Semi-supervised Learning for Ordinal Kernel Discriminant Analysis

M. Pérez-Ortiz[a,*], P.A. Gutiérrez[b], M. Carbonero-Ruz[a], C. Hervás-Martínez[b]

[a] *Department of Quantitative Methods, Universidad Loyola Andalucía, 14004-Córdoba, Spain*
[b] *Department of Computer Science and Numerical Analysis, University of Córdoba, 14070-Córdoba, Spain*

Abstract

Ordinal classification considers those classification problems where the labels of the variable to predict follow a given order. Naturally, labelled data is scarce or difficult to obtain in this type of problems because, in many cases, ordinal labels are given by an user or expert (e.g. in recommendation systems). Firstly, this paper develops a new strategy for ordinal classification where both labelled and unlabelled data are used in the model construction step (a scheme which is referred to as semi-supervised learning). More specifically, the ordinal version of kernel discriminant learning is extended for this setting considering the neighbourhood information of unlabelled data, which is proposed to be computed in the feature space induced by the kernel function. Secondly, a new method for semi-supervised kernel learning is devised in the context of ordinal classification, which is combined with our developed classification strategy to optimise the kernel parameters. The experiments conducted compare 6 different approaches for semi-supervised learning in the context of ordinal classification in a battery of 30 datasets, showing 1) the good synergy of the ordinal version of discriminant analysis and the use of unlabelled data and 2) the advantage of computing distances in the feature space induced by the kernel function.

Keywords: ordinal regression, discriminant analysis, semi-supervised learning, classification, kernel learning

1. Introduction

With the advent of the big data era and the increased popularity of machine learning, the number of scientific data-driven applications is growing at an abrupt pace. Because of this increased necessity, new related research avenues are explored every year. In this sense, the recently coined term weak

*Corresponding author
Email address: i82perom@uco.es (M. Pérez-Ortiz)

supervision [1] refers to those classification machine learning problems where the labelling information is not as accessible as in the fully-supervised problem (where a label is associated to each pattern). The problem of semi-supervised learning (i.e. learning from both labelled and unlabelled observations) is an example that has been the focus of many machine learning researchers in the past years. In many real-world applications, obtaining labelled patterns could be a challenging task, however, unlabelled examples might be available with little or no cost. The main idea behind semi-supervised learning is to take advantage from unlabelled data when constructing the machine classifier (and this is done using different assumptions on the unlabelled data: smoothness, clustering or manifold assumptions [2, 3, 4]). These learning approaches have been empirically and theoretically studied in the literature and represent a suitable solution for such circumstances, where the use of unlabelled data has been seen to improve the performance of the model and stabilise it. Semi-supervised learning has being mainly studied for binary classification [5, 6] and regression [2], although recently the main focus has shifted to multi-class problems [7, 8, 9] (and even multi-dimensional ones [10]). This paper tackles the use of unlabelled data in the context of ordinal classification [11], a learning paradigm which shares properties of both classification and regression.

Ordinal regression (also known as ordinal classification) can be defined as a relatively new learning paradigm whose aim is to learn a prediction rule for ordered categories. In contrast to multinomial classification, there exists some ordering among the elements of $\mathcal{Y}$ (the labelling space) and both standard classifiers and the zero-one loss function do not capture and reflect this ordering appropriately [11] (leading to worse models in terms of errors in the ordinal scale). Concerning regression, $\mathcal{Y}$ is a non-metric space.

An explanatory example of order among categories is the Likert scale [12], a well-known methodology used for questionnaires, where the categories correspond to the level of agreement or disagreement for a series of statements. The scheme of a typical five-point Likert scale could be: {*Strongly disagree, Disagree, Neither agree or disagree, Agree, Strongly Agree*}, where the natural order among categories can be appreciated. The major problem within this kind of classification is that the misclassification errors should not be treated equally, e.g., misclassifying a *Strongly disagree* pattern as *Strongly agree* should be more penalised than a misclassification with the *Disagree* category.

Several issues must be highlighted when developing new ordinal classifiers in order to exploit the presence of this order among categories. Firstly, this implicit data structure should be learned by the classifier in order to minimise the different ordinal classification errors [11], and, secondly, different evaluation measures or metrics should be developed in this context. The most popular approach for this type of problems are threshold models [13, 14, 15, 16]. These methods are based on the idea that, to model ordinal ranking problems from a regression perspective, one can assume that some underlying real-valued outcomes exist (also known as latent variable), which are, in practice, unobservable.

Recently, a version of the well-known Kernel Discriminant Analysis algorithm has been proposed for ordinal regression [13], showing different advantages

with respect to other ordinal classification methods, i.e. a lower computational complexity and the ability to capture the associated class distributions. In essence, the formulation seeks for the projection that allows the greater separation for the classes, but maintaining the classes ordered in the projection (to avoid serious misclassification errors). This algorithm, Kernel Discriminant Learning for Ordinal Regression (KDLOR), has shown great potential and competitiveness against other specially designed ordinal classifiers.

However, supervised ordinal regression approaches present limitations when there are few data [17, 18], which is a common situation in this setting, where most ordinal classification problems are labelled by an user or expert (a process that could be expensive or time-consuming), and the number of classes is usually relatively high (which hinders the class discrimination to a great extent). Consider, for example, the case of a film recommendation system, where most users might not have interest in labelling data, therefore unlabelled data exist and are easily available. In this sense, the paradigm of semi-supervised learning would use the unlabelled data along with the labelled data to learn more precise models. The development and analysis of semi-supervised ordinal regression algorithms is, therefore, of great interest. However, the number of works in the literature approaching this problem is very low [17, 18, 19, 20], where only two of them focus on developing ordinal and semi-supervised classifiers [17, 18] (the remainder focuses on related frameworks, such as the transductive problem [19, 20] or clustering [21], which are out of the scope of this paper).

We propose and test different approaches to deal with semi-supervised ordinal classification problems. Firstly, we extend the KDLOR algorithm to make use of unlabelled data via the smoothness and manifold assumptions, (i.e. (1) points nearby are likely to share the same label, and (2) the projection should not only match the classification task but also respect the geometric structure inferred from labelled and unlabelled data points). Secondly, this paper proposes to compute the graph Laplacian (used for the previous objective) in the feature space induced by the kernel function, as opposed to computing it in the input space. Since the final objective function is computed in the feature space, this is a crucial consideration for the proposed technique. Finally, we also propose a new method for semi-supervised kernel learning based on kernel-target alignment to use in conjunction with (ordinal) kernel methods. Kernel learning techniques are a common choice to optimise the kernel parameters and adequately fit the data using a kernel function [22, 23]. We test our proposals in a set of 30 ordinal classification datasets and compare them to other strategies, the results showing the good synergy of combining labelled and unlabelled data in the context of ordinal regression.

The rest of the paper is organised as follows: Section II shows a description of previous concepts; Section III presents the proposal of this work and Section IV describes the specific characteristics of the datasets and the experimental study; Section V analyses the results obtained; and finally, Section VI outlines some conclusions and future work.

2. Previous notions

This section introduces some of the previous work in the area of the paper.

Consider a training sample $D = \{\mathbf{x}_i, y_i\}_{i=1}^N \subseteq \mathcal{X} \times \mathcal{Y}$ generated i.i.d. from a (unknown) joint distribution $P(\mathbf{x}, y)$, where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_Q\}$. In the ordinal regression setup, the labelling space is ordered due to the data ranking structure ($\mathcal{C}_1 \prec \mathcal{C}_2 \prec \cdots \prec \mathcal{C}_Q$, where $\prec$ denotes this order information). Let N be the number of patterns in the training sample, N_q the number of samples for the q-th class and $\mathbf{X}_q$ the set of patterns belonging to class $\mathcal{C}_q$.

Furthermore, let $\mathcal{H}$ denote a high-dimensional Hilbert space. Then, for any mapping of patterns $\Phi : \mathcal{X} \to \mathcal{H}$, the inner product $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle_{\mathcal{H}}$ of the mapped inputs is known as a kernel function, giving rise to a positive semidefinite (PSD) matrix $\mathbf{K}$ for a given input set $\{\mathbf{x}_i\}_{i=1}^N$.

