Statistical Analysis of Bipartite and Multipartite Ranking by Convex Risk Minimization

Kazuki Uematsu

The Ohio State University

2/25/2010

Introduction

- The goal of ranking: Find a way to order a set of given objects or instances reflecting their underlying utility, relevance or quality.
- Ranking has gained increasing attention in the field of machine learning and information retrieval such as document retrieval and website search.

Example: OHSUMED data

A clinically-oriented MEDLINE (a bibliographic database of life sciences and biomedical information) subset stored in Oregon Health and Science University, covering all references from 270 medical journals between 1987-1991 .

- 348566 documents and 106 queries
- 16140 query-document pairs
- The relevance judgement:
  'd' (definitely relevant), 'p' (possibly relevant) and 'n' (not relevant)

Each query includes patient information and information request and is sorted by number.

- Query1
  ‘60 year old menopausal woman without hormone replacement therapy’
  ‘Are there adverse effects on lipids when progesterone is given with estrogen replacement therapy? ’
- Query 2
  ‘60 year old male with disseminated intravascular coagulation’
  ‘pathophysiology and treatment of disseminated intravascular coagulation’

• • •

query	documentID	relevance	1	2	3	4	5	6	7	8	9	10	…	40	41	42	43	44	45
1	244338	n	5	3.47	0.50	0.48	37.33	11.43	37.31	2.01	25.06	14.21		4.84	31.19	3.44	-36.63	-36.95	-35.83
1	143821	d	3	2.08	0.60	0.55	37.33	11.43	37.31	2.13	15.66	10.91		4.73	28.97	3.37	-36.18	-36.30	-35.20
1	285257	n	2	1.39	0.33	0.31	37.33	11.43	37.31	1.16	9.54	6.07		3.94	27.08	3.30	-37.35	-37.30	-36.00
1	201684	n	1	0.69	0.17	0.15	37.33	11.43	37.31	0.70	6.12	4.31		4.22	27.01	3.30	-37.54	-37.77	-36.63
1	48192	p	0	0.00	0.00	0.00	37.33	11.43	37.31	0.00	0.00	0.00		2.97	26.83	3.29	-36.32	-37.38	-36.06
1	111457	d	3	2.08	0.38	0.35	37.33	11.43	37.31	1.47	15.22	9.15		3.37	26.34	3.27	-38.49	-38.77	-37.41
1	248063	n	0	0.00	0.00	0.00	37.33	11.43	37.31	0.00	0.00	0.00		3.31	25.49	3.24	-38.76	-39.16	-37.90
1	316117	p	2	1.39	0.33	0.31	37.33	11.43	37.31	1.16	9.54	6.07		3.03	25.38	3.23	-39.04	-39.52	-37.58
1	256570	d	0	0.00	0.00	0.00	37.33	11.43	37.31	0.00	0.00	0.00	…	2.53	25.33	3.23	-35.73	-37.15	-35.66
…																			
106	337888	d	0	0.00	0.00	0.00	30.13	8.81	30.07	0.00	0.00	0.00	…	0.00	NULL	NULL	NULL	NULL	NULL
106	36526	n	0	0.00	0.00	0.00	30.13	8.81	30.07	0.00	0.00	0.00		0.00	NULL	NULL	NULL	NULL	NULL
106	63379	n	0	0.00	0.00	0.00	30.13	8.81	30.07	0.00	0.00	0.00		0.00	NULL	NULL	NULL	NULL	NULL
106	298134	p	0	0.00	0.00	0.00	30.13	8.81	30.07	0.00	0.00	0.00		0.00	NULL	NULL	NULL	NULL	NULL
106	299223	n	0	0.00	0.00	0.00	30.13	8.81	30.07	0.00	0.00	0.00	…	0.00	NULL	NULL	NULL	NULL	NULL

