Ordinal Label Proportions

Rafael Poyiadzi[1(✉)], Raúl Santos-Rodríguez[1], and Tijl De Bie[2]

[1] Department of Engineering Mathematics, University of Bristol,
Bristol, UK
{rp13102,enrsr}@bristol.ac.uk

[2] Department of Electronics and Information Systems, IDLab,
Ghent University, Ghent, Belgium
Tijl.DeBie@ugent.be

Abstract. In Machine Learning, it is common to distinguish different degrees of supervision, ranging from fully supervised to completely unsupervised scenarios. However, lying in between those, the Learning from Label Proportions (LLP) setting [19] assumes the training data is provided in the form of bags, and the only supervision comes through the proportion of each class in each bag. In this paper, we present a novel version of the LLP paradigm where the relationship among the classes is ordinal. While this is a highly relevant scenario (e.g. customer surveys where the results can be divided into various degrees of satisfaction), it is as yet unexplored in the literature. We refer to this setting as Ordinal Label Proportions (OLP). We formally define the scenario and introduce an efficient algorithm to tackle it. We test our algorithm on synthetic and benchmark datasets. Additionally, we present a case study examining a dataset gathered from the Research Excellence Framework that assesses the quality of research in the United Kingdom.

Keywords: Label Proportions · Ordinal classification
Discriminant learning

1 Introduction

According to the nature of their output, the two dominating tasks in Machine Learning are those of regression and classification. Attracting an increasing interest, Ordinal Classification (also termed Ordinal Regression) [2,5,6,12] falls somewhere in between the two. Similarly to multiclass classification tasks, the practitioner is provided with a set of data points with their corresponding labels coming from a discrete set $\mathcal{C} = \{r_1, \cdots, r_k\}$, but opposed to its nominal sibling, in ordinal classification, the labels exhibit a natural ordering: $r_1 \prec r_2 \prec \cdots \prec r_k$. There is an abundance of examples, ranging from categorizing responses to questions such as "how much do you like Greek food?" to movie ratings or grade prediction. The difference between the labels in these tasks, e.g. {*very bad, bad, good, great, excellent*} and the standard categorical labels, e.g. {*car, pedestrian, bicycle*}, is clear.

M. Berlingerio et al. (Eds.): ECML PKDD 2018, LNAI 11051, pp. 306–321, 2019.
https://doi.org/10.1007/978-3-030-10925-7_19

Let us consider opinion polls where acceptable outcomes range from strongly agree to strongly disagree on a given topic. There is a clear ordinal relationship among the outcomes. However, for privacy reasons, it is often not possible to publish each individual's opinion. On the other hand, it may be possible to aggregate the results over several demographic subsets (e.g., by region). Therefore, this is data that naturally comes in the forms of bags and where, although the ground truth labels might not be available, we might have access to the proportion of each class in each bag. We can also argue that, in many cases, individuals' opinions are not as relevant as an accurate prediction of the bag proportions. A specific example of such tasks is that of rating research papers according to quality, mapping each of them to a set of predefined categories. In the United Kingdom, publicly funded research is regularly assessed under the Research Excellence Framework (REF)[1]. In order to preserve anonymity, based on the papers submitted by the different research units, REF provides a histogram stating how many of the papers submitted were placed in each category, *without revealing which specific paper was in each of these classes*. As before, individual paper ratings are sensitive, but aggregates per submission are fine to publish. Importantly, funding levels are then based on these histograms. A difficulty for universities is that REF does not rely on a public and formal procedure to classify papers, but on the judgment of a panel. Therefore, although this can be cast as an ordinal classification task, unfortunately, the ground truth labels are not available.

As an aggregate supervision is accessible through the histograms, this problem sits in between fully supervised and unsupervised learning. In the non-ordinal case, this has been studied under the name of *learning from label proportions* [9,14,15,20], where the data is assumed to be given in the form of bags (e.g., research units) and only the proportion of each class in each bag is given (histograms). Up to the authors' knowledge, learning a classifier with this level of supervision, in the ordinal setting, has not yet been explored. We call this setting Ordinal Label Proportions (OLP). The OLP methodology developed in this work is able to efficiently make use of the ordinal constrains in order to learn from the past REF data and unveil how the expert panel operated by inferring a scoring function for papers as a function of various properties, such as journal and conference rankings, citation half life, impact factor, number of authors, or Google scholar citation count at time of submission.

The contributions of this paper are threefold. Firstly, we introduce and rigorously define the new learning paradigm of learning from Ordinal Label Proportions. Secondly, we present an efficient algorithm for learning a classifier in this setting based on discriminant learning. Thirdly, we produce a dataset for REF and present our analysis.

The paper is structured as follows. In Sect. 2 we review the related work. In Sect. 3 we introduce the basic formulation of Linear Discriminant Analysis (LDA) in the ordinal setting and show how it can be adopted to be trained

[1] https://www.ref.ac.uk/.

with label proportions. In Sect. 4 we respectively present the empirical analysis in both real and synthetic datasets. Section 5 is devoted to the conclusions.

