Dimensionality Reduction for Ordinal Classification

Mouad Zine-El-Abidine
LARIS
Université d'Angers
Angers, France
zine_mouadaix@hotmail.fr

Helin Dutagaci
Electrical-Electronics Engineering
Eskisehir Osmangazi University
Eskisehir, Turkey
hdutagaci@ogu.edu.tr

David Rousseau
LARIS
Université d'Angers
Angers, France
david.rousseau@univ-angers.fr

***Abstract*—Many unsupervised and supervised dimension reduction techniques are available for visualization and interpretation of high-dimensional data for classification tasks. While the unsupervised techniques do not employ the class information at all, most supervised algorithms are blind to the order of classes in ordinal classification problems. In this paper, we propose a novel and intuitive dimension reduction technique specifically designed for visualization of high-dimensional features in ordinal classification tasks. The technique is an iterative process, where at each iteration a search is conducted in the high-dimensional space to find the viewpoint from which the centers of adjacent classes are seen most distant from each other. The data is then projected to the lower dimensional space defined by the optimum viewpoint. The iteration is terminated when the desired dimensionality is achieved. Experimental results on various ordinal datasets demonstrate that our technique can be used as a complementary tool to the classical dimensionality reduction methods.**

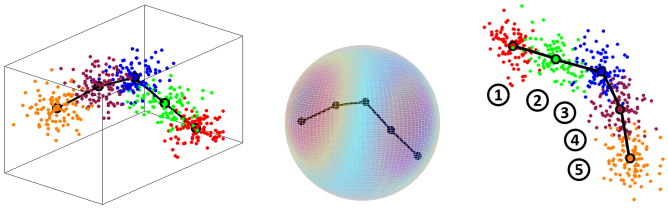
***Index Terms*—ordinal classification, dimensionality reduction, data visualization, interpretability**

I. Introduction

Ordinal classification refers to the classification problems where there is a natural order between categories [1]. Examples to ordinal classification are ranking the severity of a disease [2], prediction of movie preferences [3] or classification of images involving ordinal quantities [4]. The categories are usually represented with one-dimensional discrete values following their inherent order. It is expected that the features used to predict the ordinal categories of the instances also possess an intrinsic order in the high-dimensional space. In order to visualize and assess whether these features follow the ordinality of the categories, dimensionality reduction can be used.

Many dimensionality reduction techniques are available to transform the high-dimensional data into a low-dimensional space for interpretation of the prediction model. Principal component analysis (PCA) is the leading choice. Other unsupervised techniques include multidimensional scaling (MDS) [5], Isomap [6], non-negative matrix factorization (NMF) [7] and t-distributed stochastic neighbor embedding (t-SNE) [8]. When class information of the training data is available, supervised techniques such as Linear Discriminant Analysis (LDA) [9], Kernel Discriminant Analysis (KDA) [10], [11], or Locality Sensitive Discriminant Analysis (LSDA) [12] are more effective in retaining meaningful properties of the data that predict the categories. Although these techniques can be very useful, they do not incorporate the ordinal structure of the categories into their original formulation for ordinal classification problems.

Authors gratefully acknowledge Région des Pays de la Loire and Angers Loire Métropole, France for funding this research.



(a) 3D input data (b) View sphere (c) Projected data

Fig. 1: Ordinal classification problem. The original data is shown in (a). The class centers are enclosed with black circles. The segments joining the class centers are in black color. The objective is to find the optimum viewpoint on the view sphere (b) such that the adjacent class centers are seen as apart as possible. The colors on the sphere are indicative of the objective function defined in 2. The 2D visualization (c) of the data is obtained through projecting the data to the plane defined by the optimum viewpoint.

There is very few work on dimensionality reduction specific to ordinal classification. A variant of Kernel Discrimant Analysis has been proposed [13] to find the projection that simultaneously result in a high separation between classes and maintain the class order. In [14], a supervised method based on sufficient dimension reductions (SDR) is developed to regress the response of underlying Gaussian latent variables to ordered categorical variables. However, this method is more suitable for dimension reduction in regression problems for predictor selection and better prediction rather than visualization of the available features.

In this paper, we propose an intuitive dimensionality reduction technique, which we call *Best-view Projection* (BVP), for ordinal classification without imposing a regression model. The main motivation is to propose a complementary tool to the classical dimensionality reduction techniques, which fail for special configurations of data distribution in ordinal classification tasks. We formulate our approach as finding the best viewpoint in the feature space such that the viewer can "see" the direction of ordinality as clearly as possible.

