LEARNING FROM IMPRECISE AND FUZZY DATA:
ON THE NOTION OF DATA DISAMBIGUATION

Eyke Hüllermeier
Computational Intelligence Group
Department of Mathematics and Computer Science
Marburg University, Germany
SUPERVISED LEARNING: Algorithms and methods for discovering (alleged) dependencies and regularities in a domain of interest, expressed through appropriate models, from specific observations or examples.
FUZZY MACHINE LEARNING: Learning FUZZY MODELS from CRISP DATA!

DATA

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.34</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>0.43</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>0.82</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>0.93</td>
<td>0</td>
</tr>
<tr>
<td>22</td>
<td>0.72</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>0.82</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>0.62</td>
<td>1</td>
</tr>
</tbody>
</table>

FUZZY RULES

IF x1=high AND x2=low THEN Y=0
IF x1=low AND x2=low THEN Y=1
IF x1=high AND x2=high THEN Y=1
IF x1=low AND x2=high THEN Y=0
## Learning From Fuzzy Data

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.42</td>
<td>0</td>
<td>132</td>
<td>10.5</td>
</tr>
<tr>
<td>12</td>
<td>0.90</td>
<td>1</td>
<td>154</td>
<td>32.6</td>
</tr>
<tr>
<td>17</td>
<td>0.61</td>
<td>1</td>
<td>211</td>
<td>55.2</td>
</tr>
<tr>
<td>11</td>
<td>0.17</td>
<td>1</td>
<td>423</td>
<td>94.2</td>
</tr>
<tr>
<td>28</td>
<td>0.66</td>
<td>0</td>
<td>654</td>
<td>12.6</td>
</tr>
<tr>
<td>19</td>
<td>0.93</td>
<td>0</td>
<td>127</td>
<td>37.4</td>
</tr>
<tr>
<td>32</td>
<td>0.72</td>
<td>1</td>
<td>336</td>
<td>33.8</td>
</tr>
<tr>
<td>15</td>
<td>0.12</td>
<td>0</td>
<td>798</td>
<td>62.5</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
LEARNING FROM FUZZY DATA

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.42</td>
<td>0</td>
<td>132</td>
<td>10.5</td>
</tr>
<tr>
<td>12</td>
<td>0.90</td>
<td>1</td>
<td>154</td>
<td>16.5</td>
</tr>
<tr>
<td>▲</td>
<td>0.61</td>
<td>1</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td>11</td>
<td>▲</td>
<td>1</td>
<td>▲</td>
<td>94.2</td>
</tr>
<tr>
<td>▲</td>
<td>0.66</td>
<td>0</td>
<td>654</td>
<td>12.6</td>
</tr>
<tr>
<td>19</td>
<td>▲</td>
<td>0</td>
<td>127</td>
<td>▲</td>
</tr>
<tr>
<td>32</td>
<td>0.72</td>
<td>1</td>
<td>▲</td>
<td>▲</td>
</tr>
<tr>
<td>15</td>
<td>0.12</td>
<td>0</td>
<td>▲</td>
<td>62.5</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

HOW TO ANALYZE AND LEARN FROM SUCH DATA?
The „ontic“ view (conjunctive interpretation):
- a fuzzy set is a real data entity;
- an attribute can assume a fuzzy set as a „value“, i.e.,
- we have a (fuzzy set)-valued attribute.

EXAMPLE: Duration of sunshine in Vilamoura today.
THE ONTIC VIEW: REMARKS

- In line with the general trend of analyzing "complex" data (e.g., interval-valued, histogram-valued, functional, etc.)

- Questionable relevance for machine learning/data analysis:
  - A systematic collection of fuzzy data of that kind requires a suitable "measurement device" producing fuzzy sets (membership functions).
  - What is the meaning of a membership degree, if not related to frequency (and hence probability distributions)?
The „epistemic“ view (disjunctive interpretation):

- The true value of the attribute is precise, and a fuzzy set is used to express imprecise knowledge about this value (possibility distribution).

A FUZZY SET IS NOT THE DATA OBJECT, BUT REPRESENTS KNOWLEDGE ABOUT THIS OBJECT!
TWO INTERPRETATIONS OF A FUZZY SET

Fuzzy set could be replaced by a precise value on the basis of additional knowledge.

A further „precisiation“ of the data is not legitimate.
THE TWO INTERPRETATIONS, ONTIC AND EPISTEMIC, CALL FOR VERY DIFFERENT EXTENSIONS OF METHODS FOR DATA ANALYSIS!
The ontic view essentially calls for „lifting“ a method to a complex data space, in which data entities are fuzzy sets, and to extend the underlying operations correspondingly.

„Fuzzy observations“ are embedded as points in a (high-dimensional) space (e.g., a fuzzy metric space).

