Computational Statistics
https://doi.org/10.1007/s00180-021-01112-4

ORIGINAL PAPER

A new approach in model selection for ordinal target variables

Elena Ballante[1] · Silvia Figini[2] · Pierpaolo Uberti[3]

Received: 27 April 2020 / Accepted: 5 May 2021

Abstract
Multi-class predictive models are generally evaluated averaging binary classification indicators without a distinction between nominal and ordinal dependent variables. This paper introduces a novel approach to assess performances of predictive models characterized by an ordinal target variable and a new index for model evaluation is proposed. The new index satisfies mathematical properties and it can be applied to the evaluation of parametric and non parametric models. In order to show how our performance indicator works, empirical evidences obtained on toy examples and simulated data are provided. On the basis of the results achieved, we underline that our approach can be a more suitable criterion for model selection than the performance indexes currently suggested in the literature.

Keywords Classification · Ordinal data · Performance index · Model assessment

1 Introduction

Evaluation measures are widely used in predictive models in order to compare different algorithms, thus providing the selection of the best model.

Performance indicators can be used to assess the performance of a model in terms of accuracy, discriminatory power and stability of the results. The choice of indicators to perform model selection is essential and many approaches have been proposed over the years (see e.g. Bradley 1997; Adams and Hand 2000; Hand 2009).

✉ Silvia Figini
silvia.figini@unipv.it

Pierpaolo Uberti
pierpaolo.uberti@unige.it

1 Department of Mathematics, University of Pavia, Pavia, Italy

2 Department of Political and Social Sciences, University of Pavia, Pavia, Italy

3 Department of Economics, University of Genova, Genoa, Italy

Published online: 20 May 2021

Concerning binary target variables, different criteria to compare the performance of classification models are available (see Hand 1997, 2001; Sokolova et al. 2006; Hossin and Sulaiman 2015).

Multi-class classification models are generally evaluated averaging binary classification indicators (see Hand and Till 2001; Sokolova and Lapalme 2009; Hossin and Sulaiman 2015) and in the current literature there is not a clear distinction among them with respect to multi-class nominal and ordinal targets (e.g. Frank and Hall 2001; Pang and Lee 2005; Gaudette and Japkowicz 2009).

Concerning ordinal response variables modelling, different approaches are described in literature, both parametric (see Torra et al. 2006; Kotłowski et al. 2008; Agresti 2010) and non-parametric (see Piccarreta 2004; Galimberti et al. 2012; Ahmad and Brown 2015; Morrone et al. 2019; Hornung 2020), but for the model selection stage the tools are inadequate.

This leads us to propose a new class of measures to select the best model in predictive contexts characterized by a multi-class ordinal target variable, using the misclassification errors coupled with a measure of uncertainty on the prediction.

The paper is structured as follows: Sect. 2 reviews the metrics most used in literature; Sect. 3 shows our methodological proposal and proves mathematical properties; Sect. 4 explains how our proposed index works in two toy examples; Sect. 5 reports the empirical evidence obtained on simulated data. Conclusions and further research ideas are summarized in Sect. 6.

2 Review of the literature for ordinal dependent variables

The most popular measures of performance in ordinal predictive classification models are based on AUC (Area Under the Receiver Operating Characteristic (ROC) Curve), accuracy (expressed in terms of correct classification) and MSE (Mean Square Error), see Gaudette and Japkowicz (2009) and Huang and Ling (2007) among others. The accuracy, measured as percentage of correct predictions over total instances, is the most used evaluation metric for binary and multi-class classification problems (Sokolova et al. 2006), assuming that the costs of the different misclassifications are equal.

The AUC for multi-class classification is defined in Hand and Till (2001) as a generalization of the AUC (based on the probabilistic definition of AUC); it suffers of weaknesses also in the binary classification problem (Gigliarano et al. 2014) and it is cost-independent, assumption that can be viewed as a weakness when the target is ordinal.

The mean square error (MSE) measures the difference between prediction values and observed values in regression problems using an Euclidean distance. MSE can be used in ordinal predictive models, converting the classes of the ordinal target variable y in integers and computing the difference between them; it does not take into account the ordering in a predictive model characterized by ordinal classes in the response variable.