2.1. Discriminant Learning

This learning paradigm is one of the pioneers and leading techniques in the machine learning area, being currently used for supervised dimensionality reduction and classification. The main goal of this technique can be described as finding the optimal linear projection for the data (from which different classes can be well separated). To do so, the algorithm analyses two objectives: the maximisation of the between-class distance and the minimisation of the within-class distance, by using variance-covariance matrices ($\mathbf{S}_b$ and $\mathbf{S}_w$, respectively) and the so-called Rayleigh coefficient ($J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}} \mathbf{S}_{\mathrm{b}} \mathbf{w}}{\mathbf{w}^{\mathrm{T}} \mathbf{S}_{\mathrm{w}} \mathbf{w}}$, where $\mathbf{w}$ is the projection for the data). To achieve these objectives, the $Q-1$ eigenvectors associated to the highest eigenvalues of $\mathbf{S}_{\mathrm{w}}^{-1} \cdot \mathbf{S}_{\mathrm{b}}$ are computed.

The between-class and within-class scatter matrices ($\mathbf{S}_b$ and $\mathbf{S}_w$, respectively) are defined as follows (when considering the kernel version):

$$\mathbf{S}_{\mathrm{w}} = \frac{1}{N} \sum_{q=1}^{Q} \sum_{\mathbf{x}_i \in \mathbf{X}_q} (\Phi(\mathbf{x}_i) - \mathbf{M}_q^{\Phi})(\Phi(\mathbf{x}_i) - \mathbf{M}_q^{\Phi})^{\mathrm{T}}, \tag{1}$$

$$\mathbf{S}_{\mathrm{b}} = \frac{1}{N} \sum_{q=1}^{Q} N_q (\mathbf{M}_q^{\Phi} - \mathbf{M}^{\Phi})(\mathbf{M}_q^{\Phi} - \mathbf{M}^{\Phi})^{\mathrm{T}}, \tag{2}$$

where $\mathbf{M}_q^{\Phi} = \frac{1}{N_q} \sum_{\mathbf{x}_i \in \mathbf{X}_q} \Phi(\mathbf{x}_i)$, and $\mathbf{M}^{\Phi} = \frac{1}{N} \sum_{i=1}^{N} \Phi(\mathbf{x}_i)$. The objectives presented can be achieved by the maximisation of the so called Rayleigh coefficient. Note that, when dealing with kernel functions, $\mathbf{w}$ will have an expansion of the form:

$$\mathbf{w} = \sum_{i=1}^{N} \beta_i \Phi(\mathbf{x}_i), \beta_i \in \mathbb{R}, \tag{3}$$

where β_i represents the contribution of $\mathbf{x}_i$ to the projection $\mathbf{w}$. Then, the Rayleigh coefficient can be formulated as follows:

$$J(\boldsymbol{\beta}) = \frac{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{H} \boldsymbol{\beta}}{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{N} \boldsymbol{\beta}}, \tag{4}$$

where $\mathbf{N} = \sum_{q=1}^{Q} \mathbf{R}_q(\mathbf{I} - \mathbf{1}_{\mathbf{N}q})\mathbf{R}_q^{\mathrm{T}}$, $\mathbf{I}$ is the identity matrix, $\mathbf{1}_{\mathbf{N}q}$ is a matrix with a value of $\frac{1}{N_q}$ for all entries, $\mathbf{R}_q$ is an $N \times N_q$ matrix with $(\mathbf{R}_q)_{i,j} = \langle \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \rangle$ where $< \cdot, \cdot >$ is the scalar product and $\mathbf{x}_j \in X_q$. Moreover, $\mathbf{H} = \sum_{q=1}^{Q} N_q(\mathbf{M}_q - \mathbf{M})(\mathbf{M}_q - \mathbf{M})^{\mathrm{T}}$, where $(\mathbf{M}_q)_j = \frac{1}{N_q} \sum_{\mathbf{x}_h \in \mathbf{X}_q} k(\mathbf{x}_j, \mathbf{x}_h)$ and $\mathbf{M}_j = \frac{1}{N} \sum_{h=1}^{N} k(\mathbf{x}_j, \mathbf{x}_h)$.

Usually, a diagonal term t is added to the $\mathbf{N}$ matrix, so that very small eigenvalues are bounded away from zero to improve numerical stability.

2.2. Semi-supervised Discriminant Learning

Kernel Discriminant Analysis (KDA) seeks the optimal projection for the labelled data. This algorithm has been extended to semi-supervised learning by incorporating the manifold structure suggested by unlabelled data [5]. It is well-known that when training data are scarce, machine learning algorithms tend to overfit. To solve this, a widely used approach is to complement the sample with unlabelled data by imposing a regulariser, that controls the learning complexity of the hypothesis family and balances the model complexity and the empirical loss. In semi-supervised learning, this regulariser is used to incorporate prior knowledge about the data, i.e., the manifold structured imbued by unlabelled data. The key to semi-supervised learning is the consistency assumption [24], which for classification is the notion that nearby patterns are likely to have the same label. This regulariser is precisely what differentiates the supervised and semi-supervised versions of KDA.

Let us explain how this assumption can be included in the KDA algorithm. Given a set of examples $\mathcal{X}$, we can construct a n-nearest neighbour graph G to model the relationship of patterns in the input space. To do so, a weight matrix is defined as follows:

$$S_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_i \in N_n(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in N_n(\mathbf{x}_i), \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where $N_n(\mathbf{x}_i)$ denotes the set of n-nearest neighbours of $\mathbf{x}_i$. Using this formulation, two data points that are linked by an edge are likely to be in the same class. One of the regularisers that have been used in spectral dimensionality reduction [25] is the following:

$$R(\boldsymbol{\beta}) = \sum_{i=1}^{N} \sum_{j=1}^{N} (\mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}_i) - \mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}_j))^2 \cdot S_{ij} = 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{K}\mathbf{L}\mathbf{K}^{\mathrm{T}}\boldsymbol{\beta}, \tag{6}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{S}$ is the Laplacian matrix [26] and $\mathbf{D}$ is defined as the diagonal matrix $D_{ii} = \sum_{j=1}^{N} S_{ij}$.

With this regulariser, the semi-supervised version of KDA has been formulated as follows:

$$\max_{\boldsymbol{\beta}} \frac{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{H}\boldsymbol{\beta}}{\boldsymbol{\beta}^{\mathrm{T}}(\mathbf{N} + \mu\mathbf{K}\mathbf{L}\mathbf{K}^{\mathrm{T}})\boldsymbol{\beta}}, \tag{7}$$

where μ is a parameter to control the balance between the model complexity and the empirical loss. Note that the difference between the supervised and semi-supervised approaches lies in the inclusion of the regulariser (compare Eq. 4 and Eq. 7 to see this). This problem can also be solved by an eigenvalue formulation.

2.3. Kernel Discriminant Learning for Ordinal Regression

The idea of Kernel Discriminant Learning has also been successfully applied in the context of Ordinal Regression (KDLOR) [13]. Roughly speaking, this method searches for the optimal projection of the data that preserves the ordinal class ranking. As in the standard KDA, the objectives presented can be achieved by the maximisation of the Rayleigh coefficient, but including an extra constraint, which will force the projected classes to be ordered according to their ranks. More specifically, the original optimisation problem of KDA (exposed in Eq. 4) is transformed into the following one:

$$\begin{aligned} \min J_{\mathrm{O}}(\boldsymbol{\beta}, \rho) &= \boldsymbol{\beta}^{\mathrm{T}} \mathbf{N} \boldsymbol{\beta} - C\rho, \\ \text{s.t. } \boldsymbol{\beta}^{\mathrm{T}}(\mathbf{M}_{q+1} - \mathbf{M}_q) &\geq \rho,\ q = \{1, \ldots, Q-1\}, \end{aligned} \tag{8}$$

where C is a penalty coefficient. When the value of C is appropriately set, $\rho > 0$ is satisfied, forcing the classes to be ordered in the projection (by the use of the constraint $\boldsymbol{\beta}^{\mathrm{T}}(\mathbf{M}_{q+1} - \mathbf{M}_q) \geq \rho$ which relates the projected class means to their rank). Moreover, the distance between these means contributes positively to the optimisation, which implies that the between-class covariance matrix is not needed for, as this information is already included. This formulation is maintained in this paper for all the ordinal classification based techniques (even semi-supervised approaches).

To solve it, Lagrange multipliers can be applied, a method for optimising functions of several variables subject to constraints. The initial function and the constraints are combined in the following unique function:

$$L_O(\boldsymbol{\beta}, \rho, \boldsymbol{\alpha}) = \boldsymbol{\beta}^{\mathrm{T}} \mathbf{H} \boldsymbol{\beta} - C\rho - \sum_{q=1}^{Q-1} \alpha_q \left\{ \boldsymbol{\beta}^{\mathrm{T}}(\mathbf{M}_{q+1} - \mathbf{M}_q) - \rho \right\}.$$

This technique has been emphasized in the literature for two reasons: 1) its ability to handle nonlinear decision regions at a low computational cost [13] and 2) the fact that it computes the separating hyperplane considering the whole class distribution, whereas SVM-based methods obtain the decision hyperplane in a local way, i.e. using support vectors, which could lead to undesirable solutions in some cases [13]. For more information about this method see [13, 27].

