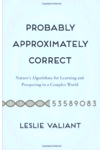
A Statistical View of Ranking: Midway between Classification and Regression

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

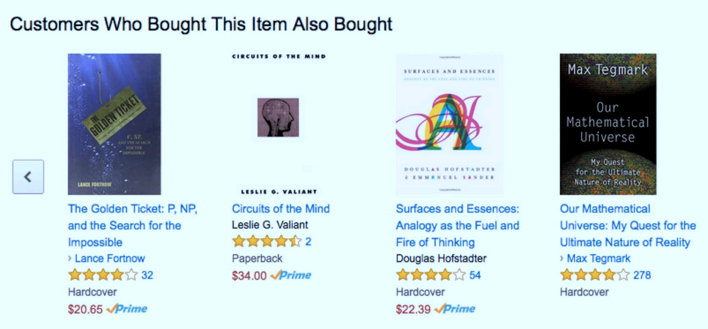
December 3, 2015
Department of Computer Science and Engineering
The Ohio State University

Ranking

- Aims to order a set of objects or instances reflecting their underlying utility, quality or relevance to queries


PROBABLY
APPROXIMATELY
CORRECT
Nature's Algorithms for Learning and Prospering in a Complex World
53589083
LESLIE VALIANT

This item: Probably Approximately Correct by Leslie Valiant



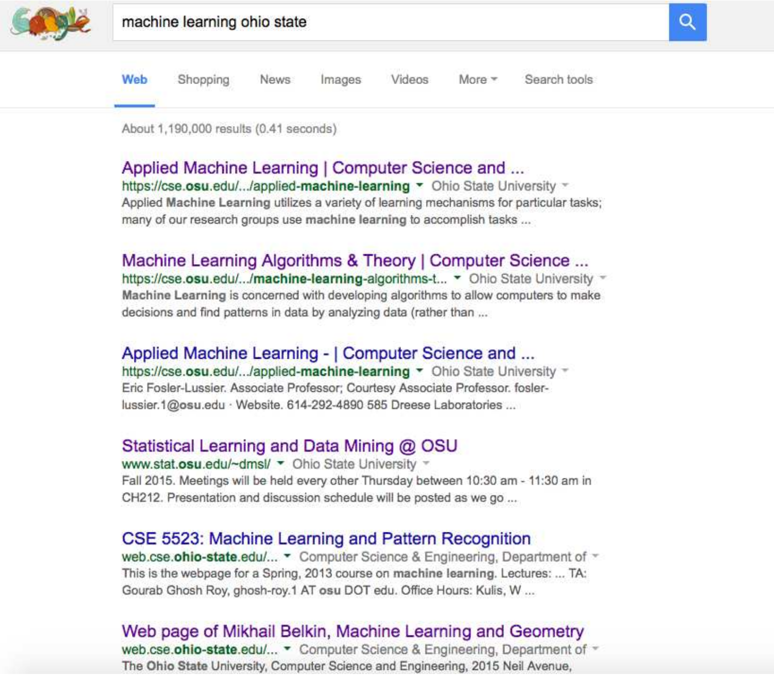
Customers Who Bought This Item Also Bought

The Golden Ticket: P, NP, and the Search for the Impossible
› Lance Fortnow
★★★★☆ 32
Hardcover
$20.65 ✓Prime

Circuits of the Mind
Leslie G. Valiant
★★★★½ 2
Paperback
$34.00 ✓Prime

Surfaces and Essences: Analogy as the Fuel and Fire of Thinking
Douglas Hofstadter
★★★★☆ 54
Hardcover
$22.39 ✓Prime

Our Mathematical Universe: My Quest for the Ultimate Nature of Reality
› Max Tegmark
★★★★☆ 278
Hardcover



machine learning ohio state

About 1,190,000 results (0.41 seconds)

Applied Machine Learning | Computer Science and ...
https://cse.osu.edu/.../applied-machine-learning ▾ Ohio State University ▾
Applied **Machine Learning** utilizes a variety of learning mechanisms for particular tasks; many of our research groups use **machine learning** to accomplish tasks ...

Machine Learning Algorithms & Theory | Computer Science ...
https://cse.osu.edu/.../machine-learning-algorithms-t... ▾ Ohio State University ▾
Machine Learning is concerned with developing algorithms to allow computers to make decisions and find patterns in data by analyzing data (rather than ...

Applied Machine Learning - | Computer Science and ...
https://cse.osu.edu/.../applied-machine-learning ▾ Ohio State University ▾
Eric Fosler-Lussier. Associate Professor; Courtesy Associate Professor. fosler-lussier.1@osu.edu · Website. 614-292-4890 585 Dreese Laboratories ...

