Cautious Ordinal Classification by Binary Decomposition

Sébastien Destercke and Gen Yang

Université de Technologie de Compiegne U.M.R. C.N.R.S. 7253 Heudiasyc Centre de recherches de Royallieu F-60205 Compiegne Cedex France
{sebastien.destercke,gen.yang}@hds.utc.fr

Abstract. We study the problem of performing cautious inferences for an ordinal classification (a.k.a. ordinal regression) task, that is when the possible classes are totally ordered. By cautious inference, we mean that we may produce partial predictions when available information is insufficient to provide reliable precise ones. We do so by estimating probabilistic bounds instead of precise ones. These bounds induce a (convex) set of possible probabilistic models, from which we perform inferences. As the estimates or predictions for such models are usually computationally harder to obtain than for precise ones, we study the extension of two binary decomposition strategies that remain easy to obtain and computationally efficient to manipulate when shifting from precise to bounded estimates. We demonstrate the possible usefulness of such a cautious attitude on tests performed on benchmark data sets.

Keywords: Ordinal regression, imprecise probabilities, Binary decomposition, Nested dichotomies.

1 Introduction

We are interested in the supervised learning problem known as *ordinal classification* [18] or *regression* [9]. In this problem, the finite set of possible labels are naturally ordered. For instance, the rating of movies can be one of the following labels: `Very-Bad`, `Bad`, `Average`, `Good`, `Very-Good` that are ordered from the worst situation to the best. Such problems are different from multi-class classification and regression problems, since in the former there is no ordering between classes and in the latter there exists a metric on the outputs (while in ordinal classification, a 5-star movie should not be considered five times better than a 1-star movie).

A common approach to solve this problem is to associate the labels to their rank, e.g., $\{1,2,3,4,5\}$ in our previous film example, and then to learn a ranking function. In the past years, several algorithms and methods [27] have been proposed to learn such a function, such as SVM techniques [26,22,23,25], monotone functions [28], binary decomposition [20], rule based models [14]. This is not the approach followed in this paper, in which our goal is to estimate the probability of the label conditionally on the observed instance. In this sense, our approach is much closer to the one proposed by Frank et Hall [18].

A common feature of all the previously cited approaches is that, no matter how reliable is the model and the amount of data it is learned from, it will always produce a

T. Calders et al. (Eds.): ECML PKDD 2014, Part I, LNCS 8724, pp. 323–337, 2014.

unique label as prediction. In this paper, we are interested in making partial predictions when information is insufficient to provide a reliable precise one. That is, if we are unsure of the right label, we may abstain to make a precise prediction and instead predict a subset of potentially optimal labels. The goal is similar to the one pursued by the use of a reject option [2,8], and in particular to methods returning subsets of possible classes [1,21]. Yet we will see in the experiments that the two approaches can provide very different results.

Besides the fact that such cautious predictions can prevent bad decisions based on wrong predictions, making such imprecise predictions in an ordinal classification setting can also be instrumental in more complex problems that can be decomposed into sets of ordinal classification problem, such as graded multi-label [7] or label ranking [6]. Indeed, in such problems with structured outputs, obtaining fully reliable precise predictions is much more difficult, hence producing partial but more reliable predictions is even more interesting [5].

To obtain these cautious predictions, we propose to estimate sets of probabilities [10] from the data in the form of probabilistic bounds over specific events, and to then derive the (possibly) partial predictions from it. As computations with generic methods using sets of probabilities (e.g., using imprecise graphical models [10]) can be quite complex, we propose in Section 2 to consider two well-known binary decompositions whose extension to probability sets keep computations tractable, namely Frank & Hall decomposition [18] and nested dichotomies decompositions [17]. In Section 3, we discuss how to perform inferences from such probability sets both with general loss functions and with the classical 0/1 loss function. We end (Section 4) by providing several experiments showing that our cautious approach can help identify hard to predict cases and provides more reliable predictions for those cases.

2 Probability Set Estimation through Binary Decomposition

The goal of ordinal classification is to associate an instance $\mathbf{x} = \mathbf{x}^1 \times \ldots \times \mathbf{x}^p$ coming from an instance space $\mathscr{X} = \mathscr{X}^1 \times \ldots \times \mathscr{X}^p$ to a single label of the space $\mathscr{Y} = \{y_1, \ldots, y_m\}$ of possible classes. Ordinal classification differs from multi-class classification in that labels y_i are ordered, that is $y_i \prec y_{i+1}$ for $i = 1, \ldots, m-1$. An usual task is then to estimate the theoretical conditional probability measure $P_{\mathbf{x}} : 2^{\mathscr{Y}} \to [0,1]$ associated to an instance $\mathbf{x}$ from a set of n training samples $(\mathbf{x}_i, \ell_{x_i}) \in \mathscr{X} \times \mathscr{Y}$, $i = 1, \ldots, n$.

In order to derive cautious inferences, we shall explore in this paper the possibility to provide a convex set $\mathscr{P}_{\mathbf{x}}$ of probabilities as an estimate rather than a precise probability $\hat{P}_{\mathbf{x}}$, with the idea that the size of $\mathscr{P}_{\mathbf{x}}$ should decrease as more data (i.e., information) become available, converging to $P_{\mathbf{x}}$.

Manipulating generic sets $\mathscr{P}_{\mathbf{x}}$ to compute expectations or make inferences can be tedious, hence it is interesting to focus on collections of assessments that are easy to obtain and induce sets $\mathscr{P}_{\mathbf{x}}$ that allow for easy computations. Here we focus on the extensions of two particularly attractive binary decomposition techniques already used to estimate a precise $\hat{P}_{\mathbf{x}}$, namely Frank et Hall [18] technique and nested dichotomies [19].

2.1 Imprecise Cumulative Distributions

In their original paper, Frank et Hall suggest [18] to estimate, for an instance $\mathbf{x}$, the probabilities that its output ℓ_x will be less or equal than y_k, $k = 1,\dots,m-1$. That is, one should estimate the $m-1$ probabilities $P_{\mathbf{x}}(A_k) := F_{\mathbf{x}}(y_k)$ where $A_k = \{y_1,\dots,y_k\}$, the mapping $F_{\mathbf{x}} : \mathscr{Y} \to [0,1]$ being equivalent to a discrete cumulative distribution. The probabilities $P_{\mathbf{x}}(\ell_{\mathbf{x}} = y_k)$ can then be deduced through the formula $P_{\mathbf{x}}(\{y_k\}) = F_{\mathbf{x}}(y_k) - F_{\mathbf{x}}(y_{k-1})$.

