Focus Article

Learning ordinal data

Xingye Qiao*

Classification is an important topic in statistical learning. The goal of classification is to build a predictive model from the training dataset for the class label of an observation. It is commonly assumed that the class labels are unordered. However, in many real applications, there exists an intrinsic ordinal relation between the class labels. Examples of these include cancer patients grouped in early, mediocre, and terminal stages, customers grouped into low, middle, and high credit levels, and experimental subjects enriched with different amounts of bacterial. In this article, we focus on the classification problem for ordinal data and introduce the theoretical setup of the problem. We review both traditional and modern methods in learning ordinal data. In particular, we emphasize the trade-off between model flexibility and interpretability. Lastly, we discuss some issues regarding ordinal data learning, including an appropriate loss function for this problem.

How to cite this article:
WIREs Comput Stat 2015, 7:341–346. doi: 10.1002/wics.1357

Keywords: classification; multivariate analysis; regression; statistical computing; support vector machine

INTRODUCTION

A classification problem is often started with a training dataset $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ where $(\mathbf{x}_i, y_i)$ s are independent and identically distributed observations from an unknown distribution $\mathbb{P}$ of $(\mathbf{X}, Y)$, $\mathbf{x}_i \in \mathbb{R}^p$ is a p-dimensional covariate vector and $y_i \in C$ is the class label, with C the set of all possible labels. The goal of any classification method is to build a classification rule (classifier) ϕ that maps a future observation $\mathbf{x}$ (whose class label is unknown) to $\phi(\mathbf{x}) \in C$ so that one may use $\phi(\mathbf{x})$ to predict the class label of $\mathbf{x}$. In the regular classification setting, C includes a set of unordered class labels, such as *A*, *B*, *C*, or *Asia*, *Europe* and *America*. There is a very large literature on classification. See Refs 1 and 2 for an overall introduction. Regular classification is an important problem of its own and there are many different treatments for different settings, ranging from the more traditional k -nearest neighbor classifier[3,4], Fisher's linear discriminant analysis[5] and logistic regression, to the more machine learning-based support vector machine (SVM),[6] neural networks, and classification trees. In particular, margin-based methods, including SVM, are very popular and many articles have studied various related issues, such as Fisher consistency,[7,8] multicategory margin-based methods,[9–12] angle margin-based methods,[13,14] the surrogate loss,[15] the use of margin based methods in probability estimation,[16] classification with a reject option,[17,18] the excess risk,[19,20] variable selection through sparsity,[21] and the high-dimensional, low-sample size asymptotics of margin-based methods.[22,23] These margin-based methods have heavily influenced the development of new approaches in statistical learning.

In this article, we focus on a more specialized case in classification where the class labels are ordered, that is, the ordinal data. Examples of ordinal data are ubiquitous. For instance, cancer patients can be categorized to Stage I (early stage), II, III, and IV (the terminal stage), and a predictive model (a classifier) may be needed to make cancer prognosis. A simple approach to ordinal data is to convert ordered labels to numerical values, such as to convert $\{I, II, III, IV\}$ to $\{1, 2, 3, 4\}$, so that a regression method can be applied to the converted dataset. In most situations, however, the scale of the difference between classes

*Correspondence to: qiao@math.binghamton.edu

Department of Mathematical Sciences, Binghamton University, Binghamton, NY, USA

Conflict of interest: The author has declared no conflicts of interest for this article.

may not be identical, unlike the equally spaced integers $\{1, 2, 3, 4\}$. An alternative is to use a predefined function $h : C \mapsto \mathbb{R}$ to convert the class labels to more general numerical values than positive integers. However, how to define this coding function is unclear. This article hence focuses on more sophisticated approaches to the ordinal data classification problem. In the sequel, we first formally define the problem and introduce some theoretical considerations, followed by an introduction to some traditional methods. Next, we introduce several machine learning methods in the literature. Some methods have motivated an appropriate loss function for this problem. Lastly, we conclude the article by stressing the issues of flexibility and interpretability.

