Imprecise and uncertain labelling
a solution based on Mixture model and belief functions.

Etienne Côme
come@inrets.fr

Institut National de Recherche sur les Transport et leurs Sécurité
Directeur : Patrice Aknin, Encadrant : Latifa Oukhellou
Université de Technologie de Compiègne
Directeur : Thierry Denœux

19 juin 2008

Introduction

Constatation

Supervised / Unsupervised frameworks
$\longrightarrow$ do not correspond with situations encountered in applications
Intermediary situations, Examples :

- few labelled samples and a lot of unlabelled samples
- difficult labelling (experts), can't assign samples to unique class
- label noise, errors in the labels
- ...

Aim of the study

Building a method to deal with imprecise and uncertain labels, based on mixture models

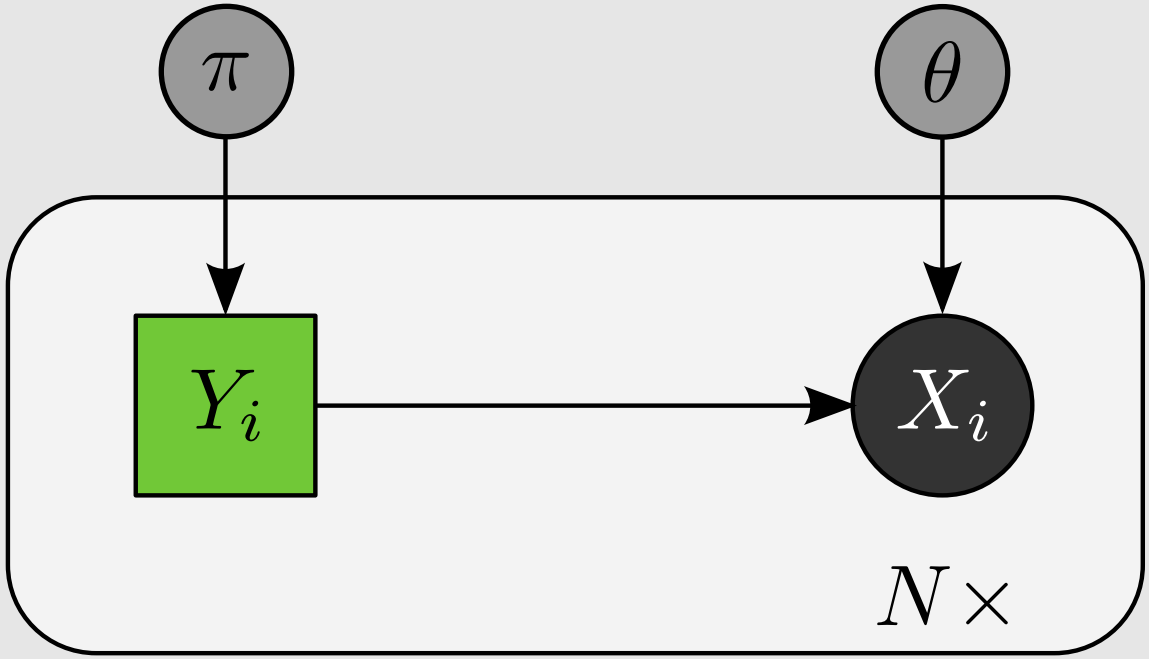
Soft labels examples

Set of classes : $\mathcal{Y} = \{c_1, ..., c_K\}$

	m_i : bba	pl_i : plausibility
Unsupervised (vacuous labels)	$m_i(\mathcal{Y}) = 1$	$pl_{ik} = 1, \quad \forall k$
Supervised (perfect labels)	$m_i(c_k) = 1$	$pl_{ik} = 1, pl_{ik'} = 0, \quad \forall k' \neq k$
Partially supervised (imprecise labels)	$m_i(C) = 1$	$\begin{cases} pl_{ik} = 1, & if \quad c_k \in C \\ pl_{ik} = 0, & if \quad c_k \notin C \end{cases}$
soft supervision (imprecise,uncertain)	m_i ?	$pl_{ik} \in [0, 1]$

Mixture model

Graphical model :

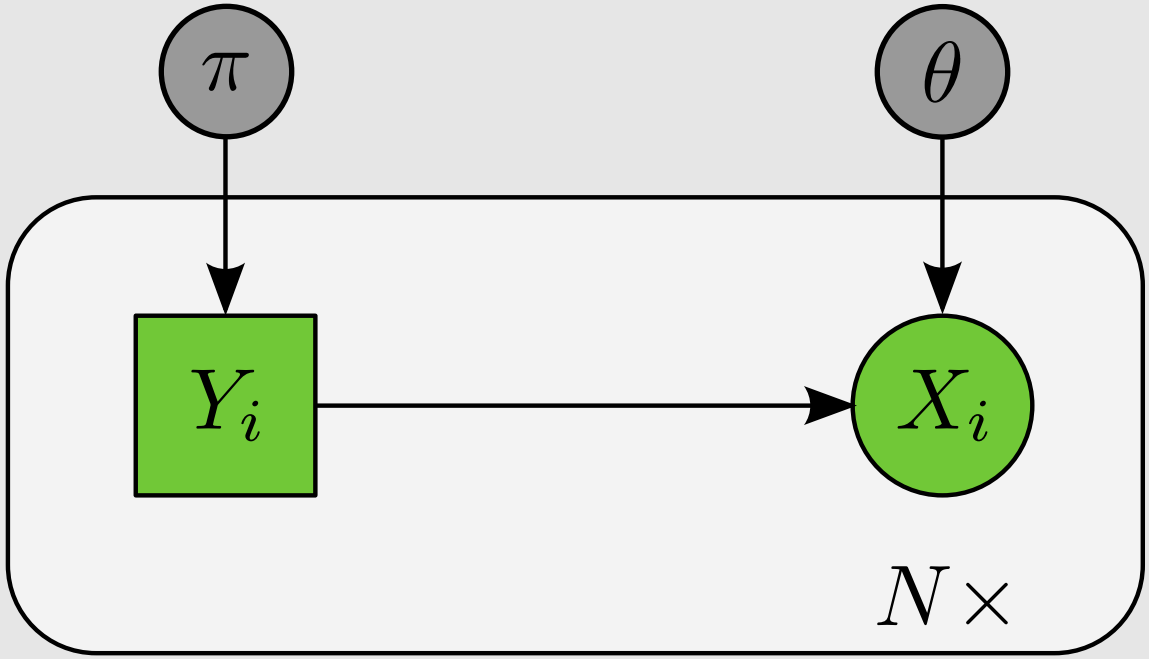


label sampling :

$$p(Y = c_k) \quad = \quad \pi_k, \quad \forall k \in \{1, \ldots, K\}. \tag{1}$$

Mixture model

Graphical model :



knowing the class, the observable variables are drawn :

$$p(\mathbf{x}|Y = c_k) = f(\mathbf{x}; \boldsymbol{\theta}_k), \quad \forall k \in \{1, \ldots, K\}. \tag{2}$$

Example : mixture of Gaussian (clustering)