Furthermore, it is well known that in imbalanced data characterized by underfitting or over-fitting the mean square error could provide trivial results (see Hossin and Sulaiman 2015).

3 A new index for model performances evaluation and comparison for ordinal target

Let $\mathbf{y} = \{y_1, \ldots, y_N\}$ be a test set for the ordinal target variable Y, where $y_i \in \{1, \ldots, M\}$ (with M number of classes ordered of the target variable) and let $\mathbb{X}$ be the $N \times p$ data matrix, where N is the number of observations and p the number of covariates.

The output of a predictive model is a matrix $P = \{p_{ij}\}$, where $0 \leq p_{ij} \leq 1$, which contains the probability that observation i belong to the class j estimated by the model under evaluation.

Standard multi-class classification rules assign the observation i to the class $j = \text{argmax}_l\{p_{i,l}\}$.

In order to introduce our proposal, the definitions of classification function and error interval are required.

Definition 1 (*Classification function*) Let observations $\{1, \ldots, N\}$ be grouped by the estimated classes $\hat{y}_i = j$. For each class, sort the observations in a non-increasing order with respect to $p_{i,j}$. The vector of indexes i of the observations is a permutation of the original vector, according to the ordering defined above. For a given model, the classification function is a piecewise constant function $f_{mod} : [0, 1] \to \{1, \ldots, M\}$ such that $f_{mod}([\frac{i-1}{N}, \frac{i}{N})) = y_i$ for $i \in \{1, \ldots, N\}$.

As a special case, the *perfect classification function*, is a piecewise constant function $f_{exact} : [0, 1] \to \{1, \ldots, M\}$ such that each estimated class corresponds to the real class identified by $\mathbf{y}$.

Note that the function f_{exact} is unique except for permutation of the observations in the same estimated class.

The error interval in each class can be derived as the interval between the first misclassified observation and the end of the observations in that estimated class.

Definition 2 (*Error Interval*)

Consider the vector of observations ordered as described in Definition 1. Suppose that the range corresponding to the estimated class j in that vector has indexes in $[n_{j-1}, n_j)$. Let $\tilde{i_j} \in \{n_{j-1}, \ldots, n_j\}$ be the index of the first misclassified observation. The error interval is defined as $[\frac{\tilde{i_j}}{N}, \frac{n_j}{N})$, i.e. the interval between the first misclassified observation and the last observation of the estimated class j; its length is defined as $e_j = \frac{n_j - \tilde{i_j}}{N}$.

If no misclassification occurs in $[n_{j-1}, n_j)$, the error interval is defined as an empty set with a length $e_j = 0$.

Consider an artificial example. Let $N = 10$ be the number of observations and each of these belongs to a class defined by a three levels target variable ($M = 3$). Suppose that a (hypothetical) predictive model returns the predictions as in Table 1.

The classification function is derived grouping the observations in the estimated class as: {3,6,7,8} in Class 1, {2,9,10} in Class 2 and {1,4,5} in Class 3. In each group the observations are sorted with respect to the probability of the estimated class. For the group 1 the probabilities are 0.828, 0.426, 0.849, 0.520 respectively, then the ordered

Table 1 Example

Observation	Probabilities			Estimated class	Real class
	Class 1	Class 2	Class 3		
1	0.288	0.174	**0.538**	3	1
2	0.325	**0.478**	0.197	2	2
3	**0.828**	0.013	0.159	1	1
4	0.310	0.106	**0.584**	3	3
5	0.120	0.262	**0.618**	3	3
6	**0.426**	0.167	0.407	1	3
7	**0.849**	0.126	0.025	1	2
8	**0.520**	0.401	0.079	1	1
9	0.147	**0.670**	0.183	2	2
10	0.142	**0.593**	0.265	2	3

The probabilities are randomly generated, the estimated class is the class with the maximum of probability assigned, the real class are generated starting from the estimated class with some classification errors artificially introduced

Table 2 Index construction

i	7	3	8	6	9	10	2	5	4	1
i	1	2	3	4	5	6	7	8	9	10
y	2	1	1	3	2	3	2	3	3	1
$\hat{y}$	1	1	1	1	2	2	2	3	3	3

group is: {7,3,8,6}. Following the same rule the group 2 becomes {9,10,2} and group 3 is {5,4,1}.