Statistical Learning and Data Mining @ OSU
www.stat.osu.edu/~dmsl/ ▾ Ohio State University ▾
Fall 2015. Meetings will be held every other Thursday between 10:30 am - 11:30 am in CH212. Presentation and discussion schedule will be posted as we go ...

CSE 5523: Machine Learning and Pattern Recognition
web.cse.ohio-state.edu/... ▾ Computer Science & Engineering, Department of ▾
This is the webpage for a Spring, 2013 course on **machine learning**. Lectures: ... TA: Gourab Ghosh Roy, ghosh-roy.1 AT **osu** DOT edu. Office Hours: Kulis, W ...

Web page of Mikhail Belkin, Machine Learning and Geometry
web.cse.ohio-state.edu/... ▾ Computer Science & Engineering, Department of ▾
The **Ohio State** University, Computer Science and Engineering, 2015 Neil Avenue,

Data for Ranking

$\boxed{\text{object}_1}$	positive
$\boxed{\text{object}_2}$	negative
$\vdots$	$\vdots$
$\boxed{\text{object}_{n-1}}$	positive
$\boxed{\text{object}_n}$	negative

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

Main Questions

- How to formulate ranking problems?
- 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 (or scoring) function from data?

Notation

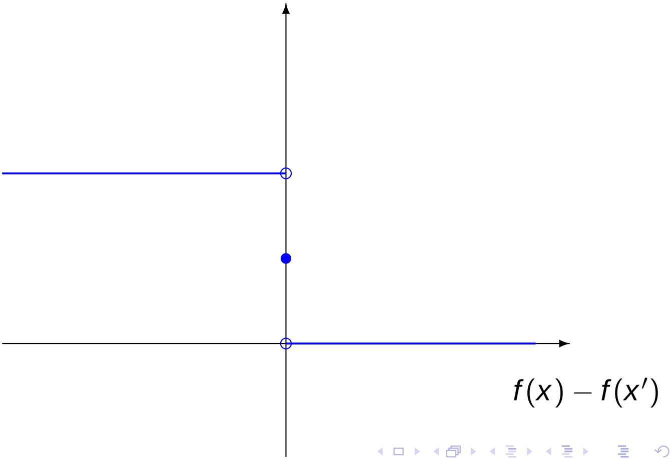
- $X \in \mathcal{X}$: an instance to rank
- $Y \in \mathcal{Y} = \{1, \cdots, k\}$: an ordinal response in multipartite ranking (bipartite ranking if $k = 2$, often with $\mathcal{Y} = \{\pm 1\}$)
- $g_{\pm}(x)$: pdfs of X given $Y = \pm 1$
- Training data: n pairs of (X, Y) from $\mathcal{X} \times \mathcal{Y}$
  e.g. $\{(x_i, +1)\}_{i=1}^{n_+} \cup \{(x_j', -1)\}_{j=1}^{n_-}$ for bipartite ranking
- $f\colon \mathcal{X} \to \mathbb{R}$: a real-valued ranking function whose scores induce ordering over the input space

$$f(x) > f(x') \Longleftrightarrow x \succ x'$$

Pairwise Ranking Loss

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

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



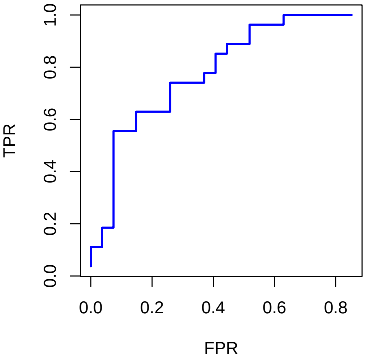
Bipartite Ranking

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

- Find f minimizing the empirical ranking error

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

- Minimizing ranking error is equivalent to maximizing AUC (area under ROC curve) of f.



Likelihood Ratio Minimizes Ranking Risk

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

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*,*

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

Remark

Connection to posterior probability in classification:

$$P(Y = 1|X = x) = \frac{\pi_+ g_+(x)}{\pi_+ g_+(x) + \pi_- g_-(x)} = \frac{f_0^*(x)}{f_0^*(x) + (\pi_-/\pi_+)}$$

Classification, Ranking and Regression

Regression: $p_+(x) = P(Y = 1|X = x)$ or $\log \frac{p_+(x)}{1-p_+(x)}$

Ranking: order-preserving transformation of $p_+(x)$ or
likelihood ratio $g_+(x)/g_-(x)$

Classification: $\operatorname{sgn}(p_+(x) - \frac{1}{2})$

Convex Surrogate Loss for Bipartite Ranking

- Exponential loss
  in RankBoost (Freund et al. 2003):