1.1 Problem Formulation

We assume that we have access to a set of observations $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ where $\mathbf{x}_i \in \mathbb{R}^d$. The true labels, $\mathbf{y} = \{y_1, \ldots, y_n\}$ with y_i being the label of observation $\mathbf{x}_i$, also exist but are hidden. Also, the class labels $\{r_1, \cdots, r_k\}$ have a natural order, i.e. $r_1 \prec r_2 \prec \ldots \prec r_k$. The set $\mathbf{X}$ is separated into distinct bags $\mathbf{X} = \bigcup_{k=1}^{K} \mathbf{B}_k$, where each $\mathbf{B}_k$ corresponds to the subset of points assigned to the k-th bag, and $\mathbf{B}_k \cap \mathbf{B}_j = \emptyset, \forall k, j \in [K]$. Moreover, for each bag $\mathbf{B}_k$ we have access to its class proportions, $\boldsymbol{\pi}_k = \{\pi_{k,1}, \ldots, \pi_{k,c}\}$, where $\sum_{h=1}^{c} \pi_{k,h} = 1$, $\pi_{k,h} \geq 0$, with $\pi_{k,h}$ corresponding to the proportion of class h in bag $\mathbf{B}_k$ and c being the number of classes ($c = 2$ being the binary classification setting).

The OLP task is then cast as minimizing the following objective:

$$\begin{gathered} \boldsymbol{d}(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}[\boldsymbol{s}]) + \lambda \boldsymbol{R}[\boldsymbol{s}] \\ s.t.\ \boldsymbol{s}(\boldsymbol{x}_i) \geq \boldsymbol{s}(\boldsymbol{x}_j), \forall i, j \in \boldsymbol{C} \end{gathered} \tag{1}$$

where $\hat{\boldsymbol{\pi}}$ and $\boldsymbol{R}$ are functionals of a scoring function $\boldsymbol{s}(.)$. The former is the estimate of the bag proportions, while the later acts as a regularizer, and λ controls the strength of the penalty term. $\boldsymbol{C}$ is the set of all pairwise ordinal relationships that should hold, i.e. $\boldsymbol{s}(\boldsymbol{x}_i) \geq \boldsymbol{s}(\boldsymbol{x}_j)$ for $y_i = r_c$ and $y_j = r_h$, with $r_c \succ r_h$. The functional $\boldsymbol{d}(.,.)$ provides a measure of distance between the true and estimated proportions.

2 Related Work

In this section we review related work for both ordinal classification and learning from label proportions.

2.1 Ordinal Classification

For a paper length discussion of the approaches to ordinal classification and their taxonomy we refer the reader to [6] and the references therein. Here, we briefly outline the main approaches.

The assumption of a natural ordering of the labels, which underlies ordinal classification is a strong one, as it states that the ordinal structure of the labels is also present in the feature space, or as stated in [8] "the ordinal class structure induces an ordinal instance structure". One could of course reduce the ordinal classification problem to a nominal one and make use of plenty of existing algorithms, but this would amount to ignoring available information about the structure of the data, that could otherwise be used to improve performance, reduce computation and in general help in building a more consistent classifier. On the other hand, the task could also be transformed to a regression problem by mapping the class labels onto the real line (while respecting the ordering) and then proceed by applying standard regression techniques. A technique in this category is that of training a regression tree [2]. However, one disadvantage of this method is that there is no principled way of choosing the map [6]. In the taxonomy of ordinal classification approaches, presented in [6], this falls under the *naive approaches* category, as they are basically the result of other standard techniques in machine learning. An alternative approach, still naive but more advanced, is that of cost-sensitive classification, where the order of the classes is taken into account in the sense that not all mistakes carry equal weight [16].

The second group of techniques is referred to as *Ordinal Binary Decompositions* (OBD). In most cases multiclass classification is tackled through the use of one-vs-one or one-vs-all voting schemes. In the OBD group, some of the schemes used are one-vs-next, one-vs-followers and one-vs-previous, which clearly make explicit the ordering of the classes (consult [6] for a longer discussion of these schemes and for their properties). One such models is presented in [5], where the original problem is decomposed into a series of binary classification tasks.

The third and final group includes the *threshold models*, which are based on the assumption of a latent continuous variable underlying the ordered discrete labels [6]. These methods have two main ingredients; a function trained to estimate the latent variable and a set of thresholds that distinguish between the classes (in the ordered setting). The reader would be right in noting the similarity of these models with the naive regression approaches. The difference between the two categories is that in the threshold models, there is no mapping from discrete (ordered) labels onto the real line (which, as previously discussed would require prior knowledge about the distances of the classes), but rather thresholds are being used, which are learned during training.

One of the first attempts was the proportional odds model [12], which extends logistic regression to the ordinal setting. The Support Vector Machine is one of the most eminent machine learning techniques due to its generalization performance, and has therefore inevitably seen many adaptations to the ordinal classification setting [2,7]. Finally, and also belonging to the category of threshold models, discriminative learning [17] will be discussed in Sect. 3.

2.2 Learning from Label Proportions

The level of supervision of bag proportions is very similar to the one of Multiple-Instance Learning [4,11], where the practitioner is provided with logical statements indicating the presence of a class in a bag. For example, in binary classification, a bag would have a positive label if it had at least one positive point in it, while it would be labeled as negative if all of the points belonging to it were negative.

Existing algorithms designed for learning from label proportions fall in three main categories. Bayesian approaches such as [9] approach the problem by generating labels consistent with bag proportions. In [14] the authors propose an algorithm that relies on the properties of exponential families and the convergence of the class mean operator, computed from the means and label proportions of each bag. Lastly, maximum-margin approaches [15,20] pose the problem as either an extension of maximum-margin clustering [18] or Support Vector Regression.