The same idea can be applied to sets of probabilities, in which case we estimate the bounds

$$\underline{P}_{\mathbf{x}}(A_k) := \underline{F}_{\mathbf{x}}(y_k) \text{ and } \overline{P}_{\mathbf{x}}(A_k) := \overline{F}_{\mathbf{x}}(y_k),$$

where $\underline{F}_{\mathbf{x}}, \overline{F}_{\mathbf{x}} : \mathscr{Y} \to [0,1]$ correspond to lower and upper cumulative distributions. These bounds induce a well-studied [15] probability set $\mathscr{P}_{\mathbf{x}}([\underline{F},\overline{F}])$. For $\mathscr{P}_{\mathbf{x}}([\underline{F},\overline{F}])$ to be properly defined, we need the two mappings $\underline{F}_{\mathbf{x}}, \overline{F}_{\mathbf{x}}$ to be increasing with $\underline{F}_{\mathbf{x}}(y_m) = \overline{F}_{\mathbf{x}}(y_m) = 1$ and to satisfy the inequality $\underline{F}_{\mathbf{x}} \leq \overline{F}_{\mathbf{x}}$. In practice, estimates $\underline{F}_{\mathbf{x}}, \overline{F}_{\mathbf{x}}$ obtained from data will always satisfy the latest inequality, however when using binary classifiers on each event A_k, nothing guarantees that they will be increasing, hence the potential need to correct the model. Algorithm 1 provides an easy way to obtain a well-defined probability set. In spirit, it is quite similar to the Frank et Hall estimates $P_{\mathbf{x}}(y_k) = \max\{0, F_{\mathbf{x}}(y_k) - F_{\mathbf{x}}(y_{k-1})\}$, where an implicit correction is performed to obtain well-defined probabilities in case $F_{\mathbf{x}}$ is not increasing.

Algorithm 1. Correction of estimates $\underline{F}_{\mathbf{x}}, \overline{F}_{\mathbf{x}}$ into proper estimates

```
Input: estimates $\underline{F}_{\mathbf{x}}, \overline{F}_{\mathbf{x}}$ obtained from data
Output: corrected estimates $\underline{F}_{\mathbf{x}}, \overline{F}_{\mathbf{x}}$
for $k=1,\dots,m-1$ do
    if $\overline{F}_{\mathbf{x}}(y_k) > \overline{F}_{\mathbf{x}}(y_{k+1})$ then $\overline{F}_{\mathbf{x}}(y_{k+1}) \leftarrow \overline{F}_{\mathbf{x}}(y_k)$;
    if $\underline{F}_{\mathbf{x}}(y_{m-k+1}) < \underline{F}_{\mathbf{x}}(y_{m-k})$ then $\underline{F}_{\mathbf{x}}(y_{m-k}) \leftarrow \underline{F}_{\mathbf{x}}(y_{m-k+1})$;
```

2.2 Nested Dichotomies

The principle of nested dichotomies is to form a tree structure using the class values $y_i \in \mathscr{Y}$. A nested dichotomy consists in recursively partitioning a tree node $C \subseteq \mathscr{Y}$ into two subsets A and B such that $A \cap B = \emptyset$ and $A \cup B = C$, until every leaf-node corresponds to a single class value ($card(C) = 1$). The root node is the whole set of classes $\mathscr{Y}$. To each branch A and B of a node C are associated conditional probabilities $P_{\mathbf{x}}(A|C) = 1 - P_{\mathbf{x}}(B|C)$. In the case of ordinal classifications, events C are of the kind $\{y_i, y_{i+1},\dots,y_j\}$ and their splits of the kind $A = \{y_i, y_{i+1},\dots,y_k\}$ and $B = \{y_{k+1},\dots,y_j\}$.

Generalizing the concept of nested dichotomies is pretty straightforward: it consists in allowing every local conditional probability to be imprecise, that is to each node C can be associated an interval $[\underline{P}_{\mathbf{x}}(A \mid C), \overline{P}_{\mathbf{x}}(A \mid C)]$, precise nested dichotomies being retrieved when $\underline{P}_{\mathbf{x}}(A \mid C) = \overline{P}_{\mathbf{x}}(A \mid C)$ for every node C. By duality of the imprecise probabilities [30, Sec.2.7.4.], we have $\underline{P}_{\mathbf{x}}(A \mid C) = 1 - \overline{P}_{\mathbf{x}}(B \mid C)$ and $\overline{P}_{\mathbf{x}}(A \mid C) = 1 - \underline{P}_{\mathbf{x}}(B \mid C)$. Such an imprecise nested dichotomy is then associated to a set $\mathscr{P}_{\mathbf{x}}$ of joint probabilities, obtained by considering all precise selection $P_{\mathbf{x}}(A \mid C) \in [\underline{P}_{\mathbf{x}}(A \mid C), \overline{P}_{\mathbf{x}}(A \mid C)]$ for each node C. Figure 1 shows examples of a precise and an imprecise nested dichotomy tree when $\mathscr{Y} = \{y_1, y_2, y_3\}$.

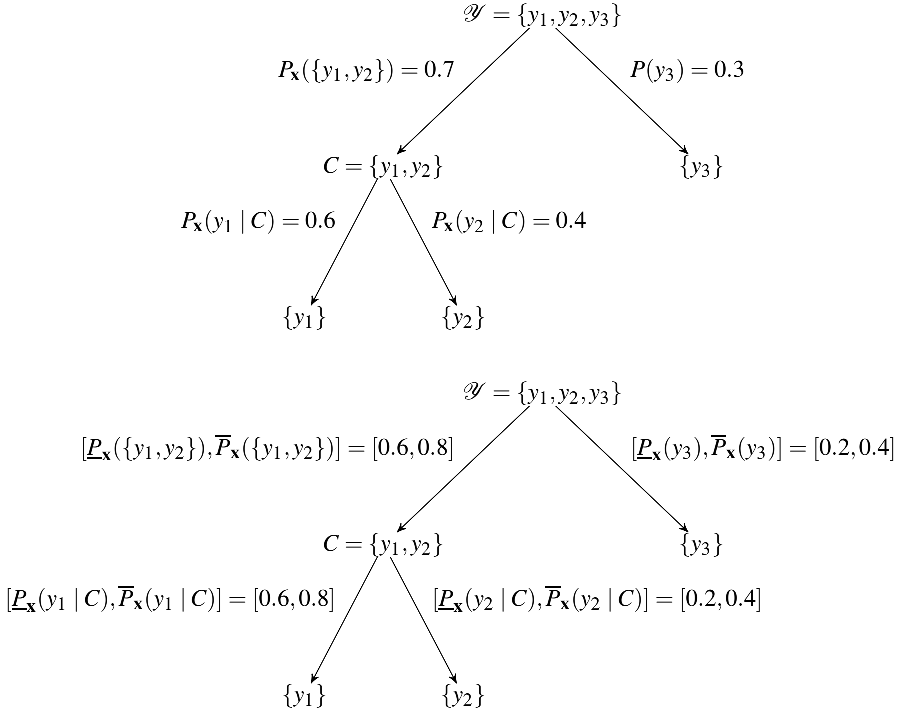
$\mathscr{Y} = \{y_1, y_2, y_3\}$
$P_{\mathbf{x}}(\{y_1, y_2\}) = 0.7$
$P(y_3) = 0.3$
$C = \{y_1, y_2\}$
$\{y_3\}$
$P_{\mathbf{x}}(y_1 \mid C) = 0.6$
$P_{\mathbf{x}}(y_2 \mid C) = 0.4$
$\{y_1\}$
$\{y_2\}$
$\mathscr{Y} = \{y_1, y_2, y_3\}$
$[\underline{P}_{\mathbf{x}}(\{y_1, y_2\}), \overline{P}_{\mathbf{x}}(\{y_1, y_2\})] = [0.6, 0.8]$
$[\underline{P}_{\mathbf{x}}(y_3), \overline{P}_{\mathbf{x}}(y_3)] = [0.2, 0.4]$
$C = \{y_1, y_2\}$
$\{y_3\}$
$[\underline{P}_{\mathbf{x}}(y_1 \mid C), \overline{P}_{\mathbf{x}}(y_1 \mid C)] = [0.6, 0.8]$
$[\underline{P}_{\mathbf{x}}(y_2 \mid C), \overline{P}_{\mathbf{x}}(y_2 \mid C)] = [0.2, 0.4]$
$\{y_1\}$
$\{y_2\}$

