Semi-Supervised Ordinal Regression Based on Empirical Risk Minimization

Taira Tsuchiya[1,2] Nontawat Charoenphakdee[1,2]

Issei Sato[1,2] Masashi Sugiyama[2,1]

1. The University of Tokyo 2. RIKEN AIP

New In ML 2019, Dec 9th, Vancouver

Ordinal Regression Problem

- Medical diagnosis [Bender+ 1997]
  - E.g., diabetic retinopathy: eye disease associated with long-standing diabetes
  - State of patient (retina) is expressed with five **discrete but ordered label** by severity.

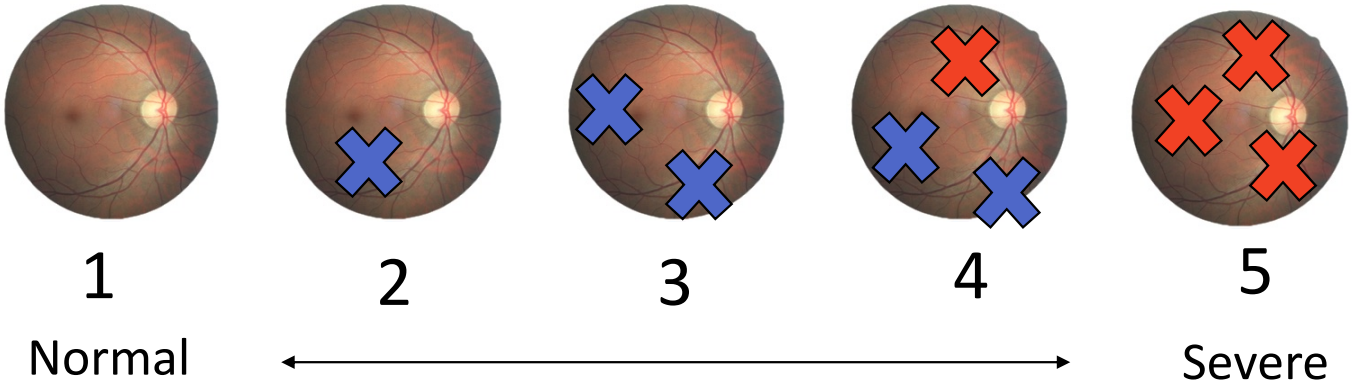


Figure 1. Example of ordinal labels for retina images[1]

Misclassifying 5 to 1 can be more harmful than 5 to 4

1. https://www.kaggle.com/c/diabetic-retinopathy-detection

Ordinal Regression Formulation

[Pedregosa+ JMLR2017]

- Estimate discrete but ordered label y from data point $\boldsymbol{x}$ by function g

$$y = g(\boldsymbol{x}),$$

where $g : \mathcal{X} \to \mathcal{Y} = \{1, \ldots, K\}$.

- Q. How to predict discrete but ordered labels?
- A. Use ***prediction function***: g to predict label

$$g(\boldsymbol{x}; f, \boldsymbol{\theta}) \coloneqq 1 + \sum_{i=1}^{K-1} \mathbb{I}\{f(\boldsymbol{x}) > \theta_i\}$$ ☹ Discrete nature

$$f : \mathcal{X} \to \mathbb{R}$$
$$\boldsymbol{\theta} = [\theta_1, \ldots, \theta_{K-1}]^\top \in \mathbb{R}^{K-1} \; (\theta_1 \leq \theta_2 \leq \cdots \leq \theta_{K-1})$$

- Goal: find g minimizing following ***task risk***

$$\mathcal{R}(g) \coloneqq \mathop{\mathbb{E}}_{X,Y} [\mathcal{L}(g(X), Y)]$$

$\mathcal{L}(g(\boldsymbol{x}), y)$: ***task loss*** (e.g., absolute $|y - g(\boldsymbol{x})|$, squared $(y - g(\boldsymbol{x}))^2$)

Task Losses and their Task Surrogate Losses

[Pedregosa+ JMLR2017]

