How to adapt the performance metrics to ordinal classification by intervals?

Giulia Binotto and Rosario Delgado

Department of Mathematics, Universitat Autònoma de Barcelona (Spain)

Abstract. In the field of Supervised Machine Learning, accurate evaluation of classification models is crucial for assessing their performance and guiding model selection. In this article, we delve into the domain of ordinal classification. Traditional metrics fail to capture the inherent ordinal structure of data, and new performance metrics have to be considered. First, we present an existing metric for ordinal classification based on a cost-sensitive approach and simplify its calculation. We extend the results to classification by intervals, where the length of the intervals, not only their order, assumes significance. We offer a comprehensive evaluation framework for this scenario. Additionally, the article addresses the challenge of classification with unbounded rightmost intervals, which further enhances the applicability of the proposed metrics.

Keywords: Ordinal classification, classification by intervals, performance metrics, unbounded intervals

1 Introduction

In the context of supervised machine learning (ML), it is not uncommon to face a classification problem where the target variable is obtained by discretization of a continuous variable into intervals, or the binning of a discrete variable. For example, consider the case where we pretend to predict the length of stay of a patient in the ICU of a hospital, assigning one of the following categories: `short` (1-2 days), `moderate` (3-7 days), `long` (8-14 days) or `extremely long` (more than 14 days). Or the goal is to predict the number of relapses for a patient with recurrent disease grouped into: `none`, `one`, `two`, or `more than two`. Naturally, discretization or binning always entails an undesirable loss of information, although since we use a classifier as a predictive model, this drawback is compensated by avoiding assumptions about other types of predictive models such as regression, for which it is essential to diagnose model assumptions.

At this point, the dilemma of which performance metric to employ for measuring classification quality emerges, given that it is an ordinal classification task with the additional particularity that the classes are intervals. On the one hand, incorporating information about the length of the intervals in the definition of the metric violates the basic concept of ordinal classification, according to which the intervals that determine the classes have not meaning in themselves [10]. On the other hand, this is precisely what fascinates us since it is what makes the problem new and intriguing at the same time. As a result, the context in which we move sets a new paradigm in ordinal classification in which the lengths of the intervals that define the classes not only can, but must, play a determinant role in the building of the behavior metric. To the best of our knowledge, no

Prepublicació Núm 6, juny 2023.
Departament de Matemàtiques.
http://www.uab.cat/matematiques

attempt has been undertaken in the literature to carry out such an approximation. It is to fill this gap that in this work we delve into the construction of a metric to evaluate the performance of supervised ML methods when the classification is ordinal by intervals. We will construct the new metric from a metric for standard ordinal classification (without intervals).

Ordinal classification

(Standard) Ordinal classification is a multiclass classification task where instances are classified into groups that have an inherent natural ordering. Without loss of generality, we can assume that the classes $1, \ldots, r$ are in this order, and this implies that we assume as more serious the error of classifying an instance of class i as belonging to class $i+2$ or $i-2$ than as belonging to class $i+1$ or $i-1$, respectively.
We denote by $C = (C_{ij})_{i,j=1,\ldots,r}$ a general confusion matrix obtained from any *validation* procedure by a supervised ML algorithm of classification, where C_{ij} is the number of instances in the test dataset that belong to class j and have been assigned to class i by the classifier. Also denote by $N = \sum_{i=1}^{r} \sum_{j=1}^{r} C_{ij}$ the total number of instances in the test dataset, and by $n_j = \sum_{i=1}^{r} C_{ij}$ the number of instances in the test dataset belonging to class j, for $j = 1, \ldots, r$, with $N = \sum_{j=1}^{r} n_j$. We assume that $n_j > 0$ for any $j = 1, \ldots, r$ (otherwise, the class would be removed and we would be left with the remaining $r-1$ classes). Also denote by $y_1, \ldots, y_N$ and $\hat{y}_1, \ldots, \hat{y}_N$, respectively, the observed and predicted classes of any of the N instances in the test dataset (then, $y_k, \hat{y}_k \in \{1, \ldots, r\}$ for all $k = 1, \ldots, N$).

Despite its practical applications, as for example for rating product reviews in sentimental analysis and opinion mining [9], ordinal classification has been less developed to date than multiclass classification in general. Nevertheless, there are different performance metrics (evaluation measures) known in the literature that allow comparing classifiers when the class variable is ordinal, and we will recall some of them in Section 2.

State-of-the-art

The authors of [2] address the problem of imbalance, when certain classes are considerably more common than others, in the context of performance measures for ordinal classification (also known as *ordinal regression*). In this case, using a metric designed for balanced datasets may result in a situation in which a classifier that assigns always the majority class outperforms highly sophisticated classification systems. To overcome this problem, they introduce macro-averaged versions of the most common ordinal classification measures, which are more resilient to imbalance and equivalent to the standard versions when the datasets are balanced.
The same problem has been considered by [12], proposing a solution based on weighted agreement measures, such as Cohen's κ, Scott's π, Gwet and Brennan-Prediger, where the weighting schemes considered are linear, quadratic, ordinal, radical and bipolar

weights, concluding from the experimental phase with real datasets that Cohen's κ and Scott's π with quadratic weights perform better than the other considered metrics.
The lack of adaptability to imbalance is not the only issue that the standard metrics for ordinal classification exhibit. In fact, as evidenced by [4], where an alternative measure that prevents this flaws is proposed, some of them, such as MAE and its derivatives, have the disadvantage of being dependent on the numbers chosen to represent the classes, whereas Kendall's τ_b outperforms this problem at the price of losing information about the absolute predictions. As a result, this metric is better suited for assessing preference learning than ordinal classification.
The authors of the papers [5, 6] introduce two new measures for ordinal classification: the maximum and the minimum of mean absolute error of all the classes, which take into account the per-class distribution of patterns as well as the magnitude of the error, and they propose using the first of them, jointly with the mean absolute error, as a pair of metrics to drive a multi-objective evolutionary algorithm, since they are competitive objectives.

Scope of the work and organization of the paper

We take the following course of action: we select a metric for ordinal classification that follows a cost-sensitive approach and is introduced in [8]. Indeed, the authors of [8] introduced the Total Misclassification Cost (TC), and in order to bound it to [0, 1], they divide it by its maximum value, which is determined by an optimization problem. To begin, in Section 2 we sketch the usual metrics for standard ordinal classification and recall the concept of the Total misclassification Cost (TC) from [8]. In Section 3 we find the maximum of TC analytically, and thereby define the Standardized Total misclassification Cost (STC) as TC divided by its maximum, which is as a modified version of the metric in [8]. Secondly, in Section 4 we adapt STC metric to the scenario when the classes are intervals, which is the ultimate goal of this work, and investigate some of its properties. Unlike standard ordinal classification, where only the order counts, each class is allocated a length, and the length of the classes play a role in this layout.
Another issue is that the rightmost interval is often unbounded, meaning it cannot be assigned a preset length. This poses a serious challenge in the task of developing the metric, as it will unavoidably be dependent on the lengths of the intervals. In Section 5 we provide a reasoned solution to this question, together with a toy example illustrating its application and a general result at the theoretical level. Finally, Section 6 is devoted to a general discussion and some closing words.

2 Metrics for standard ordinal classification

Evaluation metrics differ in how they handle the classification mistakes. In this chapter, we recall the mainstream evaluation measures usually considered for ordinal classification, which are based on the confusion matrix. This includes a cost-sensitive measure that will be used to define the STC metric in Section 3.

Mainstream metrics based on the confusion matrix

- Error Rate: Simply the fraction of incorrect predictions (that is, $1-$*Accuracy*). It has the disadvantage that all errors are treated equally and therefore does not penalize classifiers who make flagrant errors. However, we will also consider this metric as a reference. The formula is:

$$\text{Error Rate} = 1 - \frac{\sum_{i=1}^{r} C_{ii}}{N} = 1 - \frac{\sum_{\ell=1}^{N} 1_{\{y_\ell = \hat{y}_\ell\}}}{N}$$

  and ranges between 0 and 1, the first corresponding to perfect classification.

- *Mean Absolute Error* (MAE): The literature in several studies concludes that MAE is one of the best performance metrics in ordered classification. For example, in [7] the authors experimentally show that for the unbalanced dataset studied, MSE (Mean Squared Error) and MAE perform the best, but while MSE is better in situations where the severity of the errors is more important, MAE shows to be better in situations where the tolerance for small errors is lower. This is despite the fact that neither of these measures is truly ordinal by design. In [3] the authors use MAE as performance metric for monotonic ordinal classification, in order to show the usefulness of selecting the training set to obtain more accurate and efficient models.
  This metric penalizes errors (wrongly classifying an item in a category that is far from the correct one) proportionally to the distance between the categories, so the lower the metric value, the better the performance of the classifier. Its definition is the following:

$$\text{MAE} = \frac{1}{N} \sum_{i,j=1}^{r} C_{ij} \, |i-j| = \frac{1}{N} \sum_{\ell=1}^{N} |y_\ell - \hat{y}_\ell| \tag{1}$$

- *Measures related to* MAE*:* There are some measures for ordinal classification that are variations of MAE, such as:
  - *Weighted Average of Mean Absolute Error (*AMAE*):* If the class sizes are unbalanced (typical in most situations in healthcare applications, for example), computing a weighted average of MAE across all classes is more robust than MAE itself. Its definition is:

$$\text{AMAE} = \frac{1}{r} \sum_{j=1}^{r} \text{MAE}_j \,, \quad \text{where}$$

$$\text{MAE}_j = \frac{1}{n_j} \sum_{i=1}^{r} C_{ij} \, |i-j| = \frac{1}{n_j} \sum_{\ell=1}^{N} 1_{\{y_\ell = j\}} \, |y_\ell - \hat{y}_\ell| \,. \tag{2}$$

  (note that $\text{MAE} = \frac{1}{N} \sum_{j=1}^{r} n_j \, \text{MAE}_j$).