Conditional Exponential Families. The notation, as well as the overall treatment, in this section follow from [14]. For further clarification please consult the original paper. Let $\mathcal{X}$ and $\mathcal{Y}$ denote the space of the observations and the (discrete) label space respectively, and let $\boldsymbol{\phi}(\boldsymbol{x}, y) : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{H}$ be a feature map into a Reproducing Kernel Hilbert Space $\mathcal{H}$ with kernel $k((\mathbf{x}, y), (\mathbf{x}', y'))$. A conditional exponential family is stated as follows:

$$p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = exp\big(\boldsymbol{\phi}(\boldsymbol{x}, y)^T \boldsymbol{\theta} - g(\boldsymbol{\theta}|\boldsymbol{x})\big) \quad \text{with}$$
$$g(\boldsymbol{\theta}|\boldsymbol{x}) = log \sum_{y \in \mathcal{Y}} exp\big(\boldsymbol{\phi}(\boldsymbol{x}, y)^T, \boldsymbol{\theta}\big)$$

where $g(\boldsymbol{\theta}|\boldsymbol{x})$ is a log-partition function and $\boldsymbol{\theta}$ is the parameter of the distribution. Under the assumption that $\{(\boldsymbol{x}_i, y_i)_{i=1}^n\}$ are drawn independently and identically distributed by the distribution $p(x, y)$, one usually optimizes for $\boldsymbol{\theta}$ by minimizing the regularized negative conditional log-likelihood:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{n} [g(\boldsymbol{\theta}|\boldsymbol{x}_i)] - n\boldsymbol{\mu}_{XY}^T \boldsymbol{\theta} + \lambda ||\boldsymbol{\theta}||^2 \right\},$$

where $\boldsymbol{\mu}_{XY} := \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\phi}(\boldsymbol{x}_i, y_i)$. Unfortunately, in the LLP setting we cannot compute this quantity directly, as the labels are unknown.

MeanMap. In [14] the authors build upon conditional exponential families and present MeanMap, which exploits the theoretical guarantees of uniform convergence of the expectation operator to its expected value, $\mu_{xy} := \boldsymbol{E}_{(x,y) \sim p(x,y)}[\boldsymbol{\phi}(\boldsymbol{x}, y)]$. Expanding we get:

$$\mu_{xy} = \sum_{y \in \mathcal{Y}} p(y) \boldsymbol{E}_{x \sim p(x|y)}[\boldsymbol{\phi}(\boldsymbol{x}, y)] \tag{2}$$

A critical assumption is that conditioned on its label, a point is independent of its bag assignment, that is, $p(x|y,i) = p(x|y)$. Based on this we get $p(x|i) = \sum_y p(x|y)\pi_{iy}$ and subsequently

$$\mu_x^{set}[i, y'] = \boldsymbol{E}_{x \sim p(x|i)}[\boldsymbol{\phi}(\boldsymbol{x}, y')] = \sum_y \pi_{iy} \boldsymbol{E}_{x \sim p(x|y)}[\boldsymbol{\phi}(\boldsymbol{x}, y)]$$
$$= \sum_y \pi_{iy} \mu_x^{class}[y']$$

Putting these in matrix notation we get to $\boldsymbol{M}_x^{set} = \pi \boldsymbol{M}_x^{class}$. Assuming π has full column-rank, we can obtain $\mu_x^{class} = (\pi^T \pi)^{-1} \pi^T \mu_x^{set}$, to be used as an approximation of $\boldsymbol{E}_{x \sim p(x|y)}[\boldsymbol{\phi}(\boldsymbol{x}, y)]$ in Eq. 2.

Maximum Margin Approaches. The maximum margin principle has been widely used in both supervised and semi-supervised learning [3]. In [18] it was also introduced to the unsupervised setting under the name Maximum Margin Clustering (MMC).

Informally, the labels are arranged in a way such that, had an SVM been trained on the (labeled) data, it would achieve a maximum margin solution. A treatment of MMC can be found in [10]. In [20] the authors present $\propto$SVM, based on MMC with an extra term in the loss function, depending on the provided and estimated bag proportions.

In [15] the authors follow the maximum margin principle by developing a model based on the Support Vector Regression. They present *Inverse Calibration* (InvCal) that replaces the actual dataset with *super-instances* [19], one for each bag, with soft-labels corresponding to their bag-proportions.

3 Discriminant Learning with Ordinal Label Proportions

In this section we first present some necessary background on Linear Discriminant Analysis (LDA), then proceed with the adaptation to the ordinal setting and finally introduce our algorithm.

3.1 Preliminaries

LDA is one of the main approaches in supervised dimensionality reduction, but is also widely used as a classification technique. LDA aims at finding a projection of the data that both minimizes the within-class variance and maximizes the between-class variance.

Following [17], let us define the within-class and between-class scatter matrices (denoted by the w and b subscripts, respectively):

$$\boldsymbol{S}_w = \frac{1}{N} \sum_{k=1}^{K} \sum_{x \in C_k} (\boldsymbol{x} - \boldsymbol{m}_k)(\boldsymbol{x} - \boldsymbol{m}_k)^T \tag{3}$$

where the first sum runs over the K classes, and the second over the elements in each class (where C_k is used to denote the set of data-points in each class) and where $\boldsymbol{m}_k = \frac{1}{N_k} \sum_{x \in C_k} \boldsymbol{x}$ denotes the mean of each class.

