Cost-sensitive classification and imprecise probabilities: motivation and some advances

Sebastien Destercke

Heuristic and Diagnosis for Complex Systems (HEUDIASYC) laboratory,
Compiegne, France

CIMI workshop

A bit about my reasearch

- Building bridges between tools of different animals in the uncertainty zoo
- PhD in risk analysis (with E. Chojnacki and D. Dubois), focusing on information fusion, uncertainty propagation and practical uncertainty representation under severe uncertainty
- More recently, focusing on machine learning issues:
  - learning and inferring with uncertain/imprecise data
  - learning and **inferring with structured output** (this talk)
  - using imprecision in active learning

An exemple of structured/complex output

Usual classification

X_1	X_2	w_1	w_2	w_3	w_4
25	*Blue*	1	0	0	0
10	*Red*	0	1	0	0
30	*Blue*	1	0	0	0
5	*Green*	0	0	1	0
15	*Red*	0	0	0	1
...	...	...	...	...	...
5	*Red*	?	?	?	?

Multilabel classification

X_1	X_2	w_1	w_2	w_3	w_4
25	*Blue*	1	0	1	0
10	*Red*	0	1	0	0
30	*Blue*	1	0	1	1
5	*Green*	0	1	1	0
15	*Red*	1	1	0	1
...	...	...	...	...	...
5	*Red*	?	?	?	?

Introductory examples

Predict whether there is a **p**edestrian, a **b**icycle or **n**othing

Cost		Observation		
		p	b	n
Choice	p	0	1	1
	b	1	0	1
	n	1	1	0

Usual costs in classification: 0/1

Introductory examples

Predict whether there is a **p**edestrian, a **b**icycle or **n**othing

Cost		Observation		
		p	b	n
Prediction	p	0	0.5	2
	b	0.5	0	2
	n	10	10	0

Often, different mistakes have different consequences

Introductory examples

Predict the rate someone would give a movie: **v**ery **b**ad, **b**ad, **g**ood, **v**ery **g**ood

Cost		Observation			
		vb	b	g	vg
Prediction	vb	0	1	2	3
	b	1	0	1	2
	g	2	1	0	1
	vg	3	2	1	0

Predictions "further away" from truth worse

Costs

Cost in prediction problems have two main origins:

- given by the application (medical diag., intelligent vehicles, . . .)
- **induced by the output structure**

Interests of imprecise probabilities

- structured data often partially missing
- partially predicted structure may contain needed information

Challenges of imprecise probabilities

- build efficient ways to learn and **infer** with costs in such spaces
- provide **readable** and interpretable imprecise predictions

Why (not) imprecise probabilities?

Why using it?

- you are genuinely interested in having imprecise info/predictions
  - to know when collecting more info (active learning?)
  - to let the decision maker decide about its risk attitude
  - mistakes can be very costly
- you want to postpone precisiation as much as possible
  - to make minimal assumption when processing information
  - you want to postpone precisiation as much as possible

Why not using it?

- you cannot computationally afford it
  - combinatorial issues
  - big data (however, big data $\neq$ lot of data everywhere)
- you have enough data (everywhere)
- making some mistakes is not that damageable (compared to added computational burden)

Talk Outline

1. **Short reminders about IP and Decision**
2. Ordinal regression, or when costs lead to more intuitive results
3. Multilabel classification, or when including costs reduces complexity

Some notations

- Set $\mathcal{Y} = \{y_1, \dots, y_k\}$ of k disjoint states
- Space $\mathcal{A} = \{a_1, \dots, a_d\}$ of possible choices/alternatives
- Either a probability p or a (convex) set $\mathcal{P}$ of them over $\mathcal{Y}$
- Cost function $C : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ with

$$C(a, y)$$

cost of predicting a when y observed value

Decision with precise p

- With the usual 0/1 costs and $\mathcal{A} = \mathcal{Y}$,

$$\begin{aligned} y \succ y' & \text{ if } p(y) > p(y') \\ & \text{ if } p(y) - p(y') > 0 \\ & \text{ if } p(y)/p(y') > 1 \end{aligned}$$

