Robust optimal classification trees under noisy labels

Efstratios Reppas

Objective

- Understand the theoretical background of this paper
- Reproduce its experimental results

Robust optimal classification trees under noisy labels

- Victor Blanco, Alberto Japon and Justo Puerto - 2020
- Binary classification problem
- Combination of Optimal Decision Trees with SVM
- Focus on robustness by considering noisy data

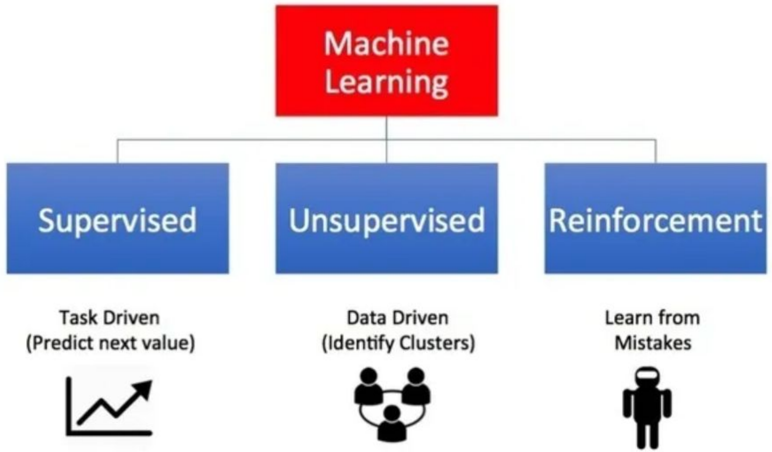
Theoretical Background

Supervised Learning

- One of the three main paradigms of machine learning

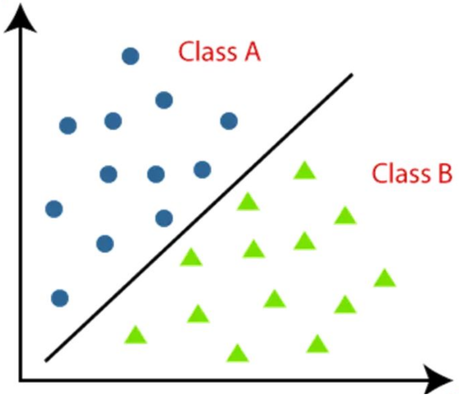
- Learn a function $h : X \rightarrow Y$ that maps inputs X (**features**) to outputs Y (**labels**).

- Labels are known for a subset of X



Binary Classification

- Subset of Supervised Learning
- The labels take binary values of $y \in \{0, 1\}$ or $\boldsymbol{y} \in \{-1, +1\}$



Mathematical Programming

- Involves problems that can expressed as minimization/maximization of a function, called the **objective function**, given set of constraints.

- General formulation:

$$\min_x f(x) \quad \text{s.t.} \quad g_i(x) \leq 0, \forall i, \quad h_j(x) = 0, \forall j$$

where f: objective function, g_i : inequality constraints, h_j: equality constraints

- Such problems can be often solved by several algorithms, off-the-shelf in many programming languages.

Categorization of Problems

- **Linear Programming**: All functions linear, variables continuous → ∈ P

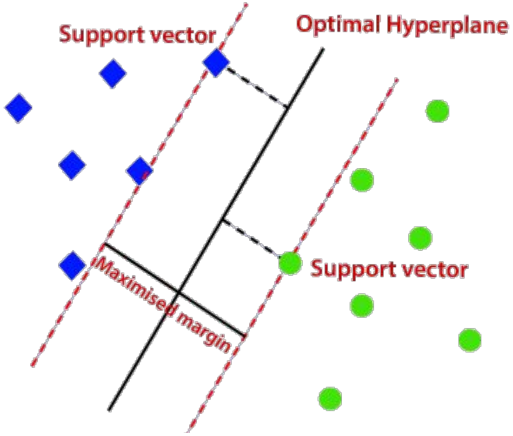
- **Mixed-Integer Programming**: Variables can be integers → ∈ NP

