A Statistical View of Ranking: Midway between Classification and Regression

Yoonkyung Lee*¹
Department of Statistics
The Ohio State University
*joint work with Kazuki Uematsu

June 4-6, 2014
Conference on Nonparametric Statistics for Big Data
Madison, WI

¹Research supported in part by NSF
Ranking

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

This item: Spline Models for Observational Data by Grace Wahba
Googling Wahba without Grace

Grace Wahba Home Page - Department of Statistics
www.stat.wisc.edu/~wahba/wahba.html

Youssef Wahba - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Youssef_Wahba
Youssef Wahba Pasha (1852-1934) (يونس فـاضل西瓜) (يوسف فاضل) Egyptian Prime Minister and jurist. Youssef Wahba was born in Cairo, Egypt in 1852 of a prominent Coptic ...

Grace Wahba - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Grace_Wahba
Grace Wahba (born August 3, 1934) is the I. J. Schoenberg Professor of Statistics at the University of Wisconsin–Madison. She is a pioneer in methods for smoothing noisy data.

Wahba's problem - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Wahba%27s_problem
In applied mathematics, Wahba's problem, first posed by Grace Wahba in 1965, seeks to find a rotation matrix (special orthogonal matrix) between two ...

Grace Wahba - Google Scholar Citations
scholar.google.com/citations?user=2PABVnoAAAJ... Google Scholar
Professor of Statistics, University of Wisconsin-Madison - Verified email at stat.wisc.edu

Wahba | Facebook
https://www.facebook.com/WahbaMusic
Wahba. 141 likes. If the Hulk had a favorite worship leader, his name would be Wahba.

Grace Wahba
Grace Wahba is the I. J. Schoenberg Professor of Statistics at the University of Wisconsin–Madison. She is a pioneer in methods for smoothing noisy data.

Born: August 3, 1934 (age 79)
Education: Cornell University, Stanford University

People also search for
Emanuel Parzen
Vladimir Vapnik
Robert Tibshirani
Bradley Efron
Jianqing Fan

See results about
Youssef Wahba (Former Prime Minister of Egypt)
Born: 1852, Cairo, Egypt
Died: 1934
Data for Ranking

\[
\begin{align*}
\text{object}_1 & \quad \text{positive} \\
\text{object}_2 & \quad \text{negative} \\
\vdots & \quad \vdots \\
\text{object}_{n-1} & \quad \text{positive} \\
\text{object}_n & \quad \text{negative}
\end{align*}
\]

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

- How to rank?
- What evaluation (or loss) criteria to use for ranking?
- What is the best ranking function given a criterion?
- How is it related to the underlying probability distribution for data?
- How to learn a ranking function from data?
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 $k = 2$)
- $f: \mathcal{X} \to \mathbb{R}$: a real-valued ranking function whose scores induce ordering over the input space
- Training data: $n$ pairs of $(X, Y)$ from $\mathcal{X} \times \mathcal{Y}$
Pairwise Ranking Loss

For a pair of “positive” $x$ and “negative” $x'$, define a loss of ranking function $f$ as

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

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

- Find $f$ minimizing the empirical ranking risk

$$R_{n_+,n_-}(f) = \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} \ell_0(f; x_i, x_j')$$

- Minimizing ranking error is equivalent to maximizing AUC (area under ROC curve).
Likelihood Ratio Minimizes Ranking Risk

Clémençon et al. (2008), Uematsu and Lee (2011), and Gao and Zhou (2012)

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:

\[
P(Y = 1 | X = x) = \frac{\pi_g(x)}{\pi_g(x) + \pi_g(x)} = \frac{f_0^*(x)}{f_0^*(x) + (\pi_1/\pi_0)}
\]
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):

  \[ \ell(f; x, x') = \log(1 + \exp(-(f(x) - f(x')))) \]
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 $\ell'(-s)/\ell'(s) = e^{s}$. $f^*(x) = \log(g_+(x)/g_-(x))$. 
Theorem

Suppose that \( \ell \) is convex, non-increasing, differentiable and \( \ell'(0) < 0 \). Then for almost every \((x, z)\),
\[
\frac{g_+(x)}{g_-(x)} > \frac{g_+(z)}{g_-(z)} \implies f^*(x) > f^*(z).
\]

Remark

For RankSVM, \( \ell(s) = (1 - s)_+ \) with singularity at \( s = 1 \) could yield ties in ranking (leading to inconsistency) while
\( \ell(s) = (1 - s)^2_+ \) is ranking-calibrated.
RankSVM Can Produce Ties

Theorem
Let $f^* = \arg \min_f E(1 - (f(X) - f(X')))$. Suppose that $f^*$ is unique up to an additive constant.

(i) For discrete $X$, a version of $f^*$ is integer-valued.
(ii) For continuous $X$, there exists an integer-valued function whose risk is arbitrarily close to the minimum risk.

Remark

- Scores from RankSVM exhibit granularity.
- Ranking with the hinge loss is not consistent!
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.
Optimal Ranking Function for Multipartite Ranking

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 $f$,

$$R_0(f^*_0; \mathbf{c}) \leq R_0(f; \mathbf{c}).$$

(ii) When $k > 3$ and let $f^*_0(x) = \frac{\sum_{i=2}^{k} c_{1i} P(Y = i|x) \sum_{j=1}^{k-1} c_{jk} P(Y = j|x)}{\sum_{j=1}^{k-1} c_{jk} P(Y = j|x)}$.