  - involves two variables $p(y), p(y')$
- With generic costs and any $\mathcal{A}$,

$$\begin{aligned} a \succ a' & \text{ if } \mathbb{E}(C(a', \cdot)) > \mathbb{E}(C(a, \cdot)) \\ & \text{ if } \sum_{y \in \mathcal{Y}} p(y)(C(a', y)) - C(a, y)) > 0 \end{aligned}$$

  - involves summation over $\mathcal{Y}$
- $\prec$ complete pre-order → getting it on $\mathcal{A}$ requires d comparisons

Decision with set $\mathcal{P}$

- With the usual 0/1 costs and $\mathcal{A} = \mathcal{Y}$,

$$\begin{aligned} y \succ y' & \text{ if } p(y) > p(y') \text{ for all } p \in \mathcal{P} \\ & \text{if } \inf_{p \in \mathcal{P}} p(y) - p(y') > 0 \\ & \text{if } \inf_{p \in \mathcal{P}} p(y)/p(y') > 1 \end{aligned}$$

  - optimizing over two variables $p(y), p(y')$
- With generic costs and any $\mathcal{A}$,

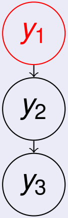
$$\begin{aligned} a \succ a' & \text{ if } \mathbb{E}(C(a', \cdot)) > \mathbb{E}(C(a, \cdot)) \\ & \text{if } \inf_{p \in \mathcal{P}} \sum_{y \in \mathcal{Y}} p(y)(C(a', y)) - C(a, y)) > 0 \end{aligned}$$

  - optimizing over k variables
- $\prec$ partial pre-order → requires at worst $\sim d^2$ comparisons

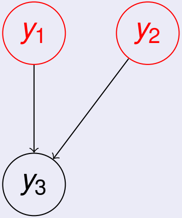
Prediction

Prediction = maximal elements of the (partial) order $\prec$

Precise decision/case



Imprecise decision/case

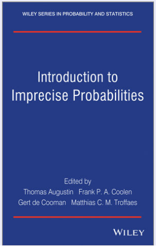


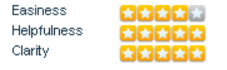
Talk Outline

1. Short reminders about IP and Decision
2. **Ordinal regression, or when costs lead to more intuitive results**
3. Multilabel classification, or when including costs reduces complexity

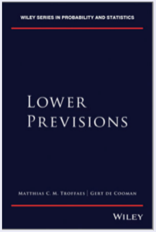
Ordinal classification setting

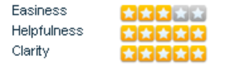
Classes $\mathcal{Y} = \{y_1, \ldots, y_n\}$ ranked, but without metric


WILEY SERIES IN PROBABILITY AND STATISTICS
Introduction to Imprecise Probabilities
Edited by
Thomas Augustin Frank P. A. Coolen
Gert de Cooman Matthias C. M. Troffaes
WILEY



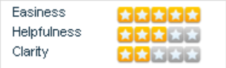
Easiness	★★★★☆
Helpfulness	★★★★★
Clarity	★★★★★


WILEY SERIES IN PROBABILITY AND STATISTICS
LOWER PREVISIONS
MATTHIAS C. M. TROFFAES | GERT DE COOMAN
WILEY



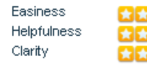
Easiness	★★★☆☆
Helpfulness	★★★★★
Clarity	★★★★★

…



Easiness	★★★★★
Helpfulness	★★★☆☆
Clarity	★★☆☆☆

…



Easiness	★★
Helpfulness	★★
Clarity	★★

Other applications:

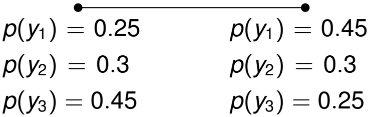
- item ranking
- disease severity diagnosis
- reliability analysis (degradation state)