- **Non-Linear Programming**: Functions can be non-linear → harder problem, when there are non-convexity issues

Support Vector Machines (SVMs)

- A very popular (binary) classification model
- Calculates a hyperplane that maximizes the margin between itself and the two classes
- Calculated as a **Quadratic Optimization Problem**:

$$
\begin{aligned}
&\underset{\mathbf{w},\, b}{\text{minimize}} && \frac{1}{2}\|\mathbf{w}\|^2 \\
&\text{subject to} && y_i(\mathbf{w}^\top \mathbf{x}_i - b) \geq 1 \quad \forall i \in \{1, \ldots, n\}
\end{aligned}
$$



RE-SVM

- Robust in noisy datasets - improved performance over standard SVM
- Allows for flipping labels during optimization
- Formulation:

$$\min_{w,w_0,e,\xi} \frac{1}{2}\|w\|^2 + c_1 \sum_{i=1}^{n} e_i + c_2 \sum_{i=1}^{n} \xi_i$$

subject to:

$$(1 - 2\xi_i) y_i (w^\top x_i + w) \geq 1 - e_i, \quad \forall i = 1, \ldots, n,$$

$$e_i \geq 0, \quad \xi_i \in \{0, 1\}, \quad \forall i = 1, \ldots, n.$$

e_i: the misclassification error for the i-th observation

ξ_i: indicates whether
observation is relabeled ($\xi i = 1$) or not ($\xi i = 0$).

c_1, c_2 : cost parameters

RE-SVM

- Robust in noisy datasets - improved performance over standard SVM
- Allows for flipping labels during optimization
- Formulation:

$$\min_{w, w_0, e, \xi} \frac{1}{2}\|w\|^2 + c_1 \sum_{i=1}^{n} e_i + c_2 \sum_{i=1}^{n} \xi_i$$

subject to:

$$(1 - 2\xi_i) y_i (w^\top x_i + w) \geq 1 - e_i, \quad \forall i = 1, \ldots, n,$$

$$e_i \geq 0, \quad \xi_i \in \{0, 1\}, \quad \forall i = 1, \ldots, n.$$

e_i: the misclassification error for the i-th observation

ξ_i: indicates whether
observation is relabeled (ξi = 1) or not (ξi = 0).

c_1, c_2 : cost parameters

$$e_i = \begin{cases} \max\{0, 1 - y_i(\omega^\top x_i + \omega_0)\}, & \text{if } \xi_i = 0, \\ \max\{0, 1 + y_i(\omega^\top x_i + \omega_0)\}, & \text{if } \xi_i = 1. \end{cases}$$

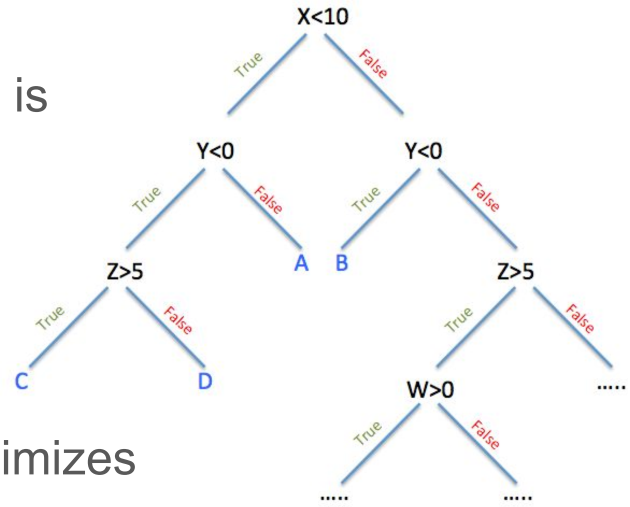
Decision Trees

- Hierarchical models that partition the feature space recursively to make decisions based on the input data.

