Imprecise probabilities and machine learning: a tradeoff between accuracy and epistemic uncertainty

M. Serrurier
IRIT, Toulouse, France

May 27, 2015
Outline

- Motivations
- Possibility distribution as a family of probability distributions
- Possibilistic loss function
- Possibilistic entropy
- Application to the learning of trees
- Conclusions
Motivations

- Probability and density estimation widely used in machine learning.
- Limitations: parameters of the distributions are estimated from limited samples of data.
  - More data $\rightarrow$ more complex models
- Model selection: best trade-off between accuracy and epistemic uncertainty
- Idea: use ability of possibility theory to represent epistemic uncertainty
Possibility theory: discrete case

- Possibility distribution $\pi$ is a mapping from $\Omega$ to $[0, 1]$
- Possibility measure: $\forall A \subseteq \Omega, \Pi(A) = \sup_{x \in A} \pi(x)$
  - $\Pi(A \cup B) = \max(\Pi(A), \Pi(B))$
  - $\Pi(A \cap B) \leq \min(\Pi(A), \Pi(B))$
- Necessity measure: $\forall A \subseteq \Omega, N(A) = 1 - \Pi(\bar{A})$.
- States of knowledge:
  - complete knowledge: $\exists x \in \Omega$ such as $\pi(x) = 1$ and $\forall y \in \Omega, y \neq x, \pi(y) = 0$
  - total ignorance: $\forall x \in \Omega, \pi(x) = 1$. 

Possibility distribution and upper bound of probability distribution

- Qualitative interpretation: description of imprecise concept (cheap, young, ...)
- Probabilistic interpretation: Upper bound of a family of probability distribution:

$$\mathcal{P}(\pi) = \{ p \in \mathcal{P}, \forall A \in \Omega, N(A) \leq P(A) \leq \Pi(A) \}.$$

- States of knowledge:
  - complete knowledge: no uncertainty.
  - total ignorance: all probability distributions are possible.
Specificity order

- Specificity $\pi \preceq \pi'$, if and only if:

$$\pi \preceq \pi' \iff \forall x \in \Omega, \pi(x) \leq \pi'(x)$$

- $\sigma$-specificity (discrete case): $\pi \preceq_\sigma \pi'$, if and only if there exists a permutation $\sigma \in S_q$ such as:

$$\pi \preceq_\sigma \pi' \iff \forall x \in \Omega, \pi(x) \leq \pi'(\sigma(x))$$

- Specificity reflects the amount of information encoded by the distribution.
Probability-possibility transformation

- ∀σ ∈ S_q we have a cumulative distribution $T^\sigma_p$ which encodes $p$:

  $\forall j \in \{1, \ldots, q\}, T^\sigma_p(C_j) = \sum_{k, \sigma(k) \leq \sigma(j)} p(C_k)$.

- Probability-possibility transformation ($T^*_p$) : the most σ specific possibility distribution which bounds the distribution

- $T^*_p$ is a cumulative function of $p$

- $T^*_p = T^\sigma^*_p$ where correspond to $\sigma^* \in S_q$ follows the probability increasing order
Probability-possibility transformation: example

\[ p \quad T_p^\sigma \]

\[ p \quad T_p^\sigma \]

\[ p \quad T_p^\sigma \]

\[ p \quad T_p^\sigma \]
α-cuts are subsets of Ω such that:

\[ A_\alpha(\pi) = \{ x \in \Omega, \pi(x) \geq \alpha \}. \]

- Specificity order based on inclusion of α-cuts
- Probability possibility transformation:
  \[ T^*_p(x) = \max_{x \in \alpha \in I_\alpha} (1 - \alpha) \]
  \[ T^*_p \text{ corresponds to the cumulative distribution function of } p \text{ according to the order the values of } p(x). \]
- α-cuts are upper bounds of \( I_{(1-\alpha)} \)
Probability-possibility transformation of a Gaussian distribution

\[ A_{0.75}^* = I_{0.25}^* \]

\[ A_{0.25}^* = I_{0.75}^* \]