- *Maximum of Mean Absolute Error* (MMAE) ([5]): Also useful if class sizes are unbalanced, the maximum of MAE across the classes is defined by:

$$\mathrm{MMAE} = \max_{j=1,\ldots,r} \mathrm{MAE}_j$$

(Analogously, the Minimum of Mean Absolute Error mMAE can be defined by $\mathrm{mMAE} = \min_{j=1,\ldots,r} \mathrm{MAE}_j$.)

The problem with MAE and its variants is that they assume that all classes are equidistant, which does not have to be true when performing an ordinal classification task. For example, classification on a scale

`very bad`, `bad`, `acceptable`, `good` or `very good`

is consequence of a subjective appreciation that will hardly correspond to equidistant numerical values. Works using these metrics, explicitly or implicitly assume that "misclassification costs are always proportional to the absolute difference between the actual and the predicted label" ([11], *expressis verbis*). But this assumption goes against the basic principle of meaninglessness of the numerical values in ordinal classification, beyond their ordering ([1]).

Association metrics

Other metrics for standard ordinal classification are *association metrics* that are based in the agreement between two raters who classify the instances into ordered categories. One of the most used is *Kendall's correlation coefficient*: to avoid the influence of arbitrary class labels, we can use a metric which is independent of the range that each class represents, and simply assesses the order relation between observed and predicted class labels. This correlation coefficient is a measure of the ordinal association (relationship between rankings) between observed and predicted classes. Its definition is:

$$\tau_b = \frac{\sum_{\ell,m=1}^{N} \gamma_{\ell m}\, \widehat{\gamma}_{\ell m}}{\sqrt{\Big(\sum_{\ell,m=1}^{N} \gamma_{\ell m}^2\Big)\Big(\sum_{\ell,m=1}^{N} \widehat{\gamma}_{\ell m}^2\Big)}}$$

with

$$\gamma_{\ell m} = \begin{cases} +1 & \text{if } y_\ell > y_m \\ 0 & \text{if } y_\ell = y_m \\ -1 & \text{if } y_\ell < y_m \end{cases} \quad \text{and} \quad \widehat{\gamma}_{\ell m} = \begin{cases} +1 & \text{if } \hat{y}_\ell > \hat{y}_m \\ 0 & \text{if } \hat{y}_\ell = \hat{y}_m \\ -1 & \text{if } \hat{y}_\ell < \hat{y}_m \end{cases}$$

Alternatively, Kendall's correlation coefficient can be written as:

$$\tau_b = \frac{Con - Dis}{\sqrt{\big(Con + Dis - Tobs\big)\big(Con + Dis - Tpred\big)}},$$

where Con is the number of concordant pairs, that is, pairs $(\ell, m) \in N \times N$ such that $\gamma_{\ell m}\,\widehat{\gamma}_{\ell m} = +1$; Dis is the number of discordant pairs, that is, pairs such that $\gamma_{\ell m}\,\widehat{\gamma}_{\ell m} = -1$; $Tobs$ is the number of tied pairs in the observed class membership, that is, such that $\gamma_{\ell m} = 0$; and $Tpred$ the number of tied pairs in the predicted class membership, that is, $\widehat{\gamma}_{\ell m} = 0$.
$\tau_b \in [-1, +1]$ and the interpretation is as follows: the higher the Kendall's correlation coefficient metric value, the better the classifier performance, with the maximum $\tau_b = +1$ corresponding to no misclassification errors at all, and $\tau_b = -1$ to a negative association or perfect inversion between observed and predicted classes.

A metric that follows a cost-sensitive approach

In [8] a new cost-sensitive metric is introduced, named **Total misclassification Cost (TC)**. The proposed measure accounts for inherent ordinal data structure, the total misclassification cost of a classifier, and the unbalanced class distribution, and shows good performance in identifying the best ordinal classifier with some real datasets and simulation studies. Its definition is:

$$\mathrm{TC} = \frac{1}{N}\sum_{j=1}^{r} \mathrm{TC}_j\,, \quad \text{with } \mathrm{TC}_j = \sum_{i=1}^{r} C_{ij}\,\gamma_{ij}\,|i-j|\,, \tag{3}$$

where

$$\gamma_{ij} = \frac{\sum_{k\neq j}^{r} n_k}{n_i} = \frac{N - n_j}{n_i} \quad (\text{which is } \geq 1 \text{ if } i \neq j).$$

(Compare with the definition of MAE (1), in which γ_{ij} is just 1, independently of i and j.)
The rationale behind this definition is as follows: this measure uses information from the class distribution and domain knowledge about the ordinal class structure, and is the sum of the misclassification costs for any class $j = 1, \ldots, r$. Indeed, for instances that are actually of class j, the misclassification cost takes into account not only the distance between predicted class i and the true class labels, $|i - j|$ (the higher, the higher cost), but the size of the classes, in the sense that the smaller the size of the class i, or of the class j, the higher the cost. That is, this measure penalizes more cases of misclassification when assigning a small class than of a large one, and penalizes more misclassification when assigning from a small class than from a large one. This is captured in (3) by γ_{ij}, which is defined as the inverse of the probability of misclassifying an object in class i given that the object is of class j and has been misclassified, if the classification is done randomly (i.e: a label with a class has been assigned, with the label chosen randomly from the available labels). This means that, if $i \neq j$, then $\gamma_{ij} = 1/p_{ij}$ with

$$\begin{aligned} p_{ij} &= P(\text{an object of class } j \text{ is classified in class } i/ \\ &\qquad \text{the object of class } j \text{ is misclassified}) \\ &= \frac{P(\text{an object of class } j \text{ is classified in class } i)}{P(\text{the object of class } j \text{ is misclassified})} = \frac{n_i/N}{(N-n_j)/N} = \frac{n_i}{N-n_j}. \end{aligned}$$

Remark 1. We can introduce the **cost matrix** $M = (m_{ij})_{i,j=1,\dots,r}$ by

$$m_{ij} = \frac{\gamma_{ij}\,|i-j|}{N} \tag{4}$$

that is, m_{ij} is the cost associated to misclassify an instance belonging to class j in class i. When is the cost matrix M symmetric? M is symmetric if for all $i \neq j$, $\gamma_{ij} = \gamma_{ji}$, which is equivalent to say that

$$\sum_{k\neq i,j} \frac{n_k}{n_i} = \sum_{k\neq i,j} \frac{n_k}{n_j} \Leftrightarrow n_i = n_j,$$

that is, matrix M is symmetric if and only if there is a perfect balance between classes in the sense that $n_i = n_j$ for all $i, j = 1, \dots, r$ (that is, $n_i = N/r$ for all $i = 1, \dots, r$). With this notation,

$$\mathrm{TC} = \sum_{i,j=1}^{r} C_{ij} m_{ij} = sum(C \odot M)$$

where $\odot$ denotes the element-wise (Hadamard) or Schur matrix product.

Remark 2. Note that in [8], the definition of TC has not N dividing. We added it in our definition by analogy with the metric MAE.

3 The STC metric

We assume that r, the number of classes, and $n_1, \dots, n_r$, the number of cases in the test set that belong to any of the classes, are fixed. Then, we introduce a new standardized total cost based on TC in the interval [0, 1]. Instead of solving an optimization problem as in [8], we consider the (finite) set of matrices:

$$\begin{aligned}\mathcal{S} = \{&A = (a_{ij})_{i,j=1,\dots,r} \text{ matrix } r \times r \text{ such that } a_{ij} \in \mathbb{N} \text{ (including 0) and} \\ &\sum_{i=1}^{r} a_{ij} = n_j\,,\ j = 1, \dots, r\}\end{aligned}$$

(when needed, we specify in the notation the dependence on r and $n_1, \dots, n_r$) with the equivalence relation $\sim$ defined by:

$$A,\, B \in \mathcal{S} \text{ belong to the same class } (A \sim B) \quad \text{if} \quad TC(A) = TC(B)$$

(where TC is defined by (3) and we use the notation $\mathrm{TC}(C)$ to make explicit the dependence on matrix C when needed). The set of the equivalence classes of $\mathcal{S}$ with this relation is the quotient set denoted by $\mathcal{S}/\sim$, where we can define a total ordering $\preceq$ by: for $[A]$, $[B] \in \mathcal{S}/\sim$,

$$[A] \preceq [B] \Longleftrightarrow \mathrm{TC}(A) \leq \mathrm{TC}(B)$$

(where we use $[A]$ to denote the equivalence class of a generic matrix $A \in \mathcal{S}$, that is, $[A] = \{M \in \mathcal{S} : M \sim A\}$). Then, since $\mathcal{S}/\sim$ is finite (because $\mathcal{S}$ is finite; indeed, for any column of a matrix $A \in \mathcal{S}$, say column j, there is a finite number of ways to accommodate n_j among the r positions in the column), and being a totally ordered set with $\preceq$, there exists a (unique) maximum of $\left(\mathcal{S}/\sim, \preceq\right)$, say class $\mathcal{S}_{max}$. Let us define

$$\mathrm{TC}_{max} = \mathrm{TC}(A_{max}) \tag{5}$$

being A_{max} any matrix representing the class $\mathcal{S}_{max}$, that is, such that $[A_{max}] = \mathcal{S}_{max}$. By definition, the value $\mathrm{TC}(A_{max})$ is uniquely defined and therefore, TC_{max} is well defined and only depends on quantities $n_1, \ldots, n_r$, and not on the confusion matrix itself, and saved us having to solve an optimization problem. Then, for any confusion matrix C,

$$\mathrm{TC}(C) \leq \mathrm{TC}_{max}$$

Definition 1. *If $C = (C_{ij})_{i,j=1,\ldots,r}$ is the confusion matrix associated at a given classifier, we define the corresponding* **Standardized Total misclassification Cost** *metric as a performance measure for ordinal classification, by*

$$\mathrm{STC} = \frac{\mathrm{TC}}{\mathrm{TC}_{max}} \in [0,\, 1]\,,$$

where TC_{max} is defined by (5). Then, STC is the proportion of TC_{max} that represents TC.

