SOFT-MARGIN TROPICAL SUPPORT VECTOR MACHINES

SAMUEL BOÏTÉ, THÉO MOLFESSIS, STÉPHANE GAUBERT, XAVIER ALLAMIGEON

Abstract. We explore the novel application of tropical geometry to the realm of classification problems, specifically through the development of soft-margin Tropical Support Vector Machines. By leveraging the structural properties of the tropical semifield, we extend traditional linear classification methods into the tropical domain. We introduce Shapley operators to describe the convex hulls of data points, providing a geometrical approach for handling overlapping data using tropical projections. We then formulate a pseudo-polynomial algorithm to compute optimal-margin separating tropical hyperplanes. We extend this framework to handle multiple classes at once, showcasing its versatility and potential for various classification tasks. Finally, we introduce the 'kernel trick' in the tropical setting, mapping data points to higher-dimensional spaces through Veronese embeddings. Using tropically polynomial separation surfaces, we emulate neural network structures and offer a promising avenue for complex classification tasks. We present a proof of concept through empirical studies on standard datasets, highlighting the strengths and potential of tropical geometry in machine learning.

1. Introduction

The tropical semifield $\mathbb{R}_{\max}$ is the set of real numbers, completed by $-\infty$ and equipped with the addition $a \oplus b = \max(a, b)$ and the multiplication $a \odot b = a + b$.

The tropical classification problem. Given a set $\mathcal{I}$ of d-dimensional data points x_i with associated labels $y_i \in \{\pm 1\}$, we define the classes C^+ (resp. C^-), consisting of the points with positive (resp. negative) labels. We want to separate our point clouds by a tropical hyperplane. We will extend this framework to d classes of points later.

Definition 1. A *tropical hyperplane of apex* $u \in \mathbb{R}^d_{max}$ splits $\mathbb{R}^d_{\max}$ depending on where $(x - u)$ reaches its maximum coordinate:

$$H_u := \left\{ x \in \mathbb{R}^d_{\max}, \quad (x-u) \text{ reaches its max coordinate at least twice} \right\}.$$

Such an hyperplane splits the space in d different *tropical sectors*, depending on the coordinate maximized. A *tropical halfspace* of configuration $I \subset [d]$ is the union of tropical sectors specified in I.

The *signed tropical hyperplane* of configuration I is the surface between two complementary tropical halfspaces:

$$H^I_u := \left\{ x \in \mathbb{R}^d_{\max}, \quad (x-u) \text{ reaches its max coordinate in } I \text{ and } I^c \right\}.$$

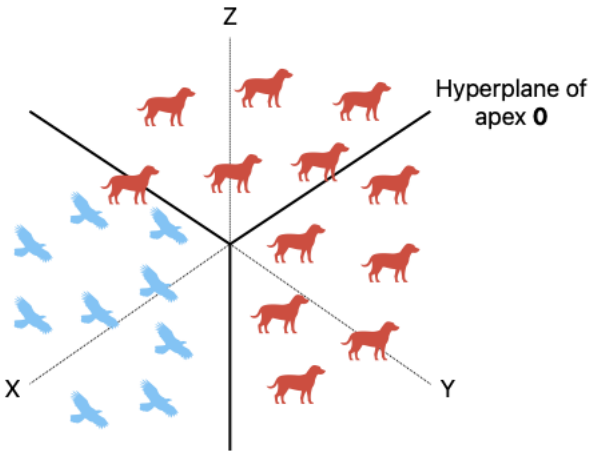

FIGURE 1.1. Binary classification problem

2. PRELIMINARIES

2.1. **General setting.** The tropical binary classification problem itself is, as we will see, a subproblem of a more general one involving the notion of Shapley operators.

Definition 2. A *Shapley operator* on $\mathbb{R}_{\max}$ is a map $T : \mathbb{R}^d_{\max} \longrightarrow \mathbb{R}^p_{\max}$ such that T is non-decreasing and that for all $x \in \mathbb{R}^n_{\max}$ and $\lambda \in \mathbb{R}_{\max}$, $T(\lambda + x) = \lambda + T(x)$. T is said to be *non-expansive* if it is 1-Lipschitz.