Inspired by the work for human skeleton visualization in [15], we determine the optimal viewpoint via maximization of the squared distances between centers of adjacent classes in the projected space. The dimensionality is reduced by one (i.e. from N to $N-1$) by projecting the features to the lower dimensional space defined by the optimum viewpoint. The process is repeated until the desired dimensionality is achieved. A major advantage of our method is that it does not require any parameters to be tuned. We provide a comparison of our BVP method with a number of classical dimensionality reduction techniques on simulated and real ordinal datasets.

II. Method

Let us consider an ordinal classification problem illustrated in Fig. 1, where the features are in 3D space and the categories are ordered. We would like to find a viewpoint on the view sphere such that when viewed from that point the adjacent class centers seem as apart as possible from each other. Our BVP method finds the optimum viewpoint that maximizes the projected square distances between adjacent class centers and projects the data points to the space defined by the optimum viewpoint.

To generalize the problem for N-dimensional space, let us first suppose that we have K classes, ordered and identified as $k = 1, 2, ..., K$. A class l is adjacent to class k if $l = k-1$ or $l = k+1$. The instances of class k are represented as N-dimensional column vectors denoted as $x_i^k \in \mathbb{R}^N$, with $i = 1, 2, ..., I_k$, where I_k is the number of instances in class k. The class centers are denoted as c_k corresponding to the arithmetic mean of the instances in class k. For the sake of simplifying the equation of the view sphere, the data is translated beforehand such that the origin of the N-dimensional space corresponds to $\frac{1}{K}\sum_{k=1}^{K} c_k$, i.e. the mean of the class centers.

Let us define the n-sphere ($n = N-1$) in the N-dimensional space as $\mathcal{S} = \{v \in \mathbb{R}^N : \|v\| = 1\}$. Given a viewpoint $v \in \mathcal{S}$, we can define an orthogonal projection $P : \mathbb{R}^N \to \mathbb{R}^N$, whose $N-1$ columns are defined by the vectors orthonormal to v, and whose last column is equal to v. Then, a point $x \in \mathbb{R}^N$ can be projected to the $N-1$-dimensional space defined by v by computing $y = Px$ and dropping the last component of y. This point, $\bar{x}(v) \in \mathbb{R}^{N-1}$ can be interpreted as point x as seen from the viewpoint v. Its component parallel to v is invisible to the viewer.

Our objective is to find the viewpoint v^* on the n-sphere such that the sum of the squared distances between the centers of the adjacent classes is maximized. If we define $\bar{c}_k(v) \in \mathbb{R}^{N-1}$ to be the projected center of class k in the $N-1$-dimensional space defined by viewpoint v, we search for v^* maximizing

$$G(v) = \sum_{k=1}^{K-1} \|\bar{c}_{k+1}(v) - \bar{c}_k(v)\|^2 \tag{1}$$

Algorithm 1: Find the optimum viewpoint

Data: Class centers: c_k, $k = 1, 2, ..., K$
Result: Optimum viewpoint: v^*
Initialize v_0 randomly such that $\|v_0\| = 1$;
MaxIter $= 100$; $\epsilon = 10^{-5}$; $\gamma_0 = 0.05$; $j = 0$;
while $j <$ *Maxiter* **do**
  Calculate $\nabla F(v_j)$ using 5 and 6;
  if $j > 0$ **then**
    Calculate γ_j using 7;
  end
  $\hat{v}_j = v_j - \gamma_j \nabla F(v_j)$;
  $v_{j+1} = \frac{\hat{v}_j}{\|\hat{v}_j\|}$;
  if $\cos^{-1}(v_{j+1}^T v_j) < \epsilon$ **then**
    $v^* = v_{j+1}$;
    break;
  end
  $j \leftarrow j+1$
end

subject to the constraint $\|v\| = 1$. Maximizing G(v) is equivalent to solving the following minimization problem:

$$\text{Minimize } F(v) = \sum_{k=1}^{K-1} [v^T(c_{k+1} - c_k)]^2 \text{subject to } \|v\| = 1 \tag{2}$$

This is an optimization problem where the search space is constrained to a smooth Riemannian manifold. We use gradient descent together with the retraction formulation for a spherical manifold in [16] to find v^*. The procedure is given in Algorithm 1. We randomly pick a viewpoint v_0 on $\mathcal{S}$ for initialization and update it as:

$$v_{j+1} = \text{Retr}_{v_j}(\eta_j) \tag{3}$$

$$\eta_j = -\gamma_j \nabla F(v_j) \tag{4}$$

The gradient of $F(v)$ is equal to:

$$\nabla F(v) = 2A^T A v \tag{5}$$

$$A = \begin{bmatrix} (c_2 - c_1)^T \\ (c_3 - c_2)^T \\ \vdots \\ (c_K - c_{K-1})^T \end{bmatrix} \tag{6}$$