M.s.p.d for N(0,1)

N(0,1)
Possibilistic distribution encoding uncertainty around Gaussian parameters

$X$

Density

$-6$ $-4$ $-2$ $0$ $2$ $4$ $6$

$0.0$ $0.2$ $0.4$ $0.6$ $0.8$ $1.0$ $N(0,1)$

$N(\mu, \sigma)$

$N(a,b)$

$\pi$

$\text{trapez}^*$

M. Serrurier IRIT, Toulouse, France

Imprecise probabilities and machine learning

May 27, 2015 11 / 37
Why using possibility distribution?

- Probability-possibility transformation $\rightarrow$ loss of information but..
- Possibility distribution can different state of knowledge from complete knowledge to total ignorance
- $T_p^\ast +$uncertainty around parameters of $p \rightarrow$ less specific possibility distribution
- $\sigma$-specificity respect the entropy order

\[ T_p^\ast \preceq_\sigma T_p^{\ast \prime} \Rightarrow H(p) \leq H(p') \]

Goal: describe loss and entropy functions that support the $\sigma$-specificity order and the probabilistic interpretation of possibility distribution
Probabilistic loss functions

- Loss functions $\mathcal{L}(f, X)$ measure adequateness between data $X = \{x_1, \ldots, x_n\}$ and a distribution $f$.

- Loss function $\mathcal{L}(f, X)$ is linear w.r.t. $X$:
  $$\mathcal{L}(f, X) = \frac{\sum_{i=1}^{n} \mathcal{L}(f, x_i)}{n}$$

- Common loss functions:
  - Log loss: $\mathcal{L}_{\log}(p|X) = -\sum_{j=1}^{q} \alpha_j \log(p_j)$.
  - Squared loss: $\mathcal{L}_{sqr}(p|X) = \frac{1}{2} * \sum_{j=1}^{q} p_j^2 - (\sum_{j=1}^{q} \alpha_j * p_j)$

- Loss functions are minimal for the frequency distribution (i.e. $p_j = \alpha_j$)
Possibilistic log-loss function

- Principle given a possibility distribution $\pi$:
  - consider $\sigma$ s.a. $\pi(C_{\sigma(1)}) \leq \ldots \leq \pi(C_{\sigma(q)})$
  - consider $BC_j = \bigcup_{i=1}^{j} C_{\sigma(i)}$ and $\overline{BC}_j$ as binary event space
  - $(\pi(C_{\sigma(j)}), 1 - \pi(C_{\sigma(j)}))$ is a probability distribution on $\Omega_j = \{BC_j, \overline{BC}_j\}$
  - apply re-scaled loss function to each Bernoulli distribution

- Poss-log loss:

  $\mathcal{L}_{\pi-l}(\pi|X)) = \\
  - \sum_{j=1}^{q} \left( \frac{cdf_j}{2} \ast \log\left( \frac{\pi_j}{2} \right) + (1 - \frac{cdf_j}{2}) \ast \log(1 - \frac{\pi_j}{2}) \right)$.

- Poss-squared loss:

  $\mathcal{L}_{\pi-s}(\pi|X) = \frac{1}{2} \sum_{j=1}^{q} \pi_j^2 + \sum_{j=1}^{q} cdf_j - \sum_{j=1}^{q} \pi_j \ast cdf_j$
Possibilistic log-loss function: properties

**Linearity**

$L_\pi$ is linear with respect to $X$

**Optimality for probability possibility transformation**

we have $\arg\min(L_\pi(\pi|X)) = T_{p^\alpha}^*$ (where $p^\alpha$ is the frequency distribution).

