Supervised Learning Using Truth Tables: Convergence Results and Algorithms

Jean Marc Brossier and Olivier Lafitte

GIPSA, INPG–LAGA, Université Sorbonne Paris Nord–Centre de Recherches Mathématiques (U de Montréal) (International Research Lab of the french CNRS)

SIAM annual meeting, July 12th, 2022, CP9

Supervised learning with two classes: the truth table and its consequences

- Problem
- The algorithm ADABOOST
- The truth table

Application for three classifiers

Presentation of the problem

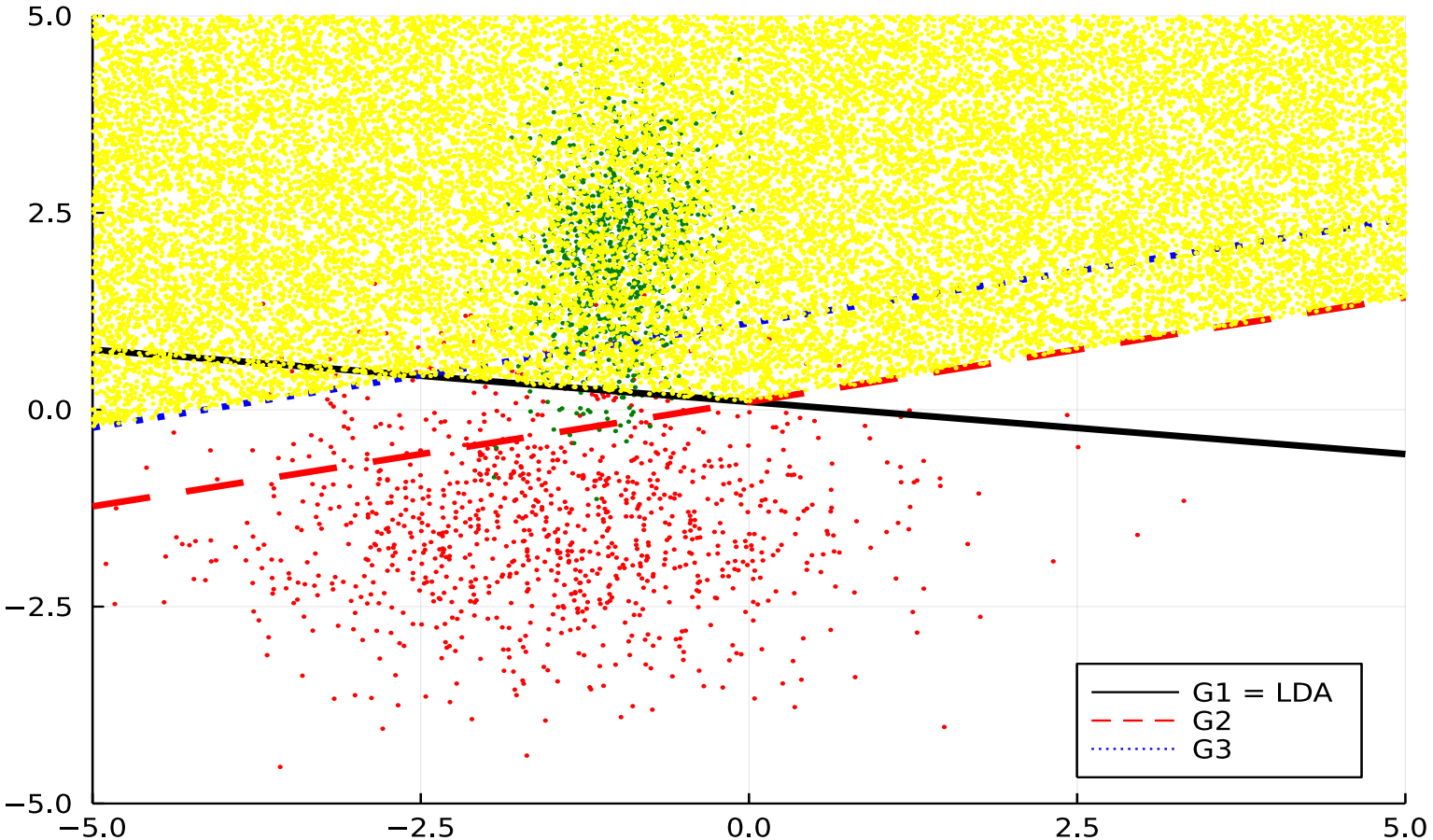
Consider the simple classification problem affecting a label $y \in \mathcal{Y} = \{-1, 1\}$ for points $x \in \mathcal{X} = \mathbf{R}^d$. One wants to determine the label ± 1 for each $x \in \mathcal{X}$, using a supervised learning method, that is based on a training set (i.i.d sample)[1] $\mathcal{S}$ of dimension n, $\mathcal{S} = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, 1 \leq i \leq n\}$.
A classifier is a function $h : \mathcal{X} \to \mathcal{Y}$. A **weak classifier** satisfies $\sharp\{i, h(x_i) = y_i\} > \frac{n}{2}$.
Each classifier is characterized by the list of signs $\text{sign}(y_i h(x_i))$.
Likelihood (risk of false decision) of h: $L_{\mathbb{P}}(h) = \mathbb{E}_{\mathbb{P}} 1_{Yh(X)<0}$.
Empirical estimate:

$$L_{\hat{\mathbb{P}}_n}(h) = \mathbb{E}_{\hat{\mathbb{P}}_n} 1_{Yh(X)<0} = n^{-1} \sum_{i=1}^{n} 1_{y_i h(x_i)<0}.$$

[1] Law of x_i is not known. Method should hold for all law $\mathbb{P}$.

Construction of a new classifier from weak classifiers

Example of the traditional work with a mixing of points of $\mathbf{R}^2$, (green and red). Use straight lines as classifiers. Example of a combination:



Brute force technique

With M weak classifiers $G_1, ..., G_M$, space of study $\mathcal{H} = Vect(G_1, ..., G_M)$.
Optimal result, for $\boldsymbol{\beta} \in \mathbf{R}^M$: minimum of

$$\mathcal{R}1_{\bullet<0}(\boldsymbol{\beta}, \mathcal{S}) := L_{\hat{\mathbb{P}}_n}(\text{sign}(\boldsymbol{\beta}^T \mathbf{G})).$$

One has

Lemma

The minimum of $\mathcal{R}1_{\bullet<0}(\boldsymbol{\beta}, \mathcal{S})$, which is piecewise constant, is obtained on a convex set

Drawback: does not construct a value of the set $\beta_1, ...\beta_M$.

Adaboost (historical)

• Algorithmic technique: Boosting (Freund and Shapire 1990, 1995).
Idea: add a ponderation on the correctly classified examples. At the beginning, all examples are equal $D_1(x_i) = \frac{1}{n}$. The ponderation of the correctly classified examples is of the form $\frac{\epsilon_m}{1-\epsilon_m}$, where $\epsilon_m = \sum_{y_i \neq h_m(x_i)} D_m(x_i)$, only if $\epsilon_m \leq \frac{1}{2}$.
New values $Z_{m+1} D_{m+1}(x_i) = D_m(x_i) \frac{\epsilon_m}{1-\epsilon_m}$ if $y_i = h_m(x_i)$ and $D_m(x_i)$ if not, with $\sum_i D_{m+1}(x_i) = 1$.
Call to *weaklearn* at each step.
• Interpretation (1997-) as the minimization of a convexified cost.

Adaboost (historical)

• Algorithmic technique: Boosting (Freund and Shapire 1990, 1995).
Idea: add a ponderation on the correctly classified examples. At the beginning, all examples are equal $D_1(x_i) = \frac{1}{n}$. The ponderation of the correctly classified examples is of the form $\frac{\epsilon_m}{1-\epsilon_m}$, where $\epsilon_m = \sum_{y_i \neq h_m(x_i)} D_m(x_i)$, only if $\epsilon_m \leq \frac{1}{2}$.
New values $Z_{m+1} D_{m+1}(x_i) = D_m(x_i) \frac{\epsilon_m}{1-\epsilon_m}$ if $y_i = h_m(x_i)$ and $D_m(x_i)$ if not, with $\sum_i D_{m+1}(x_i) = 1$.
Call to *weaklearn* at each step.
• Interpretation (1997-) as the minimization of a convexified cost.