We update step size γ_j according to the formula [17]:

$$\gamma_j = \frac{\left|(v_j - v_{j-1})^T |\nabla F(v_j) - \nabla F(v_{j-1})|\right|}{\|\nabla F(v_j) - \nabla F(v_{j-1})\|^2} \tag{7}$$

The retraction for the sphere can be chosen as [16]:

$$\text{Retr}_v(\eta) = \frac{v + \eta}{\|v + \eta\|} \tag{8}$$

The iteration is stopped when the angle between v_j and v_{j+1} is smaller than ϵ, and the optimum viewpoint v^* is set to be equal to v_{j+1}. For all experiments, ϵ is set to 10^{-5}.

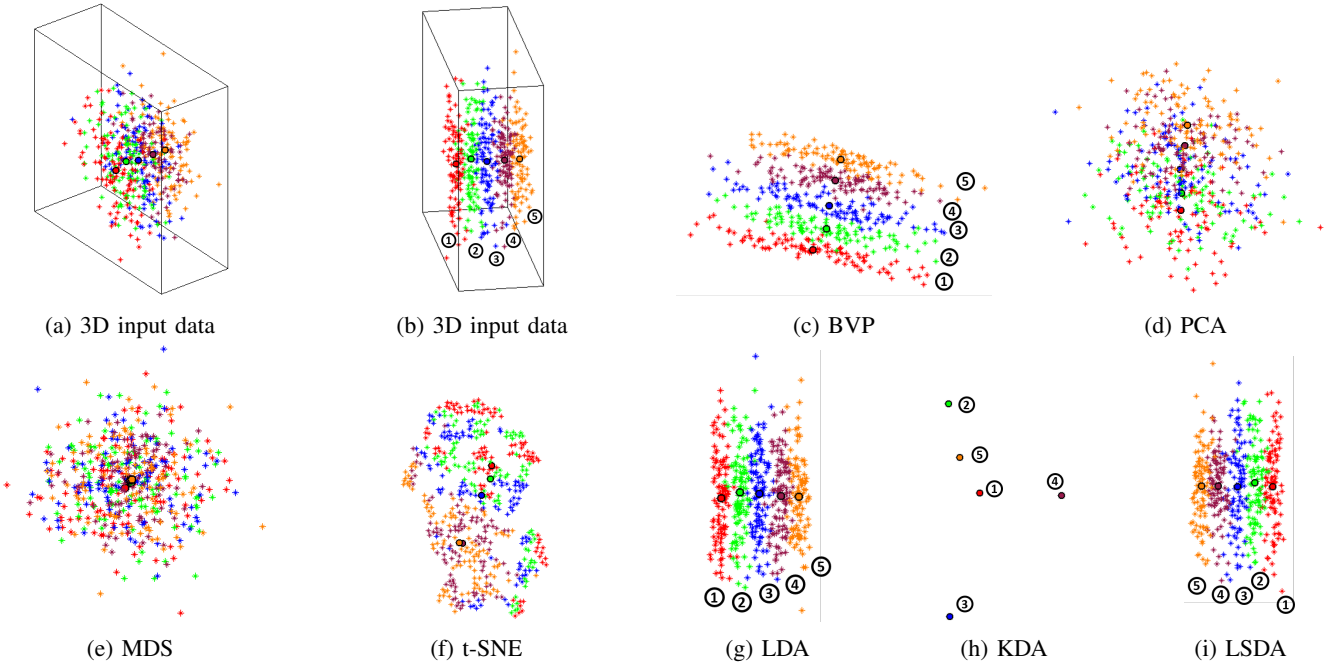


Fig. 2: Proposed best view point (BVP) algorithm in action with ordinal datasets by comparison with other classical dimensionality reduction techniques. Panels (a) and (b) are two views of the synthetic 3D ordinal data set. Panels (c) to (i) show results of dimensionality reduction from 3D to 2D. The black circles with numbers correspond to class labels.

The cluster centers are then projected to the $N-1$-dimensional space defined by v^*. The whole procedure is repeated until the desired dimensionality is achieved.

