Les bases mathématiques du perceptron

Sylvie Benzoni-Gavage

12 février 2022

1 Introduction

Le « perceptron » est un concept introduit à la fin des années 1950 par Frank Rosenblatt comme un « système nerveux hypothétique », inspiré de ce que l'on savait à l'époque de la manière dont le cerveau humain perçoit le monde extérieur [5]. Ce concept avait été mis en application dans une machine électronique conçue pour la reconnaissance d'images (bien avant l'avénement de l'imagerie numérique !). Même si le « Mark 1 perceptron » n'a pas tenu toutes ses promesses [3], il a en quelque sorte signé la naissance de ce que l'on appelle l'*intelligence artificielle* (IA).

De fait le perceptron se trouve encore aujourd'hui à la base des *algorithmes d'apprentissage* utilisés en IA. Plus précisément, il s'agit de *classifier des données* [1] selon une *valeur booléenne* : par exemple vrai ou faux, oui ou non, positif ou négatif, blanc ou noir, ... cela n'a pas d'importance pourvu qu'il n'y ait que deux valeurs possibles [2]. C'est ainsi que finalement une IA « apprend » à reconnaître si une image représente un chat ou pas, si un résultat d'analyse médicale indique que le patient est malade ou pas, etc.

On entend beaucoup parler de *données*, dans toutes sortes de domaines. L'important pour le traitement informatique des données est qu'elles soient représentées par des *valeurs numériques*. Plus précisément, chaque donnée peut être représentée par plusieurs valeurs numériques, et le type de valeurs numériques doit être le même pour tout le *jeu de données* considéré. Si l'on considère par exemple des données sous forme d'images numériques en couleur, de taille fixe en nombre de pixels, chaque donnée peut être constituée des niveaux de rouge, de vert et de bleu sur chaque pixel. S'il y a n pixels, une donnée est alors représentée par $3n$ valeurs numériques. Autrement dit, dans cet exemple les données se trouvent dans un espace de dimension $3n$.

Classifier des données revient à les séparer en deux groupes, c'est-à-dire à affecter une valeur booléenne à chacune d'entre elles. Sur l'exemple des images, une intelligence humaine est en principe capable de dire pour chaque image si elle représente un chat ou non. Alors la valeur booléenne sera oui dans le premier cas et non dans le second.

1. Il serait peut-être plus correct de dire «classer». On dit aussi «étiqueter» des données.
2. On parlera plus loin de *dichotomie*.

Un *algorithme d'apprentissage supervisé* a pour objet de classifier correctement un jeu de données par rapport aux valeurs booléennes connues, pour ensuite déterminer la valeur booléenne de nouvelles données. Toujours sur l'exemple des images, l'algorithme « s'entraîne » sur un jeu d'images connues puis il « reconnaît » si une nouvelle image représente un chat ou non.

L'idée qui sous-tend le perceptron (du moins dans la présentation que nous en faisons ici, de fait assez éloignée de celle de Rosenblatt) est de séparer les jeux de données d'entraînement de manière géométrique. Pour fixer les idées, si les valeurs booléennes de ces données sont les signes + et −, l'idée est de séparer les données affectées du signe + de celles avec le signe − par une frontière géométrique la plus simple possible dans l'espace où elles se trouvent. Comme indiqué précédemment, il s'agit par exemple d'un espace de dimension $3n$ pour des images en couleur à n pixels. Même si l'on ne voit guère a priori comment séparer géométriquement les images de chat des images de non-chat, la séparation géométrique des données est une idée simple et féconde, à la base des techniques d'IA, dont on détaille quelques aspects dans ce qui suit.

La partie 2 est consacrée d'une part à l'étude des conditions dans lesquelles un jeu de données peut être séparé géométriquement par une frontière « simple », à commencer par une frontière plane, et d'autre part à la combinatoire associée. On y explique notamment le fait qu'un espace de dimension d a une certaine *capacité* à accueillir des données séparables. Cette capacité est un seuil critique du nombre N de données séparables dont on montre qu'il est de l'ordre de $2d$ lorsque N est grand. C'est peut-être le résultat le plus marquant dans l'article fondateur [2] du théoricien de l'information Thomas Cover.

La partie 3 décrit et démontre un autre résultat de Cover : la convergence de l'algorithme d'apprentissage appelé « *perceptron learning algorithm* »(PLA), qui permet de déterminer la frontière séparant des données connues pour être séparables. C'est ce genre d'algorithme qui permet à une IA de « reconnaître » la valeur booléenne d'une nouvelle donnée : il suffit en effet de déterminer de quel côté de la frontière elle se situe.

Si les algorithmes d'aujourd'hui sont plus compliqués et de plus en plus sophistiqués, les bases mathématiques du perceptron sont accessibles à des étudiant·es de licence. L'objectif de cette note est d'en donner une présentation adaptée à des étudiant·es de mathématiques. La lectrice et le lecteur intéressé·es seront alors armé·es pour aborder des présentations plus technologiques et/ou des développements plus récents en matière d'intelligence artificielle (réseaux de neurones, apprentissage profond...).

2 Classification de données

Mathématiquement parlant, un *jeu de données* est un ensemble fini de points dans un espace de dimension finie. Par exemple cet ensemble peut être constitué d'un certain nombre de données quantitatives concernant une population donnée : chaque point correspond alors à un individu de cette population et pour chacun d'entre eux on considère par exemple l'âge, la taille, le poids (la masse), la valeur de son compte en banque. Sur cet exemple légèrement provocateur, la dimension de l'espace est 4 (car il y a quatre valeurs par individu), et l'on

voit que les données sont de nature bien différente puisqu'elles se mesurent dans des unités différentes (respectivement en années, centimètres, kilogrammes et euros, par exemple).

D'un point de vue pratique les données sont en général dans un espace affine, au sens où une valeur nulle de l'une de ces données n'a pas forcément un sens intrinsèque : penser par exemple à des températures en degrés Celsius ou Fahrenheit, unités de mesure liées par une relation affine. Cependant on ne perdra pas de généralité à considérer des jeux de données dans un espace vectoriel, qu'on identifiera simplement à $\mathbb{R}^d$ s'il est de dimension d.

Chaque donnée est alors un vecteur de $\mathbb{R}^d$ (dans l'exemple donné au début de cette partie, c'est un vecteur dont les composantes sont l'âge, la taille, la masse et la fortune de l'individu), mais on parlera souvent encore de point par abus de langage. Dans ce contexte, un jeu de N données est un sous-ensemble fini $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ de points de $\mathbb{R}^d$.

Pour des raisons diverses on s'intéresse à *classifier* les données. Cela signifie que l'on considère une valeur booléenne supposément attachée à chaque donnée. Si l'on poursuit sur notre exemple, cette valeur booléenne peut être le fait d'être citadin ou non (à condition de définir précisément ce que cela signifie), ce que l'on peut représenter par un signe $+$ ou $-$ (un individu citadin est affecté du signe $+$ et un non citadin du signe $-$, ce choix étant parfaitement arbitraire). Du point de vue mathématique, un jeu de données $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ est classifié si l'on peut associer à chaque élément $\mathbf{x}_i$ un signe $+$ ou $-$. Autrement dit, la classification d'un jeu de données $X \subset \mathbb{R}^d$ est déterminée par une application de X dans l'ensemble à deux éléments $\{-, +\}$. Une telle application s'appelle alors une *dichotomie* de X.

Le dénombrement des dichotomies relève d'un calcul rapide : pour chaque $\mathbf{x}_i$ deux choix de signe sont possibles, et les choix sont indépendants les uns des autres. Au final (quitte à s'en convaincre par une récurrence immédiate sur N), on obtient que le nombre de dichotomies d'un jeu de N données est 2^N. Ce nombre est indépendant de la dimension de l'espace dans lequel se trouvent ces données.

Si N est grand, le nombre de dichotomies est colossal. On connaît l'histoire de l'échiquier de Sissa, qui montra aux dépens de son roi que 2^{64} grains de riz dépassaient très largement toute la production du royaume (la valeur précise étant $2^{64} = 18\ 446\ 744\ 073\ 709\ 551\ 616$). Or $N = 64$ est un nombre très modeste au regard des enjeux actuels de traitement de données. Les données recueillies actuellement par les géants du numérique se comptent par milliards. Pour un jeu d'un milliard de données par exemple, le nombre de dichotomies est $2^{10^9} > (10^3)^{10^6} = (10^{100})^{300\ 000}$ (un *gogol* à la puissance trois cent mille !).

2.1 Dichotomies linéairement séparables

L'idée de base du perceptron est de s'intéresser à des dichotomies particulières de jeux de données, pour lesquelles les données affectées du signe $-$ et celles affectées du signe $+$ se séparent facilement en deux «camps». Comme sur un terrain de foot avant le coup de sifflet sonnant le début du match, on peut imaginer des dichotomies pour lesquelles les points $-$ et $+$ sont séparés par une ligne en dimension deux, et plus généralement un hyperplan en dimension quelconque.

