Ordinal Regression via Manifold Learning

Yang Liu, Yan Liu, Keith C. C. Chan
Department of Computing, The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong, P.R. China
{csygliu, csyliu, cskcchan}@comp.polyu.edu.hk

Abstract

Ordinal regression is an important research topic in machine learning. It aims to automatically determine the implied rating of a data item on a fixed, discrete rating scale. In this paper, we present a novel ordinal regression approach via manifold learning, which is capable of uncovering the embedded nonlinear structure of the data set according to the observations in the high-dimensional feature space. By optimizing the order information of the observations and preserving the intrinsic geometry of the data set simultaneously, the proposed algorithm provides the faithful ordinal regression to the new coming data points. To offer more general solution to the data with natural tensor structure, we further introduce the multilinear extension of the proposed algorithm, which can support the ordinal regression of high order data like images. Experiments on various data sets validate the effectiveness of the proposed algorithm as well as its extension.

Introduction

Ordinal regression is an important research topic in machine learning. It aims to automatically determine the implied rating of a data item on a fixed, discrete rating scale. Unlike regular regression problem, the range of the ordinal regression function should be discrete and finite. And also in contrast to the multi-class classification problem, ordinal regression not only recognizes whether the data points belong to the same group or not, but also provides the order information of different data groups.

Some algorithms have been proposed to tackle the ordinal regression problem. Herbrich et al. (2000) applied the principle of Structural Risk Minimization (SRM) to ordinal regression. Kramer et al. (2001) converted ordinal regression into a regular regression problem, which maps the ordinal variables into numeric values. In (Frank and Hall 2001), ordinal regression is transformed into a nested binary classification problem, together with the original ranking information. Crammer and Singer (2002) proposed a ranking algorithm, which aims to find the one-dimensional projection of the original data for ordinal regression. Shashua and Levin (2003) solved the ordinal regression problem using large margin techniques. Chu and Keerthi (2005) introduced two support vector approaches for ordinal regression by optimizing multiple thresholds to define the parallel discriminant hyperplanes for the ordinal scales. Li and Lin (2007) proposed a reduction framework from ordinal regression to binary classification based on extended examples. Sun et al. (2010) proposed a discriminant based ordinal regression method, which incorporates the linear discriminant analysis into the ordinal regression framework.

In this paper, we propose a novel manifold learning approach to address ordinal regression on more complex data sets with nonlinear geometry. The rationale of manifold learning is to uncover the embedded nonlinear structure of data sets based on the assumption that the high-dimensional observations lie on or close to an intrinsically low-dimensional manifold. The most representative manifold learning algorithms include isometric feature mapping (Isomap) (Tenenbaum, de Silva, and Langford 2000), locally linear embedding (LLE) (Roweis and Saul 2000), and Laplacian eigenmaps (LE) (Belkin and Niyogi 2001). Following above algorithms, more manifold learning algorithms have been developed for clustering and classification, such as locality preserving projections (LPP) (He and Niyogi 2004), neighborhood preserving embedding (NPE) (He et al. 2005), maximum variance unfolding (MVU) (Weinberger and Saul 2006), isometric projection (IsoProjection) (Cai, He, and Han 2007), discriminant LLE (DLLE) (Li et al. 2008), and discriminant Laplacian embedding (DLE) (Wang, Huang, and Ding 2010).

Although many effective manifold learning techniques have been proposed for the clustering and classification tasks, they are not directly applicable for ordinal regression problem. Figure 1 provides an illustration. The data points lie on a two-dimensional manifold with four-level ranking information. For classification or clustering task, we prefer to seek the projection function that can maximize the discriminant information, so $\mathbf{w}_1$ is the optimal axis for data mapping. However, the projection on $\mathbf{w}_1$ cannot preserve the order information of different data blocks, which is of great importance in ordinal regression. To keep both the order information and manifold geometry, the projection on $\mathbf{w}_2$ is preferred.

