Ordinal Regression as Structured Classification

Niall Twomey, Rafael Poyiadzi, Callum Mann, and Raúl Santos-Rodríguez

University of Bristol, Bristol, United Kingdom
{niall.twomey, rp13102, cm13558, enrsr}@bristol.ac.uk

Abstract. This paper extends the class of Ordinal Regression (OR) models with a structured interpretation of the problem by applying a novel treatment of encoded labels. The net effect of this is to transform the underlying problem from an OR task to a (structured) classification task which we solve with conditional random fields, thereby achieving a coherent and probabilistic model in which all model parameters are jointly learnt. Importantly, we show that although we have cast OR to classification, our method still fall within the class of decomposition methods in the OR ontology. This is an important link since our experience is that many applications of machine learning to healthcare ignores completely the important nature of the label ordering, and hence these approaches should considered naïve in this ontology. We also show that our model is flexible both in how it adapts to data manifolds and in terms of the operations that are available for practitioner to execute. Our empirical evaluation demonstrates that the proposed approach overwhelmingly produces superior and often statistically significant results over baseline approaches on forty popular OR models, and demonstrate that the proposed model significantly out-performs baselines on synthetic and real datasets. Our implementation, together with scripts to reproduce the results of this work, will be available on a public GitHub repository.

1 Introduction

OR is the task of learning to classify data-points into one of many interval classes. It can be understood as lying in between the canonical problems of classification and regression, as it is a classification task where the classes follow a pre-defined order. Model learning in these domains therefore requires particular care and attention since many assumptions underpinning standard classifiers are unsuitable in OR settings.

Let us consider Alzheimer's disease (AD) as a motivating application for this work. When assessing the current state of AD, healthcare professionals utilise one of several well-known assessment questionnaires (*c.f.* [1]). These questionnaires are designed to uncover the cognitive capacity of the persons and evaluate the risks of independent living. An emerging application area of machine learning has been to non-invasively predict questionnaire scores based on a person's behaviour and circadian patterns of Activities of Daily Living (ADL) and Instrumental ADL (IADL) in a Smart Home (SH) [2,3] or to assess the cognitive ability from

conversation analysis. These are challenging problems to model, and there has been some success in these areas already.

The standard machine learning approach is based on learning a mapping between samples and categories so that the probability of error is minimised. However, in the setting described here the categorisation of the scores into their groups is an ordinal operation (*e.g.* 'severe' diagnoses are more extreme than 'moderate' and 'mild'), and indeed classifying a person with 'severe' AD as 'mild' is more costly than predicting 'moderate'. Automated AD assessment presents an opportunity to produce valuable healthcare technology that can benefit vulnerable persons and their families, but also to benefit clinicians via the unprecedented and objective view into the effect of AD on routine and behaviour. Although in the authors' experience the vast majority of the experimental literature on ordinal medical domains ignores the ordinal nature of the data and recasts the problem into traditional binary or multiclass problems, with some notable exceptions [4].

In this work we introduce a structural interpretation of ordinal regression. The advantage of this interpretation is that significantly more flexibility is ascribed to the predictive model, and this flexibility permits the model to operate efficiently on linear and non-linear data manifolds, while the baseline methods considered were unable achieve this. Additionally, our structured interpretation captures contextual information that the other baselines cannot.

The aims and contributions of this paper are as follows: We strongly advocate the selection of ordinal techniques for ordinal problems and a review of ordinal approaches in Section 2. We extend the class of ordinal regression models in this work with a new structural interpretation of the field (Section 3), outline empirical experiments (Section 4) demonstrate its utility in our results (Section 5). We summarise and conclude in Sections 6 and 7.

2 Ordinal Regression

Within the published area of OR, there are several methodologies that are well established. We describe these with strong reference to the 'ordinal regression ontology' from [5] and then introduce the proposed approach after.

2.1 Naïve Models

Intuitively we can reduce an OR task to either a classification or a regression problem. In the case of classification, we ignore the nature of the classes, and proceed with a model that uses nominal classification. This is considered a naïve approach as the practitioner ignores prior knowledge (*e.g.* of class ordering) that could otherwise be used to increase the accuracy and predictive power of the model. For the case of regression, one may map the classes on the real line, employ regression techniques, then map back to the original classes. Unless the practitioner has a well considered way of computing the forward and backward mappings, this approach appears naïve.

A similar approach to the classification reduction, but more advanced, is that of Cost Sensitive Classification (CSC). CSC is a general treatment of models where the practitioner provides (potentially) unique penalties for each type of misclassification [6]. This is usually accomplished through the use of a cost matrix during learning. CSC can therefore be employed for OR by devising a cost matrix that depends on the *distance* between classes [7]. This would again be a sensible approach given that the practitioner has a good understanding of the *distances* between classes, and a principled way of transforming them to suitable costs.

2.2 Threshold Models

Threshold models are another approach to OR. We assume that there is a latent continuous random variable that gives rise to the observed discrete classes. With this formulation we can perform a reduction to a regression problem. As criticised earlier this would be a naïve approach due to the lack of principled way of mapping from the real line to the given classes. Approaches under the Threshold Models category, aim to surpass this limitation by learning this map, or where to 'cut' the real line from data, as opposed to assuming knowledge of it, *a priori*.

Ordered Logit: The classical ordered logit model [8] is a simple model that assumes a real-valued latent variable (y^*) is defined by

$$y^* = \mathbf{w}^\top \mathbf{x} + \epsilon \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^D$ is a data point, D is the dimensionality of the data, $\mathbf{w} \in \mathbb{R}^D$ is a weight vector, and ϵ is a noise term following the logistic distribution with zero mean and unit variance. Assuming K categories, and a set of $K+1$ thresholds $\theta_k \in \{\theta_0, \theta_1, \ldots, \theta_K\}$ (ordered by $\theta_k < \theta_{k+1}$) one can assign a response y according to the interval into which y^* falls with the function $f_k : \mathbb{R} \to \{0, 1\}$:

$$f_k(y^*) = \begin{cases} 1 & \text{if } \theta_{k-1} < y^* \leq \theta_k \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Three of the thresholds are fixed ($\theta_0 = -\infty$, $\theta_1 = 0$ and $\theta_K = \infty$) to ensure that the process is identifiable [9]. The probability over the categories is computed by integrating the probability mass that falls between the intervals

$$\begin{aligned} P(y = k|\mathbf{x}) &= P(\theta_{k-1} < y^* \leq \theta_k|\mathbf{x}) \\ &= \sigma(\theta_k - \mathbf{w}^\top \mathbf{x}) - \sigma(\theta_{k-1} - \mathbf{w}^\top \mathbf{x}) \end{aligned} \tag{3}$$

where $\sigma(\cdot)$ is the logistic function (*i.e.* cumulative distribution of the logistic distribution). The log-likelihood and its gradient with respect to the parameters ($\{\mathbf{w}, \theta_2, \theta_3, \ldots, \theta_{K-1}\}$) are easily computed and can be optimised with standard

optimisation techniques [8]. Previous work presents an approach based on the Support Vector Machine and a dataset constructed by considering all the pairwise difference vectors [10]. One of the main advantages of these models over simpler baselines (such as linear regression) is that the ordinal intervals are optimised during the learning routine and that the intervals can have arbitrary widths. It is important to understand that the primary assumption underpinning these models is that the data lies on a linear manifold, and in practice this is difficult to guarantee. Other approaches within the threshold models category include an adaptation of the online perceptron algorithm [11], as well as an approach based on a generative model, which uses Gaussian Processes [12].