Let us see that we can obtain an explicit expression for TC_{max} in general, which can be specified in particular cases. In what follows, we will use the following notation: for any $j = 1, \ldots, r$ fixed,

$$i_j = \arg \max_{\ell=1,\ldots,r} \frac{|\ell - j|}{n_\ell} \tag{6}$$

that is, fixed j, i_j is the class that maximizes the cost associated to misclassify in that class an instance that belongs to class j.

Proposition 1. *For any $r \geq 1$, fixed the number of classes and their sizes,*

$$\mathrm{TC}_{max} = \frac{1}{N} \sum_{j=1}^{r} K_j\,, \quad \text{with } K_j = \frac{n_j\,(N - n_j)}{n_{i_j}}\,|i_j - j|$$

where i_j is defined by (6).

Remark 3. Note that although i_j could not be unique, K_j is well defined.

Proof. By definition, TC_{max} will be the value of TC corresponding to the worst situation, which is that for any class $j = 1, \ldots, r$, all the n_j instances belonging to that class have been misclassified in the class with the highest cost, that is, in class i_j. In

other words, $\mathrm{TC}_{max} = \mathrm{TC}(A_{max})$ for the following matrix representing the class $\mathcal{S}_{max}$: $A_{max} = (a_{ij})_{i,j=1,\dots,r}$ given by:

$$\text{for any } j = 1, \dots, r, \quad a_{ij} = \begin{cases} 0 & \text{if } i \neq i_j \\ n_j & \text{if } i = i_j \end{cases}$$

Therefore,

$$\mathrm{TC}_{max} = \frac{1}{N} \sum_{j=1}^{r} n_j \, \gamma_{i_j \, j} \, |i_j - j| = \frac{1}{N} \sum_{j=1}^{r} n_j \, \frac{N - n_j}{n_{i_j}} \, |i_j - j| \, . \qquad \square$$

3.1 The binary case $r = 2$

Proposition 2. *In the binary case,*

$$\mathrm{TC}_{max} = 1 \, .$$

Proof. If $r = 2$, $N = n_1 + n_2$, $i_1 = 2$ and $i_2 = 1$. Then,

$$\mathrm{TC}_{max} = \frac{n_2}{N} \frac{n_1}{n_{i_1}} \, |i_1 - 1| + \frac{n_1}{N} \frac{n_2}{n_{i_2}} \, |i_2 - 2| = \frac{n_2}{N} \frac{n_1}{n_2} + \frac{n_1}{N} \frac{n_2}{n_1} = \frac{n_1 + n_2}{N} = 1 \, . \qquad \square$$

3.2 The perfectly balanced case $n_j = N/r$ for $j = 1, \dots, r$

Proposition 3. *In the perfectly balanced case,*

$$\mathrm{TC}_{max} = \frac{(r-1)^2}{2} + \frac{(r-1)\,(r-h)\,h}{r} \, ,$$

where

$$h = \begin{cases} r/2 & \textit{if } r \textit{ is even,} \\ \lfloor r/2 \rfloor + 1 & \textit{if } r \textit{ is odd.} \end{cases}$$

Therefore,

$$\mathrm{TC}_{max} = \begin{cases} \dfrac{(r-1)\,(3r-2)}{4} & \text{if } r \text{ is even,} \\ \\ \dfrac{(r-1)^2}{4\,r}\,(3r+1) & \text{if } r \text{ is odd} \, . \end{cases}$$

Proof. Since $n_j = N/r$ for $j = 1, \dots, r$, we have that

$$i_j = \arg \max_{\ell=1,\dots,r} |\ell - j| = \begin{cases} r & \text{if } j = 1, \dots, h, \\ 1 & \text{if } j = h+1, \dots, r \, , \end{cases}$$

Note that, if r is odd, we can define $i_h = 1$ instead of $i_h = r$ and this does not change the following calculations. Then, we can write

$$\begin{aligned}\mathrm{TC}_{max} &= \sum_{j=1}^{r} \frac{N - N/r}{N} \frac{N/r}{N/r} |i_j - j| = \frac{r-1}{r} \sum_{j=1}^{r} |i_j - j| \\ &= \frac{r-1}{r} \Big(\sum_{j=1}^{h} (r-j) + \sum_{j=h+1}^{r} (j-1) \Big) = \frac{r-1}{r} \Big(\sum_{i=r-h}^{r-1} i + \sum_{i=h}^{r-1} i \Big) \\ &= \frac{r-1}{r} \left((r-h)\, h + \frac{r\,(r-1)}{2} \right) = \frac{(r-1)^2}{2} + \frac{(r-1)\,(r-h)\,h}{r},\end{aligned}$$

where we have used that the sum of the n consecutive positive integers between a and b, both included, is $\frac{a+b}{2}\, n$. $\square$

Observe that, in the perfect balanced case,

$$A_{max} = \begin{pmatrix} 0 & \dots & 0 & N/r & \dots & N/r \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \underbrace{N/r \dots N/r}_{h \text{ columns}} & & & \underbrace{0 \quad 0 \quad 0}_{r-h \text{ columns}} & & \end{pmatrix}$$

As above, if r is odd, the first block could be composed by $h-1$ columns (and the second by $r-h+1$ columns, respectively) without causing any change in the value of TC_{max}.

Corollary 1. *In the perfectly balanced case,*

$$\mathrm{TC}_{max} \begin{cases} = 1 & \text{if } r = 2, \\ > 1 & \text{if } r > 2, \end{cases}$$

and TC_{max} *is a strictly increasing function of* r.

Proof. By Proposition 2, $\mathrm{TC}_{max} = 1$ if $r = 2$. On the other hand, it is sufficient to prove that for any $k \geq 1$, if we denote by $\mathrm{TC}_{max}(r)$ the value of TC_{max} when the number of classes is r, then we have that

$$\mathrm{TC}_{max}\,(2k) < \mathrm{TC}_{max}\,(2k+1) < \mathrm{TC}_{max}(2k+2)$$

and this is a mere verification, since the first inequality is equivalent to see that

$$\frac{(2\,k-1)\,(3\,(2\,k)-2)}{4} < \frac{(2\,k+1-1)^2\,(3\,(2\,k+1)+1)}{4\,(2\,k+1)}$$

which in turn is equivalent to

$$(4\,k^2-1)\,(3\,k-1) < 4\,k^2\,(3\,k+2) \Longleftrightarrow 12\,k^2+3\,k-1>0\,,$$

which is trivially true for $k \geq 1$.
As for the second inequality, it is

$$\frac{(2\,k+1-1)^2\,(3\,(2\,k+1)+1}{4\,(2\,k+1)} < \frac{(2\,k+2-1)\,(3\,(2\,k+2)-2)}{4}$$

which is equivalent to

$$4\,k^2 < (2\,k+1)^2 \Longleftrightarrow 2\ \leq 2\,k+1\,,$$

which is true for all k, finishing the proof. □

4 Adapting STC metric to classification by intervals

Assume now that the classes $1,\ldots,r$ are the subindexes of intervals $I_1,\ldots,I_r$, such that $I_i=[a_i,\,b_i)$ with $a_i<b_i$ and $a_k=b_{k-1}$ for $k=2,\ldots,r$. Then, we are assuming that they are ordered, in the sense that

$$\forall x\in I_i,\,y\in I_j\,,\quad \text{if } i<j \text{ then } x<y\,.$$

As distance between intervals we will use the Hausdorff distance, although other distances could also be considered. The Hausdorff distance between intervals I_i and I_j is defined by:

$$d(I_i,\,I_j)=\max\{|a_j-a_i|,\,|b_j-b_i|\} \tag{7}$$

We also introduce the idea of "density" of a given interval, as the number of instances of the test set belonging to that interval, divided by its length, that is, the "density" of interval I_i, $i=1,\ldots,r$ is defined by

$$\delta_i=\frac{n_i}{\ell_i}\quad\text{with}\quad \ell_i=b_i-a_i\,. \tag{8}$$

4.1 Definiton of metric $\widehat{\text{STC}}$

We substitute the number of instances belonging to a interval by its "density" in the definition of the γ_{ij} terms of the TC metric. Now, we are ready to introduce the interval-version of the cost-sensitive TC introduced in (3) for standard multi-class classification.

Definition 2. *If $C = (C_{ij})_{i,j=1,\dots,r}$ is the confusion matrix associated with a given classifier, and the ordered classes are intervals $I_1, \dots, I_r$, we define the* **Interval-Total misclassification Cost** *metric as a performance measure for ordinal classification by intervals, by*

$$\widehat{\mathrm{TC}} = \frac{1}{N} \sum_{i,j=1}^{r} C_{ij}\, \widehat{\gamma_{ij}}\, d(I_i,\, I_j) \qquad \textit{with} \quad \widehat{\gamma_{ij}} = \frac{\sum_{k\neq j}^{r} \delta_k}{\delta_i} \geq 1 \;\; \textit{(if } i \neq j\textit{)} \tag{9}$$

where the distances $d(I_i,\, I_j)$ between the intervals are given by (7)*, and "densities" δ_i are defined by* (8)*.*

The variable $\widehat{\gamma_{ij}}$ can be written as $\widehat{\gamma_{ij}} = \frac{\Delta - \delta_j}{\delta_i}$ with $\Delta = \sum_{k=1}^{r} \delta_k$.