The final sequence of observations can be written as in Table 2. The classification function and the corresponding perfect classification function are depicted in Figs. 1 and 2 respectively.

In order to define the three error intervals, as a preliminary step we identify the intervals of observations related to each estimated class: [0, 0.4) for Class 1, [0.4, 0.7) for Class 2, [0.7, 1) for Class 3. From Table 2, in the estimated Class 1 the first error corresponds to the first observation, so the error interval is [0, 0.4); in the estimated Class 2 the first error corresponds to the observation 6, then the error interval is [0.5, 0.7) and in the estimated Class 3 the first error corresponds to the observation 10 and the error interval is [0.9, 1).

Starting from Definitions 1 and 2, Definition 3 introduces a new index for model performance evaluation in predictive models characterized by an ordinal target variable.

Definition 3 (*Index*) Consider for each class $\{1, \ldots M\}$ the corresponding weight $w_j = \frac{e_j}{l_j}$, where e_j is the jth error interval length and $l_j = n_j - n_{j-1}$ is the length of the jth estimated class in the domain, such that $0 \leq w_j \leq 1$. We define the new index

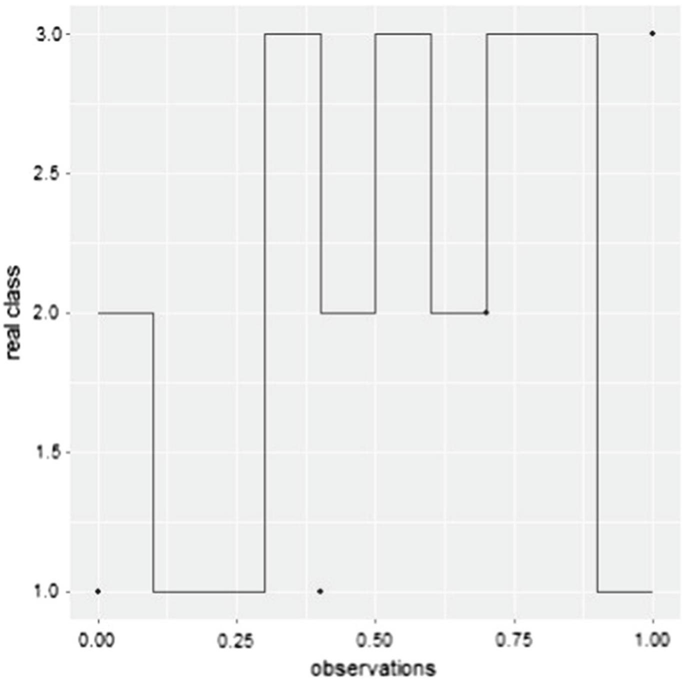

Fig. 1 Classification function

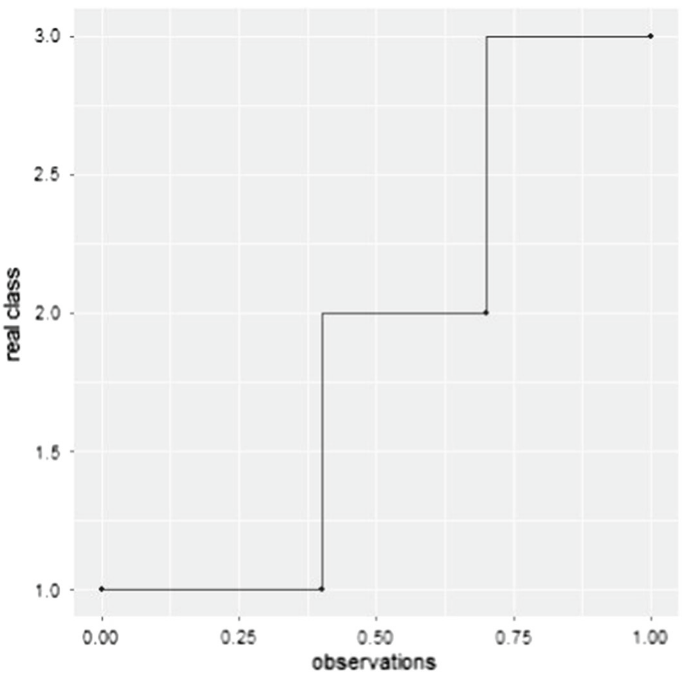

Fig. 2 Perfect classification function

as:

$$I = \sum_{j=1}^{M} w_j \int_{\frac{n_{j-1}}{N}}^{\frac{n_j}{N}} |(f_{mod}(x) - f_{exact}(x))|dx$$

i.e. the new index is defined as the weighted sum of the distance between classification function and perfect classification function.