Based on above considerations, this paper proposes an ordinal regression approach via manifold learning. To offer

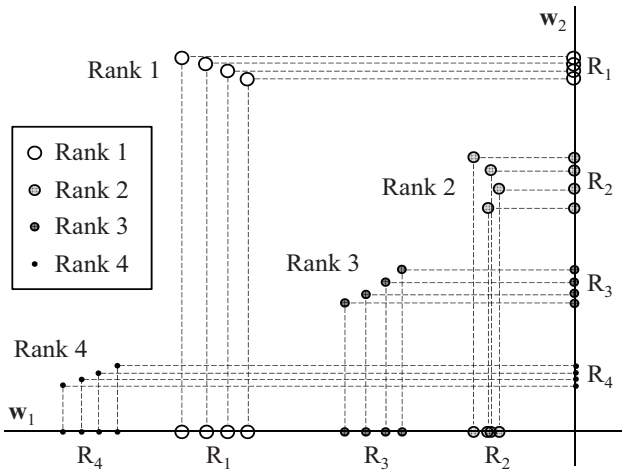

Figure 1: Illustration of ordinal regression on the data set with nonlinear geometry. For classification, the projection on $\mathbf{w}_1$ is preferred since it maximizes the discriminant information. For ordinal regression, the projection on $\mathbf{w}_2$ is preferred because it keeps the order information of different data blocks. Moreover, it preserves the manifold structure of the data set.

more general solution to the data with natural tensor structure, we further introduce the multilinear extension of the proposed algorithm, which can support the ordinal regression of high order data like images. The proposed algorithm targets several goals:

- Keeping the order information among data groups of different ranking levels.
- Maximizing the margins between two consecutive ranking levels.
- Preserving the intrinsic manifold geometry of data sets.
- Preserving the natural tensor structure of high-order data.

Ordinal Regression via Manifold Learning

Let $\{(\mathbf{x}_i, y_i)\}$ $(i = 1, ..., n)$ be the training data set, where $\mathbf{x}_i \in \mathbb{R}^d$ denotes the input data point, and $y_i \in \{1, ..., k\}$ denotes the corresponding ordered class labels. Ordinal regression aims to find a suitable function Φ which correctly maps each training data point $\mathbf{x}_i$ to its corresponding label y_i. In order to perform ordinal regression under the manifold learning framework, we formulate the objective function of the proposed algorithm as follows:

$$\begin{aligned} &\min J(\mathbf{w}, \gamma) = \sum_{i,j=1}^{n} (\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j)^2 \mathbf{A}_{ij} - C\gamma \\ &s.t.\ \ \mathbf{w}^T(\mathbf{m}_{r+1} - \mathbf{m}_r) \geq \gamma,\ r = 1, ..., k-1, \end{aligned} \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^d$ is the projection vector, and $\mathbf{A}$ is the $n \times n$ adjacency matrix constructed to model the neighborhood relationship between data points. Furthermore, $\mathbf{m}_r = \frac{1}{n_r} \sum_{y_i = r} \mathbf{x}_i$ denotes the mean vector of samples from rank r, n_r is the number of data samples in rank r, γ is the margin between the projected means of two consecutive ranks, and $C \geq 0$ is a penalty coefficient used to balance the manifold structure and order information.

Clearly, the first term of the objective function in (1) intends to preserve the manifold structure of the data set in the output space, while the second term aims to maximize the margin between the projected means of two consecutive ranks. By optimizing these two terms jointly, the projected data can be sorted according to their ranks with large margin while the intrinsic manifold structure of the data set is well preserved.

Adjacency Matrix Construction

Manifold learning uncovers the nonlinear structure by integrating the descriptions of a set of local patches using the adjacency matrix. Therefore, the effect of manifold learning largely depends on how well the adjacency matrix represents the data manifold. In this section, we present a novel method to construct the adjacency matrix of the data set.

To construct an adjacency matrix, we first need to build a neighborhood graph. Currently, there are mainly two kinds of neighborhood graphs: K graph and ε graph (Tenenbaum, de Silva, and Langford 2000; Belkin and Niyogi 2001). For the ε graph, the neighborhood radius ε is very difficult to be decided because of the variety of data density. Therefore, the K graph is more popular in practice (Hein and Maier 2007; Liu and Chang 2009). However, the K graph in existing manifold learning algorithms connect $\mathbf{x}_i$ and $\mathbf{x}_j$ if $\mathbf{x}_i$ is one of the K-nearest neighbors of $\mathbf{x}_j$, *or*, if $\mathbf{x}_j$ is one of the K-nearest neighbors of $\mathbf{x}_i$. Such an "*or*" assumption may not reflect the real neighborhood relationship between two data points. For example, some data points may be connected with outliers since the outliers consider these data points as "neighbors", even though the outliers are not really belonging to the neighborhood set of any data point.