2.4. Kernel matrix learning

Kernel matrices contain information about the similarity among patterns, and this similarity can be used to find the best mapping function Φ associated to a kernel function. The empirical ideal kernel [22], $\mathbf{K}^*$, (*i.e.*, the matrix

that would represent perfect similarity information) will submit the following structure:

$$k^*(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} +1, & \text{if } y_i = y_j, \\ -1, & \text{otherwise,} \end{cases} \tag{9}$$

where $\mathbf{K}^*_{ij} = k^*(\mathbf{x}_i, \mathbf{x}_j)$. $\mathbf{K}^*$ provides information about which patterns in the dataset should be considered as similar when performing a learning task. As we are dealing with a classification problem, patterns from the same class should be considered similar, while patterns from different classes should be considered as dissimilar as possible.

Suppose an ideal kernel matrix $\mathbf{K}^*$ and a given real kernel matrix $\mathbf{K}$. The underlying idea for kernel-target alignment (KTA) [22, 23], the strategy chosen in this paper for kernel learning, is to choose the kernel matrix $\mathbf{K}$ (among a set of different matrices) closest to the ideal matrix $\mathbf{K}^*$.

The KTA between two kernel matrices $\mathbf{K}$ and $\mathbf{K}^*$ is defined as:

$$\mathcal{A}(\mathbf{K}, \mathbf{K}^*) = \frac{\langle \mathbf{K}, \mathbf{K}^* \rangle_{\mathrm{F}}}{\sqrt{\langle \mathbf{K}^*, \mathbf{K}^* \rangle_{\mathrm{F}} \langle \mathbf{K}, \mathbf{K} \rangle_{\mathrm{F}}}}, \tag{10}$$

where $\langle \cdot, \cdot \rangle_{\mathrm{F}}$ represents the Frobenius inner product. This quantity is maximised when the kernel function is capable to reflect the properties of the training dataset used to define the ideal kernel matrix.

2.5. Weighted ordinal discriminant learning

A label propagation method [24] has been used before in the literature to estimate the class memberships of unlabelled data and complement the KDLOR method [18].

Denote the membership matrix by $\mathbf{U} = (u_{jq})_{N \times Q}$, where u_{jq} is the membership of pattern $\mathbf{x}_j$ to class $\mathcal{C}_q$. Note that the memberships of labelled data are obtained by the given labels. The main contribution of [18] (apart from the evolutionary algorithm) is the use of unlabelled data to complement the representation of the class distributions (mean and covariance matrices) in the following manner:

$$\mathbf{M}^{\Phi}_q = \frac{\sum_{j=1}^{N} u_{jq} \Phi(\mathbf{x}_j)}{\sum_{j=1}^{N} u_{jq}}, \tag{11}$$

$$\mathbf{S}_{\mathrm{w}} = \frac{1}{u} \sum_{q=1}^{Q} \sum_{j=1}^{N} u_{jq} (\Phi(\mathbf{x}_j) - \mathbf{M}^{\Phi}_q)(\Phi(\mathbf{x}_j) - \mathbf{M}^{\Phi}_q)^{\mathrm{T}}, \tag{12}$$

where $u = \sum_{q=1}^{Q} \sum_{j=1}^{N} u_{jq}$ and u_{jq} represents the membership grade of $\mathbf{x}_j$ to class $\mathcal{C}_q$. This idea is tested and compared to our proposal in the experimental section of this paper (referred as Weighted Semi-supervised Discriminant Learning for ordinal regression, WS-DL).

However, the authors argue that the proposal does not work efficiently (specially with few data), and therefore they devise an evolutionary approach that evolves the class memberships and improves the performance of the base kernel discriminant learning [18].

3. Proposals for Semi-supervised Ordinal Discriminant Learning

This section describes the different proposals presented in this paper:

- Firstly, a new objective function is proposed for kernel discriminant learning in the context of semi-supervised ordinal classification. More specifically, this new ordinal and semi-supervised classification method uses labelled and unlabelled data to construct a neighbourhood graph of the dataset, which provides a discrete approximation to the local geometry of the data manifold, and which is thereafter introduced into the original optimisation problem of kernel discriminant analysis for ordinal regression [13]. To do so, the notion of graph Laplacian is used and a smoothness penalty is introduced into the graph of the objective function in such a way that the algorithm can optimally preserve the manifold structure. This first approach considers the construction of the neighbourhood graph in the input space (as proposed in previous research [5]).

- As stated before, this previously mentioned neighbourhood graph is used to introduce non-supervised knowledge into the algorithm formulation. Because of this, secondly, we propose the construction of this neighbourhood graph in the empirical feature space induced by the kernel function [28], given that the developed method makes use of the kernel trick.

- Finally, we also introduce a new technique for ordinal semi-supervised kernel learning via kernel-target alignment. This method optimises the kernel parameters taking into account the ordinal and unlabelled nature of the data, so that it could be used in conjunction with any ordinal kernel method when unlabelled data is available (such as the classification strategy proposed in this paper).

Different versions of these proposals are tested in the experimental part of this paper.

3.1. Ordinal semi-supervised learning

This section describes a new formulation for kernel discriminant analysis in the context of ordinal and semi-supervised classification.

To include the manifold structure in the ordinal case, the same approach than in section 2.2 can be followed, including the regulariser into the optimisation formulation. In this sense, the objective function J_{O} in Equation 8 can be reformulated as follows:

$$\begin{aligned} \min J_{\mathrm{OS}}(\boldsymbol{\beta}, \rho, \sigma) = \boldsymbol{\beta}^{\mathrm{T}} \mathbf{N} \boldsymbol{\beta} - C\rho + \mu R(\boldsymbol{\beta}), \\ \text{s.t. } \boldsymbol{\beta}^{\mathrm{T}} (\mathbf{M}_{q+1} - \mathbf{M}_q) \geq \rho, \end{aligned} \tag{13}$$

where μ is a parameter associated to the contribution of unlabelled data to the model.

To solve it, Lagrange multipliers can be applied. The initial function and the constraints are combined as:

$$L(\boldsymbol{\beta}, \rho, \mu, \boldsymbol{\alpha}) = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{N}\boldsymbol{\beta} - C\rho + \mu\boldsymbol{\beta}^{\mathrm{T}}\mathbf{K}\mathbf{L}\mathbf{K}^{\mathrm{T}}\boldsymbol{\beta} - \sum_{q=1}^{Q-1} \alpha_q \left\{ \boldsymbol{\beta}^{\mathrm{T}}(\mathbf{M}_{q+1} - \mathbf{M}_q) - \rho \right\}. \tag{14}$$

The α_q coefficients are the Lagrange multipliers ($\alpha_q \geq 0$). To find the minimum or maximum of this function, L must be derived with respect to $\boldsymbol{\beta}$, ρ and μ:

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = 0 \Longrightarrow \boldsymbol{\beta} = \frac{1}{2}\mathbf{Z}^{-1} \sum_{q=1}^{Q-1} \alpha_q(\mathbf{M}_{q+1} - \mathbf{M}_q), \tag{15}$$

$$\frac{\partial L}{\partial \rho} = 0 \Longrightarrow \sum_{q=1}^{Q-1} \alpha_q = C, \tag{16}$$

$$\frac{\partial L}{\partial \mu} = 0 \Longrightarrow \boldsymbol{\beta}^{\mathrm{T}}\mathbf{K}\mathbf{L}\mathbf{K}^{\mathrm{T}}\boldsymbol{\beta} = 0. \tag{17}$$

After joining (15), (16), (17) and (13), the final function to optimise is as follows:

$$\min F(\boldsymbol{\alpha}) = \sum_{q=1}^{Q-1} \alpha_q(\mathbf{M}_{q+1} - \mathbf{M}_q)^{\mathrm{T}}(\mathbf{Z}^{-\mathbf{1}})^{\mathrm{T}} \cdot \mathbf{N} \cdot \mathbf{Z}^{-1} \sum_{q=1}^{Q-1} \alpha_q(\mathbf{M}_{q+1} - \mathbf{M}_q) \tag{18}$$

$$\text{s.t. } \alpha_q \geq 0, q \in \{1, \ldots, Q-1\} \text{ and } \sum_{q=1}^{Q-1} \alpha_q = C,$$

being $\mathbf{Z} = \mathbf{H} + \mu\mathbf{K}\mathbf{L}\mathbf{K}^{\mathrm{T}}$.