On the basis of the previous example, we can compute the value for the index introduced in Definition 3: the three integral results are (0.3, 0.1, 0.2) and the corresponding weights are (1, 0.67, 0.33), thus $I = 0.433$.

The index satisfies the following properties.

Property 1 *$I \in [0, +\infty)$. $I = 0$ if and only if $f_{mod} = f_{exact}$.*

Proof

$$I = \sum_{j=0}^{M-1} w_j \int_{\frac{n_{j-1}}{N}}^{\frac{n_j}{N}} |(f_{mod} - f_{exact})(x)|dx \geq \sum_{j=0}^{M-1} \frac{n_j - \tilde{i}_j}{N} |f_{mod} - f_{exact}| \frac{n_j - n_{j-1}}{N}$$

and

- $n_j \geq \tilde{i}_j$,
- $n_j > n_{j-1}$

by definition, than we can conclude that $I \geq 0$.

We prove also that $I = 0$ if and only if $f_{mod} = f_{exact}$.

$I = 0 \implies w_j = 0$ or $\int_{\frac{n_{j-1}}{N}}^{\frac{n_j}{N}} |(f_{mod} - f_{exact})(x)|dx = 0 \quad \forall\, j$ in $\{1, \ldots, M-1\}$.

- $w_j = 0 \iff \tilde{i}_j = n_j$, i.e there are not classification errors, so $f_{mod} = f_{exact}$ in class j.
- $\int_{\frac{n_{j-1}}{N}}^{\frac{n_j}{N}} |(f_{mod} - f_{exact})(x)|dx = 0 \iff f_{mod} = f_{exact}$ in the class j.

We can underline that $I = 0 \implies f_{mod} = f_{exact}$.

The other implication is trivial. □

Property 2 *I has a sharp upper bound $M - 1$*

The upper bound $M - 1$ is reached if and only if $M = 2$ (binary classification).

Proof

$$I = \sum_{j=0}^{M-1} w_j \int_{\frac{n_{j-1}}{N}}^{\frac{n_j}{N}} |(f_{mod} - f_{exact})(x)|dx \leq \sum_{j=0}^{M-1} 1 \cdot \int_{\frac{n_{j-1}}{N}}^{\frac{n_j}{N}} |(f_{mod} - f_{exact})(x)|dx$$

$$\leq \max_x |(f_{mod} - f_{exact})(x)| \sum_{j=0}^{M-1} \frac{n_j - n_{j-1}}{N} \leq M - 1$$

If $M = 2$ we obtain $|(f_{mod} - f_{exact})(x)| = 1\ \forall x \in [0, 1]$ so that $I = M - 1$. If $M > 2$, $|(f_{mod} - f_{exact})(x)| > 1$ for at least one class (by construction) the inequality is strict. □

Proposition 1 *$I \leq K$,*
where K is defined as

$$K = \sum_{i=1}^{M} l_i \max\{M - i, i - 1\}$$

Proof The maximum value is reached when the worst classification is obtained, i.e. when all observations are associated to the farthest class. If this happens, the error interval is as long as the class domain, so $w_j = 1\ \forall j = 1, \ldots, M$ and each integral is the area of a rectangle with basis the class domain l_j and height the maximum height reachable. □

Definition 4 (*Normalized index*)

$$I_n = \frac{1}{K} \sum_{j=0}^{M-1} w_j \int_{\frac{n_{j-1}}{N}}^{\frac{n_j}{N}} |(f_{mod} - f_{exact})(x)| dx$$

where K is the maximum defined in the Proposition 1.
So $0 \leq I_n \leq 1$.

In the previous example, $K = 1.7$ and the corresponding value of the defined normalized index is 0.255.