To construct a more robust neighborhood graph, we introduce a two-way connection criterion to construct the K graph: we connect $\mathbf{x}_i$ and $\mathbf{x}_j$ only if $\mathbf{x}_i$ is one of the K-nearest neighbors of $\mathbf{x}_j$, *and*, $\mathbf{x}_j$ is also one of the K-nearest neighbors of $\mathbf{x}_i$. The proposed criterion adopts the "*and*" hypothesis, which means it agrees to connect two data points if and only if both of them are neighbors of each other. The adjacency matrix is defined as follows:

$$\mathbf{A}_{ij} = \begin{cases} \exp(-\frac{d(\mathbf{x}_i, \mathbf{x}_j)^2}{2\sigma}), & \text{if } j \in \mathcal{N}_i \text{ and } i \in \mathcal{N}_j \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $d(\mathbf{x}_i, \mathbf{x}_j)$ denotes the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, $\mathcal{N}_i$ denotes the index set of the K nearest neighbors of $\mathbf{x}_i$, and σ is empirically set by $\sigma = \sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{x}_{i_K})^2 / n$ where $\mathbf{x}_{i_K}$ is the Kth nearest neighbor of $\mathbf{x}_i$. A similar idea has been presented by Liu and Chang (2009): they built a symmetry-favored graph by putting more trust on the connection agreed by both data points, and validated that such setting is more reliable than the traditional K graph.

After considering the two-way connection criterion, more importantly, we want to integrate order information into the neighborhood graph. Actually, we aim to preserve the local-

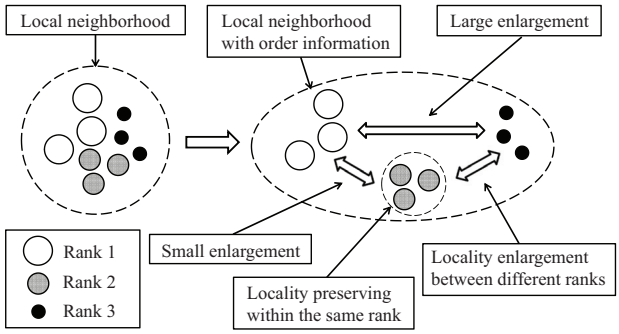

Figure 2: Schematic illustration of local neighborhood construction for ordinal regression. By considering the order information, the locality within each rank is preserved, while the locality between different ranks is enlarged.

ity within each rank, while enlarge the locality between different ranks. Moreover, we expect that the extent of enlargement can reflect the rank difference between two connected data points. Therefore, we define $d(\mathbf{x}_i, \mathbf{x}_j)$ as follows:

$$d(\mathbf{x}_i, \mathbf{x}_j) = (|y_i - y_j| + 1)||\mathbf{x}_i - \mathbf{x}_j||_2 \tag{3}$$

where $|\cdot|$ denotes the absolute value operator and $||\cdot||_2$ denotes the L_2-norm operator. If $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same rank, i.e., $y_i = y_j$, then $d(\mathbf{x}_i, \mathbf{x}_j) = ||\mathbf{x}_i - \mathbf{x}_j||_2$. If $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to different ranks, i.e., $y_i \neq y_j$, then $d(\mathbf{x}_i, \mathbf{x}_j) > ||\mathbf{x}_i - \mathbf{x}_j||_2$, which means the original Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ has been enlarged in our model. When the rank difference between two data points increases, the ratio of $d(\mathbf{x}_i, \mathbf{x}_j)$ to $||\mathbf{x}_i - \mathbf{x}_j||_2$, i.e., the extent of distance enlargement, increases accordingly. As illustrated in Figure 2, by considering the order information, the locality within each rank is kept unchanged, while the distance between different ranks are enlarged. Furthermore, the distance between rank 1 and rank 3 is enlarged more than the distance between rank 1 and rank 2.

In the adjacency matrix construction procedure, we utilize the order information from the local perspective. In the objective function (1), we aim to maximize the distance between two consecutive ranks, which considers the order information using a global manner. By incorporating the order information in our model from both local and global viewpoints, a unified manifold learning formulation is established for ordinal regression.

Optimization Procedure

To solve the optimization problem in (1), we first rewrite the objective function as follows:

$$\begin{aligned} &\min J(\mathbf{w}, \gamma) = \mathbf{w}^T \mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{w} - C\gamma \\ &s.t.\ \ \mathbf{w}^T(\mathbf{m}_{r+1} - \mathbf{m}_r) \geq \gamma,\ r = 1, ..., k-1, \end{aligned} \tag{4}$$

where $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]$ is the data matrix, $\mathbf{L} = \mathbf{D} - \mathbf{A}$ is the $n \times n$ Laplacian matrix (Belkin and Niyogi 2001), and $\mathbf{D}$ is a diagonal matrix defined as $\mathbf{D}_{ii} = \sum_{j=1}^{n} \mathbf{A}_{ij}$ $(i = 1, ..., n)$.

