Paris Cité **Apprentissage statistique**

2025/2026 **Stephan Clémençon - Ekhine Irurozki**

Travaux dirigés n° 5 : Perceptron (Novikoff Theorem)

Stéphan Clémençon <stephan.clemencon@telecom-paristech.fr>

Ekhine Irurozki <irurozki@telecom-paris.fr>

Exercice 1. On se place dans le cadre de la classification binaire, comme au TD 4. On observe cette fois-ci un n-échantillon $\mathcal{D}_n = \{(X_i, Y_i)\}_{1 \leq i \leq n}$, représentant n copies indépendantes du (X, Y). Pour des paramètres $\beta \in \mathbb{R}^d$ et $\beta_0 \in \mathbb{R}$ on note $f_{\beta,\beta_0} : x \in \mathbb{R}^d \mapsto \beta^\top x + \beta_0$ puis H_{β,β_0} l'hyperplan (affine) défini par $H_{\beta,\beta_0} = \{x \in \mathbb{R}^d : f_{\beta,\beta_0}(x) = 0\}$. On considère les classifieurs de la forme $\mathrm{sgn} \circ f_{\beta,\beta_0}$ et on suppose que les données sont linéairement séparables.

1) Soit $x \in \mathbb{R}^d$. Calculer $d(x, H_{\beta,\beta_0})$ où d est la distance euclidienne.

2) On propose comme mesure d'erreur associée à une fonction f_{β,β_0} (*Hinge loss*) :

$$\ell_H(f_{\beta,\beta_0}) = - \sum_{\{i : Y_i f_{\beta,\beta_0}(X_i) < 0\}} Y_i \, f_{\beta,\beta_0}(X_i).$$

   Comparer avec l'écriture de la perte $\ell_{0/1}$ classique utilisée habituellement.

3) Vérifier que l'on peut toujours se ramener à la situation où $\beta_0 = 0$. On supposera dorénavant que c'est le cas.

4) L'algorithme du perceptron consiste à optimiser en β la perte ℓ_H avec une approche de type descente de (sous-)gradient stochastique. On pourra introduire un paramètre de taux d'apprentissage ρ (*learning rate*) where $\rho : k \in \mathbb{N}^* \mapsto \rho \in \mathbb{R}^*$, et l'on notera β^k, le coefficient obtenu lorsque l'on a commis la k^{e} erreur.

   Discuter d'initialisations possibles pour β^0, et proposer les étapes successives pour mettre à jour β au vue des données.

On va prouver le résultat suivant : "Si les deux classes sont strictement séparables par un hyperplan sur l'échantillon d'apprentissage alors le perceptron s'arrête en un nombre fini d'étapes". On supposera dans la suite que les X_i sont presque-sûrement de norme (euclidienne) bornée par $B > 0$.

5) Pour tout $\beta \in \mathbb{R}^d$ on définit $\gamma(\beta) := \min_{1 \leq 1 \leq n} Y_i \, \beta^\top X_i$. Interpréter ce paramètre et en déduire une formulation explicite de l'hypothèse de séparabilité linéaire. On pose $\gamma^* := \max_{\|\beta\|=1} \gamma(\beta)$. Vérifier son existence. Quel est son signe ?

6) Par simplicité on initialise l'algorithme avec $\beta^0 = 0$. Montrer alors que le perceptron s'arrête en au plus $\left(\frac{B}{\gamma^*}\right)^2$ mises à jour.

Indication. Borner $\|\beta^k\|$ pour $k \geq 1$ avant la dernière mise à jour.

Enfin, pour plus de détails sur le perceptron, une bonne reference est le livre de Ripley : [Ripley, 1996].

Solution.

1) Notons $\langle ., . \rangle$ le produit scalaire canonique sur $\mathbb{R}^d$ et $\|.\|$ la norme associée. Soit $x \in \mathbb{R}^d$ et $\rho(x)$ son projeté orthogonal sur H_{β,β_0}. La distance de x à H_{β,β_0} correspond à $\|x - \rho(x)\|$, qu'il s'agit de déterminer.

   Tout d'abord, par définition du projeté orthogonal, $x - \rho(x) \in H_{\beta,\beta_0}^{\perp}$, où $H_{\beta,\beta_0}^{\perp} = \text{Vect}(\beta) = \{\lambda\,\beta : \lambda \in \mathbb{R}\}$. Il existe donc $\lambda \in \mathbb{R}$ tel que $x - \rho(x) = \lambda\,\beta$, d'où $\|x - \rho(x)\| = |\lambda|\,\|\beta\|$.