2.3 Decomposition Models

Ordinary Binary Decompositions: products of multiple binary models, or, single models capable of multiple-outputs. For example, in multi-class classification problems, one usually resorts to solving multiple smaller problems and then combining their predictions according to voting schemes such as One-Versus-One (OvO) or One-Versus-All (OvA). Considering a problem with K classes, in the former setting, one would need $K(K-1)/2$ 'small' learners, while in the latter K 'larger' learners, where the distinction between small and large refers to the average size of the data they will be dealing with. OvA is also susceptible to the problem of class-imbalance. Based on the assumption of the ordering of the classes one could construct more developed voting schemes, that reflect his prior knowledge and reduce the computational complexity of the overall algorithm. Examples of such ordinal voting schemes include, *one-vs-next*, *one-vs-followers*, and decompositions based on *Ordered Partitions* (see Section 3.2. in [5]).

These decompositions are closely related to the concept of Error Correcting Output Codes (ECOC), which is used to reduce multi-class classification problems to combinations of binary tasks [13]. In this setting, every class is assigned to an 'output code', which usually contains values in $\{-1, 0, +1\}^Q$. When considering multiple binary models, each of the Q entries of this output code is generated by one of the models. The predicted class is the one whose output code is closer to the composition of predictions. A similar line of work keeps the connection between classes and output codes, but instead of training one model per 'bit', trains a model capable of multiple outputs on the whole code. In the simplest case this could amount to the output codes being of the form of the popular one-hot embedding, but ECOC provides a framework for more delicate codes to be utilised, such as ones reflecting the prior knowledge of the classes being ordered.

Nested Binary Classifiers: A flexible ordinal model based on a decomposition of the label space can be produced with cascades of linear classifiers [14] by recasting the ordinal task into $K-1$ independent binary classification problems. The k-th binary problem re-partitions the dataset into two groups; the first group consists of all instances whose label is less than or equal to the value k,

and the second group consists of all instances with label greater than the value k.

Using an equivalent rationale to that on Equation (3), the probability distributions over each partition are unified into a probability distribution over the K categories with the following equation

$$P(Y = k|\mathbf{x}) = P(Y > k - 0.5|\mathbf{x}) - P(Y > k + 0.5|\mathbf{x}) \tag{4}$$

with the base cases $P(y > 0.5|\mathbf{x}) \triangleq 1$ and $P(y > K|\mathbf{x}) \triangleq 0$. The $K - 1$ models are learnt independently, and only the two classifiers that 'neighbour' the correct label are used in prediction.

Although this model is simple and derived from an intuitive standpoint, it also carries several disadvantages. Firstly, the $K - 1$ binary classifiers are learnt independently. While this brings gains in terms of concurrently learning each model it is unlikely that the final model will produce optimal decisions. Secondly, the mechanism for decision making shown in Equation 4 cannot guarantee consistency in classification and in general may require clipping and renormalisation for probabilistic predictions [15], and this is particularly clear if one envisages an ordinal classification task when the data lies along complex or nonlinear data manifolds.

In the taxonomy of algorithms presented in [5], the *Ordinary Binary Decompositions* category has another sub-class of methods. Therein, a first group of methods takes advantage of the ordinal nature of the classes to devise clever decompositions, while the second group transforms the problem to a multi-target one, with ordinal encodings as targets. Models must be aware of the structured nature of the output space in order to take advantage of these encodings.

3 Methods

In this section we introduce our proposed technique for ordinal regression Structured Ordinal Regression Modeling (StORM). We cast the ordinal regression task into a structured classification task. We use a simple encoding scheme for the labels which allows for a simple propagation of information through a CRF constructed from the label representation. Although classification methods in general are considered naïve on the ordinal regression ontology (*c.f.* [5] and Section 2.1) the proposed method is further developed (and hence not naïve) since the label encoding deliberately captures several desirable properties of ordinal predictors. A key advantage of the application of Conditional Random Fields (CRFs) to the encodings above is that one model is produced and optimised to produce outputs, in contrast to many approaches from the threshold and decomposition strand of the OR ontology.

3.1 Label Encoding

A key enabler of the proposed approach is the symbiotic relationship between a bespoke encoding scheme for ordinal variables on one hand and the modelling

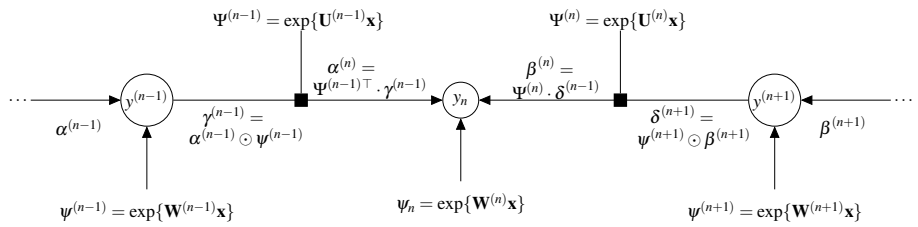


Fig. 1. A graphical illustration of marginalisation process for CRFs. Notation is defined in Section 3.2.

framework that is used to infer and predict on the space of encoded labels on the other (next section). The encoding scheme that we use has previously been introduced for capturing resemblance measures for ordinal variables [16, Ch. 8] but we believe we are the first that incorporate this representation directly into the modelling framework.

We consider an ordinal problem as having K categories, and our encoding scheme transforms these into a sequence of $K-1$ binary digits. The following function defines the value of the k-th bit of an encoded sequence:

$$\widehat{f}_K(\widehat{y}, k) = \begin{cases} 1 & \text{if } k < \widehat{y} \qquad (1 \le k \le K-1) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where the function subscript defines the support of the ordinal categories (*i.e.* K) and $\widehat{y}$ ($1 \le \widehat{y} \le K$) is the 'raw' (*i.e.* un-encoded) label. As a concrete example, for $K = 7$ and $\widehat{y} = 4$, the encoded label $\mathbf{y}$ is given as:

$$f_7(4) = (1, 1, 1, 0, 0, 0) \tag{6}$$

where we have defined the new function

$$f_K(\widehat{y}) = \left(\widehat{f}_K(\widehat{y}, 1), \widehat{f}_K(\widehat{y}, 2), \ldots, \widehat{f}_K(\widehat{y}, K-1)\right) \tag{7}$$

To motivate this encoding scheme for OR, consider two instances with $\widehat{y}^{(1)} = 3$ and $\widehat{y}^{(2)} = 5$ and their encoded values:

$$f_7(3) = (1, 1, 0, 0, 0, 0) \tag{8}$$
$$f_7(5) = (1, 1, 1, 1, 0, 0) \tag{9}$$

Recalling that these are the encoded representation of the labels of two instances, we can see that even though the raw labels are distinct that four bits of

the encoded labels are of the same value. Thus, we can split the encoded labels into three virtual segments: 1) the first two bits which are positive and identical; 2) the final two bits which are negative and identical; and 3) the middle two bits which disagree and encode the intrinsic differences between the instances. In the next section we introduce a framework for modelling sequences of data that obey the constraints of the encoding and thus capture 'shared' and 'distinct' aspects of the encoded labels above.

3.2 Conditional Random Fields

We utilise the language of probabilistic modelling and Conditional Random Fields (CRFs) in our setting. CRFs constitute a structured modelling framework [17], and in this section we motivate and introduce a generalisation of the traditional linear-chain CRFs for OR. Linear-chain CRFs incorporate weight-sharing on all positions of a sequence since, for these models, the dynamics (*i.e.* predictive response as a function of input) are stationary [18]. In other words the effect of one feature is equal at all positions of a sequence. This is a strong assumption, but in particular is inappropriate with our encoded labels since a feature may need to have diminishing (or increasing) responses depending on the position of the sequence. For the remainder of this section we assume the reader has familiarity with CRFs and recommend the following as an introducton: [19].