This optimisation problem is a convex Quadratic Programming (QP) with linear constraints. For the optimisation of the function, we reformulate it in the following canonical form of the QP problems:

$$\min F(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Q}\boldsymbol{\alpha} + \mathbf{c}^{\mathrm{T}}\boldsymbol{\alpha},$$

with the constraints $A\boldsymbol{\alpha} \leq b$ and $E\boldsymbol{\alpha} = d$.

The equation problem (18) can be solved by using the following $\mathbf{Q}$ matrix:

$$Q_{ij} = 2(\mathbf{M}_{i+1} - \mathbf{M}_i)^{\mathrm{T}}(\mathbf{Z}^{-1})^{\mathrm{T}}\mathbf{N}\mathbf{Z}^{-1}(\mathbf{M}_{j+1} - \mathbf{M}_j).$$

Since it is not necessary to use the vector $\mathbf{c}$, we can fill it with zeros. The constraints will be:

$$(-1) \cdot \boldsymbol{\alpha} \leq \mathbf{0},$$
$$\mathbf{1}^{\mathrm{T}} \cdot \boldsymbol{\alpha} = C,$$

where $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_{K-1}\}$, $A \equiv -1$, $b \equiv \mathbf{0}$, $E \equiv \mathbf{1}^{\mathrm{T}}$, $d \equiv C$ and $\mathbf{1}$ and $\mathbf{0}$ represent a vector filled with ones and zeros, respectively.

To solve ill-posed systems, we add a scalar $t > 0$ to the diagonal elements of $(\mathbf{Z}^{-1})^{\mathrm{T}}\mathbf{N}\mathbf{Z}^{-1}$, in the same way than in section 2.3.

After obtaining $\boldsymbol{\beta}$ by substituting α_q into (15), the label for the input vector $\mathbf{x}$ set can be predicted by the following decision rule:

$$f(\mathbf{x}) = \begin{cases} Q, & \text{if } \boldsymbol{\beta} \cdot \Phi(\mathbf{x}) - b_{Q-1} > 0, \\ \max_q\{\boldsymbol{\beta} \cdot \Phi(\mathbf{x}) - b_q < 0\}, & \text{otherwise.} \end{cases}$$

where $b_q = \frac{\boldsymbol{\beta}(\mathbf{M}_{q+1} + \mathbf{M}_q)}{2}$ with $q = 1, ..., Q-1$.

This approach is referred in the experiments to as Semi-supervised Discriminant Learning for ordinal regression (S-DL).

Note that the proposed algorithm involves solving the inversion of a matrix $\mathbf{Z}$ (of size $N \times N$) and a QP problem with a Hessian matrix of size $(K-1) \times (K-1)$. The additional complexity of our algorithm with respect to the original kernel discriminant learning for ordinal regression is then determined by the number of unlabelled patterns (being then both problems of size N). The linear version of this algorithm can also be used, which is a common approach in semi-supervised learning in the presence of abundant data [29].

3.2. Neighbourhood analysis in the Empirical Feature Space

Given that the decision boundary in the approach described in the previous subsection is constructed in the feature space induced by the kernel function, we consider that the distances used for constructing the similarity graph should be computed in this space as well (instead of computing them in the input space). Note that unlabelled knowledge is introduced into our problem via the smoothness and manifold assumptions, which mainly translate to the idea that points nearby are likely to share the same label. Therefore, it is, in this case, safer to assume a relationship between close patterns in the feature space (rather than in the input space) when computing a nonlinear classifier as the one used.

In this sense, the computation of distances using the information of a kernel matrix has been used in different approaches [30]. However, in this paper, we propose a slightly different idea, in order to be able to apply more complex approaches for computing the similarity matrix $\mathbf{S}$. More specifically, we make use of a concept known as the Empirical Feature Space (EFS) [28], which is isomorphic to the original feature space but Euclidean, thus being more easily tractable. Note that, although distances in the feature space can be computed via the kernel matrix, this approach (the use of the empirical feature space) will allow to change the way of performing the neighbourhood analysis (e.g. more sophisticated clustering algorithms can be used).

By definition, a kernel matrix $\mathbf{K}$ can be diagonalised as follows:

$$\mathbf{K}_{(m \times m)} = \mathbf{V}_{(m \times r)} \cdot \boldsymbol{\Lambda}_{(r \times r)} \cdot \mathbf{V}^{\top}_{(r \times m)}, \tag{19}$$

where r is the rank of $\mathbf{K}$, $\boldsymbol{\Lambda}$ is a diagonal matrix containing the r non-zero eigenvalues of $\mathbf{K}$ in decreasing order (*i.e.*, $\lambda_1, \ldots, \lambda_r$), and $\mathbf{V}$ is a matrix consisting

of the eigenvectors associated to those r eigenvalues (*i.e.*, $\mathbf{v}_1, \ldots, \mathbf{v}_r$) in such a way that $\mathbf{K} = \sum_{i=1}^{r} \lambda_i \mathbf{v}_i \mathbf{v}_i^T$. The EFS can be defined as an Euclidean space preserving the dot product information about $\mathcal{H}$ contained in $\mathbf{K}$ (*i.e.*, this space is isomorphic to the embedded feature space $\mathcal{H}$). Since distances and angles of the vectors in the feature space are uniquely determined by dot products, the training data have the same geometrical structure in both the EFS and the feature space. The map from the input space to this r-dimensional EFS is defined as $\Phi_r^e : \mathcal{X} \to \mathbb{R}^r$, where:

$$\Phi_r^e : \mathbf{x}_i \to \mathbf{\Lambda}^{-1/2} \cdot \mathbf{V}^\top \cdot (k(\mathbf{x}_i, \mathbf{x}_1), \ldots, k(\mathbf{x}_i, \mathbf{x}_m))^\top. \tag{20}$$

It can be checked that the kernel matrix of training images obtained by this transformation corresponds to $\mathbf{K}$ [28].

Furthermore, the EFS provides us with the opportunity to limit the dimensionality of the space by choosing the $j \leq r$ dominant eigenvalues (and their associated eigenvectors) to project the data, while maintaining the most important part of the structure of $\mathcal{H}$. One motivation for performing the neighbourhood analysis in the reduced dimensionality EFS is that distances have been proven to be misleading as the data dimensionality increases, making more probable that neighbours are chosen in a random fashion (this is known as the spectral properties phenomenon [31]). In this sense, distances may bear less neighbourhood information as the EFS dimensionality increases [32].

Our proposal for computing the similarity matrix $\mathbf{S}$ is the following:

$$S_{ij} = \begin{cases} 1, & \text{if } \Phi_r^e(\mathbf{x}_i) \in N_n(\Phi_r^e(\mathbf{x}_j)) \text{ or } \Phi_r^e(\mathbf{x}_j) \in N_n(\Phi_r^e(\mathbf{x}_i)), \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

To test the influence of the dimensionality of the empirical feature space we consider two approaches of this idea in the experiments: 1) the approach considering the full-rank EFS (using all dimensions), which is named as Complete Empirical feature space Semi-supervised kernel Discriminant Learning for ordinal regression (CES-DL); and 2) the reduced-rank EFS version (selecting only a subset of the dimensionality of the EFS), which is referred to as: reduced Empirical feature space Semi-supervised kernel Discriminant Learning for ordinal regression (ES-DL).

3.3. Semi-supervised ordinal kernel learning

The method presented in this section considers the optimisation of a given kernel function (and its parameters) to better fit the data. Therefore, it is not a method for classification on its own, but an approach that can be combined with different kernel-based classification strategies. Because of this and given the promising results of our proposed classification algorithm, this method is tested in combination with the approach in 3.1 and using the neighbourhood graph computed in the reduced-rank empirical feature space (which resulted in the best results in the experiments).

As stated before, both the ordinal structure of the data and its unlabelled nature can be taken into account when constructing a suitable kernel that fits our problem. In this sense, different kernel learning approaches can be considered for this purpose (such as kernel-target alignment, previously defined in section 2.4). This section presents a new approach for optimising kernel functions in the presence of both an ordinal structure and unlabelled data. This technique complements the approach in section 3.1 in the sense that it can be used to optimise the kernel parameters (even considering more complex kernel functions) and avoid cross-validation.