Definition 3. An order-preserving and additively homogeneous self-map F of $\mathbb{R}_{\max}$ is said to be *diagonal free* when $F_i(x)$ is independent of x_i for all $i \in [n]$. In other words, F is *diagonal free* if for all $i \in [n]$, and for all $x, y \in \mathbb{R}_{\max}$ such that $x_j = y_j$ for all $j \neq i$, we have $F_i(x) = F_i(y)$.

Let $T^{\pm}$ be two non-expansive Shapley operators. We define

$$V^{\pm} = \{x \in \mathbb{R}^d, \quad x \leq T^{\pm}(x)\},$$

which we want to separate with a tropical signed hyperplane.

Definition 4. For all $u \in \mathbb{R}^d_{\max}$, we define the *tropical norm* of u as the largest difference between two of its coordinates:

$$\|u\| := \max u - \min u.$$

The *tropical distance* between u and v in $\mathbb{R}^d_{\max}$ is then defined as:

$$d(u, v) := \|u - v\|.$$

Definition 5. H^I_u is said to *separate* V^+ *and* V^- *with a margin of at least* $\nu \geq 0$ when:

(1) V^+ and V^- are respectively on either side of the hyperplane H_u^I, i.e.

$$\forall x \in V^+, \qquad \operatorname{argmax}_{i\in[d]} x_i \in I,$$

$$\forall y \in V^-, \qquad \operatorname{argmax}_{i\in[d]} x_i \in I^c,$$

or vice-versa.

(2) Distance from H_u^I to $V^\pm$ is at least ν.

When ν is maximal in the previous definition, we say that H_u^I *separates* $V^\pm$ *with a margin of* ν. When ν is zero, we say that H_u^I *separates* $V^\pm$.

2.2. **Application to tropical convex hulls.** We shall see in this section that tropical projections are the right Shapley operator for describing and separating finite point clouds.

Definition 6. We define the *tropical span* of a finite set of points $A = [a_1|\dots|a_p] \in \mathbb{R}_{\max}^{d\times p}$ as the set of tropical linear combinations of these points:

$$\operatorname{Span}(A) := \left\{ A \odot \Lambda := \lambda_1 \odot a_1 \oplus \cdots \oplus \lambda_p \odot a_p, \quad \Lambda := (\lambda_i)_{i\in[p]} \in \mathbb{R}^p \right\}.$$

Given the tropical convexity of these, these spans are often called *tropically convex hulls*, meaning the smallest tropically convex sets containing them.

Definition 7. Let V a tropically convex, compact and nonempty subset of $\mathbb{R}_{\max}^d$. We define the *projection* of x on V as:

$$P_V(x) := \max\{y \in V, \quad y \le x\}.$$

When $V = \operatorname{Span}(C)$ with C in $\mathbb{R}_{\max}^{d\times p}$, we will write the projection as P_C.

The following property is derived by simple inequalities on P_C's expression with min and max:

Proposition 8. *Let $C \in \mathbb{R}_{\max}^{d\times p}$. P_C is nondecreasing and for $x \in \mathbb{R}_{max}^d$, $P_C(x) \le x$.*

Proposition 9. *(Formula 5.2.3 in* [6]*) We have the following equality:*

$$P_C = CC^\#,$$

where for $i \in [d]$, $j \in [p]$ and $(x, z) \in \mathbb{R}_{\max}^d \times \mathbb{R}_{\max}^p$:

$$(Cz)_i := \max_{j\in[p]}\{C_{ij} + z_j\}$$

$$(C^\# x)_j := \min_{i\in[d]}\{-C_{ij} + x_i\}$$

with the convention that $-(-\infty) + (-\infty) = -\infty$.

Remark 10. Tropical projections are closely tied to mean payoff games. With A and B in $\mathbb{R}_{\max}^{p\times n}$, and $T = A^\# B$:

$$\forall i \in [n], \quad T_i(x) = \min_{j\in[p]} \left\{ -A_{ji} + \max_{k\in[n]}(B_{jk} + x_k) \right\}.$$

In particular, $T_i(0)$ is the payoff, recieved by player MAX, of one round of a zero-sum game with perfect information, where player MIN starts from their state i, transitions to their opponent MAX's state j by receiving A_{ji}, who in turn chooses a MIN state k by recieving B_{jk}.

