Robust classification: basic issues and challenges

Sébastien Destercke
CNRS, Heudiasyc, Université de Technologie de Compiègne

Ales workshop

The problem illustrated

Task: recognizing obstacles for autonomous, mobile robots

Basic (supervised) learning setting

- Input features + observed output in $\mathcal{X} \times \mathcal{Y}$
- A number of observations (x_i, y_i), $i = 1, \ldots, n$
- From them, learn a model with parameters $\hat{\theta} \in \Theta$
- For new x, prediction $\hat{\theta}(x)$

X^1	$\ldots$	X^M	Y	
25	$\ldots$	*Blue*	a	Training
10	$\ldots$	*Red*	b	Training
30	$\ldots$	*Blue*	a	Training
$\ldots$	$\ldots$	$\ldots$	$\ldots$	Training
5	$\ldots$	*Green*	c	Training
15	$\ldots$	*Red*	b	Training
55	$\ldots$	*Green*	?	

Decision rule

Probabilistic case:

- A loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$
- $\ell(\hat{y}, y)$ is the loss of predicting $\hat{y}$ if y is the truth
- $y \geq y'$ if $E(\ell(y, \cdot)) \leq E(\ell(y', \cdot))$ with

$$E(\ell(y,\cdot)) = \sum_{\omega \in \mathcal{Y}} p(\omega|x)\ell(y,\omega)$$

$$\ell_{0,1} = \begin{array}{c c} & \begin{array}{cc} S & F \end{array} \\ \begin{array}{c} \hat{S} \\ \hat{F} \end{array} & \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \end{array} \qquad \ell = \begin{array}{c c} & \begin{array}{cc} S & F \end{array} \\ \begin{array}{c} \hat{S} \\ \hat{F} \end{array} & \begin{pmatrix} 0 & \alpha \\ \beta & 0 \end{pmatrix} \end{array}$$

$$S \geq F \Leftrightarrow p(S) \geq p(F) \qquad S \geq F \Leftrightarrow \beta p(S) \geq \alpha p(F)$$

Precise models and extended cost matrix

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

Cost		Observation p	Observation b	Observation n
Prediction	p	0	1	2
Prediction	b	1	0	2
Prediction	n	10	10	0

Often, different mistakes have different consequences

Classical predictions

Assuming probability $p(p) = 0.1, p(b) = 0.4, p(n) = 0.5$, we would predict

$$\begin{array}{cc} & \begin{array}{ccc} p & b & n \end{array} \\ \begin{array}{c} p \\ \mathbf{b} \\ n \end{array} & \begin{pmatrix} 0 & 1 & 2 \\ \mathbf{1} & \mathbf{0} & \mathbf{2} \\ 10 & 10 & 0 \end{pmatrix} \end{array} \times (p(p), p(b), p(n))^T = \begin{pmatrix} 1.4 \\ \mathbf{1.1} \\ 5 \end{pmatrix}$$

Or, in terms of dominance $\mathbf{b} \succ p \succ n$

Talk topic

Issue

In usual setting, a **single** class/output of lowest expected cost is predicted:

- Reasonable when many decisions of small impact (e.g., Amazon recommendation, Google ranking) $\rightarrow$ losing sometimes ok if winning in average
- Questionable when decisions are rare or mistakes of big consequences

Question(s)

- Which extensions to be informative but cautious in case of doubt?
- How to evaluate such extensions?

First solution: extend the cost matrix

Assuming probability $p(p) = 0.1, p(b) = 0.4, p(n) = 0.5$, we want the possibility to make indeterminate predictions

$$\begin{array}{cc} & \begin{array}{ccc} p & b & n \end{array} \\ \begin{array}{c} p \\ b \\ n \\ \{p,b\} \\ \{p,n\} \\ \{b,n\} \\ \{p,b,n\} \end{array} & \begin{pmatrix} 0 & 1 & 2 \\ 1 & 0 & 2 \\ 10 & 10 & 0 \\ ? & ? & ? \\ ? & ? & ? \\ ? & ? & ? \\ ? & ? & ? \end{pmatrix} \end{array} \times (p(p), p(b), p(n))^T = \begin{pmatrix} 1.4 \\ 1.1 \\ 5 \\ ? \\ ? \\ ? \\ ? \end{pmatrix}$$

How should we complete the matrix? [3, 2, 5, 1]