Then we obtain the Lagrangian equation of (4):

$$\begin{aligned} L(\mathbf{w}, \gamma, \alpha) = \ &\mathbf{w}^T \mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{w} - C\gamma \\ &- \sum_{r=1}^{k-1} \alpha_r \big(\mathbf{w}^T(\mathbf{m}_{r+1} - \mathbf{m}_r) - \gamma\big), \end{aligned} \tag{5}$$

where α_r are the Lagrange multipliers which satisfy $\alpha_r \geq 0$. The necessary conditions for the optimality are:

$$\begin{cases} \dfrac{\partial L}{\partial \mathbf{w}} = 0 \ \Rightarrow\ \mathbf{w} = \dfrac{1}{2}(\mathbf{X}\mathbf{L}\mathbf{X}^T)^\dagger \displaystyle\sum_{r=1}^{k-1} \alpha_r(\mathbf{m}_{r+1} - \mathbf{m}_r), \\ \dfrac{\partial L}{\partial \gamma} = 0 \ \Rightarrow\ C = \displaystyle\sum_{r=1}^{k-1} \alpha_r, \end{cases} \tag{6}$$

where $(\mathbf{X}\mathbf{L}\mathbf{X}^T)^\dagger$ is the Moore-Penrose pseudoinverse of the $d \times d$ matrix $\mathbf{X}\mathbf{L}\mathbf{X}^T$. If $\mathbf{X}\mathbf{L}\mathbf{X}^T$ is invertible, then the pseudoinverse and the inverse coincide, i.e., $(\mathbf{X}\mathbf{L}\mathbf{X}^T)^\dagger = (\mathbf{X}\mathbf{L}\mathbf{X}^T)^{-1}$.

Based on Eq. (6), the optimization problem in (4) could be converted to the following problem:

$$\begin{aligned} &\min \sum_{r=1}^{k-1} \alpha_r(\mathbf{m}_{r+1} - \mathbf{m}_r)^T (\mathbf{X}\mathbf{L}\mathbf{X}^T)^\dagger \sum_{s=1}^{k-1} \alpha_s(\mathbf{m}_{s+1} - \mathbf{m}_s) \\ &s.t.\ \ \alpha_r, \alpha_s \geq 0,\ r, s = 1, ..., k-1, \\ &\quad \sum_{r=1}^{k-1} \alpha_r = \sum_{s=1}^{k-1} \alpha_s = C. \end{aligned} \tag{7}$$

Since $(\mathbf{X}\mathbf{L}\mathbf{X}^T)^\dagger$ is positive semidefinite, above optimization problem is a convex quadratic programming (QP) one with linear constraints. Some standard algorithms, such as the conjugate gradient method and the interior point method, can be employed to solve it. Then we can obtain the global optimal $\mathbf{w}$ by substituting α_r into the first equation in (6).

For a new test data point, we can easily determine its rank by the following decision function:

$$f(\mathbf{x}) = \min_{r \in \{1, ..., k\}} \{r : \mathbf{w}^T\mathbf{x} - b_r < 0\}, \tag{8}$$

where b_r serves as a boundary to separate rank r and rank $r+1$, which is defined as follows:

$$b_r = \begin{cases} \dfrac{\mathbf{w}^T(n_{r+1}\mathbf{m}_{r+1} + n_r\mathbf{m}_r)}{n_{r+1} + n_r} & r = 1, ..., k-1, \\ \max\limits_{i \in \{1, ..., n\}} \{\mathbf{w}^T\mathbf{x}_i\} & r = k. \end{cases} \tag{9}$$

Multilinear Extension

In real-world applications, input data are sometimes represented as high-order tensors, such as images. In order to keep the structure of input data in the learning procedure, we introduce the multilinear formulation of the proposed algorithm.

Given the training data set $\{(\mathcal{X}_i, y_i)\}$ $(i = 1, ..., n)$, where $\mathcal{X}_i \in \mathbb{R}^{d_1 \times ... \times d_N}$, and N is the order of tensor $\mathcal{X}_i$. To take

the high-order tensors as the input directly, the objective function in (1) is reformulated as follows:

$$\begin{aligned} &\min J(\mathbf{w}_p|_{p=1}^N, \gamma) = \sum_{i,j=1}^{n} \left((\mathcal{X}_i - \mathcal{X}_j)\prod_{p=1}^{N} \times_p \mathbf{w}_p^T\right)^2 \mathbf{A}_{ij} - C\gamma \\ &s.t.\ (\mathcal{M}_{r+1} - \mathcal{M}_r)\prod_{p=1}^{N} \times_p \mathbf{w}_p^T \geq \gamma,\ r = 1, ..., k-1, \end{aligned} \tag{10}$$

where $\mathbf{w}_1, ..., \mathbf{w}_N$ are projection vectors acting on corresponding orders of input tensors, $\mathcal{M}_r$ is the mean tensor of samples from rank r, and $\mathcal{X}_i \prod_{p=1}^{N} \times_p \mathbf{w}_p^T$ denotes the multilinear multiplication between tensor $\mathcal{X}_i$ and vectors $\mathbf{w}_p$. Please see (Yan et al. 2005; He, Cai, and Niyogi 2006; Dai and Yeung 2006) for details of multilinear operations.