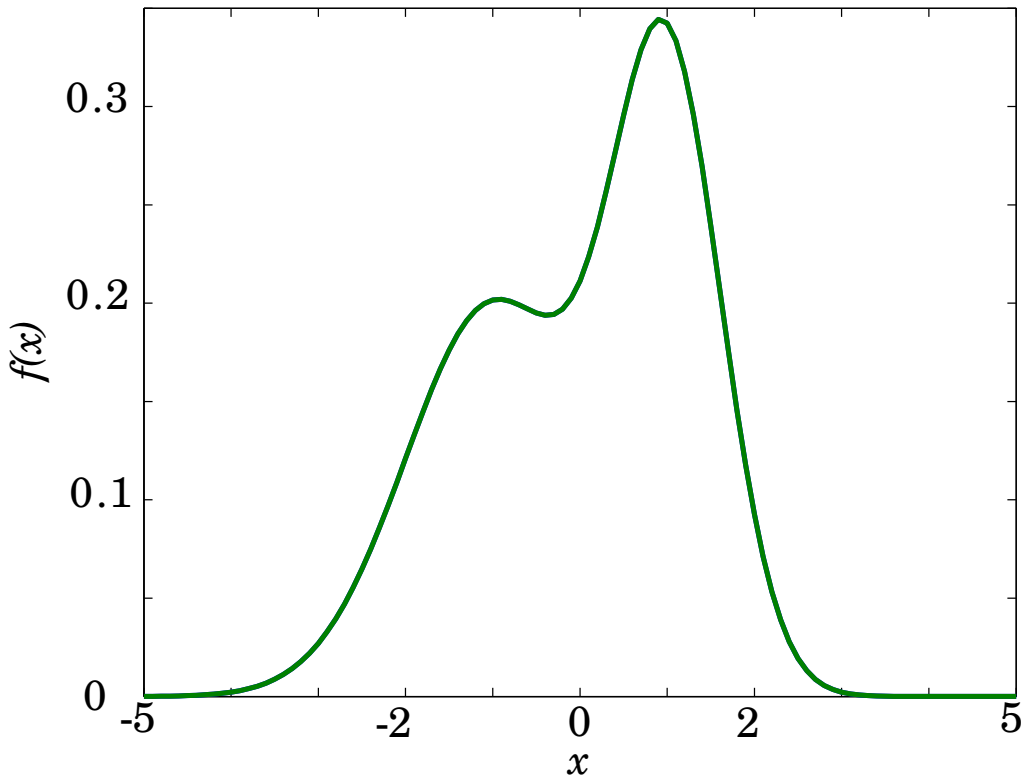


FIG.: Example of probability density function of a gaussian mixture

Example : Weibull mixture

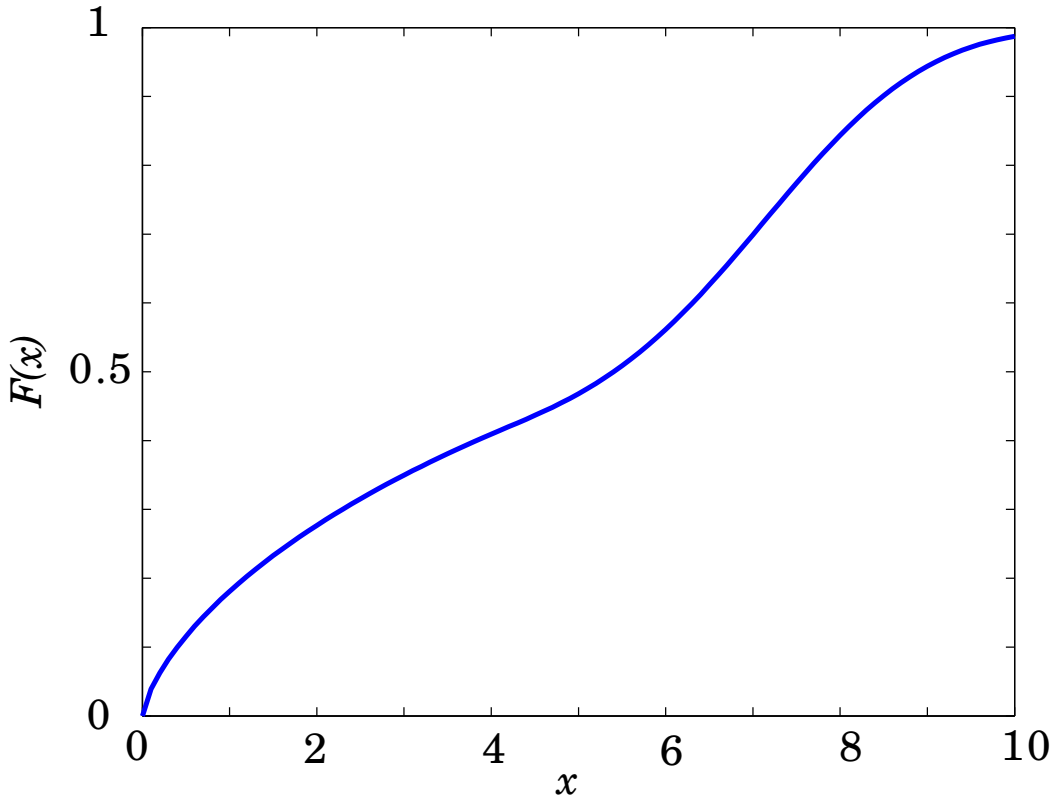


FIG.: Exemple of cumulative density function of a weibull mixture

Supervised learning with mixture models

Data

learning set : $\mathbf{X}^s = \{\mathbf{x}_i, y_i\}_{i=1}^N$,

- measurements $\mathbf{x}_i \in \mathcal{X}$,
- labels $y_i \in \mathcal{Y} = \{c_1, ..., c_K\}$,

Generative setting

parameters $\Psi = (\boldsymbol{\pi}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$, log-Likelihood :

$$L(\Psi; \mathbf{X}^s) = \sum_{i=1}^{M} \sum_{k=1}^{K} \mathbf{z}_{ik} log\left(\pi_k f(\mathbf{x}_i; \boldsymbol{\theta}_k)\right), \tag{3}$$

with $\mathbf{z}_i \in \{0,1\}^K$ binary variables encoding the class membership :
$z_{ik} = 1$ if $y_i = c_k$, and $z_{ik} = 0$ otherwise.
Analytical solutions if f is the density of a classical law.

Unsupervised learning, clustering

Data

learning set : $\mathbf{X}^{ns} = \{\mathbf{x}_i\}_{i=1}^{N}$

- measurements $\mathbf{x}_i \in \mathcal{X}$,

Goal, clustering

Find a good model of $p(\mathbf{x})$

with a latent discrete variable encoding the class membership y

$$p(\mathbf{x}) = \sum_y p(y)p(\mathbf{x}|y)$$

using bayes :

$$p(y|\mathbf{x}) = \frac{p(y)p(\mathbf{x}|y)}{\sum_y p(y)p(\mathbf{x}|y)}$$

Unsupervised learning, with mixture model

Mixture model

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}, \boldsymbol{\theta}_k) \quad (4)$$

$$L(\Psi; \mathbf{X}^{ns}) = \sum_{i=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k f_k(\mathbf{x}, \boldsymbol{\theta}_k)\right) \quad (5)$$

$f_k(., \boldsymbol{\theta}_k)$ parametric density over $\mathcal{X}$.
Parameters : $\Psi = (\boldsymbol{\pi}, \boldsymbol{\theta})$.

Optimisation !

