An Overview of Credal Classification

G. Corani and M. Zaffalon

Istituto “Dalle Molle” di Studi sull’Intelligenza Artificiale
Lugano (Switzerland)

AI*IA-14
About the speaker

- PhD in Information Engineering (Politecnico di Milano, 2005).
- Since 2006: researcher at IDSIA (www.idsia.ch), Switzerland.
- Imprecise Probability Group (http://ipg.idsia.ch)
- Research interests: Probabilistic graphical models, data mining, applied statistic, imprecise probability.
Estimating a multinomial from data: Bayesian vs. imprecise probability approach.

Naive Bayes and naive credal classifiers.

Bayesian model averaging and credal model averaging.

Comparing credal and traditional classifiers.
Estimation of a multinomial: Bayesian vs imprecise probability.
Estimation of a multinomial variable

- Categorical variable $X$, with categories $\{x_1, \ldots, x_m\}$.
- Denote by $\theta_m$ the probability of $x_m$, $\theta = \{\theta_1, \ldots, \theta_m\}$.
- Count observed on data set: $n = \{n_1, \ldots, n_m\}$.
- *Multinomial* likelihood:

$$P(n|\theta) \propto \prod_{j=1}^{m} \theta_j^{n_j}$$

- Max. likelihood estimator: $\hat{\theta}_j = \frac{n_j}{n}$.
Bayesian estimation: Dirichlet prior

- The prior expresses the beliefs about $\theta$, before analyzing the data:

\[
\pi(\theta) \propto \prod_{j=1}^{k} \theta_{j}^{s_{j}-1}.
\]

- $s > 0$ is the equivalent sample size, which can be regarded as a number of hidden instances;

- $t_{j}$ is the proportion of hidden instances in category $j$. 
Posterior distribution

- Obtained by multiplying likelihood and prior:

\[
\pi(\theta|n) \propto \prod_j \theta_j^{(n_j + s t_j - 1)}
\]

- Dirichlet posteriors are obtained from Dirichlet priors (conjugacy).

- Taking expectations from the posterior:

\[
E[\theta_j|n] = \frac{n_j + s t_j}{n + s}
\]
Example

- $n=10$, $n_1=4$, $n_2=6$, $s=1$.
- The posterior estimate depends on the prior:

<table>
<thead>
<tr>
<th>Prior 1</th>
<th>Prior 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1 = 0.5$</td>
<td>$t_1 = 0.8$</td>
</tr>
<tr>
<td>$t_2 = 0.5$</td>
<td>$t_2 = 0.2$</td>
</tr>
</tbody>
</table>

$$\hat{\theta}_1 = \frac{4 + 0.5}{10 + 1} = 0.41$$

$$\hat{\theta}_1 = \frac{4 + 0.8}{10 + 1} = 0.44$$

Uniform prior is the most common choice, but often one repeats the analysis under different priors.

- The IDM is a convex set of Dirichlet priors.

- The $t_j$ are constrained as follows:

\[
\begin{cases}
0 < t_j < 1 & \forall j \\
\sum_j t_j = 1
\end{cases}
\]

- This is a model of prior ignorance.
Learning from data with the IDM

- *Upper* and *lower* posterior estimate of $\theta_j$:

\[
E(\theta_j|\mathbf{n}) = \inf_{0<t_j<1} \frac{n_j + st_j}{n + s} = \frac{n_j}{n + s}
\]

\[
\overline{E}(\theta_j|\mathbf{n}) = \sup_{0<t_j<1} \frac{n_j + st_j}{n + s} = \frac{n_j + s}{n + s}
\]

- Automatic sensitivity analysis w.r.t the prior parameters $t_j$. 

G. Corani and M. Zaffalon (IDSIA) 

An Overview of Credal Classification 

AI*IA-14
Example

- Binary variable, with $n=10$, $n_1=4$, $n_2=6$, $s=1$.

- The estimates of $\theta_1$ are:

\[
\begin{align*}
\text{Bayes} & \quad \text{Bayes} & \quad \text{IDM} \\
(t_1 = 0.5, t_2 = 0.5) & \quad (t_1 = 0.8, t_2 = 0.2) & \\
\hat{\theta}_1 &= \frac{4 + 0.5}{10 + 1} & \hat{\theta}_1 &= \frac{4 + 0.8}{10 + 1} & \left[\frac{4}{10 + 1}, \frac{4 + 1}{10 + 1}\right] \\
& = 0.409 & & = 0.436 & & = [0.363, 0.454]
\end{align*}
\]

- The IDM estimate comprises the estimates obtained letting vary each $t_i$ within $(0, 1)$. 
On the IDM

▶ The estimates are *imprecise*, being characterized by an upper and a lower bound.

▶ The gap between upper and lower probability narrows down as the data set increases.