ID	Feature Description	Category
1	$\sum_{q_i \in q \cap d} c(q_i, d)$ in title	Q-D
2	$\sum_{q_i \in q \cap d} \log\left(c(q_i, d) + 1\right)$ in title	Q-D
3	$\sum_{q_i \in q \cap d} \frac{c(q_i,d)}{\lvert d \rvert}$ in title	Q-D
4	$\sum_{q_i \in q \cap d} \log\left(\frac{c(q_i,d)}{\lvert d \rvert} + 1\right)$ in title	Q-D
5	$\sum_{q_i \in q} \log\left(\frac{\lvert C \rvert}{df(q_i)}\right)$ in title	Q
6	$\sum_{q_i \in q} \log\left(\log\left(\frac{\lvert C \rvert}{df(q_i)}\right)\right)$ in title	Q
7	$\sum_{q_i \in q} \log\left(\frac{\lvert C \rvert}{c(q_i,C)} + 1\right)$ in title	Q
8	$\sum_{q_i \in q \cap d} \log\left(\frac{c(q_i,d)}{\lvert d \rvert} \cdot \log\left(\frac{\lvert C \rvert}{df(q_i)}\right) + 1\right)$ in title	Q-D
9	$\sum_{q_i \in q \cap d} c(q_i, d) \cdot \log\left(\frac{\lvert C \rvert}{df(q_i)}\right)$ in title	Q-D
10	$\sum_{q_i \in q \cap d} \log\left(\frac{c(q_i,d)}{\lvert d \rvert} \cdot \frac{\lvert C \rvert}{c(q_i,C)} + 1\right)$ in title	Q-D
11	BM25 of title	Q-D
12	log(BM25) of title	Q-D
...		
40	$\sum_{q_i \in q \cap d} \log\left(\frac{c(q_i,d)}{\lvert d \rvert} \cdot \frac{\lvert C \rvert}{c(q_i,C)} + 1\right)$ in 'title + abstract'	Q-D
41	BM25 of 'title + abstract'	Q-D
42	log(BM25) of 'title + abstract'	Q-D
43	LMIR.DIR of 'title + abstract'	Q-D
44	LMIR.JM of in 'title + abstract'	Q-D
45	LMIR.ABS of in 'title + abstract'	Q-D

$c(q_i, d)$ denotes the number of occurrences of query term q_i in document d. Note that while talking about a stream (e.g. title), $c(q_i, d)$ means the number of occurrences of q_i in the specific stream (e.g., title) of document d.
Inverse document frequency (IDF) of query term q_i was computed as follows,

$$idf(q_i) = \log \frac{|C| - df(q_i) + 0.5}{df(q_i) + 0.5}, \tag{1}$$

where document frequency $df(q_i)$ is the number of documents containing q_i in a stream, and $|C|$ is the total number of documents in the document collection. Note that IDF is document-independent, and all the documents associated with the same query has the same IDF value.
$|d|$ denotes the length (i.e., the number of words) of document d. When considering a specific stream, $|d|$ means the length of the stream. For example, $|d|$ of body means the length of the body of document d.
The BM25 score of a document d was computed as follows,

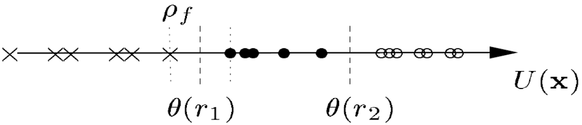
$$BM25(d, q) = \sum_{q_i \in q} \frac{idf(q_i) \cdot c(q_i, d) \cdot (k_1 + 1)}{c(q_i, d) + k_1 \cdot (1 - b + b \cdot \frac{|d|}{avgdl})} \cdot \frac{(k_3 + 1)c(q_i, q)}{k_3 + c(q_i, q)}, \tag{2}$$

where $avgdl$ denotes the average document length in the entire document corpus. k_1, k_3 and b are free parameters.