On peut aussi imaginer la séparation des données par une ligne courbe et plus généralement

par une hypersurface, ce qui est même plus naturel pour des données de nature différente (comme dans notre exemple avec des âges, des tailles, des masses et des sommes d'argent), mais restons pour le moment sur les séparations linéaires.

Définition 1. *Une dichotomie $f : X \to \{-,+\}$ d'un jeu de données $X \subset \mathbb{R}^d$ est dite* linéairement séparable *si les ensembles $X^- := f^{-1}(-)$ et $X^+ := f^{-1}(+)$ sont séparés par un hyperplan, c'est-à-dire s'il existe $\mathbf{w} \in \mathbb{R}^d$ tel que*

$$X^- \subset \{\mathbf{x} \in \mathbb{R}^d\,;\ \mathbf{w}\cdot\mathbf{x} < 0\} \quad \textit{et} \quad X^+ \subset \{\mathbf{x} \in \mathbb{R}^d\,;\ \mathbf{w}\cdot\mathbf{x} > 0\},$$

où le point $\cdot$ désigne le produit scalaire usuel de $\mathbb{R}^d$.

Remarque 1. *On pourrait également considérer des dichotomies séparables par un* hyperplan affine*, d'équation $\mathbf{w}\cdot\mathbf{x} = b$ avec b non nul.*[3] *Cependant, cela ne modifierait pas la théorie, quitte à augmenter la dimension de l'espace. En effet, il suffirait de considérer les données $\{(1,\mathbf{x}_1),\cdots,(1,\mathbf{x}_N)\}$ dans $\mathbb{R}^{d+1}$, que l'on chercherait à séparer par un hyperplan $(-b,\mathbf{w})^\perp$.*

Une caractérisation des dichotomies linéairement séparables peut s'exprimer à l'aide de la notion d'enveloppe convexe. Rappelons qu'un *convexe* de $\mathbb{R}^d$ est un sous-ensemble C tel que pour tout couple $(\mathbf{x},\mathbf{y})$ d'éléments de C le segment $[\mathbf{x},\mathbf{y}] := \{\mathbf{z} = \theta\mathbf{x} + (1-\theta)\mathbf{y}\,;\ \theta \in [0,1]\}$ d'extrémités $\mathbf{x}$ et $\mathbf{y}$ est inclus dans C. L'*enveloppe convexe* $\mathrm{conv}(Y)$ d'un sous-ensemble Y de $\mathbb{R}^d$ est le plus petit convexe contenant Y. Elle est caractérisée par le théorème de Carathéodory, qui montre que

$$\mathrm{conv}(Y) \;=\; \left\{ \mathbf{z} \in \mathbb{R}^d\,;\ \exists (\mathbf{y}_j) \in Y^{d+1}\,,\ \exists (\theta_j) \in [0,1]^{d+1}\,;\ \mathbf{z} = \sum_{j=1}^{d+1} \theta_j \mathbf{y}_j\,,\ 1 = \sum_{j=1}^{d+1} \theta_j \right\}.$$

Théorème 1. *Soit f une dichotomie d'un jeu de données X de $\mathbb{R}^d$. Elle est linéairement séparable si et seulement si les enveloppes convexes de $X^- = f^{-1}(-)$ et $X^+ = f^{-1}(+)$ sont disjointes.*

Démonstration. Supposons qu'il existe $\mathbf{x} \in \mathrm{conv}(X^-) \cap \mathrm{conv}(X^+)$. D'après le théorème de Carathéodory il existe alors $(\mathbf{x}_j^\pm) \in X_\pm^{d+1}$ et $(\theta_j^\pm) \in [0,1]^{d+1}$ tels que

$$\mathbf{x} = \sum_{j=1}^{d+1} \theta_j^- \mathbf{x}_j^- = \sum_{j=1}^{d+1} \theta_j^+ \mathbf{x}_j^+\,, \quad 1 = \sum_{j=1}^{d+1} \theta_j^\pm.$$

Si f était linéairement séparable il existerait $\mathbf{w} \in \mathbb{R}^d$ tel pour tout j, $\mathbf{w}\cdot\mathbf{x}_j^- < 0$ et $\mathbf{w}\cdot\mathbf{x}_j^+ > 0$, ce qui entraînerait à la fois $\mathbf{w}\cdot\mathbf{x} < 0$ (car au moins l'un des θ_j^- est strictement positif) et $\mathbf{w}\cdot\mathbf{x} > 0$ (car au moins l'un des θ_j^+ est strictement positif).

Par suite, si f est linéairement séparable on a nécessairement $\mathrm{conv}(X^-) \cap \mathrm{conv}(X^+) = \emptyset$.

Réciproquement, si $\mathrm{conv}(X^-) \cap \mathrm{conv}(X^+) = \emptyset$, le théorème de Hahn-Banach assure que $\mathrm{conv}(X^-)$ et $\mathrm{conv}(X^+)$ sont strictement séparables par un hyperplan, donc a fortiori X^- et X^+ aussi, ce qui signifie que la dichotomie f est linéairement séparable. □

3. C'est d'ailleurs souvent le cas en science des données, où b s'appelle un *biais*, ou encore un *seuil*.

On cherche maintenant et surtout à dénombrer les dichotomies linéairement séparables d'un jeu de données fixé. Ceci revient à compter le nombre d'éléments de $\{-,+\}^N$ obtenus comme un N-uplet $(f(\mathbf{x}_1),\dots,f(\mathbf{x}_N))$ où f est une dichotomie linéairement séparable du jeu de données $X=\{\mathbf{x}_1,\cdots,\mathbf{x}_N\}\subset\mathbb{R}^d$. Notons $C(N,d)$ ce nombre, qui dépend a priori de X.

On remarque tout d'abord que le décompte, c'est-à-dire le calcul de $C(N,d)$, risque d'être très compliqué si les vecteurs $\mathbf{x}_i$ sont «trop liés» entre eux. Imaginons par exemple que deux d'entre eux soient colinéaires et de même sens. Il n'y a alors pas de dichotomie linéairement séparable qui affecte des signes opposés à ces deux éléments de X. Afin de pas avoir à tenir compte de tous les cas particuliers possibles, on fera toujours l'hypothèse que le jeu de données est «en position générale» selon la définition suivante.

Définition 2. *Un jeu de données X dans $\mathbb{R}^d$ est dit* en position générale *si toute famille d'au plus d éléments de X est libre.*

On va voir que $C(N,d)$ ne dépend pas de X lorsque X est en position générale.

On se convainc sans trop de mal que $C(N,d)$ est strictement inférieur à 2^d en général. Considérons par exemple le cas $d=2$ et $N=3$. Alors les trois vecteurs $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ sont liés, et d'après l'hypothèse qu'ils sont en position générale, deux quelconques d'entre eux sont indépendants. On peut donc supposer sans perte de généralité (quitte à permuter les indices) que $\mathbf{x}_2$ s'écrit

$$\mathbf{x}_2=a_1\mathbf{x}_1+a_3\mathbf{x}_3$$

avec $a_1a_3\neq 0$. On peut alors toujours mettre en évidence deux dichotomies (sur les $2^3=8$ dichotomies possibles) qui ne sont pas linéairement séparables. Car on est forcément dans l'un des trois cas suivants, pour n'importe quel vecteur $\mathbf{w}$ non nul.

— Si $a_1a_3>0$, on ne peut pas avoir $(\mathbf{w}\cdot\mathbf{x}_1,\mathbf{w}\cdot\mathbf{x}_2,\mathbf{w}\cdot\mathbf{x}_3)=(-,+,-)$ ni $(-,+,-)$.
— Si $a_1<0<a_3$, on ne peut pas avoir $(\mathbf{w}\cdot\mathbf{x}_1,\mathbf{w}\cdot\mathbf{x}_2,\mathbf{w}\cdot\mathbf{x}_3)=(-,-,+)$ ni $(+,+,-)$.
— Si $a_3<0<a_1$, on ne peut pas avoir $(\mathbf{w}\cdot\mathbf{x}_1,\mathbf{w}\cdot\mathbf{x}_2,\mathbf{w}\cdot\mathbf{x}_3)=(-,+,+)$ ni $(+,-,-)$.

Théorème 2 (Cover [2])**.** *Le nombre $C(N,d)$ de dichotomies linéairement séparables d'un jeu de N données en position générale dans $\mathbb{R}^d$ est donné par*

$$C(N,d)=2\sum_{k=0}^{d-1}\binom{N-1}{k}, \tag{1}$$

où

$$\binom{N-1}{k}=\frac{(N-1)!}{k!(N-1-k)!}\ \textit{si}\ k\leqslant N-1\,,\ 0\ \textit{sinon.}$$

Remarque 2. *La formule* (1) *montre en particulier que pour tout $d\geqslant N$, le nombre $C(N,d)$ est maximal et égal à 2^N, car dans ce cas*

$$C(N,d)=2\sum_{k=0}^{N-1}\binom{N-1}{k}=2(1+1)^{N-1}$$

par la formule du binôme.