To overcome this incompatibility, we use the CRF framework with but importantly without weight sharing. We have a dataset that consists of N observations of dimensionality D, *i.e.* $\mathbf{X} \in \mathbb{R}^{N \times D}$. With K ordinal quantities the encoded labels are $\mathbf{Y} \in \{0, 1\}^{N \times (K-1)}$. In order to simplify mathematical notation for the remainder of this section we focus on one particular example/label pair ($\mathbf{x}$, $\mathbf{y}$) which can be considered as the i-th row of $\mathbf{X}$ and $\mathbf{Y}$ respectively. Of critical importance for this method is the fact that the label has been mapped from the 'one-of-K' encoding to the 'up-to-k' encoding, and hence the space of labels (and predictions) have become a sequence of binary variables for every instance. Although this might be viewed as an unnecessary complication (since no new information is introduced) we will later see the value that is introduced by this encoding.

CRFs for OR CRFs yield structured predictions over graphs. In our setting, the graph consists of $K-1$ nodes with $K-2$ edges linking the nodes together in a chain. Each node (indexed by n) contains its own set of weights as does each edge (indexed by e). We follow standard potential and marginalisation methods from the CRF literature. First, node and edge potentials are computed. The n-th node potential is given by

$$\boldsymbol{\psi}^{(n)} = \exp\{\mathbf{W}^{(n)}\mathbf{x}\} \tag{10}$$

where $\mathbf{x} \in \mathbb{R}^D$ is the feature vector and $\mathbf{W}^{(n)} \in \mathbb{R}^{2 \times D}$ is the weight vector associated with the n-th node, and $\boldsymbol{\psi}^{(n)} \in \mathbb{R}^2 \ \forall \ n$. To simplify notation we

assume that a 'bias feature' with constant value of 1 is contained in the feature vector $\mathbf{x}$. Similarly the potential of the e-th edge is given

$$\boldsymbol{\Psi}^{(e)} = \exp\{\mathbf{U}^{(e)}\mathbf{x}\} \tag{11}$$

where $\mathbf{U}^{(e)} \in \mathbb{R}^{2\times 2\times D}$ is the weight tensor associated with the e-th edge of the model and multiplication takes place on the outermost dimension, and $\boldsymbol{\Psi}^{(e)} \in \mathbb{R}^{2\times 2}\ \forall\ e$.

Inference can be performed with standard message passing which can efficiently be computed with the forward-backward dynamic program. The $n+1$ forward vector is given by

$$\boldsymbol{\alpha}^{(n+1)} = \boldsymbol{\Psi}^{(n)\top}\boldsymbol{\gamma}^{(n)} \tag{12}$$

where $\top$ represents the matrix transpose, $\boldsymbol{\gamma}^{(n)} \triangleq \boldsymbol{\alpha}^{(n)} \odot \boldsymbol{\psi}^{(n)}$ and $\odot$ represents the Hadamard product. The $n-1$ backward vector is calculated similarly with

$$\boldsymbol{\beta}^{(n-1)} = \boldsymbol{\Psi}^{(n)}\boldsymbol{\delta}^{(n)} \tag{13}$$

where $\boldsymbol{\delta}^{(n)} \triangleq \boldsymbol{\psi}^{(n)} \odot \boldsymbol{\beta}^{(n)}$, and the base cases for the forward and backward vectors are $\boldsymbol{\alpha}^{(0)} \triangleq \mathbf{1}$ and $\boldsymbol{\beta}^{(K)} \triangleq \mathbf{1}$. Note, marginalisation is often performed in the log domain with the log-sum-exp function for numerical stability but identical marginal distributions are achieved to those above.

It can be shown that the forward and backward vectors yield sufficient information for exact marginal probability estimation [19] and the probability of the n-th position of the label is given by

$$P(\mathbf{y}_n) = \boldsymbol{\alpha}^{(n)} \odot \boldsymbol{\psi}^{(n)} \odot \boldsymbol{\beta}^{(n)}/Z \tag{14}$$

where Z is the global normaliser of the sequence that can be calculated at any position, $Z = \mathbf{1}^\top(\boldsymbol{\alpha}^{(n)} \odot \boldsymbol{\psi}^{(n)} \odot \boldsymbol{\beta}^{(n)})$, and the probability across the n-th edge is

$$P(\mathbf{y}_n, \mathbf{y}_{n+1}) = \gamma^{(n)} \odot \boldsymbol{\Psi}^{(n)} \odot \delta^{(n+1)\top}/Z \tag{15}$$

Figure 1 illustrates the inference procedure along a graph. Three nodes are shown here, and each of the intermediate quantities introduced earlier are shown.

Learning Optimisation is performed by minimising the negative logarithm of the likelihood, *i.e.*

$$\mathcal{NLL} = -\sum_{n=1}^{N} \log P(\mathbf{Y}_n|\mathbf{X}_n, \Theta) \tag{16}$$

where $\Theta = \{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \ldots, \mathbf{W}^{(K-1)}, \mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \ldots, \mathbf{U}^{(K-2)}\}$ is the set of model parameters. It is easy to show that the gradient of the i-th element of the n-th weight vector is:

$$\frac{\delta \mathcal{NLL}}{\delta \mathbf{W}_i^{(n)}} = \frac{1}{N} \sum_{j=1}^{N} \left(P\left(\mathbf{Y}_{j,n} = i | \, \mathbf{X}_j, \Theta\right) - \mathbb{I}\{\mathbf{Y}_{j,n} = i\}\right) \mathbf{X}_j \tag{17}$$

where $\mathbb{I}\{\cdot\}$ is the identity function, and derivation of the above follows similar methodology for other log-linear models [17,19] and very similar expressions can be produced to produce gradients with respect to the edge weights $\mathbf{U}_{i,j}^{(e)}$. Standard gradient-based optimisation techniques can be used to minimise the negative log likelihood, *e.g.* L-BFGS.

It is interesting to note that even though log loss is optimised here that the structure of the labels can be seen to be functionally related the absolute error between labels and predictions. One can view this either as a hybrid loss function or that the proposed methodology implicitly applies misclassification costs owing to the structure of the encoded label space.

Comments on the Model Since many aspects of this model are unexplored in the field of OR we take a moment to comment on some aspects of this model in this setting

Edges: We interpret the edges of the model as driving the 'transitions' between two adjacent encoded bits. In more traditional sequence learning settings, including natural language processing, is it very typical to direct bespoke features for the edges only. We ascribe a similar interpretation of the edges in our setting, *i.e.* the n-th edge primarily drives whether the n-th bit of the encoded label is sustained or transitions whereas the node weights drive the basic identification of categories.

Predictive Distribution: CRFs facilitate several methods for producing predictions: forward filtering, Viterbi path, marginal probability distribution of the sequence [9]. Although in this work we consider the Viterbi path, we acknowledge that existing literature exists that suggests other predictive functions to be used when optimising for different performance metrics.

Errors: Not all paths are permissible with our encoding scheme, with $0 \rightarrow 1$ transitions in particular being forbidden. Several approaches can be incorporated to produce predictions consistent with the encoding. Firstly, one can explicitly forbid illegal transitions by setting $\boldsymbol{\Psi}_{0,1}^{(e)} = 0$. However, in our experimentation we recognised some evidence that there is a correspondence between invalid predictions and outlying data. This work is ongoing.

Implementation Notes Due to the specific nature of our problem, some operations can be vectorised to increase learning and inference time. Since all sequences are of the same length $(K-1)$ the message passing procedure can be vectorised across all instances. In so doing forward messages will be passed from position n to $n+1$ across all instances. Similarly, backward messages can be passed in a similar vectorised manner. This is often not possible due to the fact that most sequence learning problems have instances of different length.

