JMLR: Workshop and Conference Proceedings 39:189–204, 2014 ACML 2014

Interval Insensitive Loss for Ordinal Classification

Kostiantyn Antoniuk ANTONKOS@CMP.FELK.CVUT.CZ
Vojtěch Franc XFRANCV@CMP.FELK.CVUT.CZ
Václav Hlaváč HLAVAC@FEL.CVUT.CZ
Center for Machine Perception, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Technická 2, 166 27 Prague 6 Czech Republic

Editor: Dinh Phung and Hang Li

Abstract

We address a problem of learning ordinal classifier from partially annotated examples. We introduce an interval-insensitive loss function to measure discrepancy between predictions of an ordinal classifier and a partial annotation provided in the form of intervals of admissible labels. The proposed interval-insensitive loss is an instance of loss functions previously used for learning of different classification models from partially annotated examples. We propose several convex surrogates of the interval-insensitive loss which can be efficiently optimized by existing solvers. Experiments on standard benchmarks and a real-life application show that ordinal classifiers learned from partially annotated examples can achieve accuracy close to the accuracy of classifiers learned from completely annotated examples.

1. Introduction

We address problem of learning ordinal classifiers from partially annotated examples. The ordinal classification model (also ordinal regression, ranking) is used in problems where the set of labels is fully ordered, e.g. the label can be an age category (0-9,10-19,...,90-99) or a respondent answer to certain question (e.g. from strongly agree to strongly disagree). The ordinal classifiers are routinely used in social sciences, epidemiology, information retrieval or computer vision.

Recently, many algorithms have been proposed for discriminative learning of the ordinal classifiers from completely annotated examples. The discriminative methods learn parameters of an ordinal classifier (the form of which is assumed to be known up to the parameters) by minimizing a (regularized) empirical risk. A Perceptron-like on-line algorithm PRank has been proposed in Crammer and Singer (2001). A large-margin principle has been applied to learning ordinal classifiers in Shashua and Levin (2002). The paper Chu and Keerthi (2005) proposed the Support Vector Ordinal Regression algorithm with explicit constraints (SVOR-EXP) and the SVOR algorithm with implicit constraints (SVOR-IMC). Unlike Shashua and Levin (2002), the SVOR-EXP and SVOR-IMC guarantee that the learned ordinal classifier is statistically plausible. The same approach have been proposed independently by Rennie and Srebro (2005) who introduce so called immediate-threshold loss and all-thresholds loss functions. Minimization of a quadratically regularized immediate-threshold loss and the all-threshold loss are equivalent to the SVOR-EXP and the SVOR-IMC formulation of Shashua and Levin (2002), respectively. A generic framework proposed in Li and Lin

(2006), of which the SVOR-EXP and SVOR-IMC are special instances, allows to convert learning of the ordinal classifier into learning of two-class SVM classifier with weighted examples.

Estimating parameters of a probabilistic model by the Maximum Likelihood (ML) method is another paradigm that has been used for learning ordinal classifiers. A plug-in ordinal classifier can be then constructed by substituting the estimated model to the optimal decision rule derived for a particular loss function (see e.g. Debczynski et al. (2008) for a list of losses and corresponding decision functions suitable for ordinal classification). Parametric probability distributions designed to model the ordinal labels have been proposed in McCullagh (1980); Fu and Simpson (2002); Rennie and Srebro (2005). Besides the parametric methods, the non-parametric probabilistic approaches like the Gaussian processes have been also proposed (e.g. Chu and Ghahramani (2005)).

Properties of the discriminative and the ML based methods are complementary to each other. The ML approach can be directly applied in the presence of incomplete annotation (e.g. when label interval is given instead of a single label as considered in this paper) by using the Expectation-Maximization algorithms (Dempster et al. (1997)). However, the ML methods are sensitive to model mis-specification which complicates their application in modeling complex high-dimensional data. In contrast, the discriminative methods are known to be robust against the model mis-specification while their extension for learning from partial annotations is not straightforward. To our best knowledge, the existing discriminative approaches for ordinal classification assume the complete annotation only, i.e. that each training input is annotated by exactly one label.

In this paper, we consider the discriminative learning of the ordinal classifiers from partially annotated examples. We assume that each training input is annotated by an interval of admissible labels rather than a single label. This setting is common for example in computer vision applications. For example, consider learning of an ordinal classifier predicting a person age from an image of his/her face (e.g. Ramanathan and Chellappa (2009); Chang et al. (2011)). In this case, examples of face images can be downloaded from the Internet and the age estimated by a human annotator. However, providing a reliable year-exact age just from a facial image is difficult if not possible, hence, it is more natural and easier for the annotator to provide an interval of admissible ages. The interval annotation can be also obtained in an automated way e.g. by the method of Kotlowski et al. (2008) removing inconsistencies in the data.

To deal with the interval annotations, we propose an interval-insensitive loss function which extends an arbitrary standard loss defined for two labels to the interval setting. The interval-insensitive loss measures a discrepancy between the interval of possible labels given in the annotation and a label predicted by the classifier. Our interval-insensitive loss can be seen as the ordinal regression counterpart of the ϵ-insensitive loss used in the Support Vector Regression (Vapnik (1998)). We propose a generic convex approximation of the interval-insensitive loss that can be efficiently optimized by existing solvers for convex risk minimization. We also derive interval-insensitive variant of existing SVOR-EXP and the SVOR-IMC algorithms which makes them applicable for learning from partial annotations.

Discriminative learning from partially annotated examples has been recently studied in the context of a generic multi-class classifiers (Cour et al. (2011)), the Hidden Markov Chain based classifiers (Do and Artieres (2009)), generic structured output models (Lou

and Hamprecht (2012)), the multi-instance learning (Jie and Orabona (2010)) etc. All these methods translate learning to minimization of a partial loss evaluating discrepancy between the classifier predictions and partial annotations. In most cases the partial loss is defined as minimal value of a standard loss (i.e. loss defined on a pair of labels, e.g. 0/1-loss) over all admissible labels consistent with the partial annotation. Our interval-insensitive loss can be seen as an application of such type of partial losses in the context of the ordinal classification. It worth mentioning that the ordinal classification model allows to design reasonable convex surrogate of the partial loss in contrast to previously considered classification models which either require crude approximation (Cour et al. (2011)) or require minimization of non-convex surrogate loss hard to optimize (Do and Artieres (2009); Lou and Hamprecht (2012); Jie and Orabona (2010)).

The paper is organized as follows. Formulation of the learning problem and the proposed interval-insensitive loss is given in section 2. Algorithms minimizing convex surrogate of the interval-insensitive loss are derived in section 3. Section 4 provides experimental evaluation of the proposed methods and section 5 concludes the paper.

2. Learning ordinal classifier with the proposed interval-insensitive loss

We consider learning of a classifier for ordinal regression $h\colon \mathbb{R}^n \to \mathcal{Y} = \{1, \dots, Y\}$ of the form

$$h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta}) = 1 + \sum_{k=1}^{Y-1} [\![\langle \boldsymbol{x}, \boldsymbol{w} \rangle > \theta_k]\!] \tag{1}$$

where $\boldsymbol{w} \in \mathbb{R}^n$ and $\boldsymbol{\theta} \in \Theta = \{\boldsymbol{\theta}' \in \mathbb{R}^{Y-1} \mid \theta'_y \leq \theta'_{y+1},\ y = 1, \dots, Y-1\}$ are admisible parameters. W.l.o.g. the set of labels $\mathcal{Y} = \{1, \dots, Y\}$ is composed of natural numbers endowed with a natural order. The classifier (1) splits the space of projections $\langle \boldsymbol{x}, \boldsymbol{w} \rangle$ into Y consecutive intervals defined by thresholds $\theta_1 \leq \theta_2 \leq \dots \leq \theta_{Y-1}$. The observation $\boldsymbol{x}$ is assigned a label correspoding to the interval to which the projection $\langle \boldsymbol{w}, \boldsymbol{x} \rangle$ falls. The classifier (1) is a proper model if the label can be thought of as a rough measurement of a continuous random variable $\xi(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{w} \rangle + \text{noise}$ McCullagh (1980).

