β-risk: a New Surrogate Risk for Learning from Weakly Labeled Data

Valentina Zantedeschi[1], Rémi Emonet[1], Marc Sebban[1]

[1] Univ Lyon, UJM-Saint-Etienne, CNRS, Institut d Optique Graduate School, Laboratoire Hubert Curien UMR 5516, F-42023, SAINT-ETIENNE, France
{valentina.zantedeschi,remi.emonet,marc.sebban}@univ-st-etienne.fr

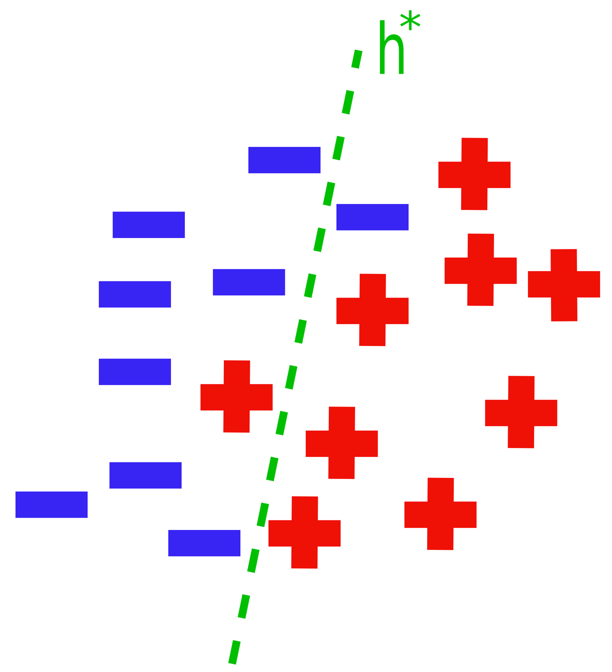
Empirical Risk Minimization

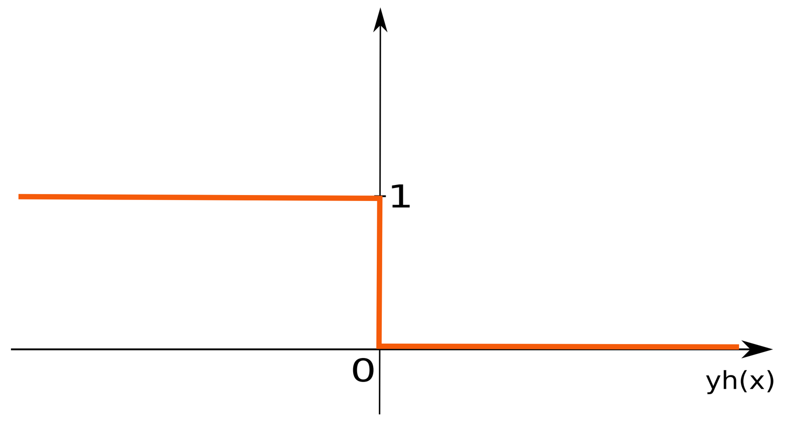
In the case of supervised binary classification, the best classifier h^* minimizes the error rate

$$h^* = \arg\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} 1_{y_i \neq h(x_i)}$$

over a sample

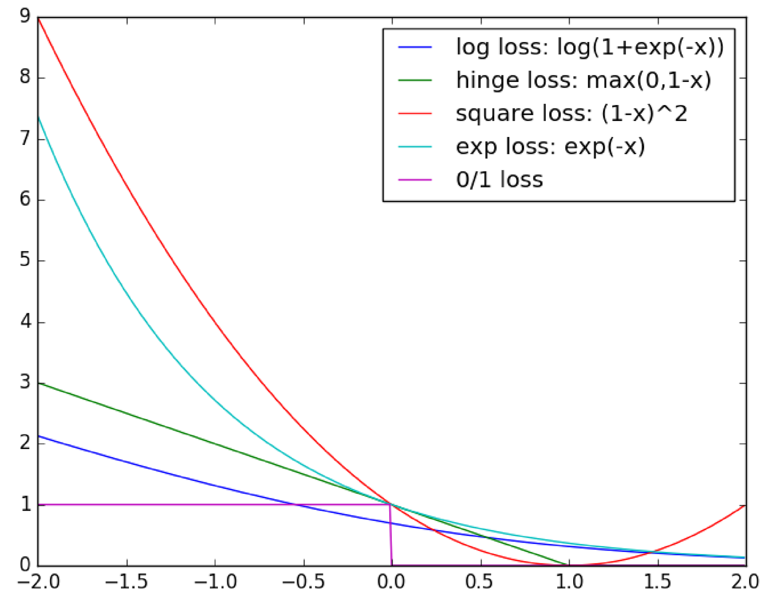
$$\mathcal{S} = [(x_i, y_i)]_{i=1}^{m} \sim P = X \times Y.$$





The 0-1 loss

- is not convex
- is not differentiable in 0
- has null gradient



The 0-1 loss is usually replaced by surrogate losses that are:

- convex
- smooth relaxations of the 0-1 loss

and the surrogate empirical risk is minimized:

$$\mathcal{R}_\phi(\mathcal{S}, h) = \frac{1}{m} \sum_{i=1}^{m} \phi(y_i h(x_i))$$

with ϕ a surrogate loss.