Note that if all the intervals have the same length, say L, therefore $\widehat{\mathrm{TC}} = L \times \mathrm{TC}$. Indeed, in this case, $\widehat{\gamma_{ij}} = \gamma_{ij}$ for any $i, j = 1, \dots, r$, and $d(I_i,\, I_j) = L\,|i - j|$, and then,

$$\widehat{\mathrm{TC}} = \frac{L}{N} \sum_{i,j=1}^{r} C_{ij}\, \gamma_{ij}\, |i - j| = L \times \mathrm{TC}\,.$$

Remark 4. In this context, the **cost-matrix** is $\widehat{M} = (\widehat{m_{ij}})_{i,j=1,\dots,r}$, defined by

$$\widehat{m_{ij}} = \frac{\widehat{\gamma_{ij}}\, d(I_i,\, I_j)}{N}$$

and then,

$$\widehat{\mathrm{TC}} = \sum_{i,j=1}^{r} C_{ij} \widehat{m_{ij}} = sum(C \odot \widehat{M})$$

When is the cost matrix $\widehat{M}$ symmetric? $\widehat{M}$ is symmetric if for all $i \neq j$, $\widehat{\gamma_{ij}} = \widehat{\gamma_{ji}}$, which is equivalent to say that

$$\sum_{k\neq i,j} \frac{\delta_k}{\delta_i} = \sum_{k\neq i,j} \frac{\delta_k}{\delta_j} \Leftrightarrow \delta_i = \delta_j,$$

that is, $\widehat{M}$ is symmetric in case of perfectly homogeneous density of the intervals. In this case, if we denote by δ the common value of δ_i, $i = 1, \dots, r$, then, $\delta = n_i/\ell_i$ for all i or, equivalently, $n_i = \delta\, \ell_i$. That is, matrix $\widehat{M}$ is symmetric if and only if the number of instances belonging to any interval is proportional to the length of the interval. We name this case *the perfectly proportional case*.

Analogously to Section 3, we can prove the following result:

Proposition 4. *For any $r \geq 1$, fixed the intervals $I_j = [a_j,\, b_j)$, $j = 1, \dots, r$, with $a_k = b_{k-1}$ for $k = 2, \dots, r$, and the number of instances of the dataset belonging to them,*

$n_1, \ldots, n_r$, respectively, then $\widehat{\mathrm{TC}}$ reaches the maximum $\widehat{\mathrm{TC}}_{max}$ in the set of confusion matrices $\mathcal{S}$, which is

$$\widehat{\mathrm{TC}}_{max} = \frac{1}{N} \sum_{j=1}^{r} \widehat{K}_j \,, \quad \text{with } \widehat{K}_j = n_j \, \frac{\sum_{k \neq j} \delta_k}{\delta_{\widehat{i_j}}} \, d(I_{\widehat{i_j}},\, I_j)$$

where $\widehat{i_j}$ is defined by

$$\widehat{i_j} = \arg \max_{\ell=1,\ldots,r} \frac{d(I_\ell,\, I_j)}{\delta_\ell} \,.$$

In the particular binary case, the following result provides the expression of $\widehat{\mathrm{TC}}_{max}$.

Proposition 5. *In the binary case,*

$$\widehat{\mathrm{TC}}_{max} = \max(\ell_1,\, \ell_2) \,.$$

Proof. If $r = 2$, $N = n_1 + n_2$, $i_1 = 2$ and $i_2 = 1$. Moreover,

$$d(I_1,\, I_2) = d(I_2,\, I_1) = \max(a_2 - a_1,\, b_2 - b_1) = \max(\ell_1,\, \ell_2)$$

(since $a_2 = b_1$). Then,

$$\widehat{K}_1 = n_1 \, \frac{\delta_2}{\delta_2} \, d(I_2,\, I_1) = n_1 \, \max(\ell_1,\, \ell_2) \,,$$

and by symmetry, $\widehat{K}_2 = n_2 \, \max(\ell_1,\, \ell_2)$. Finally,

$$\widehat{\mathrm{TC}}_{max} = \frac{1}{N} \, (n_1 + n_2) \, \max(\ell_1,\, \ell_2) = \max(\ell_1,\, \ell_2) \,. \qquad \square$$

Analogous to ordinal classification without intervals, we can define the following measure:

Definition 3. *If $C = (C_{ij})_{i,j=1,\ldots,r}$ is the confusion matrix associated at a given classifier, we introduce the corresponding* **Interval-Standardized Total misclassification Cost** *metric as a performance measure for ordinal classification by intervals, defined by*

$$\widehat{\mathrm{STC}} = \frac{\widehat{\mathrm{TC}}}{\widehat{\mathrm{TC}}_{max}} \in [0,\, 1] \,,$$

where $\widehat{\mathrm{TC}}$ is introduced in Definition 2 and $\widehat{\mathrm{TC}}_{max}$ in Proposition 4.

4.2 Properties of $\widehat{\mathrm{STC}}$

$\widehat{\mathrm{STC}}$ verifies the properties of a **norm** over the set of $r \times r$ matrices, fixed $r \geq 2$,

$$\mathcal{M} = \{A = (a_{ij})_{i,j=1,\ldots,r} \text{ matrix } r \times r \text{ such that } a_{ij} \in \mathbb{N} \text{ (including 0)}\},$$

except the *Triangle inequality* or *Subbaditivity*. However, the *Positive definiteness* property is replaced by the *Maximal agreement*. Indeed,

- $\widehat{\mathrm{STC}}(A) \geq 0$ for all $A \in \mathcal{M}$ (*Non-negativity*).
  This property holds by definition.
- $\widehat{\mathrm{STC}}(\kappa\, A) = \kappa\, \widehat{\mathrm{STC}}(A)$ for all $\kappa > 0$ and $A \in \mathcal{M}$ (*Homogeneity*).
  This is consequence of the fact that if $A \in \mathcal{S}^r_{n_1,\ldots,n_r}$, then $\kappa\, A \in \mathcal{S}^r_{\kappa\, n_1,\ldots,\kappa\, n_r}$, $\widehat{\mathrm{TC}}_{max}$ matches in the two spaces, and $\widehat{\mathrm{TC}}(\kappa\, A) = \kappa\, \widehat{\mathrm{TC}}(A)$.
- $\widehat{\mathrm{STC}}(A) = 0 \Leftrightarrow A$ is diagonal (*Maximal agreement*).
  This property holds by definition. Moreover, we have that if $A \in \mathcal{S}$,
- $\widehat{\mathrm{STC}}(A) = 1 \Leftrightarrow A \in \mathcal{S}_{max}$ (*Minimal agreement*).

We follow [1] in the consideration of desirable properties that a metric should satisfy, in relation to $\widehat{\mathrm{STC}}$.

- **Property 1 (Scale invariance)**

  The metric is invariant if we change the scale of the units in which the original data is given.

  Proof. Let denote by f the function of a change of scale, that is, a linear function of the form $f(x) = c\, x + d$ with $c > 0$. Denote by $\widehat{\mathrm{STC}}^*$ its value with the data in the new units, after applying the change of scale f. We will see easily that

  $$\widehat{\mathrm{STC}} = \widehat{\mathrm{STC}}^*.$$

  Indeed, since $c > 0$ the change of scale is monotonically increasing, and the intervals after the change of scale are $I_1^*, \ldots, I_r^*$ with $I_i^* = [f(a_i),\, f(b_i)) = [c\, a_i + d,\, c\, b_i + d)$, with length $\ell_i^* = f(b_i) - f(a_i) = c\,(b_i - a_i) = c\, \ell_i$. The Hausdorff distance between intervals I_i^* and I_j^* is:

  $$\begin{aligned} d(I_i^*,\, I_j^*) &= \max\{|f(a_j) - f(a_i)|,\, |f(b_j) - f(b_i)|\} \\ &= |c|\, \max\{|a_j - a_i|,\, |b_j - b_i|\} = c\, d(I_i,\, I_j) \end{aligned}$$

  and the densities after the change of scale are $\delta_i^* = \frac{n_i}{\ell_i^*} = \frac{\delta_i}{c}$, and then, $\widehat{\gamma_{ij}}^* = \widehat{\gamma_{ij}}$ (we use an asterisk to denote the quantities after the scale change). With this, for a given confusion matrix $A = (a_{ij})_{i,j=1,\ldots,r}$,

  $$\widehat{\mathrm{TC}}^*(A) = \frac{1}{N} \sum_{i,j=1}^{r} a_{ij}\, \widehat{\gamma_{ij}}^*\, d(I_i^*,\, I_j^*) = \frac{1}{N} \sum_{i,j=1}^{r} a_{ij}\, \widehat{\gamma_{ij}}\, c\, d(I_i,\, I_j) = c\, \widehat{\mathrm{TC}}(A).$$

On the other hand, for any $j = 1, \ldots, r$,

$$\widehat{i_j}^{\,*} = \arg\max_{\ell=1,\ldots,r} \frac{d(I^*_\ell,\, I^*_j)}{\delta^*_\ell} = \arg\max_{\ell=1,\ldots,r} c^2\, \frac{d(I_\ell,\, I_j)}{\delta_\ell} = \widehat{i_j}$$

and

$$\widehat{K}^*_j = n_j\, \frac{\sum_{k\neq j} \delta^*_k}{\delta^*_{\widehat{i_j}}}\, d(I^*_{\widehat{i_j}},\, I^*_j) = n_j\, \frac{\sum_{k\neq j} \delta_k}{\delta_{\widehat{i_j}}}\, c\, d(I_{\widehat{i_j}},\, I_j) = c\, \widehat{K}_j$$

giving that $\widehat{\mathrm{TC}}^*_{max} = \frac{1}{N}\sum_{j=1}^r \widehat{K}^*_j = c\,\widehat{\mathrm{TC}}_{max}$. Finally, then,

$$\widehat{\mathrm{STC}}^{\,*}(A) = \frac{\widehat{\mathrm{TC}}^{\,*}(A)}{\widehat{\mathrm{TC}}^*_{max}} = \frac{c\,\widehat{\mathrm{TC}}(A)}{c\,\widehat{\mathrm{TC}}_{max}} = \widehat{\mathrm{STC}}(A)\,. \qquad \square$$

- **Property 2 (Monotonicity)**

  Changing predictions farther from the true category (with the same density) should result in an increase in the metric.