Fig. 1. Precise (above) and imprecise (below) nested dichotomies

3 Inferences

In this section, we expose how inferences (decision making) can be done with our two decompositions, both with general costs and 0/1 costs. While other costs such as the absolute error cost are also natural in an ordinal classification setting [14], we chose to focus on the 0/1 cost, as it is the only one for which a theoretically sound way to compare determinate and indeterminate classifiers, i.e., classifiers returning respectively precise and (potentially) imprecise classification, has been provided [33].

We will first recall the basic of decision making with probabilities and will then present their extensions when considering sets of probabilities. Let us denote by $c_k : \mathscr{Y} \to \mathbb{R}$ the cost (loss) function associated to y_k, that is $c_k(y_j)$ is the cost of predicting y_k when y_j is true. In the case where precise estimates $P_{\mathbf{x}}(y_k)$ are obtained from the learning algorithm, obtaining the optimal prediction is

$$\hat{y} = \arg\min_{y_k \in \mathscr{Y}} \mathbb{E}_{\mathbf{x}}(c_k)$$

with $\mathbb{E}_{\mathbf{x}}$ the expectation of c_k under $P_{\mathbf{x}}$, i.e. we predict the value having the minimal expected cost.

In practice, this also comes down to build a preference relation $\succ_P$ on elements of $\mathscr{Y}$, where $y_l \succ_{P_{\mathbf{x}}} y_k$ iff $\mathbb{E}_{\mathbf{x}}(c_k) > \mathbb{E}_{\mathbf{x}}(c_l)$ or equivalently $\mathbb{E}_{\mathbf{x}}(c_k - c_l) > 0$, that is the expected cost of predicting y_l is lower than the expected cost of predicting y_k. When working with a set $\mathscr{P}_{\mathbf{x}}$ of probabilities, this can be extended by building a partial order $\succ_{\underline{\mathbb{E}}_{\mathbf{x}}}$ on elements of $\mathscr{Y}$ such that $y_l \succ_{\underline{\mathbb{E}}_{\mathbf{x}}} y_k$ iff $\underline{\mathbb{E}}_{\mathbf{x}}(c_k - c_l) > 0$ with

$$\underline{\mathbb{E}}_{\mathbf{x}}(c_k - c_l) = \inf_{P_{\mathbf{x}} \in \mathscr{P}_{\mathbf{x}}} \mathbb{E}_{\mathbf{x}}(c_k - c_l).$$

That is, we are sure that the cost of exchanging y_k with y_l will have a positive expectation (hence y_l is preferred to y_k). The final cautious prediction $\hat{Y}$ is then obtained by taking the maximal elements of the partial order $\succ_{\underline{\mathbb{E}}_{\mathbf{x}}}$, that is

$$\hat{Y} = \{y \in \mathscr{Y} : \nexists y' \neq y \text{ s.t. } y' \succ_{\underline{\mathbb{E}}_{\mathbf{x}}} y\}$$

and is known under the name maximality criterion [30,29]. In practice, getting $\hat{Y}$ requires at worst a number ${}^{m(m-1)}/_{2}$ of computations that is quadratic in the number of classes. A conservative approximation (in the sense that the obtained set of non-dominated classes includes $\hat{Y}$) can be obtained by using the notion of interval dominance [29], in which $y_l \succ_{\underline{\mathbb{E}}_{\mathbf{x}}} y_k$ if $\underline{\mathbb{E}}_{\mathbf{x}}(c_k) > -\underline{\mathbb{E}}_{\mathbf{x}}(-c_l)$, thus requiring only $2m$ computations at worst to compare all classes, yet as m is typically low in ordinal classification, we will only consider maximality here.

In particular, 0/1 costs are defined as $c_k(y_j) = 1$ if $j \neq k$ and 0 else. If we note $\mathbf{1}_{(A)}$ the indicator function of A ($\mathbf{1}_{(A)}\ (x) = 1$ if $x \in A$, 0 else), then $(c_k - c_l) = \mathbf{1}_{(y_l)}\ - \mathbf{1}_{(y_k)}$ as $c_k(y_j) - c_l(y_j) = -1$ if $j = k$, 1 if $j = l$ and 0 if $j \neq k, l$. Hence we have $y_l \succ_{\underline{\mathbb{E}}_{\mathbf{x}}} y_k$ iff $\underline{\mathbb{E}}_{\mathbf{x}}(\mathbf{1}_{(y_l)}\ - \mathbf{1}_{(y_k)}\ ) > 0$. Table 1 provides an example of the functions over which lower expectations must be computed for $0/1$ losses in the case $\mathscr{Y} = \{y_1, \dots, y_5\}$.