In the same vein that a previous work [33], we propose to consider ordinal cost matrices when computing kernel-target alignment, in order to penalise differently misalignment errors. That is, a weighting matrix $\mathbf{W}$ is defined in such a way that $\mathbf{K}^* \circ \mathbf{W}$ imposes a weighting for the different similarity or dissimilarity errors committed, where $\mathbf{A} \circ \mathbf{B}$ represents the hadamard or entrywise product between matrices $\mathbf{A}$ and $\mathbf{B}$. A common choice in ordinal classification for this cost matrix is to use the absolute errors, *i.e.*:

$$w_{ij} = \begin{cases} 1, & \text{if } y_i = y_j, \\ |r(y_i) - r(y_j)|, & \text{otherwise,} \end{cases} \tag{22}$$

where $r(y_j)$ represents the ranking of target y_j associated to pattern $\mathbf{x}_j$ (i.e. $r(\mathcal{C}_q) = q$, $q \in 1, \ldots, Q$).

The algorithm proposed in this section for semi-supervised ordinal kernel learning is based on two different steps:

1. The former is the alignment of the kernel matrix $\mathbf{K}_{\mathrm{L}}$ associated to the labelled patterns with their corresponding ideal kernel $\mathbf{K}_{\mathrm{L}}^*$. This step is used to initialise our algorithm to a viable solution adjusted to the known information. In this case, a viable solution would be a set of kernel parameters that fit the training labelled data. This alignment step is represented by $\mathcal{A}_{\mathrm{L}}$.
2. The latter is based on the adjustment of the kernel parameters using both labelled and unlabelled data. This step is referred to as $\mathcal{A}_{\mathrm{U}}$ and starts using the solution from the previous step. In this case, the ideal kernel $\mathbf{K}_{\mathrm{U}}^*$ is constructed using a different approach:

$$\mathbf{K}_{\mathrm{U}}^* = \begin{bmatrix} \mathbf{K}_{\mathrm{L}}^* \circ \mathbf{W} & \mathbf{S}_{\mathrm{LU}} \\ (\mathbf{S}_{\mathrm{LU}})^{\mathrm{T}} & \mathbf{S}_{\mathrm{UU}} \end{bmatrix}, \tag{23}$$

where $\mathbf{S}_{\mathrm{LU}}$ is the similarity matrix between labelled and unlabelled patterns (computed using (5)) and $\mathbf{S}_{\mathrm{UU}}$ is the similarity matrix between unlabelled patterns. In this case, we set the parameter n associated to the number of nearest neighbours to $\min_{q=1}^{Q} N_q$, as it resulted experimentally in a relatively good performance.

The algorithmic approach followed for optimising the kernel parameters in each step is the one proposed in [34], where the concept of kernel-target alignment

is used, optimising it through a gradient-descent strategy. In this case, the gradient-descent approach is used twice for optimising the kernel: firstly using the ideal supervised knowledge (to set an appropriate initial solution) and secondly using labelled and unlabelled data (to refine the previous solution). A Gaussian multi-scale kernel is used for this purpose, i.e. considering one kernel width per feature to better fit the data. Note again that this proposal is considered in conjunction with the reduced Empirical feature space Semi-supervised kernel Discriminant Learning for ordinal regression (ES-DL), as it was proved experimentally as the best proposal. This approach is referred to as Kernel-target alignment using the Empirical feature space for Semi-supervised ordinal Discriminant Learning (KES-DL).

4. Experiments

In this subsection, we describe the different experiments conducted, including the datasets and algorithms considered, the parameters to optimise, the performance measures and the statistical tests used for assessing the performance differences.

4.1. Datasets

The most widely used ordinal classification dataset repository is the one provided by Chu et al. [16], including different regression benchmark datasets. These datasets are not real ordinal classification ones but regression problems, which are turned into ordinal classification, the target variable being discretised into Q different bins with equal frequency. These datasets do not exhibit some characteristics of typical complex classification tasks, such as class imbalance, given that all classes are assigned the same number of patterns. Because of this, we also compare our proposals with other benchmark ordinal classification datasets.

Table 1 shows the characteristics of the 30 datasets used, including the number of patterns, attributes and classes, and also the number of patterns per class. The real ordinal classification datasets were extracted from benchmark repositories (UCI [35] and `mldata.org` [36]), and the regression ones were obtained from the website of Chu[1]. For the discretised datasets, we considered $Q = 5$ and $Q = 10$ bins to evaluate the response of the classifiers to the increase in the complexity of the problem. All nominal attributes were transformed into binary attributes and all the datasets were properly standardised.

Multiple random splits of the datasets were considered. For discretised regression datasets, 20 random splits were done and the number of training and test patterns were those suggested in [16]. For real ordinal regression problems, 30 random stratified splits with 75% and 25% of the patterns in the training and test sets were considered, respectively. All the partitions were the same for all

[1] `http://www.gatsby.ucl.ac.uk/~chuwei/ordinalregression.html`

Table 1: Characteristics of the benchmark datasets

Discretised regression datasets				
Dataset	#Pat.	#Attr.	#Classes	Class distribution
machine5 (M5)	209	7	5	≈ 42 per class
housing5 (H5)	506	14	5	≈ 101 per class
stock5 (S5)	700	9	5	140 per class
abalone5 (A5)	4177	11	5	≈ 836 per class
computer5 (C5)	8192	12	5	≈ 1639 per class
computer5' (CC5)	8192	21	5	≈ 1639 per class
cal.housing5 (CH5)	20640	8	5	4128 per class
census5 (CE5)	22784	8	5	≈ 4557 per class
machine10 (M10)	209	7	10	≈ 21 per class
housing10 (H10)	506	14	10	≈ 51 per class
stock10 (S10)	700	9	10	70 per class
abalone10 (A10)	4177	11	10	≈ 418 per class
cal.housing (CH10)	20640	8	10	2064 per class
census10 (CE10)	22784	8	10	≈ 2279 per class
Real ordinal regression datasets				
Dataset	#Pat.	#Attr.	#Classes	Class distribution
contact-lenses (CL)	24	6	3	$(15, 5, 4)$
pasture (PA)	36	25	3	$(12, 12, 12)$
squash-stored (SS)	52	51	3	$(23, 21, 8)$
squash-unstored (SU)	52	52	3	$(24, 24, 4)$
tae (TA)	151	54	3	$(49, 50, 52)$
newthyroid (NT)	215	5	3	$(30, 150, 35)$
balance-scale (BS)	625	4	3	$(288, 49, 288)$
SWD (SW)	1000	10	4	$(32, 352, 399, 217)$
car (CA)	1728	21	4	$(1210, 384, 69, 65)$
bondrate (BO)	57	37	5	$(6, 33, 12, 5, 1)$
toy (TO)	300	2	5	$(35, 87, 79, 68, 31)$
eucalyptus (EU)	736	91	5	$(180, 107, 130, 214, 105)$
LEV (LE)	1000	4	5	$(93, 280, 403, 197, 27)$
winequality-red (WR)	1599	11	6	$(10, 53, 681, 638, 199, 18)$
ESL (ES)	488	4	9	$(2, 12, 38, 100,$ $116, 135, 62, 19, 4)$
ERA (ER)	1000	4	9	$(92, 142, 181, 172,$ $158, 118, 88, 31, 18)$

the methods, and one model was trained and evaluated for each split. For every dataset, the percentage considered as unlabelled data corresponds to a stratified 80% of the training patterns, and the remaining 20% correspond to training itself (note that this is a conservative approach compared to other experimental settings in the literature where 5% of the data was considered as labelled [17]). This ratio of labelled and unlabelled data has been chosen given the low amount of patterns for some classes in the datasets considered, where we restrict the selection so that at least one pattern per class is always labelled. Previous literature has shown that only one labelled pattern was needed for performing semi-supervised learning in binary classification problems [37]. However, in multi-class environments the number of needed patterns grows linearly with the number of classes [7].

4.1.1. Performance evaluation and model selection

Different measures can be considered for evaluating ordinal regression models [38]. However, the most common one is the Mean Absolute Error (MAE) [11].

MAE is the average deviation in absolute value of the predicted rank ($\mathcal{O}(y_i^*)$) from the true one ($\mathcal{O}(y_i)$) [38]:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\mathcal{O}(y_i) - \mathcal{O}(y_i^*)|.$$

MAE values range from 0 to $Q-1$ (maximum deviation in number of categories). In this way, the well-known metric MZE considers a zero-one loss for misclassification, while MAE uses an absolute cost. We consider these costs for evaluating the datasets because they are the most common ones (e.g., see [16, 39, 13, 11]).