La démonstration du théorème 2 pourra se faire par récurrence sur N lorsque nous aurons montré la formule

$$C(N+1,d) = C(N,d) + C(N,d-1). \tag{2}$$

Notons que celle-ci est exactement la même que pour les coefficients du binôme

$$\mathrm{C}_N^d = \binom{N}{d},$$

qui vérifient en effet

$$\binom{N+1}{d} = \binom{N}{d} + \binom{N}{d-1}. \tag{3}$$

Seule l'initialisation change, ce qui montre à quel point (pour qui ne serait pas familier du raisonnement par récurrence) cette étape est importante ! Ici l'initialisation démarre à $N = 1$ (on ne considère pas une absence de données comme un cas intéressant). Quel que soit $d \geqslant 1$, on a deux dichotomies pour une seule donnée $\mathbf{x}_1$ dans $\mathbb{R}^d$, et elles sont toutes deux linéairement séparables (il suffit de considérer n'importe quel hyperplan ne contenant pas $\mathbf{x}_1$), c'est-à-dire que $C(1,d) = 2$ pour tout $d \geqslant 1$.

Remarque 3. *Parmi les autres valeurs faciles à calculer, il y a le cas $d = 1$. En effet, il y a exactement deux dichotomies linéairement séparables de N points non nuls de $\mathbb{R}$: celle qui prend le même signe qu'eux, et celle qui prend le signe opposé. Donc $C(N,1) = 2$ pour tout $N \geqslant 1$.*

Remarque 4. *Grâce aux valeurs $C(1,d) = 2$ pour tout $d \geqslant 1$ et $C(N,1) = 2$ pour tout $N \geqslant 1$, la formule de récurrence (2) permet d'établir un tableau analogue à celui de Pascal.*

N									
1	2	2	2	2	2	2	2	2	…
2	2	4	4	4	4	4	4	4	…
3	2	6	8	8	8	8	8	8	…
4	2	8	14	16	16	16	16	16	…
5	2	10	22	30	32	32	32	32	…
6	2	12	32	52	62	64	64	64	…
⋮	2	⋮					⋱		
⋮	2	$2N$	…					2^N	…
d	1	2	3	4	5	6	…	N	…

La preuve de la formule de récurrence (2) sera facilitée par le résultat préliminaire suivant.

Lemme 1. *Soit $f : X \to \{-,+\}$ une dichotomie d'un sous-ensemble fini X de $\mathbb{R}^d$. Soient $\mathbf{y}$ un vecteur non nul de $\mathbb{R}^d$ et $f_\pm$ prolongeant f à $X \cup \{\mathbf{y}\}$ par $f_\pm(\mathbf{y}) = \pm$. Alors f^- et f^+ sont toutes deux des dichotomies linéairement séparables de $X \cup \{\mathbf{y}\}$ si et seulement si f est linéairement séparable par un hyperplan contenant $\mathbf{y}$, c'est-à-dire s'il existe $\mathbf{w} \in \mathbb{R}^d$ tel que $\mathbf{w} \cdot \mathbf{y} = 0$ et*

$$X^- = f^{-1}(-) \subset \{\mathbf{x} \in \mathbb{R}^d ;\ \mathbf{w} \cdot \mathbf{x} < 0\} \quad \textit{et} \quad X^+ = f^{-1}(+) \subset \{\mathbf{x} \in \mathbb{R}^d ;\ \mathbf{w} \cdot \mathbf{x} > 0\},$$

Démonstration. Les applications f^- et f^+ sont des dichotomies par construction. Supposons qu'elles soient linéairement séparables au moyen de vecteurs $\mathbf{w}^-$ et $\mathbf{w}^+$ respectivement. Soit alors

$$\mathbf{w} := (\mathbf{w}^+ \cdot \mathbf{y})\mathbf{w}^- - (\mathbf{w}^- \cdot \mathbf{y})\mathbf{w}^+.$$

Par construction on a $\mathbf{w} \cdot \mathbf{y} = 0$ et

$$\mathbf{w} \cdot \mathbf{x} = (\mathbf{w}^+ \cdot \mathbf{y})(\mathbf{w}^- \cdot \mathbf{x}) - (\mathbf{w}^- \cdot \mathbf{y})(\mathbf{w}^+ \cdot \mathbf{x})$$

est strictement négatif pour tout $\mathbf{x} \in X^-$, comme somme de deux termes strictement négatifs, tandis que $\mathbf{w} \cdot \mathbf{x}$ est strictement positif pour tout $\mathbf{x} \in X^+$, comme somme de deux termes strictement positifs. Réciproquement, supposons qu'il existe $\mathbf{w}$ tel que $\mathbf{w} \cdot \mathbf{y} = 0$, $\mathbf{w} \cdot \mathbf{x} < 0$ pour tout $x \in X^-$ et $\mathbf{w} \cdot \mathbf{x} > 0$ pour tout $x \in X^+$. Les ensembles $X_\pm$ étant finis, il existe $\varepsilon > 0$ tel qu'en notant $\mathbf{w}_\pm = \mathbf{w} \pm \varepsilon \mathbf{y}$ on ait les mêmes inégalités strictes en remplaçant $\mathbf{w}$ par $\mathbf{w}_\pm$. De plus le signe de $\mathbf{w}_\pm \cdot \mathbf{y}$ est $\pm$ par construction. Ceci montre que f^- et f^+ sont linéairement séparables, respectivement au moyen des vecteurs $\mathbf{w}^-$ et $\mathbf{w}^+$. □

Démonstration de la formule (2). Soit $X_{N+1} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N, \mathbf{x}_{N+1}\}$ un jeu de $(N+1)$ données en position générale dans $\mathbb{R}^d$. Alors $X_N = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ est un jeu de N données en position générale dans $\mathbb{R}^d$. Une dichotomie linéairement séparable de X_{N+1} définit a fortiori une dichotomie linéairement séparable de X_N, par restriction à X_N. Il y a donc au moins autant des premières que des secondes, c'est-à-dire que $C(N+1, d) \geqslant C(N, d)$.

On veut montrer que la différence entre ces deux nombres est précisément $C(N, d-1)$.

D'après le lemme 1, si une dichotomie linéairement séparable f de X_N est donnée par $f(\mathbf{x}_j) = \operatorname{sgn}(\mathbf{w} \cdot \mathbf{x}_j)$ pour tout $j \in \{1, \ldots, N\}$, avec $\mathbf{w} \cdot \mathbf{x}_{N+1} \neq 0$, elle fournit par prolongement exactement une dichotomie linéairement séparable (f_- ou f_+) de X_{N+1}. Les autres dichotomies linéairement séparables f de X_N en fournissent deux, c'est-à-dire une de plus. Or elles sont au nombre de $C(N, d-1)$, puisqu'elles sont données par $f(\mathbf{x}_j) = \operatorname{sgn}(\mathbf{w} \cdot \mathbf{x}_j) = \operatorname{sgn}(\mathbf{w} \cdot \check{\mathbf{x}}_j)$ pour tout $j \in \{1, \ldots, N\}$, où $\check{\mathbf{x}}_j$ désigne la projection orthogonale de $\mathbf{x}_j$ sur l'espace $\mathbf{x}_{N+1}^\perp$, qui est de dimension $d-1$ dans lequel les $\check{\mathbf{x}}_j$ sont en position générale.

Ceci montre que le nombre total de dichotomies linéairement séparables de X_{N+1} est celui de X_N augmenté de $C(N, d-1)$. Autrement dit, la formule (2) est vérifiée. □

Démonstration du théorème 2. Étant données les valeurs initiales $C(1, d) = 2$ pour tout $d \geqslant 1$ et la formule de récurrence (2), on vérifie aisément (1) par récurrence sur N. Pour $N = 1$ la somme est réduite à $k = 0$ dans (1), qui donne bien $C(1, d) = 2$ pour tout $d \geqslant 1$. Soit $N \geqslant 1$ et supposons la formule (1) démontrée pour tout $d \geqslant 1$. Alors d'après (2) on a

$$C(N+1, d) = 2\sum_{k=0}^{d-1} \binom{N-1}{k} + 2\sum_{k=0}^{d-2} \binom{N-1}{k}$$

$$= 2\sum_{k=0}^{d-1} \binom{N-1}{k} + 2\sum_{k=1}^{d-1} \binom{N-1}{k-1} = 2 + 2\sum_{k=1}^{d-1} \binom{N}{k}$$

d'après la formule de récurrence satisfaite par les coefficients du binôme (3), d'où finalement

$$C(N+1,d) = 2\sum_{k=0}^{d-1} \binom{N}{k}.$$

Par suite la formule (1) est vraie pour $N+1$ au lieu de N et pour tout $d \geqslant 1$. Ceci termine la démonstration par récurrence de (1). □

Le théorème 2 est en fait équivalent à un théorème de *géométrie combinatoire* datant de 1826 [7] (repris dans le livre de Schläfli [6, 209–212] et généralisé par Winder [8]). Ce dernier a pour objet de déterminer le nombre de régions de $\mathbb{R}^d$ découpées par un ensemble de N hyperplans, qu'on appelle aussi *arrangement d'hyperplans*. Plus précisément, si les H_j sont des hyperplans de $\mathbb{R}^d$, une *région* de l'arrangement $A = \{H_1, \ldots, H_N\}$ est une *composante connexe* de l'ensemble $\mathbb{R}^d \setminus \cup_{j=1}^N H_j$.