Table 1. 0/1 cost functions comparing y_2 and y_4

	y_1	y_2	y_3	y_4	y_5
c_2	1	0	1	1	1
c_4	1	1	1	0	1
$c_2 - c_4$	0	-1	0	1	0

3.1 Inference with Imprecise Cumulative Distributions

If the probability set $\mathscr{P}_{\mathbf{x}}([\underline{F}, \overline{F}])$ is induced by the bounding cumulative distributions $[\underline{F}_{\mathbf{x}}, \overline{F}_{\mathbf{x}}]$, then it can be shown[1] that the lower expectation of any function f over $\mathscr{Y}$ can be computed through the Choquet Integral: if we denote by $()$ a reordering of elements of $\mathscr{Y}$ such that $f(y_{(1)}) \leq \ldots \leq f(y_{(m)})$, this integral reads

$$\underline{\mathbb{E}}_{\mathbf{x}}(f) = \sum_{i=1}^{m} (f(y_{(i)}) - f(y_{(i-1)})\underline{P}_{\mathbf{x}}(A_{(i)})) \tag{1}$$

[1] For details, interested readers are referred to [15]. Shortly speaking, this is due to the supermodularity of the induced lower probability.

with $f(y_{(0)}) = 0, A_{(i)} = \{y_{(i)}, \dots, y_{(m)}\}$ and $\underline{P}_{\mathbf{x}}(A_{(i)}) = \inf_{P_{\mathbf{x}} \in \mathscr{P}_{\mathbf{x}}([\underline{F},\overline{F}])} P(A_{(i)})$ is the lower probability of $A_{(i)}$. In the case of imprecise cumulative distributions, the lower probability of an event A can be easily obtained: let $C = [y_{\underline{j}}, y_{\overline{j}}]$ denote a discrete interval of $\mathscr{Y}$ such that $[y_{\underline{j}}, y_{\overline{j}}] = \{y_i \in \mathscr{Y} : \underline{j} \leq i \leq \overline{j}\}$, then $\underline{P}_{\mathbf{x}}(C) = \max\{0, \underline{F}_{\mathbf{x}}(y_{\overline{j}}) - \overline{F}_{\mathbf{x}}(y_{\underline{j}-1})\}$ with $\underline{F}_{\mathbf{x}}(y_0) = \overline{F}_{\mathbf{x}}(y_0) = 0$. Any event A can then be expressed as a union of disjoint intervals[2] $A = C_1 \cup \dots \cup C_M$, and we have [15] $\underline{P}_{\mathbf{x}}(A) = \sum_{i=1}^{M} \underline{P}_{\mathbf{x}}(C_i)$.

Table 2. Imprecise cumulative distribution

	y_1	y_2	y_3	y_4	y_5
$\overline{F}_{\mathbf{x}}$	0.15	0.5	0.55	0.95	1
$\underline{F}_{\mathbf{x}}$	0.1	0.4	0.5	0.8	1

Example 1. Consider the imprecise cumulative distributions defined by Table 2 together with a 0/1 loss and the function $c_2 - c_4$ of Table 1. The elements used in the computation of the Choquet integral (1) for this case are summarized in Table 3.

Table 3. Choquet integral components of Example 1

i	$y_{(i)}$	$f_{(i)}$	$A_{(i)}$	$\underline{P}_{\mathbf{x}}(A_{(i)})$
1	y_2	-1	$\mathscr{Y}$	1
2	y_1	0	$\{y_1, y_3, y_4, y_5\}$	0.6
3	y_3	0	$\{y_3, y_4, y_5\}$	0.5
4	y_5	0	$\{y_4, y_5\}$	0.45
5	y_4	1	$\{y_4\}$	0.25

The lower probability of $A_{(2)} = \{y_1, y_3, y_4, y_5\} = \{y_1\} \cup \{y_3, y_4, y_5\}$ is

$$\begin{aligned}\underline{P}_{\mathbf{x}}(A_{(2)}) &= \underline{P}_{\mathbf{x}}(\{y_1\}) + \underline{P}_{\mathbf{x}}(\{y_3, y_4, y_5\}) \\ &= \max\{0, \underline{F}_{\mathbf{x}}(y_1) - \overline{F}_{\mathbf{x}}(y_0)\} + \max\{0, \underline{F}_{\mathbf{x}}(y_5) - \overline{F}_{\mathbf{x}}(y_2)\} \\ &= 0.1 + 0.5,\end{aligned}$$

and the final value of the lower expectation $\underline{\mathbb{E}}_{\mathbf{x}}(c_2 - c_4) = -0.15$, meaning that y_4 is not preferred to y_2 in this case. As we also have $\underline{\mathbb{E}}_{\mathbf{x}}(c_4 - c_2) = -0.2$, y_2 and y_4 are incomparable under a 0/1 loss and given the bounding distributions $\underline{F}_{\mathbf{x}}, \overline{F}_{\mathbf{x}}$. Actually, our cautious prediction would be $\hat{Y} = \{y_2, y_4\}$, as we have $y_i \succ_{\underline{\mathbb{E}}} y_j$ for any $i \in 2,4$ and $j \in 1,3,5$.

[2] Two intervals $[y_{\underline{j}}, y_{\overline{j}}], [y_{\underline{k}}, y_{\overline{k}}]$ are said disjoint if $\overline{j} + 1 < \underline{k}$.

3.2 Inference with Nested Dichotomies

In the precise case, computations of expectations with nested dichotomies can be done by backward recursion and local computations (simply applying the law of iterated expectation). That is the global expectation $\mathbb{E}_{\mathbf{x}}(f)$ of a function $f : \mathscr{Y} \to \mathbb{R}$ can be done by computing local expectations for each node, starting from the tree leaves taking values $f(y)$. This provides nested dichotomies with a computationally efficient method to estimate expectations.

It has been shown [13] that the same recursive method can be applied to imprecise nested dichotomies. Assume we have a split $\{A, B\}$ of a node C, and a real-valued (cost) function $f : \{A,B\} \to \mathbb{R}$ defined on $\{A,B\}$. We can compute the (local) lower expectation associated with the node C by :

$$\underline{\mathbb{E}}_{\mathbf{x},C}(f) = \min \left\{ \begin{array}{l} \underline{P}_{\mathbf{x}}(A \mid C)f(A) + \overline{P}_{\mathbf{x}}(B \mid C)f(B), \\ \overline{P}_{\mathbf{x}}(A \mid C)f(A) + \underline{P}_{\mathbf{x}}(B \mid C)f(B) \end{array} \right\} \tag{2}$$

Starting from a function such as the one given in Table 1, we can then go from the leaves to the root of the imprecise nested dichotomy to obtain the associated lower expectation.

Example 2. Consider a problem where we have $\mathscr{Y} = \{y_1, y_2, y_3\}$ and the same imprecise dichotomy as in Figure 1. Figure 2 shows the local computations performed to obtain the lower expectation of $c_1 - c_3$. For instance, using Eq. (2) on node $C = \{y_1, y_2\}$, we get

$$\underline{\mathbb{E}}_{\mathbf{x},\{y_1,y_2\}}(c_1 - c_3) = \min\{-1 \cdot 0.8 + 0 \cdot 0.2, -1 \cdot 0.6 + 0 \cdot 0.4\}$$

We finally obtain $\underline{\mathbb{E}}_{\mathbf{x},\mathscr{Y}}(c_1 - c_3) = -0.44$, concluding that y_3 is not preferred to y_1. As the value $\underline{\mathbb{E}}_{\mathbf{x},\mathscr{Y}}(c_3 - c_1) = -0.04$ is also negative, we can conclude that y_1 and y_3 are not comparable. Yet we do have $\underline{\mathbb{E}}_{\mathbf{x},\mathscr{Y}}(c_2 - c_1) > 0$, meaning that y_1 is preferred to y_2, hence $\hat{Y} = \{y_1, y_3\}$.