Proposition 2 *The accuracy is a special case of the index introduced in Definition 3.*

Proof The accuracy is $acc = p_{err} = \frac{\#\{\text{misclassified observations}\}}{N}$ i.e. the proportion of misclassified observations.
Setting $M = 2$, from the Proposition 1, $K = 1$.
$max_x |f_{mod}(x) - f_{exact}(x)| = 1$, each weight is $w_j = \frac{1}{N}$ if $w_1 = w_2 = 1$ and $I_n = p_{err}$. □

Property 3 (Monotonicity) *Consider a classification C with ϵ misclassifications and N observations. Operating a transformation of the classification C in C' where an observation right classified is changed in a misclassification, the index I_n becomes higher.*

Proof In the classification C', $\epsilon'=\epsilon + 1$ are misclassified observations: the ϵ observations misclassified in C plus a new misclassification. Suppose that the new misclassification is the observation i that is classified in the class j' instead of the real class j.

Table 3 Confusion matrix Model 1

Predict	Actual		
	1	2	3
1	5	0	1
2	0	7	0
3	0	0	7

All the components in the sum of the index I_n remain unchanged except for the jth, thus obtaining I_n^j. So

$$I_n^j = w_j \int_{\frac{n_{j-1}}{N}}^{\frac{n_j}{N}} |f_{mod}(x) - f_{exact}(x)|dx$$

Looking at each of the two elements in the product:

- $w'_j \geq w_j$ Two different cases are possible: if the probability associated to the ith observations is less or equal than the probability of the first error, the error interval $w'_j = w_j$; on the other hand, the error interval become larger, thus $w'_j > w_j$.
- $|f'_{mod} - f_{exact}| > |f_{mod} - f_{exact}|$ In C' there is one misclassification more than in C, so the distance between f_{mod} and f_{exact} increases.

We can conclude that $I_n^{'j} \geq I_n^j$. □

We remark that in the Property 3 the vice versa does not hold, i.e. if $I_{mod1} \geq I_{mod2}$ we can not make conclusions on the number of misclassified observations in the two classifications.

4 Toy examples

In order to show how our index works with respect to the indexes proposed in the literature, toy examples are reported in this section with the main aim of discussing the behavior in terms of model selection of our index with respect to AUC, accuracy and MSE.

Y is a target variable characterized by $M = 3$ levels $y_i \in \{1, 2, 3\}$ and Model 1 and Model 2 are two competitive models under comparison. The numerical setting of both examples is stated in "Appendix".

4.1 First toy example

In the first toy example we take into account the ordinal structure of the target variable Y. Tables 3 and 4 are the corresponding confusion matrices for Model 1 and Model 2. It is clear that the Model 2 makes a better classification than Model 1.

For the sake of comparison, for each model the AUC, the accuracy, the MSE and our index are computed as summarized in Table 5.

Table 4 Confusion matrix Model 2

Predict	Actual		
	1	2	3
1	5	1	0
2	0	6	0
3	0	0	8

Table 5 Results

Model	Proposed index	Normalized index	AUC	Accuracy	MSE
1	0.08	0.05	0.95	0.95	0.20
2	0.04	0.03	0.95	0.95	0.05

Table 6 Confusion matrix

Predict	Actual		
	1	2	3
1	5	0	1
2	0	7	0
3	0	0	7

We remark that looking at Table 5 the values obtained for the AUC and the accuracy indexes for Model 1 and Model 2 are exactly equal, thus, in terms of model choice, Model 1 and Model 2 are not different. Our index highlights a difference in terms of performance between the two models under comparison and it selects Model 2 as the best one. Further details about the settings are given in Table 11 in "Appendix".

4.2 Second toy example

The second toy example considers the probability assigned to each observation. In practical applications where we need also to evaluate how much uncertainty is associated to a prediction, the starting point considers the probability that the new observation belongs to the estimated class.

From Table 6, both Model 1 and Model 2 assign an observation of the third class to the first one. The first classification assigns a higher probability to the misclassified observation than the second ($p = 0.866$ vs $p = 0.4004$), see Table 12 in "Appendix". Table 12 reports set probabilities and consequent assigned classes. Then we can conclude that Model 2 is better than Model 1 for data at hands.

From Table 7 both models are equivalent in terms of MSE and accuracy, thus on the basis of classical measures Model 1 and Model 2 are not different. Our index reports different values for the models under comparison and select Model 2 as the best one.