A suitable matrix

Assuming probability $p(p) = 0.1, p(b) = 0.4, p(n) = 0.5$, and the following complete matrix

$$
\begin{array}{l}
 \\
p \\
b \\
n \\
\{\mathbf{p}, \mathbf{b}\} \\
\{p, n\} \\
\{b, n\} \\
\{p, b, n\}
\end{array}
\begin{array}{c}
\begin{array}{ccc} p & b & n \end{array} \\
\begin{pmatrix}
0 & 1 & 2 \\
1 & 0 & 2 \\
10 & 10 & 0 \\
\mathbf{0.1} & \mathbf{0.1} & \mathbf{2} \\
4 & 4 & 1 \\
4 & 4 & 1 \\
3 & 3 & 1
\end{pmatrix}
\end{array}
\times (p(p), p(b), p(n)) =
\begin{pmatrix}
1.4 \\
1.1 \\
5 \\
\mathbf{1.05} \\
2.5 \\
2.5 \\
2
\end{pmatrix}
$$

Dominance relation $\{p, b\} \succ b \succ p \succ \{p, b, n\} \succ \{p, n\} \equiv \{b, n\} \succ n$

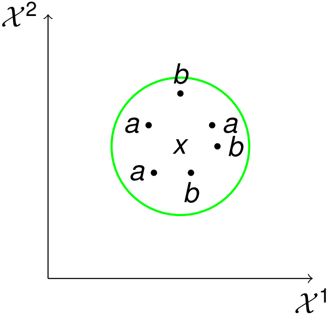
First solution: pros and cons

Imprecision is added in decision, uncertainty representation unmodified:

- +: usually rather efficient
- +: can be plugged to any existing probabilistic method
- -: gain in information=change of preferences
- -: uninformed situation not distinguished from ambiguous one

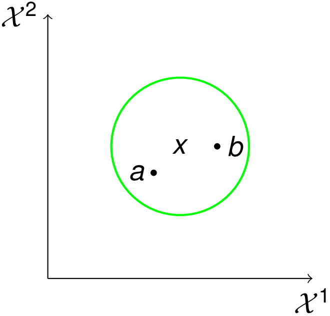
Two kinds of uncertainties

- Aleatory uncertainty: classes are really mixed $\rightarrow$ irreducible with more data (but possibly by adding features)
- Epistemic uncertainty: lack of information $\rightarrow$ reducible



Aleatory uncertainty

$P(a) \in [0.45, 0.55]$



Epistemic uncertainty

$P(a) \in [0.2, 0.8]$

From imprecise models to imprecise decision [4]

General idea: proceed through skeptic inference

- given a set $\mathcal{P}$ of possible models
- pairwise comparison: $y > y'$ only if so for every model within $\mathcal{P}$. In the imprecise probabilistic case:

$$y > y' \Leftrightarrow \min_{p \in \mathcal{P}} E_p(\ell(y', \cdot) - (\ell(y, \cdot)) > 0$$

- possible winners: y is a possibly optimal answer if there is a model for which it is optimal. In the imprecise probabilistic case:

$$\exists p \in \mathcal{P} \text{ with } y \in \arg\min_{\omega \in \mathcal{Y}} E_p(\ell(\omega, \cdot)$$

Quite different principle: we change the uncertainty representation.

Classical predictions

Assuming probability $p(p) \in [0, 0.2], p(b) \in [0.3, 0.5], p(n) \in [0.4, 0.6]$, we would predict

$$\begin{array}{c c} & \begin{array}{ccc} p & b & n \end{array} \\ \begin{array}{c} p \\ \mathbf{b} \\ n \end{array} & \begin{pmatrix} 0 & 1 & 2 \\ \mathbf{1} & \mathbf{0} & \mathbf{2} \\ 10 & 10 & 0 \end{pmatrix} \end{array} \times (p(p), p(b), p(n))^T = \begin{pmatrix} [1.2, 1.6] \\ [0.9, 1.3] \\ [4, 6] \end{pmatrix}$$

Or, in terms of dominance $\{\mathbf{b}, \mathbf{p}\} \succ n$

Second solution: pros and cons

Imprecision is added in uncertainty representation, decision unmodified:

- -: can be computationally heavy
- -: need to extend existing probabilistic method
- +: gain in information=refinement of preferences (what is said in the past remains true in the future)
- +: can distinguish lack of information from observed ambiguity

The price of cautiousness

Before being cautious, need to answer questions:

- Why do I want to be cautious? What use for that?
- How much do I want to be cautious?
- If cautious, what means an optimal trade-off between:
  - Being totally uninformative but right
  - Being fully precise but more often wrong

Again, this depends highly of the context, but how can we formalize it?

The two doctors story

In a hospital, doctors get 1$ each time diagnostic is right.

