r) = \frac{|\{x_i | f(x_i) > r\}|}{n_+}$ vs
false positive rate $\text{FPR}(r) = \frac{|\{x_j' | f(x_j') > r\}|}{n_-}$

- $\text{AUC}(f)$: area under ROC curve of $f$
- Empirical ranking risk $R_{n_+,n_-}(f) = 1 - \text{AUC}(f)$
- Minimizing ranking error is equivalent to maximizing AUC.
Theorem
Define \( f_0^*(x) = \frac{g_+(x)}{g_-(x)} \), and let \( R_0(f) = E(\ell_0(f; X, X')) \) denote the ranking risk of \( f \) under the bipartite ranking loss. Then for any ranking function \( f \),

\[
R_0(f_0^*) \leq R_0(f).
\]

Remark
Connection to classification:

\[
P(Y = 1|X = x) = \frac{\pi g_+(x)}{\pi g_+(x) + (1 - \pi)g_-(x)} = \frac{f_0^*(x)}{f_0^*(x) + (1 - \pi)/\pi},
\]

where \( \pi = P(Y = 1) \).
Bipartite Ranking Loss

$$\ell_0(f; x, x') = I(f(x) - f(x') < 0) + \frac{1}{2}I(f(x) - f(x') = 0)$$
Convex Surrogate Loss for Bipartite Ranking

- Exponential loss in RankBoost (Freund et al. 2003):
  \[ \ell(f; x, x') = \exp(-(f(x) - f(x'))) \]

- Hinge loss in RankSVM (Joachims 2002) and AUCSVM (Rakotomamonjy 2004, Brefeld and Scheffer 2005):
  \[ \ell(f; x, x') = (1 - (f(x) - f(x')))_{+} \]

- Logistic loss (cross entropy) in RankNet (Burges et al. 2005):
  \[ \ell(f; x, x') = \log(1 + \exp(-(f(x) - f(x'))) \]
Optimal Ranking Function Under Convex Loss

**Theorem**

Suppose that $\ell$ is differentiable, $\ell'(s) < 0$ for all $s \in \mathbb{R}$, and $\ell'(-s)/\ell'(s) = \exp(s/\alpha)$ for some positive constant $\alpha$. Let $f^*$ be the best ranking function $f$ minimizing $R_\ell(f) = E[\ell(f; X, X')]$. Then

$$f^*(x) = \alpha \log(g_+(x)/g_-(x)) \quad \text{up to a constant.}$$

**Remark**

- For RankBoost, $\ell(s) = e^{-s}$, and $\ell'(-s)/\ell'(s) = e^{2s}$. $f^*(x) = \frac{1}{2} \log(g_+(x)/g_-(x))$.
- For RankNet, $\ell(s) = \log(1 + e^{-s})$, and $\ell'(-s)/\ell'(s) = e^s$. $f^*(x) = \log(g_+(x)/g_-(x))$.
- See also Clémençon et al. (2008).
Theorem
Suppose that $\ell$ is convex, and the subdifferential of $\ell$ at zero contains only negative values.

(i) If $\frac{g_+(x)}{g_-(x)} > \frac{g_+(x')}{g_-(x')}$, then $f^*(x) \geq f^*(x')$ with probability 1.

(ii) If $\ell$ is differentiable and $\frac{g_+(x)}{g_-(x)} > \frac{g_+(x')}{g_-(x')}$,
    then $f^*(x) > f^*(x')$ with probability 1.

Remark
For RankSVM, $\ell(s) = (1 - s)_+$ with singularity at $s = 1$. Since it is differentiable at 0 with $\ell'(0) = -1$, (i) applies. However, there could be potential ties in ranking.
Toy Example: RankSVM

- $\mathcal{X} = \{x_1, x_2, x_3\}$ and $\frac{p_+(x_1)}{p_-(x_1)} < \frac{p_+(x_2)}{p_-(x_2)} < \frac{p_+(x_3)}{p_-(x_3)}$

- Identify $f^*$ minimizing $E(1 - (f(X) - f(X')))_+$.

- Let $s_1 = f(x_2) - f(x_1)$ and $s_2 = f(x_3) - f(x_2)$, and take the risk as a function of $s_1$ and $s_2$.

- Let $\Delta_{12} = \frac{p_-(x_1)}{p_+(x_1)} - \left(\frac{p_-(x_2)}{p_+(x_2)} + \frac{p_-(x_3)}{p_+(x_2)}\right)$ and  
  $\Delta_{23} = \frac{p_+(x_3)}{p_-(x_3)} - \left(\frac{p_+(x_2)}{p_-(x_2)} + \frac{p_+(x_1)}{p_-(x_2)}\right)$.