! non-convexe

Classical solution : EM algorithm

EM algorithm

Log-likelihood :

latent variable z : $p(x) = \frac{p(x,z)}{p(z|x)}$
log likelihood decomposition :

$$L(\Psi; \mathbf{X}^{ns}) = \underbrace{\sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}^{(q)} \ln\left(\pi_k f(\mathbf{x}_i; \boldsymbol{\theta}_k)\right)}_{Q(\Psi,\Psi^{(q)})} - \underbrace{\sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}^{(q)} \ln\left(t_{ik}\right)}_{H(\Psi,\Psi^{(q)})}, \quad (6)$$

with :

$$t_{ik}^{(q)} = \mathbb{E}_{\Psi^{(q)}}[z_{ik}|\mathbf{x}_i] = \frac{\pi_k^{(q)} f(\mathbf{x}_i; \boldsymbol{\theta}_k^{(q)})}{\sum_{k'=1}^{K} \pi_{k'}^{(q)} f(\mathbf{x}_i; \boldsymbol{\theta}_{k'}^{(q)})}. \quad (7)$$

! H propertie

$H(\Psi^{(q)}, \Psi^{(q)}) - H(\Psi^{(q+1)}, \Psi^{(q)}) >= 0$

EM algorithm

$\Rightarrow$ maximisation of $Q(\Psi, \Psi^{(q)})$ sufficient to increase the likelihood.

EM algorithm

Initialisation : $\Psi^{(0)}$

until convergence

- **Expectation :** computation of $t_{ik}^{(q)}$
- **Maximisation :** $\Psi^{(q+1)} = \arg\max_{\boldsymbol{\Psi}} Q(\Psi, \Psi^{(q)})$
  analytical solution for the proportions : $\pi_k^{(q+1)} = \frac{1}{N} \sum_{i=1}^{N} t_{ik}^{(q)}$.
  analytical solution for θ_k for classical laws
  $\sum$ outside of the logarithm.

Semi-supervised learning

Data

A mix of labelled and unlabelled samples

$$\mathbf{X}^{ss} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_M, y_M), \mathbf{x}_{M+1}, \ldots, \mathbf{x}_N\}$$

for unlabelled samples be usefull hypothesis must be made :

- model of $p(x, y)$
  $\longrightarrow$ generative methods
  (Mixture model [Nigam 00])
- decision boundary in law density area
  $\longrightarrow$ discriminative methods
  (Data dependant regularization [Bengio 05])

Semi-supervised learning with mixture models

Generative setting

log-Likelihood :

$$L(\Psi, \mathbf{X}^{ss}) = \sum_{i=1}^{M}\sum_{k=1}^{K} \mathbf{z}_{ik} \ln(\pi_k f(\mathbf{x}_i; \boldsymbol{\theta}_k)) + \sum_{i=M+1}^{N} \ln(\sum_{k=1}^{K} \pi_k f(\mathbf{x}_i; \boldsymbol{\theta}_k)), \quad (8)$$

Optimisation

Modification of the EM algorithm :

- During the E step posterior probabilities are computed only for unlabelled samples.
- During the M step the known labels are used instead of the posterior probabilities for the labelled samples.

Partially supervised learning

Data

each label is a set of possible classes :

$$\mathbf{X}^{ps} = \{(\mathbf{x}_i, C_i)\}_{i=1}^{N}$$

where $C_i \subseteq \mathcal{Y}$ is the set of possible classes for sample i, (can be $\mathcal{Y}$)

$\Rightarrow$ more general :
semi-supervisised learning is a special case of partially supervised learning.

Solutions in the literature :

- Self consistent logistic regression [Grandvallet 02]
- Mixture model [Ambroise 00]

Partially supervised learning with mixture models

Generative setting

log-Likelihood :

$$L(\Psi, \mathbf{X}^{ps}) = \sum_{i=1}^{N} \ln \left(\sum_{k=1}^{K} l_{ik} \pi_k f(\mathbf{x}_i, \boldsymbol{\theta}_k) \right), \tag{9}$$

with $\mathbf{l}_i \in \{0,1\}^K$ binary variables encoding the possibles classes :
$l_{ik} = 1$ if $c_k \in C_i$, and $l_{ik} = 0$ otherwise.

Optimisation

Modification of the EM algorithm :
During the E step the posterior probabilities are computed using :

$$t_{ik}^{(q)} = \frac{l_{ik} \pi_k^{(q)} f(\mathbf{x}_i; \boldsymbol{\theta}_k^{(q)})}{\sum_{k'=1}^{K} l_{ik} \pi_{k'}^{(q)} f(\mathbf{x}_i; \boldsymbol{\theta}_{k'}^{(q)})}. \tag{10}$$

Learning with soft supervision

Data

labels = bba m_i over $\mathcal{Y}$:

$$\mathbf{X}^{sd} = \{(\mathbf{x}_i, m_i)\}_{i=1}^{N}$$

$\Rightarrow$ all other settings are special case of this one depending on kind of bba used

existent solutions in the belief functions framework

- k-nn [Denœux 95]
- belief decision tree [Elouedi 01]
- ...

Learning with soft supervision, mixture model and generative setting

[Vannoorenberghe 05]

Previous work : attempt to deal with soft label in mixture model, modification of an EM algorithm

Problems

- No criterion
- Convergence ?
- Relationship with probabilistic formulation unclear

The criterion

Starting point :

Relation between likelihood // possibility // plausibility
Natural criterion : parameters $\boldsymbol{\Psi}$ with biggest plausibility | observations

$$\hat{\Psi} = \arg\max_{\boldsymbol{\Psi}} pl^{\boldsymbol{\Psi}}(\Psi|\mathbf{X}^{sd})$$

$$\begin{array}{rcll} pl^{\boldsymbol{\Psi}}(\Psi|\mathbf{X}^{sd}) & = & pl^{\mathcal{X}_1\times\ldots\times\mathcal{X}_N}(\mathbf{x}_1,\ldots,\mathbf{x}_N|\Psi) & , \quad \text{(GBT)} \\ & = & \prod_{i=1}^{N} pl^{\mathcal{X}_i}(\mathbf{x}_i|\Psi) & , \quad \text{(Ind. cond.)} \end{array}$$

Development of $pl^{\mathcal{X}_i}(\mathbf{x}_i|\Psi)$:

$$pl^{\mathcal{X}_i}(\mathbf{x}_i|\Psi) = \sum_{C\subseteq\mathcal{Y}} m^{\mathcal{Y}_i}(C|\Psi)\, pl^{\mathcal{X}_i|\mathcal{Y}_i}(\mathbf{x}_i|C,\Psi) \quad , \quad \text{(Th. pl. totale)}$$

Need to calculate : $m^{\mathcal{Y}_i}(C|\Psi)$ and $pl^{\mathcal{X}_i|\mathcal{Y}_i}(\mathbf{x}_i|C,\Psi)$

The criterion

$m^{\mathcal{Y}_i}(C|\Psi)$, bba over classes

2 belief functions over $\mathcal{Y}_i$: π and m_i

Assumptions : labels are independent from parameters

$\Rightarrow$ conjunctive combination

! $\boldsymbol{\pi}$ is a bayesian bba

$\Rightarrow$ after combination focal sets = singletons :

$$\begin{aligned} m^{\mathcal{Y}_i}(c_k|\Psi) &= (m_i^{\mathcal{Y}_i} \textcircled{\cap} \boldsymbol{\pi}^{\mathcal{Y}_i})(c_k) \\ &= \sum_{C \cap c_k \neq \emptyset} m_i^{\mathcal{Y}_i}(C).\boldsymbol{\pi}^{\mathcal{Y}_i}(c_k) \\ &= pl_{ik}.\pi_k \end{aligned} \quad , \quad \forall k \in \{1, ..., K\}$$

bba over $\mathcal{Y}$:

$$m^{\mathcal{Y}_i}(c_k|\Psi) = pl_{ik}.\pi_k \tag{11}$$

The criterion

$pl^{\mathcal{X}_i|\mathcal{Y}_i}(\cdot|c_k, \Psi)$, plausibility of observations | class

From the model, $pl^{\mathcal{X}_i|\mathcal{Y}_i}(\cdot|c_k, \Psi)$ plausibility measure associated with $f(\mathbf{x}_i; \boldsymbol{\theta}_k)$

Problem

$\Rightarrow$ Plausibility of a point with infinite precision is null.

