Decision in multilabel and label ranking settings: some issues

Sebastien Destercke

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

WPMSIIP

Outline

1. Basic issues
2. Decision in Multilabel
3. Decision in label ranking

The basic issue

- IP decision rule in classification with $0/1$ losses well-studied
- When looking for more complex framework, problems arise:
  - number of comparisons can explode
  - $0/1$ loss not the only natural one

This talk presents preliminary ideas about these issues

Classical classification

Goal: predict class $y \in \mathcal{W}$ for new instance x

X_1	X_2	X_3	X_4	w_1	w_2	w_3	w_4
107.1	25	*Blue*	60	1	0	0	0
-50	10	*Red*	40	0	1	0	0
200.6	30	*Blue*	58	1	0	0	0
107.1	5	*Green*	53	0	0	1	0
30	15	*Red*	62	0	0	0	1
...	...	...	...	...	...	...	...
200.4	5	*Red*	42	?	?	?	?

Multilabel classification

Goal: predict subset $y \in 2^{\mathcal{W}}$ of relevant labels for new instance x

X_1	X_2	X_3	X_4	w_1	w_2	w_3	w_4
107.1	25	*Blue*	60	1	0	1	0
-50	10	*Red*	40	0	1	0	0
200.6	30	*Blue*	58	1	0	1	1
107.1	5	*Green*	53	0	1	1	0
30	15	*Red*	62	1	1	0	1
...	...	...	...	...	...	...	...
200.4	5	*Red*	42	?	?	?	?

Multilabel classification

Goal: predict ranking/permutation/order $y \in \mathcal{L}(\mathcal{W})$ of labels for new instance x

X_1	X_2	X_3	X_4	w_1	w_2	w_3	w_4
107.1	25	*Blue*	60	4	3	1	2
-50	10	*Red*	40	1	3	2	4
200.6	30	*Blue*	58	4	1	2	3
107.1	5	*Green*	53	1	2	3	4
30	15	*Red*	62	2	3	1	4
...	...	...	...	...	...	...	...
200.4	5	*Red*	42	?	?	?	?

Why using IP in such problems?

- data (more) often incomplete (e.g., partial rankings, pairwise comparisons)
- accurate predictions more difficult to do $\rightarrow$ interest of making partial (but accurate) ones

Some notations

- A set of $\mathcal{W} = \{w_1, \ldots, w_k\}$ of k labels
- A space $\mathcal{Y}$ of predictions (built from $\mathcal{W}$)
- Convex set $\mathcal{P}$ over $\mathcal{Y}$ (learned from data for a new instance)
- Loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ with

$$L(\hat{y}, y)$$

loss of predicting $\hat{y}$ if y true value

Decision problem

- We consider the maximality criterion $\mathcal{M}$
- Under this criterion, prediction $\hat{y} \geq_{\mathcal{M}} \hat{y}'$ iff

$$\underline{E}(L(\hat{y}',\cdot) - L(\hat{y},\cdot)) = \inf_{P \in \mathcal{P}} E(L(\hat{y}',\cdot) - L(\hat{y},\cdot)) \geq 0$$

  $\geq_{\mathcal{M}}$ usually partial order over $\mathcal{Y}$
- Decision set

$$D = \{y \in \mathcal{Y} | \not\exists y' s.t. y' \geq_{\mathcal{M}} y\}$$

  maximal elements of $\geq_{\mathcal{M}}$

Decision in classification case

- Space $\mathcal{Y} = \mathcal{W}$
- "classical" 0/1 loss function $L(\hat{y}, y) = \mathbf{1}_{(\hat{y} \neq y)}$
- In this case

$$\underline{E}(L(\hat{y}', \cdot) - L(\hat{y}, \cdot)) > 0$$

$$\Leftrightarrow$$

$$\underline{E}(\mathbf{1}_{(y)} - \mathbf{1}_{(y')}) = \inf(P(\{y\}) - P(\{y'\}) > 0$$

$$\Leftrightarrow$$

$$\inf({}^{P(\{y\})}/_{P(\{y'\})}) > 1$$

- k^2 computations/comparisons at most

Outline

1. Basic issues
2. Decision in Multilabel
3. Decision in label ranking

Decision in multilabel: 0/1 loss

- Space $\mathcal{Y} = 2^{\mathcal{W}}$
- 0/1 loss function $L_{0/1}(\hat{y}, y) = \mathbf{1}_{(\hat{y} \neq y)}$
- We still have $\hat{y} \geq_{\mathcal{M}} \hat{y}'$ iff

$$\inf({}^{P(\{y\})}/_{P(\{y'\})}) > 1$$

- Getting decision $D_{0/1}$ requires 2^{2k} computations/comparisons at most!
  - $n = 10 \rightarrow 10^6$ comparisons
  - $n = 15 \rightarrow 10^9$ comparisons
- **Can we derive $D_{0/1}$ (or a good approximation) efficiently?**