- At each level of the tree, a binary decision rule is implemented based on an input feature. The classification is made in the leaf nodes

- CART, one of the baseline algorithms, uses a greedy approach to split the data at each node, by selecting the feature that minimizes the gini index:

$$Gini = 1 - \sum_j p_j^2$$



Optimal Decision Trees (OCTs)

- CART is an efficient algorithm, but produces suboptimal trees
- **OCTs** calculate trees that are optimal in some sense → minimize misclassification errors (L_t) and the number of splits (d_t)
- Their objective function:

$$\min \sum_{t \in \mathcal{L}} L_t + \alpha \sum_{t \in \mathcal{B}} d_t,$$

  where *a* controls the tradeoff between accuracy and tree depth, $\mathcal{L}$ is the set of leaf nodes and $\mathcal{B}$ is the set of intermediate nodes.

- **OCT-H** builds upon this paradigm, allowing for decision rules to be drawn based on a linear combination (hyperplane) of the features (instead of using only one at each node).

OCTSVM

OCTSVM

- The idea comes from the fact that OCTs do not include any notion of optimality regarding class separation.
- The hyperplanes of OCT-H can be drawn **anywhere between the classes**
- The authors fix this by incorporating the SVM principle of maximum margins between the hyperplane and the classes (**OCT + SVM**)
- No need for **leaf-nodes**: classification occurs at each node(*D-1* depth for same splits)
- Also allows **relabelling** for robustness
- The model is formulated as **Mixed Integer Non Linear Program (MINLP)**

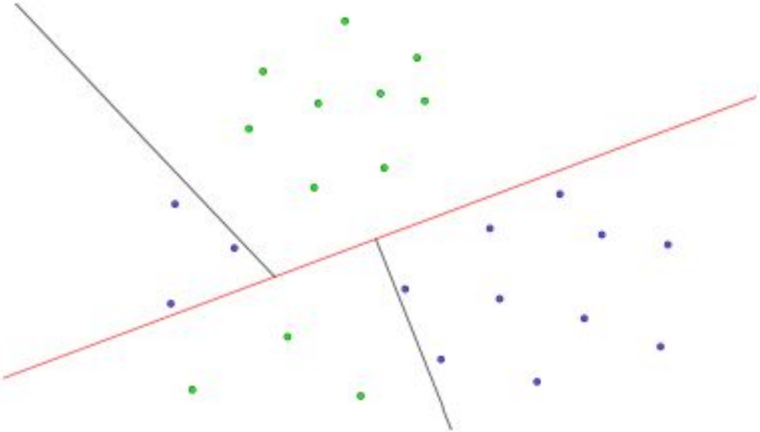


FIGURE 2. Optimal solutions for OCT-H with $D = 2$ (left) and OCTSVM with $D = 1$ (right).

Formulation as MINLP

SVM objective function

$$\frac{1}{2}||\omega_t||_2 \leq \delta \qquad \forall t = 1, \ldots, T.$$

Inverse of margin minimized ⇒ maximization of margin for each node t of the tree

minimized in the objective function

Relabelling

tradeoff parameters

$$c_1 \sum_{i=1}^{n} \sum_{t=1}^{T} e_{it} + c_2 \sum_{i=1}^{n} \sum_{t=1}^{T} \xi_{it}.$$

misclassification error

Binary {0,1} variable - indicates label flip

Same as Re-SVM, but now because we consider each observation i at each node t of the tree, we sum in their entirety.

To be minimized in the objective function.

Definition of e_{it} and ξ_{it}

Binary variable → 0 if observation i not used in node t

$$y_i(\omega_t' x_i + \omega_{t0}) - 2y_i(\beta_t' x_i + \beta_{it0}) \geq 1 - e_{it} - M(1 - z_{it}), \quad \begin{cases} \forall i = 1,\dots,N, \\ t = 1,\dots,T, \end{cases} \tag{3}$$

$$\beta_{itj} = \xi_{it}\omega_{tj}, \quad \forall i = 1,\dots,N, t = 0,\dots,T,\ j = 0,\dots,p. \tag{4}$$

$$\xrightarrow{z_{it} = 0} \quad f(\omega, \xi, e) \geq -\infty$$

Always holds

$$\xrightarrow{z_{it} = 1, \xi_{it} = 0} \quad e_{it} \geq 1 - y_i(w_t x_i + w_{t0})$$