The between-class scatter matrix is defined as:

$$\boldsymbol{S}_b = \frac{1}{N} \sum_{k=1}^{K} N_k (\boldsymbol{m}_k - \boldsymbol{m})(\boldsymbol{m}_k - \boldsymbol{m})^T \tag{4}$$

where $\boldsymbol{m} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i$ is used to denote the global mean.

The projection is found by minimizing the following generalized Rayleigh quotient:

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \boldsymbol{J}(\boldsymbol{w}), \quad \text{where} \quad \boldsymbol{J}(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w}} \tag{5}$$

This can be solved as a generalized eigenvalue problem. As with many popular techniques, such as Support Vector Machines and Principal Component Analysis, LDA can be kernelized as well and give rise to Kernel Discriminant Analysis (see for example, [1,13]).

3.2 Kernel Discriminant Learning for Ordinal Regression (KDLOR)

As mentioned earlier, in the Ordinal Classification setting the classes exhibit a natural ordering. This statement can be easily formulated as a constraint to an optimization problem. Similarly to LDA, the projection should be such that the between-class scatter is high and within-class scatter is small. This gives rise to the following problem [17]:

$$\begin{aligned} \min\ \boldsymbol{J}(\boldsymbol{w},\rho) &= \boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w} - C\rho \\ s.t.\ \boldsymbol{w}^T(\boldsymbol{m}_{k+1} - \boldsymbol{m}_k) &\geq \rho, \ \text{ for } \ k = 1, \cdots, K-1 \end{aligned} \tag{6}$$

where C can be understood as the parameter controlling the penalty on the margin between the means and where $\rho > 0$ defines the margin between the class means. Also, without loss of generality, we have assumed the class numbering (the subscript) is in accordance with the natural ordering of the classes. It can be easily seen that this problem gives rise to a projection that abides to the desired properties. We want our projection to have: (1) small within-class variance, (2) large distances between the means of the classes, and (3) a projection that respects the inherent ordering.

To solve the above problem we proceed by forming the Lagrangian as follows:

$$\mathcal{L}(\boldsymbol{w},\rho,\alpha) = \boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w} - C\rho - \sum_{k=1}^{K-1} \alpha_k \big(\boldsymbol{w}^T(\boldsymbol{m}_{k+1} - \boldsymbol{m}_k) - \rho\big) \tag{7}$$

where $\alpha_k \geq 0$ are the Lagrange multipliers. Differentiating with respect to $\boldsymbol{w}$ and ρ we get:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}} = 0 \rightarrow \boldsymbol{w} = \frac{1}{2} \boldsymbol{S}_w^{-1} \sum_{k=1}^{K-1} \alpha_k (\boldsymbol{m}_{k+1} - \boldsymbol{m}_k)$$

$$\frac{\partial \mathcal{L}}{\partial \rho} = 0 \rightarrow \sum_{k=1}^{K-1} \alpha_k = C$$

The so-called dual problem is formulated as follows:

$$\begin{aligned} \min\ f(\alpha) = \sum_{k=1}^{K} \alpha_k (\boldsymbol{m}_{k+1} - \boldsymbol{m}_k)^T \boldsymbol{S}_w^{-1} \sum_{k=1}^{K} \alpha_k (\boldsymbol{m}_{k+1} - \boldsymbol{m}_k) \\ s.t.\ \alpha_k \geq 0,\ k = 1, \cdots, K-1 \\ \sum_{k=1}^{K} \alpha_k = C \end{aligned} \tag{8}$$

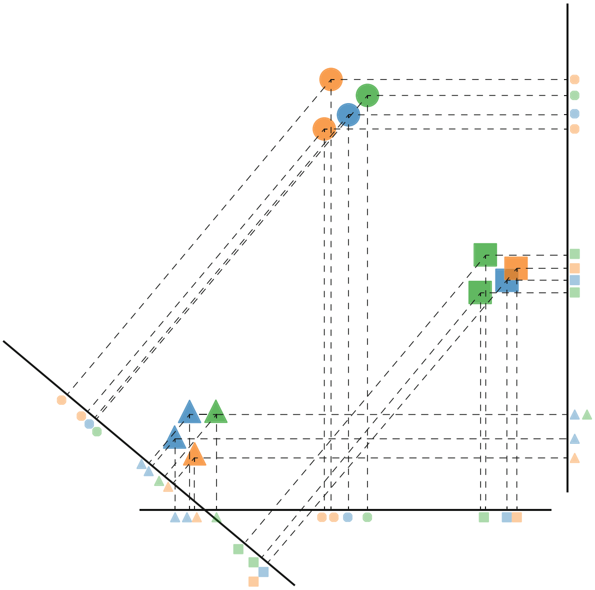

Fig. 1. Ordinary label proportions toy setup. Shape is as indication of class assignment, while colour is an indication of bag assignment. On the figure, we also see three possible projections, that all allow for perfect separation of the data. (Color figure online)

This is an example of (convex) Quadratic Programming with linear constraints and can be solved via a variety of methods. After solving this optimization program for $\boldsymbol{\alpha}^*$, the projection can be obtained using

$$\boldsymbol{w}^* = \frac{1}{2}\boldsymbol{S}_w^{-1}\sum_{k=1}^{K-1}\alpha_k^*(\boldsymbol{m}_{k+1} - \boldsymbol{m}_k) \tag{9}$$

and the derived decision rule is as follows,

$$f(x) = \min_{k\in\{1,\cdots,K\}}\{k : \boldsymbol{w^*}^T\boldsymbol{x} - b_k < 0\} \tag{10}$$

where $b_k = \boldsymbol{w}^T\frac{N_{k+1}\boldsymbol{m}_{k+1}+N_k\boldsymbol{m}_k}{N_{k+1}+N_k}$.