- Three task losses and corresponding task surrogate losses

Task loss: $\mathcal{L}(g(\boldsymbol{x}), y)$			**Task surrogate loss:** $\psi(\boldsymbol{\alpha}(\boldsymbol{x}), y)$	
Absolute	$\lvert y - g(\boldsymbol{x})\rvert$	$\longleftrightarrow$	$\sum_{i=1}^{y-1} \ell(-\alpha_i) + \sum_{i=y}^{K-1} \ell(\alpha_i)$	all threshold (AT)
Zero-one	$\mathbb{I}\{f(\boldsymbol{x}) \leq \theta_{y-1}\} + \mathbb{I}\{f(\boldsymbol{x}) > \theta_y\}$	$\longleftrightarrow$	$\ell(-\alpha_{y-1}) + \ell(\alpha_y)$	Immediate threshold (IT)
Squared	$(y - g(\boldsymbol{x}))^2$	$\longleftrightarrow$	$\left(y + \alpha_1 - \frac{3}{2}\right)^2$	Least squares (LS)

☹ Minimizing task loss can be computationally intractable

☺ Computationally efficient

$\ell(\cdot)$: binary surrogate loss

$\alpha_i(\boldsymbol{x}) := \theta_i - f(\boldsymbol{x})$

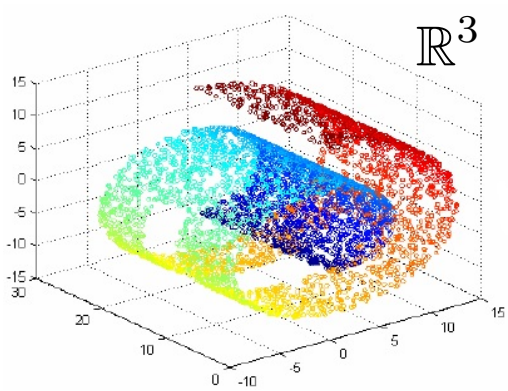
- Surrogate goal: minimize ***task surrogate risk***

$$\mathcal{R}_\psi(g) := \mathop{\mathbb{E}}_{X,Y}\left[\psi(\boldsymbol{\alpha}(X), Y)\right]$$

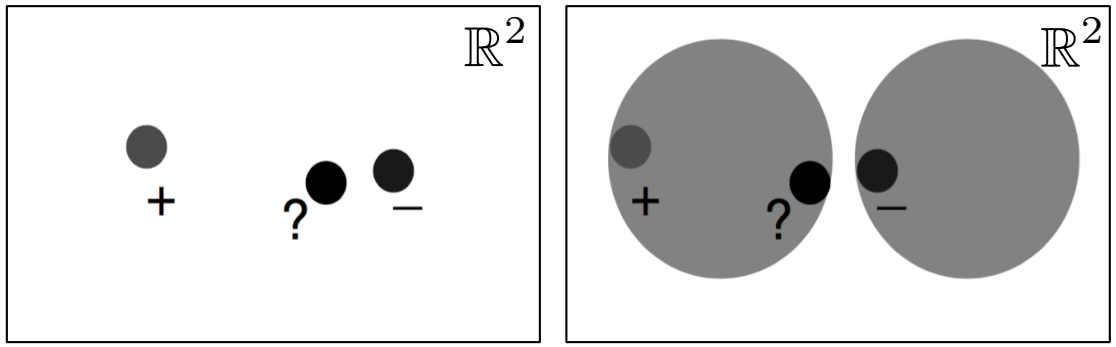
Semi-Supervised Learning

- Labeling process can be costly and time-consuming (especially in medical area.)
  -> Make use of unlabeled data (semi-supervised ordinal regression)
- Usually ***geometric assumption*** on unlabeled data is required.

[Chapelle+ 2010]



Manifold assumption



Cluster assumption
≒ Low-density separation

- ☹ Performance degrades if the geometric assumption does not hold [Sakai+ ICML2017].

Figures from http://vision.cse.psu.edu/seminars/talks/PRML/David_NDR_lecture.pdf, [Chapelle et al. 2010]

Existing Work

We are only aware of three studies for semi-supervised OR.

- Liu et al. (2011) relies on a manifold assumption for image classification.
- Seah et al. (2012) relies on a cluster assumption and cannot predict unseen unlabeled data (transductive).
- Srijith et al. (2013) relies on low-density separation principle using Gaussian process.