Discriminative methods for learning paramaters $(\boldsymbol{w}, \boldsymbol{\theta})$ of the classifier (1) from a set of completely annotated examples $\{(\boldsymbol{x}^1, y^1), \dots, (\boldsymbol{x}^m, y^m)\} \in (\mathbb{R}^n \times \mathcal{Y})^m$ has been studied e.g. in Crammer and Singer (2001); Shashua and Levin (2002); Chu and Keerthi (2005); Li and Lin (2006).

In this paper we address the problem of learning the classifier (1) from partially (or weakly) annotated examples $\{(\boldsymbol{x}^1, [y_l^1, y_r^1]), \dots, (\boldsymbol{x}^m, [y_l^m, y_r^m])\} \in (\mathcal{X} \times \mathcal{P})^m$ where $\mathcal{P} = \{[y_l, y_r] \in \mathcal{Y}^2 \mid y_l \leq y_r\}$ is a set of all possible partial annotations. The partial annotation $(\boldsymbol{x}, [y_l, y_r])$ means that a label of $\boldsymbol{x}$ is from interval $[y_l, y_r] = \{y \in \mathcal{Y} \mid y_l \leq y \leq y_r\}$.

We consider the following statistical formulation of the learning problem. The nature generates an observation $\boldsymbol{x} \in \mathbb{R}^n$, characterizing a studied object, according to some unknown p.d.f. $p(\boldsymbol{x})$. Besides the input $\boldsymbol{x}$, the object is also characterized by a hidden label $y \in \mathcal{Y}$. A human expert provided with $\boldsymbol{x}$ can give an estimate of the hidden label in form of an interval $[y_l, y_r] \in \mathcal{P}$. We assume that the expert can be modeled as stochastic process giving the estimates according to some unknown p.d.f. $p([y_l, y_r] \mid \boldsymbol{x})$. Our goal is to learn an ordinal classifier (1) whose predictions are as close as possible to annotations of

the expert. To measure distance between the expert's partial annotation $[y_l, y_r] \in \mathcal{P}$ and the classifier's precise prediction $h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta}) \in \mathcal{Y}$, we propose the *interval-insensitive loss* function $\Delta_I \colon \mathcal{P} \times \mathcal{Y} \to \mathbb{R}$ defined as follows:

$$\Delta_I(y_l, y_r, y) = \begin{cases} 0 & \text{if} \quad y \in [y_l, y_r] \\ \Delta(y, y_l) & \text{if} \quad y < y_l \\ \Delta(y, y_r) & \text{if} \quad y > y_r \end{cases} \tag{2}$$

where $\Delta \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a standard loss defined on pair of labels (we use term *standard loss* later in the text to distinguish from the interval-insensitive loss), e.g. the mean absolute error (MAE) $\Delta(y, y') = |y - y'|$ or the 0/1-loss $\Delta(y, y') = [\![y \neq y']\!]$ are most frequently used in the context of the ordinal classification. The interval-insensitive loss $\Delta(y_l, y_r, y)$ does not penalize predictions which are in the annotated interval $[y_l, y_r]$ otherwise the penalty is either $\Delta(y, y_l)$ or $\Delta(y, y_r)$ depending which border label of the interval $[y_l, y_r]$ is closer. In the special case when $\Delta(y, y') = |y - y'|$, one can think of the interval-insensitive loss $\Delta_I(y_l, y_r, y)$ as the ordinal regression counterpart of the ϵ-insensitive loss used in the Support Vector Regression Vapnik (1998).

Our task is to find Bayes classifier minimizing the expectation of the interval-insensitive loss, i.e.

$$(\boldsymbol{w}^*, \boldsymbol{\theta}^*) \in \operatorname*{Argmin}_{\boldsymbol{w} \in \mathbb{R}^n, \boldsymbol{\theta} \in \Theta} R(\boldsymbol{w}, \boldsymbol{\theta}) := \mathbb{E}_{(\boldsymbol{x}, y_l, y_r) \sim p(\boldsymbol{x}, y_l, y_r)} (\Delta_I(y_l, y_r, h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta}))) \tag{3}$$

with the expectation taken w.r.t. $p(\boldsymbol{x}, y_l, y_r) = p(\boldsymbol{x}) p(y_l, y_r \mid \boldsymbol{x})$. We adopt the (regularized) empirical risk minimization framework and learn parameters of the ordinal classifier by minimizing convex surrogate of the empirical risk

$$R_{\text{emp}}(\boldsymbol{w}, \boldsymbol{b}) = \frac{1}{m} \sum_{i=1}^{m} \Delta_I(y_l^i, y_r^i, h(\boldsymbol{x}^i; \boldsymbol{w}, \boldsymbol{b})) \tag{4}$$

that can be evaluated on the partially annotated examples $\{(\boldsymbol{x}^1, [y_l^1, y_r^1]), \ldots, (\boldsymbol{x}^m, [y_l^m, y_r^m])\} \in (\mathcal{X} \times \mathcal{P})^m$.

Before proposing algorithms implementing the risk minimization approach in the next section, we connect the problem (3) with the formulation of the partial learning considered e.g. in Cour et al. (2011); Do and Artieres (2009); Lou and Hamprecht (2012); Jie and Orabona (2010). Assume we want to minimize the expected risk with the standard loss $\Delta(y, y')$, then we can write

$$R'(\boldsymbol{w}, \boldsymbol{\theta}) = \mathbb{E}_{(\boldsymbol{x}, y) \sim p(\boldsymbol{x}, y)} \Delta(y, h(\boldsymbol{x}; \boldsymbol{w}, b))) = \mathbb{E}_{(\boldsymbol{x}, y_l, y_r) \sim p(\boldsymbol{x}, y_l, y_r)} (\Delta'(y_l, y_r, h(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{\theta}))\,,$$

where $\Delta'(y_l, y_r, y') = \sum_{y \in \mathcal{Y}} p(y \mid \boldsymbol{x}, y_l, y_r) \Delta(y, y')$. Let us assume that the loss $\Delta(y, y')$ is V-shaped Li and Lin (2006), i.e. $\Delta(y, k-1) \geq \Delta(y, k)$ if $k \leq y$ and $\Delta(y, k) \leq \Delta(y, k+1)$ if $k \geq y$. Let also assume that the annotations are consistent with the true hidden label in the sense that $y \notin [y_l, y_r]$ implies $p(y \mid \boldsymbol{x}, y_l, y_r) = 0$. Under the two assumptions we can see that

$$\Delta'(y_l, y_r, y') = \sum_{y=y_l}^{y_r} p(y \mid \boldsymbol{x}, y_l, y_r) \Delta(y, y') \geq \min_{y_l \leq y' \leq y_r} \Delta(y, y') = \Delta_I(y_l, y_r, y')\,.$$

Thus the interval-insensitive loss equals to the minimal value of the standard loss over all labels consistent with the annotation, i.e. the interval-insensitive loss is an instance of partial loss previously used for learning different classification models from partial annotations.

3. Learning with interval-insensitive loss

The goal is to derive algorithms minimizing a convex surrogate of the empirical risk $R_{\text{emp}}(\boldsymbol{w}, \boldsymbol{b})$ defined by (4). A direct application of the existing algorithms for supervised ordinal regression is not possible. The generic framework of Li and Lin (2006) is defined for the standard losses satisfying $\Delta(y, y') > 0$, $y \neq y'$, which prevents its direct extension to the interval-insensitive setting.

In section 3.2, we derive a generic algorithm which minimizes a novel convex upper bound of the interval-insensitive loss $\Delta_I(y_l, y_r, y)$ induced from any standard loss $\Delta(y, y')$. The learning is translated to a convex unconstrained minimization problem solvable by existing large-scale solvers like the Stochastic Gradient Descent or the Cutting Plane based solvers.