(Cited from Qin et al)

Ranking methods

- Itemwise ranking (Ordinal regression):
  Support Vector Ordinal Regression (Herbrich et al (1999) etc)
  Proportional Odds Model (McCullagh (1984))
  ORBoost (Lin and Li (2006))



- Pairwise ranking:
  RankBoost (Freund et al (2003))
  AUCSVM (Rakotomamonjy (2004), Brefeld and Scheffer (2005))
  SVrank (Cortes and Mohri (2007))

Notation

- K: The number of categories.
- X: An instance used for ranking
- $Y \in \{1, \cdots, K\}$: A category
- $\mathcal{X}$: The space of instances
- r: $X \to \mathbb{R}$ A ranking function
- c: $X \to \{1, \cdots, K\}$ A classification function

As a special case, the case in which $K = 2$ is called bipartite ranking. In this case, we use the following notations;

- $Y \in \{1, -1\}$
- X: an instance whose category (Y) is 1
- X': an instance whose category (Y) is -1

How is ranking different from classification?

Classification: Classification error is

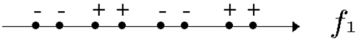
$$l_0(c; X, Y) = I_{(c(X) \neq Y)}$$

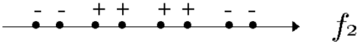
Pairwise ranking: Ranking error is defined for a pair of two instances X and X' whose categories are different (1 and -1 respectively).

$$l_0(r; X, X') = I_{(r(X) < r(X'))} + \frac{1}{2} I_{(r(X) = r(X'))}$$

- Classification error is not always a good measure for ranking problem.

- An example from Agrawal and Niyogi (2005)





Misclassification error is 2/8 in both cases, but ranking error of f_1 is 4/16 while the error of f_2 is 8/16. So f_1 is the better ranking function than f_2.

(For detailed discussion, see Cortes and Mohri (2003))

Empirical ranking risk

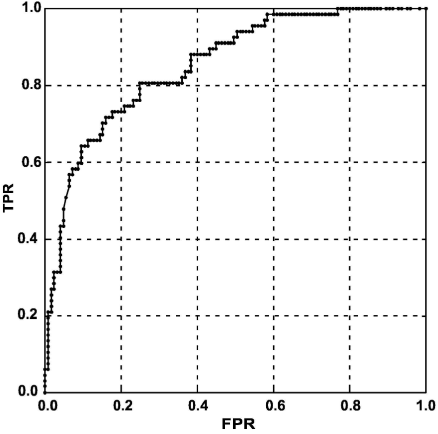
An empirical ranking risk is

$$\frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} l_0(r; x_i, x_j{}')$$

where $\{x_i\}$ is an observation of positive instance, $\left\{x_j{}'\right\}$ is an observation of negative instance and n_+, n_- are the number of observations in positive instance and negative instance respectively.

AUC (Area under the curve)

Empirical ranking risk is equal to one minus the area under Receiver Operating Characteristic (ROC) curve (true positive rate versus false positive rate). (See Hanley and McNeil (1982))



Connection to hypothesis testing

- For fixed threshold value t

$$\text{TPR}(t) = P(r(X) > t|Y = 1)$$

$$\text{FPR}(t) = P(r(X) > t|Y = -1)$$

ROC is (FPR(t), TPR(t)) $-\infty < t < \infty$.

- Consider a hypothesis testing $H_0 : Y = -1$ versus $H_a : Y = 1$ using r as a test statistics, then TPR and FPR are power and size of the test respectively.
- Hence maximizing AUC is equivalent to find MP test, and it is known that the test uses likelihood ratio.

Optimal ranking under l_0

Theorem (1)

Let $r_0^*(x) := g_+(x)/g_-(x)$ *and* $R_0(r) = E(l_0(r; X, X'))$ *denote the expected ranking risk of* r *under the bipartite ranking loss, where* X *and* X' *are, respectively, a positive instance* ($Y = 1$) *and a negative instance* ($Y = -1$) *randomly drawn from the distributions with pdf* g_+ *and* g_-*. Then for any ranking function* r*,*

$$R_0(r_0^*) \leq R_0(r).$$

Ranking with convex loss

- If pdf g_+ and g_- are known we can derive the best ranking function but they are unknown.
- Estimate the ranking function through the minimization of empirical ranking risk, but it is computationally difficult since l_0 is not convex.
- Use a convex surrogate loss as in classification.