Pour $d = 2$ par exemple, un arrangement de N droites indépendantes découpe le plan $\mathbb{R}^2$ en $2N$ régions, comme les parts d'un gâteau. On peut observer dans le tableau des valeurs de $C(N,d)$ (voir la remarque 4) que pour tout $N \geqslant 1$, on a $C(N,2) = 2N$. Ceci n'est pas un hasard.

Définition 3. *Un arrangement d'hyperplans $A = \{H_1, \ldots, H_N\}$ dans $\mathbb{R}^d$ est dit* en position générale *si ces hyperplans s'écrivent $H_j = \mathbf{x}_j^\perp$ avec $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ en position générale dans $\mathbb{R}^d$.*

Le fait qu'un arrangement d'hyperplans soit en position générale dans $\mathbb{R}^d$ implique que l'intersection de toute sous-famille de k hyperplans est de dimension minimale $d-k$. Ceci simplifie la combinatoire en ce qui concerne notamment le nombre de régions découpées par cet arrangement.

Théorème 3 (Steiner [7])**.** *Soit $A = \{H_1, \ldots, H_N\}$ un arrangement d'hyperplans en position générale dans $\mathbb{R}^d$. Alors le nombre de régions découpées par A dans $\mathbb{R}^d$ est donné par la formule* (1)*.*

On peut établir une démonstration directe de ce théorème ancien, ou bien se reposer sur le théorème 2 et le résultat suivant.

Proposition 1. *Le nombre de régions découpées par un arrangement de N hyperplans en position générale dans $\mathbb{R}^d$ est égal au nombre de dichotomies linéairement séparables d'un jeu de N données en position générale dans $\mathbb{R}^d$.*

Démonstration. Soit $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ en position générale dans $\mathbb{R}^d$ et $A = \{H_1, \ldots, H_N\}$ avec $H_j = \mathbf{x}_j^\perp$ pour tout $j \in \{1, \ldots, N\}$. Le nombre de dichotomies linéairement séparables de X est le cardinal de l'ensemble des N-uplets de la forme

$$g(\mathbf{w}) := (\mathrm{sgn}(\mathbf{w} \cdot \mathbf{x}_1), \ldots, \mathrm{sgn}(\mathbf{w} \cdot \mathbf{x}_N))$$

lorsque $\mathbf{w}$ parcourt $B := \mathbb{R}^d \setminus \cup_{j=1}^N H_j$. La fonction g étant continue sur B, elle est constante sur ses composantes connexes. Donc le nombre de N-uplets de la forme $g(\mathbf{w})$ est au plus égal au nombre de ces composantes connexes, c'est-à-dire de régions découpées par l'arrangement A. Pour montrer l'égalité entre ces deux nombres, il reste à se convaincre que g prend des valeurs distinctes sur deux régions distinctes.

On peut faire pour cela une petite récurrence sur N. Le cas $N = 1$ est facile : étant donné un hyperplan $H_1 = \mathbf{x}_j^\perp$ de $\mathbb{R}^d$, les composantes connexes de $\mathbb{R}^d \setminus H_1$ sont les demi-espaces $\{\mathbf{w}\,;\ \mathbf{w} \cdot \mathbf{x}_1 < 0\}$ et $\{\mathbf{w}\,;\ \mathbf{w} \cdot \mathbf{x}_1 > 0\}$, sur lesquels $\mathbf{w} \mapsto \operatorname{sgn}(\mathbf{w} \cdot \mathbf{x}_1)$ prend respectivement les valeurs $-$ et $+$, par définition. Elles sont donc bien distinctes. Soit $N \geqslant 1$ et supposons que la fonction $\mathbf{w} \mapsto (\operatorname{sgn}(\mathbf{w} \cdot \mathbf{x}_1), \ldots, \operatorname{sgn}(\mathbf{w} \cdot \mathbf{x}_N))$ prenne des valeurs distinctes sur toute paire de régions distinctes de $B_N := \mathbb{R}^d \setminus \cup_{j=1}^N H_j$. Soit $H_{N+1} = \mathbf{x}_{N+1}^\perp$ un nouvel hyperplan. Alors les régions de $\mathbb{R}^d \setminus \cup_{j=1}^{N+1} H_j$ sont les régions R de B_N n'intersectant pas H_{N+1}, et les ensembles $S_\pm$ définis par

$$S_- = \{\mathbf{w} \in S\,;\ \mathbf{w} \cdot \mathbf{x}_{N+1} < 0\}\,, \quad S_+ = \{\mathbf{w} \in S\,;\ \mathbf{w} \cdot \mathbf{x}_{N+1} > 0\}$$

pour S une région de B_N telle que $S \cap H_{N+1} \neq \emptyset$. D'après l'hypothèse de récurrence, pour toute paire de régions n'intersectant pas H_{N+1} il existe $j \in \{1, \ldots, N\}$ tel que $\mathbf{w} \cdot \mathbf{x}_j$ prend un signe différent entre ces deux régions. C'est aussi le cas pour une paire constituée d'une région R telle que $S \cap H_{N+1} = \emptyset$ et de S_- ou S_+, puisque celles-ci sont incluses dans une région de B_N distincte de R. Et bien sûr, pour la paire $\{S_-, S_+\}$, c'est $\mathbf{w} \cdot \mathbf{x}_{N+1}$ qui change de signe entre les deux. Donc la propriété voulue est satisfaite avec $N + 1$ hyperplans. □

Remarque 5. *Cette proposition permet de définir explicitement une bijection entre les dichotomies linéairement séparables f de $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ et les régions R de $B = \mathbb{R}^d \setminus \cup_{j=1}^N \mathbf{x}_j^\perp$, à savoir :*

$$f \mapsto R = \left\{\mathbf{w} \in \mathbb{R}^d \setminus \cup_{j=1}^N \mathbf{x}_j^\perp\,;\ \operatorname{sgn}(\mathbf{w} \cdot \mathbf{x}_j) = f(\mathbf{x}_j)\,,\ j \in \{1, \ldots, N\}\right\}.$$

Ceci montre en particulier que les régions de B sont positivement homogènes *et* convexes *: si $\mathbf{w} \in R$ et $\lambda > 0$ alors $\lambda\mathbf{w} \in R$; si $\mathbf{w}$, $\mathbf{w}' \in R$ alors $[\mathbf{w}, \mathbf{w}'] \subset R$.*

Démonstration directe du théorème 3. En faisant abstraction du résultat de la proposition 1, introduisons la notation $a(N, d)$ pour le nombre de régions découpées par un arrangement de N hyperplans en position générale dans $\mathbb{R}^d$. Comme pour $C(N, d)$, le calcul de $a(N, d)$ repose sur celui de $a(1, d)$ et sur la formule de récurrence

$$a(N + 1, d) = a(N, d) + a(N, d - 1). \tag{4}$$

Avant de prouver cette dernière, on vérifie de suite que $a(1, d) = 2$ pour tout $d \geqslant 1$, un hyperplan séparant $\mathbb{R}^d$ en exactement deux demi-espaces, comme cela a été mentionné dans la preuve de la proposition 1.

Le cœur de la démonstration du théorème 3 réside donc dans la preuve de (4). On a aussi vu dans la preuve de proposition 1 que le nombre de régions découpées par $N + 1$

hyperplans est supérieur à celui découpé par N hyperplans. Le but est de calculer la différence entre ces deux nombres, $a(N+1,d)$ et $a(N,d)$. Elle est nulle dans le cas $d=1$ mais c'est particulier car il y a un seul hyperplan de $\mathbb{R}$, le sous-espace trivial $\{0\}$. Pour $d \geqslant 2$ on a bien $a(N+1,d) > a(N,d)$.