Furthermore, if the carnality of the ordinal problem is small (in our experiments less than $K < 30$) inference can be performed the exponential domain without re-normalisation without noticeable loss of fidelity in probability estimates. This yields significant gains in terms of the computational time since neither the logarithm or exponential functions are used for marginalisation.

4 Experiments

4.1 Models

We compare four different models that are linear in their parameters. We only consider linear models so that we can compare the proposed method with baselines in the 'natural' data representations. Practitioners that wish generalise this work and consider nonlinear predictors may incorporate kernel functions (polynomial, for example) or explicitly parameterise nonlinear representations with deep network architectures. Hence, we consider the following linear models only:

1. Ordered Logit (OrdLog);
2. Nested Binary Ordinal Regression (BinNest);
3. Logistic Regression (LogReg); and
4. Structured Ordinal Regression Modeling (StORM).

These models are all log-linear and regularisation was performed on the weight parameters and we perform crossvalidation over the ℓ_2 norm of the parameters. We select the regularisation parameter on the training set using 5-fold cross validation.

4.2 Datasets

Table 1 shows the characteristics of the datasets considered in the empirical evaluation in this paper. The table presents four categories of datasets (synthetic, UCI, large and health) and these are explained in the subsequent subsections.

Synthetic For our synthetic experiments, we project data onto the four following data manifolds: 1. Linear; 2. Sine; 3. Circle; and 4. Spiral. These data manifolds lie in 2D spaces, and we illustrate the predictive distributions of all models visually in order to understand the strengths and limitations of each model. Empirical validation is performed with $K = 5$ and $K = 10$. These datasets are shown visually in our results and discussion.

Table 1. The datasets that are considered in this work.

	Dataset	Features	Train	Test	K
Synthetic	Linear	2	100	1000	5 & 10
	Sine	2	100	1000	5 & 10
	Circle	2	100	1000	5 & 10
	Spiral	2	100	1000	5 & 10
UCI	Diabetes	2	30	13	5 & 10
	Pyrimidines	27	50	24	5 & 10
	Triazines	60	100	86	5 & 10
	Wisconsin	32	130	64	5 & 10
	Machine CPU	6	150	59	5 & 10
	AutoMPG	7	200	192	5 & 10
	Boston Hous	13	300	206	5 & 10
	Stocks	9	600	350	5 & 10
	Abalone	8	1000	3177	5 & 10
Large	Bank 1	8	50	8142	5 & 10
	Bank 2	32	75	8117	5 & 10
	CompAct1	12	100	8092	5 & 10
	CompAct2	21	125	8067	5 & 10
	Cali Hous	8	150	15490	5 & 10
	Census1	8	175	16609	5 & 10
	Census2	16	200	16584	5 & 10
AD	DementiaBank	1605	200	169	4
	CASAS	278	200	118	5

UCI & Large We follow [12] with two categories of datasets. The the following datasets from the UCI machine learning repository: AutoMPG, Diabetes, Abalone, BostonHousing, MachineCup, Pyrimidines, StocksDomain, Triazines, and Wisconsin. Although many of these datasets are used to understand regression models, we incorporated equal-frequency binning on these datasets so that they can be used in ordinal tasks. Empirical validation is performed with $K = 5$ and $K = 10$. We also consider a second (larger) set of data that was also introduced in [12] as the 'large' dataset.

Healthcare Finally, we also evaluate our model on two AD datasets. DementiaBank [20] is a longitudinal dataset of multimedia interactions for the study of communication in dementia. The dataset contains transcript and audio files from interviews between patients and clinicians, and covers a range of diagnostic tests in mental health, such as Alzheimers Dementia, Parkinsons, and mild cognitive impairment. The transcripts and audio files were gathered as part of a larger protocol administered by the Alzheimer and Related Dementias Study at the University of Pittsburgh School of Medicine. We use the DementiaBank dataset in an ordinal regression setting to model the various stages of progression of AD: cognitively healthy, possible dementia, probable and dementia.

The Centre for Advanced Studies in Adaptive Systems (CASAS) research group produce models and datasets for smart-home behaviour modelling. Their datasets consist of sensor data (including Passive Infra-Red (PIR), temperature, door and object sensors) derived from naturalistic living in a SH environment. The 'cognitive assessment activity dataset' [2,21,22] consists of approximately 400 participants performing several ADLs and IADL in the SH. Cognitive clinicians graded the activities were graded by domain experts on a range of 1-5, and predicting the assigned grade from sensor data is the task that we investigate here.

4.3 Performance Evaluation

All datasets are partitioned randomly into 20 folds on the 'synthetic', 'UCI' and 'healthcare' datasets (*c.f.* Table 1). Following the protocol of [12] we also performed 100 randomised splits for the 'large' datasets. Model hyperparameters are selected with 5-fold cross validation on the training set, and the selected parameters are used for performance evaluation on the test set. We follow [23,5] in our evaluation metrics and use macro-averaged 0/1 loss, Mean Absolute Error (MAE).

Additionally, significance of results is reported with the Wilcoxon's signed rank test [24] at a (fairly stringent) significance level of $\alpha = 0.01$. We illustrate the statistical significance with critical difference diagrams [25] that are for the understanding of statistical significance when multiple classifiers are compared over multiple datasets. An example is shown in Figure 2. Four classifiers are shown here (Model 1, Model 2, Model 3, and Model 4) and the average rank of each is marked on the number line. The groups of algorithms whose results are not statistically different are connected together with a heavy horizontal line, *i.e.* the difference between Models 2 and 3 is not statistically significant, whereas the difference between Models 1 and 2 is.

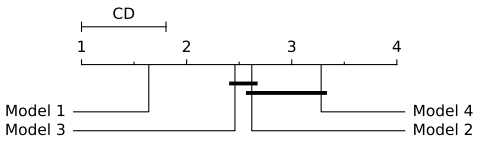
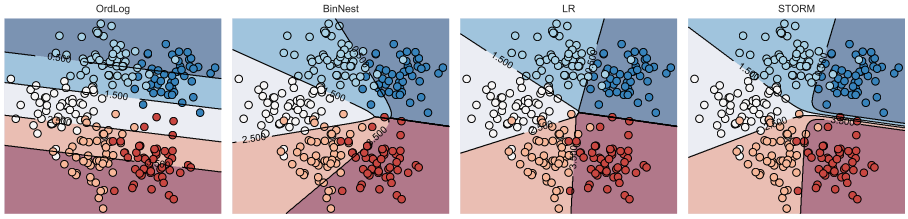


Fig. 2. Example critical difference diagram.

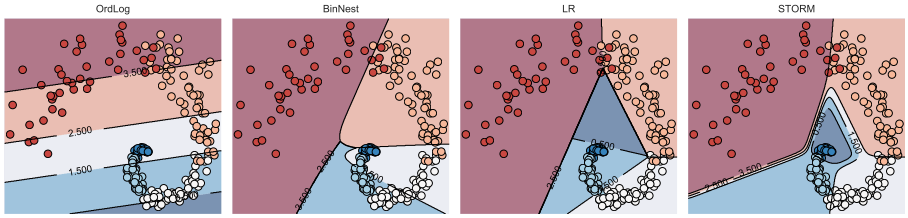
5 Results

In this section we present and discuss the results from the synthetic, UCI, large and healthcare datasets and conclude by discussing the complete results together.

5.1 Synthetic Datasets



(a) Circle dataset with 5 categories



(b) Spiral dataset with 5 categories

Fig. 3. Predictions from baseline and proposed ordinal Circle and Spiral datasets (Figures 3(a) and 3(b) datasets).