Adaboost (newer: greedy algorithm: 1998)

Assume at each step that one considers the classifier with the least (ponderated) errors.

Ensure: $w_i = 1/n$ — Algorithme ADABOOST —
  for $m = 1 \cdots k$ **do**
    $G_{j_m} \leftarrow \arg\min_{G_j} \sum_{i=1}^{n} w_i^{(m-1)} 1_{y_i G_j(x_i)<0}$
    $\epsilon_m \leftarrow \left(\sum_{i=1}^{n} w_i^{(m-1)} 1_{y_i G_{j_m}(x_i)<0}\right) / \left(\sum_{i=1}^{n} w_i^{(m-1)}\right)$
    $\alpha_m \leftarrow \log\left((1-\epsilon_m)/\epsilon_m\right)$
    for $i = 1 \cdots n$ **do**
      $w_i \leftarrow w_i \exp\left(\alpha_m 1_{y_i G_{j_m}(x_i)<0}\right)$
    end for
  end for

Many studies on convergence of this algorithm since then.

Adaboost (newer: greedy algorithm: 1998)

Assume at each step that one considers the classifier with the least (ponderated) errors.

Ensure: $w_i = 1/n$ — Algorithme ADABOOST —
  for $m = 1 \cdots k$ **do**
    $G_{j_m} \leftarrow \arg\min_{G_j} \sum_{i=1}^{n} w_i^{(m-1)} 1_{y_i G_j(x_i)<0}$
    $\epsilon_m \leftarrow \left(\sum_{i=1}^{n} w_i^{(m-1)} 1_{y_i G_{j_m}(x_i)<0}\right) / \left(\sum_{i=1}^{n} w_i^{(m-1)}\right)$
    $\alpha_m \leftarrow \log\left((1-\epsilon_m)/\epsilon_m\right)$
    for $i = 1 \cdots n$ **do**
      $w_i \leftarrow w_i \exp\left(\alpha_m 1_{y_i G_{j_m}(x_i)<0}\right)$
    end for
  end for

Many studies on convergence of this algorithm since then.

A very old algorithm: relaxation (Jacobi in the case of matrices)

Assume at each main step that one goes through all the classifiers.

Ensure: $w_i = 1/n$ — RELAXATION —
  for $m = 1 \cdots k$, $j = 1 \cdots M$ **do**
    $\epsilon_{m,j} \leftarrow \left(\sum_{i=1}^{n} w_i^{m,j-1} 1_{y_i G_j(x_i)<0}\right) / \left(\sum_{i=1}^{n} w_i^{m,j-1}\right)$
    $(w_i^{m,0} := w_i^{m-1,M})$
    $\alpha_{m,j} \leftarrow \log\left(\left(1-\epsilon_{m,j}\right)/\epsilon_{m,j}\right)$
    for $i = 1 \cdots n$ **do**
      $w_i \leftarrow w_i \exp\left(-\frac{1}{2}\alpha_{m,j}\mathsf{sign}(y_i G_j(x_i))\right)$
    end for
  end for

Convergence ensured under mild conditions.

A very old algorithm: relaxation (Jacobi in the case of matrices)

Assume at each main step that one goes through all the classifiers.

Ensure: $w_i = 1/n$ — RELAXATION —
  for $m = 1 \cdots k$, $j = 1 \cdots M$ **do**
    $\epsilon_{m,j} \leftarrow \left(\sum_{i=1}^{n} w_i^{m,j-1} 1_{y_i G_j(x_i)<0}\right) / \left(\sum_{i=1}^{n} w_i^{m,j-1}\right)$
    $(w_i^{m,0} := w_i^{m-1,M})$
    $\alpha_{m,j} \leftarrow \log\left((1-\epsilon_{m,j})/\epsilon_{m,j}\right)$
    for $i = 1 \cdots n$ **do**
      $w_i \leftarrow w_i \exp\left(-\frac{1}{2}\alpha_{m,j}\mathsf{sign}(y_i G_j(x_i))\right)$
    end for
  end for

Convergence ensured under mild conditions.

