A DISTRIBUTED CONSTRAINED-FORM SUPPORT VECTOR MACHINE

François D. Côté, Ioannis N. Psaromiligkos, and Warren J. Gross
Department of Electrical and Computer Engineering, McGill University, Montreal, Canada
francois.cote@mail.mcgill.ca; yannis@ece.mcgill.ca; warren.gross@mcgill.ca

In brief

This study is about deriving a fully distributed method for training a constrained-form SVM. We present the first algorithm of this kind.

Our algorithm draws on our recent approach, based on the Douglas-Rachford method, for solving network-structured problems.

We offer a new epigraphical projection result, which provides an efficient implementation of the kernel version of our algorithm—efficient, because it involves only closed-form steps.

Our approach is an effective strategy for fully distributed nonlinear classification.

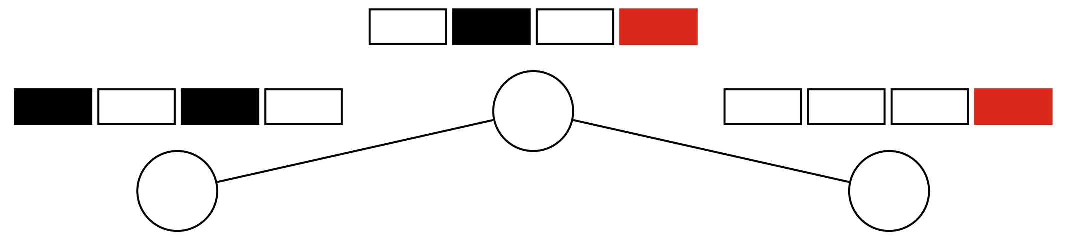
Introduction

Objective To have m networked agents learn to classify the objects of a set $\mathcal{X}$ into two classes, ± 1, training an SVM. The task involves using a dataset of ℓ labeled examples, (x_1, y_1), ..., (x_ℓ, y_ℓ), to find a function $h\colon \mathcal{X} \to \mathbb{R}$ whose sign yields the labels. This function is parameterized by a vector w in a real Hilbert space $\mathcal{H}$ and by a real number b, and is defined through a mapping ϕ. For an object x, the value of h is $\langle \phi(x), w\rangle + b$. In constrained form, the SVM entails finding h by solving, for an $\epsilon > 0$, the problem

$$\min_{(w,b)\in\mathcal{H}\times\mathbb{R}} \|w\| \quad \text{s.t.} \quad \sum_{k=1}^{\ell} \max\{0, 1 - y_k h(x_k; w, b)\} \le \epsilon.$$

Novelty While the agents know m and ϵ, they only know the dataset collectively, each agent i knowing its part as a vector a_i in $\mathbb{R}^\ell$ and a linear operator $A_i\colon \mathbb{R}^\ell \to \mathcal{H}$, such that

$$y_k h(x_k; w, b) = \Big[\sum_{i=1}^{m} (b a_i + A_i^* w)\Big]_k.$$



The dataset is known in the network as a union of subsets, possibly with overlap.

Expressing the argument of the hinge loss with a sum is not only useful analytically; it permits arbitrary data splitting [*see the figure above*].

Definition An algorithm for solving a problem with data divided among networked agents is fully distributed when

- each agent communicates only with its neighbors,
- no agent shares its part of the data, and
- all the agents agree on a solution.

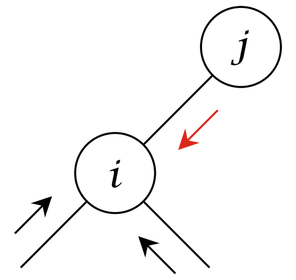
Result We derive a fully distributed method for nonlinear classification with data divided into summands.

Proposed algorithm

Network The network is characterized by each agent i's set of neighbors, $\mathcal{N}_i$.
Data Each agent i knows the summands a_i and A_i.
Approach Distributed scaled Douglas-Rachford algorithm.
Parameters All the agents know the same two arbitrary real numbers: $\gamma > 0$, and $\lambda \in (0, 2)$.
Initialization Each agent i chooses, for each j in $\mathcal{N}_i$, three quantities, $z_{1,ij,0} \in \mathcal{H}$, $z_{2,ij,0} \in \mathbb{R}$, and $z_{3,ij,0} \in \mathbb{R}^\ell$.

Main loop At iteration $n = 0, 1, \ldots$, each agent i repeats:

STEP 1 COMMUNICATE



Receive $z_{1,ji,n}$, $z_{2,ji,n}$, and $z_{3,ji,n}$ from each neighbor j.