Special case of structured output prediction, for which kernel-based learning methods are quite popular (kernels for sequences, graphs, etc.)
REGRESSION WITH INTERVAL OUTPUTS

interval-valued observation
Ontic view: Reproducing interval observations by means of an interval-valued function

\[ F^* \in \arg \min_{F \in \mathcal{F}} \sum_i D(Y_i, F(x_i)) \]
Epistemic view: Solution is a

- fuzzy set of REAL-VALUED regression functions
- instead of a single FUZZY SET-VALUED regression function
A model is deemed possible if there is a possible set of precise observations (a SELECTION) for which it is an optimal fit.

→ EXTENSION PRINCIPLE (applied to a data analysis method) ?
THE EXTENSION PRINCIPLE

- The extension principle generalizes a function

\[ f : X_1 \times X_2 \times \ldots \times X_n \rightarrow Y \]

from "crisp" to fuzzy inputs:

\[ f(x_1, x_2, \ldots, x_n) = y \]

\[ \text{?} \]

\[ F(A_1, A_2, \ldots, A_n) = Y \]

\[ \mu_Y(y) = \sup_{x=(x_1, \ldots, x_n)} \left\{ \min(A_1(x_1), \ldots, A_n(x_n)) \mid f(x) = y \right\} \]
For example, interval arithmetics: $[1, 5] \oplus [1, 3] = [-2, 4]$

All selections of (single-valued) input values are treated the same and equally contribute to the output!
THE EXTENSION PRINCIPLE

- A learning algorithm is a **mapping from data to models**:

  \[ f : D^n \rightarrow M, \ d = (d_1, \ldots, d_n) \mapsto M \]

- So, the extension now reads as follows:

  \[ F(D)(M) = F(D_1, \ldots, D_n)(M) = \sup_{\mathbf{d} = (d_1, \ldots, d_n)} \left\{ \min_i D_i(d_i) \mid f(\mathbf{d}) = M \right\} \]

- Thus, a model is plausible insofar there is a plausible selection of precise data points supporting that model:

  \[ \pi(M) = \sup_{\mathbf{d}} \left\{ \mu_D(\mathbf{d}) \mid f(\mathbf{d}) = M \right\} \]

  with \( \mu_D(\mathbf{d}) = \min(D_1(d_1), \ldots, D_n(d_n)) \).
THE EXTENSION PRINCIPLE

Questioning the equal treatment of all selections ...

In data analysis, a method inducing a model from a set of data always comes with certain MODEL ASSUMPTIONS, and under these assumptions, specific selections may appear more plausible than others!

... to be explained through some simple examples.
A plausible selection that can be fitted quite well with a **LINEAR** model!

A less plausible selection, because there is no **LINEAR** model with a good fit!
Adding **non-informative data** will have an influence on the plausibility of models!
Adding non-informative data will have an influence on the plausibility of models!
A single (imprecise) observation doesn’t tell us very much ...
... and neither does a set of them.
DATA DISAMBIGUATION

Yet, when looking at the data AS A WHOLE, and taking into account the RELATION BETWEEN THEM, some possible values become impossible!

→ constraint propagation
Yet, when looking at the data **AS A WHOLE**, and taking into account the **RELATION BETWEEN THEM**, some possible values become impossible!

→ constraint propagation
Yet, when looking at the data **AS A WHOLE**, and taking into account the **RELATION BETWEEN THEM**, some possible values become impossible!

→ **constraint propagation**
Imprecise x-values modeled as intervals.
Imprecise x-values modeled as intervals.
Scenario “red” more likely than “blue”!

The red scenario (two clusters) appears to be more plausible than the blue one (three clusters)!
The class of the red training points is not known.
This scenario allows one to fit a very simple decision tree, while other scenarios call for more complex models.
This scenario allows one to fit a very simple decision tree, while other scenarios call for more complex models.

Looking at the data from the point of view of a decision tree learner, the former scenario appears more likely than the latter.
The same does not necessarily hold under different model assumptions!

\[ \text{Data Disambiguation in Classification} \]

It all depends on how you look at the data!
Under the epistemic view, model identification and data disambiguation should be performed simultaneously:

The loss minimization approach ...
Many (supervised) learning methods are based on minimization of the **empirical risk**

\[
R_{emp}(M) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, M(x_i))
\]

or a regularized version thereof:

\[
R_{reg}(M) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, M(x_i)) + \lambda C(M)
\]

- **average loss on training data**
- **complexity term to prevent overfitting**
Consider an imprecise observation \((x, Y)\) and let \(\hat{y} = M(x)\).
How much should \(M\) be penalized for this prediction?