Since the above optimization problem is not convex, we use an iterative strategy to solve it. First we fix $\mathbf{w}_2, ..., \mathbf{w}_N$, and find the optimal $\mathbf{w}_1$. Then we fix $\mathbf{w}_1, \mathbf{w}_3, ..., \mathbf{w}_N$, and find the optimal $\mathbf{w}_2$. The rest can be deduced by analogy. Finally we fix $\mathbf{w}_1, ..., \mathbf{w}_{N-1}$, and find the optimal $\mathbf{w}_N$. Repeat above steps until the procedure converges. Concretely, in the pth step ($p = 1, ..., N$), we rewrite (10) as follows:

$$\begin{aligned} &\min J(\mathbf{w}_p, \gamma) = \sum_{i,j=1}^{n} (\mathbf{w}_p^T \mathbf{x}_i^p - \mathbf{w}_p^T \mathbf{x}_j^p)^2 \mathbf{A}_{ij} - C\gamma \\ &s.t.\ \mathbf{w}_p^T(\mathbf{m}_{r+1}^p - \mathbf{m}_r^p) \geq \gamma,\ r = 1, ..., k-1, \end{aligned} \tag{11}$$

where $\mathbf{x}_i^p$ is defined as $\mathbf{x}_i^p = \mathcal{X}_i \prod_{q=1, q\neq p}^{N} \times_q \mathbf{w}_q^T$, and similarly, $\mathbf{m}_r^p = \mathcal{M}_r \prod_{q=1, q\neq p}^{N} \times_q \mathbf{w}_q^T$. Problem (11) can be solved similarly to problem (1), therefore, we can obtain the optimal $\mathbf{w}_p$ in the pth iteration.

It is easy to prove the convergence of the entire procedure. On the one hand, the convexity of problem (11) indicates that $J(\mathbf{w}_p|_{p=1}^N, \gamma)$ is nonincreasing in each iteration. On the other hand, it is obvious that $J(\mathbf{w}_p|_{p=1}^N, \gamma) \geq -C\gamma$, where C and γ are upper bounded and $C \geq 0$, which indicates that $J(\mathbf{w}_p|_{p=1}^N, \gamma)$ is lower bounded. Therefore, the iterative procedure will finally converge to a local optimal solution.

For a new tensor data point, we can determine its rank by the following multilinear decision function:

$$f(\mathcal{X}) = \min_{r\in\{1,...,k\}} \{r : \mathcal{X}\prod_{p=1}^{N} \times_p \mathbf{w}_p^T - b_r < 0\}, \tag{12}$$

where b_r is defined as follows:

$$b_r = \begin{cases} \dfrac{(n_{r+1}\mathcal{M}_{r+1} + n_r\mathcal{M}_r)}{n_{r+1} + n_r}\prod_{p=1}^{N} \times_p \mathbf{w}_p^T & r = 1, ..., k-1, \\ \max\limits_{i\in\{1,...,n\}} \{\mathcal{X}_i \prod_{p=1}^{N} \times_p \mathbf{w}_p^T\} & r = k. \end{cases} \tag{13}$$

Computational Complexity Analysis

In this section, we analyze the computational complexity of proposed algorithm as well as its multilinear extension. The most demanding steps of proposed algorithm are calculating the matrix $\mathbf{XLX}^T$, finding the Moore-Penrose pseudoinverse of $\mathbf{XLX}^T$, and solving a QP problem with a $(k-1) \times (k-1)$ Hessian matrix. The time costs of these three steps are $O(dn^2 + d^2n)$, $O(d^3)$, and $O(k^3)$, respectively. Therefore, the total computational cost of proposed algorithm is $O(dn^2 + d^2n + d^3 + k^3)$.