Examples for surrogate losses

- RankBoost (Freund et al (2003))

$$l(r; X, X') = \exp(r(X) - r(X'))$$

- AUCSVM (Rakotomamonjy (2004) and Brefeld and Scheffer (2005)) and SVRank (Cortes and Mohri (2007))

$$l(r; X, X') = (1 - (r(X) - r(X')))_+$$

Optimal Ranking under convex loss

Theorem (2)

Suppose that l is differentiable, $l'(s) < 0$ for all $s \in \mathbb{R}$, and $l'(-s)/l'(s) = \exp(s/\alpha)$ for some positive constant α. Then the best ranking function under the loss l is of the form

$$r^*(x) = \alpha \log(g_+(x)/g_-(x)) + \beta,$$

where β is an arbitrary constant.

Application of Theorem 2

- For the RankBoost algorithm, $l(s) = \exp(-s)$, and $l'(-s)/l'(s) = \exp(2s)$. Hence $r^*(x) = \frac{1}{2}\log(g_+(x)/g_-(x))$.
- Similarly, when $l(s) = \log(1 + \exp(-s))$, the negative log likelihood in logistic regression, $l'(-s)/l'(s) = \exp(s)$, and $r^*(x) = \log(g_+(x)/g_-(x))$.
- In these cases, there is no essential difference between ranking and classification. For example RankBoost derives the same result as AdaBoost asymptotically.

More general result

Theorem (3)

Suppose that l is convex, and the subdifferential of l at zero contains only negative values.
Let $r^* := \arg\min_r R_l(r)$.

1. *If* $\frac{g_+(x)}{g_-(x)} > \frac{g_+(x')}{g_-(x')}$*, then* $r^*(x) \geq r^*(x')$ *a.e.*
2. *If* l *is differentiable and* $\frac{g_+(x)}{g_-(x)} > \frac{g_+(x')}{g_-(x')}$*, then* $r^*(x) > r^*(x')$ *a.e.*
3. *If* l *is differentiable and* l' *is one-to-one, then for* x *and* x' *such that* $\frac{g_+(x)}{g_-(x)} = \frac{g_+(x')}{g_-(x')}$, $r^*(x) = r^*(x')$.

Numerical example

- $X \sim N(1, 1)$ and $X' \sim N(-1, 1)$
- $n_+ = n_- = 200$
- $\frac{1}{2}\log(g_+(x)/g_-(x)) = x$

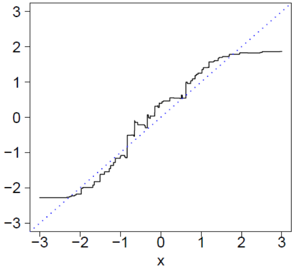


Figure 1: Ranking function given by the RankBoost algorithm. The solid line is the ranking function $\hat{f}$ centered to 0, and the dotted line is the theoretically optimal ranking function $f^*(x) = x$.

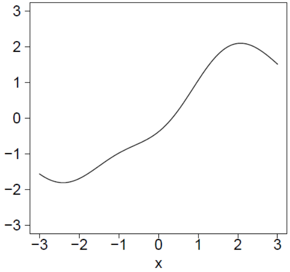


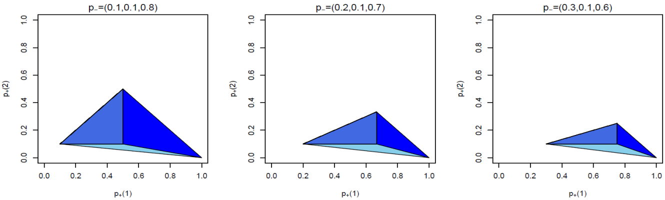
Figure 2: Ranking function given by the AUC maximizing SVM.