We also show that the state-of-the-art Support Vector Ordinal Regression (SVOR) algorithms Chu and Keerthi (2005); Li and Lin (2006), can be extended to minimize the interval-insensitive loss. First, in section 3.3, we extend the SVOR with the explicit constraints (SVOR-EXP) which originally minimizes the standard 0/1-loss. Second, in section 3.4, we extend the SVOR with the implicit constraints (SVOR-IMC) which minimizes the mean-absolute-error (MAE) being another loss frequently used in the context of ordinal classification. We name the modified algorithms IIL-SVOR-EXP and IIL-SVOR-IMP, respectively. We also show that the proposed generic upper bound provides a tighter approximation of the interval-insensitive loss than the modified IIL-SVOR-IMC.

Before introducing the generic algorithm, we show in section 3.1 that the ordinal classifier is an instance of a multi-class linear classifier. This equivalence allows to use a convex loss approximation technique known from the structured output SVMs Tsochantaridis et al. (2005).

3.1. Ordinal classifier as multi-class linear classifier

The ordinal classifier (1) can be reparametrized as a multi-class linear classifier, in the sequel denoted as multi-class ordinal (MORD) classifier, which reads

$$h'(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{b}) = \operatorname*{argmax}_{y \in \mathcal{Y}} \Big(\langle \boldsymbol{x}, \boldsymbol{w} \rangle \cdot y + b_y \Big) \tag{5}$$

where $\boldsymbol{w} \in \mathbb{R}^n$ and $\boldsymbol{b} = (b_1, \ldots, b_Y) \in \mathbb{R}^Y$ are parameters. Note that the MORD classifier has $n + Y$ parameters and any pair $(\boldsymbol{w}, \boldsymbol{b}) \in (\mathbb{R}^n \times \mathbb{R}^Y)$ is admissible. In contrast, the original ordinal classifier (1) has $n + Y - 1$, however, the admissible parameters must satisfy $(\boldsymbol{w}, \boldsymbol{\theta}) \in (\mathbb{R}^n \times \Theta)$. It is seen that the MORD reparametrization is more suitable for learning as we do not need to care about constraints $\boldsymbol{\theta} \in \Theta$. The equivalence between the standard ordinal classifier (1) and the MORD reparametrization (5) follows from Theorem 1.

Theorem 1 *The ordinal classifier (1) and the MORD classifier (5) are equivalent in the following sense. For any $\boldsymbol{w} \in \mathbb{R}^n$ and admissible $\boldsymbol{\theta} \in \Theta$ there exists $\boldsymbol{b} \in \mathbb{R}^Y$ such that*

$h(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\theta}) = h'(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{b})$, $\forall \boldsymbol{x} \in \mathbb{R}^n$. *For any* $\boldsymbol{w} \in \mathbb{R}^n$ *and* $\boldsymbol{b} \in \mathbb{R}^n$, *there exists admissible* $\boldsymbol{\theta} \in \Theta$ *such that* $h(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\theta}) = h'(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{b})$, $\forall \boldsymbol{x} \in \mathbb{R}^n$.

Proof of Theorem 1 is given in Antoniuk et al. (2013) as well as conversion formulas between these parametrizations, however they are not needed in this paper.

3.2. Generic algorithm for interval-insensitive loss minimization

We propose to learn the parameters of the MORD classifier (5) by minimizing a convex risk

$$(\boldsymbol{w}^*, \boldsymbol{b}^*) = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^n, \boldsymbol{b} \in \mathbb{R}^Y} F(\boldsymbol{w}, \boldsymbol{b}) := \frac{\lambda}{2} \Omega(\boldsymbol{w}, \boldsymbol{b}) + \frac{1}{m} \sum_{i=1}^{m} \Delta'_I(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{x}^i, y_l^i, y_r^i) \,, \tag{6}$$

where

$$\begin{aligned} \Delta'_I(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{x}, y_l, y_r) \quad = \quad & \max_{y \le y_l} \Big[\Delta(y, y_l) + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_l) + b_y - b_{y_l} \Big] \\ & + \max_{y \ge y_r} \Big[\Delta(y, y_r) + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_r) + b_y - b_{y_r} \Big] \end{aligned}$$

is a convex upper bound of the interval-insensitive loss given by formula (2) as stated in Theorem 2. The term $\Omega(\boldsymbol{w}, \boldsymbol{b})$ is a regularizer typically set to be a quadratic function $\Omega(\boldsymbol{w}, \boldsymbol{b}) = \|\boldsymbol{w}\|^2$ or $\Omega(\boldsymbol{w}, \boldsymbol{b}) = \|\boldsymbol{w}\|^2 + \|\boldsymbol{b}\|^2$. Other convex regularizers can be potentially used as well though in experiments we consider only the quadratic one.

Theorem 2 *For any* $\boldsymbol{x} \in \mathbb{R}^n$, $[y_l, y_r] \in \mathcal{P}$, $\boldsymbol{w} \in \mathbb{R}^n$ *and* $\boldsymbol{b} \in \mathbb{R}^Y$ *the inequality*

$$\Delta_I(y_l, y_r, h'(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{b})) \le \Delta'_I(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{x}, y_l, y_r)$$

holds where $h'(\boldsymbol{x}; \boldsymbol{w}, \boldsymbol{b})$ *denotes response of the MORD classifier (5).*

PROOF: Let us first consider triplet of labels (y, y_l, y_r) such that $y \notin [y_l, y_r]$. In this case, the left max-term $\max_{y \le y_l} \big[\Delta(y, y_l) + \big\langle \boldsymbol{x}, \boldsymbol{w} \big\rangle (y - y_l) + b_y - b_{y_l}\big]$ is an instance of well known margin-rescaling upper bound of Tsochantaridis et al. (2005) applied for the standard loss $\Delta(y, y_l)$ defined on labels $y \in [1, y_l - 1]$ and, in turn, it is also an upper bound of $\Delta_I(y_l, y_r, y)$. Analogically, we can see that the right max-term $\max_{y \ge y_r} \big[\Delta(y, y_r) + \big\langle \boldsymbol{x}, \boldsymbol{w} \big\rangle (y - y_r) + b_y - b_{y_r}\big]$ is margin-rescaling upper bound of the loss $\Delta(y, y_r)$ on labels $y \in [y_r + 1, Y]$ and, in turn, also upper bound of $\Delta_I(y_l, y_r, y)$. Thanks to the assumption that $\Delta(y, y) = 0$, $\forall y$, both max-terms are non-negative therefore their sum upper bounds the value of $\Delta_I(y_l, y_r, y)$ for $y \notin [y_l, y_r]$. In the case when $y \in [y_l, y_r]$ the value of $\Delta_I(y_l, y_r, y)$ is defined to be zero hence it is also upper bounded by the sum of the non-negative max-terms. ■

3.3. Modified SVOR-EXP for minimization of interval-insensitive 0/1-loss

The original SVOR-EXP algorithm for learning the parameters of the classifier (1) from completely annotated examples $\{(\boldsymbol{x}^1, y^1), \ldots, (\boldsymbol{x}^m, y^m)\} \in (\mathbb{R}^n \times \mathcal{Y})^m$ is defined as a convex quadratic program Chu and Keerthi (2005)

$$(\boldsymbol{w}^*, \boldsymbol{\theta}^*) = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^n, \boldsymbol{\theta} \in \mathbb{R}^{Y-1}} \left[\frac{\lambda}{2} ||\boldsymbol{w}||^2 + \sum_{j=1}^{r-1} \Big(\sum_{i=1}^{n^j} \xi_i^j + \sum_{i=1}^{n^{j+1}} \xi_i^{*j+1} \Big) \right] \tag{7}$$

subject to

$$\begin{aligned} \langle \boldsymbol{x}_i^j, \boldsymbol{w} \rangle - b_j \leq -1 + \xi_i^j,\ \xi_i^j \geq 0,\ \forall i = 1, \ldots, n^j\,, \\ \langle \boldsymbol{x}_i^{j+1}, \boldsymbol{w} \rangle - b_{j-1} \geq 1 - \xi_i^{*j+1},\ \xi_i^{*j+1} \geq 0,\ \forall i = 1, \ldots, n^{j+1}\,, \\ b_{j-1} \geq b_j,\ \forall j\,. \end{aligned}$$