0/1 cost problem

Consider $\mathcal{A} = \mathcal{Y} = \{y_1, y_2, y_3\}$ and $\mathcal{P}$

For any possible $p \in \mathcal{P}$

- $p(y_1) \in [0.25, 0.45]$
- $p(y_2) = 0.3$
- $p(y_3) \in [0.25, 0.45]$



$p(y_1) = 0.25$
$p(y_2) = 0.3$
$p(y_3) = 0.45$

$p(y_1) = 0.45$
$p(y_2) = 0.3$
$p(y_3) = 0.25$

either $p(y_1)$ or $p(y_3) > 0.3$

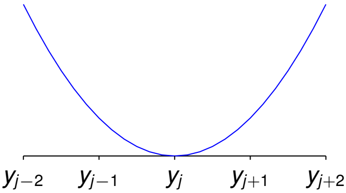
Prediction $\{y_1, y_3\}$ contains "gaps"'

First way around: usual costs (square)

Choosing the function $f(y_i) = i$ replacing y_i by its rank, we can show

- that taking the square cost

$$C_2(y_i, y_j) = (i - j)^2$$



  leads to predict ranks $i \in [\underline{\mathbb{E}}(f), \overline{\mathbb{E}}(f)]$ between lower and upper expectations

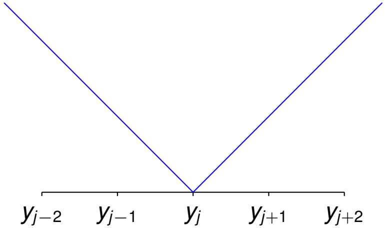
- prediction without gaps
- yet, rely on a non-ordinal concept (expectations)

First way around: usual costs (absolute)

Choosing the function $f(y_i) = i$ replacing y_i by its rank, we can show

- that taking the absolute cost

$$C_1(y_i, y_j) = |i - j|$$



  leads to predict $y_i \in [\underline{Me}_{\mathcal{P}}, \overline{Me}_{\mathcal{P}}]$ between lower and upper medians
- prediction without gaps
- relying on an ordinal concept

Numerical costs: summary

Previous costs:

- solve the issue with 0/1 costs
- extend well-known results from precise case
- yet, they still require to define a numerical cost

can we do with less assumptions?

Second way around: lower/upper median

- general V-shaped symmetric costs such that

$$C(y_i, y_j)$$

  is symmetric and strictly increasing around y_j.
- $C(y_i, y_j) - C(y_k, y_j)$ not numerically defined, yet we have

$$C(y_i, y_j) - C(y_k, y_j) \text{ is } \begin{cases} > 0 & \text{if } |i-j| > |k-j| \\ = 0 & \text{if } |i-j| = |k-j| \\ < 0 & \text{if } |i-j| < |k-j| \end{cases}$$

- using the notion of sign-preference, we can show that

$$[\underline{Me}_{\mathcal{P}}, \overline{Me}_{\mathcal{P}}]$$

  is again a natural solution

Talk Outline

1. Short reminders about IP and Decision
2. Ordinal regression, or when costs lead to more intuitive results
3. **Multilabel classification, or when including costs reduces complexity**

Problem introduction

Among a set $\mathcal{L} = \{\ell_1, \ldots, \ell_L\}$ of L labels, predict which one is relevant

Kind of problems:

- Image tagging (labels: mountains, cars, sea, animals,. . .);
- Functions of a gene, a protein, . . . ;
- Topics of documents, . . .

