From Ordinal Ranking to Binary Classification

Hsuan-Tien Lin

Learning Systems Group, California Institute of Technology

Talk at Caltech CS/IST Lunch Bunch
March 4, 2008

Benefited from joint work with Dr. Ling Li (ALT'06, NIPS'06)
& discussions with Prof. Yaser Abu-Mostafa and Dr. Amrit Pratap

Introduction to Ordinal Ranking

Which Age-Group?

2

1

2

3

4

rank: a finite ordered set of labels $\mathcal{Y} = \{1, 2, \cdots, K\}$

Hot or Not?

http://www.hotornot.com

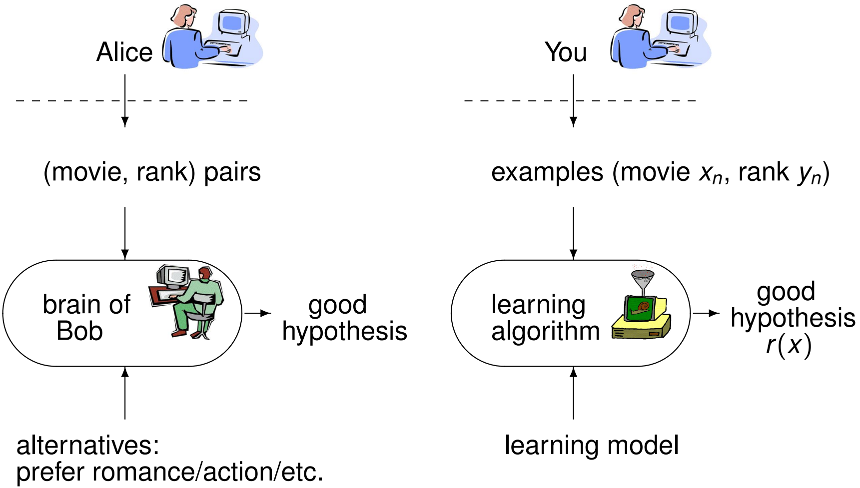
rank: natural representation of human preferences

How Much Did You Like These Movies?

http://www.netflix.com

goal: use "movies you've rated" to automatically predict your preferences (ranks) on future movies

How Machine Learns the Preference of YOU?



challenge: how to make the right-hand-side work?

Ordinal Ranking Problem

- given: N examples (input x_n, rank y_n) $\in \mathcal{X} \times \mathcal{Y}$, e.g.
  `age-group`: $\mathcal{X}$ = encoding(human pictures), $\mathcal{Y} = \{1, \cdots, 4\}$
  `hotornot`: $\mathcal{X}$ = encoding(human pictures), $\mathcal{Y} = \{1, \cdots, 10\}$
  `netflix`: $\mathcal{X}$ = encoding(movies), $\mathcal{Y} = \{1, \cdots, 5\}$
- goal: an ordinal ranker (hypothesis) $r(x)$ that "closely predicts" the ranks y associated with some **unseen** inputs x

a hot and important research problem:

- relatively new for machine learning
- connecting classification and regression
- matching human preferences—many applications in social science and information retrieval

Ongoing Heat: Netflix Million Dollar Prize (since 10/2006)

Leaderboard

Display top 3 leaders.

Rank	Team Name	Best Score	% Improvement	Last Submit Time
--	No Grand Prize candidates yet	--	--	--
Grand Prize - RMSE <= 0.8563				
1	When Gravity and Dinosaurs Unite	0.8686	8.70	2008-02-12 12:03:24
2	BellKor	0.8686	8.70	2008-02-26 23:26:28
3	Gravity	0.8708	8.47	2008-02-06 14:12:44
Progress Prize 2007 - RMSE = 0.8712 - Winning Team: KorBell				
Cinematch score on quiz subset - RMSE = 0.9514				

- a huge joint ordinal ranking problem
- given: each user u (480,189 users) rates N_u (from tens to hundreds) movies—a total of $\sum_u N_u = 100{,}480{,}507$ examples
- goal: personalized predictions $r_u(x)$ on 2,817,131 testing queries (u, x)

the first team being 10% better than original Netflix system gets a million USD

Properties of Ranks $\mathcal{Y} = \{1, 2, \cdots, 5\}$

- representing **order**:
  ★★☆☆☆ $<$ ★★★★★
  —relabeling by $(3, 1, 2, 4, 5)$ erases information