A new structuration of $\mathcal{S}, G_1, ..., G_M$: the truth table

• For a given set of classifiers, it is a **structuration** of the training set $\mathcal{S}$ (through a binary tree). It segments $\mathcal{S}$ in 2^M classes:

$\sharp\{i, sign(y_i G_1(x_i)) = sign(y_i G_2(x_i))... = sign(y_i G_M(x_i)) = 1\}$: all classifiers correctly label these points,

$\sharp\{i, sign(y_i G_1(x_i)) = sign(y_i G_2(x_i))... = sign(y_i G_M(x_i)) = -1\}$: all classifiers uncorrectly label these points,

$\sharp\{i, sign(y_i G_2(x_i)) = sign(y_i G_3(x_i))... = sign(y_i G_M(x_i)) = -1, sign(y_i G_1(x_i) = 1\}$: all classifiers but G_1 uncorrectly label these points,

$\sharp\{i, sign(y_i G_2(x_i)) = sign(y_i G_3(x_i))... = sign(y_i G_M(x_i)) = 1, sign(y_i G_1(x_i) = -1\}$: all classifiers correctly label these points except G_1,

...

$\sharp\{i, sign(y_i G_3(x_i)) = sign(y_i G_4(x_i))... = sign(y_i G_M(x_i)) = 1, sign(y_i G_1(x_i) = sign(y_i G_2(x_i)) = -1\}$: all classifiers correctly label these points except G_1 and G_2,

The elements of one class are not distinguable for $G_1, ..., G_M$.

• Classes are arranged by pairs (label l), corresponding to **all** opposite signs.

The minimum of the empirical cost

Each pair is associated to $X_l, -X_l$: l label of the class,
$X_l = \sum V_{lj}\beta_j$, $V_{lj} = \pm 1$.
For $\boldsymbol{\beta}$ such that all $X_l \neq 0$ (not being on a boundary),
$n\mathcal{R}1_{\bullet<0}(\boldsymbol{\beta}, \mathcal{S})$ is an integer in a list of 2^M elements.
The minimum value of this list defines a convex set in $\mathbf{R}^M$.

Convexification of the cost function

Replace $1_{\bullet<0}$ by $\varphi(\bullet)$, satisfying $\varphi(x) \geq 1_{x<0}$ for all x).
Convexified empirical risk:

$$\mathcal{R}_{\varphi}(\boldsymbol{\beta}, \mathcal{S}) = n^{-1} \sum_{i=1}^{n} \varphi\left(y_i \boldsymbol{\beta}^T \mathbf{G}(x_i)\right) \tag{1}$$

Existence and uniqueness of the minimum of the cost function

Theorem

- *If $2^M > n$, at least $2^M - n$ slots are empty,*
- *If at least M pairs are full (the two elements are non zero), the convexified cost function has a unique point of minimum.*
- *In this case, the relaxation algorithm converges to its unique minimum.*

For the second item, M pairs are independent, the function $\theta\varphi(x) + (1-\theta)\varphi(-x)$ is $\alpha-$convex $(\theta \in (0,1))$
The third item is a consequence of the second item.
The relaxation algorithm has the same cost as ADABOOST.

Hypotheses on φ and properties

Hypothèse

$\varphi(0) = 1$, $\varphi'(0) < 0$, φ *of class* C^1, φ *strictly positive, strictly convex, decreasing, and there exists* $\alpha_0 > 0$ *such that the even part of* φ *is* $\alpha-$*convexe.*

In particular, φ is 'classification calibrated' (Bartlett). Remark: convexified cost function $n^{-1}\sum_{j\in J}(n_j + m_j)C_{\eta_j}(X_j)$ where $\eta_j = \frac{m_j}{n_j+m_j}$, with notation $C_\eta(x) = \eta\varphi(x) + (1-\eta)\varphi(-x)$:

Lemma

For all $\eta \in (0,1)$ C_η *is* $2\min(\eta, 1-\eta)$ $\alpha-$*convex and has a unique point of minimum, denoted by* $x(\eta)$*, where* $x(\eta) > 0 \Leftrightarrow \eta > \frac{1}{2}$.

