Large-Margin Thresholded Ensembles for Ordinal Regression

Hsuan-Tien Lin and Ling Li

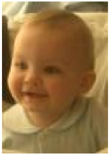
Learning Systems Group, California Institute of Technology, U.S.A.

Conf. on Algorithmic Learning Theory, October 9, 2006

Ordinal Regression

- what is the age-group of the person in the picture?

2



1

2

3

4

- rank: a finite ordered set of labels $\mathcal{Y} = \{1, 2, \cdots, K\}$
- ordinal regression:
  given training set $\{(x_n, y_n)\}_{n=1}^{N}$, find a decision function g that predicts the ranks of unseen examples well
- e.g. ranking movies, ranking by document relevance, etc.

matching human preferences:
applications in social science and info. retrieval

Properties of Ordinal Regression

- regression without metric:
  - possibly metric underlying (age), but not encoded in $\{1, 2, 3, 4\}$
  - monotonic invariance
    – relabel by $\{2, 3, 5, 7\}$ should not change results

general regression deteriorates without metric

- classification with ordered categories:
  - small mistake – classify a teenager as a child; big mistake – classify an infant as an adult
  - no shuffle invariance
    – relabel by $\{3, 1, 2, 4\}$ lose information

general classification cannot use ordering information

ordinal regression resides uniquely between classification and regression

Error Functions for Ordinal Regression

- two aspects of ordinal regression:
  determine the category – discrete nature
  or at least have a close prediction – ordering preference
- categorical prediction: classification error
  $L_C(g, x, y) = \big[g(x) \neq y \big]$
- close prediction: absolute error
  $L_A(g, x, y) = \big| g(x) - y \big|$

neither perfect; both common

Our Contributions

- new model for ordinal regression: thresholded ensemble model
  – combines thresholding and ensemble learning
- new generalization bounds for thresholded ensembles
  – theoretical guarantee of performance
- new algorithms for constructing thresholded ensembles
  – simple and efficient

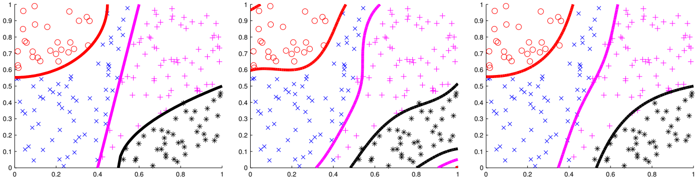
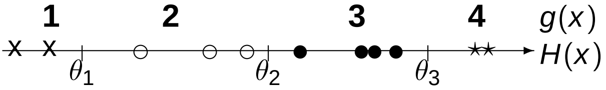


Figure: target; traditional regression; our ordinal regression

promising experimental results

Thresholded Model

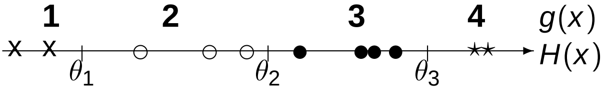
- commonly used in previous work:
  - thresholded perceptrons (PRank, Crammer and Singer, 2005)
  - thresholded SVMs (SVOR, Chu and Keerthi, 2005)
- prediction procedure:
  1. compute a potential function $H(x)$ (e.g. raw perceptron output)
  2. quantize $H(x)$ by some ordered θ to get $g(x)$



thresholded model:

$$g(x) \equiv g_{H,\theta}(x) = \min\{k : H(x) < \theta_k\}$$

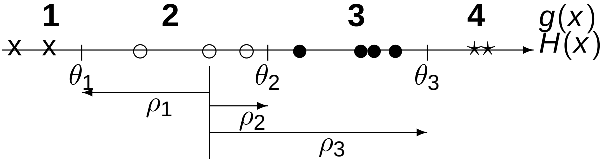
Thresholded Ensemble Model



- the potential function $H(x)$ is a weighted ensemble $H(x) \equiv H_T(x) = \sum_{t=1}^{T} w_t h_t(x)$
- intuition: combine preferences to estimate the overall confidence
- e.g. if many people, h_t, say a movie x is "good", the confidence of the movie $H(x)$ should be high

good theoretical and algorithmic properties inherited from ensemble learning for classification

Margins of Thresholded Ensembles



- margin: safe from the boundary
- normalized margin for thresholded ensemble

$$\bar{\rho}(x, y, k) = \left\{ \begin{array}{l} H_T(x) - \theta_k, \text{ if } y > k \\ \theta_k - H_T(x), \text{ if } y \le k \end{array} \right\} \Bigg/ \left(\sum_{t=1}^{T} |w_t| + \sum_{k=1}^{K-1} |\theta_k| \right)$$

negative margin $\iff$ **wrong prediction**

$$\sum_{k=1}^{K-1} \left[\bar{\rho}(x, y, k) \le 0 \right] \iff \left| g(x) - y \right| = L_A(g, x, y)$$