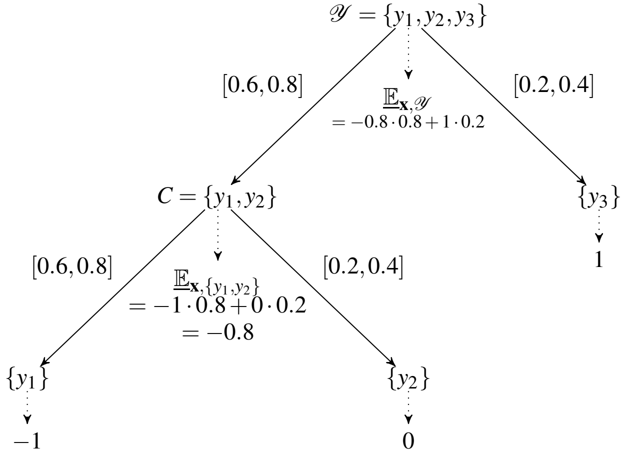
$\mathscr{Y} = \{y_1, y_2, y_3\}$
[0.6,0.8]
$\underline{\mathbb{E}}_{\mathbf{x},\mathscr{Y}}$
$= -0.8 \cdot 0.8 + 1 \cdot 0.2$
[0.2,0.4]
$C = \{y_1, y_2\}$
$\{y_3\}$
1
[0.6,0.8]
$\underline{\mathbb{E}}_{\mathbf{x},\{y_1,y_2\}}$
$= -1 \cdot 0.8 + 0 \cdot 0.2$
$= -0.8$
[0.2,0.4]
$\{y_1\}$
-1
$\{y_2\}$
0

Fig. 2. Expectation computation for $c_1 - c_3$

4 Experimentations

This section presents the experiments we achieved to compare decomposition methods providing determinate predictions and their imprecise counterpart delivering possibly indeterminate predictions.

4.1 Learning Method

In our experiments, we consider a base classifier which can be extended easily to output interval-valued probabilities, so that we can evaluate the impact of allowing for cautiousness in ordinal classification. For this reason, we use the Naive Bayesian Classifier (NBC) which has an extension in imprecise probabilities : the Naive Credal Classifier (NCC) [32].

The NCC preserves the main properties of NBC, such as the assumption of attribute independence conditional on the class. In binary problems where we have to differentiate between two complementary events A and B, NBC reads

$$P(A|x^1,\dots,x^p) = \frac{P(A)\prod_{i=1}^{p} P(x^i \mid A)}{\prod_{i=1}^{p} P(x^i \mid A)P(A) + \prod_{i=1}^{p} P(x^i \mid B)P(B)}, \tag{3}$$

where $(x_1,\dots,x_p)$ are the feature variables and A,B are the two events whose probability we have to estimate. The NCC consists in using probability bounds in Eq. 3, getting

$$\underline{P}(A|x^1,\dots,x^p) = \min\left\{\begin{array}{c} \frac{\underline{P}(A)\prod_{i=1}^{p}\underline{P}(x^i|A)}{\prod_{i=1}^{p}\underline{P}(x^i|A)\underline{P}(A)+\prod_{i=1}^{p}\underline{P}(x^i|B)\overline{P}(B)}, \\ \\ \frac{\overline{P}(A)\prod_{i=1}^{p}\underline{P}(x^i|A)}{\prod_{i=1}^{p}\underline{P}(x^i|A)\overline{P}(A)+\prod_{i=1}^{p}\underline{P}(x^i|B)\underline{P}(B)} \end{array}\right\} = 1-\overline{P}(B|x^1,\dots,x^p). \tag{4}$$

and $\underline{P}(B|x^1,\dots,x^p) = 1-\overline{P}(A|x^1,\dots,x^p)$ can be obtained in the same way. Using the Imprecise Dirichlet Model (IDM) [4], we can compute these probability estimates from the training data by simply counting occurrences :

$$\underline{P}(x^i \mid A) = \frac{occ_{i,A}}{occ_A + s}, \quad \overline{P}(x^i \mid A) = \frac{occ_{i,A} + s}{occ_A + s} \tag{5}$$

and

$$\underline{P}(A) = \frac{occ_{i,A}}{n_{A,B} + s}, \quad \overline{P}(A) = \frac{occ_{i,A} + s}{n_{A,B} + s} \tag{6}$$

where $occ_{i,A}$ is the number of instances in the training set where the attribute $\mathscr{X}^i$ is equal to x^i and the class value is in A, occ_A the number of instances in the training set where the class value is in A, $n_{A,B}$ is the number of training sample whose class is either in A or B. The hyper-parameter s that sets the imprecision level of the IDM is usually equal to 1 or 2 [31].

4.2 Evaluation

Comparing classifiers that return cautious (partial) predictions in the form of multiple classes is an hard problem. Indeed, compared to the usual setting, measures of performance have to include the informativeness of the predictions in addition to the accuracy. Zaffalon et al. [33] discuss in details the case of comparing a cautious prediction with a classical one under a 0/1 loss assumption, using a betting interpretation. They show that the discounted accuracy, which rewards a cautious prediction Y class with ${}^{1}/_{|Y|}$ if the true class is in Y, and zero otherwise, is a measure satisfying a number of appealing properties. However, they also show that discounted accuracy makes no difference between a cautious classifier providing indeterminate predictions and a random classifier: for instance, in a binary setting, a cautious classifier always returning both classes would have the same value as a classifier picking the class at random, yet the determinate classifier displays a lower variance (it always receives ${}^{1}/_{2}$ as reward, while the random one would receive a reward of 1 half of the time, and 0 the other half).