Decision in multilabel: Hamming loss

If $y = [0, 0, 1, 1]$, $0/1$ loss does not distinguish between predicting

- $\hat{y} = [1, 1, 0, 0]$ ($L_{0/1}(\hat{y}, y) = 1$) and
- $\hat{y}' = [0, 0, 1, 0]$ ($L_{0/1}(\hat{y}', y) = 1$)

$\rightarrow$ unlike usual classification, other natural "basic" losses

- Hamming loss

$$L_H(\hat{y}, y) = \frac{1}{k} \sum_{i=1,\dots,k} \mathbf{1}_{(\hat{y}_i \neq y_i)}$$

  with y_i the i^{th} component of y.
- $L_H(\hat{y}, y) = 1$
- $L_H(\hat{y}', y) = 1/4$
- under L_H, predicting $\hat{y}'$ is not so bad.

Decision in multilabel: Hamming loss

Loss minimizer $\hat{y}$ when $\mathcal{P} = P$

- $\hat{y}_i = 1$ if marginal $P(y_i = 1) > 0.5$
- $\hat{y}_i = 0$ else
- $\Rightarrow$ easy to compute

Can we obtain something similar to derive D_H with imprecise $\mathcal{P}$?

Decision in multilabel: Hamming loss

Consider two decisions $\hat{y}$ and $\hat{y}'$ such that

- for a given i, $\hat{y}_i = 1 \neq \hat{y}'_i = 0$
- $\hat{y}_j = \hat{y}'_j$ for $j \neq i$

then we can show

- $\underline{P}(y_i = 1) > 0.5 \Rightarrow \underline{E}(L_H(\hat{y}', \cdot) - L_H(\hat{y}, \cdot)) > 0$
- $\underline{P}(y_i = 0) > 0.5 \Rightarrow \underline{E}(L_H(\hat{y}, \cdot) - L_H(\hat{y}', \cdot)) > 0$

this means that the partial prediction $\hat{Y}$ such that

- $\hat{Y}_i = 1$ if $\underline{P}(y_i = 1) > 0.5$
- $\hat{Y}_i = 0$ if $\underline{P}(y_i = 0) > 0.5$
- $\hat{Y}_i \in \{0, 1\}$ else

is such that $D_H \subseteq \hat{Y}$, with inclusion possibly strict $\Rightarrow$ easy outer-approximation (requires $2k$ computation)

Decision in multilabel: Hamming loss

A short intuition of the result

y	$L_H(110, \cdot)$	$L_H(110, \cdot)$	$\ominus$
000	2	1	1
001	3	2	1
010	1	2	-1
100	1	0	1
011	2	3	-1
101	2	1	1
110	0	1	-1
111	1	2	-1

- 2-valued gamble
- value depends on the changing label

Decision in multilabel: ranking loss

The prediction is an order relation $\succ$ over labels $w_1, \ldots, w_k$. Ranking loss is

$$L_R(\hat{y}, y) = \frac{1}{|y_i = 1| \cdot |y_i = 0|} \sum_{y_i=1, y_k=0} \mathbf{1}_{((w_k, w_i) \in \succ)}$$

- $|y_i = 1|$ number of relevant labels
- $|y_i = 0|$ number of irrelevant labels
- $(w_k, w_i) \in \succ$ means $w_k \succ w_i$

$\Rightarrow$ loss assumes prediction done in another space (orders)!

If $\mathcal{P} = P$, loss minimizer given by $w_k \succ w_i$ if $P(y_k = 1) \geq P(y_i = 1)$

If $\mathcal{P}$, study what happens if $w_k \succ w_i$ when $\underline{P}(y_k = 1) \geq \overline{P}(y_i = 1)$? Need to consider partial orders. How to (easily) derive D_R?

Outline

1. Basic issues
2. Decision in Multilabel
3. Decision in label ranking

Label ranking: introduction

- Space is set of permutations $\mathcal{Y} = \mathcal{L}(\mathcal{W})$
- 0/1 loss function $L_{0/1}(\hat{y}, y) = \mathbf{1}_{(\hat{y} \neq y)}$
- We still have $\hat{y} \geq_{\mathcal{M}} \hat{y}'$ iff

$$\inf({}^{P(\{y\})}/_{P(\{y'\})}) > 1$$

- Getting decision $D_{0/1}$ requires $k!^2$ computations/comparisons at most!
  - $n = 10 \rightarrow 10^{13}$ comparisons
  - $n = 15 \rightarrow 10^{24}$ comparisons
- **Can we derive $D_{0/1}$ (or a good approximation) efficiently?**