   Ensuite, en remarquant que $x = \rho(x) + x - \rho(x) = \rho(x) + \lambda\,\beta$, il vient que $\rho(x) = x - \lambda\,\beta$. Or $\rho(x) \in H_{\beta,\beta_0}$ donc $\langle \rho(x), \beta \rangle + \beta_0 = \langle x - \lambda\,\beta, \beta \rangle + \beta_0 = 0$, puis $\langle x, \beta \rangle - \lambda\,\|\beta\|^2 + \beta_0 = 0$ et enfin $\lambda = \dfrac{\langle x, \beta \rangle + \beta_0}{\|\beta\|^2}$.

   On en conclut que $d\left(x, H_{\beta,\beta_0}\right) = \|x - \rho(x)\| = |\lambda|\,\|\beta\| = \dfrac{|\langle x, \beta \rangle + \beta_0|}{\|\beta\|}$.

2) La perte 0-1 classique s'écrit

$$\ell_{0/1}(f_{\beta,\beta_0}) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{Y_i \neq f_{\beta,\beta_0}(X_i)\}},$$

   et se veut la contrepartie empirique du risque

$$L_{0/1}(f_{\beta,\beta_0}) := \mathbb{E}\left(\mathbb{1}_{\{Y \neq X\}}\right) = \mathbb{P}\left(Y \neq X\right).$$

   En comparaison, posons $\phi : z \in \mathbb{R} \mapsto -z\,\mathbb{1}_{\{z<0\}} \in \mathbb{R}_+$. La perte $\frac{1}{n}\,\ell_H(f_{\beta,\beta_0})$ correspond alors à la contrepartie empirique du risque

$$L_H(f_{\beta,\beta_0}) := \mathbb{E}\left(\phi\left(Y\,f_{\beta,\beta_0}(X)\right)\right).$$

   La fonction ϕ est convexe, on est donc dans un cas similaire à celui étudié au TD4. La différence majeure est qu'ici, ϕ n'est pas dérivable sur tout son domaine de définition (en 0, on considèrera alors son sous-différentiel).

   Alors que la minimisation du risque empirique $\ell_{0/1}$ est typiquement NP-difficile, recourir la perte ℓ_H permet de se ramener à un problème d'optimisation convexe, que l'on sait bien résoudre.

3) En pratique, le plus simple est d'ajouter une $(d+1)^{\text{e}}$ composante aux descripteurs, toujours égale à 1 : pour tout $i \in [\![1, n]\!]$ on redéfinit $X_i = \left(1, X_{i,1}, \ldots, X_{i,d}\right)^\top$ à valeurs dans $\{1\} \times \mathbb{R}^d$ et on cherche $\beta = (\beta_0, \beta_1, \ldots, \beta_d)^\top \in \mathbb{R}^{d+1}$ qui minimise la perte.

Une approche plus mathématique du problème est développée à la fin du présent corrigé (*cf.* Compléments sur la Question 3).

On suppose dorénavant que $\beta_0 = 0$. Pour simplifier les notations, on écrira H_β au lieu de $H_{\beta,0}$.

4) Comme pour tout algorithme, le choix du paramètre initial a un impact sur la vitesse de convergence, sur le fait de se retrouver coincé dans un minimum local ou non, *etc.* . Son rôle exact dans le présent contexte sera vu plus en détail à la question 6. En pratique, il peut typiquement être choisi arbitrairement (*e.g.* $\beta^0 = 0$) ou, plus communément, aléatoirement. La littérature regorge de techniques d'initialisation permettant de maximiser les performances des algorithmes.

On propose l'algorithme suivant (sous forme de pseudo-code), appelé algorithme du Perceptron, où le taux d'apprentissage peut même être une fonction non constante du nombre de mises à jour, notée $\rho : k \in \mathbb{N}^* \mapsto \rho_k \in \mathbb{R}_+^*$.

Algorithme Perceptron

Entrée : • $\mathcal{D}_n$ et $\rho : k \in \mathbb{N}^* \mapsto \rho_k \in \mathbb{R}_+^*$ ▷ Les données et le taux d'apprentissage
• β^0 ▷ Le paramètre initial

Initialisation : • $k \leftarrow 0$ ▷ Le nombre de mises à jour
• $cv \leftarrow$ FALSE ▷ Un test de convergence de l'algorithme