	Geometric assumption	Loss agnostic	Theoretical guarantee	Inductive	Computational cost
Liu et al. [3]	Yes	No	No	☺ Yes	High
Seah et al. [4]	Yes	☺ Yes	No	No	Medium
Srijith et al. [5]	Yes	No	No	☺ Yes	High
Ours	☺ No	☺ Yes	☺ Yes	☺ Yes	☺ Low

Problem Setup

Dataset

- Suppose that we are given the following data:

$$\mathcal{X}_{\mathrm{L}} := \{(\boldsymbol{x}_j^{\mathrm{L}}, y_j)\}_{j=1}^{n_{\mathrm{L}}} \overset{\text{i.i.d.}}{\sim} \mathbb{P}(X, Y)$$

n_{L} : # of labeled data

- Also given unlabeled data drawn from marginal distribution:

$$\mathcal{X}_{\mathrm{U}} := \{\boldsymbol{x}_j^{\mathrm{U}}\}_{j=1}^{n_{\mathrm{U}}} \overset{\text{i.i.d.}}{\sim} \mathbb{P}(X) = \sum_{y=1}^{K} \pi_y \mathbb{P}(X|Y=y),\ \pi_y := \mathbb{P}(Y=y)$$

n_{U} : # of unlabeled data

Empirical risk minimization (ERM) [Vapnik 1998]

(Expected) task surrogate risk:

$$\mathcal{R}_\psi(g) := \mathop{\mathbb{E}}_{X,Y}[\psi(\boldsymbol{\alpha}(X), Y)]$$

unbiased

[Estimate risk by finite samples]
Empirical task surrogate risk:

$$\widehat{\mathcal{R}}_\psi(g) := \frac{1}{n_{\mathrm{L}}} \sum_{j=1}^{n_{\mathrm{L}}} \psi(\boldsymbol{\alpha}(x_j), y_j)$$

↑ Minimize this

Proposed Risk Estimator

Idea: use a large portion of U data to estimate risk (inspired by Sakai+ 2017)

(Our goal)

Task surrogate risk : $\mathcal{R}_{\psi}(g) := \mathop{\mathbb{E}}_{X,Y}[\psi(\boldsymbol{\alpha}(X), Y)]$

unbiased

Risk estimator :

$$\widehat{\mathcal{R}}^{\backslash k}_{\psi,\mathrm{LU}}(g) := \sum_{y\in\mathcal{Y}^{\backslash k}} \frac{\pi_y}{n_y} \sum_{j=1}^{n_y} \psi(\boldsymbol{\alpha}(\boldsymbol{x}^y_j), y)$$

$$+ \frac{1}{n_{\mathrm{U}}} \sum_{j=1}^{n_{\mathrm{U}}} \psi(\boldsymbol{\alpha}(\boldsymbol{x}^{\mathrm{U}}_j), k)$$

$$- \sum_{y\in\mathcal{Y}^{\backslash k}} \frac{\pi_y}{n_y} \sum_{j=1}^{n_y} \psi(\boldsymbol{\alpha}(\boldsymbol{x}^y_j), k)$$

* Holds for any $k \in \{1, \ldots, K\}$
* k : removed class

☺ Use unlabeled data

☹ Data labeled as class k is not used

-> Combine with the supervised risk via convex sum

$$\mathcal{R}^{\backslash k}_{\psi,\text{SEMI-}\gamma}(g) := \gamma\mathcal{R}^{\backslash k}_{\psi,\mathrm{LU}}(g) + (1-\gamma)\mathcal{R}_{\psi}(g)$$

Theoretical Analysis (informal)

Prop. (Fisher-consistency)

Task surrogate loss: AT, IT, or LS (Bayes optimal risk)

$\Longrightarrow$ Every minimizer g of $\mathcal{R}^{\backslash k}_{\psi,\text{SEMI-}\gamma}(g)$ reaches $\mathcal{R}^* := \inf_g \mathcal{R}(g)$

☺ Minimizing $\mathcal{R}^{\backslash k}_{\psi,\text{SEMI-}\gamma}(g)$ $\Longleftrightarrow$ minimizing $\mathcal{R}(g)$