This is why a decision maker that wants to value cautiousness should consider modifying discounted accuracy by a risk-adverse utility function [33]. Here, we consider the u_{65} function: Let $(\mathbf{x}_i, \ell_i)$, $i = 1, \ldots, n$ be the set of test data and Y_i our (possibly imprecise) predictions, then u_{65} is

$$u_{65} = \frac{1}{n} \sum_{i=1}^{n} -0.6 d_i^2 + 1.6 d_i,$$

where $d_i = \mathbf{1}_{(Y_i)}(\ell_i)/|Y_i|$ is the discounted accuracy. It has been shown in [33] that this approach is consistent with the use of F_1 measures [12,1] as a way to measure the

Table 4. Data set details

Name	#instances	#features	#classes
autoPrice	159	16	5
bank8FM	8192	9	5
bank32NH	8192	33	5
boston housing	506	14	5
california housing	20640	9	5
cpu small	8192	13	5
delta ailerons	7129	6	5
elevators	16599	19	5
delta elevators	9517	7	5
friedman	40768	11	5
house 8L	22784	9	5
house 16H	22784	17	5
kinematics	8192	9	5
puma8NH	8192	9	5
puma32H	8192	33	5
stock	950	10	5
ERA	1000	5	9
ESL	488	5	9
LEV	1000	5	5

quality of indeterminate classifications. In fact, it is shown in [33] that u_{65} is less in favor of indeterminate classifiers than the use of F_1 measure.

4.3 Results

In this section, our method is tested on 19 datasets of the UCI machine learning repository [16], whose details are given in Table 4. As there is a general lack of benchmark data sets for ordinal classification data, we used regression problems that we turned into ordinal classification by discretizing the output variable, except for the data sets LEV that has 5 ordered classes and ESL, ERA that have 9 ordered classes. The results reported in this section are obtained with a discretization into five classes of equal frequencies. We also performed experiments on the other data sets, using 7 and 9 discretized classes, obtaining the same conclusions.

The results in this section are obtained from a 10-fold cross validation. To build the dichotomy trees, we selected at each node the split $A = \{y_i, y_{i+1}, \dots, y_k\}$ and $B = \{y_{k+1}, \dots, y_j\}$ of $C = \{y_i, y_{i+1}, \dots, y_j\}$ that maximised the u_{65} measure on the binarized data set. We use ordinal logistic regression (logreg) as a base line classifier to compare our results. For each decomposition method, Frank & Hall (F) and nested dichotomies (N), we compared the naive Bayes classifier (B) with its indeterminate counterpart (NCC), picking an hyper-parameter $s = 2$ for the IDM in Eqs. (5)- (6). The naive Bayes classifier was used in a classical way to provide determinate predictions

Table 5. u_{65} Results (and method rank) obtained on the different methods. Log= logistic regression, B = Naive Bayes classifier, C = Naive credal classifier, A= Alonso *et al.* prediction method, F = Frank & Hall, N = Nested Dichotomies.

	Log	B/F	B/F/A	C/F	B/N	B/N/A	C/N
autoPrice	52.2 (5)	58.5 (3)	39.7 (7)	53.8 (4)	59.1 (1)	51.3 (6)	58.6 (2)
bank8FM	68.2 (2)	67.4 (3)	37.3 (7)	68.3 (1)	63.9 (5)	54.9 (6)	64.8 (4)
bank32NH	43.3 (4)	43.6 (3)	30.2 (7)	47.8 (1)	42.9 (5)	40.2 (6)	46.7 (2)
boston hous.	55.6 (4)	55.1 (5)	34.1 (7)	55.8 (3)	56.1 (2)	43.9 (6)	57.4 (1)
california hous.	47.6 (5)	48.2 (4)	32.9 (7)	48.6 (2)	48.3 (3)	43.5 (6)	48.7 (1)
cpu small	58.8 (3)	57 (5)	40.9 (7)	57.1 (4)	60.8 (2)	54.1 (6)	61.1 (1)
delta ail.	50.2 (6)	53.5 (4)	31.8 (7)	53.8 (3)	54.2 (2)	52.1 (5)	54.9 (1)
elevators	42.7 (2)	39.0 (5)	30.5 (7)	39.2 (4)	42.6 (3)	37.9 (6)	42.9 (1)
delta elev.	46.5 (6)	49.9 (5)	34.3 (7)	50.4 (4)	50.8 (3)	53.2 (1)	51.2 (2)
friedman	53.2 (5)	63.8 (2)	32 (7)	64.5 (1)	62.2 (4)	47.3 (6)	63 (3)
house 8L	39.9 (6)	49.6 (2)	34.9 (7)	49.8 (1)	49.4 (4)	43.9 (5)	49.6 (3)
house 16H	41.4 (6)	47.5 (4)	35.3 (7)	47.6 (3)	50.0 (2)	43.9 (5)	50.2 (1)
kinematics	37.7 (5)	44.9 (3)	28.8 (7)	46.2 (1)	44.4 (4)	37.5 (6)	45.4 (2)
puma8NH	30.3 (6)	46.5 (4)	29.7 (7)	47.6 (3)	47.7 (2)	42.9 (5)	48.3 (1)
puma32H	30.5 (6)	48.6 (3)	29.7 (7)	50.9 (1)	47.7 (4)	40.6 (5)	49.9 (2)
stock	61.2 (6)	72.4 (3)	41.7 (7)	71.5 (4)	75.1 (1)	61.2 (5)	74.2 (2)
ERA	23.2 (5)	23.2 (4)	14.1 (7)	28.5 (1)	22.5 (6)	26.8 (2)	26.6 (3)
ESL	12.7 (7)	55.7 (4)	28.1 (6)	53.4 (5)	57 (2)	63 (1)	56.5 (3)
LEV	46.3 (6)	60.5 (2)	44.9 (7)	60.4 (3)	59.8 (4)	61.6 (1)	59.6 (5)
Avg. rank	5	3.6	6.9	2.6	3.1	4.7	2.1

and (B/A) with the F_1 measure of Alonso *et al.* [1] to produce indeterminate predictions (details about this latter method can be found in [1]).

Table 5 show the obtained results in terms of u_{65} (that reduces to classical accuracy for the three determinate methods) as well as the rank of each classifier. Using Demsar's approach by applying the Friedman statistic on the ranks of algorithm performance for each dataset, we obtain a value of 68.16 for the Chi Square, and a 26.8 statistic for the F-distribution. Since the statistic is 1.7 for a p-value of 0.05, we can safely reject the null hypothesis, meaning that the performances of the classifiers are significatively different. This shows that in average the introduced indeterminacy (or cautiousness) in the predictions is not too important and is compensated by more reliable predictions. We use Nemenyi test as a post-hoc test, and obtain that two classifiers are significantly different (with p-value 0.05) if the difference between their mean rank is higher than 2.06.

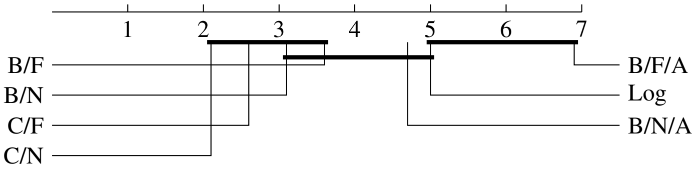

Fig. 3. Post-hoc test results on algorithms. Thick lines links non-significantly different algorithms.