5 Empirical evaluation on simulated data

In order to show how our proposal works in model selection, this section reports the empirical results achieved on a simulated database.

Table 7 Results

Model	Proposed index	Normalized index	AUC	Accuracy	MSE
1	0.083	0.051	0.956	0.950	0.200
2	0.017	0.010	0.983	0.950	0.200

Table 8 Simulated data structure

y	1	2	3	4	5
x1	N(2,1.5)	N(3,1)	N(4,1.5)	N(5,1)	N(6,1)
x2	N(1,2.5)	N(5,2)	N(7,2.5)	N(8.5,2)	N(9.5,2)
x3			U(0,3)		

The simulated database is composed of three covariates obtained by a Monte Carlo simulation and an ordinal target variable with $M = 5$, as reported in Table 8. The sample size is $N = 7500$. The database is exactly balanced in terms of response variable: 1500 observations are generated for each level of y.

Five different models are under comparison:

- Ordinal logistic regression (Ord Log),
- Conditional inference tree (Tree),
- Support vector machine (SVM),
- Ordinal Random forest (RFor),
- k- Nearest Neighbour with k=20 (kNN-20),
- k- Nearest Neighbour with k=50 (kNN-5),
- Naive Bayes (NaiveB),
- Classification tree for ordinal response (OrdTree).

For each model AUC, accuracy, MSE and our index are computed using a 10-fold cross validation. More specifically, the database is randomly partitioned into 10 equal sized sub-samples (of 750 observations), each one retained as validation data and the remaining 9 sub-samples are used as training data. The process is then repeated 10 times, with each of the sub-samples used exactly once for validation. The resulting metrics are averaged and than reported in Table 9.

For the sake of clarity, Table 10 shows the resulting ranks for the models, using the results obtained for the four metrics under comparison.

We can see that the k-nearest neighbor with $k = 5$ is classified as the best model according to all the indexes employed for model choice except for the AUC metric, but the values of AUC are extremely similar to the best model (the difference is less than 0.001). Furthermore, from Table 9 k-nearest neighbor outperforms the other models (with both choices of k). The Naive Bayes is ranked as the second-best model after kNN with respect to all performance indicators except for MSE (with minimum differences from Ord Log and SVM).

The classification tree for ordinal responses (OrdTree) as presented in Galimberti et al. (2012) show lower performances of the other methods, but performs better than the standard classification tree in terms of MSE and the proposed index.

Table 9 Model comparison

Model	Proposed index	Normalized index	AUC	Accuracy	MSE
Ord log	0.450	0.141	0.864	0.581	0.580
Tree	1.569	0.491	0.875	0.586	0.643
SVM	0.446	0.137	0.869	0.592	0.581
RFor	0.469	0.143	0.875	0.589	0.643
kNN-20	0.003	0.0009	0.999	0.976	0.025
kNN-5	0.002	0.0006	0.999	0.993	0.008
NaiveB	0.434	0.132	0.877	0.604	0.594
OrdTree	0.494	0.150	0.818	0.580	0.635

Table 10 Results in terms of ranking

Model	Proposed index/normalized	AUC	Accuracy	MSE
Ord log	5	7	7	3
Tree	7	4	6	8
SVM	4	6	4	4
RFor	6	5	5	7
kNN-20	2	1	2	2
kNN-5	1	2	1	1
NaiveB	3	3	3	5
OrdTree	7	8	8	6

When the performance differences between models are macroscopic all the indexes agree in model selection. The interest of a new metric comes out when other indexes can not individuate differences between performances, then the natural structure of data and prediction probabilities become fundamental for the selection of the best model.

6 Conclusions

A new performance indicator is proposed to compare predictive classification models characterized by ordinal target variable.

Our index is based on the definition of a classification function and an error interval. A normalized version of the index is derived. The empirical evidence at hands underlined that our index discriminates better among different models with respect to classical measures available in literature.

Our index can be used coupled with other metrics for assessing model performances for model selection.

From a computational point of view a further idea of research will consider the implementation of our index in a new R package. In terms of application we think that our index could be directly incorporate in the process of assessment for predictive analytics.

Acknowledgements This paper has been supported by IRCCS Mondino Foundation. The paper has been written by Elena Ballante under the supervision of Prof. Figini and Prof. Uberti.