3.3 Discriminant Learning for Ordinal Label Proportions (DL-OLP)

The algorithm presented in the previous subsection (KDLOR) is suitable for the fully supervised learning setting. Though, when the level of supervision is that of learning with label proportions, KDLOR cannot be employed as the class means are required for both the main problem Eq. 8 and for the computation of the within-class scatter matrix in Eq. 3. To that end, we choose to estimate the class means building upon [14]. Figure 1 demonstrates the setting through a toy example, where we have three (well) separated clusters (as shown by shape) with the only supervision available being the label proportions (as shown by colour). In the Figure we also see three possible projections, which all fully separate the clusters.

Key to this derivation is the underlying (but often realistic) assumption of conditional independence of a data point and its bag assignment, given its class assignment. Formally, $p(x|y,i) = p(x|y)$, which gives us:

$$p(x|i) = \sum_y p(x|y,i)p(y|i) = \sum_y p(x|y)\pi_{iy} \tag{11}$$

where $\pi_{iy} = p(y|i)$. Let $\boldsymbol{\mu}_k$ denote the mean of x in bag k and $\boldsymbol{m}_c$ the mean of class c. Following Eq. 11,

$$\boldsymbol{\mu}_k := \boldsymbol{E}_{x\sim p(x|i)}[\boldsymbol{\phi}(\boldsymbol{x})] = \sum_y \pi_{iy}\boldsymbol{m}_y \tag{12}$$

Putting these in matrix form, with $\boldsymbol{M}^{bag}$ and $\boldsymbol{M}^{class}$ denoting the matrices of means for the bags and classes, respectively, we have $\boldsymbol{M}^{bag} = \boldsymbol{\pi}\boldsymbol{M}^{class}$, from which we can obtain a least squares estimate of $\boldsymbol{M}^{class}$:

$$\hat{\boldsymbol{M}}^{class} = \boldsymbol{\pi}^{+}\boldsymbol{M}^{bag} \tag{13}$$

where the + superscript denotes the Moore–Penrose pseudo-inverse. (For the sake of clarity, it should be noted that $\boldsymbol{\pi}$ denotes a *matrix*, and *not* a vector).

Having shown how to estimate the class means, let us make explicit how to compute the within-class scatter matrix, as it requires a sum over data points in each class.

$$\begin{aligned}\boldsymbol{S}_w &= \frac{1}{N}\sum_{k=1}^{K}\sum_{x\in C_k}(\boldsymbol{x}-\boldsymbol{m}_k)(\boldsymbol{x}-\boldsymbol{m}_k)^T \\ &= \frac{1}{N}\sum_{k=1}^{K}\sum_{x\in C_k}\boldsymbol{x}\boldsymbol{x}^T - 2\boldsymbol{x}\boldsymbol{m}_k^T + \boldsymbol{m}_k\boldsymbol{m}_k^T \\ &= \frac{1}{N}\sum_{x\in\mathcal{X}}\boldsymbol{x}\boldsymbol{x}^T + \frac{1}{N}\sum_{k=1}^{K}N_k\boldsymbol{m}_k\boldsymbol{m}_k^T - \frac{2}{N}\sum_{k=1}^{K}\boldsymbol{m}_k\sum_{x\in C_k}\boldsymbol{x} \\ &= \frac{1}{N}\sum_{x\in\mathcal{X}}\boldsymbol{x}\boldsymbol{x}^T - \frac{1}{N}\sum_{k=1}^{K}N_k\boldsymbol{m}_k\boldsymbol{m}_k^T\end{aligned} \tag{14}$$

The procedure is finally summarized in Algorithm 1. When new instances are observed, one can plug in $\boldsymbol{w}^*$ into Eq. 10 to obtain the corresponding prediction.

4 Experiments

In our experiments we consider both synthetic and real-world datasets. We first describe the datasets and then present the empirical results. In our experiments

Algorithm 1. LDA for OLP

```
Input: $\boldsymbol{\pi}, \boldsymbol{X}$, bag assignments
Output: $\boldsymbol{w}^*$
Compute the means of each bag, $\boldsymbol{M}^{bag}$.
Compute the means of each class using Eq. 13.
Compute the within-class scatter matrix, $\boldsymbol{S}_w$ using Eq. 14.
Solve the problem defined by Eq. 8 for $\boldsymbol{\alpha}^*$.
Obtain the projection $\boldsymbol{w}^*$ using Eq. 9
```

we want to test the relevance of the ordinal constraints as well as the trade-off in accuracy when using aggregated proportions instead of the actual labels. To that end we compare against: KDLOR - trained with the actual labels, MeanMap - trained with bag proportions. Additionally, we use Clustering as a baseline. For clustering, we first run k-means, with the number of components corresponding to the number of classes. Then, in order to classify each cluster, we consider a voting scheme, where each data point's vote is its corresponding bag's proportions.

Regarding the first two types of experiments (Synthetic and Benchmark datasets) we make the following notes.