In agreement with the idea of data disambiguation, we look at the smallest possible loss, namely

\[
L^*(Y, \hat{y}) = \min \left\{ L(y, \hat{y}) \mid y \in Y \right\},
\]

and the value for which it is obtained:

\[
y^* = \arg \min \left\{ L(y, \hat{y}) \mid y \in Y \right\}.
\]

Given the model \(M\), this value appears to be the most plausible in \(Y\).
On the basis of the generalized loss function $L^*$, we define

$$
R_{emp}(M) = \frac{1}{N} \sum_{i=1}^{N} L^*(Y_i, M(x_i))
$$

how well the "crisp" model fits the imprecise data.
Note similarity to eps-insensitive loss in support vector regression
GENERALIZATION TO THE CASE OF FUZZY DATA

\[ \mathcal{L}(Y, \hat{y}) = \int_{0}^{1} L^*([Y]_{\alpha}, \hat{y}) \, d\alpha \]

\[ \bar{R}_{emp}(M) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(Y_i, M(\mathbf{x}_i)) \]
FUZZIFICATION OF L1 LOSS

→ close connection to Huber loss!
FUZZIFICATION OF L1 LOSS

→ overestimation is worse than underestimation!
Let $\mathcal{Y} = \{-1, +1\}$ and consider a class of scoring classifiers $M : \mathcal{X} \rightarrow \mathbb{R}$. A margin loss is a function of the form

$$L(y, s) = f(y s),$$

where $f : \mathbb{R} \rightarrow \mathbb{R}$ is a non-increasing function.
Let \( \mathcal{Y} = \{-1, +1\} \) and consider a class of scoring classifiers \( M : \mathcal{X} \rightarrow \mathbb{R} \).

A margin loss is a function of the form

\[
L(y, s) = f(ys),
\]

where \( f : \mathbb{R} \rightarrow \mathbb{R} \) is a non-increasing function.

**Hinge loss:**

\[
L(y, s) = \max (1 - ys, 0)
\]

**Log-loss:**

\[
L(y, s) = \log \left( 1 + \exp \left( -ys \right) \right)
\]

**Exponential loss:**

\[
L(y, s) = \exp \left( -s \right)
\]
Suppose the output is a fuzzy subset $Y$ with membership degrees

$$
\mu_Y(\lambda) = \begin{cases} 
1 & \text{if } \lambda = y \\
1 - w & \text{if } \lambda = \bar{y}
\end{cases},
$$

where $y, \bar{y} \in \{-1, +1\}$ such that $y\bar{y} = -1$, and $w$ can be interpreted as a degree of confidence in $y$. 
FUZZY MARGIN LOSSES

Suppose the output is a fuzzy subset $Y$ with membership degrees

$$\mu_Y(\lambda) = \begin{cases} 
1 & \text{if } \lambda = y \\
1 - w & \text{if } \lambda = \bar{y}
\end{cases},$$

where $y, \bar{y} \in \{-1, +1\}$ such that $y\bar{y} = -1$, and $w$ can be interpreted as a degree of confidence in $y$.

Then, the fuzzy loss function is given by

$$\mathcal{L}(Y, s) = f_w(ys) = w \cdot f(ys) + (1 - w) \cdot f(|ys|).$$
FUZZY MARGIN LOSSES

HINGE LOSS
FUZZY MARGIN LOSSES

LOG-LOSS
Two classes, both normally distributed, sample size 200.
FIRST EXPERIMENT

- Class information was partly removed from the training instances.
- More specifically, each of the 200 instances was declared „unlabeled“ with a fixed probability $\gamma$.
- Thus, we are in a **semi-supervised setting**, in which approximately $200(1-\gamma)$ of the instances are labeled.
First Experiment

- Class information was partly removed from the training instances.
- More specifically, each of the 200 instances was declared "unlabeled" with a fixed probability $\gamma$.
- Thus, we are in a semi-supervised setting, in which approximately $200(1-\gamma)$ of the instances are labeled.

- In our approach, the unlabeled instances are considered as being labeled with the fuzzy set that assigns a membership degree of 1 to both the positive and the negative class.
- Then, a model is trained using the fuzzy log-loss.
- Standard logistic regression is used for comparison.
FIRST EXPERIMENT

classification error

probability of missing label

0.16
0.165
0.17
0.175
0.18
0.185
0.19
0.195
0.2

0
0.2
0.4
0.6
0.8
SECOND EXPERIMENT

- The label of each example is switched with a fixed probability $\gamma$.
- This noise level is supposed to be known, whereas for each individual training example, it is not known whether the observed label corresponds to the original one or has been switched.
SECOND EXPERIMENT

- The label of each example is switched with a fixed probability $\gamma$.
- This noise level is supposed to be known, whereas for each individual training example, it is not known whether the observed label corresponds to the original one or has been switched.
- We model the label information in terms of a fuzzy set with a membership degree of 1 to the observed and of $\gamma$ to the other label.
- Standard logistic regression simply uses the observed label information, which is the best it can do.
SECOND EXPERIMENT
LABEL RANKING