Funding Open access funding provided by Universit[Pleaseinsertintopreamble] degli Studi di Pavia within the CRUI-CARE Agreement.

Declaration

Conflict of interest The authors declare that they have no conflict of interest.

A Toy example settings

In order to clarify the toy examples, numerical settings are reported. Tables 11 and 12 contain the hypothetical output of the two models described in Sect. 4: a progressive ID of observations, probabilities assigned for each class (p_1, p_2, p_3) by Model 1 and Model 2, the resulting estimated class for each model and the real class assigned arbitrary by the author.

Table 11 First toy example

Observation	Model 1			Model 2			Estimated class Model 1	Estimated class Model 2	Real class
	p_1	p_2	p_3	p_1	p_2	p_3			
1	0.114	0.473	0.413	0.114	0.473	0.413	2	2	2
2	0.068	0.184	0.747	0.068	0.184	0.747	3	3	3
3	**0.750**	**0.125**	**0.125**	**0.125**	**0.750**	**0.125**	**1**	**2**	**3**
4	0.587	0.212	0.201	0.587	0.212	0.201	1	1	1
5	0.0583	0.623	0.319	0.0583	0.623	0.319	2	2	2
6	0.371	0.063	0.565	0.371	0.063	0.565	3	3	3
7	0.329	0.179	0.491	0.329	0.179	0.491	3	3	3
8	0.114	0.444	0.442	0.114	0.444	0.442	2	2	2
9	0.936	0.014	0.050	0.936	0.014	0.050	1	1	1
10	0.116	0.229	0.655	0.116	0.229	0.655	3	3	3
11	0.376	0.398	0.226	0.376	0.398	0.226	2	2	2
12	0.435	0.438	0.128	0.435	0.438	0.128	2	2	2
13	0.452	0.226	0.321	0.452	0.226	0.321	1	1	1
14	0.740	0.173	0.087	0.740	0.173	0.087	1	1	1
15	0.180	0.796	0.0243	0.180	0.796	0.0243	2	2	2

Table 11 continued

Observation	Model 1			Model 2			Estimated class Model 1	Estimated class Model 2	Real class
	p_1	p_2	p_3	p_1	p_2	p_3			
16	0.343	0.392	0.265	0.343	0.392	0.265	2	2	2
17	0.049	0.073	0.878	0.049	0.073	0.878	3	3	3
18	0.522	0.076	0.403	0.522	0.076	0.403	1	1	1
19	0.012	0.194	0.794	0.012	0.194	0.794	3	3	3
20	0.128	0.380	0.491	0.128	0.380	0.491	3	3	3

Table 12 Second toy example

Observation	Model 1			Model 2			Estimated class	Real class
	p_1	p_2	p_3	p_1	p_2	p_3		
1	0.114	0.473	0.413	0.114	0.473	0.413	2	2
2	0.068	0.184	0.747	0.068	0.184	0.747	3	3
3	**0.866**	**0.012**	**0.121**	**0.400**	**0.300**	**0.300**	**1**	**3**
4	0.587	0.212	0.201	0.587	0.212	0.201	1	1
5	0.0583	0.623	0.319	0.0583	0.623	0.319	2	2
6	0.371	0.063	0.565	0.371	0.063	0.565	3	3
7	0.329	0.179	0.491	0.329	0.179	0.491	3	3
8	0.114	0.444	0.442	0.114	0.444	0.442	2	2
9	0.936	0.014	0.050	0.936	0.014	0.050	1	1
10	0.116	0.229	0.655	0.116	0.229	0.655	3	3
11	0.376	0.398	0.226	0.376	0.398	0.226	2	2
12	0.435	0.438	0.128	0.435	0.438	0.128	2	2
13	0.452	0.226	0.321	0.452	0.226	0.321	1	1
14	0.740	0.173	0.087	0.740	0.173	0.087	1	1
15	0.180	0.796	0.0243	0.180	0.796	0.0243	2	2
16	0.343	0.392	0.265	0.343	0.392	0.265	2	2
17	0.049	0.073	0.878	0.049	0.073	0.878	3	3
18	0.522	0.076	0.403	0.522	0.076	0.403	1	1
19	0.012	0.194	0.794	0.012	0.194	0.794	3	3
20	0.128	0.380	0.491	0.128	0.380	0.491	3	3