4.2. Methodologies tested

Different methodologies are tested in the experimental part of this paper. Firstly, we consider a supervised approach to analyse the difference between the supervised and semi-supervised framework. Secondly, we also take into account one of the proposals for semi-supervised ordinal learning in the literature [18], which is also based on kernel discriminant analysis. Other methods in the literature have not been included in the experiments because they are based on other classification paradigms (such as Gaussian processes [17]) or focused on other slightly different settings such as the transductive one [19, 20] or ordinal clustering [21]. Note that the ordinal version of Gaussian processes has been already compared to the ordinal version of kernel discriminant analysis, resulting in worst performance for real ordinal datasets with a much higher computational cost [11]. Finally, we include four different versions of our proposals (S-DL, CES-DL, ES-DL and KES-DL), where the main differences revolve around the space in which the neighbourhood information is computed for the inclusion of unlabelled data and the use of a kernel learning strategy to optimise the kernel parameters. More specifically, the methodologies tested are the following:

- Standard kernel Discriminant Learning for ordinal regression (referred in the experiments to as DL). Unlabelled data is ignored in this case.
- Weighted Semi-supervised kernel Discriminant Learning for ordinal regression (named as WS-DL). Unlabelled data is introduced in the model using the approach in section 2.5 [18].
- Semi-supervised Discriminant Learning for ordinal regression (referred to as S-DL). Unlabelled data is included in the model using the proposal in section 3.1. In this case, the neighbourhood information is computed in the input space, as proposed in [5].
- Complete Empirical feature space Semi-supervised ordinal Discriminant Learning (named as CES-DL). In this case, the neighbourhood graph is constructed in the full-rank empirical feature space induced by the kernel function (see section 3.2 for more information). The rest is optimised using the formulation in 3.1 (i.e. the proposed ordinal and semi-supervised classification framework).

- Reduced Empirical feature space Semi-supervised kernel Discriminant Learning for ordinal regression (named as ES-DL). In this case, the neighbourhood graph is constructed in the reduced-rank empirical feature space, i.e. removing noise when computing distances (see section 3.2 for more information). The rest is optimised using the formulation in section 3.1.

- Kernel-target alignment using the Empirical feature space for Semi-supervised Discriminant Learning in ordinal regression (KES-DL). The new strategy devised for kernel learning and presented in section 3.3 is used in this case and combined with the approach in section 3.1 using the reduced-rank empirical feature space.

All model hyperparameters were selected using a nested five-fold cross-validation over the training set. Once the lowest cross-validation error alternative was obtained, it was applied to the complete training set and test results were extracted. The criteria for selecting the best configuration was MAE. The parameter configurations explored are now specified. The Gaussian kernel function was considered for all the methods. The following values were considered for the width of the kernel, $\sigma \in \{10^{-1}, 10^{0}, 10^{1}\}$ (this hyperparameter was also cross-validated for the case of the label propagation method [24], i.e. WS-DL and MS-DL). The cost parameter C of all methods was fixed to 1. An additional parameter t was also considered to avoid ill-posed matrices. The value considered was 10^{-8}. The parameter a associated to the label propagation method is fixed to 0.99 as done in [24]. The k parameter for the k-nearest neighbour analysis is cross-validated using the values {3,5,7}. The parameter μ associated to the contribution of unlabelled data to the model was cross-validated within the values {0.5, 0.25, 0.1, 0.01}. Finally, the parameter r was fixed to $0.5m$ in the case of ES-DL (as suggested in [32], although this parameter could be cross-validated as well for greater improvement).

4.2.1. Results

Table 2 presents the results obtained for all the methodologies tested. From this table, several conclusions can be extracted: Firstly, that the combination of labelled and unlabelled data results in a more precise and robust model (compare the results for example of S-DL and DL, where S-DL wins a total of 21 out of 30 times compared to DL). Secondly, that the computation of distances in the reduced-rank EFS is satisfactory and leads to better performance. To see this, compare ES-DL against CES-DL (ES-DL wins a total of 28 times with respect to CES-DL), where the full-rank EFS is used, or to S-DL, where the similarity matrix is computed in the input space (ES-DL winning 20 times). Thirdly, it is important to note that our proposal works better than the weight-based proposal in [18] (compare ES-DL against WS-DL, where ES-DL wins in 27 cases). Note however, that the authors of [18] also propose to include an

evolutionary approach which we did not use in this paper[2]. Furthermore, recall that the label propagation method used in [18] is not advisable when there are very few training data, which is the case of some of the datasets used in this paper. Concerning the kernel learning approach (KES-DL), the results are similar to the ES-DL algorithm in most cases, showing this the feasibility of optimising the kernel parameters using both labelled and unlabelled sources of information. Finally, it could be noted that there are a few exceptions, e.g. car and contact-lenses, where the sole use of labelled data is enough to construct an accurate model.

Table 2: MAE mean and standard deviations (Mean ± SD) obtained by all the methodologies compared.

Dataset	DL	WS-DL	S-DL	CES-DL	ES-DL	KES-DL
ERA	1.859 ± 0.138	2.01 ± 0.196	1.845 ± 0.146	1.847 ± 0.146	*1.843 ± 0.147*	**1.765 ± 0.147**
ESL	0.577 ± 0.107	0.633 ± 0.056	*0.449 ± 0.068*	0.465 ± 0.068	0.452 ± 0.064	**0.429 ± 0.178**
LEV	0.57 ± 0.06	*0.543 ± 0.051*	0.555 ± 0.05	0.564 ± 0.049	0.556 ± 0.05	**0.511 ± 0.046**
SWD	0.591 ± 0.052	*0.548 ± 0.054*	0.585 ± 0.061	0.589 ± 0.048	0.586 ± 0.047	**0.547 ± 0.057**
abalone	0.994 ± 0.053	0.93 ± 0.052	0.798 ± 0.037	0.946 ± 0.047	*0.797 ± 0.044*	**0.796 ± 0.041**
abalone10	2.033 ± 0.082	2.084 ± 0.148	2.05 ± 0.121	1.768 ± 0.152	*1.750 ± 0.144*	*1.767 ± 0.163*
balance-scale	0.278 ± 0.037	0.335 ± 0.054	0.244 ± 0.038	0.243 ± 0.035	0.239 ± 0.037	**0.156 ± 0.045**
bondrate	0.678 ± 0.164	0.744 ± 0.076	0.671 ± 0.14	0.769 ± 0.293	*0.662 ± 0.115*	**0.638 ± 0.160**
calhousing-10	2.193 ± 0.234	2.536 ± 0.233	2.306 ± 0.27	2.046 ± 0.21	**2.005 ± 0.207**	*2.018 ± 0.313*
calhousing-5	1.085 ± 0.119	1.136 ± 0.125	1.099 ± 0.103	*0.966 ± 0.078*	**0.965 ± 0.091**	0.992 ± 0.112
car	**0.137 ± 0.026**	0.421 ± 0.019	0.233 ± 0.021	0.316 ± 0.035	0.271 ± 0.032	*0.223 ± 0.031*
census1-10	2.302 ± 0.278	2.461 ± 0.256	2.341 ± 0.237	2.147 ± 0.239	**2.017 ± 0.187**	*2.142 ± 0.341*
census1-5	1.101 ± 0.084	1.127 ± 0.113	1.096 ± 0.074	*1.004 ± 0.110*	**0.937 ± 0.149**	1.02 ± 0.112
computer1-5	0.972 ± 0.211	1.018 ± 0.159	0.937 ± 0.146	0.882 ± 0.146	**0.855 ± 0.137**	*0.862 ± 0.192*
computer2-5	0.896 ± 0.325	1.099 ± 0.118	1.007 ± 0.171	0.956 ± 0.199	**0.834 ± 0.158**	*0.885 ± 0.303*
contact-lenses	**0.689 ± 0.213**	0.806 ± 0.334	0.767 ± 0.308	0.739 ± 0.222	*0.694 ± 0.304*	0.717 ± 0.33
eucalyptus	0.932 ± 0.058	1.059 ± 0.142	0.883 ± 0.09	0.966 ± 0.131	1.005 ± 0.081	*0.907 ± 0.098*
housing	0.793 ± 0.152	0.821 ± 0.094	0.705 ± 0.127	0.623 ± 0.082	**0.615 ± 0.082**	*0.622 ± 0.081*
housing10	1.508 ± 0.408	1.942 ± 0.181	1.494 ± 0.153	1.388 ± 0.185	**1.224 ± 0.179**	*1.359 ± 0.290*
machine	0.949 ± 0.177	0.948 ± 0.14	*0.621 ± 0.132*	0.635 ± 0.151	**0.576 ± 0.140**	0.642 ± 0.293
machine10	1.938 ± 0.382	2.364 ± 0.333	*1.542 ± 0.291*	1.543 ± 0.436	**1.314 ± 0.436**	1.795 ± 0.579
newthyroid	0.198 ± 0.088	0.241 ± 0.044	*0.100 ± 0.041*	0.112 ± 0.069	**0.078 ± 0.057**	0.121 ± 0.075
pasture	0.659 ± 0.028	0.596 ± 0.181	0.581 ± 0.129	0.604 ± 0.174	*0.556 ± 0.226*	**0.485 ± 0.161**
squash-stored	0.538 ± 0.064	0.523 ± 0.124	0.515 ± 0.126	*0.503 ± 0.140*	0.536 ± 0.127	**0.495 ± 0.132**
squash-unstored	0.523 ± 0.042	0.521 ± 0.144	**0.436 ± 0.130**	0.477 ± 0.145	0.474 ± 0.123	*0.464 ± 0.127*
stock	0.27 ± 0.035	0.291 ± 0.033	**0.190 ± 0.022**	0.206 ± 0.029	*0.199 ± 0.031*	0.218 ± 0.021
stock10	0.603 ± 0.085	0.646 ± 0.06	**0.413 ± 0.032**	0.425 ± 0.038	*0.415 ± 0.038*	0.557 ± 0.066
tae	0.673 ± 0.096	0.725 ± 0.128	0.702 ± 0.113	0.675 ± 0.111	*0.660 ± 0.091*	**0.636 ± 0.084**
toy	0.181 ± 0.059	0.544 ± 0.141	0.312 ± 0.047	*0.154 ± 0.049*	**0.152 ± 0.043**	0.157 ± 0.084
winequality-red	0.562 ± 0.029	0.567 ± 0.042	*0.536 ± 0.057*	0.574 ± 0.062	0.544 ± 0.036	**0.524 ± 0.074**
Average	0.909	1.007	0.867	0.838	**0.794**	*0.815*
Ranking	4.533	5.400	3.267	3.633	**2.067**	*2.100*
Friedman's test	Confidence interval $C_0 = (0, F_{(\alpha=0.05)} = 2.28)$. F-val.$_{MAE}$: $29.10 \notin C_0$.					