D'après l'argument du raisonnement par récurrence dans la preuve de la proposition 1, la différence $a(N+1,d) - a(N,d)$ est précisément égale au nombre de régions S de $B_N = \mathbb{R}^d \setminus \cup_{j=1}^{N} H_j$ telles que $S \cap H_{N+1} \neq \emptyset$, car chacune d'entre elles ajoute exactement une région dans $B_{N+1} = \mathbb{R}^d \setminus \cup_{j=1}^{N+1} H_j$ par rapport à celles qu'on a dans B_N (la région S de B_N étant séparée en S_- et S_+ dans B_{N+1}), tandis que les autres régions de B_N donnent une et une seule région dans B_{N+1}.

Il reste donc à dénombrer les régions S de B_N telles que $S \cap H_{N+1} \neq \emptyset$. Or elles sont en bijection avec les régions de $\check{B}_N := B_N \cap H_{N+1} = H_{N+1} \setminus \cup_{j=1}^{N} (H_j \cap H_{N+1})$: en effet, S est une région de B_N telle que $S \cap H_{N+1} \neq \emptyset$ si et seulement $\check{S} := S \cap H_{N+1}$ (convexe comme intersection de deux convexes, et donc connexe) est une région de $\check{B}_N$. Comme H_{N+1} est de dimension $d-1$ dans lequel les $(H_j \cap H_{N+1})$ en sont des hyperplans en position générale, le nombre de régions de $\check{B}_N$ est égal à $a(N,d-1)$.

Ceci prouve la formule de récurrence (4). On en déduit alors que $a(N,d)$ est égal à $C(N,d)$ donné par la formule (1), puisque $a(1,d) = C(1,d) = 2$ pour tout $d \geqslant 1$ et $a(N,d)$ vérifie la même formule de récurrence que $C(N,d)$. □

Les problèmes d'arrangements d'hyperplans ont été généralisés de multiples façons au XXème siècle (pour plus de détails on pourra consulter le séminaire Bourbaki de Pierre Cartier [1] ou le livre légèrement plus récent [4]) et sont toujours une thématique active de la géométrie combinatoire.

2.2 Dichotomies nonlinéairement séparables

Nous allons maintenant quitter le royaume de la géométrie linéaire. Il n'y a en effet aucune raison pour qu'une dichotomie sur un jeu de données « réel » soit linéairement séparable. C'est tout à fait improbable sur l'exemple donné au début, même si l'on se concentre sur des données physiques comme la taille et le poids des individus par exemple : un indicateur largement utilisé, même s'il est contestable et contesté, est l'indice de masse corporelle, qui compare en effet le poids à la taille élevée au carré ; ceci suggère de classifier un jeu de données correspondant à la taille et au poids d'une population non pas en les séparant linéairement mais plutôt en les séparant par une courbe parabolique. De manière générale, on peut chercher à séparer des données par une hypersurface.

Un premier point de vue consisterait à le faire en restant dans l'espace $\mathbb{R}^d$ où se trouvent les données X. Ceci reviendrait à se donner une équation implicite $\varphi(\mathbf{x}) = 0$ pour chaque hypersurface considérée, où φ appartiendrait à un ensemble de fonctions régulières de $\mathbb{R}^d$ dans $\mathbb{R}$. On chercherait alors à généraliser le dénombrement effectué au § 2.1 en dénombrant les applications $f : X \to \{-,+\}$ telles qu'il existe une telle fonction φ pour laquelle

$$f^{-1}(-) \subset \{\mathbf{x} \in \mathbb{R}^d ;\ \varphi(\mathbf{x}) < 0\} \quad \text{et} \quad f^{-1}(+) \subset \{\mathbf{x} \in \mathbb{R}^d ;\ \varphi(\mathbf{x}) > 0\}.$$

Un autre point de vue, plus abordable, consiste à changer d'espace de données, afin de se ramener au dénombrement effectué au § 2.1. Sur l'exemple en dimension 2 où chaque donnée consiste en un couple (h, m) pour la taille (plus précisément la hauteur) h et le poids (en fait la masse) m d'un individu, on peut considérer l'application

$$\begin{array}{rccc} \Phi : & \mathbb{R}^2 & \to & \mathbb{R}^2 \\ & (h, m) & \mapsto & (h^2, m) \end{array}$$

et étudier les jeux de données $\{\mathbf{x}_1 = (h_1, m_1), \cdots, \mathbf{x}_N = (h_N, m_N)\}$ au travers de leurs images $\{\mathbf{y}_1 := \Phi(\mathbf{x}_1) = (h_1^2, m_1), \cdots, \mathbf{y}_N := \Phi(\mathbf{x}_N) = (h_N^2, m_N)\}$. En généralisant ce point de vue on peut donner les définitions suivantes.

Définition 4. *Soit $\Phi : \mathbb{R}^n \to \mathbb{R}^d$. Une dichotomie $f : X \to \{-, +\}$ d'un jeu de données $X \subset \mathbb{R}^n$ est dite* Φ-séparable *s'il existe $\mathbf{w} \in \mathbb{R}^d$ tel que*

$$\Phi(f^{-1}(-)) \subset \{\mathbf{y} \in \mathbb{R}^d ;\ \mathbf{w} \cdot \mathbf{y} < 0\} \quad et \quad \Phi(f^{-1}(+)) \subset \{\mathbf{y} \in \mathbb{R}^d ;\ \mathbf{w} \cdot \mathbf{y} > 0\}.$$

On notera que les dimensions de l'espace de départ $\mathbb{R}^n$ et de l'espace d'arrivée $\mathbb{R}^d$ peuvent être différentes. Ceci permet par exemple de traiter la séparation par des hyperplans affines, en considérant

$$\begin{array}{rccc} \Phi : & \mathbb{R}^n & \to & \mathbb{R}^{n+1} \\ & \mathbf{x} & \mapsto & (1, \mathbf{x}) \end{array}$$

comme indiqué dans la remarque 1. On donnera d'autres exemples plus loin.

Le dénombrement des dichotomies Φ-séparables sur un jeu de données dans $\mathbb{R}^n$ repose donc sur le dénombrement des dichotomies linéairement séparables sur son image par Φ dans $\mathbb{R}^d$, et relève du théorème 2 pourvu que ce jeu image soit en position générale. Ceci conduit à la définition suivante, avant d'énoncer un théorème.

Définition 5. *Soit $\Phi : \mathbb{R}^n \to \mathbb{R}^d$. Un jeu de données $X \subset \mathbb{R}^n$ est dit* en position Φ-générale *dans $\mathbb{R}^n$ si son image $Y = \Phi(X) \subset \mathbb{R}^d$ est en position générale dans $\mathbb{R}^d$.*

Le théorème 2 implique donc le résultat suivant.

Théorème 4 (Cover [2])**.** *Soient $\Phi : \mathbb{R}^n \to \mathbb{R}^d$ et $X \subset \mathbb{R}^n$ un jeu de données en position Φ-générale dans $\mathbb{R}^n$. Le nombre de dichotomies Φ-séparables de X est égal à $C(N, d)$, donné par la formule* (1)*.*

Si l'on considère le cas de la dimension $n = 2$ par exemple, on peut s'intéresser à des dichotomies séparables par une droite affine, par un cercle, ou encore par une conique. Ceci correspond à des dichotomies Φ-séparables respectivement pour $\Phi(x, y) = (1, x, y)$, $\Phi(x, y) = (1, x, y, \sqrt{x^2 + y^2})$, $\Phi(x, y) = (1, x, y, x^2, xy, y^2)$. La figure reproduite ci-dessous de l'article original de Cover [2] donne trois exemples de dichotomies [4] sur un même jeu de cinq données, séparables de l'une de ces trois manières (dans le cas (c) il faut voir la courbe comme une hyperbole, bien qu'elle ressemble à deux paraboles).

4. Les triangles correspondant à un signe et les ronds à un autre.

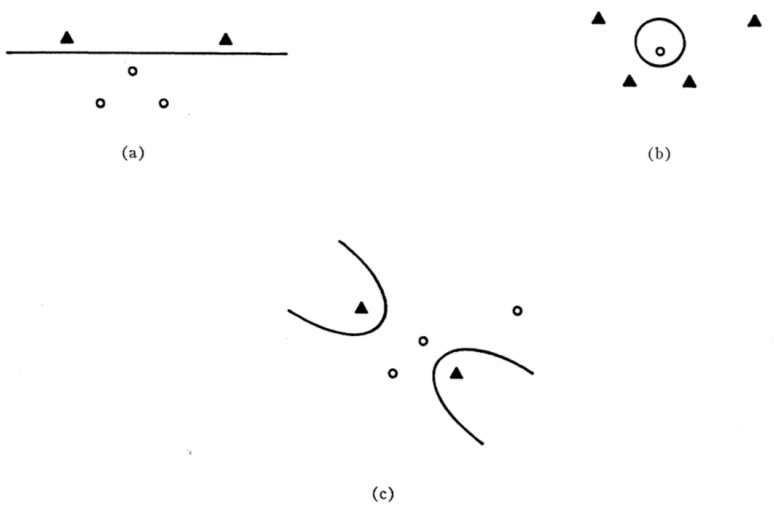

Fig. 1. Examples of ϕ-separable dichotomies of five points in two dimensions. (a) Linearly separable dichotomy. (b) Spherically separable dichotomy. (c) Quadrically separable dichotomy.