Solution

Punctual observations ?
Always a finite precision $\longrightarrow$ considers a small region around the point
Approximation, product of function with an infinitesimal $\mathrm{d}x_{i1} \ldots \mathrm{d}x_{ip}$:

$$pl^{\mathcal{X}_i|\mathcal{Y}_i}(\mathbf{x}_i|c_k, \Psi) = f(\mathbf{x}_i; \boldsymbol{\theta}_k)\mathrm{d}x_{i1} \ldots \mathrm{d}x_{ip}. \quad (12)$$

The criterion

Using all of this

$$pl^{\mathcal{X}_i}(\mathbf{x}_i|\Psi) = \left(\sum_{k=1}^{K} pl_{ik}\pi_k f(\mathbf{x}_i;\boldsymbol{\theta}_k)\right) \mathrm{d}x_{i1}\ldots\mathrm{d}x_{ip}. \quad (13)$$

Final criterion : generalized log-likelihood

$$L(\Psi, \mathbf{X}^{sd}) = \sum_{i=1}^{N} \ln\left(\sum_{k=1}^{K} pl_{ik}\pi_k f(\mathbf{x}_i,\boldsymbol{\theta}_k)\right) + Cst. \quad (14)$$

Remark

- natural expression ;
- uniquely depends of singletons plausibilities ;
- extends all the probabilist criterions (semi supervised, partially supervised ,...).

Optimisation

Information available on samples classes

at iteration q information on class labels of all samples comes from three sources :

1. label $m_i^{\mathcal{Y}}$;
2. proportions $\boldsymbol{\pi}^{(q)}$ (bayesian bba over $\mathcal{Y}$) ;
3. observation $\mathbf{x}_i$ and current parameters estimates $\boldsymbol{\theta}^{(q)}$ which lead to a bba over $\mathcal{Y}$ by the GBT : $pl^{\mathcal{Y}_i|\mathcal{X}_i}(\{c_k\}|\mathbf{x}_i, \Psi) = f(\mathbf{x}_i; \boldsymbol{\theta}_k)\mathrm{d}x_{i1} \ldots \mathrm{d}x_{ip}$.

t_{ik} : conjunctive combination of the three sources

$$t_{ik}^{(q)} = \frac{pl_{ik}\pi_k f(\mathbf{x}_i; \boldsymbol{\theta}^{(q)})}{\sum_{k'=1}^{K} pl_{ik'}\pi_{k'} f(\mathbf{x}_i; \boldsymbol{\theta}^{(q)})},$$

Optimisation

log likelihood decomposition :

Classical as in EM :

$$L(\Psi; \mathbf{X}^{ns}) = \underbrace{\sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}^{(q)} \ln\left(pl_{ik}\pi_k f(\mathbf{x}_i;\boldsymbol{\theta}_k)\right)}_{Q(\Psi,\Psi^{(q)})} - \underbrace{\sum_{i=1}^{N}\sum_{k=1}^{K} t_{ik}^{(q)} \ln\left(t_{ik}\right)}_{H(\Psi,\Psi^{(q)})},$$

with : $t_{ik}^{(q)} = \frac{pl_{ik}\pi_k f(\mathbf{x}_i;\boldsymbol{\theta}^{(q)})}{\sum_{k'=1}^{K} pl_{ik'}\pi_{k'} f(\mathbf{x}_i;\boldsymbol{\theta}^{(q)})}$

Increasing

H same form as in classical EM.

$\Rightarrow$ Optimisation of Q sufficient.

$\Rightarrow$ Same algorithm except for the E step where the t_{ik} are computed using the soft labels

Optimisation

An EM algorithm

Initialisation : $\Psi^{(0)}$

While :

- **Expectation :** $t_{ik}^{(q)} = \frac{pl_{ik}\pi_k f(\mathbf{x}_i;\boldsymbol{\theta}^{(q)})}{\sum_{k'=1}^{K} pl_{ik'}\pi_{k'} f(\mathbf{x}_i;\boldsymbol{\theta}^{(q)})}$,
- **Maximisation :** $\Psi^{(q+1)} = \arg\max_{\boldsymbol{\Psi}} Q(\Psi, \Psi^{(q)})$

  Analytical solution for the proportions : $\pi_k^{(q+1)} = \frac{1}{N}\sum_{i=1}^{N} t_{ik}^{(q)}$.

  Analytical solution for the θ_k if classical laws.

Convergence

Experimentation (1) : from unsupervised to supervised learning

from unsupervised to supervised learning

Simulation of different sets of labels with varying label precision
More precise labels :

- Better results ?
- Simpler optimisation problem ?

Information measure : non specificity

Average non specificity μ_s of learning set labels :

$$NS(m_i) = \sum_{C \subseteq \mathcal{Y}} m_i^{\mathcal{Y}}(C) log(card(C))$$

between :

$$\begin{array}{ccc} 0 & \longrightarrow & ln(card(\mathcal{Y})) \\ \text{perfect labels} & \longrightarrow & \text{vacuous labels} \\ \text{supervised learning} & \longrightarrow & \text{unsupervised learning} \end{array}$$

Labels simulations :

- for each sample i a non-specificity $NS(m_i)$ is drawn, with a uniform law in $[\mu_s - 0.05, \mu_s + 0.05]$
- a plausibility pl_i with $NS(m_i)$ as non-specificity is build.
- each plausibility value is assigned to a class with the following rules :
  - the true class has the biggest plausibility
  - random assignation for the other classes

$\Rightarrow$ No error in the labelling
(Assumption removed in the second experiment)

Data simulation

Mixture of 2 Gaussian : $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_1 = I$, $\pi_1 = \pi_2 = 0.5$, dimension 10 ;
Distance between the two centres $\delta \in \{1, 2, 4\}$;
Learning set size $N = 1000$.