Weakly Label Learning

Most of the time, datasets are huge and, due to time and budget limitations, some of their labels are incorrect, not unique or missing.

Aim: Learning with datasets that lack of complete and sure information on the labels.

It covers several settings:

- Semi-Supervised Learning
- Learning with Label Proportions
- Multi-Instance Learning
- Noise-Tolerant Learning
- Multi-Expert Learning
- Unsupervised Learning

Website

https://vzantedeschi

Empirical Surrogate β-risk Minimization

Definition

For any $\mathcal{S}$, ϕ and h, and for any non-negative real coefficients β_i^{-1} and β_i^{+1} defined for each instance $x_i \in \mathcal{S}$ such that $\beta_i^{-1} + \beta_i^{+1} = 1$, the empirical surrogate risk $\mathcal{R}_\phi(\mathcal{S}, h)$ can be rewritten as

$$\mathcal{R}_\phi(\mathcal{S}, h) = \mathcal{R}_\phi(\mathcal{S}, h, \beta)$$

where

$$\mathcal{R}_\phi(\mathcal{S}, h, \beta) = \frac{1}{m} \sum_{i=1}^{m} \sum_{\sigma \in \{-1,+1\}} \beta_i^{\sigma} F_\phi(\sigma h(x_i)) + \frac{1}{m} \sum_{i=1}^{m} \beta_i^{-y_i}(-y_i h(x_i))$$

and the first term is the empirical surrogate β-risk.

β_i^{-1} and β_i^{+1} can be interpreted as:

- the degree of confidence in the label +1 (resp. -1) for x_i
- the probability of the label +1 (resp. -1) for x_i
- the injected side information

Examples of instantiation

- Multi-Expert Learning: average of votes
- Learning with Label Proportions: proportions of labels per bag
- Noise-Tolerant Learning: probabilities of the labels

Example: Soft-Margin β-SVM

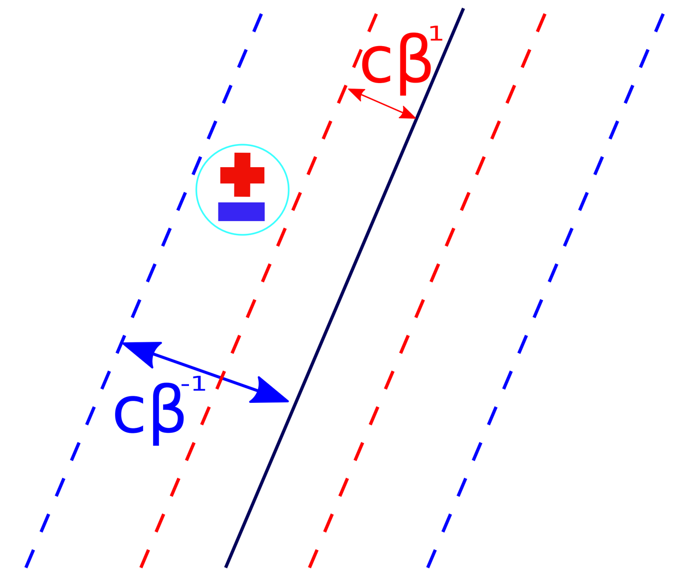
Direct generalization of a standard soft-margin SVM:

$$\arg\min_{\theta, \xi, b} \frac{1}{2} \|\theta\|_2^2 + c \sum_{i=1}^{m} (\beta_i^{-1}\xi_i^{-1} + \beta_i^{+1}\xi_i^{+1})$$

$$s.t.\ \sigma(\theta^T \mu(x_i) + b) \geq 1 - \xi_i^{\sigma} \quad \forall i = 1..m, \sigma \in \{-1, 1\}$$

$$\xi_i^{\sigma} \geq 0 \ \forall i = 1..m, \sigma \in \{-1, 1\}$$

where $\theta \in X'$ is the vector defining the margin hyperplane and b its offset, $\mu : X \to X'$ a mapping function and $c \in \mathbb{R}$ a tuned hyper-parameter.