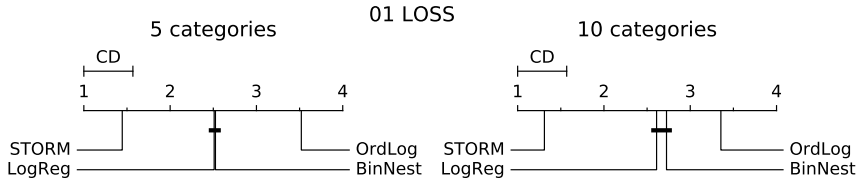
Results We first present our results on synthetic datasets visually since these datasets are in two dimensions. The upper two subfigures of Figure 3 present the results from the four classifiers considered (OrdLog, BinNest, LogReg, and StORM) on the Circle and Spiral (we show the Linear and Sine predictions in the supplementary material). The dots represent instances in a two dimensional space, and the fill colour of each depicts the ground-truth label; darkest blue representing class 1 and darkest red representing class K. Additionally, the background colour in these figures represents the predicted ordinal quantities obtained from each model. The colour scheme is shared between the background and fill colours.

Figure 3(a) and 3(b) show the predictions obtained when the ordinal data lies on a Circle and Spiral manifolds respectively on the 5-category dataset. Clearly, due to the limitations of the OrdLog model it cannot perform optimally

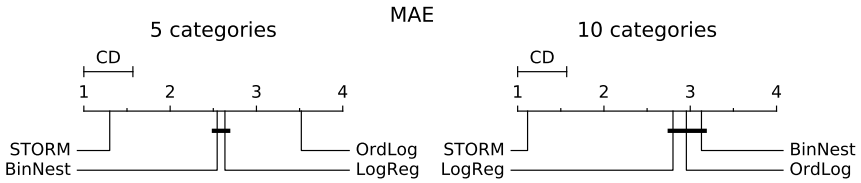
here. Additionally the BinNest model does not adapt to the dynamics of the data manifold in this setting either since the greedy learning routine cannot resolve these manifolds (particularly with the Spiral dataset). The non-ordinal LogReg model and the proposed StORM are better able to adapt to the challenges with these data manifolds, with StORM adapting most efficiently.

We have observed noteworthy behaviour with the StORM on all synthetic experiments, namely that the space of low-valued predictions tend to be 'consumed' the domain of higher-valued predictions. This phenomenon is illustrated clearly in Figure 3(b) (right) with the spiral dataset and the StORM model, but can also be observed in Figures 3(a). This is achieved due to the encoding of the labels and is a fundamental property of StORM. However, this is also a feature of many OrdLog models, but cannot be guaranteed by the other baselines we consider, *e.g.* BinNest or LogReg.

Following [23,5], we quantified performance using two metrics: macro-averaged mean absolute error and macro-averaged mean 0/1 loss. Figure 4 shows the critical difference diagram [25] for the mean zero-one loss (Figure 4(a)) and mean absolute error (Figure 4(b)). (For a description on how to read and interpret critical difference diagrams we refer the reader to Section 4.3 and more generally to [25].) We can see from this figure that the proposed approach is ranked best and that its performance is significantly better than those of the baselines on all performance metrics considered.



(a) Macro-averaged 0/1 loss over synthetic datasets



(b) Macro-averaged mean absolute loss over synthetic datasets

Fig. 4. Critical difference diagrams for synthetic datasets over the 20 train/test permutations.

Versatile Queries Since the language of probabilistic graphical models underpin the proposed method, StORM may be queried in a variety of ways. In particular, here we will demonstrate how non-standard queries can be made by visualising the predictive distribution on an edge transition.

The probability distribution over the transition between the i-th and $(i+1)$-th positions is given by marginalising over all other positions, *i.e.*

$$P(Y_i, Y_{i+1}) = \sum_{Y_1} \sum_{Y_2} \cdots \sum_{Y_{i-1}} \sum_{Y_{i+2}} \cdots \sum_{Y_{K-1}} \sum_{Y_K} P(Y_1, Y_2 \ldots Y_K) \tag{18}$$

and this can efficiently computed with forward and backward vectors (see Equation 15). Figure 5 depicts the probability distribution over the transition between positions 3 and 4 on a variation of the 10-category Spiral dataset. Regions shaded in blue and red represent regions of low and high predicted probability respectively. The left figure shows $P(Y_3 = 0, Y_4 = 0)$, the middle figure shows $P(Y_3 = 1, Y_4 = 0)$, and the right figure shows $P(Y_3 = 1, Y_4 = 1)$. These probability distributions can be interpreted as $P(Y < 4)$ in the left figure captures, $P(Y = 4)$ in the middle captures and the right figure shows $P(Y > 4)$. To understand why, we can consider at the third and fourth tags of encoded labels for several labels, and observe that the third and fourth tags for $\widehat{y} < 4$ are both 0, for $\widehat{y} > 4$ are both 1, and for $\widehat{y} = 4$ we observe the pair $(1, 0)$ corresponding to Figure 5.

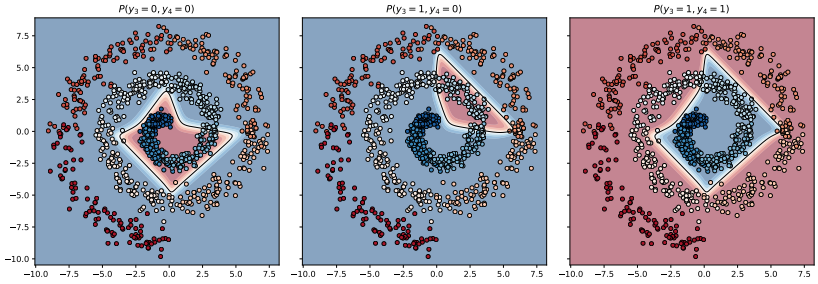
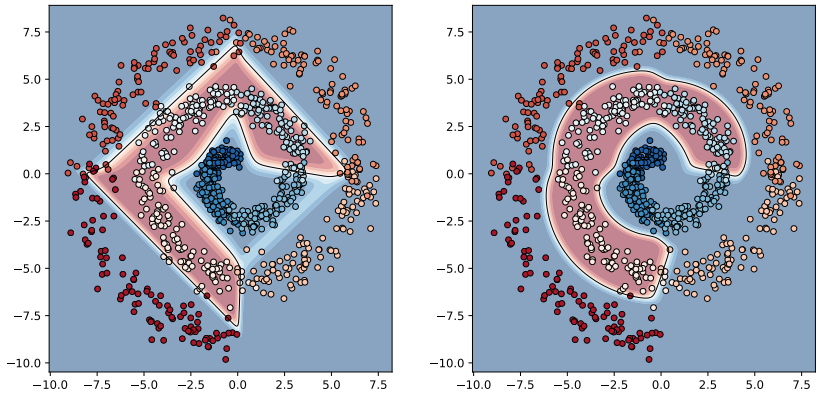


Fig. 5. $P(y < 4)$ (left), $P(y = 4)$ (middle), and $P(y > 4)$ (right).

Although the model itself is linear in its parameters the predictive distribution has adapted to the nonlinear data manifold. In settings with large K (*i.e.* many ordinal categories) one can easily execute more general queries (*e.g.* $P(4 \leq Y < 7)$). As discussed in Section 1, this is a common task in clinical settings, *e.g.* AD patients will first be graded on a large scale before these are reduced into important intervals. The predictive distribution of $P(4 \leq Y \leq 7)$ may be indicative of a particular grade (*e.g.* 'moderate' AD) and can be computed in our model. We demonstrate this visually for the a variant of the spiral

dataset in Figure 6(a). Although the focus of this work is on linear settings, we demonstrate the effect of Nyström kernel approximation [26] of the Radial Basis Function (RBF) kernel in Figure 6(b). We notice that the nature of the data manifold is better captured with this representation and the predictive distribution curves alongside the manifold.