For the multilinear extension, we assume that the sample tensors are of uniform size in each order, i.e., $d_1 = d_2 = ... = d_N = d$. In each iteration, $\mathbf{x}_i^p$ $(i = 1, ..., n)$ should be calculated first. Together with the normal procedure in proposed algorithm, the complexity of each iteration is $O(nd^N + dn^2 + d^2n + d^3 + k^3)$. Therefore, the total computational cost of the multilinear extension of proposed algorithm is $O\big(tN(nd^N + dn^2 + d^2n + d^3 + k^3)\big)$, where t is the number of loops needed for the algorithm convergence.

Experiments

In this section, we show the performance of proposed algorithms on three data sets: the UMIST face data set (Graham and Allinson 1998), the 100k MovieLens data sets[1], and the USPS digit data set (Hull 1994). For the proposed algorithms, the nearest neighbor number K is fixed at 10 and the ten-fold cross validation is employed to determine the parameter C. The following evaluation criterion is used to quantify the accuracy of predicted ordinal scales $\{\hat{y}_1, \hat{y}_2, ..., \hat{y}_n\}$ with respect to true targets $\{y_1, y_2, ..., y_n\}$ (Sun et al. 2010):

- The mean absolute error (MAE) - the average deviation of the prediction from true order, i.e., $\frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i|$, in which we treat the ordinal scales as consecutive integers.

UMIST Face Data Set

In order to intuitively illustrate that the proposed algorithm is capable of preserving both order information and manifold structure, we first conduct an experiment using the data samples from the UMIST face data set (Graham and Allinson 1998).

Suppose that we have three kinds of face images: a man with glasses, a man without glasses, and a woman without glasses. Now we are required to find the face image of a man with glasses. Clearly, these three classes can be ranked according to the requirement: the images of a man with glasses are ranked first since they totally match the requirement; the images of a man without glasses are ranked second since they partially match the requirement; and the images of a woman without glasses are ranked last since they do not match the requirement at all.

The data set in this experiment is composed of three classes of human faces from UMIST face data set. There are 26, 38, and 20 images in rank 1, 2, and 3, respectively. Each image is grayscale and downsampled to the resolution of 56×46. For each rank, 10 images are used for training and the rest are used for test.

Figure 3 shows the projection results of training and test data. It is clear that the projected training data are arranged

[1] http://www.grouplens.org/node/73

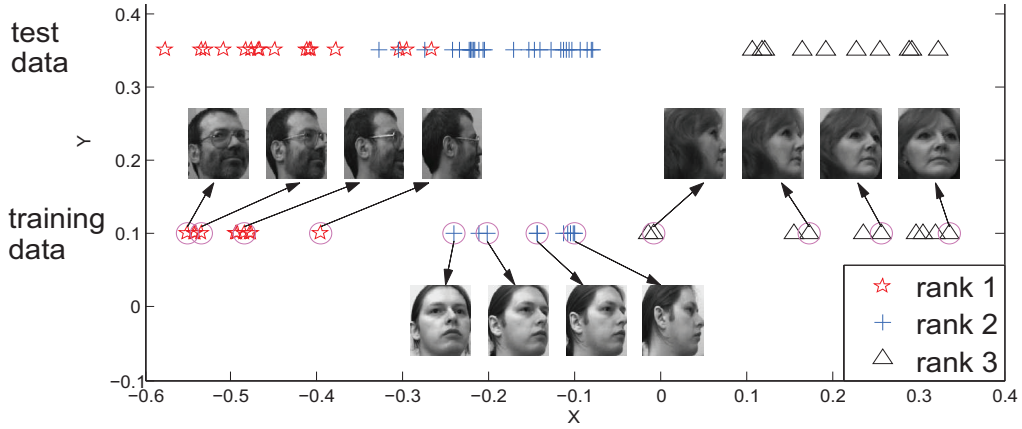

Figure 3: Projection results of training and test data points on UMIST face data set using proposed algorithm. The order information is correctly kept in both training and test projections. Furthermore, the manifold structure within each rank, i.e., the pose variation of each person, is well preserved.

Table 1: The mean absolute error (MAE) of six algorithms on MovieLens data set.

Methods	ORML	LDA	KDLOR
MAE	$\mathbf{1.285 \pm 0.176}$	1.459 ± 0.130	1.368 ± 0.172
Methods	IsoProjection	NPE	LPP
MAE	1.713 ± 0.366	1.790 ± 0.451	1.676 ± 0.173

orderly: the data points in rank 1 (corresponding to images of a man with glasses) are located in the interval of $(-0.6, -0.3)$ of the X-axis; the data points in rank 2 (corresponding to images of a man without glasses) are located in the interval of $(-0.3, -0.05)$; and the data points in rank 3 (corresponding to images of a woman without glasses) are located in the interval of $(-0.05, 0.4)$. Furthermore, from the selected data points (which are marked by the circles) and their corresponding face images we can see that the manifold structure within each rank, i.e., the pose variation of each person, is preserved in the projected space smoothly. For the test data, although there are overlaps between rank 1 and rank 2, most of the samples are sorted correctly according to their ranks, which means that the proposed algorithm provides a faithful prediction on the test data.