Evaluation. In our experiments (except REF) we consider the test data to be provided *without* bag proportions. In the case of the test data set being provided in the form of a bag, one could do the training in the exact same manner as presented in the paper and during testing, after the predictions have been generated, sort the data points of each bag according to their scores and re-arrange predictions to account for the provided bag proportions. For the synthetic and benchmark datasets we have access to the ground truth (the true labels) and our evaluation is based on those.

Results. These should be read as follows. The first column is the name of the dataset. The rest of the values should be read as *mean*(*one standard deviation*).

4.1 Synthetic Dataset

Data. In our experiments we consider one synthetic dataset configuration as shown in Fig. 2. The data samples $(100, 1000)$ were generated as coming from three Gaussian distributions with means lying on a line on equal intervals and identity covariance matrix.

Model Setup. In our experiments we focus on problems involving three classes of equal size, and three bags, again, of equal size – but different proportions. The proportions used are: {(0.25, 0.25, 0.50, 0.50, 0.25, 0.25, 0.25, 0.50, 0.25)}. The data is first generated according to the desired total size and then separated into the bags, respecting the desired bag proportions.

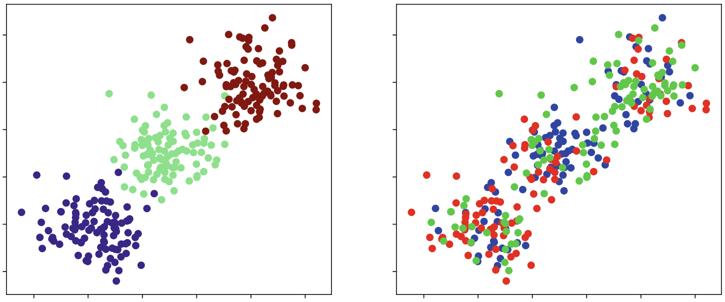

Fig. 2. Configuration used for the synthetic dataset. On the left, colour indicates class, while on the right, colour indicates bag assignment. (Color figure online)

Results. Results for the synthetic dataset are shown in the first two rows of Tables 1 and 2. Table 1 shows results on Zero-One accuracy while Table 2 shows results on Mean Absolute Error (MAE) loss. This simple example is only intended to show the difference between the two evaluation metrics. In terms of MAE, as expected the best performing method is the inherently ordinal KLDOR, while DL-OLP outperforms MeanMap. The Clustering baseline works particularly well due to the underlying distribution of the data.

4.2 Benchmark Datasets

Data. In our experiments we also consider various benchmark datasets for ordinal classification. The datasets can be found in the repository provided[2].

Model Setup. For these datasets, three classes were chosen and then three bags were created by randomly picking points from the classes to fulfill the desired bag proportions. The bag proportions used are: $\{(0.25, 0.25, 0.50, 0.50, 0.25, 0.25, 0.25, 0.50, 0.25)\}$.

Results. Results for the benchmark datasets are shown in Tables 1 and 2. Again, in most cases, in terms of MAE, as expected the best performing method is the inherently ordinal KLDOR, while DL-OLP makes better use of the label proportions than MeanMap by respecting the ordinal constraints.

4.3 REF Dataset

The final experiment is a real-world case study illustrating the effectiveness of DL-OLP in a problem of actual importance, both for accurate prediction and for interpretation of the prediction model (i.e., the weight vector).

[2] The benchmark datasets are also available at http://www.gagolewski.com/resources/data/ordinal-regr.

Table 1. Zero-one accuracy

	KDLOR	MeanMap	Clustering	DL-OLP
Synthetic100	0.98(0.014)	0.49(0.056)	0.97(0.018)	0.97(0.026)
Synthetic1000	0.98(0.003)	0.53(0.052)	0.98(0.006)	0.97(0.002)
Cali/Housing	0.67(0.015)	0.35(0.032)	0.29(0.011)	0.53(0.018)
Cement-Strength	0.93(0.004)	0.72(0.012)	0.66(0.013)	0.87(0.026)
Fireman	0.83(0.011)	0.70(0.008)	0.64(0.019)	0.81(0.023)
Kinematics	0.66(0.016)	0.62(0.012)	0.68(0.015)	0.62(0.022)
Skill	0.66(0.028)	0.55(0.042)	0.59(0.04)	0.59(0.012)
Stockord	0.68(0.042)	0.40(0.049)	0.59(0.061)	0.80(0.078)

Table 2. Mean absolute error

	KDLOR	MeanMap	Clustering	DL-OLP
Synthetic100	0.023(0.016)	0.66(0.090)	0.26(0.018)	0.031(0.026)
Synthetic1000	0.024(0.003)	0.61(0.031)	0.024(0.006)	0.028(0.009)
Cali/Housing	0.35(0.016)	0.69(0.036)	0.77(0.025)	0.51(0.022)
Cement-Strength	0.073(0.004)	0.28(0.012)	0.35(0.013)	0.14(0.034)
Fireman	0.17(0.011)	0.32(0.010)	0.38(0.029)	0.20(0.023)
Kinematics	0.40(0.016)	0.44(0.015)	0.36(0.017)	0.42(0.013)
Skill	0.35(0.034)	0.52(0.057)	0.44(0.068)	0.42(0.011)
Stockord	0.36(0.086)	0.61(0.056)	0.42(0.086)	0.16(0.057)

Data. In the United Kingdom, the Research Excellence Framework (REF) assesses the research of all university departments. This is done by asking all departments to send in a list of the 4 best publications for each academic staff member. Then a panel decides how many of these submissions are awarded 4* (world-leading), 3* (internationally recognized), 2* (nationally leading), and 1* (nationally recognized) or unclassified). At the end of the process, the outcome is a histogram for each department stating how many research outputs were in each category, but not revealing which paper was in each of these classes. Funding levels are then based on this histogram. As the panel do not reveal how they classify papers, in deciding which papers to submit, the departments have to guess which papers the panel would rank most highly. This also affects the individuals' and departmental publication strategy. In this experiments we aim to reverse engineer the panel's thinking, and work out a classifier that mimics the way the panel operates.

