Likelihood-Based Naive Credal Classifier

Alessandro Antonucci, Marco Cattaneo, Giorgio Corani
IDSIA, Switzerland & Ludwig-Maximilians-Universität München, Germany
alessandro@idsia.ch cattaneo@stat.uni-muenchen.de giorgio@idsia.ch

Credal Classifiers (CCs)

Classification Class variable C with generic value $c \in \mathscr{C}$
Features $\mathbf{F} := (F_1, \ldots, F_m)$, values $f_i \in \mathscr{F}_i$, $i = 1, \ldots, m$
Given data, which class label for the instance $\mathbf{F} = \tilde{\mathbf{f}}$?

Bayesian Classifiers Learn joint distribution $P(C, \mathbf{F})$
Assign to $\mathbf{f}$ the most probable class label $\arg\max_{c' \in \mathscr{C}} P(c', \tilde{\mathbf{f}})$
This defines a *classifier*, i.e., a map: $(\mathscr{F}_1 \times \ldots \times \mathscr{F}_m) \to \mathscr{C}$

Credal Classifiers Learn joint credal set $\mathbf{P}(C, \mathbf{F})$
Set of optimal classes (e.g., according to *maximality*)

$$\{c' \in \mathscr{C} \mid \nexists c'' \in \mathscr{C}, \forall P \in \mathbf{P} : P(c''|\tilde{\mathbf{f}}) > P(c'|\tilde{\mathbf{f}})\}$$

This defines a *credal classifier*, i.e.,

$$(\mathscr{F}_1 \times \ldots \times \mathscr{F}_m) \to 2^{\mathscr{C}}$$

May return more than a single class label!

CCs Performance Evaluation

Accuracy is not a sufficient descriptor for CCs performances!

- *determinacy*: % of instances classified with a single class
- *single/set accuracy*: accuracy over instances classified with single class/multiple classes
- *indeterminate output size*: average # of classes when classification indeterminate

Learning Credal Sets (IDM)

Bayesian learning extended to imprecision

Set of Dirichlet modelling prior near-ignorance

Bounds of the posterior is

$$P(c) \in \left[\frac{n(c)}{N+s}, \frac{n(c)+s}{N+s}\right]$$

$n(\cdot)$ counting function

Real parameter s as equivalent sample size

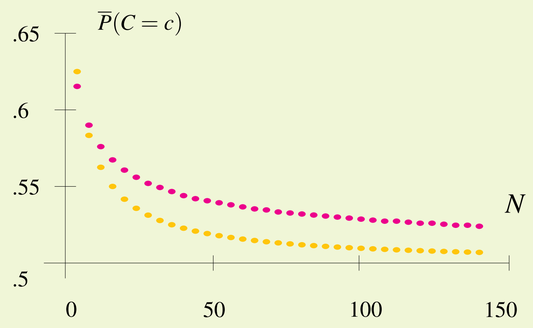
Typical choices $s \in [1, 2]$

If $s \to +\infty$, vacuous intervals

If $s \to 0$, ML estimator

IDM vs. Likelihood

Learning credal sets from N data $\mathscr{D}$ about Boolean C, with $n(c) = N/2$, by likelihood-based ($\alpha = .85$) and IDM ($s = 2$) approaches



Learning Credal Sets (Likelihood)

Frequentist (ML) learning extended to imprecision

With complete (multinomial) data, likelihood is unimodal

Replace the single ML estimator, with the set of models whose likelihood is behind a threshold (α times the ML)

$$\mathbf{P}^{\alpha}(C) = \{P(C) \in \mathbf{P}(C) : P(\mathscr{D}) \geq \alpha P_{\mathrm{ML}}(\mathscr{D})\}$$

The starting credal set $\mathbf{P}(C)$ can be vacuous
(or any other more informative set)

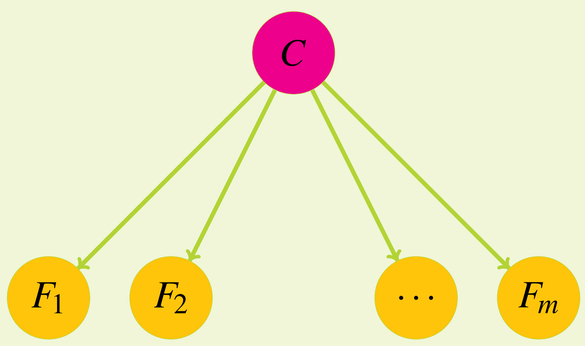
For each $\alpha \in [0, 1]$, $\mathbf{P}^{\alpha}(C) \subseteq \mathbf{P}(C)$. In particular:

$\mathbf{P}^{\alpha=0}(C) = \mathbf{P}(C)$ (no threshold, no learning)

$\mathbf{P}^{\alpha=1}(C) = \{P_{\mathrm{ML}}(C)\}$ (highest threshold, ML "precise" learning)

Naive Classifiers

given class C, features $(F_1, \ldots, F_m)$ are conditionally independent



Often unrealistic, but good for classification!