  > general multiclass classification cannot properly use ordering information

- **not** carrying numerical information:
  ★★★★★ not 2.5 times better than ★★☆☆☆
  —relabeling by $(2, 3, 5, 9, 16)$ shouldn't change results

  > general metric regression deteriorates without correct numerical information

ordinal ranking resides uniquely between multiclass classification and metric regression

Cost of Wrong Prediction

- ranks carry no numerical meaning: how to say "closely predict"?
- artificially quantify the **cost** of being wrong


infant (1)


child (2)


teen (3)


adult (4)

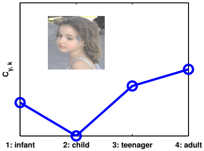
- small mistake—classify a child as a teen;
  big mistake—classify an infant as an adult
- cost vector $\mathbf{c}$ of example $(x, y, \mathbf{c})$:
  $\mathbf{c}[k]$ = cost when predicting (x, y) as rank k
  e.g. for ( , 2), a reasonable cost is $\mathbf{c} = (2, 0, 1, 4)$

closely predict: small cost

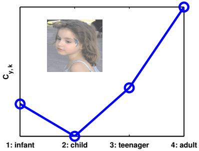
Reasonable Cost Vectors

For an ordinal example $(x, y, \mathbf{c})$, the cost vector $\mathbf{c}$ should

- respect the rank y: $\mathbf{c}[y] = 0$; $\mathbf{c}[k] \geq 0$
- respect the ordinal information: V-shaped or even convex



V-shaped: pay more when predicting further away



convex: pay **increasingly** more when further away

$\mathbf{c}[k] = [\![y \neq k]\!]$	$\mathbf{c}[k] = \lvert y - k \rvert$	$\mathbf{c}[k] = (y - k)^2$
classification: V-shaped only	absolute: convex	squared (Netflix): convex
$(1, 0, 1, 1)$	$(1, 0, 1, 2)$	$(1, 0, 1, 4)$

Our Contributions

a new framework that works with any reasonable cost, and ...

- reduces ordinal ranking to binary classification **systematically**
- unifies and **clearly explains** many existing ordinal ranking algorithms
- makes the design of new ordinal ranking algorithms **much easier**
- allows **simple and intuitive** proof for new ordinal ranking theorems
- leads to **promising experimental results**

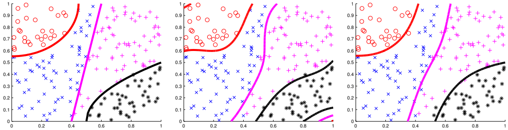
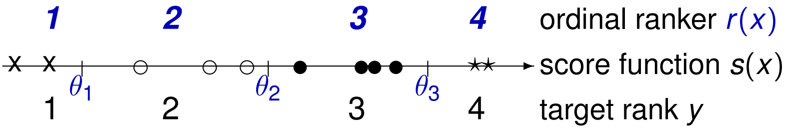


Figure: answer; traditional method; our method

Reduction from Ordinal Ranking to Binary Classification

Thresholded Model

- If we can first compute the score $s(x)$ of a movie x, how can we construct $r(x)$ from $s(x)$?



quantize $s(x)$ by some **ordered** threshold θ

- commonly used in previous work:
  - thresholded perceptrons (PRank, Crammer and Singer, 2002)
  - thresholded hyperplanes (SVOR, Chu and Keerthi, 2005)
  - thresholded ensembles (ORBoost, Lin and Li, 2006)

thresholded model: $r(x) = \min \{k \colon s(x) < \theta_k\}$

Key of Reduction: Associated Binary Questions

getting the rank using a thresholded model

1. is $s(x) > \theta_1$? Yes
2. is $s(x) > \theta_2$? No
3. is $s(x) > \theta_3$? No
4. is $s(x) > \theta_4$? No

generally, how do we query the rank of a movie x?

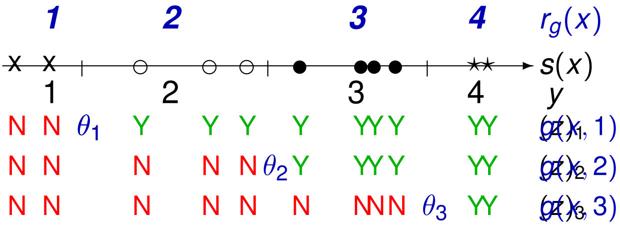
1. is movie x better than rank 1? Yes
2. is movie x better than rank 2? No
3. is movie x better than rank 3? No
4. is movie x better than rank 4? No

associated binary questions $g(x, k)$:
is movie x better than rank k?