Influence of labelling on results quality, $\delta = 1$

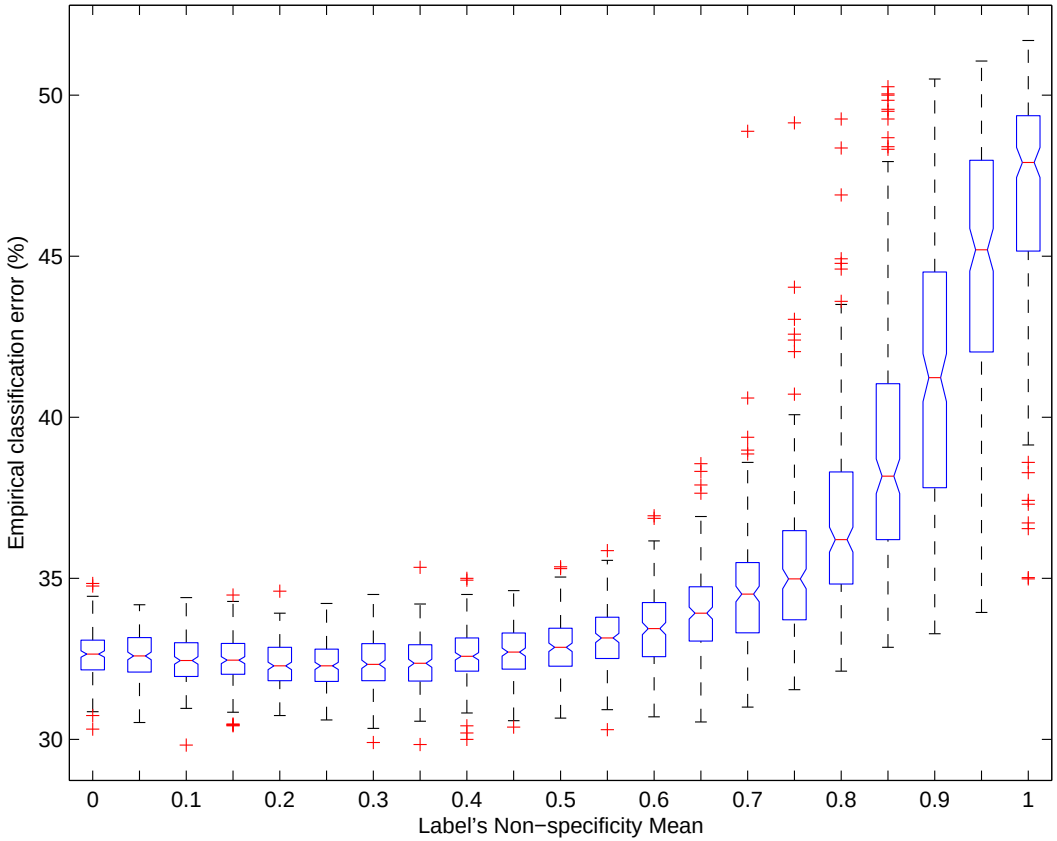


FIG.: Classification error on test set / labels non-specificity : $\delta = 1, \quad N = 1000$

Influence of labelling on results quality, $\delta = 2$

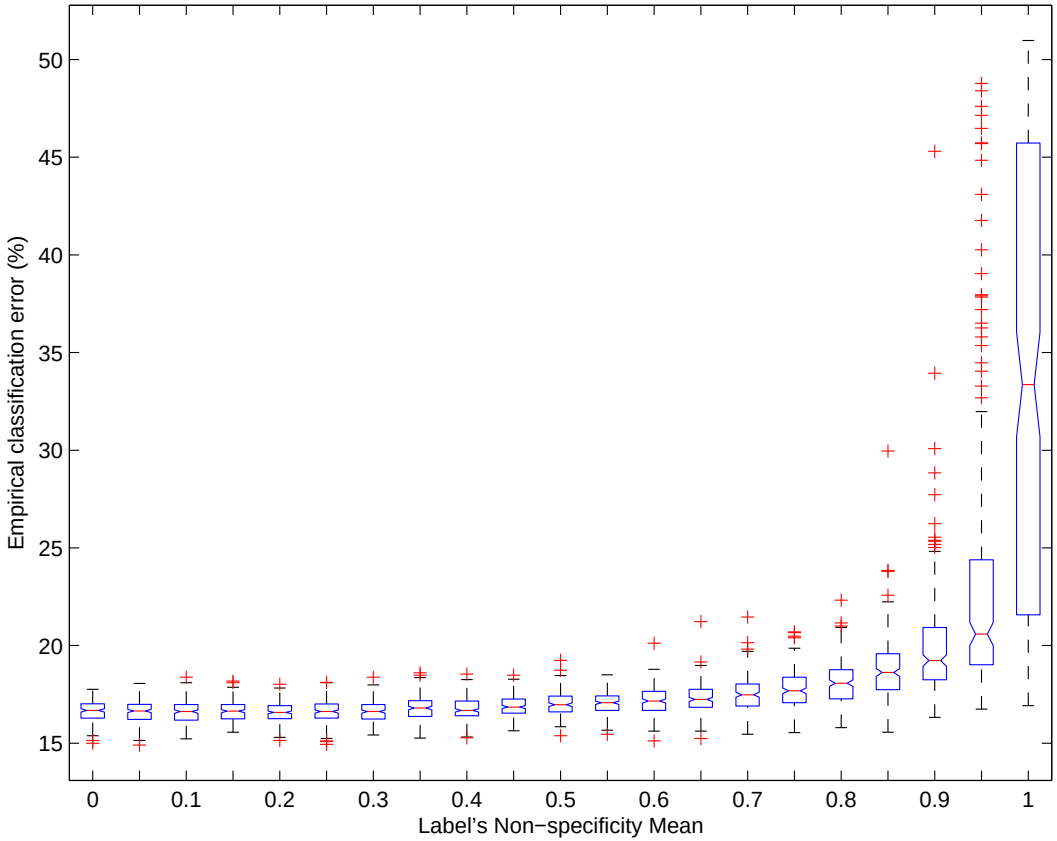


FIG.: Classification error on test set / labels non-specificity : $\delta = 2 \quad N = 1000$

Influence of labelling on results quality, $\delta = 4$

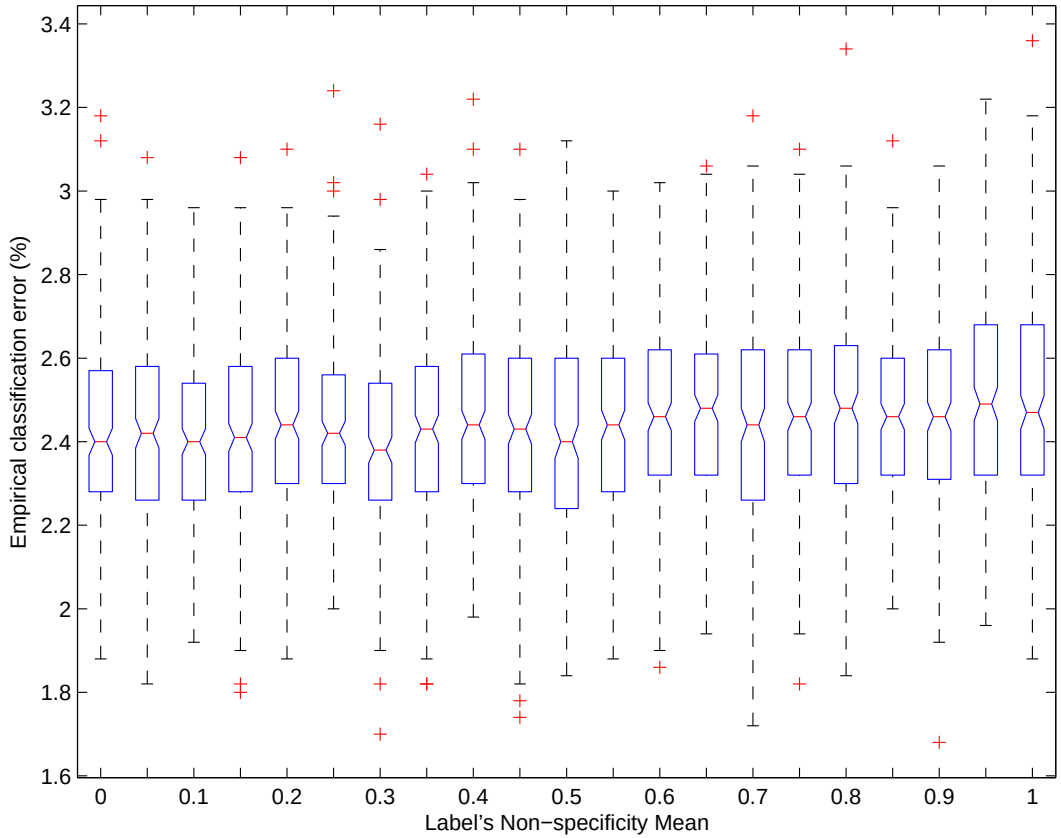
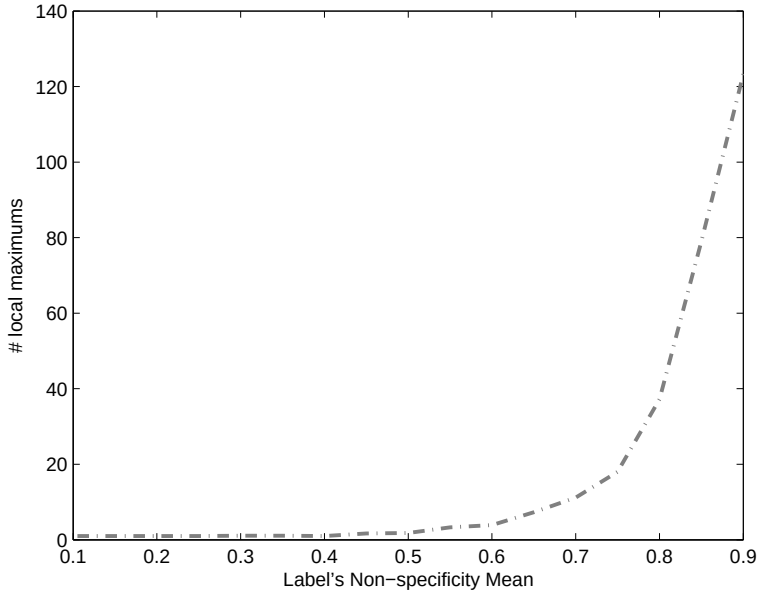