Thm. (Estimation error bound)

$$\widehat{g}^{\backslash k} := \arg\min_{g\in\mathcal{G}} \widehat{\mathcal{R}}^{\backslash k}_{\psi,\text{LU}}(g)$$
$$g^* := \arg\min_{g\in\mathcal{G}} \mathcal{R}^{\backslash k}_{\psi,\text{LU}}(g)$$

$$\underbrace{\mathcal{R}_\psi(\widehat{g}^{\backslash k}) - \mathcal{R}_\psi(g^*)}_{\text{Estimation error of empirical risk minimizer}} = \mathcal{O}_p\left(\underbrace{\sum_{y\in\mathcal{Y}^{\backslash k}} \frac{\pi_y}{\sqrt{n_y}}}_{\text{L data not labeled as class } k} + \underbrace{\frac{1}{\sqrt{n_\text{U}}}}_{\text{U data}}\right)$$

☺ Estimation error -> 0 w/ optimal rate w/o strong assumption

Thm. (Variance reduction)

For fixed $g \in \mathcal{G}$, if we select γ properly

$$\mathcal{R}^{\backslash k}_{\psi,\text{SEMI-}\gamma}(g) := \gamma\mathcal{R}^{\backslash k}_{\psi,\text{LU}}(g) + (1-\gamma)\mathcal{R}_\psi(g)$$

$$\text{Var}[\widehat{\mathcal{R}}^{\backslash k}_{\psi,\text{SEMI-}\gamma}(g)] < \text{Var}[\widehat{\mathcal{R}}_\psi(g)]$$

☺ Unlabeled data reduce variance

What Class to Pick Out?

We can arbitrarily select k to estimate risk by our risk $\widehat{\mathcal{R}}^{\setminus k}_{\psi,\mathrm{LU}}(g)$.

Q. How can we choose "best" k ?

Strategy 1: pick k with that has smallest number of data to reduce variance:

$$k = \arg\min_{y \in \mathcal{Y}} n_y$$

Strategy 2: pick k that minimizes upper bound of estimation error:

$$\mathcal{R}_\psi(\widehat{g}^{\setminus k}) - \mathcal{R}_\psi(g^*) = \mathcal{O}_p\left(\sum_{y \in \mathcal{Y}^{\setminus k}} \frac{\pi_y}{\sqrt{n_y}} + \frac{1}{\sqrt{n_{\mathrm{U}}}}\right)$$

Minimize here

$$\Longrightarrow \quad k = \arg\max_{y \in \mathcal{Y}} n_y$$

Experiments 1: Variance Reduction

- Setup
  - Model: linear-in-parameter model
  - **Task surrogate loss: AT, IT, LS**
  - # of trials: 20
- Experiment 1: Variance reduction

See the ratio of variance of empirical risk for random regressor g_{rand}

$$\mathrm{Var}[\widehat{\mathcal{R}}^{\setminus k}_{\psi,\text{SEMI-}\gamma}(g_{\mathrm{rand}})]/\,\mathrm{Var}[\widehat{\mathcal{R}}_{\psi}(g_{\mathrm{rand}})]$$ <- good if smaller than 1

Surr.	car	era	lev	swd	winequality	bank	calhousing	census
AT	0.108	0.115	0.044	0.041	0.058	0.313	0.122	0.234
IT	0.109	0.116	0.042	0.048	0.058	0.360	0.140	0.196
LS	0.157	0.081	0.070	0.065	0.071	1.303	0.275	0.736

Surr.	computer	autompg	abalone	boston	machinecpu	stocks	triazines	wisconsin
AT	0.122	0.165	0.100	0.110	0.097	0.099	0.227	0.181
IT	0.124	0.165	0.104	0.126	0.130	0.078	0.284	0.222
LS	0.387	0.223	0.195	0.065	1.021	0.159	0.176	0.198

