Credal Classification

Giorgio Corani
giorgio@idsia.ch

ULB – Machine Learning Group
June 2012

About the speaker

- PhD in Information Engineering (Politecnico di Milano, 2005).
- Visiting period at MLG group (lazy learning).
- Since 2006: researcher at IDSIA (`www.idsia.ch`), Switzerland.
- Research interests: probabilistic graphical models, data mining, statistical modelling in general.

Estimating a multinomial

- Variable X, with sample space $\{x_1, \ldots, x_m\}$.
- The vector of probabilities to be estimated is $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_m\}$.
- A complete data set D of n observations is available, whose sufficient statistics are the counts $\{n_1, \ldots, n_m\}$.

Bayesian approach

- The prior distribution $P(\boldsymbol{\theta})$ models our *a priori* beliefs about $\boldsymbol{\theta}$.
- $P(\boldsymbol{\theta})$ is a Dirichlet distribution with parameters $\{\alpha_1, \ldots, \alpha_m\}$; $\alpha = \sum \alpha_i$.
- The posterior $P(\boldsymbol{\theta}|D)$ is obtained multiplying the prior by the multinomial likelihood; it is again a Dirichlet.
- Taking expectation from the posterior $P(\boldsymbol{\theta}|D)$:

$$\hat{\theta}_j = E[\theta_j]_{post} = \frac{n_j + \alpha_j}{n + \alpha}$$

Non-informative priors

- Non informative prior: all α_j are equal.
- This is a model of prior indifference.
- Nothing prevents adopting a non-uniform prior, yet this is uncommon.

Modelling prior-ignorance: the IDM (Walley, 1996)

- The IDM is a convex set of Dirichlet prior.
- It avoids stating that a priori the states are equally probable: there is no reason for such a strong statement.
- The IDM represents prior ignorance.
- The *credal* set of priors is multiplied by the likelihood, yielding a *credal* set of posteriors.

Example

- Let us consider a binary variable, with n=10, n_1=4, n_2=6.
- The estimates of θ_1 are:

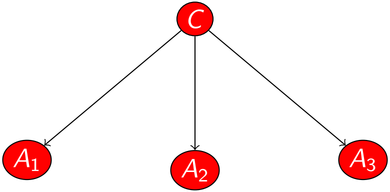
Bayes $(\alpha_1 = 0.5, \alpha_2 = 0.5)$	Bayes $(\alpha_1 = 0.8, \alpha_2 = 0.2)$	IDM
$\hat{\theta}_1 = \frac{4+0.5}{10+1}$	$\hat{\theta}_1 = \frac{4+0.8}{10+1}$	$\hat{\theta}_1 \in \left[\frac{4}{10+1}, \frac{4+1}{10+1}\right]$

- The interval estimate of the IDM comprises the point estimates obtained using different Dirichlet priors, letting each α_1 range within $(0, 1)$.

Credal classifiers

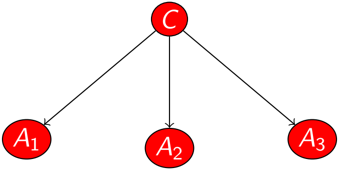
- Credal classifiers learn using a set of priors.
- They can identify the instances whose most probable class varies with the prior (prior-dependent).
- On prior-dependent instances, Bayesian classifiers are unreliable; credal classifiers return instead more classes (indeterminate classifications).
- In this way, they robustly deal e.g. with small data sets.

Naive Bayes (NBC)



- *Naively* assumes the features to be independent given the class.
- NBC is highly biased, but achieves good accuracy, especially on small data sets, thanks to low variance (Friedman,1997).

Naive Bayes (NBC)



- Learns from data the joint probability of class and features, decomposed as the marginal probability of the classes and the conditional probability of each feature given the class.

Naive Credal Classifier (NCC)

- Uses the IDM to specify a set of joint prior distributions; this is updated with the likelihood, yielding a set of posteriors.
- When classifying an instance, class c' credal-dominates c'' if for each prior of the IDM:

$$P(c'|\mathbf{a}) > P(c''|\mathbf{a})$$

  where **a** represents the set of observed features.

- Credal-dominance is checked by solving an optimization problem.
- NCC eventually returns the *non-dominated* classes.

.

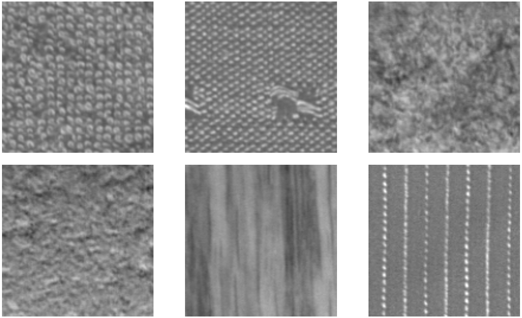
NCC and prior-dependent instances

- NCC returns more classes on the instances recognized as prior-dependent; a single class on the safe instances.