More on Associated Binary Questions

$g(x, k)$: is movie x better than rank k?
e.g. thresholded model $g(x, k) = \text{sign}(s(x) - \theta_k)$

- $K - 1$ binary classification problems w.r.t. each k



- let $((x, k), (z)_k)$ be binary examples
  - (x, k): extended input w.r.t. k-th query
  - $(z)_k$: binary label Y/N

if $g(x, k) = (z)_k$ for all k, we can compute $r_g(x)$ from $g(x, k)$ such that $r_g(x) = y$

Computing Ranks from Associated Binary Questions

$g(x, k)$: is movie x better than rank k?

Consider $\big(g(x,1), g(x,2), \cdots, g(x,K-1)\big)$,

- consistent answers: (Y, Y, N, N, $\cdots$, N)
- extracting the rank from consistent answers:
  - minimum index searching: $r_g(x) = \min\{k\colon g(x,k) = \text{N}\}$
  - counting: $r_g(x) = 1 + \sum_k [\![g(x,k) = \text{Y}]\!]$
- two approaches equivalent for consistent answers
- noisy/inconsistent answers? e.g. (Y, N, Y, Y, N, N, Y, N, N)
  —counting is simpler to analyze, and is robust to noise

are all associated binary questions of the same importance?

Importance of Associated Binary Questions

- given a movie x with rank $y = 2$ and $\mathbf{c}[k] = (y-k)^2$

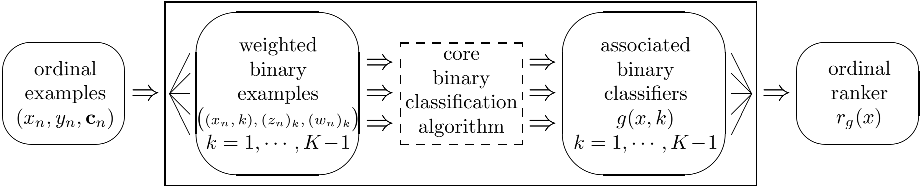
$g(x,1)$: is x better than rank 1?	No	Yes	Yes	Yes
$g(x,2)$: is x better than rank 2?	No	No	Yes	Yes
$g(x,3)$: is x better than rank 3?	No	No	No	Yes
$g(x,4)$: is x better than rank 4?	No	No	No	No
$r_g(x)$	1	2	3	4
$\mathbf{c}\left[r_g(x)\right]$	1	0	1	4

- 1 more for answering question 2 wrong;
  but 3 more for answering question 3 wrong
- $(w)_k \equiv \left|\mathbf{c}[k+1] - \mathbf{c}[k]\right|$: the importance of $\left((x,k),(z)_k\right)$
- per-example error bound (Li and Lin, 2007; Lin, 2008):
  for **consistent answers** or **convex costs**

$$\mathbf{c}\left[r_g(x)\right] \le \sum_{k=1}^{K-1} (w)_k \left[\!\left[(z)_k \ne g(x,k)\right]\!\right]$$

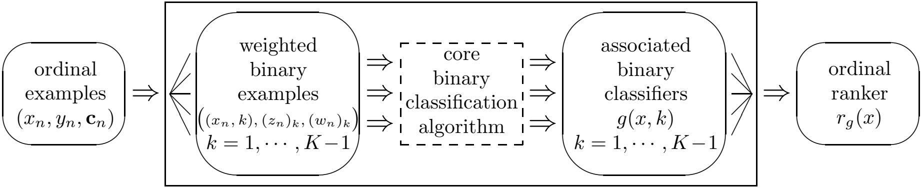
accurate binary answers $\Longrightarrow$ correct ranks

The Reduction Framework



1. transform ordinal examples $(x_n, y_n, \mathbf{c}_n)$ to weighted binary examples $\big((x_n, k), (z_n)_k, (w_n)_k\big)$
2. use your favorite algorithm on the weighted binary examples and get $K-1$ binary classifiers (i.e., one big joint binary classifier) $g(x, k)$
3. for each new input x, predict its rank using $r_g(x) = 1 + \sum_k [\![g(x, k) = \text{Y}]\!]$

Properties of Reduction



- performance guarantee:
  accurate binary answers $\Longrightarrow$ correct ranks
- wide applicability:
  systematic; works with any reasonable **c** and any binary classification algorithm
- up-to-date:
  allows new improvements in binary classification to be immediately inherited by ordinal ranking

If I have seen further it is by standing on the shoulders of Giants—I. Newton

Theoretical Guarantees of Reduction (1/3)

- is reduction a reasonable approach? **YES!**

error transformation theorem (Li and Lin, 2007)

For **consistent answers** or **convex costs**,
 if g makes test error Δ in the induced binary problem,
 then r_g pays test cost at most Δ in ordinal ranking.