Using auxiliary variables $\theta_0 = -\infty$ and $\theta_Y = \infty$, we can rewrite (7) as an equivalent problem

$$\begin{aligned} (\boldsymbol{w}^*, \boldsymbol{\theta}^*) \quad = \quad & \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^n, \boldsymbol{\theta} \in \Theta} F'(\boldsymbol{w}, \boldsymbol{\theta}) := \\ & \frac{\lambda}{2} ||\boldsymbol{w}^2|| + \sum_{i=1}^{m} \Big[\max(1 - \langle \boldsymbol{x}^i, \boldsymbol{w} \rangle + \theta_{y^i - 1}, 0) + \max(1 + \langle \boldsymbol{x}^i, \boldsymbol{w} \rangle - \theta_{y^i}, 0) \Big]\,. \end{aligned}$$

It is not hard to see that objective function $F'(\boldsymbol{w}, \boldsymbol{\theta})$ is a convex approximation of $\frac{\lambda}{2} ||\boldsymbol{w}^2|| + \sum_{i=1}^{m} \left([\![\langle \boldsymbol{x}_i, \boldsymbol{w} \rangle \leq \theta_{y^i - 1}]\!] + [\![\langle \boldsymbol{x}_i, \boldsymbol{w} \rangle \geq \theta_{y^i}]\!] \right)$, i.e., it upper bounds the empirical risk with the 0/1-loss.

We propose an extension of the SVOR-EXP to minimize the interval-insensitive variant of the 0/1-loss by changing the objective function to

$$F_I'(\boldsymbol{w}, \boldsymbol{\theta}) = \frac{\lambda}{2} \|\boldsymbol{w}\|^2 + \sum_{i=1}^{m} \Big(\max(1 - \langle \boldsymbol{x}^i, \boldsymbol{w} \rangle + \theta_{y_l^i - 1}, 0) + \max(1 + \langle \boldsymbol{x}^i, \boldsymbol{w} \rangle - \theta_{y_r^i}, 0) \Big)\,,$$

which is a convex surrogate of the actually desired $\frac{\lambda}{2} \|\boldsymbol{w}\|^2 + \sum_{i=1}^{m} ([\![\langle \boldsymbol{x}_i, \boldsymbol{w} \rangle \leq \theta_{y_l^i - 1}]\!] + [\![\langle \boldsymbol{x}_i, \boldsymbol{w} \rangle \geq \theta_{y_r^i}]\!])$. The interval-insensitive SVOR algorithm with explicit constraints (IIL-SVOR-EXP) for learning from partially annotated examples $\{(\boldsymbol{x}^1, y_l^1, y_r^1), \ldots, (\boldsymbol{x}^m, y_l^m, y_r^m)\} \in (\mathbb{R}^n \times \mathcal{P})^m$ then reads

$$(\boldsymbol{w}^*, \boldsymbol{\theta}^*) = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^n, \boldsymbol{\theta} \in \Theta} F_I'(\boldsymbol{w}, \boldsymbol{\theta})\,,$$

which can be easily rewritten as a convex quadratic program similar to (7).

3.4. Modified SVOR-IMC for minimization of interval-insensitive MAE loss

The original SVOR-IMC algorithm for learning the ordinal classifier (1) from completely annotated examples $\{(\boldsymbol{x}^1, y^1), \ldots, (\boldsymbol{x}^m, y^m)\} \in (\mathbb{R}^n \times \mathcal{Y})^m$ leads to a convex quadratic program Chu and Keerthi (2005)

$$(\boldsymbol{w}^*, \boldsymbol{\theta}^*) = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^n, \boldsymbol{\theta} \in \mathbb{R}^{Y-1}} \left[\frac{\lambda}{2} ||\boldsymbol{w}||^2 + \sum_{j=1}^{r-1} \Big(\sum_{k=1}^{j} \sum_{i=1}^{n^k} \xi_{ki}^j + \sum_{k=j+1}^{r} \sum_{i=1}^{n^k} \xi_{ki}^{*j} \Big) \right] \tag{8}$$

subject to

$$\begin{aligned} \langle \boldsymbol{x}_i^k, \boldsymbol{w} \rangle - \theta_j \leq -1 + \xi_{ki}^j,\ \xi_{ki}^j \geq 0,\ k = 1, \ldots, j,\ i = 1, \ldots, n^k\,, \\ \langle \boldsymbol{x}_i^k, \boldsymbol{w} \rangle - \theta_j \geq 1 - \xi_{ki}^{*j},\ \xi_{ki}^{*j} \geq 0,\quad k = j+1, \ldots, r,\ i = 1, \ldots, n^k\,. \end{aligned}$$

The authors of Chu and Keerthi (2005); Li and Lin (2006) proved that the optimal parameters are admissible, i.e. $(\boldsymbol{w}^*, \boldsymbol{\theta}^*) \in (\mathbb{R}^n, \Theta)$ holds, hence the explicit constraints $\boldsymbol{\theta} \in \Theta$ are not needed in this case. It is also shown that the sum of slack variables in (8) upper bounds the average of the MAE loss $\Delta(y, y') = |y - y'|$ computed on the training examples.

We now derive a modification of the SVOR-IMC algorithm which optimizes the interval-insensitive variant of the MAE loss. Let us first rewrite the problem (8) as an equivalent unconstrained minimization problem

$$(\boldsymbol{w}^*, \boldsymbol{\theta}^*) = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^n, \boldsymbol{\theta} \in \mathbb{R}^{Y-1}} F''(\boldsymbol{w}, \boldsymbol{\theta})$$

with the objective defined as

$$\begin{aligned} F''(\boldsymbol{w}, \boldsymbol{\theta}) \quad = \quad & \frac{\lambda}{2}||\boldsymbol{w}||^2 + \sum_{i=1}^{m} \Big[\sum_{y=1}^{y^i - 1} \max(0, 1 - \langle \boldsymbol{x}^i, \boldsymbol{w} \rangle + \theta_{y-1}) + \\ & \sum_{y=y^i}^{Y-1} \max(0, 1 + \langle \boldsymbol{x}^i, \boldsymbol{w} \rangle - \theta_y) \Big] , \end{aligned}$$

where we use the convention $\sum_{i=m}^{n} a_i = 0$ if $m > n$.

In the case of partially annotate examples $\{(\boldsymbol{x}^1, y_l^1, y_r^1), \ldots, (\boldsymbol{x}^m, y_l^m, y_r^m)\} \in (\mathbb{R}^n \times \mathcal{P})^m$, we propose to change the objective of the SVOR-IMC to

$$\begin{aligned} F_I''(\boldsymbol{w}, \boldsymbol{\theta}) \quad = \quad & \frac{\lambda}{2}||\boldsymbol{w}||^2 + \sum_{i=1}^{m} \Big[\sum_{y=1}^{y_l^i - 1} \max(0, 1 - \langle \boldsymbol{x}^i, \boldsymbol{w} \rangle + \theta_{y-1}) + \\ & \sum_{y=y_r^i}^{Y-1} \max(0, 1 + \langle \boldsymbol{x}^i, \boldsymbol{w} \rangle - \theta_y) \Big] . \end{aligned}$$

We denote the modified algorithm minimizing $F_I''(\boldsymbol{w}, \boldsymbol{\theta})$ as the interval-insensitive SVOR with implicit constraints (IIL-SVOR-IMC).