The best performing method is in **bold** face and the second one in *italics*.

To quantify whether a statistical difference exists among the algorithms compared, a procedure is employed to compare multiple classifiers in multiple datasets [40]. Table 2 also shows the result of applying the non-parametric statistical Friedman's test (for a significance level of $\alpha = 0.05$) to the mean MAE rankings. It can be seen that the test rejects the null-hypothesis that all of the algorithms perform similarly in mean ranking for this metric.

On the basis of this rejection and following the guidelines in [40], we consider the best performing methods in MAE (i.e., S-DL, ES-DL and KES-DL) as control methods for the following tests. We compare these three methods to the rest according to their rankings. It has been noted that the approach of comparing all classifiers to each other in a post-hoc test is not as sensitive as

[2]We consider that using an evolutionary algorithm might not be a fair comparison given the nature and the very high computational cost associated to such a method.

the approach of comparing all classifiers to a given classifier (a control method). One approach to this latter type of comparison is the Holm's test. The test statistics for comparing the i-th and j-th method using this procedure is:

$$z = \frac{R_i - R_j}{\sqrt{\frac{J(J+1)}{6T}}},$$

where J is the number of algorithms, T is the number of datasets and R_i is the mean ranking of the i-th method. The z value is used to find the corresponding probability from the table of the normal distribution, which is then compared with an appropriate level of significance α. Holm's test adjusts the value for α in order to compensate for multiple comparisons. This is done in a step-up procedure that sequentially tests the hypotheses ordered by their significance. We will denote the ordered p-values by $p_1, p_2, \ldots, p_q$ so that $p_1 \leq p_2 \leq \ldots \leq p_q$. Holm's test compares each p_i with $\alpha^*_{\text{Holm}} = \alpha/(J-i)$, starting from the most significant p value. If p_1 is below $\alpha/(J-1)$, the corresponding hypothesis is rejected and we allow to compare p_2 with $\alpha/(J-2)$. If the second hypothesis is rejected, the test proceeds with the third, and so on.

Table 3 presents the results of applying the Holm's test, where different conclusions can be drawn. First, the base proposal (i.e., the S-DL algorithm) significantly improves the result of DL (the sole use of labelled data) and other algorithms in the ordinal semi-supervised literature (e.g. WS-DL). However, it also presents a significant lower performance than other proposals of this paper (more specifically, KES-DL and ES-DL). The computation of pattern similarities in the input space might be, in general, beneficial but it can also be improved by the use of the EFS (analyse the results obtained for ES-DL and KES-DL). When comparing ES-DL and KES-DL no differences are found. In this case, our recommendation would be to use ES-DL for large-scale data instead of KES-DL, as the kernel optimisation phase is time-consuming (or limiting the number of parameters to optimise via kernel-target alignment).

Previous research has shown that the performance gap between semi-supervised approaches and standard ones grows as the number of unlabelled patterns increases and the number of labelled ones decreases [17]. To test this, we analyse the performance gap obtained under two circumstances: 1) 20% labelled and 80% unlabelled data (i.e. the results included in Table 2) and 2) 10% labelled and 90% unlabelled data. The methods chosen for this comparison are the supervised approach DL (which ignores unlabelled data) and ES-DL (the best performing method of the ones tested, which makes use of unlabelled data to complement the model). For the sake of simplicity, we only consider the performance gap in mean for the 30 datasets considered, i.e. $\sum_{i=1}^{30} MAE_i^{DL} - MAE_i^{ES-DL}$, where the subscript i refers to the dataset. The results of these experiments are the following: When using a 20%-80% labelled-unlabelled ratio the performance gap is 0.119, whereas using the ratio 10%-90% it is 0.240, which indicates that greater improvement could be expected from our proposal with respect to the supervised approach in circumstances where the ratio of unlabelled patterns grows with respect to labelled ones.

Table 3: Results of the Holm test using S-DL, ES-DL and KES-DL as control methods: corrected α values, compared method and resulting p-values, ordered by number of comparison (i).

Control alg.: S-DL			MAE	
i	$\alpha^*_{0.05}$	$\alpha^*_{0.10}$	Method	p_i
1	0.01000	0.02000	WS-DL	0.00001_{++}
2	0.01250	0.02500	DL	0.00874_{++}
3	0.01667	0.03333	ES-DL	0.01298_{--}
4	0.02500	0.05000	KES-DL	0.01573_{--}
5	0.05000	0.10000	CES-DL	0.44782
Control alg.: ES-DL			MAE	
i	$\alpha^*_{0.05}$	$\alpha^*_{0.10}$	Method	p_i
1	0.01000	0.02000	WS-DL	0.00000_{++}
2	0.01250	0.02500	DL	0.00000_{++}
3	0.01667	0.03333	CES-DL	0.00118_{--}
4	0.02500	0.05000	S-DL	0.01298_{--}
5	0.05000	0.10000	KES-DL	0.94499
Control alg.: KES-DL			MAE	
i	$\alpha^*_{0.05}$	$\alpha^*_{0.10}$	Method	p_i
1	0.01000	0.02000	WS-DL	0.00000_{++}
2	0.01250	0.02500	DL	0.00000_{++}
3	0.01667	0.03333	CES-DL	0.00150_{--}
4	0.02500	0.05000	S-DL	0.01573_{--}
5	0.05000	0.10000	ES-DL	0.94499

Statistically significant win ($++$) or lose ($--$) for $\alpha = 0.05$

5. Conclusions

This paper presents a new classification strategy for incorporating semi-supervised information into the ordinal version of discriminant learning. This source of knowledge is included via the smoothness and manifold assumptions, commonly used for semi-supervised learning. To do so, a neighbourhood analysis of the data is conducted, via distances in the input space and the feature space induced by a kernel function. Finally, a kernel learning strategy is also proposed for optimising the kernel parameters using both labelled and unlabelled sources of information in the context of ordinal classification problems. Our experiments show (1) that in the presence of unlabelled data, a semi-supervised approach is usually preferred over the fully-supervised one (even when very few data is available), (2) that the ordinal version of discriminant learning can be successfully adapted to deal with unlabelled data, (3) that the analysis of distances in the feature space is usually preferred for semi-supervised kernel algorithms when performing a neighbourhood analysis and (4) that a kernel function (or in this case, the kernel parameters) can be easily optimised using supervised and unsupervised knowledge.