Therefore, $(T^m(0))_i$ is the *value* of such a game after m consecutive rounds, having started in the MIN state i. The *escape rate* of the game defined by T is defined by:

$$\chi_T := \lim_{m\to+\infty} \frac{T^m(0)}{m}.$$

It is well known that this limit does exist and coincides with the *mean payoff* of the game [4]. Under certain conditions, χ_T is the unique eigenvalue of the Shapley operator T.

Let $P^{\pm}$ the projections associated with finite point clouds $C^{\pm}$.

Operators $P^{\pm}$ can be slightly tweaked to make them *diagonal-free* (DF), a good property that will be of interest to us in the future. In the modified game, the MIN player is prevented from replying to his opponent eye-for-an-eye:

$$\left[P^{\pm}_{\mathrm{DF}}\right] := \max_{j\in[p]} \left\{ C^{\pm}_{ij} + \min_{k\neq i}(-C^{\pm}_{kj} + v_k) \right\}.$$

Proposition 11. *$P^{\pm}_{DF}$ are diagonal-free Shapley operators that equivalently describe $V^{\pm}$.*

3. Separating finite overlapping data

Assuming that the data overlap, we want to transform $V^{\pm}$ slightly so as to make them separable by a tropical hyperplane.

3.1. **Measuring data overlap.**

Lemma 12. *Intersection between convex hulls can be described using the Shapley operator:*

$$V^+ \cap V^- = \left\{ v \in \mathbb{R}^d_{\max}, \quad v \leq T(v) \right\},$$

where $T = T^+ \wedge T^-$.

As we have a Shapley operator, the analogy with game theory applies, and the notions of Collatz-Wielandt numbers, spectral radus and game value are identical [1]. We can then state the following theorem:

Theorem 13. *(Allamigeon, Gaubert et al. [4]) $V^+ \cap V^-$ contains a Hilbert ball of positive radius if and only if the spectral radius of F, defined as*

$$\rho(T) = \sup\{\mu \in \mathbb{R}, \quad \exists z \in \mathbb{R}^d, \quad T(z) = \mu + z\}$$

is strictly positive. In this case, $\rho(F)$ is the inner radius inrad$(V^+ \cap V^-)$, i.e supremum of the radii of the Hilbert balls contained in it.

We can then apply the following algorithm to compute the eigenpair in pseudo-polynomial time.

3.2. **Efficiently finding the eigenpair.** We now describe the following projective Krasnoselskii-Mann iteration algorithm. Given an initial point a^0, we iteratively compute:

$$\begin{cases} z^{k+1} & = \frac{a^k + T(a^k)}{2} \\ a^{k+1} & = z^{k+1} - \max_{i\in[d]} z_i^{k+1} \cdot \mathbf{1}_d \end{cases}$$

The following convergence theorem follows from [5].

Corollary 14. As T is a non-expansive Shapley operator, the Krasnoselskii-Mann algorithm converges in pseudo-polynomial time.

The eigenpair we search is $\left(a^\infty, 2 \cdot \max_{i\in[d]} z_i^\infty\right)$.

3.3. **Separating overlapping finite data.** We place ourselves in the case of separation of finite tropical convex envelopes. We have seen that the corresponding Shapley operator is the diagonal-free operator.

We now define a process for separating overlapping data. Assuming $V^+ \cap V^- \neq \emptyset$, let (a, λ) the eigenpair of T approximated by the iteration algorithm.

Proposition 15. *For all sector $i \in [d]$, we choose one class $\pm_i$ such that $T(a)_i = T^{\pm_i}(a)_i = \lambda + a_i$, (we note $\mp_i$ the other one) and we project all points of $C^{\pm_i}$ located in sector i onto H_a. Then the intersection of new convex hulls $W^\pm$ is of empty interior.*

Proof. Let $Q^\pm$ be the *diagonal-free* operators over transformed point clouds, and $Q = Q^+ \wedge Q^-$. We prove that $Q(a) = a$. We use the following notations for $j \in [p]$:

$$\begin{cases} D_{ij}^\pm := C_{ij}^\pm - a_i \\ m_{kj}^\pm := W_{kj}^\pm - a_k \\ c_{ij}^\pm := W_{ij}^\pm - \max_{k\neq i} m_{kj}^\pm \end{cases}$$

We also write $s^\pm(j)$ the sector of $C_{\cdot j}^\pm$, and $d^\pm(j)$ the second argmax of $D_{\cdot j}^\pm$.