MovieLens Data Set

In this subsection, we compare the proposed method with five algorithms: linear discriminant analysis (LDA) (Fisher 1936), kernel discriminant learning for ordinal regression (KDLOR) (Sun et al. 2010), IsoProjection (Cai, He, and Han 2007), NPE (He et al. 2005), and LPP (He and Niyogi 2004), on the 100k MovieLens data sets, which contains $100,000$ ratings (5 levels: from 1 to 5) for 1682 movies by 943 users. Here LDA is a classical discriminant approach, KDLOR is a competitive ordinal regression method, and IsoProjection, NPE, and LPP are three representative manifold learning algorithms.

In our evaluation, we randomly select 10 users who have rated more than 100 movies. For each of these 10 users, 100 movies are used for training and the rest are used for test. For each movie, its 19 genres are used as the original features and its corresponding rating is used as the label. Therefore, each data point is a 19-dimensional vector.

Table 1 lists the MAE of the six aforementioned algorithms. We name our algorithm as ORML, short for ordinal regression via manifold learning. In this experiment, we set $K = 10$ for IsoProjection, NPE, and LPP, and calculate the average results over the 10 users. By jointly optimizing the order information and the manifold structure, ORML performs better than the discriminant approach, the ordinal regression method, and the manifold learning algorithms.

USPS Digit Data Set

To further evaluate the performance of ORML as well as its multilinear extension, which is denoted as ORML/M, we conduct an experiment on the United State Postal Service (USPS) data set (Hull 1994), which has already been shown containing underlying manifold structure (Zhou et al. 2004). The USPS data set of hand written digital characters comprises 11000 normalized grayscale images of size 16×16, with 1100 images for each of the ten classes: from 0 to 9.

In this experiment, our aim is ranking the data according to the true digit shown in the images. We compare ORML and ORML/M with the following nine agorithms: LDA, multilinear LDA (MLDA) (Yan et al. 2005), IsoProjection, multilinear isometric embedding (MIE) (Liu, Liu, and Chan 2009), NPE, tensor NPE (TNPE) (Dai and Yeung 2006), LPP, tensor LPP (TLPP) (He, Cai, and Niyogi 2006), and KDLOR, where MLDA, MIE, TNPE, and TLPP are the multilinear extensions of LDA, IsoProjection, NPE, and LPP, respectively.

For IsoProjection, MIE, NPE, TNPE, LPP, and TLPP, we set $K = 10$. For each class, p (= 10, 20, 50, 100) images are randomly selected for training and the rest are used for test. We repeat the experiments for 20 times and report the average results. As shown in Table 2, the proposed algorithms out perform other methods in most of the cases. In addition, by considering the tensor structure of image data, ORML/M performs better than ORML in general.

Table 2: The mean absolute error (MAE) of eleven algorithms on USPS data set with different training/test partition sizes.

Train/Test	ORML	LDA	IsoProjection	NPE	LPP	KDLOR
100/10900	2.729 ± 0.220	3.091 ± 0.228	2.960 ± 0.167	2.943 ± 0.174	3.048 ± 0.277	$\mathbf{2.503 \pm 0.105}$
200/10800	$\mathbf{2.466 \pm 0.082}$	2.953 ± 0.131	2.849 ± 0.150	2.846 ± 0.133	2.903 ± 0.195	2.483 ± 0.157
500/10500	$\mathbf{1.998 \pm 0.091}$	2.659 ± 0.166	2.731 ± 0.141	2.634 ± 0.132	2.684 ± 0.120	2.169 ± 0.087
1000/10000	$\mathbf{1.750 \pm 0.041}$	2.425 ± 0.091	2.650 ± 0.090	2.471 ± 0.162	2.600 ± 0.107	1.817 ± 0.034
Train/Test	ORML/M	MLDA	MIE	TNPE	TLPP	
100/10900	$\mathbf{2.652 \pm 0.234}$	2.883 ± 0.186	2.878 ± 0.224	2.990 ± 0.256	2.932 ± 0.290	
200/10800	$\mathbf{2.328 \pm 0.287}$	2.809 ± 0.180	2.757 ± 0.201	2.795 ± 0.314	2.836 ± 0.214	
500/10500	$\mathbf{2.097 \pm 0.217}$	2.584 ± 0.190	2.704 ± 0.232	2.547 ± 0.236	2.655 ± 0.278	
1000/10000	$\mathbf{1.630 \pm 0.204}$	2.461 ± 0.141	2.593 ± 0.172	2.420 ± 0.195	2.391 ± 0.257	