New Large-Margin Bounds for the Model

- core results: if (x_n, y_n) i.i.d. from $\mathcal{D}$, with prob. $> 1 - \delta$, $\forall \Delta > 0$,

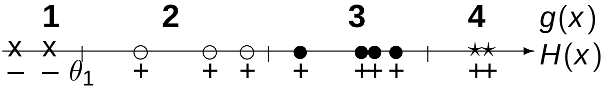
$$\mathcal{E}_{(x,y)\sim\mathcal{D}} L_A(g, x, y) \le \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K-1} [\bar{\rho}(x_n, y_n, k) \le \Delta] + O\left(K \sqrt{\frac{1}{N} \left(\frac{\log^2 N}{\Delta^2} + \log \frac{1}{\delta} \right)} \right)$$

$$\mathcal{E}_{(x,y)\sim\mathcal{D}} L_C(g, x, y) \le \frac{2}{N} \sum_{n=1}^{N} \sum_{k=y_n-1}^{y_n} [\bar{\rho}(x_n, y_n, k) \le \Delta] + O\left(\sqrt{\frac{1}{N} \left(\frac{\log^2 N}{\Delta^2} + \log \frac{1}{\delta} \right)} \right)$$

- sketch of the proof (to be illustrated with L_A):
  1. reduce ordinal regression examples to dependent binary examples
  2. extract i.i.d. binary examples; apply existing classification bounds
  3. bound the deviation caused by the i.i.d. extraction

large-margin thresholded ensembles could generalize

Reduction to Binary Classification



- $K - 1$ binary classification problems w.r.t. each θ_k
- encode (x, y, k) as $\big((X)_k, (Y)_k\big) = \big((x, \mathbf{1}_k), \text{sign}(y - k - 0.5)\big)$:

  $\bar{\rho}(x, y, k) \propto (Y)_k \big(H_T(x) - \langle \theta, \mathbf{1}_k \rangle\big) =$ bin. classifier margin $\rho_C\big((X)_k, (Y)_k\big)$

- key observation:

$$
\begin{aligned}
\mathcal{E}_{(x,y)\sim\mathcal{D}} L_A(g, x, y) &= \mathcal{E}_{(x,y)\sim\mathcal{D}} \sum_{k=1}^{K-1} \big[\bar{\rho}(x, y, k) \le 0\big] \\
&= (K-1)\mathcal{E}_{(x,y)\sim\mathcal{D}, k\sim\mathcal{K}} \big[\bar{\rho}(x, y, k) \le 0\big] \\
&= (K-1)\mathcal{E}_{((X)_k, (Y)_k)\sim\hat{\mathcal{D}}} \big[\rho_C\big((X)_k, (Y)_k\big) \le 0\big]
\end{aligned}
$$

ordinal regression problem $\Longrightarrow$
one big joint binary classification problem

Extraction of Independent Examples

$$\mathcal{E}_{(x,y)\sim\mathcal{D}} L_A(g,x,y) \quad = \quad (K-1)\mathcal{E}_{((X)_k,(Y)_k)\sim\hat{\mathcal{D}}}\left[\rho_C\left((X)_k,(Y)_k\right) \le 0\right]$$

- testing distribution $\hat{\mathcal{D}}$ of $\left((X)_k,(Y)_k\right)$: derived from $(x,y,k) \sim \mathcal{D} \times \mathcal{K}$
- extended training examples $\hat{\mathcal{S}} = \left\{\left((X_n)_k,(Y_n)_k\right)\right\}$: **not** i.i.d. from $\hat{\mathcal{D}}$; cannot be directly used in existing bounds
- i.i.d. subset of $\hat{\mathcal{S}}$: randomly choose k_n for each n
- apply ensemble learning bound (Schapire et al., 1998): if (x_n,y_n,k_n) i.i.d. from $\mathcal{D} \times \mathcal{K}$, with prob. $> 1-\delta$, $\forall\Delta > 0$,

$$\mathcal{E}_{(x,y)\sim\mathcal{D}} L_A(g,x,y) \le \frac{K-1}{N}\sum_{n=1}^{N}\left[\bar{\rho}(x_n,y_n,k_n) \le \Delta\right] + O\left(K\sqrt{\frac{1}{N}\left(\frac{\log^2 N}{\Delta^2} + \log\frac{1}{\delta}\right)}\right)$$

can we obtain a deterministic RHS?