The data is online for all university departments: all the papers each department submitted, and the histogram saying how many were in each class for each department. We use the 2008 submission for all Computer Science departments in the country to compile our REF dataset[3]. For each book, book chapter, conference or journal paper, we collect the following features whenever available:

1. `Number of citations`: total number of citations for a given paper at the time of submission from Google Scholar (integer).
2. `Number of authors`: total number of authors for a given paper (integer).
3. `IsMultidisciplinary`: whether the outcome is categorized by the author as a multidisciplinary piece of research (boolean).
4. `Ranking ERA`: ranking provided by the Excellence in Research for Australia (ERA)[4] for journals and conferences (ordinal categorical).
5. `JCR total citations`: total number of citations for every article published in the journal from the Journal Citation Reports (JCR)[5] (integer).
6. `JCR impact factor`: frequency with which an average article from a journal is cited in a particular year (positive real).
7. `JCR Immediacy index`: frequency with which the average article from a journal is cited within the same year as publication (positive real).
8. `JCR total number of articles of the journal`: total number of articles published in the publication for a specific year (integer).
9. `JCR Cited half-life`: median age of the articles in the journal that were cited in the year (positive real).
10. Additionally, we compute a feature based on the product of the `JCR impact factor` and the `JCR cited half-life`, as this was traditionally thought to be a good proxy for the behaviour of the panel.

This leads to a total of 10 features for the 4966 research outputs over the 81 Computer Science departments.

For non-journal papers, the JCR measures (features 5–10) are not available, and feature 4 is not available for contributions other than conference and journal papers. There are many possible approaches for filling out these missing values. For the sake of simplicity, we set all missing feature values to zero. While this is clearly not very sophisticated or well-founded (and alternative approaches are subject of ongoing investigations), we will show later in the discussion that it nonetheless leads to informative and interpretable results.

Model Setup. For the REF dataset, we consider each department as being one bag and each paper's (hidden) rating to range between 4* and no-stars. We therefore have a total of 81 bags and 5 classes. Even though we do not include experiments using the kernel extension of the algorithm for interpretability purposes, we do provide some guidelines as to how one could use it. The standard approach to learning parameters of a model is through k-fold cross-validation (where one splits the dataset in k folds, trains the model on $k-1$ of them, and tests on the k-th one). In the LLP setting one does not have access to the true labels, so the standard CV procedure cannot be employed. One can though

[3] https://www.rae.ac.uk/pubs/2008/01/.

[4] http://www.arc.gov.au/excellence-research-australia.

[5] http://jcr.incites.thomsonreuters.com/.

adapt it to this setting by using the bags as folds. In order to evaluate the performance on the 'test-bag', the practitioner can consider how well the estimated bag proportions match the provided ones.

Discussion. For the REF data, the ground truth (the actual rating for each paper) is not available and therefore our evaluation is limited. Therefore, we focus on discussing our model parameters. With regards to MAE on the bag proportions, our approach achieves a value of 12.14, outperforming clustering which achieves a value of 19.76.

Based on the empirical long-tail distribution of `Number of Citations`, it was believed sensible to apply a log-transform on the feature (the difference between 0 and 10 citations is a lot more important than the difference between 1000 and 1010 citations). The rest of the features are standardized. The weight vector obtained by DL-OLP is as follows (numbers are given in the same order as the numbered list of features shown before):

$$[0.179, -0.025, 0.158, 0.026, -0.020, -0.095, -0.074, 0.032, -0.110, -0.081].$$

Not only does DL-OLP allow one to predict the histogram for a given REF submission well – the weight vector also provides insight into what the panel values. The (positive) dominating features are the `Number of citations`, together with multidisciplinary nature of the submission. The importance of the number of citations comes at no surprise. The latter is in accordance with the widely held belief that multidisciplinary contributions are valued more highly by the REF panels.

It is worth noting that the number of authors has a negative weight. Thus, a large number of authors is perceived as a lowering the quality of a submission (perhaps as one's contribution to it is considered inversely proportional to the number of authors). However, many authors are justified from a REF optimization perspective if necessitated by the paper's multidisciplinary nature. To further illustrate this point, we proceeded by combining these two features into one, by multiplying `Number of authors` with +1 in the case of `isMultidisciplinary` being true, and −1 otherwise. The new weight vector is shown below (N/A is inserted for the third feature, as it is now combined with the second):

$$[0.102, 0.153, \mathrm{N/A}, -0.019, -0.019, -0.051, -0.033, 0.002, -0.104, -0.036].$$

Interestingly, the new feature now dominates the weight vector, with a positive effect, with the number of citations coming second.