- a one-step extension of the per-example error bound
- conditions: general and minor
- performance guarantee in the absolute sense:

accuracy in binary classification $\Longrightarrow$ correctness in ordinal ranking

What if the induced binary problem is "too hard" and even the best g_* can only commit a big Δ?

Theoretical Guarantees of Reduction (2/3)

- is reduction a promising approach? **YES!**

regret transformation theorem (Lin, 2008)

For a general class of **reasonable costs**,
if g is ϵ-close to the optimal binary classifier g_*,
then r_g is ϵ-close to the optimal ordinal ranker r_*.

- error guarantee in the relative setting:

  regardless of the absolute hardness of the induced binary prob.,
  optimality in binary classification $\Longrightarrow$ optimality in ordinal ranking

- reduction does not introduce additional hardness

It is sufficient to go with reduction plus binary classification, but is it necessary?

Theoretical Guarantees of Reduction (3/3)

- is reduction a principled approach? **YES!**

equivalence theorem (Lin, 2008)

For a general class of **reasonable costs**,
ordinal ranking is learnable by a learning model
if and only if binary classification is learnable by the associated learning model.

- a surprising equivalence:

  ordinal ranking is **as easy as** binary classification

- "without loss of generality", we can just focus on binary classification

reduction to binary classification:
systematic, reasonable, promising, and principled

Usefulness of the Reduction Framework

Unifying Existing Algorithms

ordinal ranking	cost	binary classification algorithm
PRank (Crammer and Singer, 2002)	absolute	modified perceptron rule
kernel ranking (Rajaram et al., 2003)	classification	modified hard-margin SVM
SVOR-EXP	classification	modified soft-margin SVM
SVOR-IMC (Chu and Keerthi, 2005)	absolute	modified soft-margin SVM
ORBoost-LR	classification	modified AdaBoost
ORBoost-All (Lin and Li, 2006)	absolute	modified AdaBoost

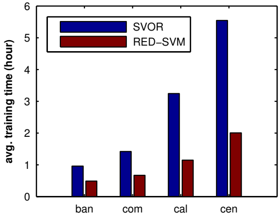
- if the reduction framework had been there, development and implementation time could have been saved
- correctness proof significantly simplified (PRank)
- algorithmic structure revealed (SVOR, ORBoost)

variants of existing algorithms can be designed quickly by tweaking reduction

Designing New Algorithms (1/2)

ordinal ranking	cost	binary classification algorithm
Reduction-C4.5	absolute	standard C4.5 decision tree
Reduction-AdaBoost	absolute	standard AdaBoost
Reduction-SVM	absolute	standard soft-margin SVM

SVOR (modified SVM) v.s. Reduction-SVM (standard SVM):



advantages of core binary classification algorithm inherited in the new ordinal ranking one

Designing New Algorithms (2/2)

AdaBoost (Freund and Schapire, 1997)

for $t = 1, 2, \cdots, T$,

1. find a simple g_t that matches best with the current "view" of $\{(X_n, Y_n)\}$
2. give a larger weight v_t to g_t if the match is stronger
3. update "view" by emphasizing the weights of those (X_n, Y_n) that g_t doesn't predict well

prediction:
majority vote of $\left\{\left(v_t, g_t(x)\right)\right\}$

AdaBoost.OR (Lin, 2008)

for $t = 1, 2, \cdots, T$,

1. find a simple r_t that matches best with the current "view" of $\{(x_n, y_n)\}$
2. give a larger weight v_t to r_t if the match is stronger
3. update "view" by emphasizing the costs $\mathbf{c}_n$ of those (x_n, y_n) that r_t doesn't predict well

prediction:
weighted median of $\left\{\left(v_t, r_t(x)\right)\right\}$

AdaBoost.OR:
an extension of Reduction-AdaBoost;
a parallel of AdaBoost in ordinal ranking