SETTING AND CLASSIC METHODS

Similar to regular (multiclass) classification problems, an ordinal classification problem may have the goal of minimizing the misclassification error for a future (random) observation $(\mathbf{X}, Y)$, also known as the generalization error, defined as $\text{pr}(\phi(\mathbf{X}) \neq Y)$. $P_c(\mathbf{x}) := pr(Y = c|\mathbf{X} = \mathbf{x})$ is denoted as the conditional probability for class c. The Bayes optimal classifier is the one with the smallest conditional prediction error given any $\mathbf{X} = \mathbf{x}$, that is $\phi_{\text{Bayes}}(\mathbf{x}) = argmin_{c \in C} pr(Y \neq c|\mathbf{X} = \mathbf{x}) = \text{argmax}_{c \in C} pr(Y = c|\mathbf{X} = \mathbf{x}) = \text{argmax}_{c \in C} P_c(\mathbf{x})$.

Hence if one knew the true underlying distribution $\mathbb{P}$, they can calculate the class conditional probability $P_c(\mathbf{x})$ (regardless of the order information) for each class c and make the optimal decision. However, as the distribution is unknown, one may seek to estimate the class conditional probabilities. There is a large literature on this topic (see, for example, Refs 16, 24–27) but we will omit it here and instead focus on approaches that do not rely on probability estimation. We stress, however, that the k -nearest neighbor method could be viewed as a means to estimate the class conditional probability, and hence falls into this category.

Remark

The generalization error is associated with the $0-1$ loss which counts loss 1 if there is a misclassified observation. Note that no ordinal information is incorporated into the $0-1$ loss. Hence one may want to consider more specific loss functions that incorporate the ordinal information. We will discuss two such examples in the next section. In the literature, Huhn and Hullermeier[28] have conducted an investigation to show that incorporating ordinal information often helps boosting the classification performance.

In addition to methods that rely on probability estimation, ordered logistic regression is another traditional method for ordinal data.[29] Different from multinomial logistic regression (which is a generalization of logistic regression from binary classification to the multicategory case), ordered logistic regression fully incorporates the order information. For an illustration, suppose that there are four ordered classes, $1, \ldots, 4$. The log-odd (or logit) of probability p of an event is defined as $\text{logit}(p) = \log[p/(1-p)]$. Consider the event that an observation belongs to the meta-class $\{1, \ldots, j\}$, that is $\{Y \in \{1, \ldots, j\}\}$, whose probability is $\sum_{k=1}^{j} P_k(\mathbf{x})$. The ordered logistic regression model assumes that the log-odds are

$$\begin{aligned}
\text{logit}\left(P_1(\mathbf{x})\right) &= \log\left(\frac{P_1(\mathbf{x})}{P_2(\mathbf{x}) + P_3(\mathbf{x}) + P_4(\mathbf{x})}\right) \\
&= \mathbf{x}^T \beta + b_1 \\
\text{logit}\left(P_1(\mathbf{x}) + P_2(\mathbf{x})\right) &= \log\left(\frac{P_1(\mathbf{x}) + P_2(\mathbf{x})}{P_3(\mathbf{x}) + P_4(\mathbf{x})}\right) \\
&= \mathbf{x}^T \beta + b_2 \\
\text{logit}\left(P_1(\mathbf{x}) + P_2(\mathbf{x}) + P_3(\mathbf{x})\right)& \\
&= \log\left(\tfrac{P_1(\mathbf{x}) + P_2(\mathbf{x}) + P_3(\mathbf{x})}{P_4(\mathbf{x})}\right) = \mathbf{x}^T \beta + b_3
\end{aligned}$$

That is, it models the logit of the sum of class condition probabilities using linear functions of the covariates $\mathbf{x}$. However, note that the coefficient vectors β in all three equations above are the same; only their intercept terms are different. As a consequence, the boundaries that divide classes 1 and 2, 2 and 3, and 3 and 4 are hyperplanes parallel to each other. This lack of flexibility may be undesirable in some applications. Moreover, it is unclear whether the linear function is the right way to model the logits.

