Cautious label-wise ranking with constraint satisfaction

Sébastien Destercke, Yonatan Carlos Carranza Alarcon

DA2PL 2018

Some announcements: SUM 2019

- When: 16-18 december 2019
- Where: Compiègne
- What: (scalable) undertainty management
- How: papers (long/short/abstracts) but also tutorials/surveys of particular areas

Where is Compiegne?

Our approach in a nutshell

What?

Cautious label-ranking by rank-wise decomposition

How?

- Rank-wise decomposition
- For each label, predict set of ranks using imprecise probabilities
- Use CSP to:
  - resolve inconsistencies
  - remove impossible assignments

Why?

- weak information in structured settings more prone to be of use
- few rank-wise approaches (except score-based) for this problem

Introduction and decomposition

Ranking data - preferences

To each instance x correspond an ordering over possible labels

Blog theme

A blog x can be about

$$\text{Politic} \succ \text{Literature} \succ \text{Movies} \succ \ldots$$

Cutomer preferences

A customer x may prefer

$$\text{White wine} \succ \text{Red wine} \succ \text{Beer} \succ \ldots$$

Classification problem/data

$$\mathcal{C} = \{c_1, c_2, c_3\}$$

X_1	X_2	X_3	X_4	Y
107.1	25	*Blue*	60	c_3
-50	10	*Red*	40	c_1
200.6	30	*Blue*	58	c_2
107.1	5	*Green*	60	c_4
...	...	...	...	...

Label ranking problem/data

$$\mathcal{W} = \{w_1, w_2, w_3\}$$

X_1	X_2	X_3	X_4	Y
107.1	25	*Blue*	60	$w_1 \succ w_3 \succ w_2$
-50	10	*Red*	40	$w_2 \succ w_1 \succ w_3$
200.6	30	*Blue*	58	$w_1 \succ w_2 \succ w_3$
107.1	5	*Green*	60	$w_3 \succ w_1 \succ w_2$
...	...	...	...	...

Potentially huge output space (K! with K labels) $\rightarrow$ naive extension (one ranking=one class) doomed to fail

One solution: rank-wise decomposition

$\mathbb{D}$

X_1	X_2	X_3	X_4	Y
107.1	25	*Blue*	60	$\lambda_1 \succ \lambda_3 \succ \lambda_2$
−50	10	*Red*	40	$\lambda_2 \succ \lambda_3 \succ \lambda_1$
200.6	30	*Blue*	58	$\lambda_2 \succ \lambda_1 \succ \lambda_3$
107.1	5	*Green*	33	$\lambda_1 \succ \lambda_2 \succ \lambda_3$
...	...	...	...	...

$\mathbb{D}_1$ ↙ $\mathbb{D}_2$ ↓ ↘ $\mathbb{D}_3$

X_1	X_4	Y
107.1	60	1
−50	40	3
200.6	58	2
107.1	33	1
...	...	...

X_1	X_4	Y
107.1	60	3
−50	40	1
200.6	58	1
107.1	33	2
...	...	...

X_1	X_4	Y
107.1	60	2
−50	40	2
200.6	58	3
107.1	33	3
...	...	...

For each label, solve an ordinal regression problem

Predicting candidate ranks

Learning with IP: a crash course

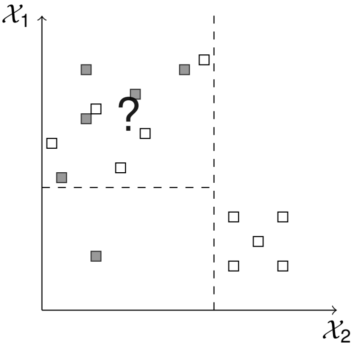
Classical case:

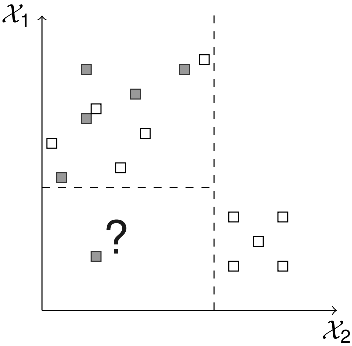
- input space $\mathcal{X}$ and output space $\mathcal{Y}$
- set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of data
- given x, estimate $P(y|x)$ using $\mathcal{D}$
- $P(y|x)$ = information about y when observing x

However, estimate $\hat{P}(y|x)$ of $P(y|x)$ can be pretty bad if

- data are noisy, missing, imprecise
- estimation is based on little data

Replace the estimate $\hat{P}(y|x)$ by **a set** $\mathcal{P}$ of estimates





Ambiguity

$$P(\square|?) \in [0.49, 0.51]$$
$$P(\blacksquare|?) \in [0.49, 0.51]$$

Lack of information

$$P(\square|?) \in [0, 0.7]$$
$$P(\blacksquare|?) \in [0.3, 1]$$

Decision with probability sets

Probability sets

If $\ell_\omega : \mathcal{Y} \to \mathbb{R}$ loss function of choice $\omega \in \mathcal{Y}$, then

$$\omega \succeq \omega' \Leftrightarrow \inf_{P \in \mathcal{P}(y|x)} \mathbb{E}(\ell_{\omega'} - \ell_\omega) \geq 0 = \underline{\mathbb{E}}(\ell_{\omega'} - \ell_\omega)$$

$$\Leftrightarrow \inf_{P \in \mathcal{P}(y|x)} \sum_{y \in \mathcal{Y}} P(y|x) \left(\ell_{\omega'}(y) - \ell_{\omega'}(y)\right) \geq 0$$

$\Rightarrow$ if insufficient information, we can have $\omega \not\succeq \omega'$ and $\omega' \not\succeq \omega$
That is, we can have $\underline{\mathbb{E}}(\ell_{\omega'} - \ell_\omega) < 0$ and $\underline{\mathbb{E}}(\ell_\omega - \ell_{\omega'}) < 0$

$\Rightarrow$ Possibly optimal decisions = maximal element(s) of $\succeq$

Our choice of $\ell, \mathcal{P}$

What?