Figure 3 summarises the average ranks of the different methods and shows which one are significatively different from the others. We can see that, although techniques using probability sets (C/N and C/F) have the best average rank, they are not significantly different from their determinate Bayesian counterpart (B/N and B/F) under u_{65} measure. This is not surprising, since the goal of such classifiers is not to outperform Bayesian methods, but to provide more reliable predictions when not enough information is available. It should also be recalled that the u_{65} measure is only slightly favourable to indeterminate classifiers, and that other measures such as F_1 and u_{80} would have given better scores to indeterminate classifiers.

An interesting result is that Alonso *et al.* [1] method, that use a precise probabilistic models and produce indeterminate predictions through the use of specific cost functions (the F_1 measure in our case), performs quite poorly, in particular when applied with the Frank and Hall decomposition (B/F/A). This can be explained by the fact that Alonso *et al.* [1] method will mainly produce indeterminate classifications when the labels having the highest probabilities will have close probability values, i.e., when there will be some ambiguity as to the modal label. However, it is well known that the naive Bayes classifier tends to overestimate model probabilities, therefore acting as a good classifier for 0/1 loss functions, but as a not so good probability density estimator. This latter feature can clearly be counter-productive when using Alonso *et al.* [1] method, that relies on having good probability estimates. On the other hand, indeterminate classification using probability sets can identify situations where information is lacking, even if the underlying estimator is poor. Our results indicate that, while the two methods both produce indeterminate classifications, they do so in very different ways (and therefore present different interests).

Table 6 shows the mean imprecision of indeterminate predictions for all the methods producing such predictions. This sheds additional light on the bad performances of the B/F/A method, which tends to produce rather imprecise predictions without necessarily counterbalancing them with an higher reliability or accuracy. For the other methods, the mean imprecision is comparable.

Table 6. Mean imprecision of predictions (rank)

	B/F/A	C/F	B/N/A	C/N
autoPrice	2.22 (3)	2.25 (4)	1.03 (1)	1.93 (2)
bank8FM	2.06 (4)	1.06 (1)	1.55 (3)	1.08 (2)
bank32NH	2.11 (4)	1.78 (2)	2.01 (3)	1.72 (1)
boston housing	2.23 (4)	1.36 (2)	1.11 (1)	1.51 (3)
california housing	2.17 (4)	1.04 (2)	1.6 (3)	1.04 (1)
cpu small	2.38 (4)	1.03 (1)	1.2 (3)	1.04 (2)
delta ailerons	2.54 (4)	1.03 (1)	1.62 (3)	1.06 (2)
elevators	2.47 (4)	1.03 (1)	1.39 (3)	1.04 (2)
delta elevators	2.47 (4)	1.05 (2)	1.63 (3)	1.04 (1)
friedman	2.06 (3)	1.06 (1)	2.17 (4)	1.06 (2)
house 8L	2.24 (4)	1.01 (1)	1.43 (3)	1.02 (2)
house 16H	2.28 (4)	1.02 (1)	1.25 (3)	1.03 (2)
kinematics	2.12 (3)	1.21 (2)	2.36 (4)	1.2 (1)
puma8NH	2.16 (4)	1.12 (2)	1.89 (3)	1.1 (1)
puma32H	2.47 (4)	1.43 (1)	1.91 (3)	1.5 (2)
stock	2.21 (4)	1.15 (3)	1.04 (1)	1.14 (2)
ERA	4.02 (4)	2.81 (3)	2.24 (1)	2.32 (2)
ESL	3.62 (4)	2.27 (3)	1.39 (1)	1.84 (2)
LEV	2.05 (4)	1.18 (2)	1.52 (3)	1.12 (1)

Figures 4 displays the non-discounted accuracy (that is, we count 1 each time the true class is in the prediction, whether its determinate or not) on those instances where the use of NCC returned an indeterminate classification. On those instances, the accuracy

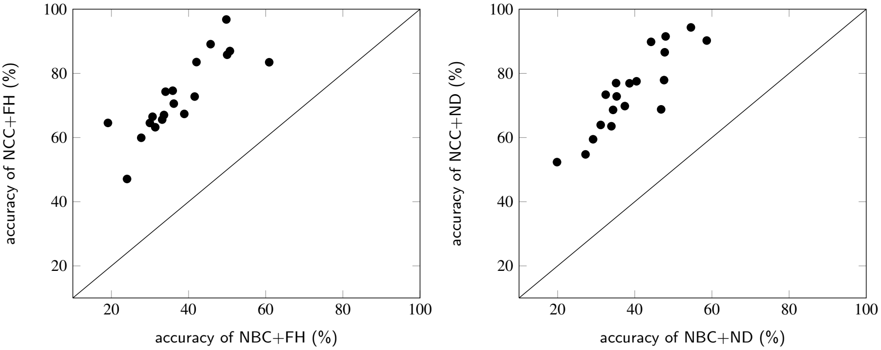

Fig. 4. Non-discounted accuracy of the NBC vs NCC methods for both decompositions on indeterminate instances

of the determinate version (NBC) is on average 10 % lower than the accuracy displayed in Table 5. In contrast, the non-discounted accuracy of the indeterminate version on these instances is much higher, meaning that the indeterminacy actually concerns hard-to-classify instances.

5 Conclusions

In this paper, we have proposed two methods to learn cautious ordinal classifiers, in the sense that they provide indeterminate predictions when information is insufficient to provide a reliable determinate one. More precisely, these methods extend two well-known binary decomposition methods previously used for ordinal classification, namely Frank & Hall decomposition and nested dichotomies. The extension consists in allowing one to provide interval-valued probabilistic estimates rather than precise ones for each binary problem, the width of the interval reflecting our lack of knowledge about the instances.

Our experiments on different data sets show that allowing for cautiousness in ordinal classification methods can increase the reliability of the prediction, while not providing too indeterminate predictions. More specifically, indeterminacy tends to focus on those instances that are hard to classify for determinate classifiers. We could probably improve both the efficiency of inferences, e.g., by studying extensions of labelling trees to imprecise trees [3], or their accuracy by using more complex classifiers, e.g., credal averaging techniques [11]. Yet, as the number m of labels in ordinal classification is usually small, and as the advantages of using binary decompositions are usually lower when using complex estimation methods, the benefits of such extensions would be limited.

In these experiments, we have focused on the 0/1 loss and its extensions to indeterminate classification u_{65}, which is more favourable to determinate classifier than the F_1 measure proposed by Alonso et al. [1].The reason for this is that 0/1 loss is the only one to which the results of Zaffalon *et al.* [33] that allows to compare determinate and indeterminate classifiers apply. Yet, our approaches can easily handle generic losses (in contrast with the multi-class naive credal classifier [32]), as shows Section 3 and Eqs (1)- (2). Also, there are loss functions such as the absolute error that are at least as natural to use in an ordinal classification problem as the 0/1 loss function. Our future efforts will therefore focus on determining meaningful ways to compare cost-sensitive determinate and indeterminate classifiers. Another drawback of using 0/1 loss function [24], shown by Examples 1 and 2, is that we may obtain indeterminate predictions containing non-consecutive labels. We expect that considering other losses such as L_1 loss could solve this issue.