If $c_{1k}c_{ji} = c_{1i}c_{jk} - c_{1j}c_{ik}$ for all $1 < j < i < k$, then for any ranking function $f$,

$$R_0(f^*_0; \mathbf{c}) \leq R_0(f; \mathbf{c}).$$

Remark
Let $c_{ji} = (s_i - s_j)w_i w_j l(i > j)$ for some increasing scale $\{s_j\}_{j=1}^k$ and non-negative weight $\{w_j\}_{j=1}^k$. e.g. $c_{ji} = (i - j) l(i > j)$.
Ordinal Regression

- Ordinal regression is commonly used to analyze data with ordinal response in practice.

\[
\begin{align*}
&y = 1 \quad 2 \quad \cdots \quad k - 1 \quad k \\
&-\infty = \theta_0 \quad \theta_1 \quad \theta_2 \quad \cdots \quad \theta_{k-2} \quad \theta_{k-1} \quad \theta_k = \infty
\end{align*}
\]

- 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 \( \theta_k = \infty \).
Convex Loss in Ordinal Regression

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

  \[ \ell(s) = \log(1 + \exp(-s)) \]

- **Support Vector Ordinal Regression** (*Herbrich et al.* 2000):
  \[ \ell(s) = (1 - s)_+ \]
Ordinal Regression Boosting (ORBoost)

- The optimal ranking function $f^*$ under $\ell(s) = \exp(-s)$ is

$$f^*(x) = \frac{1}{2} \log \frac{\sum_{i=2}^{k} P(Y = i|x) \exp(\theta^*_{i-1})}{\sum_{j=1}^{k-1} P(Y = j|x) \exp(-\theta^*_j)}$$

where $\theta^*_j$ are constants depending only on $P_{X,Y}$.

- When $k = 3$,

$$f^*(x) = \frac{1}{2} \log \frac{P(Y = 2|x) + \exp(\theta^*_2 - \theta^*_1)P(Y = 3|x)}{\exp(\theta^*_2 - \theta^*_1)P(Y = 1|x) + P(Y = 2|x)}$$

up to a constant. 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 \(-\infty = \theta_0 < \theta_1 < \ldots < \theta_{k-1} < \theta_k = \infty\).

Given \(\{\theta_j\}_{j=1}^{k-1}\), maximizing the log likelihood amounts to ordinal regression with \(\ell(s) = \log(1 + \exp(-s))\).

When \(k = 3\), given \(\theta_1\) and \(\theta_2\), the minimizer of the deviance risk \(f^*\) satisfies

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

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

When \(\theta_2 > \theta_1\), \(f^*(x)\) preserves the ordering of \(r(x)\).
Support Vector Ordinal Regression

- SVOR with Implicit constraints in *Chu and Keerthi* (2007)

\[
\ell(f, \{\theta_j\}_{j=1}^{k-1} ; x, y) = \sum_{j=1}^{y-1} (1-(f(x)-\theta_j))_+ + \sum_{j=y}^{k-1} (1-(\theta_j-f(x)))_+.
\]

- When \( k = 3 \), \( f^*(x) \) is a step function of

\[
r(x) = \frac{p_2(x) + p_3(x)}{p_1(x) + p_2(x)} \text{ (i.e. } f_0^* \text{ with } c_{12} = c_{13} = c_{23}).
\]

<table>
<thead>
<tr>
<th>( r(x) )</th>
<th>((0, \frac{1}{2}))</th>
<th>((\frac{1}{2}, 1))</th>
<th>((1, 2))</th>
<th>((2, \infty))</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f^*(x) )</td>
<td>( \theta_1 - 1 )</td>
<td>( \min(\theta_1 + 1, \theta_2 - 1) )</td>
<td>( \max(\theta_1 + 1, \theta_2 - 1) )</td>
<td>( \theta_2 + 1 )</td>
</tr>
</tbody>
</table>
Numerical Illustration

- Simulation setting:
  \[ X \mid Y = 1 \sim N(-2, 1), \ X \mid Y = 2 \sim N(0, 1) \text{ and} \]
  \[ X \mid Y = 3 \sim N(2, 1) \]

- When \( c_{12} = c_{23} = c_{13} = 1 \),
  \[
  f_0^*(x) = \frac{P(Y = 2 \mid X = x) + P(Y = 3 \mid X = x)}{P(Y = 1 \mid X = x) + P(Y = 2 \mid X = x)} = \frac{e^{2x} + e^2}{e^{-2x} + e^2}.
  \]

- Generate 500 observations in each category.
- Apply pairwise ranking risk minimization with exponential loss, proportional odds model, ORBoost and SVOR.
Figure: Theoretical ranking function (dotted line) and estimated ranking function (solid line) for pairwise ranking risk minimization with exponential loss, ORBoost, proportional odds model and SVOR with implicit constraints.
The data set consists of 100,000 ratings (on a scale of 1 to 5) for 1,682 movies by 943 users (GroupLens-Research).

Contains content information about the movies (release date and genres) and demographic information about the users (age, gender and occupation).

Transform five categories into three categories: “Low” (1-3), “Middle” (4) and “High” (5) and check the analytical results in $k = 3$. 
Figure: Scatter plots of ranking scores from ORBoost, regression, proportional odds model, and SVOR against pairwise ranking scores with matching cost $c_{13}$ for MovieLens data with three categories. The solid lines indicate theoretical relation between ranking scores.
Concluding Remarks

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

- For pairwise multipartite ranking, the optimal ranking depends on the ratio of conditional probability weighted by misranking costs.

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

- Our study bridges traditional methods such as proportional odds model in statistics with ranking algorithms in machine learning.
Behold the sower in the field,
With her arm she scatters the seeds.
Some seeds are trodden in the pathway;
Some seeds fall on stony ground.

But some seeds fall on fallow ground
They grow and multiply a thousand fold.

– From Pete Seeger’s “Sower Of Seeds”
References