- For $s_1^* = f^*(x_2) - f^*(x_1)$ and $s_2^* = f^*(x_3) - f^*(x_2)$
  
  (i) if $\Delta_{12} > 0$ and $\Delta_{23} > 0$, $(s_1^*, s_2^*) = (1, 1)$.
  (ii) if $\Delta_{23} < 0$ and $p_+(x_2) > p_-(x_2)$, $(s_1^*, s_2^*) = (1, 0)$.
  (iii) if $\Delta_{12} < 0$ and $p_+(x_2) < p_-(x_2)$, $(s_1^*, s_2^*) = (0, 1)$. 
Toy Example: Trichotomy of Probability Space

- \( p_+(x_1) = p_+(x_2) = 0.2, \) 
  \( p_+(x_3) = 0.6 \)
- Partition of \( p_\) based on \((s^*_1, s^*_2)\)
  (i) \( \Delta_{12} > 0: \)
  \( p_-(x_1) > 1/2 \)
  (ii) \( \Delta_{23} > 0: \)
  \( 2p_-(x_1) + 5p_-(x_2) > 2 \)
  (iii) \( p_+(x_2) > p_-(x_2): \)
  \( p_-(x_2) < 0.2 \)
- \( f^* \) is not unique at the intersections.
Theorem

Let \( f^* = \arg \min_f E(1 - (f(X) - f(X')))_+ \).

(i) \( f^* \) may not be unique.

(ii) When \( \mathcal{X} \) is countable, there exists a step function with each step size equal to 1 among \( f^* \).

(iii) When \( \mathcal{X} \) is uncountable, there exists a step function with each step size equal to 1 whose risk is arbitrarily close to the minimum risk.
Numerical Illustration

- Simulation setting:
  \[ X \sim N(1, 1) \text{ and } X' \sim N(-1, 1) \]
  \[ \log(g_+(x)/g_-(x)) = 2x \]

- Generate \( \{x_i, \ i = 1, \ldots, n\} \) and \( \{x'_j, \ j = 1, \ldots, n\} \)

- \( n \ (= n_+ = n_-) \): sample size for each category

- Apply RankBoost and RankSVM.
RankBoost

- $n = 2000$
- The dotted line: $f^*(x) = x$ (the theoretically optimal ranking function)

- The solid line: the estimated ranking function with 200 weak rankings

- Weak ranking: $f_\theta(x) = I(x > \theta)$ with $\theta$ equal to a sample point
For $f^*$, we do not have an explicit form, but can consider step functions taking integers and find the risk minimizer given number of steps.

For $f(x) = \sum_{i=1}^{k+1} i \cdot I(a_{i-1} < x \leq a_i)$ with $a_0 = -\infty$ and $a_{k+1} = \infty$, we can compute the risk analytically and identify the jump discontinuities $a_i$ of the step function with minimal risk.

$$g_+(a_i)(1 - G_-(a_{i-1})) = g_-(a_i)G_+(a_{i+1})$$

for $i = 1, \ldots, k$

where $G_+$ and $G_-$ are the cdf of $X$ and $X'$. 
Finite Sample Version of RankSVM

- AUC maximizing SVM (Brefeld and Scheffer 2005)

\[
\min_{f \in \mathcal{F}} \quad C \sum_{i,j} (1 - (f(x_i) - f(x'_j)))_+ + \|f\|^2
\]

- When \( f \in \mathcal{H}_K \) with a kernel \( K \),

\[
\hat{f}(x) = \sum_{i,j} c_{ij} (K(x_i, x) - K(x'_j, x))
\]

- Gaussian kernel \( K(x, x') = \exp(-{(x - x')^2}/2\sigma^2) \) is used \((\sigma^2 = 0.15)\)
Figure: The solid lines are the estimated ranking functions, and the dotted lines are step functions with minimal risk.
Extension to Multipartite Ranking

- In general ($k \geq 2$), for a pair of $(x, y)$ and $(x', y')$ with $y > y'$, define a loss of ranking function $f$ as

$$\ell_0(f; x, x', y, y') = c_{y'y} l(f(x) < f(x')) + \frac{1}{2} c_{y'y} l(f(x) = f(x'))$$

where $c_{y'y}$ is the cost of misranking a pair of $y$ and $y'$. Waegeman et al. (2008)

- Again, $\ell_0$ is invariant under order-preserving transformations.
Multipartite Ranking Risk

- \( R_0(f; c) \), the risk of \( f \) is

\[
\sum_{1 \leq j < i \leq k} c_{ji} P(f(X) < f(X')| Y = i, Y' = j) P( Y = i, Y' = j).
\]

\[
+ \frac{1}{2} c_{ji} P(f(X) = f(X')| Y = i, Y' = j) P( Y = i, Y' = j).
\]

- If \( c_{ji} = I(j < i) \), then \( R_0(f; c) \) is the probability of misranking a pair of objects by \( f \). Minimizing \( R_0(f; c) \) is equivalent to maximizing the expected pairwise AUC.
Optimal Ranking Function When $k = 3$ or $k = 4$

Theorem
(i) When $k = 3$, let

$$f_0^*(x) = \frac{c_{12}P(Y = 2|x) + c_{13}P(Y = 3|x)}{c_{13}P(Y = 1|x) + c_{23}P(Y = 2|x)}.$$ 