**Specificity order**

$\forall \sigma \in S_q, T_{p^\alpha}^\sigma \preceq \pi_1 \preceq \pi_2 \Rightarrow$

$L_\pi(T_{p^\alpha}^\sigma|X) \leq L_\pi(\pi_1|X) \leq L_\pi(\pi_2|X)$
Possibilistic loss function and entropy: continuous case

- Direct extension of the discrete case based on $\alpha$-cuts
- Poss-log loss:

$$\mathcal{L}_{\pi-l}(\pi|x) = -\int_{\mathbb{R}} \log(1 - \pi(x)/2)\,dx$$

$$- 0.5 \cdot \int_{A_{\pi_x}} \log(\pi(x)/2) - \log(1 - \pi(x)/2)\,dx$$

- Poss-squared loss:

$$\mathcal{L}_{\pi-s}(\pi|x) = \int_{A_{\pi(x)}} \pi(t)\,dt - |A_{\pi(x)}| - \frac{1}{2} \int_{\mathbb{R}} \pi(t)^2\,dt$$

- Same properties than in the discrete case
Special cases

- **Gaussian distribution**
  - Confidence intervals obtained by mood regions
  - Analytic formulas for the possibility distribution that encodes such family
  - Poss-log loss has to be approximated

- **Triangular and trapezoidal possibility distributions**
  - Triangular distribution upper bound any unimodal distribution at a given confidence threshold
  - Poss-squared loss easy to compute.
Approximation with Poss-squared loss
### Entropy

- **Probabilistic case:**
  - The entropy is the loss function value of the frequency distribution (i.e. $H(p^\alpha) = L(p^\alpha | X)$)
  - Entropy is maximal for the uniform distribution
  - Entropy is minimal when all the data pertain to the same class

- **Expected properties in the possibilistic case**
  - The possibilistic entropy applied to probability possibility transformations respects the specificity order.
  - The possibilistic entropy increases when uncertainty around the considered probability distribution increases.
The possibility cumulative entropy is the entropy of a possibility distribution $\pi$ with respect to a probability distribution $p$

$$H_{\pi \rightarrow I}(p, \pi) =$$
$$= - \sum_{j=1}^{q} \frac{T_p^*(C_j)}{2} \cdot \log\left(\frac{\pi(C_j)}{2}\right) + \left(1 - \frac{T_p^*(C_j)}{2}\right) \cdot \log\left(1 - \frac{\pi(C_j)}{2}\right)$$

(1)
Given $X$ and his associated frequency distribution $p^\alpha$ we have

$$H_{\pi}(p^\alpha, \pi) = \frac{L_{\pi}(\pi|X)}{q*\log(q)}$$

### Specificity order

$$T_p^* \preceq T_{p'}^* \Rightarrow H_{\pi}(p, T_p^*) \leq H_{\pi}(p', T_{p'}^*)$$

### Increase with uncertainty

$$T_p^* \preceq \pi \preceq \pi' \Rightarrow H_{\pi}(p, T_p^*) \leq H_{\pi}(p, \pi) \leq H_{\pi}(p, \pi')$$
Given $p(c)$ estimated from $n$ pieces of data, we compute the upper bound $p^*_{\gamma,n}$ of the $(1 - \gamma)$% confidence interval with Agresti-Coull method.

Possibility distribution as an upper bound of the frequency distribution

$$\pi_{p,n}(C_j) = P^*_{\gamma,n}(\bigcup_{i=1}^{j} C_{\sigma(i)})$$

where $\sigma \in S_q$ follows the probability order.

Possibilistic entropy of a frequency distribution estimated from $n$ pieces of data:

$$\mathcal{H}^*_{\pi-1}(p, n, \gamma) = \mathcal{H}_{\pi-1}(p, \pi^\gamma_{p,n})$$
Example

- \( p(C_1) = 0.5, \ p(C_2) = 0.2 \) and \( p(C_3) = 0.3 \).
- \( n = 10 \ (\gamma = 0.05) \)
- \( \pi_{p,10}^{0.05}(C_1) = P_{0.05,10}^*(C_1 \cup C_2 \cup C_3) = 1 \)
- \( \pi_{p,10}^{0.05}(C_2) = P_{0.05,10}^*(C_2) = 0.52 \)
- \( \pi_{p,10}^{0.05}(C_3) = P_{0.05,10}^*(C_2 \cup C_3) = 0.76 \)
- \( \mathcal{H}^*_{\pi^{-1}}(p, 50, 0.05) = 0.38 \).
Possibilistic cumulative entropy for a limited set of data: properties