It is interesting to compare the modified IIL-SVOR-IMC algorithm with the generic algorithm (6) instantiated for the interval-insensitive MAE loss. In particular, the objective function of the generic algorithm, using the same quadratic regularizer as the IIL-SVOR-IMC, reads

$$\begin{aligned} F(\boldsymbol{w}, \boldsymbol{b}) = \tfrac{\lambda}{2}||\boldsymbol{w}^2|| \quad + \quad & \tfrac{1}{m} \sum_{i=1}^{m} \max_{y \le y_l^i} \Big[y_l^i - y + \big\langle \boldsymbol{x}^i, \boldsymbol{w} \big\rangle (y - y_l^i) + b_y - b_{y_l^i} \Big] \\ + \quad & \tfrac{1}{m} \sum_{i=1}^{m} \max_{y \ge y_r^i} \Big[y - y_r^i + \big\langle \boldsymbol{x}^i, \boldsymbol{w} \big\rangle (y - y_r^i) + b_y - b_{y_r^i} \Big] . \end{aligned} \tag{9}$$

We can prove that the generic algorithm uses a tighter upper bound of the interval-insensitive MAE compared to the IIL-SVOR-IMC:

Theorem 3 *Let $(\boldsymbol{w}^*, \boldsymbol{b}^*)$ be a minimizer of $F(\boldsymbol{w}, \boldsymbol{b})$ defined by (9) and let $(\boldsymbol{w}'^*, \boldsymbol{\theta}'^*)$ be a minimizer of $F''_I(\boldsymbol{w}, \boldsymbol{\theta})$ defined by (9). The the inequality*

$$F(\boldsymbol{w}^*, \boldsymbol{b}^*) \leq F''_I(\boldsymbol{w}'^*, \boldsymbol{\theta}'^*)$$

holds true.

PROOF: W.l.o.g we consider the generic algorithm and the IIL-SVOR-IMC in the case of a single training example. Then the corresponding objectives read $F(\boldsymbol{w}, \boldsymbol{b}) = \frac{\lambda}{2}||\boldsymbol{w}^2|| + \max_{y \leq y_l} \left[y_l - y + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_l) + b_y - b_{y_l} \right] + \max_{y \geq y_r} \left[y - y_r + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_r) + b_y - b_{y_r} \right]$ and $F'(\boldsymbol{w}, \boldsymbol{\theta}) = \frac{\lambda}{2}||\boldsymbol{w}^2|| + \sum_{y=1}^{y_l - 1} \max(0, 1 - \langle \boldsymbol{x}, \boldsymbol{w} \rangle + \theta_{y-1}) + \sum_{y=y_r}^{Y-1} \max(0, 1 + \langle \boldsymbol{x}, \boldsymbol{w} \rangle - \theta_y)$. Let $(\boldsymbol{w}^*, \boldsymbol{b}^*)$ be a minimizer of $F(\boldsymbol{w}, \boldsymbol{b})$ and $(\boldsymbol{w}'^*, \boldsymbol{\theta}'^*)$ be a minimizer of $F'(\boldsymbol{w}, \boldsymbol{\theta})$. Let us further introduce a function $G(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{\theta}) = F(\boldsymbol{w}, \boldsymbol{b}) + \infty \cdot [\![\theta_y = b_y - b_{y+1}, \forall y = 1, \ldots, Y-1]\!]$ and denote its minimizer by $(\boldsymbol{w}^*_G, \boldsymbol{b}^*_G, \boldsymbol{\theta}^*_G)$. For a fixed $(\boldsymbol{w}, \boldsymbol{\theta})$ we say that $\boldsymbol{b}$ is admissible if $G(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{\theta}) < \infty$. By construction of $G(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{\theta})$ we have $G(\boldsymbol{w}^*_G, \boldsymbol{b}^*_G, \boldsymbol{\theta}^*_G) = f(\boldsymbol{w}^*, \boldsymbol{b}^*)$ and $G(\boldsymbol{w}^*_G, \boldsymbol{b}^*_G, \boldsymbol{\theta}^*_G) \leq \min_{\boldsymbol{w}, \boldsymbol{b}} G(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{\theta}'^*)$. Since for fixed $\boldsymbol{\theta}$ the value of $G(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{\theta})$ does not depend on any admissible $\boldsymbol{b}$, we see that $\min_{\boldsymbol{w}, \boldsymbol{b}} G(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{\theta}'^*) \leq H(\boldsymbol{w}', \boldsymbol{\theta}'^*), \forall \boldsymbol{w}'$, where we used

$H(\boldsymbol{w}, \boldsymbol{\theta}) = \frac{\lambda}{2}||\boldsymbol{w}^2|| + \max_{y < y_l} \left[\max(0, y_l - y + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_l) + \sum_{y'=y}^{y_l - 1} \theta_{y'}) \right] +$

$\max_{y > y_r} \left[\max(0, y - y_r + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_r) - \sum_{y'=y_r}^{y-1} \theta_{y'}) \right]$. Since $\max_{y < y_l} \left[\max(0, y_l - y + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_l) + \sum_{y'=y}^{y_l - 1} \theta_{y'}) \right] \leq \max_{y < y_l} \left[\sum_{\hat{y}=y}^{y_l - 1} \max(0, 1 - \langle \boldsymbol{x}, \boldsymbol{w} \rangle + \theta_{\hat{y}}) \right] = \sum_{\hat{y}=1}^{y_l - 1} \max(0, 1 - \langle \boldsymbol{x}, \boldsymbol{w} \rangle + \theta_{\hat{y}})$ and $\max_{y > y_r} \left[\max(0, y - y_r + \langle \boldsymbol{x}, \boldsymbol{w} \rangle (y - y_r) - \sum_{y'=y_r}^{y-1} \theta_{y'}) \right] \leq \sum_{\hat{y}=y_r}^{Y-1} \max(0, 1 + \langle \boldsymbol{x}, \boldsymbol{w} \rangle - \theta_{\hat{y}})$ we get $H(\boldsymbol{w}, \boldsymbol{\theta}'^*) \leq F'(\boldsymbol{w}, \boldsymbol{\theta}'^*), \forall \boldsymbol{w}$. Hence, we have $F(\boldsymbol{w}^*, \boldsymbol{b}^*) = G(\boldsymbol{w}^*_G, \boldsymbol{b}^*_G, \boldsymbol{\theta}^*_G) \leq \min_{\boldsymbol{w}, \boldsymbol{b}} G(\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{\theta}^*_G) \leq H(\boldsymbol{w}^*_G, \boldsymbol{\theta}^*_G) \leq F'(\boldsymbol{w}'^*, \boldsymbol{\theta}'^*), \ \forall \boldsymbol{w}$, and, therefore, $F(\boldsymbol{w}^*, \boldsymbol{b}^*) \leq F'(\boldsymbol{w}'^*, \boldsymbol{\theta}'^*)$. ■

4. Experiments

We evaluate the proposed algorithms on i) a set of standard benchmarks that have been previously used for benchmarking ordinal classification methods and ii) on a real-life application of predicting age from a photograph. The experiments on the standard benchmarks are reported in Section 4.1 and the real-life problem is described in Section 4.2.

4.1. Standard benchmarks

We perform experiments on six datasets previsously used to benchmarks ordinal regression algorithms (Li and Lin (2006); Chu and Keerthi (2005)). The data were produced by discretising metric regression problems into $Y = 10$ bins. The data are randomly partitioned to training/test part. The partitioning was repeated 20 times. The features are normalized

	CPU	BOSTON	ABALONE	BANK	COMPUTER	CALIFORNIA
dimension	27	13	8	32	21	8
train	50	300	1,000	3,000	4,000	5000
test	24	206	3,177	5,192	4,192	15,640

Table 1: Table shows the number of dimensions, number of training and testing examples for the six datasets used in the experiments.

to zero mean and unit variance coordinate wise. Table 1 summarizes dimensions of the datasets.

We compare the original SVOR-IMC algorithm learning from completely annotated examples with the two proposed algorithms learning from partial annotation. Namely, we consider an instance of the generic algorithm optimizing the MAE loss, denoted as IIL-Generic-MAE, and the modification of the SVOR-IMC denoted as IIL-SVOR-IMC. All compared methods, SVOR-IMC, ILL-Generic-MAE and ILL-SVOR-IMC optimize different surrogates of the MAE loss.