The transformation defined above consists in setting $W_{ij}^{\pm_i} = C_{ij}^{\pm_i} - \mathbb{1}_{i=s^{\pm_i}(j)} \cdot d(C_{\cdot j}^{\pm_i}, H_a)$ and $W_{ij}^{\mp_i} = C_{ij}^{\mp_i}$. It leads by construction to $m_{s^\pm(j)j}^\pm \geq D_{d^\pm(j)j}^\pm$ with equality if the class of $C_{\cdot j}^\pm$ is the one chosen in its sector, which can be noted $\pm = \pm_{s^\pm(j)}$.

We can also write that $Q^\pm(a)_i = \max_{j\in[p]} c_{ij}^\pm$.

Let $i \in [d]$. If $C_{\cdot,j}^\pm$ isn't in the i-th sector, then for $k \neq s^\pm(j)$, we have $m_{kj}^\pm \leq D_{d^\pm(j)j}^\pm \leq m_{s^\pm(j)j}^\pm$. Therefore, in that case, $c_{ij}^\pm = D_{ij}^\pm - m_{s^\pm(j)j}^\pm + a_i \leq a_i$.

Otherwise, $s^\pm(j) = i$ and $\max_{k\neq i} m_{kj}^\pm = D_{d^\pm(j)j}^\pm$, thus $c_{ij}^\pm = m_{s^\pm(j)j}^\pm - D_{d^\pm(j)j}^\pm + a_{s^\pm(j)} \geq a_{s^\pm(j)} = a_i$, with equality if $\pm = \pm_{s^\pm(j)}$.

Lastly, using the proof of Theorem 22 in [2], we know that there exists $j^{\pm_i}, j^{\mp_i} \in [p]$ such that $C_{\cdot j^{\pm}}^\pm$ are in sector i. Therefore, $c_{ij^{\pm_i}}^{\pm_i} = a_i$ and $c_{ij^{\mp_i}}^{\mp_i} \geq a_i$. Hence $Q(a)_i = Q^{\pm_i}(a)_i = a_i$, and finally $Q(a) = a$.

□

Example 16. Here is what the transformation yields with a toy inseparable dataset:

Remark 17. We would have to deal with the branching points that don't contribute to increasing the interior of the intersection.

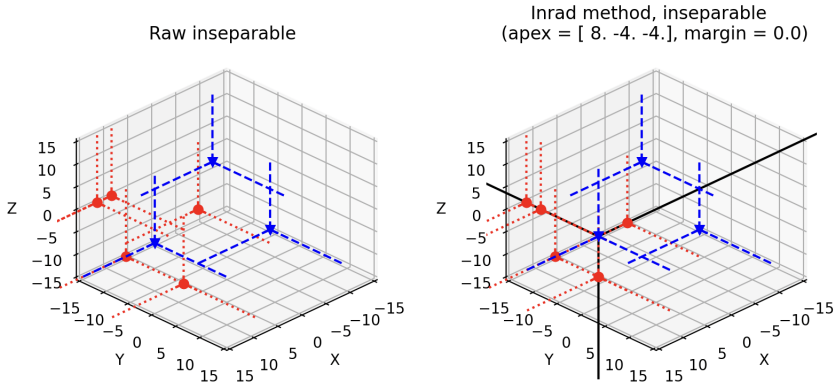

FIGURE 3.1. Separating convex hulls

4. Optimal-margin separating hyperplanes

Assuming that data is now separable, we prove that the previous method can be extended to give us separating hyperplanes with maximal margin.

Remark 18. When V^+ and V^- are disjoint, the spectral radius λ of T is strictly negative. From [3], we may see $-\lambda$ as the eigenvalue an operator T^{dual} in dual space, as T is itself a finitely-generated Shapley operator. This would give us a complementary inner radius interpretation, and makes H_a a great candidate in the separable case.