SOME MACHINE LEARNING METHODS

In this section, we introduce several machine learning-oriented methods for learning ordinal data. This review is not exclusive and omission of any reference does not reflect the lack of importance of the work.

Herbrich et al.[30] used the principle of structural risk minimization to formulate an ordinal learning problem based on a novel loss function. This loss function is defined based on pairs of inter-class observations. In particular, a loss is generated if the learning algorithm mistakenly assigns the wrong ranks between the two observations from two different classes. However, a drawback of this approach, despite its novelty, is that for relatively large sample with n observations, the computational burden can be very heavy. For example, suppose there are k classes with n_0 observations in each class (hence a total of $n = kn_0$ observations), the empirical loss entails $k\left(k-1\right)/2 \times n_0^2$ pairs, that is, the actual number of terms involved in the optimization problem is in the order of $O(n^2)$.

Frank and Hall[31] converted an ordinal classification problem to several nested binary classification problems. For each $j \in \{1, \ldots, k-1\}$, each observation is classified between the meta-class that includes classes 1 to j and the meta-class that includes classes $j+1$ to k. The final classification result can be inferred from the $k-1$ binary predictions. For example, suppose $k=4$, then three binary classification problems with 1 versus $\{2, 3, 4\}$, $\{1, 2\}$ versus $\{3, 4\}$ and $\{1, 2, 3\}$ versus 4 are considered. An observation which is classified to all three meta-classes that contain class 3 would be unarguably classified to class 3. In contrast to the work of Herbrich et al.[30], this method is very straightforward, and requires no modification to existing binary classification methods. However, because the three classification boundaries are trained separately, sometimes the boundaries may cross with each other. In particular, it is possible that an observation is classified to class 1 when comparing with meta-class $\{2, 3, 4\}$, and is classified to meta-class $\{3, 4\}$ when comparing with meta-class $\{1, 2\}$. In this case, no sensible classification result can be obtained for this observation.

There is another group of methods which enjoys great computational convenience. These methods typically consider mapping observations to the real line using a function $g(\mathbf{x})$ trained from the data. Ordinal classification is then conducted by thresholding the mapped observations $g(\mathbf{x}_i)$ using a series of $(k-1)$ cut-off numbers, $b_1 \le b_2 \le \ldots \le b_{k-1}$. For example, Shashua and Levin[32] generalized the SVM formulation for ordinal regression. In particular, a common classification direction vector β is found which maps observation $\mathbf{x}$ to $\mathbf{x}^T\beta$. With the thresholds $b_1 \le b_2 \le \ldots \le b_{k-1}$, this method has induced $k-1$ parallel separating hyperplanes defined by $\{\mathbf{x} : \mathbf{x}^T\beta = b_j\}$ for $j = 1, \ldots, k-1$, which divide the sample space into k parts, one for each class. An optimization problem is solved to encourage large margins from observations to these hyperplanes. For example, for the jth hyperplane, observations in classes j and $j+1$ are treated as if they are in a new binary SVM problem, with class j the new negative class and class $j+1$ the new positive class. The Hinge loss is evaluated for each subproblem using the slack variable technique as in the regular binary SVM method. The $k-1$ binary SVM subproblems are jointly trained with the constraint that they have the same coefficient vectors β. A kernel version is also possible for more flexibility.

However, Chu and Keerthi[33] pointed out that the thresholds obtained in Shashua and Levin[32] may not satisfy the natural ordering constraint $b_1 \le b_2 \le \ldots \le b_{k-1}$. This may be because of the fact that class j' ($j' \le j-1$ or $j' \ge j+2$) does not directly contribute to the fitting of the jth hyperplane. In this case, similar to the crossing-boundary issue in Frank and Hall[31], the classification result may be difficult to infer. Hence, Chu and Keerthi[33] proposed an improvement which can implicitly enforce ordered thresholds, and hence provide more sensible separating hyperplanes.