III. Results

Since we present our dimensionality reduction method as a complementary visualization tool for ordinal datasets rather than reducing the dimension of inputs of classification techniques, we provide visual results only. We compare the best-view projection method with three unsupervised methods (1) PCA, 2) MDS, 3) T-SNE) and three supervised methods (4) LDA, 5) KDA, 6) LSDA). The comparison is based on: 1) whether the classes are well-separated, 2) whether the ordinality between classes is preserved, and 3) whether the distribution of the data is informative in the low-dimensional space. First, we present results with simulated data where the dimensionality is reduced from 3 to 2. Then, we provide comparisons with real ordinal datasets of higher initial dimensions.

A. Simulated Data

We created two 3-dimensional ordinal datasets and employed our best-view projection method and other six algorithms to reduce the dimensionality to 2. Fig. 2a and 2b show the first dataset, where there are five ordinal classes and the within class distribution is Gaussian. 100 instances were generated for each class. The class labels are given in Fig. 2a. In this example, the direction of ordinality in the original space is not aligned with the principal axes of variation of the whole data; hence PCA fails to appropriately reduce the dimensionality as seen in Fig. 2d. MDS tries to place data points into 2D space such that the pairwise distances are preserved as much as possible. It does not take into account the class labels, and it fails when the instances of adjacent classes are close to each other in the original space (Fig. 2e). In t-SNE, local neighborhood of points are embedded to capture the local structure of the data together with clusters at several scales. We experimented throughly with varying the perplexity parameter, which is a measure of the effective number of neighbors used in the algorithm. We give the best result, with perplexity 40, in Fig. 2f. Although, t-SNE manages to group together samples of the same class in local clusters, the global ordinality present in the data is lost in the resulting 2D space.

Notice that PCA, MDS and t-SNE are unsupervised techniques, which are great for revealing important global or local structure of data. However, the class distribution is not necessarily aligned with that structure in many cases, as in this example. The supervised techniques, LDA (Fig. 2g) and LSDA (Fig. 2i) are able to reduce the dimensionality to 2 with good class separation while preserving the ordinality. Our BVP algorithm performs very similarly to LDA and LSDA (Fig. 2c). What it does is essentially to rotate of the 3D data given in Fig. 2a until it finds the best view that separates the adjacent cluster centers as well as possible. For KDA, we used a Gaussian kernel and set the regularization parameter to 0.1. KDA, a supervised technique, compressed the instances of each class to a small local region in the 2D space through class-based Gaussian kernels for this particular case (Fig. 2h). Although the classes seem well-separated, the ordinality relationship is lost and the within class distribution of the data is not

observable in the new space.

The second simulated dataset is shown in Fig. 3a and Fig. 3b. Instances of each class belong to a 3D partial swiss roll. 200 instances were generated for each class. The classes are separated by an offset in 3D in accordance to their ordinality. Similar to the first dataset, PCA is not effective (Fig. 3d) since the principal axis of global data distribution is not aligned with the direction of ordinality. MDA is successful in this case Fig. (3e) due to the fact that there is greater separation between instances from different classes in the original space as compared to the first dataset. For t-SNE, the best configuration was obtained with perplexity 30. In accordance with its objective, t-SNE gathered the data in local clusters in the 2D space, preserving the separability and ordinality to some degree Fig. 3f; however, the global nature of the data is not observable.

For this dataset, we observe that LDA failed to reduce the dimensionality properly (Fig. 3g). LDA searches for a projection that minimizes the distances of instances of each class to its center, and in this case, the class centers are located closer to the instances of adjacent classes in the original space. The result is a 2D configuration where the class separation is lost. Our BVP method does well in this case (Fig. 3c) as does LSDA (Fig. 3c), projecting the data such that the separation and ordinality between classes are preserved together with an informative distribution in the reduced space. KDA does not reveal such global information (Fig. 3g).

B. Real Ordinal Data

We tested our dimensionality reduction technique on real ordinal classification datasets [18], [19]. Due to lack of space, we only give visual comparisons with the two supervised techniques; LDA and LSDA. The datasets and their properties are given in Table I.

TABLE I: Real ordinal datasets used for the experiments [18], [19] (I is the total number of instances, Q is the dimensionality of the original data and K is the number of classes).