To measure benefit of learning from partially annotated examples we used the following protocol. Let m_{max} denote the maximal number of examples in the training split (different for each data set). We trained IIL-SVOR-IMC and IIL-Generic-MAE from $m = m_{\text{comp}} + m_{\text{part}}$ examples where we fixed the number of completely annotated examples to $m_{\text{comp}} = 0.1m_{\text{max}}$ and we varied the number of partially annotated examples m_{part} from 0 to $0.9m_{\text{max}}$. We consider two case of partial annotations:

Case 1: the partial annotation was set randomly with uniform distribution to be one of the following intervals $[y-1, y], [y, y+1], [y-1, y+1]$ where y is the true label.

Case 2: the partial annotation was set randomly with uniform distribution to be one of the following intervals $[y-1, y], [y, y+1], [y-1, y+1], [y-2, y], [y, y+2], [y-2, y+2], [y-2, y+1], [y-1, y+2]$ where y is the true label.

In the case of the original SVOR-IMC we just varied the number of fully annotated examples m. The regularization constants were found by cross-validation on the training splits. The reported results are averages and standard deviations computed over the 20 random splits.

Figures 1 and 2 show results for the case 1 and 2, respectively. The x-axis denotes the varied number of training examples m and the y-axis corresponding estimate of the (complete) MAE loss on test examples.

The results show that adding the partially annotated examples steadily decreases the test MAE of the classifiers learned by ILL-Generic-MAE and ILL-SVOR-IMC. The classifier learned by the SVOR-IMC from complete annotations has consistently lower test MAE compared to the IIL-SVOR-IMC and IIL-Generic-MAE using the partial annotations. Both surrogates of the interval-insensitive loss, i.e. the IIL-Generic-MAE and the IIL-SVOR-IMC, work equally well in terms of the accuracy. The gain from using the complete annotation is relatively low especially in the case 1. Not surprisingly the gap between complete and partial algorithms increases with the increased uncertainty in the partial annotation (compare the cases 1 and 2).

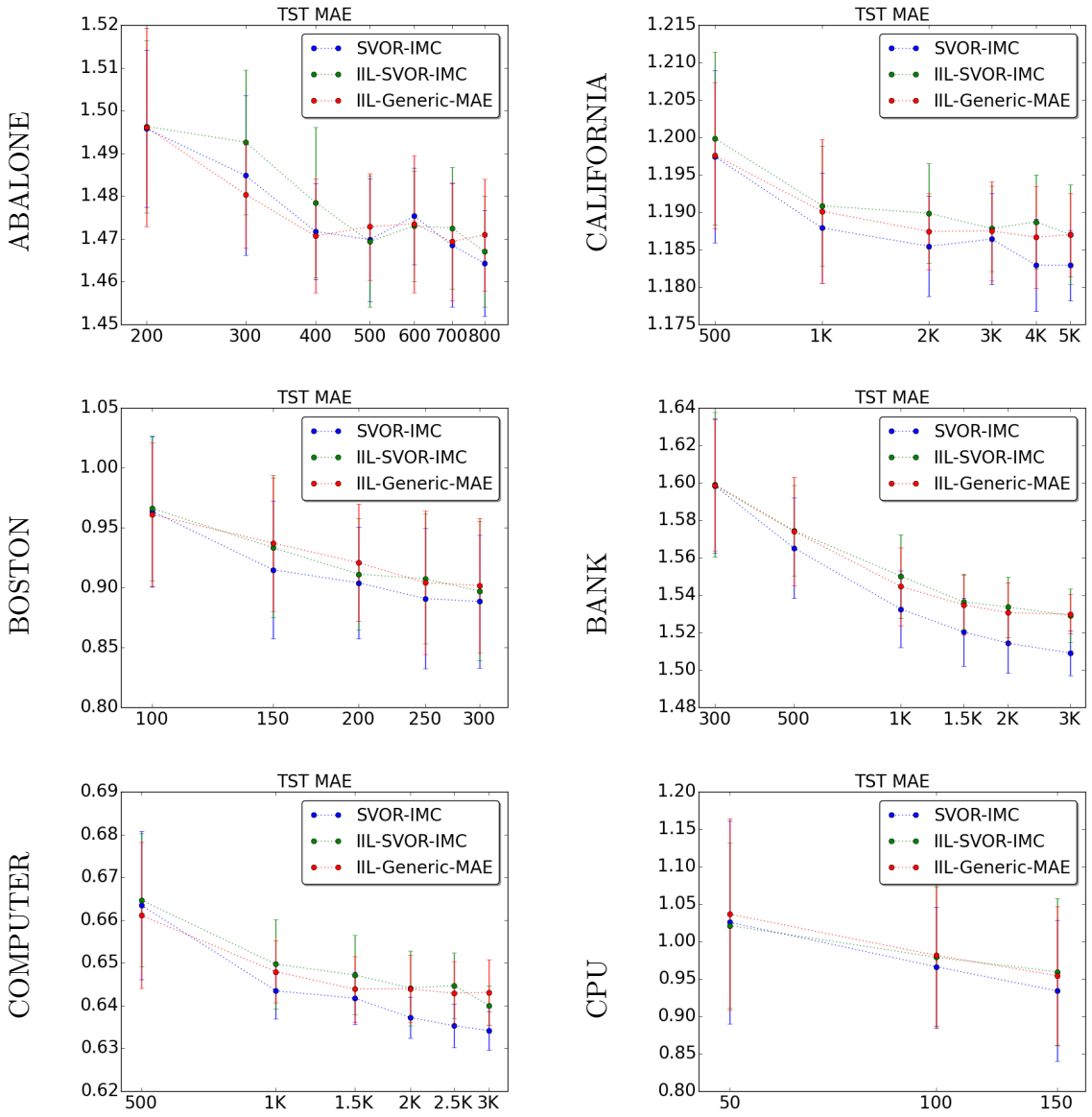

Figure 1: Experiment with partially annotated examples – Case 1 (see text for description). The x-axis corresponds to the number of partially (in case of IIL-SVOR-IMC and IIL-Generic-MAE) and completely (in case of SVOR-IMC) annotated exaples. The y-axis corresponds to the test estimate of the MAE loss. We show means (line) and standard deviations (error bars) of the mean absolute error (MAE) for a corresponding dataset.