Under these conditions, the relaxation algorithm defined above converges to the unique point of minimum of $\mathcal{R}_\varphi(\boldsymbol{\beta}, \mathcal{S})$, one has explicit equations, and a calculation of κ_φ.

Margins

Definition

For a given $\mathcal{S}, G_1, ..., G_M$ (with existence and uniqueness of a point of minimum through M pairs), one calls margin $\kappa_\varphi(\mathcal{S}, G_1, ..., G_M)$ the smallest coefficient κ such that, for any set $\mathcal{S}', G'_1, ..., G'_M$, with $\max_{\{s_p\}} \left| \frac{\sharp\{i, sign(y_i G_p(x_i)) = s_p\}}{\sharp\mathcal{S}} - \frac{\sharp\{i, sign(y'_i G'_p(x'_i)) = s_p\}}{\sharp\mathcal{S}'} \right| \leq \kappa$, the sign of the resulting classifier is identical.

Theorem

The real $\kappa_\varphi(\mathcal{S}, G_1, ..., G_M)$ is well defined and is strictly positive.

The theorem is a consequence of the C^1 behavior of the point of minimum of $\sum_{\{s_p\}} \alpha_{s_p} \varphi(X_{s_p})$ for $\sum_{\{s_p\}} \alpha_{s_p} = 1$ positive coefficients with M pairs non zero with respect to the parameters (α_{s_p}).

The truth table for three classifiers

	n_0	m_0	n_1	m_1	n_2	m_2	n_3	m_3
G_1	-1	$+1$	$+1$	-1	-1	$+1$	-1	$+1$
G_2	-1	$+1$	-1	$+1$	$+1$	-1	-1	$+1$
G_3	-1	$+1$	-1	$+1$	-1	$+1$	$+1$	-1
$\boldsymbol{\beta}^T\mathbf{G}$	$-X_0$	X_0	$-X_1$	X_1	$-X_2$	X_2	$-X_3$	X_3

$X_0 = \beta_1 + \beta_2 + \beta_3, X_1 = -\beta_1 + \beta_2 + \beta_3, X_2 = \beta_1 - \beta_2 + \beta_3, X_3 = \beta_1 + \beta_2 - \beta_3.$

$\beta_1 = \frac{1}{2}(X_1 + X_2), \beta_2 = \frac{1}{2}(X_1 + X_3), \beta_3 = \frac{1}{2}(X_1 + X_2)$.

Three classifiers $n\mathcal{R}_\varphi(\boldsymbol{\beta}, \mathcal{S}) =$

$$n_0\varphi(-X_0) + m_0\varphi(X_0) + n_1\varphi(-X_1) + m_1\varphi(X_1) +$$
$$n_2\varphi(-X_2) + m_2\varphi(X_2) + n_3\varphi(-X_3) + m_3\varphi(X_3).$$

New adaboost is a decision tree.