Label ranking: measuring accuracy

Many possible measures

- Kendall's L_k:

$$L_k(\hat{y}, y) = \frac{C - D}{k(k-1)/2}$$

 with $C = |(w_i, w_j) \in y \wedge \hat{y}|$ (concording pairs) and $D = |(w_i, w_j) \in (y \wedge \neg\hat{y}) \vee (\neg y \wedge \hat{y})|$ (discording pairs)

- Spearman rank L_s:

$$L_s(\hat{y}, y) = 1 - \frac{6D(\hat{y}, y)}{k(k^2 - 1)}$$

 where $D(\hat{y}, y) = \sum_{i=1,\dots,k} (\hat{y}(w_i) - y(w_i))^2$ with $y(w_i)$ rank of w_i

Accuracy: example

$$y = w_1 \leq w_2 \leq w_3$$

$$\hat{y} = w_1 \leq w_3 \leq w_2$$

- $A_{0/1} = 0$
- $\tau = {}^{1}\!/_{3}$

$$\hat{y} = w_3 \leq w_2 \leq w_1$$

- $A_{0/1} = 0$
- $\tau = -1$

Reduce the problem: pairwise decomposition

Use marginal information $P(\{w_i \geq w_j\})$ to infer ranking.

One way: $S(w_i) = \sum_{j=1}^{k} P(\{w_i \geq w_j\})$ and order according to S (always gives an ordering without cycles)

$\Rightarrow$ minimize Spearman L_s loss if estimates $P(\{w_i \geq w_j\})$ are perfect

- **+**: make the problem tractable (n^2 item of info vs $n!$)
- **-**: loss of information compared to complete space $\mathcal{Y}$

Pairwise comparison: the precise case

$\geq$	w_1	w_2	w_3	w_4	$\sum$	S
w_1	0	0.3	0.4	0.2		0.9
w_2	0.7	0	0.6	0.3		1.6
w_3	0.6	0.4	0	0.4		1.4
w_4	0.8	0.7	0.6	0		2.1

Prediction: $w_1 \leq w_3 \leq w_2 \leq w_4$

Pairwise comparison: the imprecise case

$\geq$	w_1	w_2	w_3	w_4	$\sum$	S
w_1	0	[0.2,0.4]	[0.3,0.5]	[0.1,0.3]		[0.6,1.2]
w_2	[0.6,0.8]	0	[0.5,0.7]	[0.2,0.4]		[1.3,1.9]
w_3	[0.5,0.7]	[0.3,0.5]	0	[0.3,0.5]		[1.1,1.7]
w_4	[0.7,0.9]	[0.6,0.8]	[0.5,0.7]	0		[1.8,2.4]

Prediction: $w_1 \leq w_4$ and $w_1 \leq w_2$

$\Rightarrow$ **provides partial prediction, related to D_s, the maximal set under Spearman loss? the same way as Hamming in multilabel?**

Reduce the problem: use parametric models

Plackett-Luce model order object iteratively (first, second, ...)

v_{w_i}: "probability" of w_i first in a race with $w_i, w_{i+1} \dots w_k$

Probability w_1 first: $\frac{v_{w_1}}{\sum_{j=1}^{k} v_{w_j}}$

Probability w_2 second= being first among $w_2 \dots w_k$

$$P(y) = \prod_{i=1}^{m} \frac{v_{w_i}}{v_{w_i} + v_{w_{i+1}} + \dots + v_{w_k}}$$

$\Rightarrow$ if precise model, loss minimizer same for $L_{0/1}, L_s, L_k$.

Reduce the problem: use parametric models

Assume parameters v_{w_i} known to lie in $[\underline{v}_{w_i}, \overline{v}_{w_i}]$ define $\mathcal{P}$

if $\hat{y}$ and $\hat{y}'$ such that only w_i, w_j swapped between the two

${}^{P(\hat{y})}/{}_{P(\hat{y}')}$ only depends on $[\underline{v}_{w_i}, \overline{v}_{w_i}]$, $[\underline{v}_{w_j}, \overline{v}_{w_j}]$ (thanks Alessandro!)

Easy to obtain partial order outer-approximating $D_{0/1}$

$$w_i \succ w_j \text{ if } \underline{v}_{w_i} > \overline{v}_{w_j}$$

Questions:

- is it equal to $D_{0/1}$?
- does loss matter when consider imprecise model?
- how to learn imprecise $[\underline{v}_{w_i}, \overline{v}_{w_i}]$?

Conclusions

- extracting maximal sets of solutions for structured data hard!
- need for efficient (approximate) solution $\rightarrow$ results from precise case can help (sometimes)
- yet other interesting problems:
  - ordinal classification
  - graded multilabel classification
  - predicting graphs (semantic trees, relational graphs, ...)