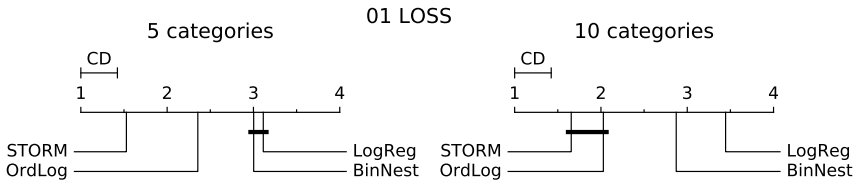
(a) Linear features. (b) Nyström features.

Fig. 6. Versatile querying of the StORM on Spiral.

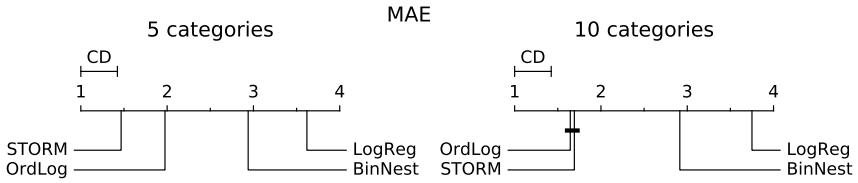
5.2 Predictive Performance on UCI Datasets

Predictive performance was also evaluated on several datasets from the UCI machine learning datasets repository [27]. Figure 7 presents the critical difference diagrams for the 0/1 loss (Figure 7(a)) and mean absolute error (Figure 7(b)) over 5 (left) and 10 (right) categories. Figure 7 illustrates that the StORM model is the best performing model over all metrics with 5 categories, and its performance is significantly better on all metrics with the sole exception of mean squared error. Figure 7 also shows that other models are competitive with StORM on the 10-category datasets. StORM is never significantly less performant than the winning model, but is significantly better than LogReg and BinNest baselines.

We test the performance of StORM with larger datasets (in terms of number of instances and features) with the 'large' dataset from [12], and the results of these are shown in Figure 8. These experiments were repeated over 100 randomised folds with 5 and 10 categories. StORM significantly outperforms baseline approaches.

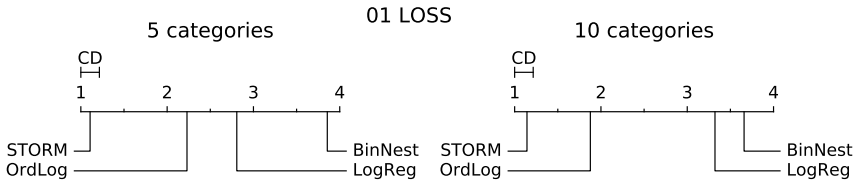


(a) Macro-averaged 0/1 loss over UCI datasets

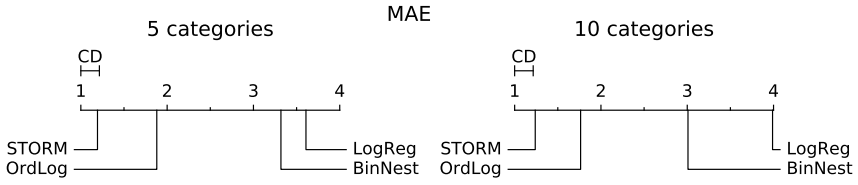


(b) Macro-averaged mean absolute loss over UCI datasets

Fig. 7. Critical difference diagrams for UCI datasets.



(a) Macro-averaged 0/1 loss over large datasets



(b) Macro-averaged mean absolute loss over large datasets

Fig. 8. Critical difference diagrams for large datasets.

5.3 Healthcare Datasets

Finally, we present results on the healthcare datasets. For the CASAS dataset we were unable to produce the same feature representations that were used in the original paper since some of the data is withheld to preserve anonymity. We extracted the duration of the activity, the number of unique sensors, the most commonly triggered sensor, and the number of sensors from each category (presence, door, object *etc.*) that were triggered. The task of this dataset is to estimate the 'incompleteness' of an AD with 5 meaning the task was not completed and 1 good completion.

With DementiaBank, we analysed the transcripts of the interviews conducted with the participants and defined an ordinal task on the following order: cognitively healthy, possible dementia, probable dementia, dementia. The transcripts also included annotations of pausing and verbal disfluency. Data representation consisted of counting occurrences and normalising features.

(a) 0-1 loss on healthcare datasets. (b) MAE on healthcare datasets.

Fig. 9. Critical difference diagrams on the healthcare.

Figure 9 presents the critical difference diagrams for the healthcare datasets. Note, that in all cases the critical difference in these figures is larger than in the synthetic and UCI datasets due to the smaller number of datasets here. These experiments have produced a much more competitive set of results with no one model consistently out-performing the others in a statistically meaningful manner. The StORM model is the best performing model over all tests conducted, even though its performance is not significantly better than ordinal regression.

In Figure 10 we show feature embeddings of the CASAS dataset. We show two diagnostic categories from opposite ends of cognitive spectrum: young volunteers and volunteers with dementia. This visualisation highlights two challenges with this dataset: 1) the class distribution is unequal (much fewer dementia data are available), and 2) there is significant overlap between the classes in this visualisation. As a result it is not surprising that difference in performance is not significant since the task is challenging.

We present the raw classification tables for the healthcare datasets in Table 2. We can observe here that on average the 0/1 and MAE losses are much lower

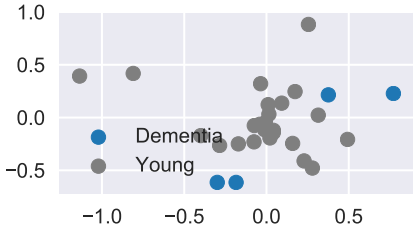


Fig. 10. Embedding of CASAS features.

Table 2. Table of results for the healthcare datasets.

Dataset	Model	0/1 Loss	MAE
CASAS	BinNest	0.426 ± 0.076	0.698 ± 0.138
	LogReg	0.436 ± 0.066	0.788 ± 0.157
	OrdLog	0.499 ± 0.064	0.759 ± 0.11
	StORM	0.42 ± 0.077	0.674 ± 0.151
DBank	BinNest	0.234 ± 0.045	0.382 ± 0.07
	LogReg	0.228 ± 0.042	0.389 ± 0.075
	OrdLog	0.356 ± 0.038	0.456 ± 0.053
	StORM	0.234 ± 0.038	0.366 ± 0.063

on DementiaBank than on the CASAS dataset. However, the losses are, on average, rather high, due to the challenging learning task.

6 Discussion

The main results presented here show that the proposed method (StORM) is a robust and a winning model for the prediction of ordinal quantities in most of the settings considered here. On the synthetic datasets (which primarily are used for the understanding of the model in comparison to baselines) we showed visually that our approach is able to adapt to non-linear and challenging data manifolds. Although it is highly unlikely that one will encounter manifolds of the exact form of Figure 3 in real datasets, we also find it highly unlikely that strictly linear manifolds will be encountered in real-life scenarios. We are confident in the utility of the proposed methodology given its robust adaptation to the variation of challenging data manifolds. Although the absolute performance of all models is slightly disappointing on the healthcare datasets, this is illustrative of data representation challenges that still remain. Indeed, on these datasets some of the most important and discriminatory features (including health records) are witheald to preserve the anonymity of the participants, which further exacerbates the classification task. Yet, StORM is the best performing model.