D'après la combinatoire fournie par le théorème 4, le nombre total de dichotomies de cinq données séparables dans le plan :

— par une droite est $C(5,3) = 22$,

— par un cercle est $C(5,4) = 30$,

— par une conique est $C(5,4) = 32$,

c'est-à-dire que toutes les dichotomies de cinq données sont séparables par une conique.

De manière générale, on peut s'intéresser aux données séparables par une *variété algébrique*, ce qui correspond à $\Phi : \mathbb{R}^n \to \mathbb{R}^d$ avec $d = D(n,r)$ la dimension de l'espace des polynômes à n variables et de degré inférieur ou égal à r, le degré de la variété, et les composantes de Φ sont les monômes à n variables et de degré inférieur ou égal à r.

Comme on l'a vu sur les exemples précédents, $D(2,1) = 2$, $D(2,2) = 6$. On peut montrer facilement par récurrence sur n, en utilisant la formule de Pascal (3), que quels que soient $n \in \mathbb{N}^*$ et r, $D(n,r) = \binom{n+r}{r}$.

On peut alors chercher le degré r minimal pour que toutes les dichotomies sur un jeu de N données dans $\mathbb{R}^n$ soient séparables par une variété de degré r. Il s'agit d'assurer que $D(n,r)$ soit au moins égal à N.

La lecture du tableau ci-dessous permet de trouver le degré minimal pour n compris entre 1 et 10 et d'assez grandes valeurs de N. Par exemple, on lit que le degré minimal est $r = 30$ pour $n = 2$, $N = 466$ $(= D(2,29) + 1)$. Ou encore $r = 30$ pour $n = 10$ et $N = 635\ 745\ 397$ $(= D(10,29) + 1)$.

r \ n	1	2	3	4	5	6	7	8	9	10
1	2	3	4	5	6	7	8	9	10	11
2	3	6	10	15	21	28	36	45	55	66
3	4	10	20	35	56	84	120	165	220	286
4	5	15	35	70	126	210	330	495	715	1 001
5	6	21	56	126	252	462	792	1 287	2 002	3 003
6	7	28	84	210	462	924	1 716	3 003	5 005	8 008
7	8	36	120	330	792	1 716	3 432	6 435	11 440	19 448
8	9	45	165	495	1 287	3 003	6 435	12 870	24 310	43 758
9	10	55	220	715	2 002	5 005	11 440	24 310	48 620	92 378
10	11	66	286	1 001	3 003	8 008	19 448	43 758	92 378	184 756
11	12	78	364	1 365	4 368	12 376	31 824	75 582	167 960	352 716
12	13	91	455	1 820	6 188	18 564	50 388	125 970	293 930	646 646
13	14	105	560	2 380	8 568	27 132	77 520	203 490	497 420	1 144 066
14	15	120	680	3 060	11 628	38 760	116 280	319 770	817 190	1 961 256
15	16	136	816	3 876	15 504	54 264	170 544	490 314	1 307 504	3 268 760
16	17	153	969	4 845	20 349	74 613	245 157	735 471	2 042 975	5 311 735
17	18	171	1 140	5 985	26 334	100 947	346 104	1 081 575	3 124 550	8 436 285
18	19	190	1 330	7 315	33 649	134 596	480 700	1 562 275	4 686 825	13 123 110
19	20	210	1 540	8 855	42 504	177 100	657 800	2 220 075	6 906 900	20 030 010
20	21	231	1 771	10 626	53 130	230 230	888 030	3 108 105	10 015 005	30 045 015
21	22	253	2 024	12 650	65 780	296 010	1 184 040	4 292 145	14 307 150	44 352 165
22	23	276	2 300	14 950	80 730	376 740	1 560 780	5 852 925	20 160 075	64 512 240
23	24	300	2 600	17 550	98 280	475 020	2 035 800	7 888 725	28 048 800	92 561 040
24	25	325	2 925	20 475	118 755	593 775	2 629 575	10 518 300	38 567 100	131 128 140
25	26	351	3 276	23 751	142 506	736 281	3 365 856	13 884 156	52 451 256	183 579 396
26	27	378	3 654	27 405	169 911	906 192	4 272 048	18 156 204	70 607 460	254 186 856
27	28	406	4 060	31 465	201 376	1 107 568	5 379 616	23 535 820	94 143 280	348 330 136
28	29	435	4 495	35 960	237 336	1 344 904	6 724 520	30 260 340	124 403 620	472 733 756
29	30	465	4 960	40 920	278 256	1 623 160	8 347 680	38 608 020	163 011 640	635 745 396
30	31	496	5 456	46 376	324 632	1 947 792	10 295 472	48 903 492	211 915 132	847 660 528

Du point de vue algorithmique, on peut aussi chercher à séparer les dichotomies par des hypersurfaces affines par morceaux, au lieu de variétés algébriques. On ne détaillera pas ce point de vue ici.

2.3 Séparation de dichotomies aléatoires

On peut introduire de l'aléa à la fois dans le choix des jeux de données et dans le choix des dichotomies.

Pour simplifier, considérons des dichotomies aléatoires sur un jeu de données fixé et concentrons nous sur celles qui sont linéairement séparables, puisque les séparations plus générales vues au §2.2 s'y ramènent, au prix d'augmenter la dimension de l'espace.

Étant donné un jeu de N données X fixé, en position générale dans $\mathbb{R}^d$, on sait qu'il y a 2^N dichotomies de X en tout, et qu'il y en a $C(N,d)$ qui sont linéairement séparables. Autrement dit, la probabilité pour qu'une dichotomie de X prise au hasard soit linéairement séparable est

$$P(N,d) = \frac{C(N,d)}{2^N} = \frac{1}{2^{N-1}} \sum_{k=0}^{d-1} \binom{N-1}{k}, \tag{5}$$

d'après (1), pour autant que toutes les dichotomies soient équiprobables. Comme on l'a déjà observé, $C(N,d) = 2^N$ et donc $P(N,d) = 1$ pour $d \geqslant N$. Cependant on peut remarquer une

autre transition pour $N = 2d$. En effet,

$$\sum_{k=0}^{d-1} \binom{2d-1}{k} = \frac{1}{2} \sum_{k=0}^{2d-1} \binom{2d-1}{k}$$

par la symétrie des coefficients du binôme

$$\binom{2d-1}{k} = \binom{2d-1}{2d-1-k},$$

et

$$\sum_{k=0}^{2d-1} \binom{2d-1}{k} = 2^{2d-1}$$

donc

$$P(2d, d) = \frac{1}{2}.$$

Le théorème 6 ci-après met en évidence le comportement asymptotique de $P(N, d)$ lorsque N tend vers $+\infty$ avec $d \sim \frac{N}{2} + \alpha\frac{\sqrt{N}}{2}$ pour $\alpha \in \mathbb{R}$ ou $d \sim \frac{N}{2}(1 \pm \varepsilon)$ pour $\varepsilon > 0$. Sa démonstration repose sur le théorème de Moivre-Laplace, que l'on rappelle ci-dessous dans la version utile pour ce qui suit. Commençons par un petit rappel sur la loi binomiale.

Définition 6. *On dit qu'une variable aléatoire S_n à valeurs dans $\mathbb{N}$ suit la* loi binomiale *d'ordre n et de paramètre $1/2$ si pour tout $k \in \mathbb{N}$, la probabilité que S_n vaille k est donnée par*

$$\mathbb{P}(S_n = k) = \frac{1}{2^n}\binom{n}{k},$$

toujours avec la convention que $\binom{n}{k} = 0$ pour $k > n$.

Une telle variable aléatoire est obtenue par exemple en sommant le nombre de fois où une pièce non truquée tombe sur pile au cours de n tirages.

Théorème 5 (Moivre-Laplace)**.** *Soit S_n une variable aléatoire suivant une loi binomiale d'ordre n et de paramètre $1/2$, alors la variable $Z_n := \frac{S_n - n/2}{\sqrt{n}/2}$ converge en loi vers une loi normale centrée et réduite, ce qui signifie en particulier que pour tout $\alpha \in \mathbb{R}$,*

$$\mathbb{P}(Z_n < \alpha) \to \Phi(\alpha) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha} \mathrm{e}^{-x^2/2}\,\mathrm{d}x,$$

$$\mathbb{P}(Z_n > \alpha) \to 1 - \Phi(\alpha) = \frac{1}{\sqrt{2\pi}} \int_{\alpha}^{+\infty} \mathrm{e}^{-x^2/2}\,\mathrm{d}x$$

quand n tend vers $+\infty$. Le même résultat vaut pour $\check{Z}_n := \frac{S_{n-1} - n/2}{\sqrt{n}/2}$.