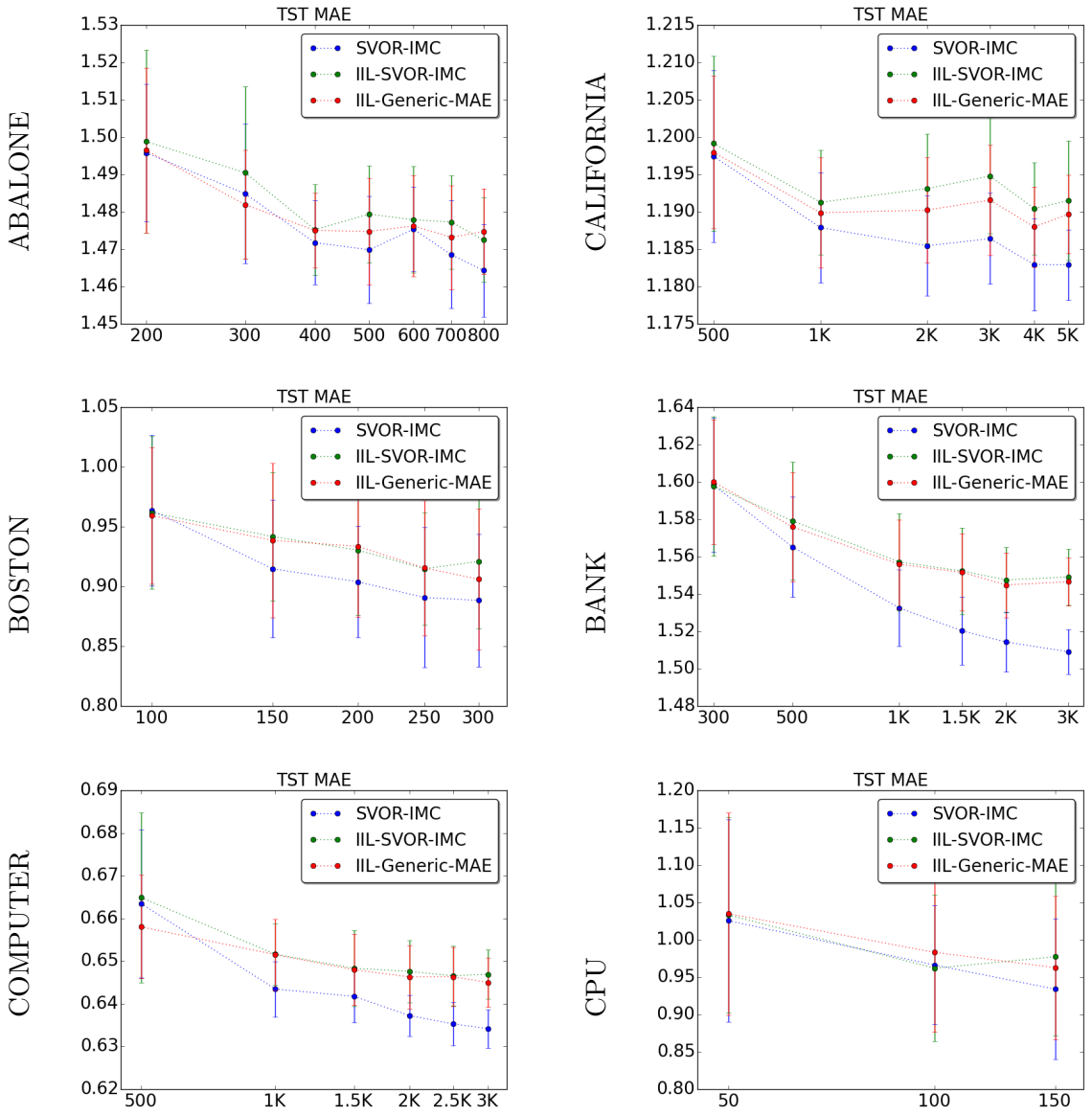

Figure 2: Experiment with partially annotated examples – Case 2 (see text for description).

4.2. Visual age estimation

We consider problem of estimating an apparent age of a person from an image of his/her face. We use the MORPH database (Ricanek and Tesafaye (2006)) which contains 55,134 face images with exact age annotation ranging from 16 to 77 years [1]. The faces were first localized by AdaBoost based face detector and consequently we found facial landmarks by

1. Because the age category 70+ is severely under-represented (only 9 examples in total) we removed faces with age higher than 70.

SVOR-IMC	MAE	6.23 ± 0.04	6.27 ± 0.08	6.21 ± 0.04	5.98 ± 0.05
	comp+part	$10\% + 0\%$	$40\% + 0\%$	$70\% + 0\%$	$100\% + 0\%$
IIL-SVOR-IMC	MAE	6.31 ± 0.10	6.38 ± 0.04	6.25 ± 0.03	6.03 ± 0.02
	comp+part	$10\% + 0\%$	$10\% + 30\%$	$10\% + 60\%$	$10\% + 90\%$

Table 2: The estimation error of the ordinal classification based visual age predictor trained the MORPH database by the SVOR-IMC and the proposed IIL-SVOR-IMC. For each method, the first line shows that mean and the standard deviation of the MAE. The second line shows the percentage of completely and partially annotated examples used for training.

a Deformable Part Model based detector (Uricar et al. (2012)). The detected landmarks are used to transform the input face into a canonical pose by an affine transform. Finally, the canonical image is described by a pyramid-of-LBP descriptors (Sonnenburg and Franc (2010)) which is $n = 159,488$-dimensional binary sparse vector serving as an input of the ordinal classifier. The data were randomly split 3 times into training/validation/test part in the ratio 60/20/20. The validation part was used to tune the regularization constant λ. The reported results are averages and standard deviations computed of the error measures computed over the 3 splits.

Because the IIL-SVOR-IMC and IIL-Generic-MAE have been shown to give comparable results (c.f. the previous section) here we test only the former surrogate loss. To measure benefit of learning from partially annotated examples we used the same protocol as described in Section 4.1. I.e. we trained the proposed IIL-SVOR-IMC from $m = m_{\text{comp}} + m_{\text{part}}$ examples where the number of completely annotated examples was fixed to $m_{\text{comp}} = 0.1 m_{\max}$ and the amount of the partially annotated examples m_{part} was varied from 0 to $0.9 m_{\max}$ ($m_{\max}$ denotes the total number of training examples). The partial annotation was generated by rounding the exact age into 5-year intervals, i.e. interval annotation is from $\{[15, 20], [21, 25], \ldots, [61, 70]\}$. In the case of SVOR-IMC we just varied the total number of fully annotated examples.

The results are summarized in Table 2. It is seen that the proposed IIL-SVOR-IMC achieves almost the same accuracy as the SVOR-IMC for all used amounts of training examples. In particular, when using all training examples the MAE of the SVOR-IMC is 5.98 ± 0.05 compared to 6.03 ± 0.02 of the IIL-SVOR-IMC. Note, that the IIL-SVOR-IMC is trained only from 10% of fully annotated examples while 90% of had the cheaper interval annotation. We point out that the state-of-the-art approach using the SVM based ordinal classifier trained from the completely annotated examples achieves MAE of 6.07% on the very same database (Chang et al. (2011)).

5. Conclusions

We have proposed an interval-insensitive loss function suitable for risk minimization based learning of ordinal classifiers from partially annotated examples. We derived a generic algorithm optimizing a convex surrogate of the proposed interval-insensitive loss. We also derived other surrogate losses for learning from partial annotations by extending the ex-

isting state-of-the-art SVOR-EXP and SVOR-IMC methods. We provided theoretical and experimental comparison of the proposed methods. The experiments conducted on standard benchmarks and a real-life problem of visual age estimation show that learning ordinal classifiers from partially annotated examples is competitive with the so-far used methods requiring completely annotated examples.

Acknowledgments

AK was supported by Technology Agency of the Czech Republic under Project TE01020197. VF was supported by the project ERC-CZ LL1303. VH was supported EU FP7 project 609763.

Appendix A

PROOF: of Theorem 3. W.l.o.g we consider the generic algorithm and the IIL-SVOR-IMC in the case of a single training example. Then the corresponding objectives read

$$F(\boldsymbol{w},\boldsymbol{b}) = \frac{\lambda}{2}||\boldsymbol{w}^2|| + \max_{y\leq y_l}\left[y_l - y + \langle \boldsymbol{x},\boldsymbol{w}\rangle(y-y_l) + b_y - b_{y_l}\right] + \max_{y\geq y_r}\left[y - y_r + \langle \boldsymbol{x},\boldsymbol{w}\rangle(y-y_r) + b_y - b_{y_r}\right]$$

and

$$F'(\boldsymbol{w},\boldsymbol{\theta}) = \frac{\lambda}{2}||\boldsymbol{w}^2|| + \sum_{y=1}^{y_l-1}\max(0, 1-\langle \boldsymbol{x},\boldsymbol{w}\rangle + \theta_{y-1}) + \sum_{y=y_r}^{Y-1}\max(0, 1+\langle \boldsymbol{x},\boldsymbol{w}\rangle - \theta_y)\,.$$