Chu and Ghahramani[34] introduced a Bayesian approach where the latent score $g(\mathbf{x}_i)$ for the ith observation is modeled from a Gaussian process while the covariance between the scores is defined by Mercer kernel functions. Given the score, the probability that an observation belongs to a class is modeled by the difference of two Gaussian cumulative distribution functions, which generalize the *probit* function in the binary case to the multiclass setting.

Lastly, Cardoso and da Costa[35] introduced another simple adaption of existing methods in order to handle ordinal data. In particular, for each training data observation $(y_i, \mathbf{x}_i)$, they propose the replicated and augmented data $\left\{\left(\tilde{y}_i^j, \tilde{\mathbf{x}}_i^j\right), j = 1, \ldots, k-1\right\}$, where $\tilde{y}_i^j = \text{sign}\left(y_i - j - 1/2\right) \in \{-1, +1\}$ and $\tilde{\mathbf{x}}_i^j = \left(\mathbf{x}_i; \mathbf{e}_{j-1}\right)$ is the augmented vector with $\mathbf{e}_\ell \in \mathbb{R}^{k-2}$ a $(k-2)$-dimensional vector with $h > 0$ on the ℓth element and 0 elsewhere. Hence the dataset has been replicated for $k-1$ times and the $\mathbf{x}$ vector has been augmented by $k-2$ additional covariates. Here $\tilde{y}^j$ is the binary class label (-1 or $+1$) for a binary subproblem involving meta-classes $\{1, \ldots, j\}$ and $\{j+1, \ldots, k\}$. To solve all these binary subproblems together, an existing binary classification method can be applied to the replicated and augmented dataset, which now includes $(k-1)n$ data points with $p+k-2$ dimensions. A final decision for classification is made from pooling decisions from the binary classification for the replicated and augmented data.

The method of Cardoso and da Costa[35] shares the similar idea to that of Frank and Hall.[31] In particular, Cardoso and da Costa[35] tried to solve $(k-1)$ binary classification problems, like Frank and Hall,[31] except that it was done by a single classifier. Each of the additional covariates counts for a threshold. A major difference is that the method of Cardoso and da Costa[35] calls for a common direction among all the boundaries while the boundaries in Frank and Hall[31] are trained independently and hence are more flexible.

It is interesting to note that the method of Cardoso and da Costa[35] is also similar to ordered logistic regression. In particular, both methods implicitly assume that the boundaries are parallel in the feature space. In parameter estimation, the ordered logistic regression is model-based, while Cardoso and da Costa[35] can be more flexible.

The idea of both Frank and Hall and Cardoso and da Costa[31,35] is to solve the ordinal classification problem via several binary classification problems (whether separately or jointly). We provide a theoretical insight to this framework. As was discussed earlier, the generalization error is associated with the $0-1$ loss which does not consider the ordinal information. Let us consider an improved loss function that does. For each class label y, $\widetilde{y}^j = \mathrm{sign}(y-j-1/2)$ is denoted as the converted class code which is the binary class label (-1 or $+1$) for the jth binary subproblem involving meta-classes $\{1, \ldots, j\}$ and $\{j+1, \ldots, k\}$. If we aggregate the $0-1$ losses for these $k-1$ binary problems together, then one can show that the sum of these $0-1$ losses is the same as the distance loss, that is,

$$L\left(\boldsymbol{x}, y, \phi\right) = \left|y - \phi\left(\boldsymbol{x}\right)\right| = \sum_{j=1}^{k-1} 1_{\{\widetilde{y}^j \neq \mathrm{sign}\left[\phi\ (\boldsymbol{x}) - j - 1/2\right]\}}$$
$$= \sum_{j=1}^{k-1} L_{0-1}\left(\boldsymbol{x}, \widetilde{y}^j, \widetilde{\phi}^j\right),$$