References:

Application to classification

From Naive Bayes to Naive Credal Classifier.
Classification

- Determine the type of Iris (class) on the basis of length and width (features) of sepal and petal.

(a) setosa  (b) virginica  (c) versicolor

- Classification: to predict the class $C$ of a given object, on the basis of features $\mathbf{A} = \{A_1, \ldots, A_k\}$.

- Prediction: the *most probable class* a posteriori.
Naive Bayes (NBC)

- *Naively* assumes the features to be independent given the class.
- NBC achieves good accuracy especially on small data sets thanks to low variance of the parameter estimates (Friedman, 1997).
Naive Bayes (NBC)

The joint probability of class and features decomposes as the marginal probability of the classes and the conditional probability of each feature given the class.
Training naive Bayes

- Requires learning different multinomials:
  - the marginal probability of the class \( P(C) \);
  - the conditional probability each feature \( A_i \) given the class \( P(A_i|C) \).

- Usually one adopts the uniform prior.
Classifying an instance

- Assume two classes $c_0$ and $c_1$

- Given the observed features $a_1, a_2, \ldots, a_k$ the posterior probability of $c_0$ is:

$$P(c_0|a_1, a_2, \ldots, a_k) = \frac{P(c_0) \prod_{i=1}^{k} P(a_i|c_0)}{P(c_0) \prod_{i=1}^{k} P(a_i|c_0) + P(c_1) \prod_{i=1}^{k} P(a_i|c_1)}$$

- Under 0-1 one returns the most probable class.
Prior-dependent classifications

- An instance is prior-dependent if the most probable class depends on the prior used to learn the multinomials.

- Prior-dependent classifications are more common on small data sets.

- They are overlooked by Bayesian classifiers.

- They are instead detected by credal classifiers.
Naive Credal Classifier (NCC)

- Uses the IDM to specify the marginal and the conditional prior probabilities of naive Bayes.
- The IDM probabilities factorizes yielding a set of joint prior distributions over class and features.
- When performing classification, the posterior probability of class $c$ ranges within an interval.
- The interval highlights the sensitivity of the posterior probability of the class on the prior used to learn the multinomials.
NCC and prior-dependent instances

- Given feature observation $\mathbf{a}$, class $c'$ credal-dominates $c''$ if for each posterior of the credal set:

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

- If an class credal-dominates all the others, the instance is safe: the most probable class does not vary with the prior.

- If the instance is prior-dependent, there are multiple non-dominated classes returned by NCC.

Naive Bayes and Naive Credal Classifier:

Texture recognition
Texture recognition

- The OUTEX data sets (Ojala, 2002): 4500 images, 24 classes (textiles, carpets, woods ..).
Features: Local Binary Patterns (Ojala, 2002)

- The gray level of each pixel is compared with that of its neighbors, resulting in a binary judgment (0: more intense/1: less intense).
- Such booleans are collected in a string for each pixel.
Local Binary Patterns (2)

▶ Each string of boolean is assigned to a category, according to the pattern of 0’s and 1’s. There are 18 categories.
▶ For each image there are 18 features: the % of pixels assigned to each category.
Results (Corani et al., BMVC 2010)

- Accuracy of NBC: 92% (SVMs: 92.5%).
- NBC is highly accurate on the safe instances, but almost random on the prior-dependent ones.

<table>
<thead>
<tr>
<th></th>
<th>Safe</th>
<th>Prior-dependent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Amount %</strong></td>
<td>95%</td>
<td>5%</td>
</tr>
<tr>
<td><strong>NBC: accuracy</strong></td>
<td>94%</td>
<td>56%</td>
</tr>
<tr>
<td><strong>NCC: accuracy</strong></td>
<td>94%</td>
<td>85%</td>
</tr>
<tr>
<td><strong>NCC: non-dom. classes</strong></td>
<td>1</td>
<td>2.4</td>
</tr>
</tbody>
</table>
Indeterminacy decreases with the sample size

As \( n \) grows there is a decrease of both:

- the % of indet. classification;
- the number of classes returned when indeterminate.
Tree-augmented credal classifier (TAN)

- TAN improves over naive Bayes by relaxing the independence assumption.
Tree-augmented credal classifier (TAN)

- Credal TAN detects instances unreliable classified by Bayesian TAN because of prior-dependence.

References:
- Zaffalon, Reliable Computing, 2003
Credal ensemble of classifiers

Bayesian Model Averaging and Credal Model Averaging
Model uncertainty.

- Consider logistic regression with features $A_1, A_2, \ldots, A_k$.

- Which is the optimal set of features?

- Given $k$ features, we can design $2^k$ regression models, each with a different feature set.