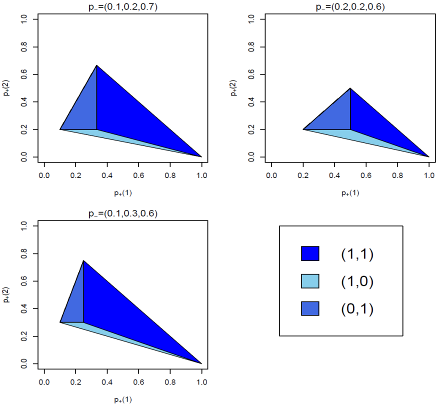
Toy example for possible ties under hinge loss

- Assume that $\mathcal{X} = \{x_1, x_2, x_3\}$.
- x_1, x_2 and x_3 are ordered such that for pmf p_+ and p_-

$$\frac{p_+(x_1)}{p_-(x_1)} > \frac{p_+(x_2)}{p_-(x_2)} > \frac{p_+(x_3)}{p_-(x_3)}$$

  (Note that $\sum_{i=1}^{3} p_+(x_i) = \sum_{i=1}^{3} p_-(x_i) = 1$)
- In this setting, we can derive the optimal ranking r^* function under hinge loss.
  Let $\Delta_{12} = r^*(x_1) - r^*(x_2)$ and
  $\Delta_{23} = r^*(x_2) - r^*(x_3)$





$$c_{12} = \frac{p_+(x_1)}{p_-(x_1)} - \left(\frac{p_+(x_2)}{p_-(x_2)} + \frac{p_+(x_3)}{p_-(x_2)} \right)$$

$$c_{23} = \frac{p_-(x_3)}{p_+(x_3)} - \left(\frac{p_-(x_2)}{p_+(x_2)} + \frac{p_-(x_1)}{p_+(x_2)} \right)$$

$$(1,1) \Leftrightarrow c_{12} > 0, c_{23} > 0$$

$$(1,0) \Leftrightarrow c_{23} < 0, p_-(x_2) > p_+(x_2)$$

$$(0,1) \Leftrightarrow c_{12} < 0, p_-(x_2) < p_+(x_2)$$

Extension to multipartite ranking problem

- If $K \geq 3$ there is difference between Ordinal regression and Pairwise ranking method.
- Examples for an evaluation metric in information retrieval:
  Assume that data is $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ and $r(x_{[n]}) \leq \cdots \leq r(x_{[1]})$.
  - MRR (Mean Reciprocal Rank, $\sum_{i=1}^{n} \frac{g(y_{[i]})}{i}$)
  - NDCG (Normalized discounted cumulative gain, $\sum_{i=1}^{n} \frac{2^{y[i]}-1}{\log_2(i+1)}$)

  where g is an nondecreasing function.

Multipartite version of AUC

Let n_j be the number of observations in category $j \in \{1, \ldots, K\}$.

- Hand and Till (2001)
  $\hat{U}_{\text{ovo}}(r) = \frac{2}{k(k-1)} \sum_{l'<l} \hat{A}_{l'l}$
  where $\hat{A}_{l'l} = \frac{1}{n_{l'} n_l} \sum_{y_i=l'} \sum_{y_j=l} I_{(r(x_i)<r(x_j))}$

- Waegeman et al (2008)
  $\hat{U}_{\text{pairs}}(r) = \frac{1}{\sum_{l'<l} n_{l'} n_l} \sum_{y_i<y_j} I_{(r(x_i)<r(x_j))}$
  $\hat{U}_{\text{cons}}(r) = \frac{1}{k-1} \sum_{l=1}^{k-1} \hat{B}_l$
  where
  $\hat{B}_l = \frac{1}{\sum_{i=1}^{l} n_i \sum_{j=l+1}^{k} n_j} \sum_{y_i \leq l} \sum_{y_j>l} I_{(r(x_i)<r(x_j))}$

A tripartite ranking risk

- Suppose $K = 3$, and let a pairwise ranking loss of (X, Y) and (X', Y') (where $Y > Y'$) using r be

  $$l_0(r; X, X', Y, Y') = c_{Y'Y} I_{(r(X)<r(X'))}$$

  then the expected risk of r is
  $R_0(r; \mathrm{c})$
  $= \sum_{1\le j<i\le 3} E[l_0(r; X, X', i, j)] P(Y = i, Y' = j)$