Test on UCI data sets:

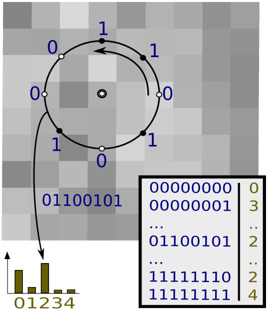
- the % of indeterminate classifications tend to decrease on larger data sets;
- NBC is unreliable on prior-dependent instances, while NCC remains reliable on them thanks to indeterminate classifications.

The OUTEX data sets (Ojala, PAMI 2002)



- 4500 images, 24 different textures (carpets, woods, etc.).x
- The goal: identifying the class of each image.

Local Binary Patterns



- This technique compares the gray level of each selected pixel with that of the surrounding points.
- For each pixel, the pattern of gray is represented by a string of 0 and 1.
- By processing the strings obtained in many different pixels, we eventually extract 18 features for each image.

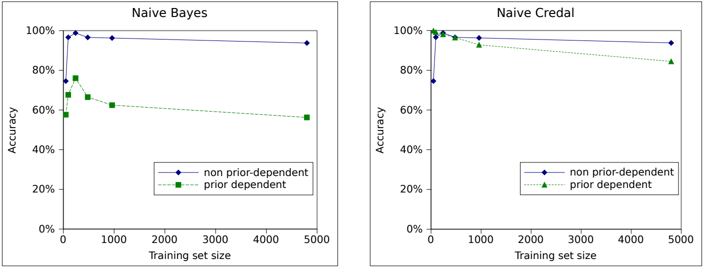
Cross-validation results

- NBC accuracy is 92% on average, decomposed as 94% on the safe instances and 56% on the prior-dependent ones.

	Safe instances	*Prior-dependent*
Amount%	95%	5%
NBC: accuracy	94%	56%
NCC: accuracy	94%	85%
NCC: non-dom. classes	1	2.4

Sensitivity on n

- Smaller training sets generated by stratified downsampling.



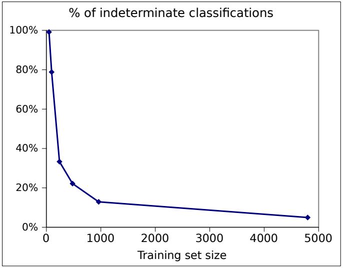
At any sample size

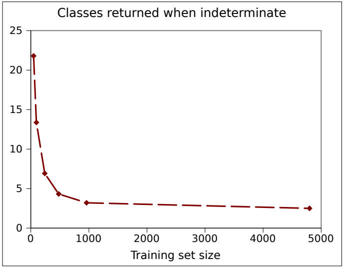
- the accuracy of NBC drops on prior-dependent instances;
- indeterminate classifications preserve the reliability of NCC.

Different training set sizes (II)

As n grows:

- the % of indet. classification decreases;
- the avg. number of classes returned when indeterminate decreases.





- On larger data sets, the choice of the prior has less importance.

Rejection rule

- Refuses to classify an instance, if the the most probable class does not achieve a probability threshold.
- But in the previous example half of the prior-dependent instances is classified by NCC with probability $> 90\%$.
- In general, the instances seen as uncertain by the rejection rule and the credal classifier overlap only partially.
- Moreover, rejection rule is formally justified only is a cost is set for not deciding, unlike credal classifiers.

Comparing indeterminate and traditional classifiers

- Designing a synthetic score for comparing indeterminate and traditional classifiers is very challenging.
- For instance, on the prior-dependent instances, do you prefer a 85% accuracy returning two classes, or 55% returning a single class?
- To the best of my knowledge, the most convincing approach is to account for the utility of the decision maker (Zaffalon et al., ISIPTA '11).
- These utility-based metric are numerically close to information retrieval scores such as the F_1 or F_2 metric.

Improvements over NCC

- Conservative treatment of missing data.
- Lazy NCC.
- Credal TAN (tree-augmented networks).
- Best so far: credal model averaging, extending the AODE model (Webb et al., Machine Learning, 2005): accepted at ECAI '12.

Future works

- Developing scoring rules for indeterminate classifications in cost-sensitive settings.
- Further develop the credal model averaging approach.
- Discriminative learning of credal classifiers.
- Credal regression.