Same as RE-SVM constraint

$$\xrightarrow{z_{it} = 1, \xi_{it} = 1} \quad e_{it} \geq 1 + y_i(w_t x_i + w_{t0})$$

Applies RE-SVM in each node of the tree

OCT constraint

- Adding the minimal split objective as in OCTs (the misclassification error has already been implemented as part of RE-SVM), we have the following in the objective function:

$$c_3 \sum_{t=1}^{T} d_t.$$

- d_t is defined as a binary variable that expresses whether a split is applied at node t (0 else).

Definition of d_t

- $\|\omega_t\|_2 \leq M d_t$

$d_t = 0$: $\|\omega_t\|_2 \leq 0 \Rightarrow \|\omega_t\|_2 = 0 \ \ \forall t = 1, \ldots, T.$

$d_t = 1$: $\|\omega_t\|_2 \leq \infty \ \ \forall t = 1, \ldots, T.$ always holds

Is 1 only if weights are non zero

- $d_t \leq d_{p(t)}$

If a node doesn't have a split, then all of its children must also have no splits (implements reduction of nodes)

Parent node of t

Final Objective Function

- Adding the minimal split objective as in OCTs (the misclassification error has already been implemented as part of RE-SVM), we have the following:

$$\delta + c_1 \sum_{i=1}^{n} \sum_{t=1}^{T} e_{it} + c_2 \sum_{i=1}^{n} \sum_{t=1}^{T} \xi_{it} + c_3 \sum_{t=1}^{T} d_t.$$

Sanity constraints

(recall) Binary variable that indicates that the observation i is used in node t

Set of tree levels

- $$\sum_{t \in u} z_{it} = 1, \qquad \forall i = 1, \ldots, N,\ u \in U.$$

An observation is used exactly once in all levels

- $$z_{it} \leq z_{ip(t)}, \qquad \forall i = 1, \ldots, N, t = 2, \ldots, T.$$

If an observation i is not in a node t, then it cannot be in its successors

Sanity constraints

Binary variable that indicates in which half-space the observation i lies regarding the hyperplane at node t

$$\omega_t' x_i + \omega_{t0} \geq -M(1-\theta_{it}), \qquad \forall i = 1, \ldots, N, t = 1, \ldots, T,$$
$$\omega_t' x_i + \omega_{t0} \leq M\theta_{it}, \qquad \forall i = 1, \ldots, N, t = 1, \ldots, T.$$

$$w_t x_i + w_{t0} \geq 0 \Rightarrow \theta_{it} = 1$$

Positive half-space

$$w_t x_i + w_{t0} \leq 0 \Rightarrow \theta_{it} = 0$$

Negative half-space

Sanity Constraints

Set of nodes from left splits (even indexed)

Set of nodes from right splits (odd indexed)

$$z_{ip(t)} - z_{it} \leq \theta_{ip(t)} \;\; \forall i = 1, \ldots, N, t \in \tau_{bl}$$

$$z_{ip(t)} - z_{it} \leq 1 - \theta_{ip(t)} \;\; \forall i = 1, \ldots, N, t \in \tau_{br}$$

respectively

$$\theta_{ip(t)} = 0, z_{ip(t)} = 1 \Rightarrow z_{it} = 1$$

Observation i is inherited to the left child if it lies on the negative half-space