- If $c_{12} = c_{13} = c_{23} = 1$, the risk above is the same as $E[\hat{U}_{\text{pairs}}(r)]$

Theorem (4)

Let

$$r_0^*(x) := \frac{c_{13}P(Y = 3|x) + c_{12}P(Y = 2|x)}{c_{13}P(Y = 1|x) + c_{23}P(Y = 2|x)}$$

and $R(r; \mathrm{c})$ *denote the pairwise ranking risk of* r *above. Then for any ranking function* r,

$$R_0(r_0^*; \mathrm{c}) \leq R_0(r; \mathrm{c}).$$

Note: Under some conditions, this result can be extended to the case where $K \geq 4$.

Relation to Ordinal regression

- A loss function in ordinal regression is represented as

$$l(r, \{\theta_i\}_{i=0}^{K}; x, y) = l(r(x) - \theta_y) + l(\theta_{y+1} - r(x))$$

- If $l(s) = I_{(s<0)}$, then the ranking function r^* minimizing $E[l(r, \{\theta_i\}_{i=0}^{K}; X, Y)]$ is

$$r^*(x) = \text{argmax}_j P(Y = j|X)$$

(Dembczyński, Kotłowski and Słowiński (2008))

- Hence the multipartite ranking problem is essentially the same as multiclass classification problem.

Relation to Proportional odds model ($K = 3$)

- $\log \frac{P(Y \leq j|x)}{P(Y > j|x)} = r(x) - \theta_j$
  where $-\infty = \theta_0 < \theta_1 \leq \theta_2 < \theta_3 = \infty$
  $\Leftrightarrow P(Y = j|x) = \frac{1}{1+\exp(r(x)-\theta_j)} - \frac{1}{1+\exp(r(x)-\theta_{j-1})}$
- $r^*(x)$ maximizing the likelihood satisfies
  $$e^{r^*(x)} = \frac{q(x)-1+\sqrt{(q(x)-1)^2+4e^{\theta_1-\theta_2}q(x)}}{2e^{-\theta_2}}$$
  where $q(x) = \frac{p_2(x)+p_3(x)}{p_1(x)+p_2(x)}$
- $e^{r^*(x)}$ is order perserved transformation of $q(x)$.
- This implies that proportional odds model minimizes $E[\hat{U}_{\text{pairs}}(r)]$.

Relation to Ordinal Regression Boosting

$l(s) = e^{-s}$ and it is shown that for $K \geq 3$

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

where $\hat{\theta}_i$ satisfies

$$e^{2\theta_i} = \frac{E_X \left[P(Y = i|X) \left(\frac{\sum_{j=1}^{k-1} P(Y=j+1|X) \exp(\theta_j)}{\sum_{j=1}^{k-1} P(Y=j|X) \exp(-\theta_j)} \right)^{1/2} \right]}{E_X \left[P(Y = i + 1|X) \left(\frac{\sum_{j=1}^{k-1} P(Y=j|X) \exp(-\theta_j)}{\sum_{j=1}^{k-1} P(Y=j+1|X) \exp(\theta_j)} \right)^{1/2} \right]}$$

When $K = 3$

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

Hence this is the best if we set $c_{12} = c_{23} = 1, c_{13} = \exp(\hat{\theta}_2 - \hat{\theta}_1)$. If $\hat{\theta}_2 > \hat{\theta}_1$, the cost of misranking the pair of instances in category 1 and category 3 is higher than the other two types of misranking.

Numerical example for tripartite ranking

Setting:

$$X|Y=1 \sim N(-2,1), X|Y=2 \sim N(0,1) \text{ and } X|Y=3 \sim N(2,1)$$
$$P(Y=1)=P(Y=2)=P(Y=3)=1/3$$

In ORBoost, the number of observation genearted for each category is 2000.