Conclusion

In this paper, we present a novel ordinal regression approach via manifold learning. By taking manifold structure into consideration, the geometry of the data sets is well preserved. By keeping the order information among data groups of different ranking levels, the proposed algorithm provides faithful ordinal regression on new coming data points. By preserving the tensor form of input data, the proposed algorithm offers more general solution to the data with naturally high-order structure. Experiments on several data sets demonstrate that the ORML and ORML/M achieve good performance on the ordinal regression task.

Acknowledgments

This work was supported by grant PolyU 5245/09E.

References

Belkin, M., and Niyogi, P. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS 14*, 585–591.

Cai, D.; He, X.; and Han, J. 2007. Isometric projection. In *AAAI*.

Chu, W., and Keerthi, S. S. 2005. New approaches to support vector ordinal regression. In *ICML*, 145–152.

Crammer, K., and Singer, Y. 2002. Pranking with ranking. In *NIPS 14*, 641–647.

Dai, G., and Yeung, D.-Y. 2006. Tensor embedding methods. In *AAAI*, 330–335.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* 7:179–188.

Frank, E., and Hall, M. 2001. A simple approach to ordinal classification. In *ECML*, 145–156.

Graham, D. B., and Allinson, N. M. 1998. Characterizing virtual eigensignatures for general purpose face recognition. In *Face Recognition: From Theory to Applications, NATO ASI Series F, Computer and Systems Sciences Vol. 163*, 446–456.

He, X., and Niyogi, P. 2004. Locality preserving projections. In *NIPS 16*.

He, X.; Cai, D.; Yan, S.; and Zhang, H.-J. 2005. Neighborhood preserving embedding. In *ICCV*, 1208–1213.

He, X.; Cai, D.; and Niyogi, P. 2006. Tensor subspace analysis. In *NIPS 18*. 499–506.

Hein, M., and Maier, M. 2007. Manifold denoising. In *NIPS 19*.

Herbrich, R.; Graepel, T.; and Obermayer, K. 2000. Large margin rank boundaries for ordinal regression. In *Advances in Large Margin Classifiers*, 115–132. MIT Press.

Hull, J. J. 1994. A database for handwritten text recognition research. *IEEE Trans. Pattern Anal. Mach. Intell.* 16(5):550–554.

Kramer, S.; Widmer, G.; Pfahringer, B.; and De Groeve, M. 2001. Prediction of ordinal classes using regression trees. *Fundam. Inf.* 47:1–13.

Li, L., and Lin, H.-T. 2007. Ordinal regression by extended binary classification. In *NIPS 19*, 865–872.

Li, X.; Lin, S.; Yan, S.; and Xu, D. 2008. Discriminant locally linear embedding with high-order tensor data. *IEEE Trans. SMC-B* 38(2):342–352.

Liu, W., and Chang, S.-F. 2009. Robust multi-class transductive learning with graphs. In *CVPR*, 381–388.

Liu, Y.; Liu, Y.; and Chan, K. C. C. 2009. Multilinear isometric embedding for visual pattern analysis. In *ICCV Subspace Workshop*, 212–218.

Roweis, S., and Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500):2323–2326.

Shashua, A., and Levin, A. 2003. Ranking with large margin principle: two approaches. In *NIPS 15*, 961–968.

Sun, B.-Y.; Li, J.; Wu, D. D.; Zhang, X.-M.; and Li, W.-B. 2010. Kernel discriminant learning for ordinal regression. *IEEE Trans. Knowl. and Data Eng.* 22:906–910.

Tenenbaum, J. B.; de Silva, V.; and Langford, J. C. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290(5500):2319–2323.

Wang, H.; Huang, H.; and Ding, C. H. Q. 2010. Discriminant laplacian embedding. In *AAAI*, 618–623.

Weinberger, K. Q., and Saul, L. K. 2006. An introduction to nonlinear dimensionality reduction by maximum variance unfolding. In *Proc. 21st AAAI*.

Yan, S.; Xu, D.; Yang, Q.; Zhang, L.; Tang, X.; and Zhang, H.-J. 2005. Discriminant analysis with tensor representation. In *CVPR*, volume 1, 526–532.

Zhou, D.; Weston, J.; Gretton, A.; Bousquet, O.; and Schölkopf, B. 2004. Ranking on data manifolds. In *NIPS 16*.