References

Adams NM, Hand DJ (2000) Improving the practice of classifier performance assessment. Neural Comput 12:305–311. https://doi.org/10.1162/089976600300015808

Agresti A (2010) Analysis of ordinal categorical data, vol 656. Wiley, Hoboken

Ahmad A, Brown G (2015) Random ordinality ensembles: ensambles methods for multi-valued categorical data. Inf Sci 296:75–94. https://doi.org/10.1016/j.ins.2014.10.064

Bradley AP (1997) The use of the area under the ROC curve in evaluation of machine learning algorithms. Pattern Recognit 30:1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Frank E, Hall M (2001) A simple approach to ordinal classification. Eur Conf Mach Learn 2167:145–156. https://doi.org/10.1007/3-540-44795-4_13

Galimberti G, Soffritti G, Di Maso M (2012) Classification trees for ordinal responses in R: the rpartscore package. J Stat Softw 47:1–25

Gaudette L, Japkowicz N (2009) Evaluation methods for ordinal classification. In: Gao Y, Japkowicz N (ed) Advances in artificial intelligence, pp 207–210. https://doi.org/10.1007/978-3-642-01818-3_25

Gigliarano C, Figini S, Muliere P (2014) Making classifier performance comparisons when ROC curves intersect. Comput Stat Data Anal 77:300–312. https://doi.org/10.1016/j.csda.2014.03.008

Hand DJ (1997) Construction and assessment of classification rules. Wiley series in probability and statistics. Wiley, Hoboken

Hand DJ (2001) Measuring diagnostic accuracy of statistical prediction rules. Stat Neerl 55:3–16. https://doi.org/10.1111/1467-9574.00153

Hand DJ (2009) Measuring classifier performance: a coherent alternative to the area under the ROC curve. Mach Learn 77(1):103–123. https://doi.org/10.1007/s10994-009-5119-5

Hand DJ, Till RJ (2001) A simple generalisation of the area under the roc curve for multiple class classification problems. Mach Learn 45(2):171–186. https://doi.org/10.1023/A:1010920819831

Hornung R (2020) Ordinal forests. J Classif 37:4–17. https://doi.org/10.1007/s00357-018-9302-x

Hossin M, Sulaiman MN (2015) A review on evaluation metrics for data classification evaluations. Int J Data Min Knowl Manag Process 5(2):171–186. https://doi.org/10.5121/ijdkp.2015.5201

Huang J, Ling CX (2007) Constructing new and better evaluation measures for machine learning. In: IJCAI international joint conference on artificial intelligence (IJCAI'07), pp 859–864. https://doi.org/10.5555/1625275.1625414

Kotłowski W, Dembczyński K, Greco S, Słowiński R (2008) Stochastic dominance-based rough set model for ordinal classification. Inf Sci 178(21):4019–4037. https://doi.org/10.1016/j.ins.2008.06.013

Morrone A, Piscitelli A, D'Ambrosio A (2019) How disadvantages shape life satisfaction: an alternative methodological approach. Social Indic Res Int Interdiscip J Quality-of-Life Meas 141(1):477–502. https://doi.org/10.1007/s11205-017-1825-8

Pang B, Lee L (2005) Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05), pp 115–124. https://doi.org/10.3115/1219840.1219855

Piccarreta R (2004) Ordinal classification trees based on impurity measures. In: Bock HH, Chiodi M, Mineo A (eds) Advances in multivariate data analysis. Studies in classification, data analysis, and knowledge organization, 2004. https://doi.org/10.1007/978-3-642-17111-6_4

Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Inf Process Manag 45(4):427–437

Sokolova M, Japkowicz N, Szpakowicz S (2006). Beyond accuracy, f-score and ROC: a family of discriminant measures for performance evaluation. In: Advances in artificial intelligence. Lecture notes in computer science (AI 2006), vol 4304, pp 1015–1021

Torra V, Domingo-Ferrer J, Mateo-Sanz JM, Ng M (2006) Regression for ordinal variables without underlying continuous variables. Inf Sci 176(4):465–474. https://doi.org/10.1016/j.ins.2005.07.007