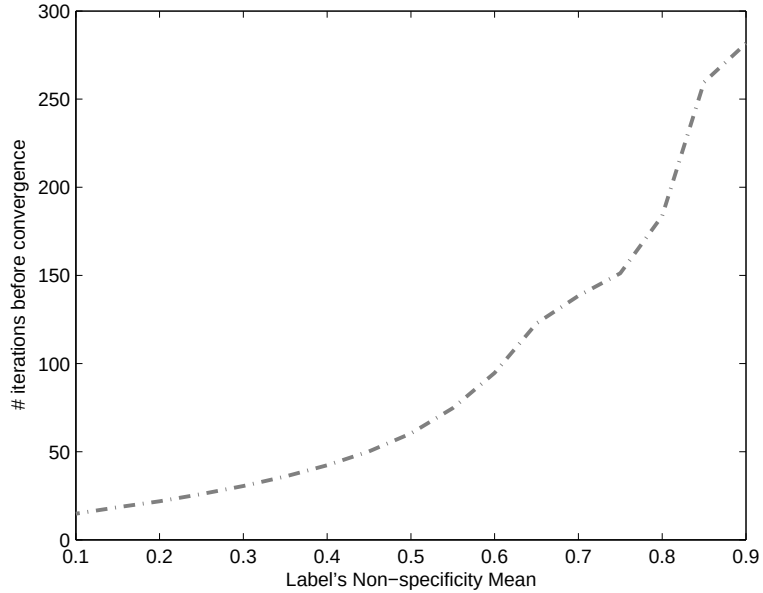
FIG.: Classification error on test set / labels non-specificity $\delta = 4$, $N = 1000$

Influence of labels precision on optimisation

$$\delta = 2, N = 1000$$



(a) # Locals maximums



(b) # iterations before convergence

Conclusions on experiment (1)

More precise labels :

- Better classification performance if the classes overlap
- Simplify the optimisation problem
  - less local maxima
  - faster EM convergence

Experiment (2) : Label noise

Context :

- Labelling done by an expert.
- Some mistake are made during the labeling process.
- But a value, $p_i \in [0, 1]$ = expert's doubt, is supplied.

How to deal with such additional information ?

Solution with "soft" labels

Discounting the hard label by p_i :

$$\left\{ \begin{array}{lcl} m_i(C_k) & = & 1 - p_i \\ m_i(\mathcal{Y}) & = & p_i \end{array} \right. \Leftrightarrow \left\{ \begin{array}{lcll} pl_{ik} & = & 1 & \\ pl_{ik'} & = & p_i, & \forall k' \neq k \end{array} \right. \quad (15)$$

Simulations

- Data = benchmark classification problem (UCI),
- Expert's doubt p_i simulated with a *Beta* law,
- With a probability p_i, the true label is flipped.

Rmq : Expectation of p_i = asymptotic labelling error rate.

Comparison :

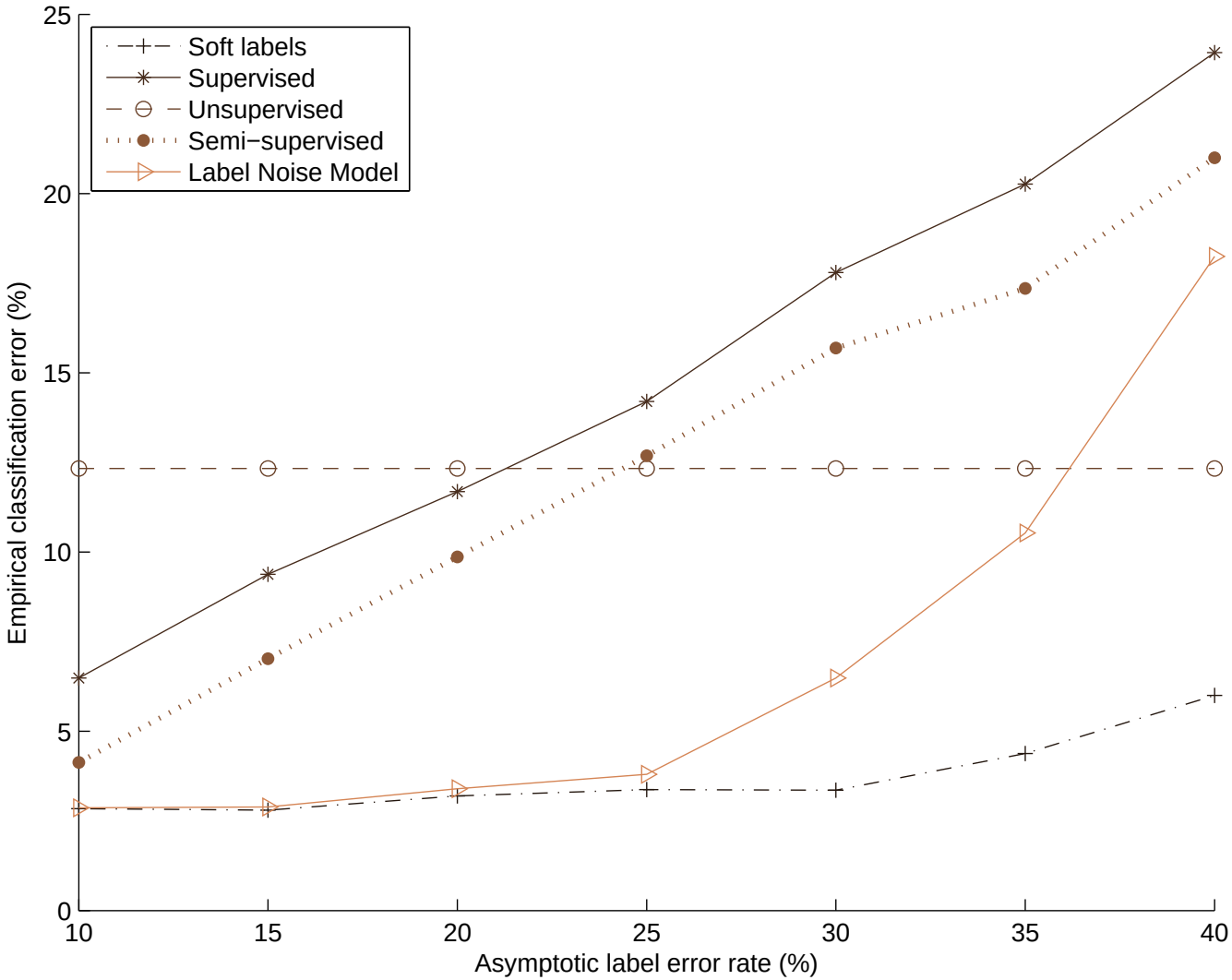
Supervised learning, Unsupervised learning, adaptative Semi-supervised learning, label noise model.
Performance Criterion : classification error