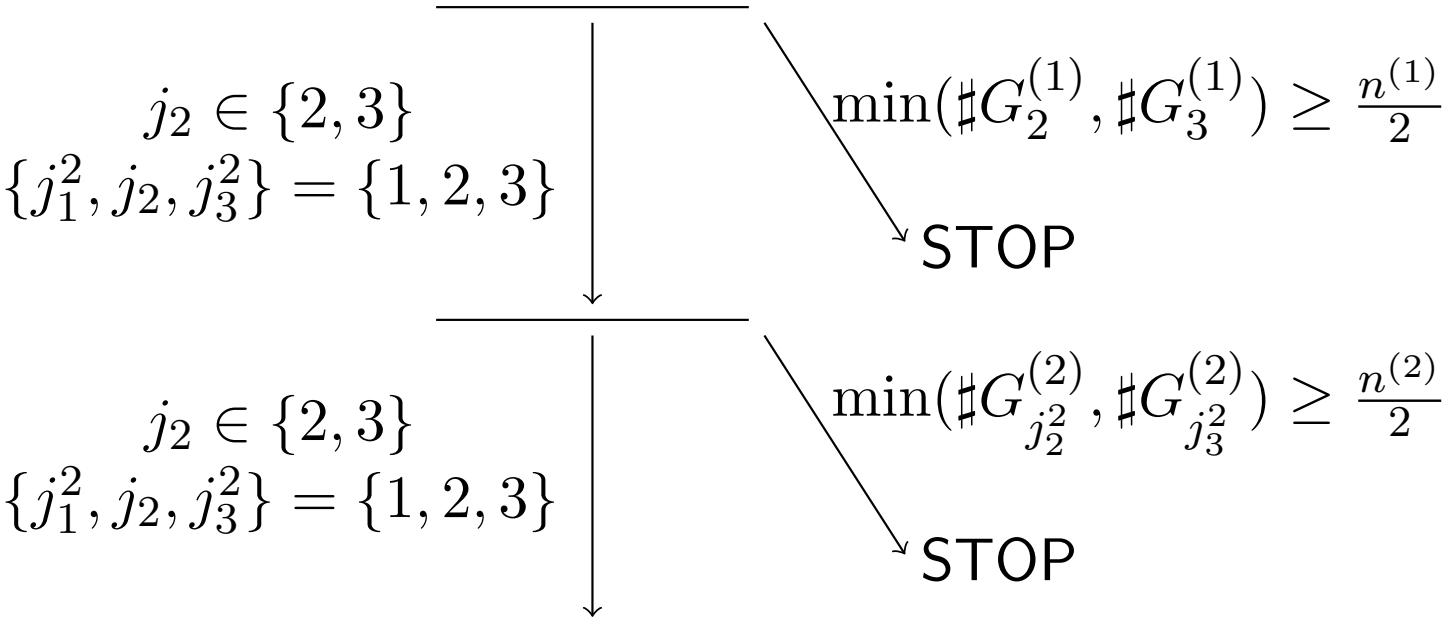
Assume $\sharp G_1 \leq \sharp G_2 \leq \sharp G_3$ (it is an ordering of the classifiers)

$$\underset{j_1=1}{\sharp G_1 < \sharp G_2} \quad \underset{j_1=1}{\sharp G_1 = \sharp G_2 < \sharp G_3} \quad \underset{j_1=1}{\sharp G_1 = \sharp G_2 = \sharp G_3}$$

Update the coefficients and consider $\beta_1^1 = -\frac{1}{2}\ln\frac{\sharp G_1}{n-\sharp G_1}$. Update by considering n_j^1, m_j^1 multiplied by $e^{\pm\beta_1^1}$. One obtains $\sharp G_1^1 = \frac{1}{2}(\sum n_j^1 + m_1^j) = \frac{n^1}{2}$. The next step reads as

$$\underset{j_2=2}{\sharp G_2^1 < \tfrac{n^1}{2} \leq \sharp G_3^1} \quad \underset{stop}{\sharp G_2^1 = \tfrac{n^1}{2} < \sharp G_3^1} \quad \underset{stop}{\tfrac{n^1}{2} < \sharp G_3^1 < \sharp G_2^1} \quad \underset{stop}{\tfrac{n^1}{2} < \sharp G_3^1 = \sharp G_2^1} \quad \underset{j_2=3}{\sharp G_3^1 < \tfrac{n^1}{2} < \sharp G_2^1}$$

Decision tree



The value of the cost function decreases,
but β^m does not converge to its unique point of minimum.

Conclusion

- A novel method for combining weak classifiers in supervised learning is described, which fully characterizes the set of weak classifiers by a truth table.
- Convexification of the risk function with any calibrated C^2 classification function φ, yields a minimization problem in $\mathbb{R}^M$, whose unique solution is easily studied using a classical minimization algorithm that amounts to iteratively solving equations in $\mathbb{R}$ with a Newton method.
- The complexity of this method depends only linearly on the number M of weak classifiers (no dependency on n and d).
- In the case of two well-known φ's, the Boosting function (for all M) or the Logistic function.
- This framework is then used to study the quality of the training set, thus setting criteria for the stability of the results under such operations.