STEP 2 OPTIMIZE

Find the quantity (w_i, b_i) in $\mathcal{H}\times\mathbb{R}$ and the family $(\Delta v_{ij})_{j\in\mathcal{N}_i}$ of vectors in $\mathbb{R}^\ell$ that minimize

$$\gamma\|w_i\|^2 + \frac{1}{2}\sum_{j\in\mathcal{N}_i}\Big(\|w_i - z_{1,ji,n}\|^2 + (b_i - z_{2,ji,n})^2 + \|\Delta v_{ij} + z_{3,ji,n}\|^2\Big)$$

subject to

$$\sum_{k=1}^{\ell} \max\Big\{0, \frac{1}{m} - \Big[b_i a_i + A_i^* w_i + \sum_{j\in\mathcal{N}_i} \Delta v_{ij}\Big]_k\Big\} \le \frac{\epsilon}{m}$$

and assign the minimizers to $(w_{i,n}, b_{i,n})$ and $(\Delta v_{ij,n})_{j\in\mathcal{N}_i}$.

STEP 3 UPDATE

For each j in $\mathcal{N}_i$, compute

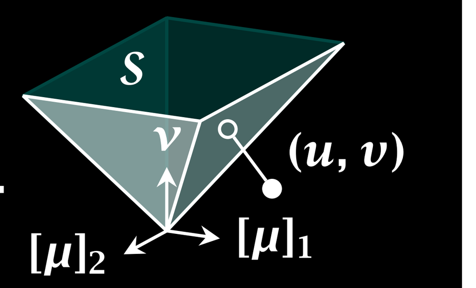
$$z_{1,ij,n+1} = z_{1,ij,n} + \lambda\Big(w_{i,n} - \frac{1}{2}(z_{1,ij,n} + z_{1,ji,n})\Big),$$

$$z_{2,ij,n+1} = z_{2,ij,n} + \lambda\Big(b_{i,n} - \frac{1}{2}(z_{2,ij,n} + z_{2,ji,n})\Big), \quad \text{and}$$

$$z_{3,ij,n+1} = z_{3,ij,n} + \lambda\Big(\Delta v_{ij,n} - \frac{1}{2}(z_{3,ij,n} - z_{3,ji,n})\Big).$$

Key result

Epigraphical projection



The set $\mathcal{S}$, given by $\{(\mu, \nu) \in \mathbb{R}^\ell \times \mathbb{R} :\ 0 \le [\mu]_k \le \nu\ \forall k\}$, is an epigraph. To efficiently project onto it, we provide the following result:

Proposition Let $\bar{u}$ be a vector u in $\mathbb{R}^\ell$ with entries sorted in ascending order. Let $v \in \mathbb{R}$. Define

$$q_k = \max\{0, (v + [\bar{u}]_k + \cdots + [\bar{u}]_\ell)/(\ell + 2 - k)\}, \quad k = 1, \ldots, \ell.$$

Then, at most one of $q_1 \le [\bar{u}]_1$, $[\bar{u}]_1 < q_2 \le [\bar{u}]_2$, ..., $[\bar{u}]_{\ell-1} < q_\ell \le [\bar{u}]_\ell$ holds. For the one that is true, define $\nu' = q_k$; if none holds, set $\nu' = \max\{0, v\}$. The projection of (u, v) onto $\mathcal{S}$ is given by

$$[\mu]_k = \text{median}\{0, [u]_k, \nu'\} \quad \forall k \quad \text{and} \quad \nu = \nu'.$$

Convergence

Proposition Suppose that the following conditions hold:

- The network is connected.
- There exists a (w, b) such that the inequality in the problem holds strictly.

Then, provided that a solution to the problem exists and that $\mathcal{H}$ is finite dimensional, the sequence $(w_{i,0}, b_{i,0})$, $(w_{i,1}, b_{i,1})$, ... converges to a solution for every i.

Nonlinear classification

Dimensionality reduction To make our method useful for the nonlinear, and generally infinite-dimensional case, we assume that neighbors i and j share ℓ_{ij} possibly unlabeled objects, $\tilde{x}_{ij,k} \in \mathcal{X}$ for $k = 1, \ldots, \ell_{ij}$, allowing them to form an operator $R_{ij}\colon \mathbb{R}^{\ell_{ij}} \to \mathcal{H}$. This operator serves to modify our algorithm so that $z_{1,ij,n}$ and $z_{1,ji,n}$ belong to $\mathbb{R}^{\ell_{ij}}$.

Modified algorithm Converges, but to an approximation.
STEP 2′ Replace $\|w_i - z_{1,ji,n}\|^2$ with $\|R_{ij}^* w_i - z_{1,ji,n}\|^2$.
STEP 3′ Replace $w_{i,n}$ with $r_{ij,n} = R_{ij}^* w_{i,n}$.

Kernel method By viewing the problem in the second step of the modified algorithm through duality, we see that ϕ occurs only in inner products and thus kernel evaluations, $K(x_1, x_2) = \langle\phi(x_1), \phi(x_2)\rangle$ for some $x_1, x_2 \in \mathcal{X}$. The most popular kernel is the Gaussian kernel,

$$K(x_1, x_2) = \exp(-\|x_1 - x_2\|^2/C), \quad C > 0;$$

however, our algorithm works with any K.