Parameters

- Expectation of *Beta* law : $(0.1 \rightarrow 0.4) = (\%$ false labels $)$
- Variance of *Beta* law keep constant at 0.2

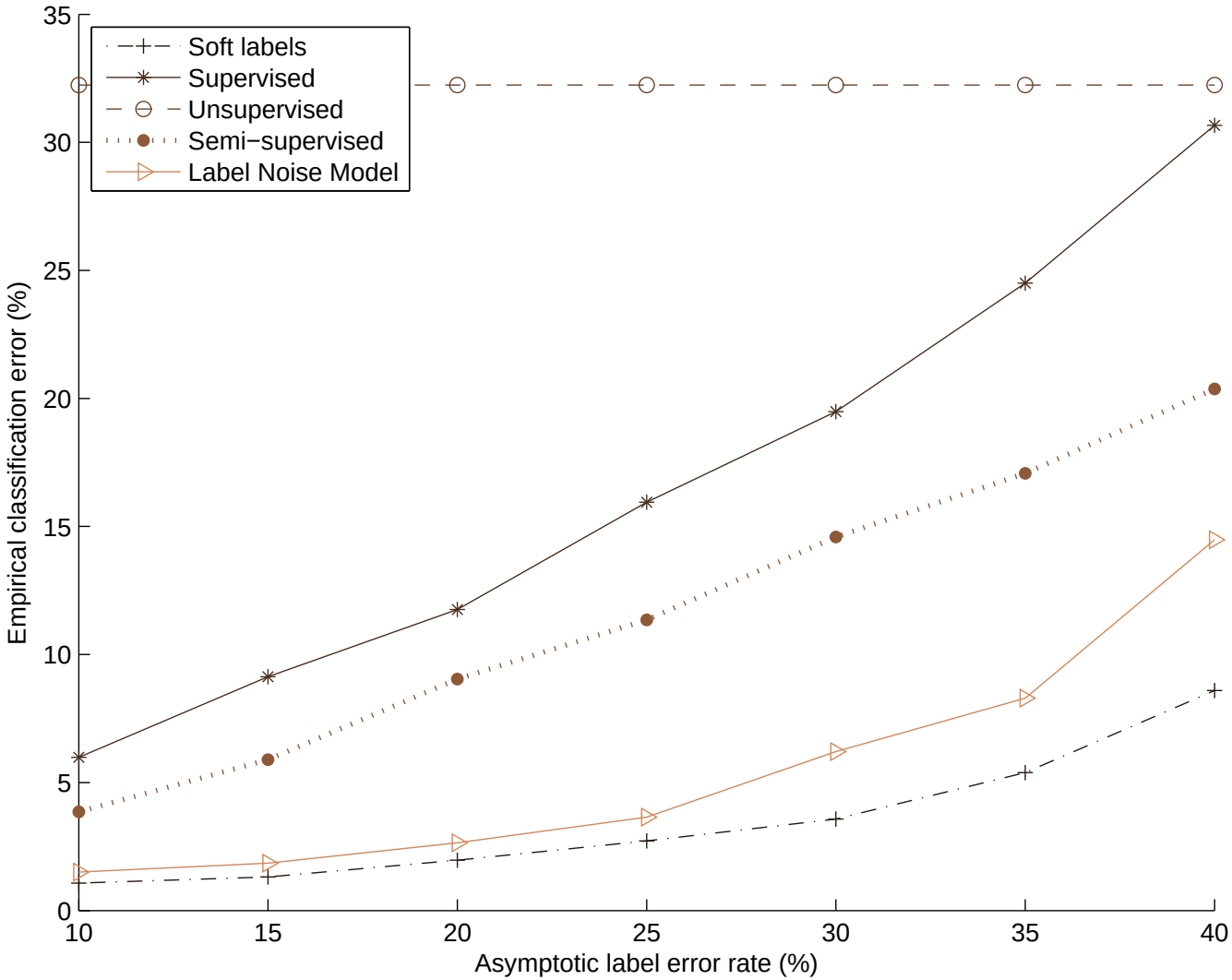
Results are averaged over 100 simulated learning set and performances are assessed by cross-validation (10 block)

Results *Iris*



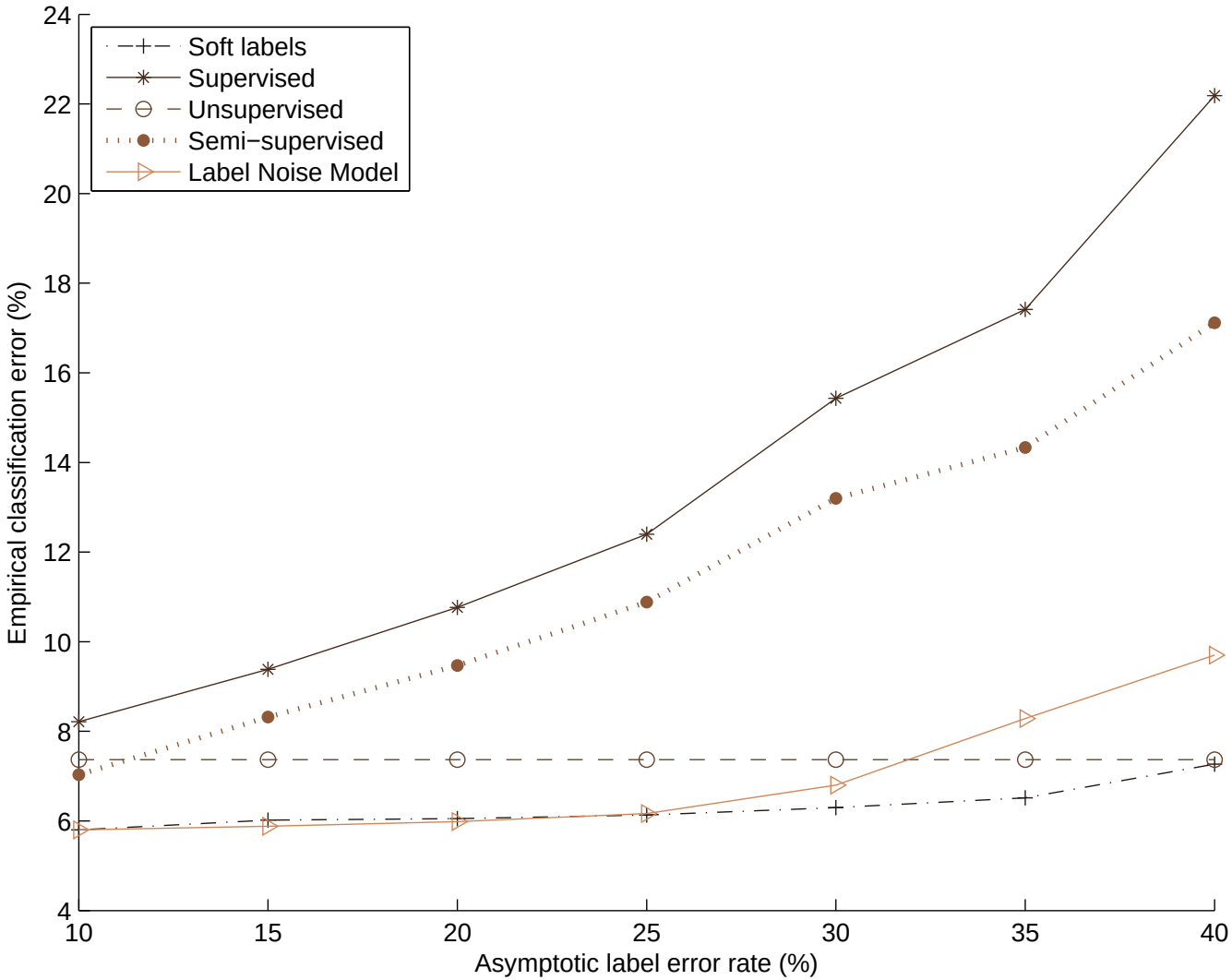
Classification error / Expert's labelling error rate