  Proof. Indeed, if we change the prediction of an item of a fixed class j_0 from $i_0 \neq j_0$ to $k_0 \neq j_0$ such that

  $$d(I_{k_0},\, I_{j_0}) > d(I_{i_0},\, I_{j_0}) \quad \text{while} \quad \delta_{k_0} = \delta_{i_0}\,,$$

  then $m_{k_0\, j_0} > m_{i_0\, j_0}$ and $\widehat{\mathrm{TC}}$ increases in

  $$m_{k_0\, j_0} - m_{i_0\, j_0} = \frac{\widehat{\gamma_{i_0\, j_0}}}{N}\, \big(d(I_{k_0},\, I_{j_0}) - d(I_{i_0},\, I_{j_0})\big) > 0$$

  (note that $\widehat{\gamma_{i_0\, j_0}} = \widehat{\gamma_{k_0\, j_0}}$). $\square$

- **Property 3 (Imbalance)**

  Distancing items from (respectively, to) a low-density class has more effect on the metric than distancing items from (respectively, to) a high-density class.

  Proof. Indeed, if we consider classes j_0 and j_1, with $\delta_{j_0} < \delta_{j_1}$, which is the effect to misclassify an item of any of these classes to class i_0, assuming that $d(I_{i_0},\, I_{j_0}) = d(I_{i_0},\, I_{j_1})$? (denote these distances by Δ).
  The effect in $\widehat{\mathrm{TC}}$ of misclassify an item of class j_0 (respect., of class j_1) into class i_0 is an increase of quantity:

  $$m_{i_0\, j_0} = \frac{1}{N}\, \widehat{\gamma_{i_0\, j_0}}\, d(I_{i_0},\, I_{j_0}) \quad (\text{respect.}, \quad m_{i_0\, j_1} = \frac{1}{N}\, \widehat{\gamma_{i_0\, j_1}}\, d(I_{i_0},\, I_{j_1})\,)$$

and the difference is:

$$m_{i_0\,j_0} - m_{i_0\,j_1} = \frac{\Delta}{N}\left(\widehat{\gamma_{i_0\,j_0}} - \widehat{\gamma_{i_0\,j_1}}\right) = \frac{\Delta}{N}\,\frac{\delta_{j_1} - \delta_{j_0}}{\delta_{i_0}} > 0\,.$$

As a consequence, $\widehat{\mathrm{TC}}$ increases more if the misclassified item belongs to class j_0 than if it belongs to class j_1.

If, instead, we consider classes i_0 and i_1, with $\delta_{i_0} < \delta_{i_1}$, which is the effect to misclassify an item of a class j_0 to any of these classes, assuming that $d(I_{i_0},\, I_{j_0}) = d(I_{i_1},\, I_{j_0}) = \Delta$? The effect in $\widehat{\mathrm{TC}}$ of misclassify an item of class j_0 into class i_0 (respect., into class i_1) is an increase of quantity:

$$m_{i_0\,j_0} = \frac{1}{N}\,\widehat{\gamma_{i_0\,j_0}}\,d(I_{i_0},\, I_{j_0}) \quad (\text{respect.,} \quad m_{i_1\,j_0} = \frac{1}{N}\,\widehat{\gamma_{i_1\,j_0}}\,d(I_{i_1},\, I_{j_0})\,)$$

and the difference is:

$$m_{i_0\,j_0} - m_{i_1\,j_0} = \frac{\Delta}{N}\left(\widehat{\gamma_{i_0\,j_0}} - \widehat{\gamma_{i_1\,j_0}}\right) = \frac{\Delta}{N}\,\Big(\sum_{k\neq j_0}^{r}\delta_k\Big)\,\Big(\frac{1}{\delta_{i_0}} - \frac{1}{\delta_{i_1}}\Big) > 0\,.$$

This is, in this case it is also fulfilled that $\widehat{\mathrm{TC}}$ increases more if the item is misclassified in class i_0 than in class i_1. $\square$

5 What length to assign to the rightmost interval?

As was said in the introduction, a problem occurs when binning a discrete variable or discretizing a continuous variable since frequently the rightmost interval is unconstrained, which means it cannot be assigned a preset length. To use the $\widehat{\mathrm{STC}}$ measure described in Section 4, however, this interval must have a length defined in advance. How can this issue be solved?

A initial naïve approximation may be to take the maximum of the observations in the given database as the upper limit of the interval. Nothing, however, prevents this sample limit from being surpassed in the future. Furthermore, this approach has a concerning lack of definition, such that one might take the greatest value plus one unit, or plus two, and so on.

After ruling out the first approximation as unsatisfactory, we investigated an alternative strategy based on two pillars. The first is the simple but powerful notion of decoupling the calculation of the number of observations in the rightmost interval from its length. That is, regardless of the length we assign it to compute the metric $\widehat{\mathrm{STC}}$, the number of observations in the interval will be the same, equal to all those that exceed its set lower limit. The second pillar is the concept of determining the length of the rightmost interval, given the lengths of the other intervals are known, in such a way that $\widehat{\mathrm{TC}}_{max}$ is minimized. This is backed by the rationale that minimizing $\widehat{\mathrm{TC}}_{max}$ would maximize the impact of a confusion matrix improvement/worsening on $\widehat{\mathrm{STC}}$, enhancing its capacity to discern differences among classifiers. Let us provide an example to illustrate this.

A toy example

Classifiers A and B give the following confusion matrices for the same validation dataset with $N = 15$ instances, which are classified into $r = 3$ intervals, I_1, I_2 and I_3 with corresponding lengths $\ell_1 = \ell_2 = 1$ and $\ell_3 = x$, undetermined. The validation set is perfectly balanced, with $n_1 = n_2 = n_3 = 5$ instances in each interval.

$$C_A = \begin{array}{c} \\ \text{observed} \\ \end{array} \begin{array}{c} \begin{array}{ccc} I_1 & I_2 & I_3 \end{array} \\ \begin{pmatrix} 3 & 2 & \mathbf{1} \\ 2 & 2 & \mathbf{2} \\ 0 & 1 & 2 \end{pmatrix} \end{array} \qquad C_B = \begin{array}{c} \\ \text{observed} \\ \end{array} \begin{array}{c} \begin{array}{ccc} I_1 & I_2 & I_3 \end{array} \\ \begin{pmatrix} 3 & 2 & \mathbf{2} \\ 2 & 2 & \mathbf{1} \\ 0 & 1 & 2 \end{pmatrix} \end{array}$$

We are going to calculate the $\widehat{\text{STC}}$ metric for both matrices. First, we calculate $\widehat{\text{TC}}$ using (9). Indeed,

$$\widehat{\text{TC}}(C) = \sum_{i,j=1}^{3} C_{ij}\widehat{m_{ij}} = sum(C \odot \widehat{M})$$

where $\widehat{M} = (\widehat{m_{ij}})_{i,j=1,\dots,3}$ and $\widehat{m_{ij}} = \frac{\widehat{\gamma_{ij}}\, d(I_i, I_j)}{15}$, with $d(I_1, I_2) = 1$, $d(I_1, I_3) = 1 + \max(1, x)$, $d(I_2, I_3) = \max(1, x)$, and $\delta_1 = n_1 = 5$, $\delta_2 = n_2 = 5$, $\delta_3 = n_3/x = 5/x$. Using that $\widehat{\gamma_{ij}} = \frac{\sum_{k\neq j}^{3} \delta_k}{\delta_i}$ we have that matrix M is

$$M = \frac{1}{15} \begin{pmatrix} 0 & 1 + \frac{1}{x} & 2\,(1 + \max(1, x)) \\ 1 + \frac{1}{x} & 0 & 2\,\max(1, x) \\ (1 + x)\,(1 + \max(1, x)) & (1 + x)\,\max(1, x) & 0 \end{pmatrix},$$

$$\widehat{\text{TC}}(C_A) = sum(C_A \odot \widehat{M}) = \frac{1}{15}\left(6 + \frac{4}{x} + (7 + x)\,\max(1, x)\right),$$
$$\widehat{\text{TC}}(C_B) = sum(C_B \odot \widehat{M}) = \widehat{\text{TC}}(C_A) + \frac{2}{15}$$