- $\ell = L_1$ norm between ranks, loss of predicting rank j if k is true

$$\ell_j(k) = |j - k|$$

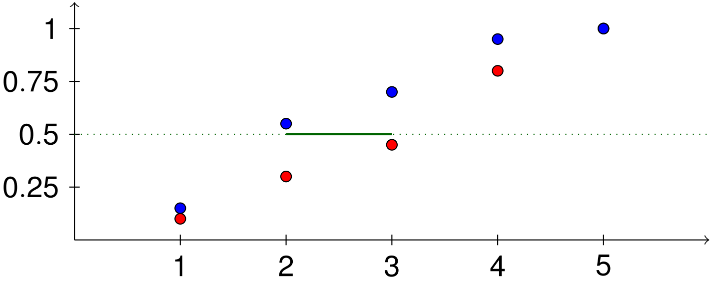
- $\mathcal{P}$ described by lower/upper cumulative distributions $\underline{F}, \overline{F}$

Why?

- prediction is guaranteed to be an “interval" of ranks (dedicated CSP models)
- it corresponds to the set of possible medians (very easy to get)

An example of rank prediction

Rank j	1	2	3	4	5
$\overline{F}_j$	0.15	0.55	0.7	0.95	1
$\underline{F}_j$	0.1	0.3	0.45	0.8	1



Predicted rank for label: $\{2, 3\}$

Making a final cautious prediction

Inconsistency and assignment reductions

Inconsistency

Consider four labels $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, then the predicted possible ranks

$$\hat{R}_1 = \{1,3\}, \hat{R}_2 = \{1,3\}, \hat{R}_3 = \{1,3\}, \hat{R}_4 = \{2,4\}$$

are inconsistent $\rightarrow \lambda_1, \lambda_2, \lambda_3$ should **all** take different values

Removal of impossible solutions

Consider the predictions

$$\hat{R}_1 = \{1,2\}, \hat{R}_2 = \{1,2,3\}, \hat{R}_3 = \{2\}, \hat{R}_4 = \{1,2,3,4\}.$$

As λ_3 has to take value 2, λ_1 has to take value $\{1\}$, ...until we get

$$\hat{R}'_1 = \{1\}, \hat{R}'_2 = \{3\}, \hat{R}'_3 = \{2\}, \hat{R}'_4 = \{4\}$$

Dealing with the issue: CSP modelling

- A possible assignment $\hat{R}_i \subseteq \{1, \ldots, K\}$
- Need to find if each of them can take a different value
- Exactly what the *all different* constraint does in CSP
- So, just apply standard librairies

Bonus: if all $\hat{R}_i$ intervals, efficient (polynomial) algorithms exist

Experiments

Setting

Material and method

- Classification and regression data sets turned into ranking
- Binary decomposition + Naive imprecise classifier

Measuring results quality

Completeness (CP)

$$CP(\hat{R}) = \frac{k^2 - \sum_{i=1}^{k} |\hat{R}_i|}{k^2 - k}$$

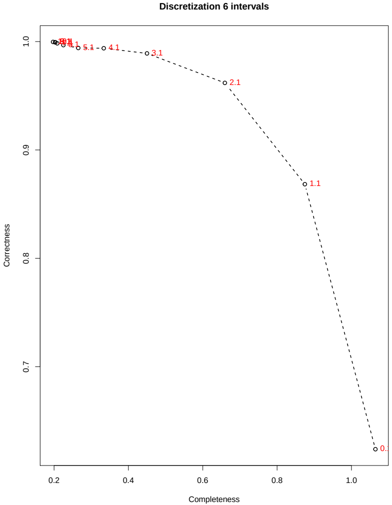
Max if one ranking possible
Min if all rankings possible

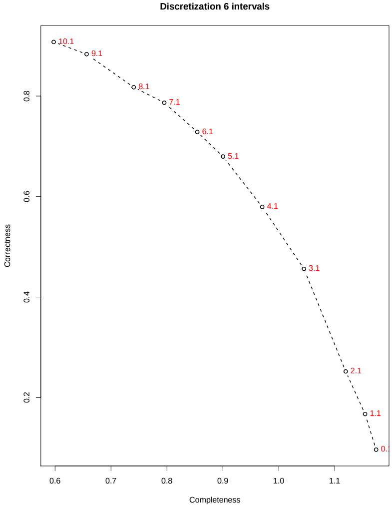
Correctness (CR)

$$CR(\hat{R}) = 1 - \frac{\sum_{i=1}^{k} \min_{\hat{r}_i \in \hat{R}_i} |\hat{r}_i - r_i|}{0.5k^2}$$

Equivalent to Spearman footrule if one ranking predicted

An example of results





Why rank-wise approaches?

Yes, why?

- different expressiveness when it comes to represent partial predictions, e.g., the set-valued prediction

$$\{\lambda_1 \succ \lambda_2 \succ \lambda_3, \lambda_2 \succ \lambda_2 \succ \lambda_1\}$$

  between three labels is perfectly representable by imprecise ranks, but not through pairwise information or partial orders (i.e., interval-valued scores)
- not (entirely) clear how to make score-based methods imprecise (IP-SVM?) + need to turn them into imprecise ranks?