**Encoding**

\[ p \in \mathcal{P}(\pi_p, n) \]
\[ \forall n > 0, \pi^*_p \preceq \pi^\gamma_{p,n} \text{ and } \lim_{n \to \infty} \pi^\gamma_{p,n} = \pi^*_p \]

**Increases when uncertainty increases**

given \( n' \leq n \) we have

\[ \forall \gamma \in ]0, 1[, \mathcal{H}^*_\pi(p, n, \gamma) \leq \mathcal{H}^*_\pi(p, n', \gamma) \]

**Stability**

given \( p \) and \( p' \) we have

\[ \forall \gamma \in ]0, 1[, \ T^*_p \preceq T^*_{p'} \Rightarrow \mathcal{H}^*_\pi(p, n, \gamma) \leq \mathcal{H}^*_\pi(p', n, \gamma) \]
Possibilistic cumulative entropy

The graph shows the possibilistic cumulative entropy for different values of the parameter $n$. The entropy is plotted against the parameter $p$ for $n=1, 10, 2, 20, 5, 50$.
Upper entropy on credal sets

Without specificity term

With specificity term
Learning of decision trees is based on entropy of frequency distributions.

When we go deeper downward the tree, the examples become rarer and the faithfulness of entropy decreases.

Log entropy-based gain: splitting a node always decreases the weighted entropy of the leaves obtained.
Algorithm

- Principal: recursively choose the attribute that maximize the gain function
- Log gain function:

\[
G(Z, A) = \mathcal{H}(p_Z) - \sum_{k=1}^{r} \frac{|v_k|}{n} \mathcal{H}(p_{v_k})
\]

- Possibilistic gain function:

\[
G_\gamma(Z, A) = \mathcal{H}_\pi^{-1}(p_Z, n, \gamma) - \sum_{k=1}^{r} \frac{|v_k|}{n} \mathcal{H}_\pi^{-1}(p_{v_k}, |v_k|, DS(\gamma, r)).
\]
Advantages of the approach

- Significant choices of split
- Statistically relevant stopping criterion
- Reasonable estimator of the performances of a decision tree
- Provide well sized and well balanced trees
Online algorithm

Algorithm :
1. browse recursively the tree to the corresponding leaf
2. add $x$ to the set of examples
3. search the attribute with the best $G_\gamma^\pi$
4. if the gain is positive, create a new node with the corresponding attribute, else do nothing.

Advantages :
- Incremental algorithm
- Built new leaves only when the split gives a statistical significant gain
- Only consider the leaf concerned by the new example
Experiments: Entropy vs size