where $\widetilde{\phi}^j$ is defined as $\mathrm{sign}(\phi - j - 1/2)$. Note that $\widetilde{\phi}^j$ can be viewed as the induced decision of ϕ for the jth binary sub-problem involving meta-classes $\{1,\ldots,j\}$ and $\{j+1, \ldots, k\}$. Hence an ordinal classification method which considers the $k-1$ binary problems altogether may be viewed as an attempt to minimize the distance loss defined as $|y-\phi(\boldsymbol{x})|$. Compared to the $0-1$ loss, the distance loss has incorporated the ordinal information. In particular, misclassifying a class 1 observation to class 10 would cost much more than misclassifying it to class 2, where the latter case seems to be a less severe misclassification to make. In contrast, a simple $0-1$ loss treats these two types of misclassification equally. This could be problematic in multicategory classification especially if the classes are imbalanced in the sample size.[36]

The distance loss can be easily generalized to a weighted version, counting for different levels of cost for misclassifying an observation from meta-class $\{1, \ldots, j\}$ to $\{j+1, \ldots, k\}$.

CONCLUSION

When modeling ordinal data, especially high-dimensional ones, there seem to be two concerns that compete with each other. On one end, the model is supposed to be simple and easy to interpret. The classic ordered logistic regression would be a good example of this kind. However, the simple linear model may not reflect the true relation between the covariate and the class conditional probability, and the assumption that the coefficients are identical among different classes may not be ideal in some cases. Hence, on the other end, for improved classification performance, some flexibility is desired. Many of the machine learning approaches mentioned above allow for flexible boundaries that are beyond the linear model. However, if the model gets too flexible, such as in Frank and Hall[31] and Shashua and Levin[32], the classification boundaries for two adjacent binary classification problems may cross with each other, leading to ill-classification results that are difficult to handle. This in turn leads to a difficulty in interpretation. Several aforementioned works have been dedicated to strategies that can help avoid crossings of the classification boundaries (or unnatural ordering of the thresholds). In the ideal world, one would wish to amplify the flexibility to the greatest extend while keeping the boundaries noncrossing to maintain certain straightforwardness and interpretability [see a related work in Qiao[37]]. We see that the general framework to consider $(k-1)$ binary subproblems can be viewed as to minimize the expected distance loss, which is specific and desirable for the ordinal classification problem.

ACKNOWLEDGMENTS

This work was partially supported by a grant from Simons Foundation (246649 to Xingye Qiao).