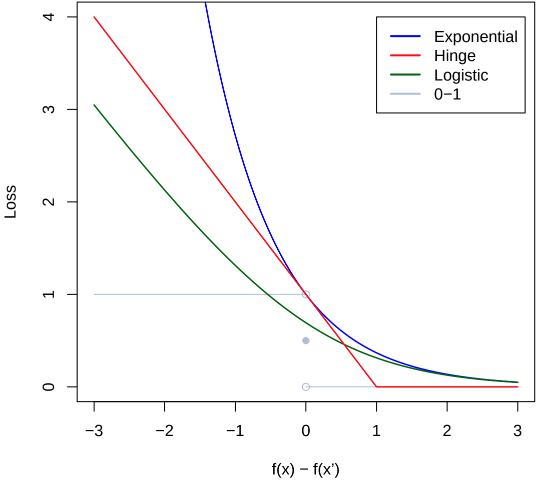
$$\ell(f;x,x') = \exp(-(f(x) - f(x')))$$

- 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):

$$\ell(f;x,x') = \log(1 + \exp(-(f(x) - f(x')))$$



Bartlett et al. (2006), *Convexity, Classification, and Risk Bounds*

Is classification-calibration sufficient for ranking consistency?

Optimal Ranking Function Under Convex Loss

Theorem

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

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

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* $\ell'(-s)/\ell'(s) = e^s$. $f^*(x) = \log(g_+(x)/g_-(x))$.

Ranking-Calibrated Loss

Theorem

Suppose that ℓ 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.

Toy Example: RankSVM

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

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

  For f^*, the optimal increments s_1^* and s_2^* are:

  (i) if $\Delta_{12} > 0$ and $\Delta_{23} > 0$, $(s_1^*, s_2^*) = (1, 1)$
  (ii) if $\Delta_{23} < 0$ and $g_+(x_2) > g_-(x_2)$, $(s_1^*, s_2^*) = (1, 0)$
  (iii) if $\Delta_{12} < 0$ and $g_+(x_2) < g_-(x_2)$, $(s_1^*, s_2^*) = (0, 1)$

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 $\mathcal{X}$, a version of f^* is integer-valued.*

(ii) *For continuous $\mathcal{X}$, 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!*

Numerical Illustration

- Simulation setting:
  $X \sim N(1,1)$ and $X' \sim N(-1,1)$
  $\log(g_+(x)/g_-(x)) = 2x$ with 'Bayes' ranking error of
  $P(X < X') = \Phi(-\sqrt{2}) \approx 0.07865$

- Generate $\{(x_i,+1)\}_{i=1}^n \cup \{(x'_j,-1)\}_{j=1}^n$
  where n: sample size for each category

- Apply AUC maximizing SVM (Brefeld and Scheffer 2005) with a Gaussian kernel K

$$\min_{f \in \mathcal{H}_K} \quad C \sum_{i,j} \Big(1 - (f(x_i) - f(x'_j))\Big)_+ + \|f\|^2$$

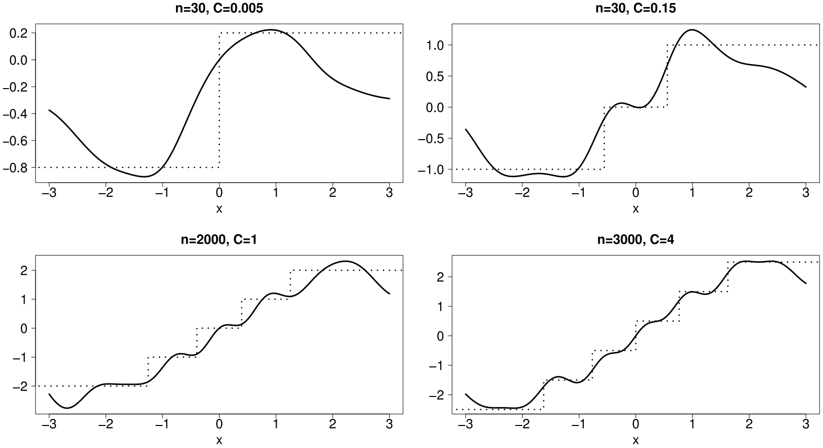


Figure: The solid lines are the estimated ranking functions, and the dotted lines are step functions with minimal risk.

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

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

  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* $f_0^*(x) = \dfrac{c_{12}P(Y = 2|x) + c_{13}P(Y = 3|x)}{c_{13}P(Y = 1|x) + c_{23}P(Y = 2|x)}$.
Then for any ranking function f,

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

(ii) When $k > 3$ *and let* $f_0^*(x) = \dfrac{\sum_{i=2}^{k} c_{1i}P(Y = i|x)}{\sum_{j=1}^{k-1} c_{jK}P(Y = j|x)}$.
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,