Proving New Theorems

Binary Classification
(Bartlett and Shawe-Taylor, 1998)

For SVM, with prob. $> 1 - \delta$,

$$\text{expected test error} \leq \underbrace{\frac{1}{N}\sum_{n=1}^{N} [\![\bar{\rho}(X_n, Y_n) \leq \Phi]\!]}_{\text{ambiguous training predictions w.r.t. criteria } \Phi} + \underbrace{O\left(\frac{\log N}{\sqrt{N}}, \frac{1}{\Phi}, \sqrt{\log \frac{1}{\delta}}\right)}_{\text{deviation that decreases with stronger criteria or more examples}}$$

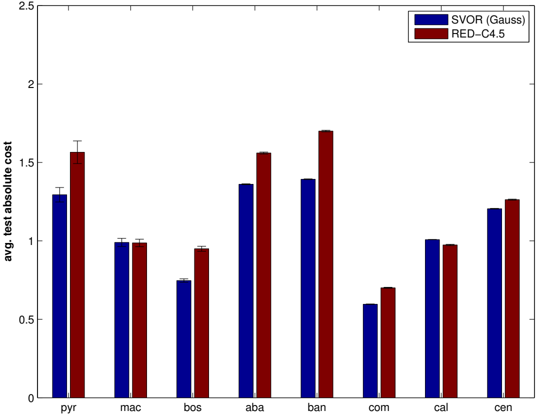
Ordinal Ranking
(Li and Lin, 2007)

For SVOR or Red.-SVM, with prob. $> 1 - \delta$,

$$\text{expected test cost} \leq \underbrace{\frac{\beta}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1} (w_n)_k [\![\bar{\rho}((x_n, k), (z_n)_k) \leq \Phi]\!]}_{\text{ambiguous training predictions w.r.t. criteria } \Phi} + \underbrace{O\left(\frac{\log N}{\sqrt{N}}, \frac{1}{\Phi}, \sqrt{\log \frac{1}{\delta}}\right)}_{\text{deviation that decreases with stronger criteria or more examples}}$$

new test cost bounds with any c$[\cdot]$

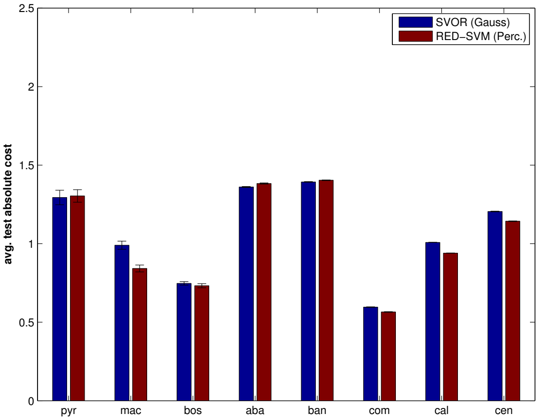
Reduction-C4.5 v.s. SVOR



- C4.5: a (too) simple binary classifier —decision trees
- SVOR: state-of-the-art ordinal ranking algorithm

even simple Reduction-C4.5 sometimes beats SVOR

Reduction-SVM v.s. SVOR



- SVM: one of the most powerful binary classification algorithm
- SVOR: state-of-the-art ordinal ranking algorithm extended from modified SVM

Reduction-SVM without modification often better than SVOR* and faster

Can We Win the Netflix Prize with Reduction?

- possibly
  - a principled view of the problem
  - now easy to apply known binary classification techniques or to design suitable ordinal ranking approaches
    e.g., AdaBoost.OR "boosted" some simple r_t and reduced the test cost from 1.0704 to 1.0343
- but not yet
  - need 0.8563 to win
  - the problem has its own characteristics
    - huge data set: computational bottleneck
    - allows real-valued predictions: $r(x) \in \mathbb{R}$ instead of $r(x) \in \{1, \cdots, K\}$
    - encoding(movie), encoding(user): important

many interesting research problems arose during "CS156b: Learning Systems"

Conclusion

- reduction framework: simple, intuitive, and useful for ordinal ranking
- algorithmic reduction:
  - unifying existing ordinal ranking algorithms
  - designing new ordinal ranking algorithms
- theoretic reduction:
  - new bounds on ordinal ranking test cost
- promising experimental results:
  - some for better performance
  - some for faster training time

reduction keeps ordinal ranking up-to-date with binary classification