Statistical Consistency of Multipartite Ranking

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

August 3-8, 2013
Joint Statistical Meetings, Montreal, Canada

[1] Research supported in part by NSF

Ranking

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

(Source: Google images of “ranking”)

Google JSM

Web Images Maps Shopping More Search tools

About 9,950,000 results (0.18 seconds)

Jimmy Swaggart Ministries | Family Worship Center | SonLife ...
www.**jsm**.org/
Jimmy Swaggart Ministries website for Evangelist Jimmy Swaggart, Family Worship Center and the SonLife Broadcasting Network.
SBN - Watch SBN - Contact - Family Worship Center

2013 **Joint Statistical Meetings** - American Statistical Association
www.amstat.org/meetings/**jsm**/2013/
JSM (the **Joint Statistical Meetings**) is the largest gathering of statisticians held in North America. It is held jointly with the American Statistical Association, the ...
Housing Information - Registration - Program - JSM 2013 Online Program

Joint Statistical Meetings - American Statistical Association
www.amstat.org/meetings/**jsm**.cfm
Meetings & Events > **Joint Statistical Meetings**. **Joint Statistical Meetings** Conference on Statistical Practice ASA Biopharmaceutical Section FDA-Industry ...

JSM 2013 Online Program - American Statistical Association
www.amstat.org/meetings/**jsm**/2013/onlineprogram/
JSM (the Joint Statistical Meetings) is the largest gathering of statisticians held in North America. It is held jointly with the American Statistical Association, the ...

JSM Apartments
www.**jsm**apts.com/
Leases apartments on and near the University of Illinois campus.

Data for Ranking

object_1	positive
object_2	negative
$\vdots$	$\vdots$
object_{n-1}	positive
object_n	negative

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

Main Questions

- How to rank?
- What 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?

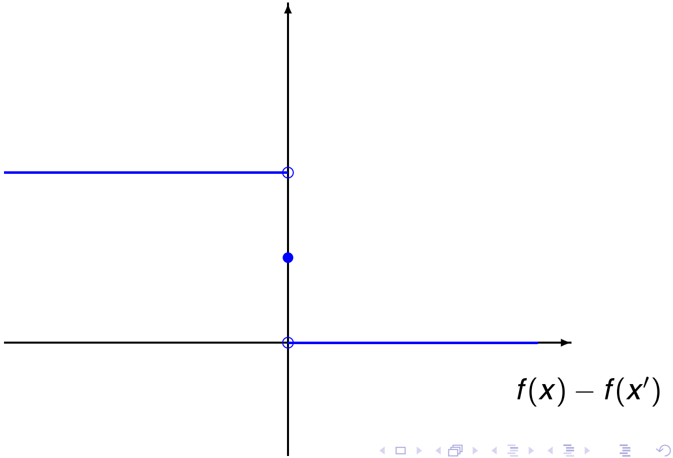
Notation

- $X \in \mathcal{X}$: 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).
- Use convex surrogate loss for risk minimization. (e.g. RankBoost, RankSVM, and RankNet)

Likelihood Ratio Minimizes Ranking Risk

Clémençon et al. (2008) and Uematsu and Lee (2011)

Theorem

Define $f_0^*(x) = g_+(x)/g_-(x)$*, 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*,*

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

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

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

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

- Again, ℓ_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* $f_0^*(x) = \frac{\sum_{i=2}^{k} c_{1i}P(Y=i|x)}{\sum_{j=1}^{k-1} c_{jK}P(Y=j|x)}$.
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*,*

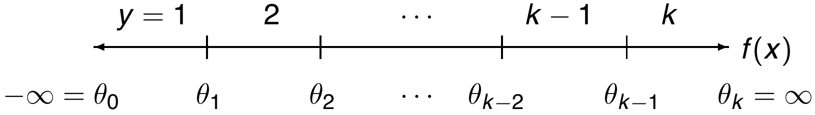
$$R_0(f_0^*; \mathbf{c}) \leq R_0(f; \mathbf{c}).$$

Remark

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

Ordinal Regression

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



- 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):

$$\ell(s) = \exp(-s)$$

- Proportional Odds model (*McCullagh* 1980, *Rennie* 2006):

$$\ell(s) = \log(1 + \exp(-s))$$

- 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 θ^*_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 θ_1 and θ_2, the minimizer of the deviance risk f^* satisfies

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

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

- When $\theta_2 > \theta_1$, $f^*(x)$ preserves the ordering of $q(x)$.