and therefore, $\widehat{\text{TC}}(C_B) - \widehat{\text{TC}}(C_A) = 2/15$, which is independent of x. Since $\widehat{\text{STC}}(C) = \frac{\widehat{\text{TC}}(C)}{\widehat{\text{TC}}_{max}}$, with $\widehat{\text{TC}}_{max}$ independent of the values of matrix C, given the column sums are fixed, it is obvious that the value of x that minimizes $\widehat{\text{TC}}_{max}$ at the same time maximizes the relative difference between $\widehat{\text{TC}}(C_A)$ and $\widehat{\text{TC}}(C_B)$.
By Proposition 6 below we obtain that in this example, the value of x that minimizes $\widehat{\text{TC}}_{max}$ is $\tilde{x} = 1/\sqrt{2}$ and $\widehat{\text{TC}}_{max}(\tilde{x}) = (2\sqrt{2}+7)/3$. Then, we assign a length of $\tilde{x} = 1/\sqrt{2}$

to interval I_3 for the purposes of computing the metric, and

$$\widehat{\mathrm{STC}}(C_A) = \frac{\widehat{\mathrm{TC}}(C_A)}{\widehat{\mathrm{TC}}_{max}} = \frac{\frac{1}{15}\left(6 + \frac{4}{\tilde{x}} + (7+\tilde{x})\,\max(1,\tilde{x})\right)}{\frac{2\sqrt{2}+7}{3}} = \frac{73 + \frac{11}{\sqrt{2}}}{205}$$

$$\widehat{\mathrm{STC}}(C_B) = \frac{\widehat{\mathrm{TC}}(C_B)}{\widehat{\mathrm{TC}}_{max}} = \frac{\widehat{\mathrm{TC}}(C_A) + \frac{2}{15}}{\widehat{\mathrm{TC}}_{max}} = \frac{87 + \frac{3}{\sqrt{2}}}{205}$$

and $\widehat{\mathrm{STC}}(C_B) - \widehat{\mathrm{STC}}(C_A) = \frac{14 - \frac{8}{\sqrt{2}}}{205} \approx 0.0407$. With any other value of x, the difference between the two would be less.

The general perfectly balanced $r = 3$ case

The following result, whose proof is in Appendix A, states the length of the rightmost interval that minimizes $\widehat{\mathrm{TC}}_{max}$ in a particular case that, precisely, we have used in the toy example. This particular case fully covers the situation of having $r = 3$ intervals, in the perfectly balanced setting where the number of cases in each interval is the same. Although it is a particular scenario, the balanced case is of great interest in practice since one of the most used methods for the discretization of continuous variables, if not the most, consists of dividing into intervals in such a way the number of cases in each interval be the same, used by default in the functions that implement the discretization algorithms in R (it is the case of `method="frequency"` in function `arules::discretize`, and of `method="quantile"` in function `bnlearn::discretize`).

Proposition 6. *If $r = 3$, $\ell_1 = 1$ and $\ell_2 = L > 0$ known, and if $n_1 = n_2 = n_3$, the value of $x = \ell_3$ that minimizes $\widehat{\mathrm{TC}}_{max}$ is given by:*

$$\tilde{x} = \begin{cases} \sqrt{\frac{L}{L+1}} & \text{if } L \leq 1, \text{ and } \widehat{\mathrm{TC}}_{max}(\tilde{x}) = \frac{1}{3}\left(2\sqrt{\frac{L+1}{L}} + 2\,L + 4 + \frac{1}{L}\right) \\ \frac{L}{\sqrt{L+1}} & \text{if } 1 < L \leq \frac{1+\sqrt{5}}{2}, \text{ and } \widehat{\mathrm{TC}}_{max}(\tilde{x}) = \frac{1}{3}\left(2\sqrt{L+1} + 3\,L + 3 + \frac{1}{L}\right) \\ \frac{L\,(\sqrt{5}-1)}{2} & \text{if } \frac{1+\sqrt{5}}{2} < L \leq \frac{3+\sqrt{5}}{2}, \text{ and } \widehat{\mathrm{TC}}_{max}(\tilde{x}) = \frac{L}{6}\left((\sqrt{5}+1)\,L + \sqrt{5} + 7\right) \\ \sqrt{L} & \text{if } \frac{3+\sqrt{5}}{2} < L, \text{ and } \widehat{\mathrm{TC}}_{max}(\tilde{x}) = \frac{L}{3}\left(2\sqrt{L} + L + 3\right). \end{cases}$$

Note that taking ℓ_1 as a reference, L is the ratio ℓ_2/ℓ_1, and that the Proposition 6 provides the optimal value for $x = \ell_3/\ell_1$ in function of L. Figure 1 is a visual illustration of the result, which states that for large values of L (greater than $\frac{3+\sqrt{5}}{2} \approx 2.62$), $\tilde{x}$ is $\sqrt{L}$.

6 Conclusions

In this article, we address the task of accurate evaluation of classification models in the domain of Supervised Machine Learning. Specifically, we focus on the challenges posed by ordinal and interval classification, as well as the consideration of unbounded rightmost intervals.

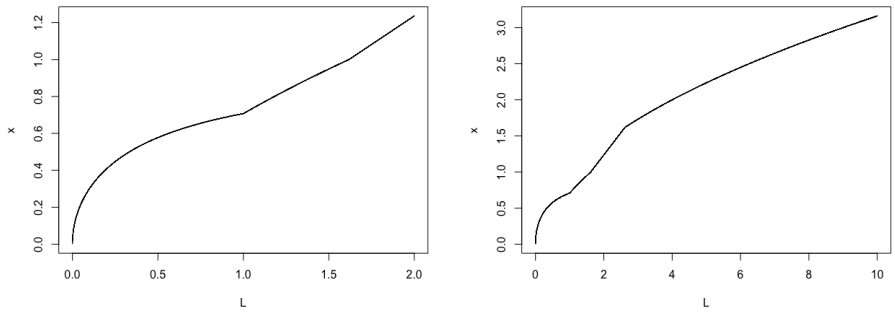

Fig. 1. Plot of $\tilde{x}$ as function of L, for $L \in [0, 2]$ (left) and $L \in [0, 10]$ (right).

First, we identify the limitations of traditional metrics in capturing the inherent ordinal structure of data in ordinal classification. To overcome this, we recall an existing metric based on a cost-sensitive approach and offer a simplified calculation method.
We extend our results to the case where the classification is made by intervals. We recognize that, in addition to the order, the length of intervals plays an important role. We define a new metric, based on the one for standard ordinal classification, and investigate some of its properties.
Furthermore, we address the challenge of classification with unbounded rightmost intervals. By incorporating this consideration into our metrics, we enhance their applicability and provide a more robust evaluation framework that aligns with the complexities of practical classification tasks.
Moving forward, our research opens avenues for further exploration, such as investigating the scalability and robustness of the proposed metrics on larger datasets and exploring their applicability in specific domains.

Authorship contribution statement

The two authors have contributed equally.

Funding

The authors are supported by Ministerio de Ciencia e Innovación, Gobierno de España, project ref. PID2021-123733NB-I00

Declaration of competing interest

The authors declare that they have not competing interests in relation to the research reported in this paper.

References

1. Amigó, E., Gonzalo, J., Mizzaro, S., Carrillo-de-Albornoz, J. (2020) An effectiveness metric for ordinal classification: formal properties and experimental results. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3938–3949.
2. Baccianella, S., Esuli, A., Sebastiani, F. (2009) Evaluation measures for ordinal regression. IEEE 2009 Ninth International Conference on Intelligent Systems Design and Applications.
3. Cano, J.R., García, S. (2017) Training set selection for monotonic ordinal classification. Data & Knowledge Engineering, vol. 112, pp. 94–105.
4. Cardoso J.S., Sousa R. (2011) Measuring the performance of ordinal classification. International Journal of Pattern Recognition and Artificial Intelligence, vol. 25(8), pp. 1173-1195.
5. Cruz-Ramírez, M., Hervás-Martínez, C., Sánchez-Monedero, J., Gutiérrez, P.A. (2011) A preliminary study of ordinal metrics to guide multi-objective evolutionary algorithm. IEEE 2011 11th International Conference on Intelligent Systems Design and Applications.
6. Cruz-Ramírez, M., Hervás-Martínez, C., Sánchez-Monedero, J., Gutiérrez, P.A. (2014) Metrics to guide a multi-objective evolutionary algorithm for ordinal classification. Neurocomputing, vol. 135, pp. 21–31.
7. Gaudette, L., Japkowicz, N. (2009) Evaluation methods for ordinal classification. Lecture Notes in Computer Science 5549, pp. 207–210.
8. George, N.I., Lu, T-P., Chang, Ch-W. (2016) Cost-sensitive Performance Metric for Comparing Multiple Ordinal Classifiers. Artif Intell Res., vol. 5(1), pp. 135–143. doi:10.5430/air.v5n1p135
9. Pang B., Lee L. (2008) Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1/2), pp. 1–135.
10. Stevens S.S. (1946) On the theory of scales of measurement. Science, New Series, 103(2684), pp. 677-680.
11. Waegeman, W., De Baets, B., Boullart, L. (2006) A comparison of different ROC measures for ordinal regression. In Proceedings of the CML 2006 workshop on ROC Analysis in Machine Learning.
12. Yilmaz A.E., Demirhan H. (2023) Weighted Kappa measures for ordinal multi-class classification performance. Applied Soft Computing, 134, 110020.