Dataset	I	Q	K	**Class Distribution**
contact-lenses	24	6	3	(15,5,4)
pasture	36	25	3	(12,12,12)
squash-stored	52	51	3	(23,21,8)
newthyroid	215	5	3	(30,150,35)
car	1728	21	4	(1210,384,69,65)
bondrate	57	37	5	(6,33,12,5,1)

Figs. 4 through 9 show the dimensionality reduction results obtained by BVP in comparison to LDA and LSDA. For all cases, BVP was able to provide a glimpse of the data distribution and relative relations of the classes with respect to each other. LDA showed good performance in some cases (Figs. 4b and 7b). In the other cases, LDA pulled the instances close to the class centers, causing loss of information on within class distribution. Not originally designed for dimensionality reduction for visualization, LSDA did not retain class separability in 2 dimensions for the real datasets, except for the dataset newthyroid (Fig. 7c). These results demonstrate that for many ordinal datasets, classical dimensionality reduction techniques may fail to provide a proper visualization of the high-dimensional features, and our method can be used as an alternative tool to map high-dimensional data to 2D for interpretation.

IV. Conclusion

The paper presents a new and intuitive technique for the visualization of high-dimensional data for ordinal classification. We provided visual comparisons with various dimensionality reduction techniques on both simulated and real datasets and demonstrated that our technique is capable of retaining separability and class order in cases where the other techniques failed. This visualisation step is important for ordinal classification to guaranty that the latent space, on which a final classification is to be performed by a machine, is interpretable to a human eye [20]. As future work, the objective function can be extended to involve the pairwise distances of instances of adjacent classes. Another direction is to use this technique to find the optimum dimension to provide input to ordinal classification methods.

V. Acknowledgement

Authors are grateful to the EU project H2020 INVITE under grand agreement Number 817970

References

[1] E. Frank and M. Hall, *A Simple Approach to Ordinal Classification*. Berlin, Heidelberg: Springer-Verlag, 2001, p. 145–156.

[2] X. Liu, Y. Zou, Y. Song, C. Yang, J. You, and B. V. K Vijaya Kumar, "Ordinal regression with neuron stick-breaking for medical diagnosis," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[3] J. Bobadilla, F. Ortega, A. Hernando, and J. Bernal, "Generalization of recommender systems: Collaborative filtering extended to groups of users and restricted to groups of items," *Expert Systems with Applications*, vol. 39, no. 1, pp. 172–186, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417411009675

[4] J. Xie and C. Pun, "Deep and ordinal ensemble learning for human age estimation from facial images," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2361–2374, 2020.

[5] I. Borg and P. Groenen, *Modern Multidimensional Scaling: Theory and Applications*. Springer, 2005.

[6] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000. [Online]. Available: https://science.sciencemag.org/content/290/5500/2319

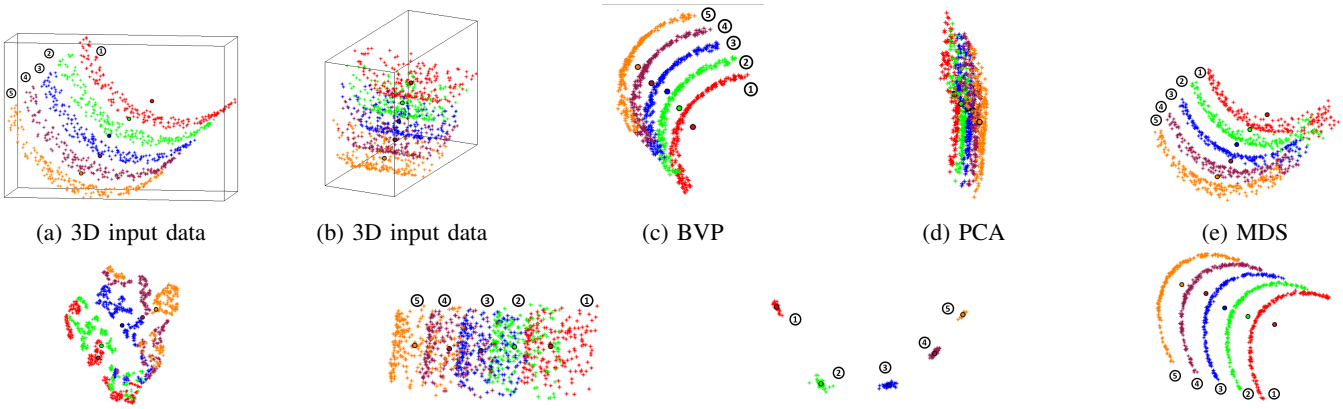
[7] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[8] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

[9] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[10] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. R. Mullers, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (Cat. No.98TH8468)*, 1999, pp. 41–48.