- Model uncertainty: different logistic models (with different feature sets) can fit similarly well the data. Considering only a single model leads to overconfident conclusions.
Bayesian Model Averaging (BMA)

- **Bayesian Model Averaging (BMA)** combines the inferences of multiple models, using as weights the posterior probability of the models.

- The inferences yielded by BMA depends on the prior set over the models.

- Often the BMA analysis is repeated under different prior over the models.
Credal model averaging (CMA)

- Credal model averaging substitutes the single prior over the models by a set of priors.

- A priori the probability of each model ranges within an interval rather being fixed to a number.

- CMA automates sensitivity analysis with respect to the prior over the models.

- CMA detects prior-dependent instances, whose most probable class depends on the prior over the models.
Predicting presence or absence of marmot burrows

- Study area: 9500 cells (100m² each) within the Stelvio National Park (Italy).
- Presence of burrows in about 4.5% of the cells.
- Features available for each cell:
  - altitude, slope, aspect (the direction in which the slope faces), soil temperature, soil cover, etc.
BMA accuracy

- BMA is highly accurate on the safe instances but almost a random guesser on the prior-dependent ones!

Corani and Mignatti, IJAR 2014.
Comparing Credal and Traditional Classifiers
Discounted-accuracy

\[d-\text{acc} = \frac{1}{N} \sum_{i=1}^{N} \frac{(\text{accurate})_i}{|Z_i|}\]

- \(\text{accurate}_i\): whether the set of classes returned on instance \(i\) contains the actual class;
- \(|Z_i|\): the number of classes returned on the \(i\)-th instance.
- For a traditional classification, \(d-\text{acc}\) equals the 0-1 accuracy.
Some properties

\[ d\text{-}acc = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{(accurate)}_i}{|Z_i|} \]

- The d-acc of an accurate classification is positive; the d-acc of an inaccurate classification is 0.

- Among two accurate classifications, d-acc is higher for the more informative classification.
Doctor random and doctor vacuous.

- Possible diseases: \{A,B\}.
- Doctor \textit{random}: uniformly random diagnosis.
- Doctor \textit{vacuous}: return both categories.

<table>
<thead>
<tr>
<th>Disease</th>
<th>Random</th>
<th>Vacuous</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class.</td>
<td>d-acc</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>A</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>B</td>
<td>1</td>
</tr>
</tbody>
</table>

\[
\overline{\text{d-acc}} = 0.5 \quad 0.5
\]
Doctor random vs doctor vacuous.

- Assume that the hospital profits a quantity of money proportional to the discounted-accuracy.

- After $n$ visits, the profits are:
  - doctor *vacuous*: $n/2$, with *no variance*.
  - doctor *random*: expected $n/2$, with *variance* $n/4$. 
Introducing utility

- Expected utility increases with the expected value of the rewards but decreases with their variance.

- Any risk-adverse manager prefers doctor vacuous over doctor random.

- And you prefer doctor vacuous over doctor random!

- Idea: quantify such preference through a utility function (Zaffalon et al., IJAR 2012).
How to design the utility function

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
<th>Utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
<td>0</td>
</tr>
</tbody>
</table>

- Utility corresponds to accuracy for traditional classification consisting of a single class.
How to design the utility function

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
<th>Utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B, C</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>A, B</td>
<td>0.65-0.80</td>
</tr>
</tbody>
</table>

- A wrong indeterminate classification has utility 0.
- The utility of an accurate but indeterminate classification consisting of two classes has to be larger than 0.5 ...
- ... otherwise doctor random and doctor vacuous yield the same utility.
How to design the utility function

- Let this utility range between 0.65 (function $u_{65}$) or 0.8 (function $u_{80}$), then perform sensitivity analysis.

- Utility of credal classifiers and accuracy of determinate classifiers can be now compared.

- Such utility functions are numerically close to F1 and F2 metric, also used to score indeterminate classifications (Del Coz et al., JMLR, 2009).
Comparing naive Bayes and naive Credal

Wins on 55 UCI data sets:

<table>
<thead>
<tr>
<th>Utility</th>
<th>naive Bayes</th>
<th>naive Credal</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_{65}$</td>
<td>18</td>
<td>37</td>
</tr>
<tr>
<td>$u_{80}$</td>
<td>13</td>
<td>42</td>
</tr>
</tbody>
</table>

- Similar results if we compare Bayesian model averaging vs. credal model averaging, or credal TAN vs Bayesian TAN.
Other application of imprecise probability.

- Imprecise Bayesian networks (credal networks: see our IJCAI ’13 tutorial)

- Representing expert knowledge.

- Credal classification trees (Abellan et al., Comp. Statistics & Data Analysis, 2013)

- Imprecise signed-rank test (Benavoli et al, ICML 2014)
Interested in working with imprecise probability?

- We are happy to collaborate with other research groups.


- We will also host the next conference on Probabilistic Graphical Models (PGM 2016).