StORM is shown to have higher performance in a statistically meaningful way on the synthetic, UCI and large datasets across all categories. In particular, we see that StORM achieves very good results in the large datasets (Figure 8). However, in the case of the UCI datasets we see that for the 10 category dataset StORM is still the highest-performing model but that the the baseline ordinal regression model performs well. It is worth pointing out that many of the datasets within the UCI group were converted into an ordinal task from a regression task. Although the converted datasets still constitute legitimate ordinal challenges, we believe the process of conversion is relatively 'arbitrary' and that the groupings given do not necessarily constitute meaningful groups of data. We believe this to be the reason for the absence of statistically meaningful results on the 10 category UCI datasets. However, on the large datasets we see that StORM is comfortably the best model amongst the baselines. We believe that this is driven primarily by the scale of the datasets here: StORM is better able to capture the training data distribution with larger datasets. This makes sense intuitively. Since StORM has a larger number of parameters these models witll typically require more data.

7 Conclusion

In this paper we proposed a structured propabilistic architecture for ordinal regression that is based on a structured encoding of the target variables and undirected graphical models. We have shown empirically that the proposed method (structural ordinal regression modelling) performs significantly better than three baseline methods over several synthetic, UCI and healthcare datasets. Additionally, our proposed framework has several appealing properties: inference can be vectorised over the whole dataset to speed up optimisation, locally and globally consistent abstract queries can be executed on the data, and our model preserves several desirable monotonic features for ordinal model. Future work will investigate non-linear representation methods with the proposed system and to compare the proposed techniques against more baseline methods.

Acknowledgements

This research was conducted under the 'Continuous Behavioural Biomarkers of Cognitive Impairment' project funded by the UK Medical Research Council Momentum Awards under Grant MC/PC/16029.

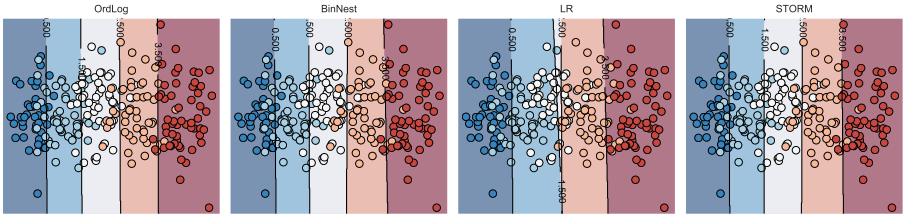
References

1. Sidney Katz. Assessing self-maintenance: activities of daily living, mobility, and instrumental activities of daily living. *Journal of the American Geriatrics Society*, 31(12):721–727, 1983.
2. Prafulla N Dawadi et al. Automated assessment of cognitive health using smart home technologies. *Technology and health care*, 21(4):323–343, 2013.

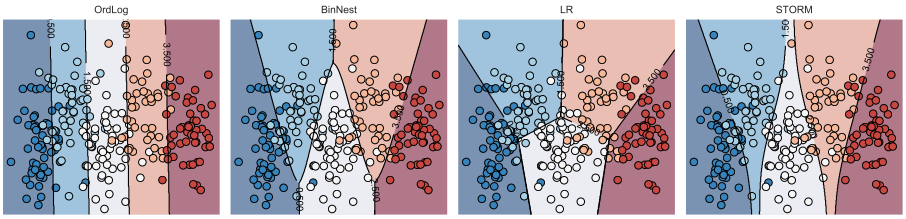
3. Prabitha Urwyler et al. Cognitive impairment categorized in community-dwelling older adults with and without dementia using in-home sensors that recognise activities of daily living. *Scientific Reports*, 7:42084, 2017.
4. Orla M Doyle et al. Predicting progression of alzheimers disease using ordinal regression. *PloS one*, 9(8):e105542, 2014.
5. Pedro Antonio Gutiérrez et al. Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146, 2016.
6. Han-Hsing Tu and Hsuan-Tien Lin. One-sided support vector regression for multiclass cost-sensitive classification. In *ICML*, pages 1095–1102, 2010.
7. Sotiris B Kotsiantis et al. A cost sensitive technique for ordinal classification problems. In *Hellenic Conference on Artificial Intelligence*, pages 220–229. Springer, 2004.
8. Peter McCullagh. Regression models for ordinal data. *Journal of the royal statistical society. Series B (Methodological)*, pages 109–142, 1980.
9. Murphy Kevin. Machine learning: a probabilistic perspective, 2012.
10. Ralf Herbrich et al. Support vector learning for ordinal regression. *IET Conference Proceedings*, pages 97–102(5), January 1999.
11. Koby Crammer and Yoram Singer. Online ranking by projecting. *Neural Computation*, 17(1):145–175, 2005.
12. Wei Chu and Zoubin Ghahramani. Gaussian processes for ordinal regression. *Journal of machine learning research*, 6(Jul):1019–1041, 2005.
13. Thomas G Dietterich et al. Solving multiclass learning problems via error-correcting output codes. *Journal of artificial intelligence research*, 2:263–286, 1994.
14. Eibe Frank and Mark Hall. A simple approach to ordinal classification. *Machine Learning: ECML 2001*, pages 145–156, 2001.
15. Jaime S Cardoso et al. Learning to classify ordinal data: The data replication method. *Journal of Machine Learning Research*, 8(Jul):1393–1429, 2007.
16. Floriana Esposito, Donato Malerba, V Tamma, and HH Bock. Similarity and dissimilarity. In *Analysis of Symbolic Data*, pages 139–197. Springer, 2000.
17. John D. Lafferty et al. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.
18. Niall Twomey, Tom Diethe, and Peter Flach. On the need for structure modelling in sequence prediction. *Machine Learning*, 104(2-3):291–314, 2016.
19. Charles Sutton et al. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373, 2012.
20. James T Becker et al. The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6):585–594, 1994.
21. Aaron S Crandall and Diane J Cook. Smart home in a box: A large scale smart home deployment. In *Intelligent Environments (Workshops)*, pages 169–178, 2012.
22. Prafulla Dawadi et al. An approach to cognitive assessment in smart home. In *Proceedings of the 2011 workshop on Data mining for medicine and healthcare*, pages 56–59. ACM, 2011.
23. Stefano Baccianella et al. Evaluation measures for ordinal regression. In *Intelligent Systems Design and Applications, 2009. ISDA'09. Ninth International Conference on*, pages 283–287. IEEE, 2009.
24. Alessio Benavoli et al. Should we really use post-hoc tests based on mean-ranks? *The Journal of Machine Learning Research*, 17(1):152–161, 2016.

25. Janez Demsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.
26. Christopher KI Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Advances in neural information processing systems*, pages 682–688, 2001.
27. Arthur Asuncion et al. Uci machine learning repository, 2007.

Appendix



(a) Linear dataset with 5 categories



(b) Sine dataset with 5 categories

Fig. 11. Predictions from baseline and proposed ordinal on the Linear and Sine datasets (Figures 11(a) and 11(b)).

Here we present additional visualisations and results tables for the interpretation and reproduction of the main results of this paper.

7.1 Supplementary Figures

Figure 11 shows the predictions of the baseline and proposed methods on the Linear and Sine datasets.

7.2 Supplementary Tables

Tables 3 and 4 present the results on the synthetic datasets on the 5 and 10 category splits respectively, Tables 5 and 6 present the results on the UCI datasets

on the 5 and 10 category splits respectively. The first two columns show depict the dataset and prediction model and the remining columns show the scores on 0/1 loss and MAE.