Problem setting

- $\mathcal{Y}$: set of binary vectors of size L
- $y^j \in \{0, 1\}$ j*th* value of $y \in \mathcal{Y}$
- $y^j = 1$ means j*th* label relevant

We will consider two costs and sets of predictions:

- Hamming costs where $\mathcal{A} = \mathcal{Y}$
- Ranking costs where $\mathcal{A} =$ sets of rankings over $\mathcal{L}$

Some issues

Computational

Comparing 2 alternatives

- for 0/1 costs and $\mathcal{A} = \mathcal{Y}$, may be doable
- for other costs and $\mathcal{A}$, naive summation prohibitive

Building orders if $\mathcal{A} = \mathcal{Y}$

- 2^L comparisons for complete orders
- 2^{2L} comparisons for partial ones

Doable only if L small (< 15) and comparisons computationally cheap

Representational

Providing a (big) set of binary vectors as prediction not very user friendly

0/1 cost and problem structure

Under 0/1 cost and $L = 6$, if

y=	1	1	0	1	0	0

is observed, cost $C(a, y)$ of predicting

a=	0	1	0	1	0	0

same as $C(a', y)$ of predicting

a'=	0	0	1	0	1	1

the 0/1 cost does not integrate any notion of structure. But is a not better than a'?

The Hamming cost

- $\mathcal{A} = \mathcal{Y}$
- $C_H(a, y)$ hamming distance between a and y:

$$C_h(a, y) = \sum_{j \in \{1, \dots, L\}} \mathbf{1}_{(a^j \neq y^j)}$$

  count the number of mistakes
- reflect the structure of the problem

Example

Under the Hamming loss, if

y=	1	1	0	1	0	0

is observed, we have cost $C(a, y) = 1$

a=	0	1	0	1	0	0

and $C(a', y) = 6$

a'=	0	0	1	0	1	1

Predicting with Hamming cost

- if $\mathcal{P}$ probability set over $\mathcal{Y}$ and
- $[\underline{P}(y^j = 1), \overline{P}(y^j = 1)]$ the marginal probability bounds
- the prediction A such that

$$A^j = \begin{cases} 1 & \text{if } \underline{P}(y^j = 1) > 1/2 \\ 0 & \text{if } \overline{P}(y^j = 1) < 1/2 \\ * & \text{else} \end{cases}$$

  includes all the maximal elements (and possibly more) obtained using Hamming cost.
- Computing A requires only $2L$ estimations and comparisons
- Provides an easily readable and computable outer-approximation

Example

Predicting

A= 1 * 0 1 * 0

includes the predictions

1 1 0 1 0 0

1 0 0 1 0 0

1 1 0 1 1 0

1 0 0 1 1 0

The ranking cost

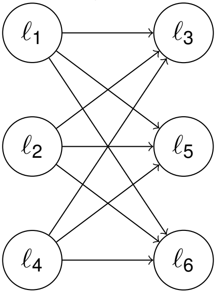
- $\mathcal{A} =$ rankings/set of permutations over $\mathcal{L}$
- $|\mathcal{A}| = L!$, computationally worse than before
- Aim at ranking labels from most to least relevant
- Multilabel observations seen as bipartite dominance graph encoding partial information about ranking
- $C_R(a, y)$ number of discordant pairs between a and y:

$$C_R(a, y) = \sum_{i,j \in \{1,\ldots,L\}^2} \mathbf{1}_{((\ell_i \succ \ell_j) \wedge (y^j = 1, y^i = 0))}$$

Example

Consider $L = 6$ and ℓ_1, ℓ_2, ℓ_4 are relevant

y=



a=

$\ell_1 \succ \ell_4 \succ \ell_2 \succ \ell_3 \succ \ell_5 \succ \ell_6$

$$C_R(a, y) = 0$$

$\ell_4 \succ \ell_2 \succ \ell_1 \succ \ell_5 \succ \ell_6 \succ \ell_3$

$$C_R(a, y) = 0$$

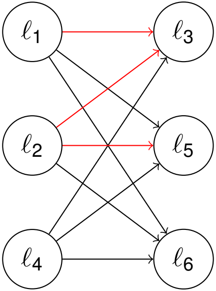
$\ell_4 \succ \ell_3 \succ \ell_1 \succ \ell_5 \succ \ell_2 \succ \ell_6$