Observation i is inherited to the right child if it lies on the positive half-space

$$
\begin{aligned}
\min\ & \delta + c_1 \sum_{i=1}^{n}\sum_{t=1}^{T} e_{it} + c_2 \sum_{i=1}^{n}\sum_{t=1}^{T} \xi_{it} + c_3 \sum_{t=1}^{T} d_t && \text{(OCTSVM)} \\
\text{s.t.}\ & \frac{1}{2}||\omega_t||_2 \leq \delta, \quad \forall t = 1,\ldots,T, \\
& y_i(\omega_t' x_i + \omega_{t0}) - 2y_i(\beta_t' x_i + \beta_{it0}) \geq 1 - e_{it} - M(1 - z_{it}), && \forall i = 1,\ldots,N, t = 1,\ldots,T, \\
& \beta_{itj} = \xi_{it}\omega_{tj}, \qquad \forall i = 1,\ldots,N, t = 0,\ldots,T,\ j = 0,\ldots,p, \\
& \|\omega_t\|_2 \leq M d_t, \qquad \forall t = 1,\ldots,T, \\
& d_t \leq d_{p(t)}, \qquad \forall t = 1,\ldots,T, \\
& \sum_{t\in u} z_{it} = 1, \qquad \forall i = 1,\ldots,N,\ u \in U, \\
& z_{it} \leq z_{ip(t)}, \qquad \forall i = 1,\ldots,N, t = 2,\ldots,T, \\
& \omega_t' x_i + \omega_{t0} \geq -M(1 - \theta_{it}), \qquad \forall i = 1,\ldots,N, t = 1,\ldots,T, \\
& \omega_t' x_i + \omega_{t0} \leq M\theta_{it}, \qquad \forall i = 1,\ldots,N, t = 1,\ldots,T, \\
& z_{ip(t)} - z_{it} \leq \theta_{ip(t)}, \qquad \forall i = 1,\ldots,N, t \in \tau_{bl}, \\
& z_{ip(t)} - z_{it} \leq 1 - \theta_{ip(t)}, \qquad \forall i = 1,\ldots,N, t \in \tau_{br}, \\
& e_{it} \in \mathbb{R}^+, \beta_{it} \in \mathbb{R}^p, \beta_{it0} \in \mathbb{R}, \xi_{it}, z_{it}, \theta_{it} \in \{0,1\}, \forall i = 1,\ldots,N, t = 1,\ldots,T, \\
& \omega_t \in \mathbb{R}^p, \omega_{t0} \in \mathbb{R}, d_t \in \{0,1\}, \forall t = 1,\ldots,T.
\end{aligned}
$$

Experimental Section

Overview

- The authors test their model alongside other aforementioned paradigms: **CART, OCT, OCT-H**
- They test on 8 different UCI Machine Learning Repository datasets
- To test for noise robustness, they apply different amounts of noise (label flips) in these datasets
- They compare the models in terms of average accuracy, after 10 evaluations using cross-validation

Algorithms

- **CART**: implemented off-the-shelf in sklearn python library (*DecisionTreeClassifier(criterion='gini')*)

- **OCT**: implemented here, based on Bertsimas et al. 2017

- **OCT-H**: a less constrained version of OCT, so we implement it by removing the constraint of using a single variable per node from OCT

- **OCTSVM**: Implemented by coding in the IMNLP formulation using the Gurobi library

To be continued…

See progress in the code in my Github repository

References

[1] V. Blanco, A. Jap´on, and J. Puerto, “Robust optimal classification trees under noisy labels,” arXiv:2012.08560 [cs.LG], Dec. 2020.

[2] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees, 1st ed. Belmont, CA, USA: Wadsworth, 1984.

[3] V. Vapnik, The Nature of Statistical Learning Theory, 1st ed. New York,NY, USA: Springer, 1995.

[4] V. Blanco, A. Jap´on, and J. Puerto, “A mathematical programming approach to SVM-based classification with label noise,” Computers & Industrial Engineering, vol. 172, p. 108611, 2022.

References

[5] H. A. Taha, Operations Research: An Introduction, 10th ed. Upper Saddle River, NJ, USA: Pearson, 2017.

[6] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009.

[7] D. Bertsimas and J. Dunn, “Optimal classification trees,” Machine Learning, vol. 106, no. 7, pp. 1039–1082, 2017