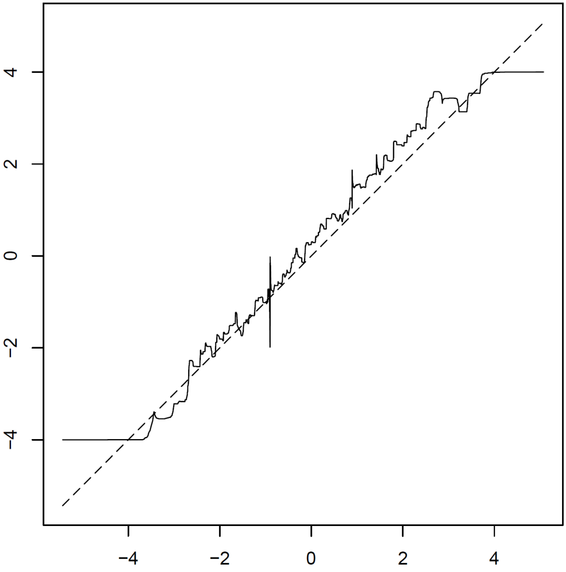
Theoretical result:

$\theta_1 = -1$ and $\theta_2 = 1$

$$r^*(x) = \frac{1}{2}\log\frac{\sum_{j=1}^{2} p_{j+1}(x)\exp(\theta_j)}{\sum_{j=1}^{2} p_j(x)\exp(-\theta_j)} = \frac{1}{2}\log\frac{e^{\theta_1}+e^{-2(1-x)+\theta_2}}{e^{-\theta_2}+e^{-2(1+x)-\theta_1}} = x$$

where $p_j(x) = P(x|Y=j)$

ORBoost (Solid, iteration=2000) and theoretical rank (dashed)



S. Agarwal and P. Niyogi.
Stability and Generalization of Bipartite Ranking Algorithms .
In *Lecture notes in computer science*, 3559:32–47 2005.

U. Brefeld and T. Scheffer.
AUC maximizing support vector learning.
In *Proceedings of the ICML 2005 workshop on ROC Analysis in Machine Learning*, 2005.

W. Chu and S.S. Keerthi.
Support vector ordinal regression.
Neural computation, 19(3):792–815, 2007.

C. Cortes and M. Mohri.
AUC optimization vs. error rate minimization.
In *Advances in Neural Information Processing Systems*, volume 16. 2004.

C. Cortes and M. Mohri.
An alternative ranking problem for search engines.
In *Lecture notes in computer science*, 4525:1–22 2007.

K. Dembczyński, W. Kotłowski, and R. Słowiński.
Ordinal Classification with Decision Rules.
In *Lecture Notes in Computer Science*, volume 20. 2008.

Y. Freund, R. Iyer, R.E. Schapire, and Y. Singer.
An efficient boosting algorithm for combining preferences.
Journal of Machine Learning Research, 4:933–969, 2003.

D. J. Hand and R. J. Hill.
A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems.
Machine Learning, 45(2):171–186, 2001.

J. A. Hanley and B. J. McNeil.
The meaning and use of the area under a receiver operating characteristic (ROC) curve.
Radiology, 143(1):29–36, 1982.

R. Herbrich, T. Graepel, and K. Obermayer.
Large margin rank boundaries for ordinal regression.
In *Advances in Large Margin Classifiers*, pages 115–132. 2000.

H. Lin, and L. Li.
Large-Margin Thresholded Ensembles for Ordinal Regression: Theory and Practice.
In *Lecture Notes in Computer Science*, pages 319–333. 2006.

P. McCullagh.
Regression Models for Ordinal Data.
Journal of Royal Statistical Society (B), pages 109–142. 1980.

T. Qin, T. Liu, J. Xu and H. Li.
LETOR: A Benchmark Collection for Research on Learning to Rank for Information Retrieval.
http://research.microsoft.com/en-us/people/taoqin/letor3.pdf.

A. Rakotomamonjy.
Optimizing area under ROC curve with SVMs.
In *Proceedings of the First Workshop on ROC Analysis in Artificial Intelligence*, 2004.

W. Waegeman, B. D. Baets and L.Boullart.
ROC analysis in ordinal regression learning.
Pattern Recognition Letters, 29(1):1–9, 2008.