size vs log loss

size vs possibilistic log-loss
Experiments: entropy vs accuracy

Accuracy vs log-loss

Accuracy vs possibilistic log-loss
<table>
<thead>
<tr>
<th>Data set</th>
<th>Log Tree</th>
<th>PrunTree</th>
<th>ΠTree</th>
<th>O-ΠTree</th>
<th>J48</th>
</tr>
</thead>
<tbody>
<tr>
<td>soybean</td>
<td>89.4±5.0</td>
<td>89.4±5.0</td>
<td>94.0±2.8</td>
<td>89.0±3.8</td>
<td>91.7±3.1</td>
</tr>
<tr>
<td>lymph</td>
<td>72.9±11.8</td>
<td>72.9±11.8</td>
<td>78.3±7.9</td>
<td>78.3±8.2</td>
<td>75.8±11.0</td>
</tr>
<tr>
<td>zoo2</td>
<td>97.0±4.8</td>
<td>97.0±4.8</td>
<td>97.0±4.8</td>
<td>96.0±5.1</td>
<td>92.6±7.3</td>
</tr>
<tr>
<td>ilpd</td>
<td>67.9±5.5</td>
<td>67.4±5.6</td>
<td>69.9±5.3</td>
<td>66.8±4.7</td>
<td>68.1±5.6</td>
</tr>
<tr>
<td>yeast</td>
<td>52.0±4.1</td>
<td>57.0±3.3</td>
<td>57.1±3.4</td>
<td>56.7±3.6</td>
<td>56.6±3.7</td>
</tr>
<tr>
<td>waveform</td>
<td>75.2±1.5</td>
<td>75.3±1.5</td>
<td>77.4±1.5</td>
<td>72.6±1.8</td>
<td>75.2±1.9</td>
</tr>
<tr>
<td>diabetes</td>
<td>68.7±5.7</td>
<td>70.4±4.7</td>
<td>74.3±4.4</td>
<td>70.4±3.4</td>
<td>74.4±5.2</td>
</tr>
<tr>
<td>banknote</td>
<td>98.3±1.1</td>
<td>98.3±1.1</td>
<td>98.3±1.0</td>
<td>97.4±2.1</td>
<td>98.5±1.0</td>
</tr>
<tr>
<td>ecoli</td>
<td>78.9±7.7</td>
<td>80.4±7.4</td>
<td>82.4±7.9</td>
<td>83.6±7.2</td>
<td>82.8±5.7</td>
</tr>
<tr>
<td>vehicle</td>
<td>71.6±4.7</td>
<td>71.6±4.0</td>
<td>74.1±4.1</td>
<td>69.1±3.1</td>
<td>72.2±4.3</td>
</tr>
<tr>
<td>ionosphere</td>
<td>90.3±4.7</td>
<td>90.3±4.7</td>
<td>91.1±3.6</td>
<td>87.7±4.0</td>
<td>89.7±4.3</td>
</tr>
<tr>
<td>segment</td>
<td>96.8±0.6</td>
<td>96.7±0.7</td>
<td>96.9±1.2</td>
<td>94.7±1.4</td>
<td>96.7±1.2</td>
</tr>
<tr>
<td>pendigits</td>
<td>96.5±0.5</td>
<td>96.4±0.5</td>
<td>96.4±0.2</td>
<td>93.2±1.0</td>
<td>96.5±0.6</td>
</tr>
<tr>
<td>spambase</td>
<td>91.8±1.2</td>
<td>91.7±1.3</td>
<td>94.0±1.3</td>
<td>90.5±1.2</td>
<td>92.8±1.0</td>
</tr>
<tr>
<td>breast-wv2</td>
<td>92.9±2.4</td>
<td>92.9±2.4</td>
<td>93.9±3.1</td>
<td>94.7±1.6</td>
<td>94.1±2.5</td>
</tr>
<tr>
<td>wine2</td>
<td>92.5±8.7</td>
<td>92.5±8.7</td>
<td>93.7±7.3</td>
<td>94.3±8.3</td>
<td>93.2±5.9</td>
</tr>
</tbody>
</table>
Size comparison

Number leaves of LogTree, PrunTree, ΠTree and O-ΠTree, J4.8 comparison for different databases.
Regression trees

- Same approach than for decision trees with poss-log los function
- Use possibilistic cumulative entropy of the possibility distribution that encodes the family of Gaussian distribution that has parameters inside the confidence interval based on mood confidence region.
- Online algorithm also works
- Promising results
Conclusions

- Possibility loss functions and entropies
  - Agrees with the probabilistic view of possibility theory
  - Reflects both the entropy of a probability distribution and the uncertainty around the parameters
  - Can be used for upper estimate densities without assuming a particular shape

- Application to decision and regression trees:
  - Provides well balanced and well sized trees
  - Avoids over-fitting
  - Simple and efficient online algorithm

- Easy extension to:
  - Bayesian networks
  - Density estimation
  - Bandwidth selection in knn