E.g., NBC (naive Bayes classifier [3])

Can be extended to the imprecise case with both concepts of strong independence and epistemic irrelevance. Same inferences are obtained!

Naive Credal Classifier (with IDM, [4])

IDM to learn joint credal set $\mathbf{P}(C, \mathbf{F})$ under naive assumption

Efficient algorithm for maximality-based classification!

Optimization problem

$$\left[\frac{n(c') + st(c')}{n(c'') + st(c'')}\right]^{1-m} \prod_{j=1}^{m} \frac{n(c', \tilde{f}_j)}{n(c'', \tilde{f}_j) + st(c'', \tilde{f}_j)},$$

with the IDM constraints on the Dirichlet priors

$$\mathscr{T} := \left\{ \mathbf{t} \;\middle|\; \begin{array}{l} \sum_{c \in \mathscr{C}} t(c) = 1 \\ \sum_{f_j \in \mathscr{F}_i} t(c, f_j) = t(c), \forall j \\ t(c, f_j) > 0, \forall (c, f_j) \in \mathscr{C} \times \mathscr{F}_j, \forall j \end{array} \right\}.$$

Reject c'' if the optimum is bigger than one

Coping with zero-counts If $n(c', \tilde{f}_j) = 0$, class c' cannot dominate any other class (feature problem). IDM constraints $\mathscr{T}$ can be rewritten by an ε-contamination in the form $\varepsilon |\mathscr{C}|^{-1} \leq t(c) \leq (1 - \varepsilon) + \varepsilon |\mathscr{C}|^{-1}$, and similarly for $t(c, f_j)$ [2].

This is called NCC$_{\varepsilon}$ (standard NCC for $\varepsilon = 0$, NBC for $\varepsilon = 1$).

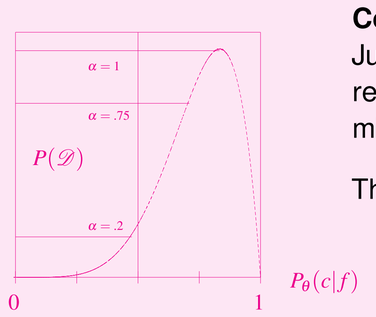
Likelihood-based NCC [1]

Start with the "vacuous" credal set $\mathbf{P}(C, \mathbf{F})$ including all the possible NBC specifications

Refine this set with likelihood-based learning + maximality:

$$\inf_{P \in \mathbf{P} : P(\mathscr{D}) \geq \alpha P_{\mathrm{ML}}(\mathscr{D})} \frac{P(c', \tilde{\mathbf{f}})}{P(c'', \tilde{\mathbf{f}})} > 1$$

This condition for c' dominating c'' can be checked with no need of stochastic approaches: **an analytical formula for the upper envelope of the likelihood** is derived (see paper)



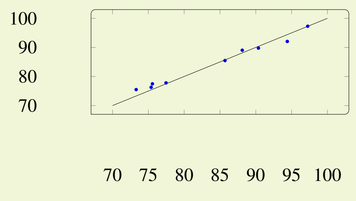
Coping with zero-counts Just add to the likelihood record $(C = *, \mathbf{F} = \mathbf{f})$ with C missing-at-random!

This is called LNCC$_{\alpha}$.

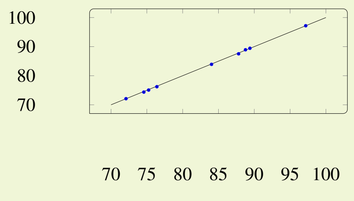
Experiments

- *Discounted accuracy*: rewards a set-valued classification with $1/k$ or 0, depending on whether the set contains or not the correct class; *Single Accuracy* is the accuracy of the classifier when it return a single class.
- The performance of the two classifiers is very close, when the determinacy is comparable!

Single Accuracy



Discounted-accuracy



$\mathscr{D}$	NCC	Det	Sgl-acc	D-acc	NBC-I
1	IDM $\varepsilon = 0.05$	96.1	75.6	74.6	58.6
1	LIK $\alpha = 0.95$	95.7	75.7	74.6	57.5
2	IDM $\varepsilon = 0.05$	95.1	73.3	72.1	45.6
2	LIK $\alpha = 0.95$	93.9	73.8	71.9	50.2
3	IDM $\varepsilon = 0.05$	95.3	85.7	60.3	84.0
3	LIK $\alpha = 0.75$	95.4	85.5	61.1	83.9

The scatter plots compare NCC$_{\varepsilon}$ (x-axis) and LNCC$_{\alpha}$ (y-axis).

References

[1] M. Cattaneo. Likelihood-based inference for probabilistic graphical models: Some preliminary results. In *PGM 2010*, pages 57–64. HIIT Publications, 2010.

[2] G. Corani and A. Benavoli. Restricting the IDM for classification. In *IPMU 2010*, pages 328–337. Springer, 2010.

[3] P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Mach. Learn.*, 29:103–130, 1997.

[4] M. Zaffalon. Statistical inference of the naive credal classifier. In *ISIPTA '01*, pages 384–393, 2001.