Let (a, λ) the eigenpair of T approximated by previous algorithm, verifying $T(a) = \lambda + a$. Let's define the sectors:

$$I^{\pm} := \{i \in [d], \quad T^{\pm}(a)_i > \lambda + a_i\}.$$

Proposition 19. *H_a^I, given the sectors defined above, separates V^+ and V^- with a margin of $-\lambda$. Moreover, this margin is optimal in the case where $T^{\pm}$ are of the form $T^{\pm}(x) = P_{V^{\pm}}(x) = \sup_{v \in V^{\pm}} (v_i + \min(-v + x))$, which is in particular the case when separating finite point clouds.*

Proof. As $T^{\pm}$ is non-expansive, let's first remark that for $x^{\pm} \in V^{\pm}$:

$$x_i^{\pm} \leq T^{\pm}(x^{\pm})_i = \left(T^{\pm}(x^{\pm}) - T^{\pm}(a)\right)_i + T^{\pm}(a)_i,$$

hence

$$x_i^{\pm} \leq \max(x^{\pm} - a) + T^{\pm}(a)_i. \tag{4.1}$$

For instance, let $i \in [d] \setminus I^+$. Then $T^+(a)_i = \lambda + a_i$, so for $x^+ \in V^+$, using equation 4.1:

$$x_i^+ - a_i \leq \max(x^+ - a) + \lambda.$$

In particular, $x_i^+ - a_i < \max(x^+ - a)$ and any element of V^+ can't belong to any of sectors in $[d] \setminus I^+$ with respect to H_a, from which the sectors $I^{\pm}$ are well-defined.

Finally,

$$d(H_a^I, x^+) = \max(x^+ - a) - \max(x^+ - a)_{[d]\setminus I^+} \geq -\lambda,$$

and the margin comes from the fact that this applies to any element of V^+.

Let's finally prove that the margin is maximal in the case where $T^\pm$ are of the form $T^\pm(x) = P_{V^\pm}(x) = \sup_{v \in V^\pm}(v_i + \min(-v + x))$. Let $i \in [d] \setminus I^+$. Then, for $\varepsilon > 0$, we can find $v \in V^+$ such that

$$T^+(a)_i - \varepsilon \leq v_i - \max(v - a) \leq T^+(a)_i,$$

giving us

$$\lambda - \varepsilon \leq v_i - a_i - \max(v - a) \leq \lambda.$$

Maximizing over all $i \in [d] \setminus I^+$ yields that v is at most at distance $-\lambda + \varepsilon$ of H_a^I, hence the optimality.

□

Example 20. Here is what the algorithm gives with a toy separable dataset:

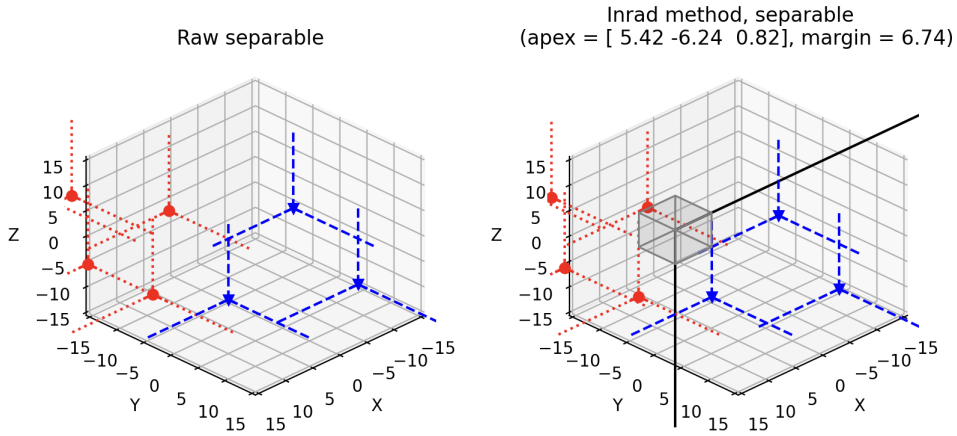

FIGURE 4.1. Optimal-margin separable hyperplane

5. Multi-class optimal-margin separation

In this section, we consider convex hulls of point clouds of D classes, noted V^k for $k \in [D]$, and described by Shapley operators T^k. We give a sufficient condition for these classes to be tropically separable in their whole, meaning that there exists a tropical signed hyperplane such that each V^k belong to sectors of $I^k \subset [d]$ with $I^k \cap I^l = \emptyset$ for $k \neq l \in [D]$. We then adapt the previous results to this case.