2 Doctors pretty sure that patients have either Pneumonia (P) or Bronchitis (B)

Doctor 1

- Flip a coin each time
- Diagnose the result
- Gets 0.5$ in average

Doctor 2

- Tells you he does not know b/w P and B
- Should his reward be 0.5 $, same as doc 1? higher? lower?

Main solution so far for 0/1 loss

$$u(\hat{Y}, y) = \begin{cases} 0 & \text{if } y \notin \hat{Y} \\ \frac{\alpha}{|\hat{Y}|} + \frac{1-\alpha}{|\hat{Y}|^2} & \text{otherwise} \end{cases}$$

with $u(\hat{Y}, y) = 1$ if $|\hat{Y}| = 1$ and $\hat{Y} = y$

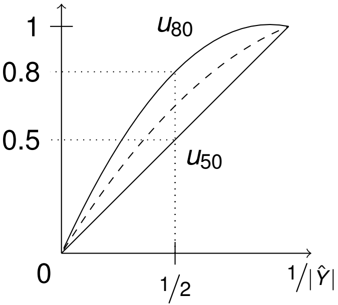
- Discounted accuracy: $\alpha = 1$

$$u(\hat{Y}, y) = \frac{1}{|\hat{Y}|}$$

  $\rightarrow$ no reward to cautiousness (cautiousness$\equiv$randomness)
- u_{65}: $\alpha = 1.6$, moderate reward to cautiousness
- u_{80}: $\alpha = 2.2$, big reward to cautiousness
- $u_{'\infty'}$: $\rightarrow 1$ if $y \in \hat{Y}$, no penalty for being cautious

Solutions exists for generic losses too.

Boldness averseness illustrated



2 classes predicted,
good one in it

Data sets and results

#	a	b	c	d	e	f	g	h	i	j
Names	Breats	Iris	Wine	Auto	Seed	Glass	Forest	Derma	Diabete	Segment
Instances	106	150	178	205	210	214	325	366	769	2310
Features	10	4	13	26	7	9	27	34	8	19
Labels	6	3	3	7	3	7	4	6	2	7

		$SR = 50\%$		$\epsilon = 20\%$			$SR = 50\%$		$\epsilon = 20\%$	
#	*stats*	$\epsilon = 10\%$	$\epsilon = 40\%$	SR=30%	SR=75%	#	$\epsilon = 10\%$	$\epsilon = 40\%$	SR=30%	SR=75%
	precise	56.4	56.4	56.9	57.0		80.7	80.7	80.1	82.6
a	u_{65}	61.4	49.0	55.2	55.9	*f*	52.7	39.6	45.9	46.3
	precise	97.1	97.1	96.3	95.8		87.4	87.4	87.2	87.3
b	u_{65}	97.3	96.6	96.9	96.1	*g*	88.5	88.9	88.2	88.4
	precise	62.7	62.7	61.3	62.9		98.9	98.9	99.1	99.0
c	u_{65}	86.2	82.0	84.9	85.9	*h*	96.9	78.3	92.2	92.8
	precise	80.0	80.0	79.6	79.8		79.2	79.2	79.7	79.7
d	u_{65}	82.8	61.0	74.9	74.0	*i*	79.7	79.6	80.0	79.5
	precise	93.1	93.1	93.6	94.0		89.3	89.3	89.2	89.3
e	u_{65}	92.4	91.6	92.2	92.2	*j*	61.7	50.1	56.7	56.3

- Adding even little imprecision harmful
- Adding little imprecision good, a lot bad
- Adding imprecision, even quite a lot, actually pays off

Conclusions

- Many ways to add cautiousness in learning problems
- Not all of them equivalent, at least from a principled standpoint (but also from a practical one)
- Important to answer the questions: why and how much do we want to be cautious?

References I

[1] Juan José del Coz, Jorge Díez, and Antonio Bahamonde.
Learning nondeterministic classifiers.
Journal of Machine Learning Research, 10(Oct):2273–2293, 2009.

[2] Thien M Ha.
The optimum class-selective rejection rule.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(6):608–615, 1997.

[3] Gen Yang, Sébastien Destercke, and Marie-Hélène Masson.
The costs of indeterminacy: How to determine them?
IEEE transactions on cybernetics, 47(12):4316–4327, 2017.

[4] M. Zaffalon.
The naive credal classifier.
J. Probabilistic Planning and Inference, 105:105–122, 2002.

[5] Marco Zaffalon, Giorgio Corani, and Denis Mauá.
Evaluating credal classifiers by utility-discounted predictive accuracy.
International Journal of Approximate Reasoning, 53(8):1282–1301, 2012.