Support Vector Ordinal Regression

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

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

- When $k = 3$, $f^*(x)$ is a step function of $r(x) = \dfrac{p_2(x) + p_3(x)}{p_1(x) + p_2(x)}$ (i.e. f_0^* with $c_{12} = c_{13} = c_{23}$).

$r(x)$	$(0, \frac{1}{2})$	$(\frac{1}{2}, 1)$	$(1, 2)$	$(2, \infty)$
$f^*(x)$	$\theta_1 - 1$	$\min(\theta_1 + 1, \theta_2 - 1)$	$\max(\theta_1 + 1, \theta_2 - 1)$	$\theta_2 + 1$

Numerical Illustration

- Simulation setting:
  $X|Y=1 \sim N(-2,1)$, $X|Y=2 \sim N(0,1)$ and
  $X|Y=3 \sim N(2,1)$
- When $c_{12} = c_{23} = c_{13} = 1$,

$$f_0^*(x) = \frac{P(Y=2|X=x) + P(Y=3|X=x)}{P(Y=1|X=x) + P(Y=2|X=x)} = \frac{e^{2x} + e^2}{e^{-2x} + e^2}.$$

- Generate 500 observations in each category.
- Apply pairwise ranking risk minimization with exponential loss, proportional odds model, ORBoost and SVOR.

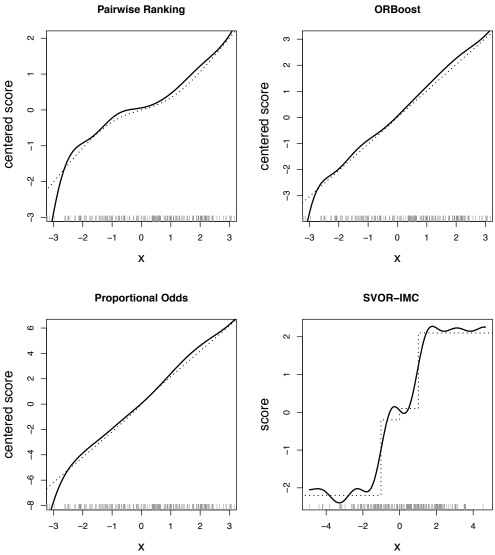


Figure: Theoretical ranking function (dotted line) and estimated ranking function (solid line) for pairwise ranking risk minimization with exponential loss, ORBoost, proportional odds model and SVOR with implicit constraints.

Application to Movie-Lens Data

- 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 $k = 3$.

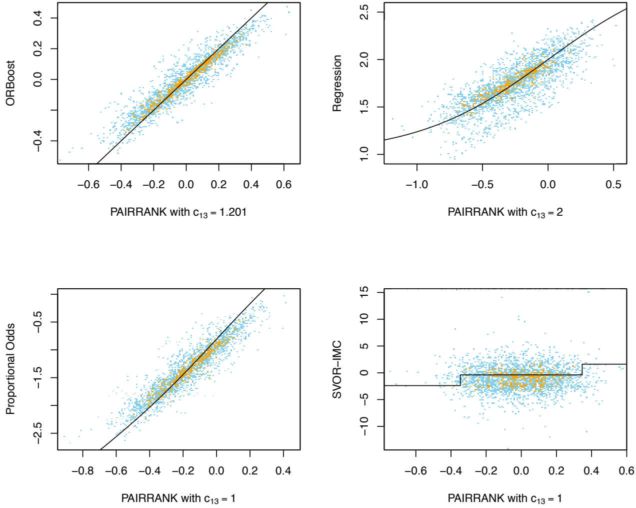


Figure: Scatter plots of ranking scores from ORBoost, regression, proportional odds model, and SVOR against pairwise ranking scores with matching cost c_{13} for MovieLens data with three categories. The solid lines indicate theoretical relation between ranking scores.