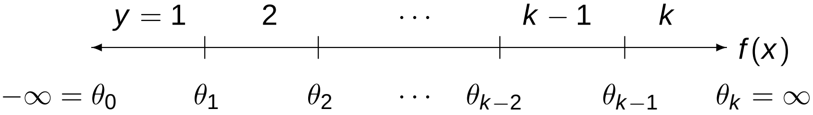
$$R_0(f_0^*; \mathbf{c}) \leq R_0(f; \mathbf{c}).$$

Remark

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

Ordinal Regression

- Ordinal regression is commonly used to analyze data with ordinal responses 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):

$$\ell(s) = (1 - s)_+$$

Optimal Ranking Function with Ordinal Regression

Letting $p_j(x) = P(Y = j|X = x)$, when $k = 3$,

$$f_0^*(x) = \frac{c_{12}p_2(x) + c_{13}p_3(x)}{c_{13}p_1(x) + c_{23}p_2(x)}$$

- Ordinal Regression Boosting (ORBoost):

$$f^*(x) = \frac{1}{2}\log\frac{p_2(x) + \exp(\theta_2^* - \theta_1^*)p_3(x)}{\exp(\theta_2^* - \theta_1^*)p_1(x) + p_2(x)} = \frac{1}{2}\log f_0^*(x)$$

  with $c_{12} = c_{23} = 1$ and $c_{13} = e^{\theta_2^* - \theta_1^*}$.

- Proportional Odds Model:
  $f^*(x)$ preserves the ordering of
  $r(x) = \frac{p_2(x) + p_3(x)}{p_1(x) + p_2(x)} = f_0^*(x)$ with $c_{12} = c_{23} = c_{13} = 1$.

- Support Vector Ordinal Regression (SVOR):
  $f^*(x)$ is a non-decreasing step function of $r(x)$.

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.

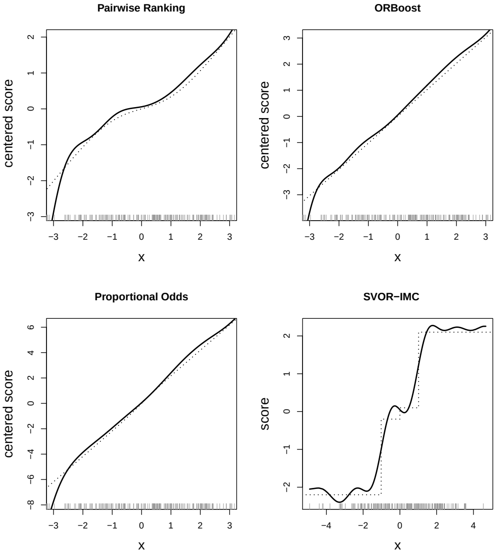


Figure: Theoretical ranking function (dotted line) and estimated ranking function (solid line) for pairwise ranking risk minimization with exponential loss, ORBoost, proportional odds model and SVOR with implicit constraints.

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 $k = 3$.

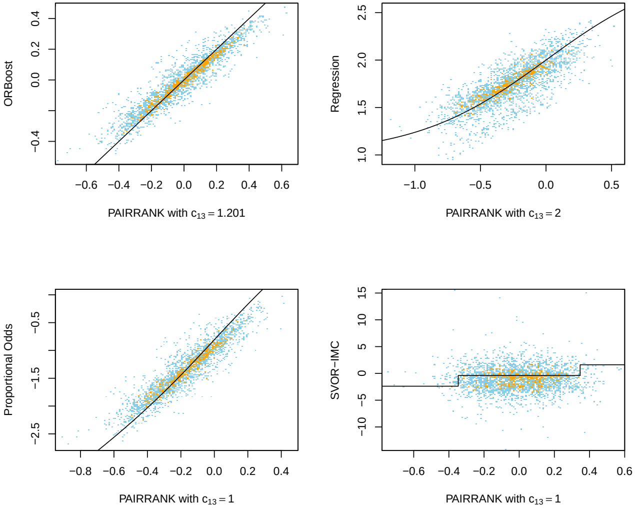


Figure: Scatter plots of ranking scores from ORBoost, regression, proportional odds model, and SVOR against pairwise ranking scores with matching cost c_{13} for MovieLens data with three categories. The solid lines indicate theoretical relation between ranking scores.

Concluding Remarks

- Provide a statistical view of ranking by identifying the optimal ranking function given loss criteria.
- Illustrate the connection between ranking and classification/ordinal regression in the framework of convex risk minimization.
- Ranking requires more information than classification.

References

Kazuki Uematsu and Yoonkyung Lee.
On theoretically optimal ranking functions in bipartite ranking.
Technical Report 863, Department of Statistics, The Ohio State University, 2011.
Revision submitted.

Kazuki Uematsu and Yoonkyung Lee.
Statistical optimality in multipartite ranking and ordinal regression.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(5):1080–1094, May 2015.

Acknowledgments

Kazuki Uematsu

DMS-12-09194
DMS-15-13566