... mapping instances to **TOTAL ORDERS** over a fixed set of alternatives/labels:

\[(28, 0, 187, 0.4) \rightarrow \text{Green Party} \succ \text{SPD} \succ \text{CDU} \succ \text{FPD}\]

\text{instance } x \in \mathcal{X} \quad \text{(e.g., features of a person)}

\text{ranking of labels/alternatives}
\[
\mathcal{Y} = \{ y_1, y_2, \ldots, y_k \}
\]
\[
\mathcal{Y} = \{ A, B, C, \ldots \}\]
## Label Ranking: Training Data

### Training

<table>
<thead>
<tr>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>Preferences</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.34</td>
<td>0</td>
<td>10</td>
<td>174</td>
<td>(A \succ B, C \succ D)</td>
</tr>
<tr>
<td>1.45</td>
<td>0</td>
<td>32</td>
<td>277</td>
<td>(B \succ C)</td>
</tr>
<tr>
<td>1.22</td>
<td>1</td>
<td>46</td>
<td>421</td>
<td>(B \succ D, A \succ D, C \succ D, A \succ C)</td>
</tr>
<tr>
<td>0.74</td>
<td>1</td>
<td>25</td>
<td>165</td>
<td>(C \succ A, C \succ D, A \succ B)</td>
</tr>
<tr>
<td>0.95</td>
<td>1</td>
<td>72</td>
<td>273</td>
<td>(B \succ D, A \succ D)</td>
</tr>
<tr>
<td>1.04</td>
<td>0</td>
<td>33</td>
<td>158</td>
<td>(A \succ B, A \succ C)</td>
</tr>
</tbody>
</table>

Instances are associated with pairwise preferences between labels.

- rank of A between 1 and 2
- rank of B between 2 and 4
- rank of C between 2 and 4
- rank of D between 1 and 4
Performance in terms of Kendall’s tau on synthetic data: missing-at-random (above) and top-rank setting (below).

<table>
<thead>
<tr>
<th></th>
<th>complete ranking</th>
<th></th>
<th>30% missing labels</th>
<th></th>
<th>60% missing labels</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LWD</td>
<td>PL</td>
<td>LWD</td>
<td>PL</td>
<td>LWD</td>
<td>PL</td>
</tr>
<tr>
<td>authorship</td>
<td>.933±.016</td>
<td>.936±.015</td>
<td>.925±.018</td>
<td>.833±.030</td>
<td>.891±.021</td>
<td>.601±.054</td>
</tr>
<tr>
<td>glass</td>
<td>.840±.075</td>
<td>.841±.067</td>
<td>.819±.078</td>
<td>.669±.064</td>
<td>.721±.072</td>
<td>.395±.068</td>
</tr>
<tr>
<td>iris</td>
<td>.960±.036</td>
<td>.960±.036</td>
<td>.932±.051</td>
<td>.896±.069</td>
<td>.876±.068</td>
<td>.787±.111</td>
</tr>
<tr>
<td>pendigits</td>
<td>.940±.002</td>
<td>.939±.002</td>
<td>.924±.002</td>
<td>.770±.004</td>
<td>.709±.005</td>
<td>.434±.007</td>
</tr>
<tr>
<td>segment</td>
<td>.953±.006</td>
<td>.950±.005</td>
<td>.914±.009</td>
<td>.710±.013</td>
<td>.624±.020</td>
<td>.381±.020</td>
</tr>
<tr>
<td>vehicle</td>
<td>.853±.031</td>
<td>.859±.028</td>
<td>.836±.032</td>
<td>.753±.032</td>
<td>.767±.037</td>
<td>.520±.050</td>
</tr>
<tr>
<td>vowel</td>
<td>.876±.021</td>
<td>.851±.020</td>
<td>.821±.022</td>
<td>.612±.027</td>
<td>.536±.034</td>
<td>.327±.033</td>
</tr>
<tr>
<td>wine</td>
<td>.938±.050</td>
<td>.947±.047</td>
<td>.933±.054</td>
<td>.919±.059</td>
<td>.921±.062</td>
<td>.863±.094</td>
</tr>
</tbody>
</table>

58
**SUMMARY AND CONCLUSION**

- **Learning from fuzzy data** is gaining increasing attention.

- Different **interpretations** of fuzzy data exist and suggest different ways of extending machine learning and data analysis methods:
  
  - ontic interpretation → data reproduction
  - epistemic interpretation → **data disambiguation** (extension principle)

- We proposed a method based on **generalized (fuzzy) loss functions and risk minimization**: A fuzzy set properly „modulates“ the loss associated with an individual observation → **data modeling**

- Our framework covers several existing approaches as **special cases** (Huber loss, instance weighting, semi-supervised learning), but also supports the systematic development of **new methods**.

Preprint version: arXiv:1305.0698