Effect of Differential Ranking Cost

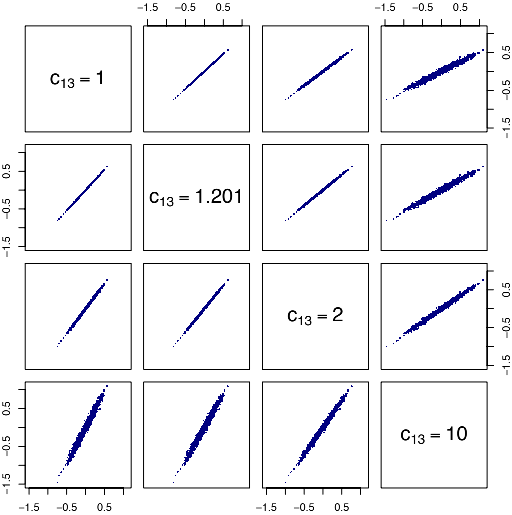


Figure: Scatter plots of pairwise ranking scores (centered to zero) with different ranking cost c_{13} for MovieLens data when $c_{12} = c_{23} = 1$.

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.
- The solution to some ordinal regression methods can be viewed as a special case of the optimal function in multipartite ranking.
- Our study bridges traditional methods such as proportional odds model in statistics with ranking algorithms in machine learning.

References

- S. Agarwal, and P. Niyogi. (2009) "Generalization Bounds for Ranking Algorithms via Algorithmic Stability." *Journal of Machine Learning Research,* 10, 441- 474.
- W. Chu, and S.S. Keerthi. (2007) "Support vector ordinal regression." *Neural computation,* 19(3), 792- 815.
- S. Clémençon, G. Lugosi and N. Vayatis. (2008) "Ranking and empirical minimization of U-statistics." *Annals of Statistics,* 36, 844- 874.
- R. Herbrich, T. Graepel, and K. Obermayer. (2000) "Large margin rank boundaries for ordinal regression." *Advances in Large Margin Classifiers,* 115- 132.
- P. Li, C. Burges, and Q. Wu. (2007) "McRank: Learning to Rank Using Multiple Classification and Gradient Boosting" *Advances in Neural Information Processing Systems,* 20, 897- 904.
- H. Lin, and L. Li. (2006) "Large-Margin Thresholded Ensembles for Ordinal Regression: Theory and Practice." *Lecture Notes in Computer Science,* 4264, 319- 333.

- P. McCullagh. (1980) “Regression Models for Ordinal Data.” *Journal of Royal Statistical Society (B),* 42(2), 109- 142.
- J.D.M Rennie. (2006) “A comparison of McCullagh’s proportional odds model to modern ordinal regression algorithms.” http://people.csail.mit.edu/jrennie/writing/proportionalOdds.pdf
- A. Shashua, and A. Levin. (2003) “Ranking with large margin principle: Two approaches.” *Advances in Neural Information Processing Systems,* 15, 937- 944.
- K. Uematsu, and Y. Lee. (2011) “On Theoretically Optimal Ranking Functions in Bipartite Ranking.” *Technical Report No. 863, Department of Statistics, The Ohio State University.*
- K. Uematsu, and Y. Lee. (2013) “Statistical Optimality in Multipartite Ranking and Ordinal Regression” *Technical Report No. 873, Department of Statistics, The Ohio State University.*
- W. Waegeman, B. D. Baets and L.Boullart. (2008) “ROC analysis in ordinal regression learning.” *Pattern Recognition Letters,* 29(1), 1- 9.