Appendix A: Proof of Proposition 6

To find the value of x that minimizes $\widehat{\mathrm{TC}}_{max}$ we use its expression, as given in Proposition 4, that is,

$$\widehat{\mathrm{TC}}_{max} = \frac{1}{3\,n}\sum_{j=1}^{3}\widehat{K}_j\,, \quad \text{with } \widehat{K}_j = n\,\frac{\sum_{k\neq j}\delta_k}{\delta_{\widehat{i_j}}}\,d(I_{\widehat{i_j}},\,I_j)$$

where $\widehat{i_j} = \arg\max\limits_{\ell=1,2,3}\dfrac{d(I_\ell,\,I_j)}{\delta_\ell}$ and n denotes the common number of instances at any interval ($n = n_i$, $i = 1, 2, 3$). Note that

$$d(I_1,\,I_2) = \max(1,\,L), \quad d(I_2,\,I_3) = \max(L,\,x), \quad d(I_1,\,I_3) = L + \max(1,\,x)$$

so a first division in cases will be according to the value of $\max(1,\,L)$.

Case a) $L \leq 1$ (then, $d(I_1,\,I_2) = \max(1,\,L) = 1$).

If $j = 1$, $\widehat{i_1} = \arg\max\limits_{\ell=1,2,3}\dfrac{d(I_\ell,\,I_1)}{\delta_\ell}$. Taking into account that

$$\frac{d(I_\ell,\,I_1)}{\delta_\ell} = \begin{cases} \frac{L}{n} & \text{if } \ell = 2 \\ \frac{x}{n}\left(L + \max(1,\,x)\right) = \begin{cases} \frac{x}{n}\,(L+1) & \text{if } 0 < x \leq 1 \\ \frac{x}{n}\,(x+L) & \text{if } x > 1 \end{cases} & \text{if } \ell = 3 \end{cases}$$

we have that

$$\widehat{i_1} = \begin{cases} 2 & \text{if } 0 < x \leq \frac{L}{L+1} \\ 3 & \text{if } x > \frac{L}{L+1}\,. \end{cases}$$

Then,

$$\widehat{K}_1 = n\,\frac{\sum_{k=2,3}\delta_k}{\delta_{\widehat{i_1}}}\,d(I_{\widehat{i_1}},\,I_1) = \begin{cases} n\,\frac{(x+L)}{x} & \text{if } 0 < x \leq \frac{L}{L+1} \\ n\,(x+L)\,\frac{(L+1)}{L} & \text{if } \frac{L}{L+1} < x \leq 1 \\ n\,\frac{(x+L)^2}{L} & \text{if } 1 < x\,. \end{cases}$$

Analogously,

$$\widehat{i_2} = \begin{cases} 1 & \text{if } 0 < x \leq 1 \\ 3 & \text{if } x > 1 \end{cases}\,, \qquad \widehat{i_3} = 1\ \forall x > 0\,,$$

and

$$\widehat{K}_2 = n\,\frac{\sum_{k=1,3}\delta_k}{\delta_{\widehat{i_2}}}\,d(I_{\widehat{i_2}},\,I_2) = \begin{cases} n\,(1+\frac{1}{x}) & \text{if } 0 < x \leq 1 \\ n\,x\,(x+1) & \text{if } x > 1\,, \end{cases}$$

$$\widehat{K}_3 = n\,\frac{\sum_{k=1,2}\delta_k}{\delta_{\widehat{i_3}}}\,d(I_{\widehat{i_3}},\,I_3) = \begin{cases} n\,\frac{(L+1)^2}{L} & \text{if } 0 < x \leq 1 \\ n\,(x+L)\,\frac{(L+1)}{L} & \text{if } x > 1\,. \end{cases}$$

Finally,

$$\widehat{\mathrm{TC}}_{max} = \frac{1}{3\,n}\sum_{j=1}^{3}\widehat{K}_j = \begin{cases} \frac{1}{3}\left(\frac{(L+1)}{x} + L + 4 + \frac{1}{L}\right) & \text{if } 0 < x \leq \frac{L}{L+1} \\ \frac{1}{3}\left((1+\frac{1}{L})\,x + \frac{1}{x} + 2\,L + 4 + \frac{1}{L}\right) & \text{if } \frac{L}{L+1} < x \leq 1 \\ \frac{1}{3}\left((1+\frac{1}{L})\,x^2 + (4+\frac{1}{L})\,x + 2\,L + 1\right) & \text{if } x > 1\,. \end{cases}$$

Note that $\widehat{\mathrm{TC}}_{max}$ is a function of x independent of n, which is continuous for $x > 0$, and piecewise differentiable. Its minimum at each interval of definition is given by:

$$\min \widehat{\mathrm{TC}}_{max}(x) = \begin{cases} \frac{2}{3}\left(L + 3 + \frac{1}{L}\right) & \text{reached at } \tilde{x} = \frac{L}{L+1},\ \text{if } 0 < x \leq \frac{L}{L+1} \\ \frac{1}{3}\left(2\sqrt{\frac{L+1}{L}} + 2L + 4 + \frac{1}{L}\right) & \text{reached at } \tilde{x} = \sqrt{\frac{L}{L+1}},\ \text{if } \frac{L}{L+1} < x \leq 1 \\ \frac{2}{3}\left(L + 3 + \frac{1}{L}\right) & \text{reached at } \tilde{x} = 1,\ \text{if } x > 1\,. \end{cases}$$

Since $\frac{1}{3}\left(2\sqrt{\frac{L+1}{L}} + 2\,L + 4 + \frac{1}{L}\right) < \frac{2}{3}\left(L + 3 + \frac{1}{L}\right)$ holds always, we have that $\min \widehat{\mathrm{TC}}_{max}(x) = \frac{1}{3}\left(2\sqrt{\frac{L+1}{L}} + 2\,L + 4 + \frac{1}{L}\right)$ and is reached at $\tilde{x} = \sqrt{\frac{L}{L+1}}$.

Case b) $L > 1$ (then, $d(I_1,\, I_2) = \max(1,\, L) = L$).

If $j = 1$, $\widehat{i_1} = \arg\max_{\ell=1,2,3} \dfrac{d(I_\ell,\, I_1)}{\delta_\ell}$ with

$$\frac{d(I_\ell,\, I_1)}{\delta_\ell} = \begin{cases} \frac{L^2}{n} & \text{if } \ell = 2 \\ \frac{x}{n}\left(L + \max(1,\, x)\right) = \begin{cases} \frac{x}{n}\,(L+1) & \text{if } 0 < x \leq 1 \\ \frac{x}{n}\,(x+L) & \text{if } x > 1 \end{cases} & \text{if } \ell = 3 \end{cases}$$

and therefore, distinguishing when $L^2 < x\,(L+1)$ if $x \leq 1$, and when $L^2 < x\,(x+L)$ if $x > 1$, we have that

$$\begin{cases} \text{If } 1 < L \leq \frac{1+\sqrt{5}}{2}, & \widehat{i_1} = \begin{cases} 2 & \text{if } 0 < x \leq \frac{L^2}{L+1} \\ 3 & \text{if } x > \frac{L^2}{L+1} \end{cases} \\ \text{If } L > \frac{1+\sqrt{5}}{2}, & \widehat{i_1} = \begin{cases} 2 & \text{if } 0 < x \leq \frac{L}{2}\,(\sqrt{5}-1) \\ 3 & \text{if } x > \frac{L}{2}\,(\sqrt{5}-1)\,. \end{cases} \end{cases}$$

Then, using that $\widehat{K}_1 = n\,\dfrac{\sum_{k=2,3}\delta_k}{\delta_{\widehat{i_1}}}\,d(I_{\widehat{i_1}},\, I_1)$ we have that

$$\begin{cases} \text{If } 1 < L \leq \frac{1+\sqrt{5}}{2}, & \widehat{K}_1 = \begin{cases} n\,L\,\frac{x+L}{x} & \text{if } 0 < x \leq \frac{L^2}{L+1} \\ n\,(x+L)\,\frac{L+1}{L} & \text{if } \frac{L^2}{L+1} < x \leq 1 \\ n\,\frac{(x+L)^2}{L} & \text{if } x > 1 \end{cases} \\ \text{If } L > \frac{1+\sqrt{5}}{2}, & \widehat{K}_1 = \begin{cases} n\,L\,\frac{x+L}{x} & \text{if } 0 < x \leq \frac{L}{2}\,(\sqrt{5}-1) \\ n\,\frac{(x+L)^2}{L} & \text{if } x > \frac{L}{2}\,(\sqrt{5}-1)\,. \end{cases} \end{cases}$$

With the same approach, we can get that

$$\widehat{i_2} = \begin{cases} 1 & \text{if } 0 < x \leq 1 \\ 3 & \text{if } x > 1 \end{cases}, \qquad \widehat{K}_2 = \begin{cases} n\,L\,(1+\frac{1}{x}) & \text{if } 0 < x \leq 1 \\ n\,L\,(x+1) & \text{if } 1 < x \leq L \\ n\,x\,(x+1) & \text{if } x > L\,, \end{cases}$$

$$\begin{cases} \text{If } 1 < L \leq \frac{1+\sqrt{5}}{2}, & \widehat{i_3} = \begin{cases} 1 & \text{if } 0 < x \leq \frac{L}{L-1} \\ 2 & \text{if } x > \frac{L}{L-1} \end{cases} \\ \text{If } \frac{1+\sqrt{5}}{2} < L \leq 2, & \widehat{i_3} = \begin{cases} 2 & \text{if } 0 < x \leq L\,(L-1) \\ 1 & \text{if } L\,(L-1) < x \leq \frac{L}{L-1} \\ 2 & \text{if } x > \frac{L}{L-1} \end{cases} \\ \text{If } L > 2, & \widehat{i_3} = 2 \quad \forall x > 0\,, \end{cases}$$

$$\begin{cases} \text{If } 1 < L \leq \frac{1+\sqrt{5}}{2}, & \widehat{K}_3 = \begin{cases} n\,\frac{(L+1)^2}{L} & \text{if } 0 < x \leq 1 \\ n\,(L+1)\,\frac{x+L}{L} & \text{if } 1 < x \leq \frac{L}{L-1} \\ n\,(L+1)\,x & \text{if } x > \frac{L}{L-1} \end{cases} \\ \text{If } \frac{1+\sqrt{5}}{2} < L \leq 2, & \widehat{K}_3 = \begin{cases} n\,L\,(L+1) & \text{if } 0 < x \leq L\,(L-1) \\ n\,(L+1)\,\frac{x+L}{L} & \text{if } L\,(L-1) < x \leq \frac{L}{L-1} \\ n\,(L+1)\,x & \text{if } x > \frac{L}{L-1} \end{cases} \\ \text{If } L > 2, & \widehat{K}_3 = \begin{cases} n\,L\,(L+1) & \text{if } 0 < x \leq L \\ n\,(L+1)\,x & \text{if } x > L\,. \end{cases} \end{cases}$$