Results *Wine*



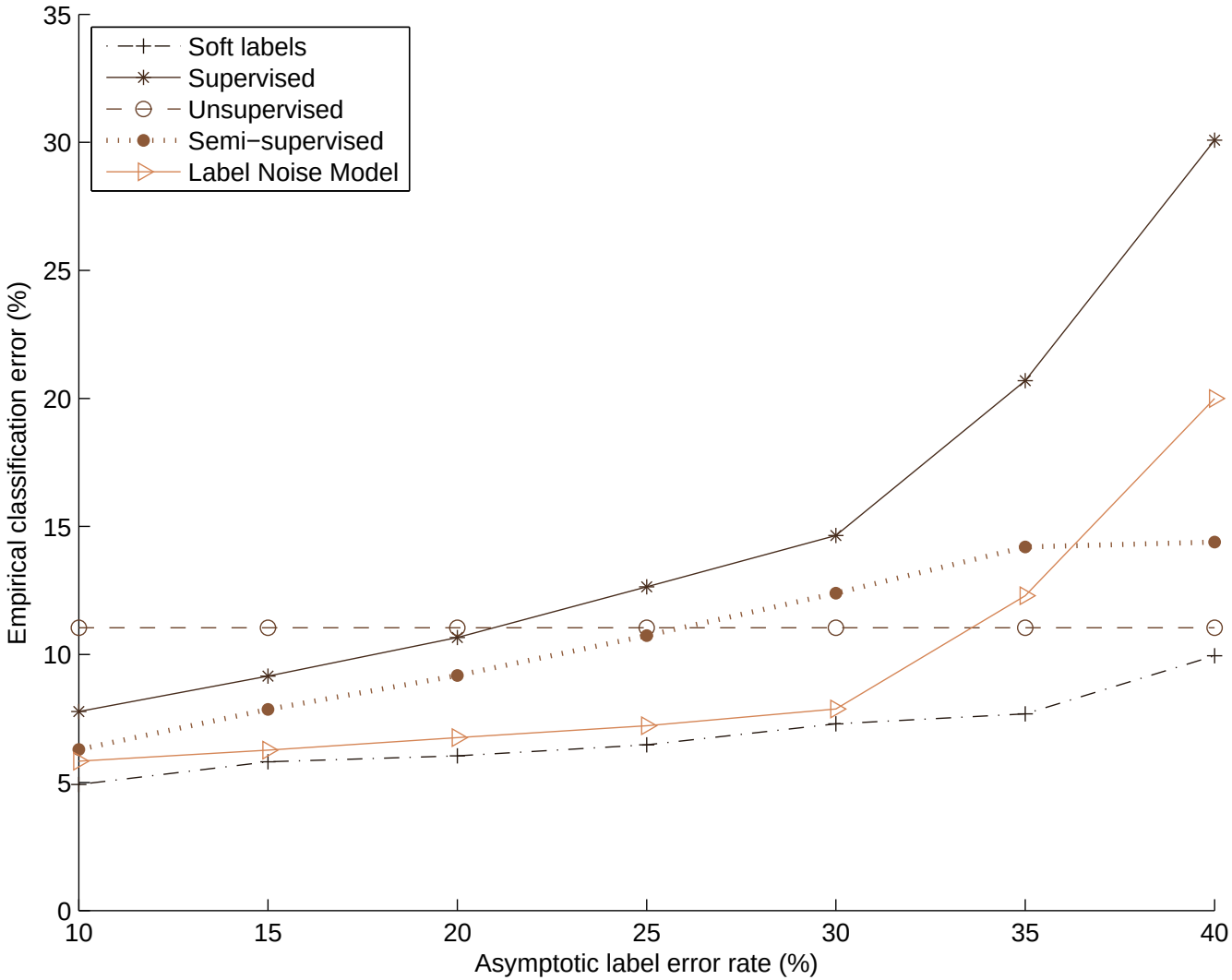
Classification error / Expert's labelling error rate

Results *Crabs*



Classification error / Expert's labelling error rate

Results *Breast Cancer Wisconsin*



Classification error / Expert's labelling error rate

Conclusions on experiment (2)

- Information on label reliability useful in the presence of label noise
- Our solution based on mixture model with soft label can deals with such additional information efficiently

Conclusion

Results

- Definition of a generalized likelihood criterion
- Demonstration of EM convergence if likelihood bounded and smooth
- Relations with probabilist criterion
- Validation with experiments

Advantages of the method

- Can benefit from all the developments done in the probabilist framework (parsimonious model, model selection criterion BIC, ...)
- Can deal with any type of information on labels

Disadvantages of the method

- Generative settings : distributional assumptions must hold for the method to work

Future work

Performance assessment

- Validation of classification results (how to extend misclassification rate to soft labels),

Extension to other Latent variable models

- **Latent Traits Analysis :**
  Continuous latent variables, discrete observations ;
- **Latent Profile Analysis :**
  Discrete latent variables, discrete observations ;
- **Independent Factor Analysis :**
  Mixed continuous, discrete latent variables, continuous observations.


C. Ambroise & G. Govaert.
EM algorithm for partially known labels.
In Proceedings of the 7^{th} international conference, IFCS, pages 161–166. Springer, 2000.


Y. Bengio & Y. Grandvalet.
Semi-supervised Learning by Entropy Minimization.
Advances in Neural Information Processing Systems, vol. 17, 2005.


T. Denœux.
A k-Nearest Neighbor Classification rule Based on Dempster Shafer Theory.
IEEE Tansaction on System Man and Cybernetics, vol. 25, no. 5, pages 804–813, may 1995.


Z. Elouedi, K. Mellouli & P. Smets.
Belief Decision Trees : Theoretical Foundations.
International Journal of Approximate Reasoning, vol. 28, 2001.


Y. Grandvallet.
Logistic regression for partial labels.
In 9th International Conference IPMU'02, volume III, pages 1935–1941, 2002.


K. Nigam, A. McCallum, S. Thrun & T. Mitchell.
Text Classification from labelled and unlabelled documents using EM.
Machine Learning, vol. 39, no. 2-3, pages 103–134, 2000.


P. Smets.
Belief functions : The disjunctive rule of combination and the generalized Bayesian theorem.
International Journal of Approximate Reasoning, vol. 9, no. 1, pages 1–35, August 1993.


P. Smets.
Belief functions on real numbers.
International Journal of Approximate Reasoning, vol. 40, no. 3, pages 181–223, 2005.


P. Vannoorenberghe & P. Smets.
Partially supervised learning by a credal em approach.
Springer, 2005.