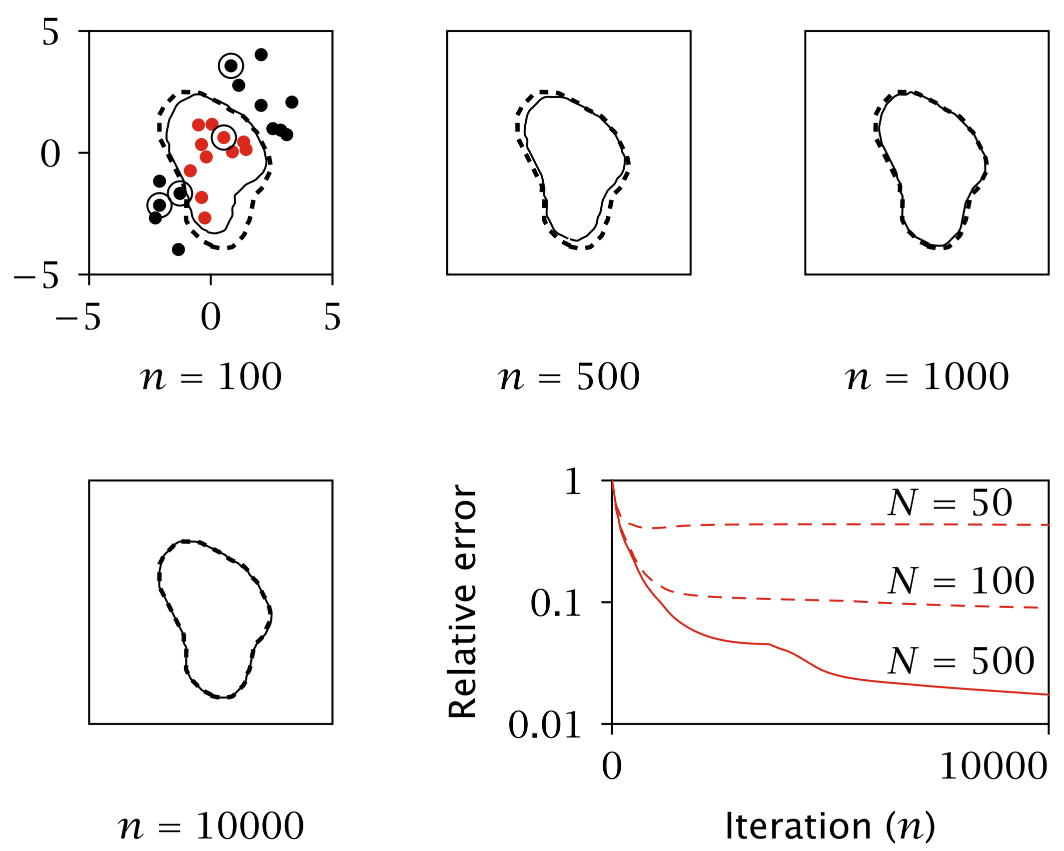
Solving the dual problem We approximate the solution to the dual problem in Step 2 by using two warm-started projection-gradient iterations. These iterations depend on a parameter δ_i, which is any number in $(0, 2/L_i)$, where L_i is the Lipschitz constant of the gradient. The projection is onto a set $\mathcal{S}$ and can be determined simply by observation [*see the box above*].

Obtaining h Part of the solution to the dual problem is a vector $\mu_{i,n}$ in $\mathbb{R}^\ell$, and together with related quantities, $\tilde{\mu}_{ij,n} \in \mathbb{R}^{\ell_{ij}}$ for $j \in \mathcal{N}_i$, it leads not only to $(r_{ij,n})_{j\in\mathcal{N}_i}$, $b_{i,n}$, and $(\Delta v_{ij,n})_{j\in\mathcal{N}_i}$, but also to the local $h(x; w_{i,n}, b_{i,n})$:

$$\sum_{k=1}^{\ell} [a_i]_k[\mu_{i,n}]_k K(x, x_k) + \sum_{j\in\mathcal{N}_i}\sum_{k=1}^{\ell_{ij}} [\tilde{\mu}_{ij,n}]_k K(x, \tilde{x}_{ij,k}) + b_{i,n}.$$

Simulations

Simple 2D data We consider a network of six agents and a dataset of 24 points in $\mathbb{R}^2$ from two equiprobable classes. One class corresponds to a normal distribution, and the other to a mixture of two normal distributions. Each agent knows a subset of four labeled points. The agents share N unlabeled points drawn uniformly in an area surrounding the labeled ones. We set ϵ, γ and λ to 1 and δ_i to $1.99/L_i$. The agents use a Gaussian kernel with $C = 1.8$. We observe agent 1's decision boundary and the relative error between $(w_{1,n}, b_{1,n})$ and the centralized result [*see the plots below*].



Our algorithm's convergence behavior for simple 2D data. Despite knowing only the circled data, an agent's decision boundary (——) agrees with the centralized result (- - -) when the agents share enough random points (see ——).

Conclusion

Training a constrained-form SVM in a fully distributed way is possible. We have illustrated that our strategy, with its Douglas-Rachford and projection-gradient underpinnings, can efficiently train a nonlinear classifier that agrees closely with the centralized result.

Take-home message — Despite having been recognized as a "notoriously complex problem," training a nonlinear classifier in a fully distributed way can be done efficiently, using a sequence of closed-form steps.