Table 3. Table of results for the synthetic collection of datasets with 5 categories.

dataset	model	0/1 Loss	MAE
Circle	BinNest	0.14 ± 0.01	0.174 ± 0.013
	LogReg	0.058 ± 0.006	0.094 ± 0.011
	OrdLog	0.519 ± 0.014	0.544 ± 0.014
	StORM	0.08 ± 0.006	0.09 ± 0.009
Sine	BinNest	0.128 ± 0.013	0.129 ± 0.013
	LogReg	0.169 ± 0.013	0.173 ± 0.013
	OrdLog	0.164 ± 0.009	0.164 ± 0.009
	StORM	0.128 ± 0.011	0.128 ± 0.011
Linear	BinNest	0.17 ± 0.01	0.17 ± 0.01
	LogReg	0.32 ± 0.018	0.322 ± 0.018
	OrdLog	0.17 ± 0.011	0.171 ± 0.011
	StORM	0.168 ± 0.011	0.168 ± 0.011
Spiral	BinNest	0.299 ± 0.008	0.593 ± 0.011
	LogReg	0.063 ± 0.01	0.066 ± 0.012
	OrdLog	0.65 ± 0.011	0.922 ± 0.016
	StORM	0.058 ± 0.009	0.06 ± 0.01

Table 4. Table of results for the synthetic collection of datasets with 10 categories.

dataset	model	0/1 Loss	MAE
Circle	BinNest	0.442 ± 0.013	0.772 ± 0.064
	LogReg	0.338 ± 0.012	0.641 ± 0.069
	OrdLog	0.738 ± 0.011	1.211 ± 0.028
	StORM	0.356 ± 0.01	0.417 ± 0.021
Sine	BinNest	0.148 ± 0.009	0.151 ± 0.009
	LogReg	0.474 ± 0.011	0.727 ± 0.017
	OrdLog	0.193 ± 0.012	0.194 ± 0.012
	StORM	0.142 ± 0.009	0.144 ± 0.009
Linear	BinNest	0.237 ± 0.048	0.243 ± 0.067
	LogReg	0.693 ± 0.01	1.31 ± 0.01
	OrdLog	0.198 ± 0.01	0.199 ± 0.01
	StORM	0.198 ± 0.011	0.198 ± 0.011
Spiral	BinNest	0.68 ± 0.01	2.765 ± 0.043
	LogReg	0.299 ± 0.012	0.983 ± 0.04
	OrdLog	0.905 ± 0.007	2.474 ± 0.017
	StORM	0.067 ± 0.009	0.112 ± 0.016

Table 5. Table of results for the UCI collection of datasets with 5 categories.

dataset	model	0/1 Loss	MAE
Abalone	BinNest	0.625 ± 0.008	0.887 ± 0.039
	LogReg	0.656 ± 0.014	0.985 ± 0.085
	OrdLog	0.664 ± 0.008	0.924 ± 0.046
	StORM	0.615 ± 0.02	0.799 ± 0.053
AutoMPG	BinNest	0.445 ± 0.028	0.549 ± 0.061
	LogReg	0.498 ± 0.057	0.648 ± 0.173
	OrdLog	0.39 ± 0.019	0.394 ± 0.019
	StORM	0.361 ± 0.033	0.37 ± 0.036
BostonHousing	BinNest	0.481 ± 0.076	0.618 ± 0.121
	LogReg	0.564 ± 0.088	0.742 ± 0.149
	OrdLog	0.392 ± 0.048	0.48 ± 0.055
	StORM	0.337 ± 0.041	0.412 ± 0.052
Diabetes	BinNest	0.739 ± 0.09	0.907 ± 0.135
	LogReg	0.723 ± 0.058	0.949 ± 0.106
	OrdLog	0.643 ± 0.101	0.766 ± 0.12
	StORM	0.622 ± 0.129	0.771 ± 0.179
MachineCup	BinNest	0.496 ± 0.08	0.657 ± 0.149
	LogReg	0.553 ± 0.026	0.775 ± 0.113
	OrdLog	0.45 ± 0.08	0.509 ± 0.124
	StORM	0.414 ± 0.079	0.456 ± 0.087
Pyrimidines	BinNest	0.643 ± 0.081	0.791 ± 0.152
	LogReg	0.608 ± 0.079	0.803 ± 0.143
	OrdLog	0.605 ± 0.087	0.73 ± 0.117
	StORM	0.503 ± 0.08	0.673 ± 0.154
StocksDomain	BinNest	0.489 ± 0.098	0.603 ± 0.157
	LogReg	0.39 ± 0.048	0.402 ± 0.052
	OrdLog	0.3 ± 0.023	0.304 ± 0.023
	StORM	0.146 ± 0.016	0.15 ± 0.016
Triazines	BinNest	0.774 ± 0.029	1.397 ± 0.111
	LogReg	0.763 ± 0.026	1.528 ± 0.084
	OrdLog	0.763 ± 0.026	1.292 ± 0.074
	StORM	0.711 ± 0.045	1.287 ± 0.123
Wisconsin	BinNest	0.803 ± 0.032	1.535 ± 0.229
	LogReg	0.797 ± 0.027	1.829 ± 0.223
	OrdLog	0.738 ± 0.051	1.159 ± 0.112
	StORM	0.813 ± 0.05	1.432 ± 0.154

Table 6. Table of results for the UCI collection of datasets with 10 categories.

dataset	model	0/1 Loss	MAE
Abalone	BinNest	0.831 ± 0.025	2.389 ± 0.126
	LogReg	0.808 ± 0.004	2.306 ± 0.122
	OrdLog	0.796 ± 0.004	1.671 ± 0.122
	StORM	0.747 ± 0.015	1.572 ± 0.133
AutoMPG	BinNest	0.728 ± 0.043	1.342 ± 0.244
	LogReg	0.786 ± 0.05	2.268 ± 0.582
	OrdLog	0.57 ± 0.044	0.85 ± 0.091
	StORM	0.544 ± 0.046	0.753 ± 0.115
BostonHousing	BinNest	0.719 ± 0.02	1.426 ± 0.041
	LogReg	0.786 ± 0.035	1.999 ± 0.245
	OrdLog	0.577 ± 0.031	0.858 ± 0.081
	StORM	0.559 ± 0.054	0.84 ± 0.109
Diabetes	BinNest	0.808 ± 0.086	1.712 ± 0.417
	LogReg	0.844 ± 0.03	1.731 ± 0.18
	OrdLog	0.838 ± 0.084	1.473 ± 0.278
	StORM	0.787 ± 0.14	1.619 ± 0.317
MachineCup	BinNest	0.708 ± 0.088	1.699 ± 0.69
	LogReg	0.792 ± 0.046	2.802 ± 0.533
	OrdLog	0.569 ± 0.094	0.984 ± 0.31
	StORM	0.562 ± 0.084	0.989 ± 0.269
Pyrimidines	BinNest	0.715 ± 0.093	1.169 ± 0.302
	LogReg	0.735 ± 0.092	1.763 ± 0.438
	OrdLog	0.658 ± 0.083	1.045 ± 0.162
	StORM	0.613 ± 0.091	1.086 ± 0.313
StocksDomain	BinNest	0.639 ± 0.051	1.059 ± 0.229
	LogReg	0.813 ± 0.07	2.048 ± 0.389
	OrdLog	0.579 ± 0.022	0.657 ± 0.023
	StORM	0.296 ± 0.019	0.308 ± 0.021
Triazines	BinNest	0.853 ± 0.02	2.647 ± 0.199
	LogReg	0.885 ± 0.013	2.929 ± 0.206
	OrdLog	0.85 ± 0.028	2.301 ± 0.235
	StORM	0.816 ± 0.047	2.371 ± 0.367
Wisconsin	BinNest	0.901 ± 0.033	3.292 ± 0.55
	LogReg	0.898 ± 0.011	4.268 ± 0.293
	OrdLog	0.874 ± 0.035	2.469 ± 0.236
	StORM	0.918 ± 0.039	3.051 ± 0.28