$$C_R(a, y) = 3$$

Example

Consider $L = 6$ and ℓ_1, ℓ_2, ℓ_4 are relevant

y=



a=

$\ell_1 \succ \ell_4 \succ \ell_2 \succ \ell_3 \succ \ell_5 \succ \ell_6$	$\ell_4 \succ \ell_2 \succ \ell_1 \succ \ell_5 \succ \ell_6 \succ \ell_3$	$\ell_4 \succ \ell_3 \succ \ell_1 \succ \ell_5 \succ \ell_2 \succ \ell_6$
$C_R(a, y) = 0$	$C_R(a, y) = 0$	$C_R(a, y) = 3$

Predicting with ranking cost

- if $\mathcal{P}$ probability set over $\mathcal{Y}$ and
- $[\underline{P}(y^j = 1), \overline{P}(y^j = 1)]$ the marginal probability bounds
- predicting the partial order $\prec$ such that

  $$\ell_i \prec \ell_j \text{ iff } \overline{P}(y^i = 1) < \underline{P}(y^j = 1)$$

  has linear extensions including all the maximal elements (and possibly more) obtained using ranking cost.
- Computing $\prec$ requires $2L$ estimations and at most L^2 comparisons
- Provides an easily readable and computable outer-approximation
- Drawback: outer-approximation can be of bad quality → go beyond interval orders?

Multilabel case: conclusions

Costs:

- allow to encode that some predictions are closer to the observation
- can consider the case predictions are different from observations
    - observations seen as degraded information
    - use of techniques providing outputs different form observations

"Decomposable" costs

- can lead to efficient and readable inferences
- can pinpoint peculiar values to estimate

Structured output: other problems

- Predicting rankings
  - preferences over objects
  - any relation "more xxx than"
- Predicting partial orders
  - preferences with incomparability
  - acyclic graphs (causal networks?)
- Many other structured outputs
  - hierarchical classes
  - grammar trees
  - (ontic) histograms or fuzzy sets, . . .

Other issues and challenges

Learning and evaluating

- How to efficiently learn models?
  - decomposing the problem
  - directly making the prediction (without estimation step?)
  - use of parametric/simplified models
- How can we define an "optimal" IP model?
  - what makes a IP model "better" than another?
  - how to evaluate IP models and imprecise predicitons with costs?
  - how to define this notion so that optimal model is easy to obtain?

Conclusions

- $\neq$ costs for $\neq$ mistakes in most, if not all practical application
- costs an integral part of many recent machine learning problems
- structured output prediction present technically challenging problems where IP may be useful
- beyond costs for mistakes, need to study cost (value) of information

Some selected references I

[1] Jaime Alonso, Juan José Del Coz, Jorge Díez, Oscar Luaces, and Antonio Bahamonde.
Learning to predict one or more ranks in ordinal regression tasks.
In *Machine Learning and Knowledge Discovery in Databases*, pages 39–54. Springer, 2008.

[2] Alessandro Antonucci and Giorgio Corani.
The multilabel naive credal classifier.

[3] Weiwei Cheng, Eyke Hüllermeier, Willem Waegeman, and Volkmar Welker.
Label ranking with partial abstention based on thresholded probabilistic models.
In *Advances in neural information processing systems*, pages 2501–2509, 2012.

Some selected references II

[4] Sébastien Destercke.
Multilabel predictions with sets of probabilities: the hamming and ranking loss cases.
Pattern Recognition, 2015.

[5] Sébastien Destercke and Gen Yang.
Cautious ordinal classification by binary decomposition.
In *Machine Learning and Knowledge Discovery in Databases*, pages 323–337. Springer, 2014.

[6] Marie-Hélène Masson, Sébastien Destercke, and Thierry Denoeux.
Modelling and predicting partial orders from pairwise belief functions.
Soft Computing, pages 1–12, 2014.