REFERENCES

1. Duda R, Hart P, Stork D. *Pattern Classification*. New York, NY: John Wiley & Sons; 2001.
2. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer; 2009.
3. Fix E, Hodges JL. Discriminatory analysis, nonparametric discrimination: consistency properties. Technical Report 4, Project 21-49-004, Contract AF41(128)-31, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.
4. Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Trans Inf Theory* 1967, 130:21–27.
5. Fisher R. The use of multiple measurements in taxonomic problems. *Ann Hum Genet* 1936, 70:179–188.
6. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995, 200:273–297.
7. Lin Y. A note on margin-based loss functions in classification. *Stat Probab Lett* 2004, 680:73–82.
8. Tewari A, Bartlett PL. On the consistency of multiclass classification methods. *J Mach Learn Res* 2007, 8:1007–1025.
9. Lee Y, Lin Y, Wahba G. Multicategory support vector machines. *J Am Stat Assoc* 2004, 990:67–81.
10. Huang H, Liu Y, Du Y, Perou CM, Hayes DN, Todd MJ, Marron JS. Multiclass distance weighted discrimination. *J Comput Graph Stat* 2013, 220:953–969.
11. Weston J, Watkins C. Support vector machines for multi-class pattern recognition. In: *Proceedings of European Symposium on Artificial Neural Networks*, Bruges, Belgium, D-Facto Public, 21–23 April, 1999, 219–224.
12. Crammer K, Singer Y. On the learnability and design of output codes for multiclass problems. *Mach Learn* 2002, 47:201–233.
13. Zhang C, Liu Y, Wang J, Zhu H. Reinforced angle-based multicategory support vector machines. *J Comput Graph Stat*. Forthcoming.
14. Lange K, Wu TT. An MM algorithm for multicategory vertex discriminant analysis. *J Comput Graph Stat* 2008, 170:527–544.
15. Nguyen X, Wainwright MJ, Jordan MI. On surrogate loss functions and f-divergences. *Ann Stat* 2009, 37:876–904.
16. Wang J, Shen X, Liu Y. Probability estimation for large-margin classifiers. *Biometrika* 2008, 950:149.
17. Yuan M, Wegkamp MH. Classification methods with reject option based on convex risk minimization. *J Mach Learn Res* 2010, 11:111–130.
18. Bartlett PL, Wegkamp MH. Classification with a reject option using a hinge loss. *J Mach Learn Res* 2008, 9:1823–1840.
19. Zhang T. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann Stat* 2004, 32:56–85.
20. Bartlett PL, Jordan MI, McAuliffe JD. Convexity, classification, and risk bounds. *J Am Stat Assoc* 2006, 101:138–156.
21. Zhu J, Rosset S, Hastie T, Tibshirani R. 1-norm support vector machines. In: Thrun S, Saul LK, Schölkopf B, eds. *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference*. Cambridge, MA: MIT Press; 2003.
22. Hall P, Marron JS, Neeman A. Geometric representation of high dimension, low sample size data. *J R Stat Soc Series B Stat Methodol* 2005, 670:427–444.
23. Qiao X, Zhang H, Liu Y, Todd M, Marron J. Weighted distance weighted discrimination and its asymptotic properties. *J Am Stat Assoc* 2010, 1050:401–414.
24. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: *Advances in Large Margin Classifiers*, vol. 10. Cambridge, MA: MIT Press; 1999, 61–74.
25. Bartlett PL, Tewari A. Sparseness vs estimating conditional probabilities: some asymptotic results. *J Mach Learn Res* 2007, 8:775–790.
26. Wu Y, Zhang HH, Liu Y. Robust model-free multiclass probability estimation. *J Am Stat Assoc* 2010, 1050:424–436.
27. Zhang C, Liu Y, Wu Z. On the effect and remedies of shrinkage on classification probability estimation. *Am Stat* 2013, 670:134–142.
28. Huhn JC, Hullermeier E. Is an ordinal class structure useful in classifier learning? *Int J Data Mining Model Manage* 2008, 10:45–67.
29. McCullagh P, Nelder JA, McCullagh P. *Generalized Linear Models*, vol. 2. London, UK: Chapman and Hall; 1989.
30. Herbrich R, Graepel T, Obermayer K. Large margin rank boundaries for ordinal regression. In: Bartlett PJ, Schölkopf B, Schuurmans D, Smola AJ, eds. *Advances in Large Margin Classifiers*. Cambridge, MA: MIT Press; 2000, 115–132.
31. Frank E, Hall M. A simple approach to ordinal classification. In: De Raedt L, Flach P, eds. *Machine Learning: ECML 2001*. Lecture Notes in Computer Science, vol. 2167. Berlin and Heidelberg: Springer; 2001, 145–156.
32. Shashua A, Levin A. Ranking with large margin principle: two approaches. In: Becker STS, Obermayer K, eds. *Advances in Neural Information Processing Systems 15*. Cambridge, MA: MIT Press; 2003, 937–944.
33. Chu W, Keerthi S. New approaches to support vector ordinal regression. In: *Proceedings of the 22nd International Conference on Machine learning*, ICML '05, New York, NY, USA, ACM, 7–11 August, 2005, 145–152.

34. Chu W, Ghahramani Z. Gaussian processes for ordinal regression. *J Mach Learn Res* 2005, 6:1019–1041.

35. Cardoso J, da Costa J. Learning to classify ordinal data: the data replication method. *J Mach Learn Res* 2007, 80:1393–1429.

36. Qiao X, Liu Y. Adaptive weighted learning for unbalanced multicategory classification. *Biometrics* 2009, 65:159–168.

37. Qiao X. Noncrossing ordinal classification. arXiv preprint arXiv:1505.03442, 2015.