Deviation from the Extraction

$$\mathcal{E}_{(x,y)\sim\mathcal{D}} L_A(g,x,y) \le \frac{K-1}{N}\sum_{n=1}^{N}\left[\bar{\rho}(x_n,y_n,k_n)\le\Delta\right] + O\left(K\sqrt{\frac{1}{N}\left(\frac{\log^2 N}{\Delta^2}+\log\frac{1}{\delta}\right)}\right)$$

- let $b_n = \left[\bar{\rho}(x_n,y_n,k_n)\le\Delta\right]$: binary independent r.v. with mean

$$\mu_n = \frac{1}{K-1}\sum_{k=1}^{K-1}\left[\bar{\rho}(x_n,y_n,k)\le\Delta\right]$$

- extended Chernoff bound: with prob. $> 1-\delta$,

$$\frac{K-1}{N}\sum_{n=1}^{N} b_n \le \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K-1}\left[\bar{\rho}(x_n,y_n,k)\le\Delta\right] + O\left(\sqrt{\frac{1}{N}\log\frac{1}{\delta}}\right)$$

connection between bound and algorithm design? boosting

Boosting for Large-Margin Thresholded Ensembles

- existing algorithm (RankBoost, Freund et al., 2003): construct H_T iteratively with some margin concepts, but no θ
- our work:
  - RankBoost-AE: extended RankBoost for ordinal regression
    – obtain θ by minimizing training L_A using dynamic programming
  - ORBoost: new boosting formulation for ordinal regression

ORBoost:
simpler and faster than existing approaches;
connects well to large-margin bounds

ORBoost: Ordinal Regression Boosting

- inspired from AdaBoost: operationally

$$\min \sum_{n=1}^{N} \exp\bigl(-\rho(x_n, y_n)\bigr) \approx \max \operatorname{softmin}_n \rho(x_n, y_n)$$

- ORBoost:

$$\min \sum_{n=1}^{N} \sum_{k} \exp\bigl(-\rho(x_n, y_n, k)\bigr) \geq \text{const.} \cdot \sum_{n=1}^{N} \sum_{k} \bigl[\rho(x_n, y_n, k) \leq \Delta\bigr]$$

ORBoost-LR

- $k \in \{y_n - 1, y_n\}$
- connects to bound on L_C

ORBoost-All

- $k \in \{1, 2, \cdots, K - 1\}$
- connects to bound on L_A

algorithmic derivation based on theoretical bounds

Advantages of ORBoost

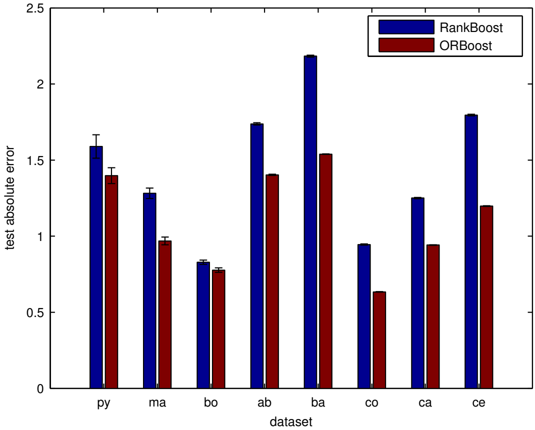
- ensemble learning:
  combine simple preferences to approximate complex targets
- thresholding:
  adaptively estimating scales to perform ordinal regression
- benefits inherited from AdaBoost
  - simple implementation
  - if h_t good enough: guarantee on rapidly minimizing

$$\sum_{n,k} \left[\bar{\rho}(x_n, y_n, k) \le \Delta \right]$$

– decision function g improves with T

ORBoost not very vulnerable to overfitting in practice

ORBoost v.s. RankBoost

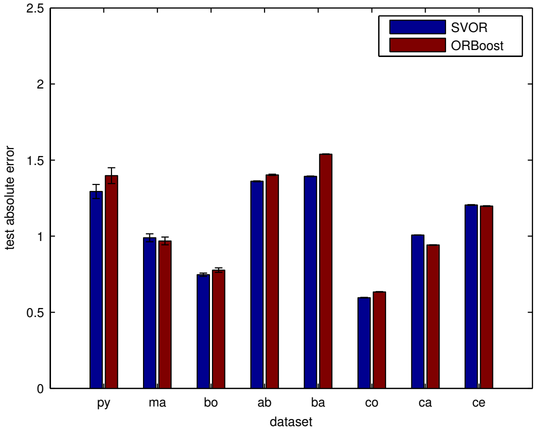


Results (ORBoost-All)

- significantly better than RankBoost (best existing boosting approach)
- simpler to implement and less vulnerable to overfitting

ORBoost: promising boosting approach for ordinal regression

ORBoost v.s. SVOR



Results (ORBoost-All)

- comparable to SVOR (state-of-the-art algorithm)
- much faster in training (1 hour v.s. 2 days on 6000 examples)

ORBoost: could be especially useful for large-scale tasks

Conclusion

- thresholded ensemble model: useful for ordinal regression
  - theoretical reduction: new large-margin bounds
  - algorithmic reduction: new learning algorithms
- ORBoost:
  - simplicity and better performance over existing boosting algorithm
  - comparable performance to state-of-the-art algorithms
  - fast training and not very vulnerable to overfitting
- broader reduction view: many more bounds/algorithms and more general error functions (Li and Lin, NIPS 2006)

Thank you. Questions?