Application to Semi-Supervised Learning

Given a set $\mathcal{X}_l$ of labeled instances of size m_l and a set $\mathcal{X}_u$ of unlabeled instances of size m_u:

Iterative Algorithm

1. initialize βs:
   $\forall i = 1..m_l$ and $\forall \sigma$ in $\{-1, 1\}$, ${}^{0}\beta_i^{\sigma} = 1$ if $\sigma = y_i, 0$ otherwise
   $\forall i = m_l{+}1..m_u$ and $\forall \sigma$ in $\{-1, 1\}$, ${}^{0}\beta_i^{\sigma} = 0.5$
2. iteratively learn:
   - an optimal separator
     $h^{t+1} = P_1(\mathcal{X}_l \cup \mathcal{X}_u, {}^{t}\beta) = \arg\min_h \ c_1 \mathcal{R}_\phi^{{}^t\beta}(\mathcal{X}_l, h) + c_2 \mathcal{R}_\phi^{{}^t\beta}(\mathcal{X}_u, h) + \mathcal{N}(h)$
   - βs only for $\mathcal{X}_u$
     ${}^{t+1}\beta = P_2(\mathcal{X}_u, h^{t+1}) = \text{argmin}\ \beta \mathcal{R}_\phi^{\beta}(\mathcal{X}_\sqcap, h^{t+1})$
     $s.t.\ \sum_{i=m_l}^{m_u} \beta_i^{-y_i}(-y_i\, h^{t+1}(x_i)) = 0$
     $\beta_i^{-1} + \beta_i^{+1} = 1,\ \beta_i^{-1} \geq 0, \beta_i^{+1} \geq 0 \ \forall i = m_l..m_u$.

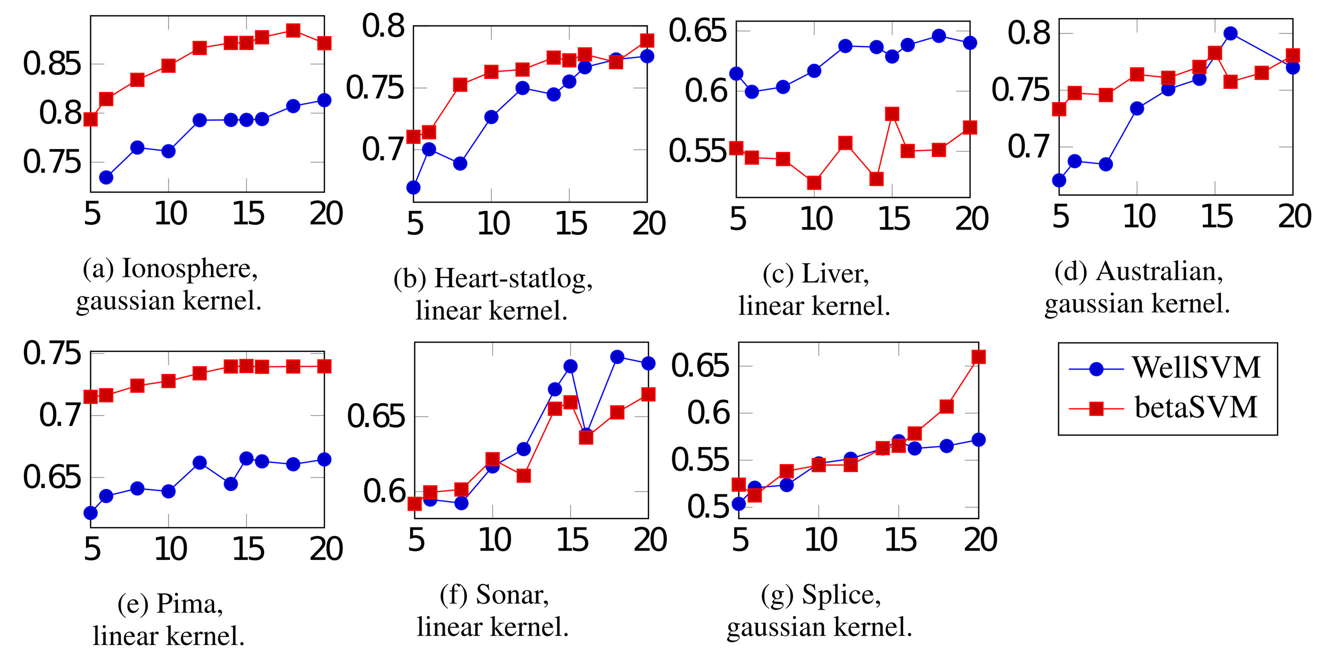


Figure 1: Comparison of the mean accuracies of WellSVM and β-SVM versus the percentage of labeled data on different UCI datasets.

References

WellSVM: Li Yu-Feng, Tsang Ivor W, Kwok James T, Zhou Zhi-Hua. *Convex and scalable weakly labeled SVMs.* The Journal of Machine Learning Research, 2013.