Then, using that $\widehat{\mathrm{TC}}_{max} = \dfrac{1}{3\,n}\displaystyle\sum_{j=1}^{3} \widehat{K}_j$, we obtain the following:

Case b.1) $1 < L \leq \frac{1+\sqrt{5}}{2}$

$$\widehat{\mathrm{TC}}_{max} = \begin{cases} \frac{1}{3}\left((L^2+L)\,\frac{1}{x} + (3\,L+2+\frac{1}{L})\right) & \text{if } 0 < x \leq \frac{L^2}{L+1} \\ \frac{1}{3}\left(\frac{L+1}{L}\,x + \frac{L}{x} + (3\,L+3+\frac{1}{L})\right) & \text{if } \frac{L^2}{L+1} < x \leq 1 \\ \frac{1}{3}\left(\frac{x^2}{L} + (3+L+\frac{1}{L})\,x + (3\,L+1)\right) & \text{if } 1 < x \leq L \\ \frac{1}{3}\left((1+\frac{1}{L})\,x^2 + (4+\frac{1}{L})\,x + (2\,L+1)\right) & \text{if } L < x \leq \frac{L}{L-1} \\ \frac{1}{3}\left((1+\frac{1}{L})\,x^2 + (4+L)\,x + L\right) & \text{if } x > \frac{L}{L-1}\,, \end{cases}$$

which is a continuous function of $x > 0$ independent of n and piecewise differentiable. Its minimum at each interval of definition is given by:

$$\min \widehat{\mathrm{TC}}_{max}(x) = \begin{cases} \frac{2}{3}\big(2\,L + 2 + \frac{1}{L}\big) & \text{reached at } \tilde{x} = \frac{L^2}{L+1},\ \text{if } 0 < x \le \frac{L^2}{L+1} \\ \frac{1}{3}\big(2\sqrt{L+1} + 3L + 3 + \frac{1}{L}\big) & \text{reached at } \tilde{x} = \frac{L}{\sqrt{L+1}},\ \text{if } \frac{L^2}{L+1} \le x \le 1 \\ \frac{2}{3}\big(2\,L + 2 + \frac{1}{L}\big) & \text{reached at } \tilde{x} = 1,\ \text{if } 1 \le x \le L \\ \frac{1}{3}\big(L^2 + 7\,L + 2\big) & \text{reached at } \tilde{x} = L,\ \text{if } L \le x \le \frac{L}{L-1} \\ \frac{2}{3}\,\frac{L}{(L-1)^2}\,\big(L^2 + L - 1\big) & \text{reached at } \tilde{x} = \frac{L}{L-1},\ \text{if } x > \frac{L}{L-1}\,, \end{cases}$$

and comparing the above values with each other, we obtain that in case **b.1)**, $\min \widehat{\mathrm{TC}}_{max}(x) = \dfrac{1}{3}\big(2\sqrt{L+1} + 3L + 3 + \dfrac{1}{L}\big)$, which is reached at $\tilde{x} = \frac{L}{\sqrt{L+1}}$.

Case b.2) $\frac{1+\sqrt{5}}{2} < L \le 2$

$$\widehat{\mathrm{TC}}_{max} = \begin{cases} \frac{1}{3}\left(L\,(L+1)\,\frac{1}{x} + L\,(L+3)\right) & \text{if } 0 < x \le 1 \\ \frac{1}{3}\left(L\,x + \frac{L^2}{x} + L\,(L+3)\right) & \text{if } 1 < x \le \frac{L}{2}\,(\sqrt{5}-1) \\ \frac{1}{3}\left(\frac{x^2}{L} + (L+2)\,x + L\,(L+3)\right) & \text{if } \frac{L}{2}\,(\sqrt{5}-1) < x \le L\,(L-1) \\ \frac{1}{3}\left(\frac{x^2}{L} + (L+3+\frac{1}{L})\,x + (3\,L+1)\right) & \text{if } L\,(L-1) < x \le L \\ \frac{1}{3}\left((1+\frac{1}{L})\,x^2 + (4+\frac{1}{L})\,x + (2\,L+1)\right) & \text{if } L < x \le \frac{L}{L-1} \\ \frac{1}{3}\left((1+\frac{1}{L})\,x^2 + (4+L)\,x + L\right) & \text{if } x > \frac{L}{L-1}\,, \end{cases}$$

which again is a continuous function of $x > 0$ independent of n and piecewise differentiable, whose minimum at each interval of definition is given by:

$$\min \widehat{\mathrm{TC}}_{max}(x) = \begin{cases} \frac{2\,L}{3}\,(L+2) & \text{reached at } \tilde{x} = 1,\ \text{if } 0 < x \le 1 \\ \frac{L}{6}\big((\sqrt{5}+1)\,L + \sqrt{5} + 7\big) & \text{reached at } \tilde{x} = \frac{L}{2}\,(\sqrt{5}-1),\ \text{if } 1 \le x \le L\,(L-1) \\ \frac{2\,L}{3}\,(L^2+1) & \text{reached at } \tilde{x} = L\,(L-1),\ \text{if } L\,(L-1) \le x \le L \\ \frac{1}{3}\big(L^2 + 7\,L + 2\big) & \text{reached at } \tilde{x} = L,\ \text{if } L \le x \le \frac{L}{L-1} \\ \frac{2}{3}\,\frac{L}{(L-1)^2}\,\big(L^2 + L - 1\big) & \text{reached at } \tilde{x} = \frac{L}{L-1},\ \text{if } x > \frac{L}{L-1}\,. \end{cases}$$

Comparing the above values with each other, we obtain that, in case **b.2)**, $\min \widehat{\mathrm{TC}}_{max}(x) = \dfrac{L}{6}\big((\sqrt{5}+1)\,L + \sqrt{5} + 7\big)$ and it is reached at $\tilde{x} = \frac{L}{2}\,(\sqrt{5}-1)$.

Case b.3) $L > 2$

$$\widehat{\mathrm{TC}}_{max} = \begin{cases} \frac{1}{3}\left(L\,(L+1)\,\frac{1}{x} + L\,(L+3)\right) & \text{if } 0 < x \le 1 \\ \frac{1}{3}\left(L\,x + \frac{L^2}{x} + L\,(L+3)\right) & \text{if } 1 < x \le \frac{L}{2}\,(\sqrt{5}-1) \\ \frac{1}{3}\left(\frac{x^2}{L} + (L+2)\,x + L\,(L+3)\right) & \text{if } \frac{L}{2}\,(\sqrt{5}-1) < x \le L \\ \frac{1}{3}\left((1+\frac{1}{L})\,x^2 + (4+L)\,x + L\right) & \text{if } x > L\,, \end{cases}$$

which is continuous as function of $x > 0$, independent of n, and piecewise differentiable, whose minimum at each interval of definition is given by:

$$\min \widehat{\mathrm{TC}}_{max}(x) =$$
$$\begin{cases} \frac{2\,L}{3}\,(L+2) \quad \text{reached at } \tilde{x}=1, & \text{if } 0 < x \leq 1 \\ \begin{cases} \frac{L}{6}\big((\sqrt{5}+1)\,L+\sqrt{5}+7\big) & \text{at } \tilde{x}=\frac{L}{2}\,(\sqrt{5}-1),\ \text{if } 2 < L \leq \frac{3+\sqrt{5}}{2} \\ \frac{L}{3}\big(2\,\sqrt{L}+L+3\big) & \text{at } \tilde{x}=\sqrt{L},\ \text{if } L > \frac{3+\sqrt{5}}{2} \end{cases} & \text{if } 1 < x \leq \frac{L}{2}\,(\sqrt{5}-1) \\ \frac{L}{6}\big((\sqrt{5}+1)\,L+\sqrt{5}+7\big) \quad \text{reached at } \tilde{x}=\frac{L}{2}\,(\sqrt{5}-1), & \text{if } \frac{L}{2}\,(\sqrt{5}-1) < x \leq L \\ \frac{2\,L}{3}\,(L+3) \quad \text{reached at } \tilde{x}=L, & \text{if } x > L\,. \end{cases}$$

Comparing the above values with each other in case **b.3)**, if $2 < L \leq \frac{3+\sqrt{5}}{2}$ we obtain the same as in case **b.2)**: $\min \widehat{\mathrm{TC}}_{max}(x) = \dfrac{L}{6}\big((\sqrt{5}+1)\,L+\sqrt{5}+7\big)$ and it is reached at $\tilde{x} = \frac{L}{2}\,(\sqrt{5}-1)$, while, if $L > \frac{3+\sqrt{5}}{2}$, we obtain instead that $\min \widehat{\mathrm{TC}}_{max}(x) = \dfrac{L}{3}\big(2\,\sqrt{L}+L+3\big)$, which is reached at $\tilde{x} = \sqrt{L}\,.$ $\square$