We now consider the Shapley operator :

$$T := \bigvee_{1 \leq k < l \leq D} T^k \wedge T^l$$

We remark that in the case $D = 2$, T is the same as previously defined.

Let (a, λ) the eigenpair of T approximated by the Kranoselskii-Mann algorithm, verifying $T(a) = \lambda + a$. Let's define the sectors:

$$I^k := \{i \in [d], \quad T^k(a)_i > \lambda + a_i\}.$$

Given the definition of T, it is clear that $I^k \cap I^l = \emptyset$ for $k \neq l \in [D]$, and we have the following result :

Proposition 21. *If $\lambda < 0$, the signed tropical hyperplane H_a^I, given the sectors defined above, separates V^k and V^l for all $k \neq l \in [D]$ with a margin of $-\lambda$. Moreover, this margin is optimal in the case where all T^k are of the form $T^k(x) = P_{V^k}(x) = \sup_{v \in V^k} (v_i + \min(-v + x))$, which is in particular the case when separating finite point clouds.*

Proof. Let $k \in [D]$ and $i \in [d] \setminus I_k$. Using the same reasoning as in the proof of Proposition 19, we obtain that for all $x^k \in V^k$,

$$d(H_a^I, x^k) = \max(x^k - a) - \max(x^k - a)_{[d]\setminus I_k} \geq -\lambda,$$

hence the margin. Let's now prove the optimality in the case where all T^k are of the form $T^k(x) = P_{V^k}(x) = \sup_{v \in V^k} (v_i + \min(-v + x))$. For all $i \in [d]$, there are two distinct classes $k \neq l \in [D]$ such that $T^k(a)_i \wedge T^l(a)_i = \lambda + a_i$. As $I^k \cap I^l = \emptyset$, we can suppose by symmetry that $i \in [d] \setminus I_k$. Then using the same argument as in the proof of optimality in Proposition 19, we know that for all $\varepsilon > 0$ there is a point $v^k \in V^k$ such that $\max(v^k - a) - (v_i^k - a_i) \leq -\lambda + \varepsilon$, which means that

$$d(H_a^I, x^k) = \max(x^k - a) - \max(x^k - a)_{[d]\setminus I_k} \leq -\lambda + \varepsilon.$$

As this holds for every sector $i \in [d]$, we have proven the optimality. (???) □

6. Kernel trick: tropically polynomial decision boundaries

In the classical setting, the kernel trick consists in mapping the training points in a higher-dimensional space in which we expect the data to become easily linearly separable. In this paragraph, we generalize this idea to integer combinations of features in the tropical setting.

If $\mathcal{A} \subset \mathbb{Z}^d$ is a set of vectors, we define the *veronese embedding* of x as

$$\mathrm{ver}(x) = (\langle x, \alpha \rangle)_{\alpha \in \mathcal{A}} \in \mathbb{R}^{\mathcal{A}},$$

allowing us to map our point space into a larger space, made up of various integer combinations of features. Noting $s \in \mathbb{N}$ a scale parameter, and Δ_d the d-dimensional simplex, we take

$$\mathcal{A}_s := (s\Delta_d) \cap \mathbb{Z}^d.$$

Proposition 22. *Applying the previous method to point clouds $ver(C_i)$ for each class i yields a classifier whose decision boundaries, when seen in the initial vector space, are tropical polynomials.*

Feedforward neural networks with rectified linear units are, modulo trivialities, nothing more than tropical rational maps [7], i.e differences of tropical polynomials. Thus, the framework we have chosen is close to that of dense neural networks, and should yield similar results for classic datasets.