La première partie de l'énoncé correspond au théorème de Moivre-Laplace tel qu'il figure au programme du lycée. La seconde partie s'en déduit grâce à l'argument suivant. Étant donné $\delta > 0$, il existe $n_0 \geqslant 1$ tel que pour $n \geqslant n_0$,

$$1 < \frac{\sqrt{n}}{\sqrt{n-1}} \leqslant 1 + \delta\,, \quad \frac{1}{\sqrt{n-1}} \leqslant \delta.$$

Ainsi, pour tout $n \geqslant n_0$, on a en supposant pour fixer les idées que α est positif (le cas où α est négatif se traitant de manière analogue) :

$$\alpha < \frac{1}{\sqrt{n-1}} + \frac{\sqrt{n}}{\sqrt{n-1}}\alpha \leq \delta + (1+\delta)\,\alpha,$$

et comme

$$\check{Z}_n < \alpha \quad \Leftrightarrow \quad Z_{n-1} < \frac{1}{\sqrt{n-1}} + \frac{\sqrt{n}}{\sqrt{n-1}}\alpha\,,$$

on en déduit

$$\mathbb{P}(Z_{n-1} < \alpha) \leqslant \mathbb{P}(\check{Z}_n < \alpha) \leqslant \mathbb{P}(Z_{n-1} < \delta + (1+\delta)\,\alpha),$$

d'où

$$\Phi(\alpha) \leqslant \liminf_{n\to\infty} \mathbb{P}(\check{Z}_n < \alpha) \leqslant \limsup_{n\to\infty} \mathbb{P}(\check{Z}_n < \alpha) \leqslant \Phi(\delta + (1+\delta)\,\alpha).$$

En faisant tendre δ vers zéro on obtient finalement

$$\liminf_{n\to\infty} \mathbb{P}(\check{Z}_n < \alpha) = \limsup_{n\to\infty} \mathbb{P}(\check{Z}_n < \alpha) = \Phi(\alpha)$$

puisque Φ est continue, ce qui montre que $\mathbb{P}(\check{Z}_n < \alpha)$ converge vers $\Phi(\alpha)$. La limite pour $\mathbb{P}(\check{Z}_n > \alpha)$ s'obtient de la même manière.

Théorème 6. *Soient $\alpha \in \mathbb{R}$ et $\varepsilon > 0$. Lorsque N tend vers $+\infty$,*

- *si $d \sim \frac{N}{2} + \alpha\frac{\sqrt{N}}{2}$ alors $P(N,d) \to \Phi(\alpha)$;*
- *si $d \sim \frac{N}{2}(1+\varepsilon)$ alors $P(N,d) \to 1$;*
- *si $d \sim \frac{N}{2}(1-\varepsilon)$ alors $P(N,d) \to 0$.*

Démonstration. Par définition et avec les notations introduites précédemment on a

$$P(N,d) = \frac{1}{2^{N-1}} \sum_{k=0}^{d-1} \binom{N-1}{k} = \sum_{k=0}^{d-1} \mathbb{P}(S_{N-1} = k) = \mathbb{P}(S_{N-1} < d) = \mathbb{P}\left(\check{Z}_N < \frac{d - N/2}{\sqrt{N}/2}\right).$$

On voit ainsi que si $d = \frac{N}{2} + \alpha\frac{\sqrt{N}}{2}$ exactement, $P(N,d) = \mathbb{P}(\check{Z}_N < \alpha)$, ce qui tend vers $\Phi(\alpha)$ lorsque N tend vers $+\infty$ d'après le théorème 5. Plus généralement, si $d = d(N) \sim \frac{N}{2} + \alpha\frac{\sqrt{N}}{2}$ alors pour tout $\delta > 0$ il existe N_0 tel que pour tout $N \geqslant N_0$,

$$\frac{d(N) - N/2}{\sqrt{N}/2} \in [\alpha - \delta, \alpha + \delta],$$

d'où

$$\mathbb{P}\left(\check{Z}_N < \alpha - \delta\right) \leqslant P(N, d(N)) = \mathbb{P}\left(\check{Z}_N < \frac{d(N) - N/2}{\sqrt{N}/2}\right) \leqslant \mathbb{P}\left(\check{Z}_N < \alpha + \delta\right).$$

En passant à la limite inf et à la limite sup comme on l'a fait plus haut on en déduit

$$\Phi(\alpha - \delta) \leqslant \liminf_{N\to\infty} P(N, d(N)) \leqslant \limsup_{N\to\infty} P(N, d(N)) \leqslant \Phi(\alpha + \delta).$$

En faisant tendre δ vers zéro on en conclut que $P(N, d(N))$ converge vers $\Phi(\alpha)$.

Pour traiter le cas $d \sim \frac{N}{2}(1+\varepsilon)$, on raisonne de manière similaire. Étant donné $\beta > 0$ il existe N_0 tel que pour tout $N \geqslant N_0$,

$$\frac{d(N) - N/2}{\sqrt{N}/2} \geqslant \beta,$$

donc

$$P(N, d(N)) = \mathbb{P}\left(\check{Z}_N < \frac{d(N) - N/2}{\sqrt{N}/2}\right) \geqslant \mathbb{P}\left(\check{Z}_N < \beta\right).$$

On en déduit en passant à la limite inf que

$$\liminf_{N\to\infty} P(N, d(N)) \geqslant \Phi(\beta).$$

En faisant tendre β vers $+\infty$, comme $\Phi(\beta)$ tend vers 1, on déduit que

$$\limsup_{N\to\infty} P(N, d(N)) = \liminf_{N\to\infty} P(N, d(N)) = 1.$$

Le cas $d = d_-(N) \sim \frac{N}{2}(1-\varepsilon)$ est tout à fait similaire, en utilisant des inégalités opposées, de sorte

$$\limsup_{N\to\infty} P(N, d_-(N)) \leqslant \Phi(\beta)$$

pour tout $\beta < 0$, d'où la convergence de $P(N, d_-(N))$ vers zéro en faisant tendre β vers $-\infty$. □

Le fait que $N = 2d$ soit une valeur critique de la taille d'un jeu de données par rapport à sa séparabilité a conduit à définir $2d$ comme la « capacité » de l'espace $\mathbb{R}^d$.

3 Un algorithme d'apprentissage

Nous allons maintenant présenter l'algorithme à la base des méthodes d'apprentissage automatique [5] pour classifier les données.

5. « machine learning »

Cet algorithme concerne les dichotomies de jeux de données dont on sait qu'elles sont linéairement séparables. L'objectif de cet algorithme est de calculer, pour une telle dichotomie f, un vecteur $\mathbf{w}$ qui définisse un hyperplan de séparation comme dans la définition 1.

L'idée de départ est très simple. Prenons un vecteur $\mathbf{w}$ « au hasard »(c'est le seul endroit où l'on parle d'aléa dans cette partie). Pour un vecteur $\mathbf{x}$ appartenant au jeu de données, on peut avoir directement $f(\mathbf{x})\,\mathrm{sgn}(\mathbf{w}\cdot\mathbf{x}) > 0$, auquel cas $\mathbf{w}$ est un candidat à garder. Si au contraire $f(\mathbf{x})\,\mathrm{sgn}(\mathbf{w}\cdot\mathbf{x}) \leqslant 0$, on peut « corriger » $\mathbf{w}$ dans la direction $\mathbf{x}$ en considérant le vecteur modifié

$$\widehat{\mathbf{w}} = \mathbf{w} + \eta\, f(\mathbf{x})\,\mathbf{x},$$

où η est un paramètre strictement positif. Bien sûr, avec une seule donnée $\mathbf{x}$ il suffit de choisir $\eta > -f(\mathbf{x})\,\mathrm{sgn}(\mathbf{w}\cdot\mathbf{x})/\|\mathbf{x}\|^2$ pour assurer que $f(\mathbf{x})\,\mathrm{sgn}(\widehat{\mathbf{w}}\cdot\mathbf{x})$ soit strictement positif. Si l'on imagine effectuer des corrections successives en parcourant le jeu de données, il n'est pas évident que l'on obtienne au final un vecteur qui classifie correctement toutes les données. C'est pourtant ce qui se passe avec l'algorithme suivant, appelé en anglais « perceptron learning algorithm » (PLA), valable avec un paramètre $\eta > 0$ quelconque, à condition de parcourir éventuellement plusieurs fois le jeu de données.