☺ Unlabeled data indeed help reduce the variance

Experiments 2: Task Loss-agnostic Property

- Experiment 2: comparison with supervised learning

Task surrogate loss: AT, evaluation metrics: mean squared error

Dataset	SV-Linear	SEMI1-Linear	SEMI2-Linear	SV-Kernel	SEMI1-Kernel	SEMI2-Kernel
car	0.529 (0.09)	0.566 (0.06)	0.449 (0.08)	0.401 (0.12)	0.614 (0.14)	**0.345 (0.02)**
era	0.541 (0.11)	0.464 (0.11)	0.410 (0.12)	0.411 (0.15)	**0.320 (0.06)**	**0.303 (0.04)**
lev	0.493 (0.08)	0.309 (0.16)	0.176 (0.10)	0.168 (0.08)	0.192 (0.14)	**0.127 (0.02)**
swd	0.625 (0.06)	0.670 (0.05)	0.379 (0.09)	**0.253 (0.01)**	0.509 (0.13)	**0.252 (0.01)**
winequality	0.424 (0.08)	0.457 (0.11)	0.034 (0.02)	**0.020 (0.01)**	0.291 (0.19)	**0.018 (0.00)**
bank	0.651 (0.07)	0.640 (0.06)	0.639 (0.05)	**0.438 (0.11)**	0.464 (0.10)	**0.406 (0.02)**
calhousing	0.655 (0.06)	0.619 (0.10)	0.519 (0.08)	0.463 (0.09)	0.466 (0.11)	**0.418 (0.04)**
census	0.635 (0.09)	0.624 (0.08)	0.515 (0.10)	0.447 (0.11)	0.451 (0.12)	**0.405 (0.02)**
computer	0.609 (0.16)	0.602 (0.16)	0.499 (0.16)	**0.471 (0.18)**	0.522 (0.19)	**0.412 (0.04)**
autompg	0.439 (0.22)	0.467 (0.21)	0.359 (0.17)	**0.346 (0.24)**	**0.267 (0.02)**	**0.263 (0.03)**
abalone	0.612 (0.13)	0.470 (0.15)	0.370 (0.11)	0.347 (0.10)	**0.307 (0.00)**	**0.307 (0.00)**
boston	0.536 (0.15)	0.431 (0.16)	0.387 (0.15)	0.450 (0.25)	**0.275 (0.16)**	**0.229 (0.03)**
machinecpu	0.494 (0.26)	0.536 (0.29)	**0.375 (0.16)**	0.431 (0.19)	0.587 (0.19)	**0.346 (0.07)**
stocks	0.520 (0.12)	0.488 (0.15)	0.329 (0.10)	0.346 (0.10)	**0.320 (0.12)**	**0.274 (0.03)**
triazines	0.783 (0.06)	0.774 (0.06)	0.733 (0.09)	0.349 (0.00)	0.635 (0.23)	**0.349 (0.00)**
wisconsin	0.578 (0.07)	0.557 (0.07)	0.568 (0.08)	0.512 (0.09)	**0.454 (0.01)**	0.483 (0.06)

☺ Estimation error bound based strategy works well

☺ Works well for most of datasets, all evaluation metrics

Same tendency for (IT, mean zero-one error), (LS, mean squared error)

Conclusion

- Proposed a novel semi-supervised ordinal regression framework based on unbiased risk estimator

Evaluation metrics agnostic

	Geometric assumption	Loss agnostic	Theoretical guarantee	Inductive	Computational cost
Liu et al. [3]	Yes	No	No	☺ Yes	High
Seah et al. [4]	Yes	☺ Yes	No	No	Medium
Srijith et al. [5]	Yes	No	No	☺ Yes	High
Ours	☺ No	☺ Yes	☺ Yes	☺ Yes	☺ Low

Use unlabeled data to estimate risk

- Consistency
- Estimation error bound
- Variance reduction

References

[1] Fabian Pedregosa, Francis Bach, and Alexandre Gramfort. On the consistency of ordinal regression methods. Journal of Machine Learning Research, 18:1 – 35, 2017.

[2] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. Semi-Supervised Learning. MIT Press, 2006.

[3] Yang Liu, Yan Liu, Shenghua Zhong, and Keith CC Chan. Semi-supervised manifold ordinal regression for image ranking. In ACMMM, pages 1393–1396, 2011.

[4] Chun-Wei Seah, Ivor W Tsang, and Yew-Soon Ong. Transductive ordinal regression. IEEE Transactions on Neural Networks and Learning Systems, 23(7):1074–1086, 2012.

[5] PK Srijith, Shirish Shevade, and S Sundararajan. Semi-supervised Gaussian process ordinal regression. In ECML-PKDD, pages 144–159, 2013.

[6] Tomoya Sakai, Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. Semi- supervised classification based on classification from positive and unlabeled data. In ICML, volume 70, pages 2998–3006, 2017.