Let $(\boldsymbol{w}^*,\boldsymbol{b}^*)$ be a minimizer of $F(\boldsymbol{w},\boldsymbol{b})$ and $(\boldsymbol{w}'^*,\boldsymbol{\theta}'^*)$ be a minimizer of $F'(\boldsymbol{w},\boldsymbol{\theta})$. Let us further introduce a function $G(\boldsymbol{w},\boldsymbol{b},\boldsymbol{\theta}) = F(\boldsymbol{w},\boldsymbol{b}) + \infty\cdot[\![\theta_y = b_y - b_{y+1}\,, \forall y = 1,\ldots,Y-1]\!]$ and denote its minimizer by $(\boldsymbol{w}^*_G,\boldsymbol{b}^*_G,\boldsymbol{\theta}^*_G)$. For a fixed $(\boldsymbol{w},\boldsymbol{\theta})$ we say that $\boldsymbol{b}$ is admissible if $G(\boldsymbol{w},\boldsymbol{b},\boldsymbol{\theta}) < \infty$. By construction of $G(\boldsymbol{w},\boldsymbol{b},\boldsymbol{\theta})$ we have $G(\boldsymbol{w}^*_G,\boldsymbol{b}^*_G,\boldsymbol{\theta}^*_G) = f(\boldsymbol{w}^*,\boldsymbol{b}^*)$ and $G(\boldsymbol{w}^*_G,\boldsymbol{b}^*_G,\boldsymbol{\theta}^*_G) \leq \min_{\boldsymbol{w},\boldsymbol{b}} G(\boldsymbol{w},\boldsymbol{b},\boldsymbol{\theta}'^*)$. Since for fixed $\boldsymbol{\theta}$ the value of $G(\boldsymbol{w},\boldsymbol{b},\boldsymbol{\theta})$ does not depend on any admissible $\boldsymbol{b}$, we see that $\min_{\boldsymbol{w},\boldsymbol{b}} G(\boldsymbol{w},\boldsymbol{b},\boldsymbol{\theta}'^*) \leq H(\boldsymbol{w}',\boldsymbol{\theta}'^*), \forall \boldsymbol{w}'$, where we used

$$\begin{aligned} H(\boldsymbol{w},\boldsymbol{\theta}) \;=\; & \tfrac{\lambda}{2}||\boldsymbol{w}^2|| + \max_{y<y_l}\Big[\max(0, y_l - y + \langle \boldsymbol{x},\boldsymbol{w}\rangle(y-y_l) \\ & + \sum_{y'=y}^{y_l-1}\theta_{y'})\Big] + \max_{y>y_r}\Big[\max\Big(0, y - y_r + \langle \boldsymbol{x},\boldsymbol{w}\rangle(y-y_r) - \sum_{y'=y_r}^{y-1}\theta_{y'}\Big)\Big]\,. \end{aligned}$$

Since $\max_{y<y_l}\big[\max(0, y_l - y + \langle \boldsymbol{x},\boldsymbol{w}\rangle(y-y_l) + \sum_{y'=y}^{y_l-1}\theta_{y'})\big] \leq \max_{y<y_l}\big[\sum_{\hat{y}=y}^{y_l-1}\max(0, 1-\langle \boldsymbol{x},\boldsymbol{w}\rangle + \theta_{\hat{y}})\big] = \sum_{\hat{y}=1}^{y_l-1}\max(0, 1-\langle \boldsymbol{x},\boldsymbol{w}\rangle + \theta_{\hat{y}})$ and $\max_{y>y_r}\big[\max(0, y - y_r + \langle \boldsymbol{x},\boldsymbol{w}\rangle(y-y_r) - \sum_{y'=y_r}^{y-1}\theta_{y'})\big] \leq \sum_{\hat{y}=y_r}^{Y-1}\max(0, 1+\langle \boldsymbol{x},\boldsymbol{w}\rangle + \theta_{\hat{y}})$ we get $H(\boldsymbol{w},\boldsymbol{\theta}'^*) \leq F'(\boldsymbol{w},\boldsymbol{\theta}'^*),\ \forall \boldsymbol{w}$. Hence, we have $F(\boldsymbol{w}^*,\boldsymbol{b}^*) = G(\boldsymbol{w}^*_G,\boldsymbol{b}^*_G,\boldsymbol{\theta}^*_G) \leq \min_{\boldsymbol{w},\boldsymbol{b}} G(\boldsymbol{w},\boldsymbol{b},\boldsymbol{\theta}^*_G) \leq H(\boldsymbol{w}^*_G,\boldsymbol{\theta}^*_G) \leq F'(\boldsymbol{w}'^*,\boldsymbol{\theta}'^*),\ \forall \boldsymbol{w}$, and, therefore, $F(\boldsymbol{w}^*,\boldsymbol{b}^*) \leq F'(\boldsymbol{w}'^*,\boldsymbol{\theta}'^*)$. ■

References

K. Antoniuk, V. Franc, and V. Hlavac. Mord: Multi-class classifier for ordinal regression. In *Machine Learning and Knowledge Discovery in Databases - ECML PKDD*, pages 96–111, 2013.

K. Chang, C. Chen, and Y. Hung. Ordinal hyperplane ranker with cost sensitivities for age estimation. In *Computer Vision and Pattern Recognition*, pages 585–592, 2011.

W. Chu and Z. Ghahramani. Preference learning with gaussian processes. In *International Conference on Machine Learning*, pages 137–144, 2005.

W. Chu and S. S. Keerthi. New approaches to support vector ordinal regression. In *International Conference on Machine Learning (ICML)*, pages 145–152, 2005.

T. Cour, B. Sapp, and B. Taskar. Learning from partial labels. *Journal of Machine Learning Research*, 12:1225–1261, 2011.

K. Crammer and Y. Singer. Pranking with ranking. In *Advances in Neural Information Processing Systems (NIPS)*, pages 641–647. MIT Press, 2001.

K. Debczynski, W. Kotlowski, and R. Slowinski. Ordinal classification with decision rules. In *Mining Complex Data, LNCS*, volume 4944, pages 169–181, 2008.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 1(39), 1997.

T. Do and T. Artieres. Large margin training for hidden markov models with partially observed states. In *International Conference on Machine Learning (ICML)*, pages 265–272, 2009.

L. Fu and D.G. Simpson. Conditional risk models for ordinal response data: simultaneous logistic regression analysis and generalized score test. *Journal of Statistical Planning and Inference*, pages 201–217, 2002.

L. Jie and F. Orabona. Learning from candidate labeling sets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1504–1512, 2010.

W. Kotlowski, K. Dembczynski, S. Greco, and R. Slowinski. Stochastic dominance-based rough set model for ordinal classification. *Journal of Information Sciences*, 178(21): 4019–4037, 2008.

L. Li and H. Lin. Ordinal regression by extended binary classification. In *Advances in Neural Information Processing Systems (NIPS)*, pages 865–872, 2006.

X. Lou and F. Hamprecht. Structured learning from partial annotations. In *International Conference on Machine Learning (ICML)*, 2012.

P. McCullagh. Regression models for ordinal data. *Journal of the Royal Statical Society*, 42(2):109–142, 1980.

N. Ramanathan and R. Chellappa. Computational methods for modeling facial aging: A survey. *Journal of Visual Languages and Computing*, pages 131–144, 2009.

J.D.M. Rennie and N. Srebro. Loss functions for preference levels: Regression with discrete ordered labels. In *Proc. of the IJCAI Multidisciplinary Workshop on Advances in Preference Handling*, pages 180–186, 2005.

K. Ricanek and T. Tesafaye. Morph: A longitudial image database of normal adult age-progression. In *In proc. of Automated Face and Gesture Recognition*, pages 341–345, 2006.

A. Shashua and A. Levin. Ranking with large margin principle: Two approaches. In *In Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 937–944, 2002.

S. Sonnenburg and V. Franc. Coffin: A computational framework for linear svms. In *International Conference on Machine Learning*, pages 999–1006, Madison, USA, June 2010. Omnipress.

I. Tsochantaridis, T. Joachims, T. Hofmann, Y. Altun, and Y. Singer. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005.

M. Uricar, V. Franc, and V. Hlavac. Detector of facial landmarks learned by the structured output SVM. In Gabriela Csurka and José Braz, editors, *International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 1, pages 547–556, 2012.

V. Vapnik. *Statistical learning theory*. Adaptive and Learning Systems. New York, USA, 1998.