As future work, we plan to explore more options for the construction of the neighbourhood graph (e.g. density-based clustering algorithms for detecting outliers) and adapt the label propagation strategy for the specific case of ordinal classification.

Acknowledgements

This work has been partially subsidized by the TIN2014-54583-C2-1-R project of the Spanish Ministerial Commission of Science and Technology (MINECO), FEDER funds and the P11-TIC-7508 project of the "Junta de Andalucía" (Spain).

References

[1] J. Hernández-González, I. Inza, J. A. Lozano, Weak supervision and other non-standard classification problems: a taxonomy, Pattern Recognition Letters 69 (2016) 49 – 55.

[2] X. Zhu, Semi-supervised learning literature survey, Tech. Rep. 1530, Computer Sciences, University of Wisconsin-Madison (2005).
URL http://pages.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf

[3] O. Chapelle, B. Schölkopf, A. Zien, Semi-Supervised Learning, 1st Edition, The MIT Press, 2010.

[4] J. Wang, X. Shen, W. Pan, On efficient large margin semisupervised learning: Method and theory, Journal of Machine Learning Research 10 (2009) 719–742.

[5] D. Cai, X. He, J. Han, Semi-supervised discriminant analysis, in: IEEE 11th International Conference on Computer Vision, 2007, pp. 1–7.

[6] I. Cohen, F. G. Cozman, N. Sebe, M. C. Cirelo, T. S. Huang, Semisupervised learning of classifiers: theory, algorithms, and their application to human-computer interaction, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (2004) 1553– 1566.

[7] J. Ortigosa-Hernández, I. Inza, J. A. Lozano, Semisupervised multiclass classification problems with scarcity of labeled data: A theoretical study, IEEE Transactions on Neural Networks and Learning Systems Accepted (99) (2016) 1–13. doi:10.1109/TNNLS.2015.2498525.

[8] R. Xu, G. C. Anagnostopoulos, D. C. Wunsch, Multiclass cancer classification using semisupervised ellipsoid ARTMAP and particle swarm optimization with gene expression data, IEEE/ACM Trans. Comput. Biology Bioinform. 4 (1) (2007) 65–77.

[9] R. G. Soares, H. Chen, X. Yao, Semisupervised classification with cluster regularization, IEEE Transactions on Neural Networks and Learning Systems 23 (11) (2012) 1779–1792.

[10] J. Ortigosa-Hernández, J. D. Rodríguez, L. Alzate, M. Lucania, I. Inza, J. A. Lozano, Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers, Neurocomputing 92 (2012) 98 – 115.

[11] P. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernández-Navarro, C. Hervás-Martínez, Ordinal regression methods: survey and experimental study, IEEE Transactions on Knowledge and Data Engineering 28 (1) (2016) 127–146.

[12] R. Likert, A technique for the measurement of attitudes, Archives of Psychology 22 (140).

[13] B.-Y. Sun, J. Li, D. D. Wu, X.-M. Zhang, W.-B. Li, Kernel discriminant learning for ordinal regression, IEEE Transactions on Knowledge and Data Engineering 22 (2010) 906–910.

[14] A. Shashua, A. Levin, Ranking with large margin principle: Two approaches, 2003.

[15] P. McCullagh, J. A. Nelder, Generalized Linear Models, 2nd Edition, Monographs on Statistics and Applied Probability, Chapman & Hall/CRC, 1989.

[16] W. Chu, Z. Ghahramani, Gaussian processes for ordinal regression, Journal of Machine Learning Research 6 (2005) 1019–1041.

[17] P. Srijith, S. Shevade, S. Sundararajan, Semi-supervised gaussian process ordinal regression, in: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases, Vol. 8190 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2013, pp. 144–159.

[18] Y. Wu, Y. Sun, X. Liang, K. Tang, Z. Cai, Evolutionary semi-supervised ordinal regression using weighted kernel fisher discriminant analysis, in: IEEE Congress on Evolutionary Computation (CEC), 2015, pp. 3279–3286.

[19] C.-W. Seah, I. W. Tsang, Y.-S. Ong, Transductive ordinal regression., IEEE Transactions Neural Networks and Learning Systems 23 (7) (2012) 1074–1086.

[20] Y. Liu, Y. Liu, S. Zhong, K. C. Chan, Semi-supervised manifold ordinal regression for image ranking, in: Proceedings of the 19th ACM International Conference on Multimedia, ACM, New York, NY, USA, 2011, pp. 1393–1396.

[21] Y. Xiao, B. Liu, Z. Hao, A maximum margin approach for semisupervised ordinal regression clustering, IEEE Transactions on Neural Networks and Learning Systems 27 (5) (2016) 1003–1019.

[22] N. Cristianini, J. Kandola, A. Elisseeff, J. Shawe-Taylor, On kernel-target alignment, in: Advances in Neural Information Processing Systems 14, MIT Press, 2002, pp. 367–373.

[23] C. Cortes, M. Mohri, A. Rostamizadeh, Algorithms for learning kernels based on centered alignment, Journal of Machine Learning Research 13 (2012) 795–828.

[24] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, B. Schölkopf, Learning with local and global consistency, in: S. Thrun, L. Saul, B. Schölkopf (Eds.), Advances in Neural Information Processing Systems 16, MIT Press, 2004, pp. 321–328.

[25] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, in: Advances in Neural Information Processing Systems 14, MIT Press, 2001, pp. 585–591.

[26] F. R. K. Chung, Spectral Graph Theory, American Mathematical Society, 1997.

[27] M. Pérez-Ortiz, P. A. Gutiérrez, C. R. García-Alonso, L. Salvador-Carulla, J. A. Salinas-Perez, C. Hervás-Martínez, Ordinal classification of depression spatial hot-spots of prevalence, in: 11th International Conference on Intelligent Systems Design and Applications, 2011, pp. 1170–1175.

[28] B. Schölkopf, S. Mika, C. J. C. Burges, P. Knirsch, K.-R. Müller, G. Rätsch, A. J. Smola, Input space versus feature space in kernel-based methods, IEEE Transactions on Neural Networks 10 (1999) 1000–1017.

[29] V. Sindhwani, S. S. Keerthi, Large scale semi-supervised linear svms, in: Proceedings of the 29th Annual International Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2006, pp. 477–484.

[30] B. Schölkopf, The kernel trick for distances, TR MSR 2000-51, Microsoft Research, Redmond, WA, advances in Neural Information Processing Systems (2000).

[31] M. Ledoux, The Concentration of Measure Phenomenon, Mathematical surveys and monographs, American Mathematical Society, 2005.

[32] M. Pérez-Ortiz, P. A. Gutiérrez, P. Tino, C. Hervás-Martínez, Oversampling the minority class in the feature space, IEEE Transactions on Neural Networks and Learning Systems (Accepted on 19th July 2015) (99) (2016) 1–1. doi:10.1109/TNNLS.2015.2461436.

[33] M. Pérez-Ortiz, P. Gutiérrez, M. Cruz-Ramírez, J. Sánchez-Monedero, C. Hervás-Martínez, Kernelising the proportional odds model through kernel learning techniques, Neurocomputing 164 (C) (2015) 23–33.

[34] M. Pérez-Ortiz, P. Gutiérrez, J. Sánchez-Monedero, C. Hervás-Martínez, A study on multi-scale kernel optimisation via centered kernel-target alignment, Neural Processing Letters (2015) 1–27doi:10.1007/s11063-015-9471-0.

[35] A. Asuncion, D. Newman, UCI machine learning repository (2007).
URL http://www.ics.uci.edu/~mlearn/MLRepository.html

[36] PASCAL, Pascal (pattern analysis, statistical modelling and computational learning) machine learning benchmarks repository (2011).
URL http://mldata.org/

[37] V. Castelli, T. M. Cover, On the exponential value of labeled samples, Pattern Recognition Letters 16 (1) (1995) 105–111.

[38] S. Baccianella, A. Esuli, F. Sebastiani, Evaluation measures for ordinal regression, in: Proceedings of the Ninth International Conference on Intelligent Systems Design and Applications, 2009, pp. 283–287.

[39] W. Chu, S. S. Keerthi, Support vector ordinal regression, Neural Computation 19 (3) (2007) 792–815.

[40] J. Demšar, Statistical comparisons of classifiers over multiple data sets, Journal of Machine Learning Research 7 (2006) 1–30.