A final observation is that the JCR features (with the exception of the 8'th feature, total number of articles in the journal), are negative. This may seem counter intuitive at first, but there are two logical explanations. The first is that journal-level metrics are only proxies for a paper's actual impact, which is better quantified by the actual number of citations (which does have a large positive weight). To test this explanation, we also ran our method after removing the first feature, yielding the following weights (again with N/A for the first as this one is removed):

$$[\text{N/A}, -0.004, 0.078, 0.019, -0.000, -0.068, -0.072, 0.090, 0.003, -0.055].$$

We do indeed see that the JCR measures (features 5–10) all increase, yet, most of them remain negative. This can be explained by our approach of filling in missing values for non-journal papers. Actually, these have JCR measures that are set to zero (i.e. the minimum possible). The result thus implies that the REF panels do value strong non-journal contributions, contrary to popular belief.

5 Conclusion

In this paper we have introduced a new learning task which we have called Ordinal Label Proportions. We have also presented a method to tackle the problem based on discriminant learning and a sound estimation of class means. The method aims to find a projection that minimizes the within-class scatter while also respecting the natural ordering of the classes. Our approach compares favourably with MeanMap, that does not exploit the ordinal nature of the data. Moreover, even though DL-OLP has the benefit of training with the true labels, instead of label proportions, only a minor setback is observed empirically. In the future we wish to examine more real world data sets that exhibit the characteristics of Ordinal Label Proportions.

Acknowledgements. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC Grant Agreement no. 615517, from the FWO (project no. G091017N, G0F9816N), and from the European Union's Horizon 2020 research and innovation programme and the FWO under the Marie Sklodowska-Curie Grant Agreement no. 665501. Additionally, this study was supported by EPSRC and MRC through the SPHERE IRC (EP/K031910/1) and CUBOID (MC/PC/16029) respectively.

References

1. Bishop, C.: Pattern Recognition and Machine Learning. Springer, Boston (2006). https://doi.org/10.1007/978-1-4615-7566-5
2. Chu, W., Keerthi, S.S.: Support vector ordinal regression. Neural Comput. **19**(3), 792–815 (2007)
3. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, Cambridge (2000)
4. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. **89**(1–2), 31–71 (1997)
5. Frank, E., Hall, M.: A simple approach to ordinal classification. In: De Raedt, L., Flach, P. (eds.) ECML 2001. LNCS (LNAI), vol. 2167, pp. 145–156. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44795-4_13

6. Gutiérrez, P.A., Perez-Ortiz, M., Sanchez-Monedero, J., Fernandez-Navarro, F., Hervas-Martinez, C.: Ordinal regression methods: survey and experimental study. IEEE Trans. Knowl. Data Eng. **28**(1), 127–146 (2016)
7. Herbrich, R., Graepel, T., Obermayer, K.: Support vector learning for ordinal regression. In: International Conference on Artificial Neural Networks. IET (1999)
8. Huhn, J.C., Hullermeier, E.: Is an ordinal class structure useful in classifier learning? Int. J. Data Min. Model. Manag. **1**(1), 45–67 (2008)
9. Kuck, H., de Freitas, N.: Learning about individuals from group statistics (2012). arXiv preprint: arXiv:1207.1393
10. Li, Y.F., Tsang, I.W., Kwok, J., Zhou, Z.H.: Tighter and convex maximum margin clustering. In: Artificial Intelligence and Statistics, pp. 344–351 (2009)
11. Maron, O., Ratan, A.L.: Multiple-instance learning for natural scene classification. In: Proceedings of the Fifteenth International Conference on Machine Learning, pp. 341–349. Morgan Kaufmann Publishers Inc., San Francisco (1998)
12. McCullagh, P.: Regression models for ordinal data. J. R. Stat. Soc. Ser. B (Methodol.) **42**, 109–142 (1980)
13. Mika, S., Rätsch, G., Weston, J., Schölkopf, B., Müller, K.R.: Fisher discriminant analysis with kernels. In: Proceedings of the 1999 IEEE Signal Processing Society Workshop, Max-Planck-Gesellschaft, vol. 9, pp. 41–48. IEEE (1999)
14. Quadrianto, N., Smola, A.J., Caetano, T.S., Le, Q.V.: Estimating labels from label proportions. J. Mach. Learn. Res. **10**, 2349–2374 (2009)
15. Rueping, S.: SVM classifier estimation from group probabilities. In: Proceedings of the 27th International Conference on Machine Learning, pp. 911–918 (2010)
16. Santos-Rodríguez, R., Guerrero-Curieses, A., Aláiz-Rodríguez, R., Cid-Sueiro, J.: Cost-sensitive learning based on Bregman divergences. Mach. Learn. **76**, 14 (2009)
17. Sun, B.Y., Li, J., Wu, D.D., Zhang, X.M., Li, W.B.: Kernel discriminant learning for ordinal regression. IEEE Trans. Knowl. Data Eng. **22**(6), 906–910 (2010)
18. Xu, L., Neufeld, J., Larson, B., Schuurmans, D.: Maximum margin clustering. In: Advances in Neural Information Processing Systems, pp. 1537–1544 (2005)
19. Yu, F.X., Choromanski, K., Kumar, S., Jebara, T., Chang, S.F.: On learning from label proportions (2014). arXiv preprint: arXiv:1402.5902
20. Yu, F.X., Liu, D., Kumar, S., Jebara, T., Chang, S.F.: $\propto$ SVM for learning with label proportions (2013). arXiv preprint: arXiv:1306.0886