Then for any ranking function $f$,

$$R_0(f_0^*; c) \leq R_0(f; c).$$

(ii) When $k = 4$, let

$$f_0^*(x) = \frac{c_{12}P(Y = 2|x) + c_{13}P(Y = 3|x) + c_{14}P(Y = 4|x)}{c_{14}P(Y = 1|x) + c_{24}P(Y = 2|x) + c_{34}P(Y = 3|x)}.$$ 

If $c_{12}c_{34} + c_{14}c_{23} = c_{13}c_{24}$, then for any ranking function $f$,

$$R_0(f_0^*; c) \leq R_0(f; c).$$

Remark
With proper conditions on $c_{ij}$, results for general $k$ can be derived.
A typical form of loss in ordinal regression for $f$ with thresholds $\{\theta_j\}_{j=1}^{k-1}$:

$$\ell(f, \{\theta_j\}_{j=1}^{k-1}; x, y) = \ell(f(x) - \theta_{y-1}) + \ell(\theta_y - f(x)),$$

where $\theta_0 = -\infty$ and $\theta_k = \infty$.

Support Vector Ordinal Regression (Herbrich et al. 1999):

$$\ell(s) = (1 - s)_+$$
Relation to Ordinal Regression

- **ORBoost (Lin and Li 2006):** \( \ell(s) = \exp(-s) \)

\[
f^*(x) = \frac{1}{2} \log \frac{\sum_{j=1}^{k-1} P(Y = j+1|x) \exp(\theta_j^*)}{\sum_{j=1}^{k-1} P(Y = j|x) \exp(-\theta_j^*)}
\]

where \( \theta_j^* \) are constants depending on \( P_{X,Y} \).

- When \( k = 3 \),

\[
f^*(x) = \frac{1}{2} \log \frac{P(Y = 2|x) + \exp(\theta_2^* - \theta_1^*)P(Y = 3|x)}{\exp(\theta_2^* - \theta_1^*)P(Y = 1|x) + P(Y = 2|x)} + \frac{1}{2}(\theta_1^* + \theta_2^*).
\]

Hence, \( f^* \) preserves the ordering of \( f_0^* \) with \( c_{12} = c_{23} = 1 \) and \( c_{13} = e^{\theta_2^* - \theta_1^*} \).
Proportional Odds Model

- Cumulative logits (McCullagh 1980)

\[
\log \frac{P(Y \leq j|x)}{P(Y > j|x)} = f(x) - \theta_j,
\]

where \(-\infty = \theta_0 < \theta_1 < \ldots < \theta_{k-1} < \theta_k = \infty\).

- The log likelihood of \(\{y_i\}_{i=1}^n\) given \(\{x_i\}_{i=1}^n\):

\[
\sum_{i=1}^n \log \left( \frac{1}{1 + \exp(f(x_i) - \theta_{y_i})} - \frac{1}{1 + \exp(f(x_i) - \theta_{y_i-1})} \right)
\]

\[
= \sum_{i=1}^n \log(1 - \exp(\theta_{y_i-1} - \theta_{y_i})) - \left( \log(1 + \exp(f(x_i) - \theta_{y_i})) + \log(1 + \exp(\theta_{y_i-1} - f(x_i))) \right)
\]

- Given \(\{\theta_j\}_{j=1}^{k-1}\), maximizing the log likelihood amounts to ordinal regression with \(\ell(s) = \log(1 + e^{-s})\). (Rennie 2006)
When $k = 3$, given $\theta_1$ and $\theta_2$, let $f^*$ be the minimizer of the 'logit' risk. We can show

$$\exp(f^*(x)) = \frac{q(x) - 1 + \sqrt{(q(x) - 1)^2 + 4 \exp(\theta_1 - \theta_2)q(x)}}{2 \exp(-\theta_2)},$$

where $q(x) = \frac{P(Y = 2|x) + P(Y = 3|x)}{P(Y = 1|x) + P(Y = 2|x)}$.

$q(x) = f_0^*(x)$ when $c_{12} = c_{23} = c_{13} = 1$.

When $\exp(\theta_2 - \theta_1) > 1$, $\exp f^*(x)$ is increasing in $q(x)$. Hence, $f^*(x)$ preserves the ordering of $q(x)$. 
Concluding Remarks

- Provide a statistical view of ranking by identifying the optimal ranking function given loss criteria.

- Illustrate the connection between ranking and classification in the framework of convex risk minimization.

  - In bipartite ranking, likelihood ratio of two categories provides the optimal ranking.

  - There is great similarity between classification and ranking when boosting and logistic regression are used.

  - However, the optimal ranking may not be unique under the hinge loss, and there could be potential ties in ranking.
Our study offers theoretical understanding of popular ranking methods for both bipartite and multipartite ranking.

Bridge traditional methods such as proportional odds model with algorithmic ranking methods in machine learning.
References