In addition to that, we plan to apply the methods developed in this paper to more complex problems that can be reduced as a set of ordinal classification problems, such as graded multi-label [7] or label ranking [6].

Acknowledgements. This work was carried out and funded by the French National Research Agency, through the project ANR-13-JS03-0007 RECIF. It was also supported by the French Government, through the LABEX MS2T and the program "Investments for the future" managed by the National Agency for Research (Reference ANR-11-IDEX-0004-02).

References

1. Alonso, J., del Coz, J.J., Díez, J., Luaces, O., Bahamonde, A.: Learning to predict one or more ranks in ordinal regression tasks. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part I. LNCS (LNAI), vol. 5211, pp. 39–54. Springer, Heidelberg (2008)
2. Bartlett, P., Wegkamp, M.: Classification with a reject option using a hinge loss. The Journal of Machine Learning Research 9, 1823–1840 (2008)
3. Bengio, S., Weston, J., Grangier, D.: Label embedding trees for large multi-class tasks. In: NIPS, vol. 23, p. 3 (2010)
4. Bernard, J.: An introduction to the imprecise dirichlet model for multinomial data. International Journal of Approximate Reasoning 39(2), 123–150 (2005)
5. Cheng, W., Hüllermeier, E., Waegeman, W., Welker, V.: Label ranking with partial abstention based on thresholded probabilistic models. In: Advances in Neural Information Processing Systems 25 (NIPS 2012), pp. 2510–2518 (2012)
6. Cheng, W., Hüllermeier, E.: A nearest neighbor approach to label ranking based on generalized labelwise loss minimization
7. Cheng, W., Hüllermeier, E., Dembczynski, K.J.: Graded multilabel classification: The ordinal case. In: Proceedings of the 27th International Conference on Machine Learning (ICML 2010), pp. 223–230 (2010)
8. Chow, C.: On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory 16(1), 41–46 (1970)
9. Chu, W., Keerthi, S.S.: Support vector ordinal regression. Neural Computation 19(3), 792–815 (2007)
10. Corani, G., Antonucci, A., Zaffalon, M.: Bayesian networks with imprecise probabilities: Theory and application to classification. In: Holmes, D.E., Jain, L.C. (eds.) Data Mining: Foundations and Intelligent Paradigms. ISRL, vol. 23, pp. 49–93. Springer, Heidelberg (2012)
11. Corani, G., Zaffalon, M.: Credal model averaging: an extension of bayesian model averaging to imprecise probabilities. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part I. LNCS (LNAI), vol. 5211, pp. 257–271. Springer, Heidelberg (2008)
12. José del Coz, J., Bahamonde, A.: Learning nondeterministic classifiers. The Journal of Machine Learning Research 10, 2273–2293 (2009)
13. De Cooman, G., Hermans, F.: Imprecise probability trees: Bridging two theories of imprecise probability, vol. 172, pp. 1400–1427
14. Dembczyński, K., Kotłowski, W., Słowiński, R.: Learning rule ensembles for ordinal classification with monotonicity constraints. Fundamenta Informaticae 94(2), 163–178 (2009)
15. Destercke, S., Dubois, D., Chojnacki, E.: Unifying practical uncertainty representations - i: Generalized p-boxes. Int. J. Approx. Reasoning 49(3), 649–663 (2008)
16. Frank, A., Asuncion, A.: UCI machine learning repository (2010), `http://archive.ics.uci.edu/ml`
17. Frank, E., Kramer, S.: Ensembles of nested dichotomies for multi-class problems. In: ICML 2004, p. 39 (2004)
18. Frank, E., Hall, M.: A simple approach to ordinal classification. In: Flach, P.A., De Raedt, L. (eds.) ECML 2001. LNCS (LNAI), vol. 2167, pp. 145–156. Springer, Heidelberg (2001)
19. Frank, E., Kramer, S.: Ensembles of nested dichotomies for multi-class problems. In: Proceedings of the Twenty-first International Conference on Machine Learning, p. 39. ACM (2004)
20. Fürnkranz, J., Hüllermeier, E., Vanderlooy, S.: Binary decomposition methods for multipartite ranking. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009, Part I. LNCS (LNAI), vol. 5781, pp. 359–374. Springer, Heidelberg (2009)

21. Ha, T.M.: The optimum class-selective rejection rule. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(6), 608–615 (1997)
22. Herbrich, R., Graepel, T., Obermayer, K.: Large margin rank boundaries for ordinal regression. In: Advances in Neural Information Processing Systems, pp. 115–132 (1999)
23. Joachims, T.: Training linear svms in linear time. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 217–226. ACM (2006)
24. Kotlowski, W., Slowinski, R.: On nonparametric ordinal classification with monotonicity constraints. IEEE Trans. Knowl. Data Eng. 25(11), 2576–2589 (2013)
25. Li, L., Lin, H.T.: Ordinal regression by extended binary classification. In: Advances in Neural Information Processing Systems, pp. 865–872 (2006)
26. Shashua, A., Levin, A.: Ranking with large margin principle: Two approaches. In: Advances in Neural Information Processing Systems, pp. 937–944 (2002)
27. Sousa, R., Yevseyeva, I., da Costa, J.F.P., Cardoso, J.S.: Multicriteria models for learning ordinal data: A literature review. In: Yang, X.-S. (ed.) Artificial Intelligence, Evolutionary Computing and Metaheuristics. SCI, vol. 427, pp. 109–138. Springer, Heidelberg (2013)
28. Tehrani, A.F., Cheng, W., Hüllermeier, E.: Preference learning using the choquet integral: The case of multipartite ranking. IEEE T. Fuzzy Systems 20(6), 1102–1113 (2012)
29. Troffaes, M.: Decision making under uncertainty using imprecise probabilities. Int. J. of Approximate Reasoning 45, 17–29 (2007)
30. Walley, P.: Statistical reasoning with imprecise Probabilities. Chapman and Hall, New York (1991)
31. Walley, P.: Inferences from multinomial data: learning about a bag of marbles. Journal of the Royal Statistical Society. Series B (Methodological), 3–57 (1996)
32. Zaffalon, M.: The naive credal classifier. J. Probabilistic Planning and Inference 105, 105–122 (2002)
33. Zaffalon, M., Corani, G., Mauá, D.: Evaluating credal classifiers by utility-discounted predictive accuracy. International Journal of Approximate Reasoning (2012)