Définition 7 (PLA)**.** *Soit f une dichotomie linéairement séparable sur un jeu de données $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ de $\mathbb{R}^d$. Pour tout $k \in \mathbb{N}$ on pose $\mathbf{x}_{kN+j} = \mathbf{x}_j$ pour tout $j \in \{1, \ldots, N\}$, ce qui définit une suite $(\mathbf{x}_j)_{j\in\mathbb{N}^*}$, périodique de période N.*

Soient $\eta > 0$ et $\mathbf{w}_0 \in \mathbb{R}^d$. On définit par récurrence la suite $(\mathbf{w}_j)_{j\in\mathbb{N}}$ telle que pour tout $j \in \mathbb{N}^*$:

- *si $f(\mathbf{x}_j)\,\mathrm{sgn}(\mathbf{w}_{j-1}\cdot\mathbf{x}_j) > 0$ alors $\mathbf{w}_j = \mathbf{w}_{j-1}$,*
- *si $f(\mathbf{x}_j)\,\mathrm{sgn}(\mathbf{w}_{j-1}\cdot\mathbf{x}_j) \leqslant 0$ alors $\mathbf{w}_j = \mathbf{w}_{j-1} + \eta\, f(\mathbf{x}_j)\,\mathbf{x}_j$,*

Le résultat suivant (« perceptron convergence theorem ») montre que l'algorithme PLA converge en un nombre fini d'itérations, c'est-à-dire qu'il existe un entier $J \in \mathbb{N}^*$ tel que pour tout $j \geqslant J$, $f(\mathbf{x}_j)\,\mathrm{sgn}(\mathbf{w}_{j-1}\cdot\mathbf{x}_j) > 0$, et donc $\mathbf{w}_j = \mathbf{w}_J$. L'hyperplan $\mathbf{w}_J^{\perp}$ sépare alors la dichotomie considérée.

Théorème 7. *Soient f une dichotomie linéairement séparable sur un jeu de données $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ de $\mathbb{R}^d$, $\eta > 0$, $\mathbf{w}_0 \in \mathbb{R}^d$, et $(\mathbf{w}_j)_{j\in\mathbb{N}}$ donnée par l'algorithme PLA (définition 7). Alors il existe $J \in \mathbb{N}^*$ tel que pour tout $j \geqslant J$, $f(\mathbf{x}_j)\,\mathrm{sgn}(\mathbf{w}_{j-1}\cdot\mathbf{x}_j) > 0$.*

Démonstration. Le but est de montrer que l'ensemble

$$A := \{j \in \mathbb{N}^*\,;\ f(\mathbf{x}_j)\,\mathrm{sgn}(\mathbf{w}_{j-1}\cdot\mathbf{x}_j)) \leqslant 0\}$$

est fini. Il suffira alors en effet de poser $J = 1 + \max A$. Notons n_j pour tout $j \in \mathbb{N}^*$ le nombre d'éléments de A au plus égaux à j, et $n_0 = 0$.

Par hypothèse, on sait qu'il existe $\mathbf{w} \in \mathbb{R}^d$ (non nul) tel que $f(\mathbf{x}_j)\,\mathrm{sgn}(\mathbf{w}\cdot\mathbf{x}_j) > 0$ pour tout $j \in \mathbb{N}^*$. Grâce à l'inégalité de Cauchy-Schwarz

$$\mathbf{w}\cdot\mathbf{w}_j \leq \|\mathbf{w}\|\,\|\mathbf{w}_j\|,$$

on va montrer que la suite d'entiers $(n_j)_{j\in\mathbb{N}}$ est majorée, ce qui entraîne nécessairement que A est fini.

On pose, pour tout $j \in \mathbb{N}^*$,

$$\delta_j := f(\mathbf{x}_j)\,\frac{\mathbf{w}\cdot\mathbf{x}_j}{\|\mathbf{w}\|}.$$

Ce nombre, strictement positif par définition de $\mathbf{w}$, est appelé *marge* de $\mathbf{x}_j$ par rapport à l'hyperplan $\mathbf{w}^\perp$. Étant donné que la suite $(\mathbf{x}_j)_{j\in\mathbb{N}^*}$ est périodique, la suite des δ_j est minorée par un nombre strictement positif. Notons le δ. Notons par ailleurs $R = \max_{j\in\mathbb{N}^*}\|\mathbf{x}_j\| = \max_{1\leqslant j\leqslant N}\|\mathbf{x}_j\|$ (qui est fini).

Montrons par récurrence sur j que pour tout $j \in \mathbb{N}$,

$$\mathbf{w}\cdot\mathbf{w}_j \geqslant \mathbf{w}\cdot\mathbf{w}_0 + \eta\,\delta\,\|\mathbf{w}\| n_j\,, \quad \|\mathbf{w}_j\|^2 \leqslant \|\mathbf{w}_0\|^2 + \eta^2\,R^2\,n_j.$$

Ces inégalités sont des égalités pour $j = 0$. Soit $j \in \mathbb{N}^*$. Supposons les inégalités ci-dessus vérifiées jusqu'à l'indice $j-1$. Par définition de $\mathbf{w}_j$ et n_j on a :

- si $f(\mathbf{x}_j)\,\mathrm{sgn}(\mathbf{w}_{j-1}\cdot\mathbf{x}_j) > 0$ alors $\mathbf{w}\cdot\mathbf{w}_j = \mathbf{w}\cdot\mathbf{w}_{j-1}$, $\|\mathbf{w}_j\| = \|\mathbf{w}_j\|$, et $n_j = n_{j-1}$,
- si $f(\mathbf{x}_j)\,\mathrm{sgn}(\mathbf{w}_{j-1}\cdot\mathbf{x}_j) \leqslant 0$ alors $n_j = n_{j-1} + 1$,

$$\mathbf{w}\cdot\mathbf{w}_j = \mathbf{w}\cdot\mathbf{w}_{j-1} + \eta\,f(\mathbf{x}_j)\,\mathbf{w}\cdot\mathbf{x}_j \geqslant \mathbf{w}\cdot\mathbf{w}_{j-1} + \eta\,\delta\,\|\mathbf{w}\|,$$

$$\|\mathbf{w}_j\|^2 = \|\mathbf{w}_{j-1}\|^2 + \eta^2\,\|\mathbf{x}_j\|^2 + 2\eta\,f(\mathbf{x}_j)\,(\mathbf{w}_{j-1}\cdot\mathbf{x}_j) \leqslant \|\mathbf{w}_{j-1}\|^2 + \eta^2\,R^2.$$

L'hypothèse de récurrence implique donc les inégalités voulues.

Pour conclure, on remarque que la seconde implique

$$\|\mathbf{w}_j\| \leqslant \|\mathbf{w}_0\| + \eta\,R\,\sqrt{n_j}.$$

Par suite, d'après l'inégalité de Cauchy-Schwarz on a

$$\mathbf{w}\cdot\mathbf{w}_0 + \eta\,\delta\,\|\mathbf{w}\| n_j \leqslant \mathbf{w}\cdot\mathbf{w}_j \leqslant \|\mathbf{w}\|\,\|\mathbf{w}_j\| \leqslant \|\mathbf{w}\|\,(\|\mathbf{w}_0\| + \eta\,R\,\sqrt{n_j}),$$

ce qui entraîne

$$\delta\,n_j \leqslant 2\|\mathbf{w}_0\|/\eta + R\,\sqrt{n_j},$$

d'où finalement

$$\sqrt{n_j} \leqslant \frac{R + \sqrt{R^2 + 8\delta\|\mathbf{w}_0\|/\eta}}{2\delta}\,.$$

On déduit que le cardinal de l'ensemble A est majoré par

$$n = \left\lfloor \frac{(R + \sqrt{R^2 + 8\delta\|\mathbf{w}_0\|/\eta})^2}{4\delta^2} \right\rfloor,$$

ce qui prouve le théorème avec $J = n + 1$. □

Le nombre n obtenu ci-dessus donne le nombre d'itérations en fonction notamment de la marge δ. Il tend vers l'infini comme $1/\delta^2$ quand cette marge tend vers zéro, ce qui peut être très pénalisant. Ce n'est cependant pas la seule limitation du perceptron. Les algorithmes plus récents ont des performances bien meilleures.

Références

[1] Pierre Cartier. Les arrangements d'hyperplans : un chapitre de géométrie combinatoire. In *Bourbaki Seminar, Vol. 1980/81*, volume 901 of *Lecture Notes in Math.*, pages 1–22. Springer, Berlin-New York, 1981.

[2] T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3) :326–334, 1965.

[3] Mikel Olazaran. A sociological study of the official history of the perceptrons controversy. *Social Studies of Science*, 26(3) :611–659, 1996.

[4] Peter Orlik and Hiroaki Terao. *Arrangements of hyperplanes*, volume 300 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1992.

[5] Frank Rosenblatt. The perceptron : A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6) :386–408, 1958.

[6] Ludwig Schläfli. *Gesammelte mathematische Abhandlungen. Band I.* Verlag Birkhäuser, Basel, 1950.

[7] J. Steiner. Einige Gesetze über die Theilung der Ebene und des Raumes. *J. Reine Angew. Math.*, 1 :349–364, 1826.

[8] R. O. Winder. Partitions of N-space by hyperplanes. *SIAM J. Appl. Math.*, 14 :811–818, 1966.