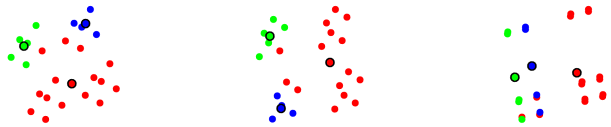
[11] D. Cai, X. He, and J. Han, "Speed up kernel discriminant analysis," *The VLDB Journal*, vol. 20, no. 1, pp. 21–33, 2011.



(a) 3D input data (b) 3D input data (c) BVP (d) PCA (e) MDS

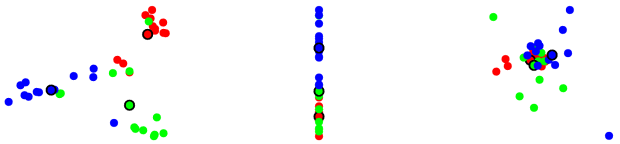
(f) t-SNE (g) LDA (h) KDA (i) LSDA

Fig. 3: Same as Fig. 2 with an ordinal dataset composed of partial swiss rolls.



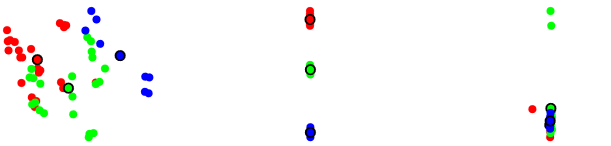
(a) BVP (b) LDA (c) LSDA

Fig. 4: 2D visualization of the dataset contact-lenses



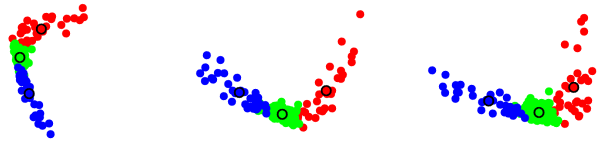
(a) BVP (b) LDA (c) LSDA

Fig. 5: 2D visualization of the dataset pasture



(a) BVP (b) LDA (c) LSDA

Fig. 6: 2D visualization of the dataset squash-stored

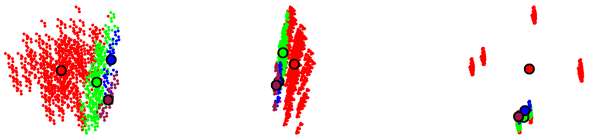


(a) BVP (b) LDA (c) LSDA

Fig. 7: 2D visualization of the dataset newthyroid

(a) BVP (b) LDA (c) LSDA

Fig. 8: 2D visualization of the dataset bondrate



(a) BVP (b) LDA (c) LSDA

Fig. 9: 2D visualization of the dataset car

[12] D. Cai, X. He, K. Zhou, J. Han, and H. Bao, "Locality sensitive discriminant analysis," in *International Joint Conference on Artificial Intelligence (IJCAI'07)*, 2007.

[13] B. Sun, J. Li, D. D. Wu, X. Zhang, and W. Li, "Kernel discriminant learning for ordinal regression," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 6, pp. 906–910, 2010.

[14] L. Forzani, R. Garcia Arancibia, P. Llop, and D. Tomassi, "Supervised dimension reduction for ordinal predictors," *Computational Statistics & Data Analysis*, vol. 125, pp. 136–155, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016794731830080X

[15] B. Kwon, J. Huh, K. Lee, and S. Lee, "Optimal camera point selection toward the most preferable view of 3-d human pose," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–21, 2020.

[16] N. Boumal, P. Absil, and C. Cartis, "Global rates of convergence for nonconvex optimization on manifolds," *IMA Journal of Numerical Analysis*, vol. 39, no. 1, pp. 1–33, Jan. 2019.

[17] J. Barzilai and J. M. Borwein, "Two-Point Step Size Gradient Methods," *IMA Journal of Numerical Analysis*, vol. 8, no. 1, pp. 141–148, 01 1988. [Online]. Available: https://doi.org/10.1093/imanum/8.1.141

[18] J. Sánchez-Monedero, P. A. Gutiérrez, and M. Pérez-Ortiz, "Orca: A matlab/octave toolbox for ordinal regression," *Journal of Machine Learning Research*, vol. 20, no. 125, pp. 1–5, 2019. [Online]. Available: http://jmlr.org/papers/v20/18-349.html

[19] P. A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernández-Navarro, and C. Hervás-Martínez, "Ordinal regression methods: Survey and experimental study," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 1, pp. 127–146, 2016.

[20] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019.