We therefore consider:

- The *Iris flower dataset*, which includes measurements such as sepal length, sepal width, petal length, and petal width, across three species of Iris (Iris setosa, Iris virginica, and Iris versicolor). This dataset is commonly used for classification tasks and testing clustering algorithms.
- The *Breast Cancer Wisconsin dataset*, which contains features computed from a digitized image of a fine needle aspirate of a breast mass. The features describe characteristics of the cell nuclei present in the image, and the dataset is primarily used for binary classification tasks to distinguish between malignant and benign tumors.
- The *Wine quality and type dataset*, which comprises physicochemical tests (like alcohol content, acidity, sugar level, etc.) and sensory information (quality score) for various samples of red and white wines. This dataset is often employed for regression tasks to predict wine quality or for classification tasks to differentiate between wine types.
- The *FIFA 2022 cards dataset*, which includes skill ratings of players featured in the game and playing positions. This dataset can be used to predict positions on the field.

These datasets have the characteristic of having features that are comparable with each other, which makes them suitable for the tropical framework – although this should work in any case with the kernel trick, since we're describing objects analogous to DNNs.

So as not to bother with branches in the non-separable case, we assign the sectors to the majority population, which in particular settles these borderline cases.

Dataset	Groups	d	p
Iris	Setosa, Virginica, Versicolor	4	150
Cancer	Malign, Benign	31	570
Wine	Red (bad), Red (good), White (bad), White (good)	10	6500
FIFA	Striker, Center back, Center mid, Goal	34	19240

FIGURE 6.1. Datasets description

The sets A_s grow exponentially with the dimension and polynomially with s, and encode the complexity of the mimicked neural network. s will therefore be a relevant hyperparameter of overfitting or underfitting.

Dataset	Hyperplane		$\mathcal{A}_1$		$\mathcal{A}_2$		$\mathcal{A}_3$		$\mathcal{A}_4$	
	1v1	*1vR*	*1v1*	*1vR*	*1v1*	*1v1*	*1v1*	*1vR*	*1v1*	*1vR*
Iris	70%	83%	70%	87%	93%	93%	90%	87%	87%	87%
Cancer	90%	89%	90%	89%	90%	91%	88%	86%	⋯	⋯
Wine	67%	57%	67%	56%	72%	74%	78%	85%	87%	92%
FIFA	78%	69%	78%	69%	79%	77%	⋯	⋯	⋯	⋯

FIGURE 6.2. Accuracies for one-vs-one and one-vs-rest classifiers

In the table, the dots indicate overfitting, or that the dimension has become too large for the calculations to succeed.

The addition of complexity by our kernel method seems particularly effective for the wine dataset.

We note that irrelevant features are driven out of the decision process by a correspondingly large coordinate in the apex, preventing this feature from being maximal. This could guide us towards heuristics to remove superfluous features, combat overfitting and reduce training time.

References

[1] Marianne Akian, Stephane Gaubert, and Alexander Guterman. Tropical polyhedra are equivalent to mean payoff games. *International Journal of Algebra and Computation*, 22(01):1250001, Feb 2012.

[2] Marianne Akian, Stéphane Gaubert, Yang Qi, and Omar Saadi. Tropical linear regression and mean payoff games: or, how to measure the distance to equilibria. *SIAM J. Discret. Math.*, 37:632–674, 2021.

[3] Marianne Akian, Stephane Gaubert, and Sara Vannucci. Ambitropical geometry, hyperconvexity and zero-sum games, 2021.

[4] Xavier Allamigeon, Stephane Gaubert, Ricardo D. Katz, and Mateusz Skomra. Condition numbers of stochastic mean payoff games and what they say about nonarchimedean semidefinite programming. 2018.

[5] R. Cominetti, J. A. Soto, and J. Vaisman. On the rate of convergence of krasnosel'skiĭ-mann iterations and their connection with sums of bernoullis. *Israel Journal of Mathematics*, 199(2):757–772, Aug 2013.

[6] Diane Maclagan and Bernd Sturmfels. *Introduction to Tropical Geometry*. American Mathematical Society, Apr 2015.

[7] Liwen Zhang, Gregory Naitzat, and Lek-Heng Lim. Tropical geometry of deep neural networks, 2018.