Procédure :

tant que $cv =$ FALSE **faire**
  $I \leftarrow \{1, \ldots, n\}$ ▷ Les observations que l'on peut tirer
  $m \leftarrow$ FALSE ▷ Un booléen indiquant si on a rencontré une donnée mal classée
  tant que $m =$ FALSE **faire**
    tirer i uniformément dans I
    si $Y_i \langle \beta^k, X_i \rangle \leq 0$ **alors** ▷ Si l'observation i est mal classée
      $m \leftarrow$ TRUE ▷ On l'indique
      $\beta^{k+1} \leftarrow \beta^k + \rho_{k+1} Y_i X_i$ ▷ On actualise le paramètre
      $k \leftarrow k + 1$ ▷ On passe à la mise à jour suivante
    sinon si $\#I = 1$ **alors** ▷ Sinon, si c'était la dernière observation disponible
      $m \leftarrow$ TRUE ▷ On arrête de force la boucle
      $cv \leftarrow$ TRUE ▷ On indique que tout est bien classé (on a convergé)
    sinon $I \leftarrow I \backslash \{i\}$ ▷ Sinon on l'enlève et on recommence

---

5) Notons $\mathcal{P} := \left\{i \in [\![1, n]\!] : Y_i = +1\right\}$ et $\mathcal{N} := \left\{i \in [\![1, n]\!] : Y_i = -1\right\}$ les ensembles d'observations de chacune des deux classes et prenons $\beta \in \mathbb{R}^d$.

L'hyperplan vectoriel H_β sépare les données ssi le classifieur $\operatorname{sgn} \circ f_{\beta,0}$ ne commet aucune erreur (empirique), *i.e.*

$$\min_{i \in \mathcal{P}} \langle \beta, X_i \rangle > 0 \quad \text{et} \quad \max_{i \in \mathcal{N}} \langle \beta, X_i \rangle \leq 0. \tag{i}$$

Remarque. Les ensembles H_β et $H_{\beta/\|\beta\|}$ coïncident. On peut donc dorénavant supposer sans perte de généralité que $\|\beta\| = 1$.

Interprétation de $\gamma(\beta)$. En remarquant que $\min_{i \in \mathcal{N}} Y_i \langle \beta, X_i \rangle = \min_{i \in \mathcal{N}} -\langle \beta, X_i \rangle = -\max_{i \in \mathcal{N}} \langle \beta, X_i \rangle$, on obtient que

$$\gamma(\beta) := \min_{1 \leq i \leq n} Y_i \langle \beta, X_i \rangle = \left(\min_{i \in \mathcal{P}} Y_i \langle \beta, X_i \rangle \right) \wedge \left(\min_{i \in \mathcal{N}} Y_i \langle \beta, X_i \rangle \right) = \left(\min_{i \in \mathcal{P}} \langle \beta, X_i \rangle \right) \wedge \left(- \max_{i \in \mathcal{N}} \langle \beta, X_i \rangle \right).$$

Ainsi, H_β sépare les données on a $\gamma(\beta) \geq 0$. Ce dernier est nul si l'une des données de label -1 se trouve sur l'hyperplan séparateur. D'après la question 1, comme $\|\beta\| = 1$, il correspond dans ce cas à la plus petite distance des observations à la frontière, appelée la *marge*.
Inversement, lorsque H_β ne sépare pas les données, $\gamma(\beta) \leq 0$ et $-\gamma(\beta)$ donne la plus grande distance des points mal classés à la frontière.

Séparabilité linéaire. Nous avons vu que si H_β sépare les données alors $\gamma(\beta) \geq 0$, mais que la réciproque est fausse : si $\gamma(\beta) = 0$ on ne peut pas conclure à ce stade.
Malheureusement, il existe des situations exceptionnelles dans lesquelles le seul H_β séparant les données donne $\gamma(\beta) = 0$. Par exemple, quand $d = 2$, considérez $X_1 = (0, 0)$, $X_2 = (1, 1)$ et $X_3 = (1, -1)$ de labels $Y_1 = +1$ puis $Y_2 = Y_3 = -1$. Alors le seul hyperplan *vectoriel* qui sépare X_1 de X_2 et X_3 est la droite passant exactement par X_2 et X_3, *i.e.* de vecteur normal $\beta := (0, 1)^\intercal$ qui vérifie $\gamma(\beta) = 0$.
Le problème ne se pose pas si l'on cherche un hyperplan séparateur *affine* (*cf.* Compléments sur la Question 3).
Si l'on veut absolument chercher un hyperplan vectoriel, on peut toujours supposer que les situations exceptionnelles se produisent avec probabilité nulle. Cela sera le cas, par exemple, si l'on suppose que la loi des descripteurs admet une densité. On énoncera alors les résultats en précisant qu'ils sont presque-sûrs.
Vue la transformation suggérée à la question 3 (qui permet de ne pas se restreindre aux hyperplans vectoriels), nous utiliserons la définition suivante : les données sont linéairement séparables ssi il existe $\beta \in \mathbb{R}^d$ de norme $\|\beta\| = 1$ tel que $\gamma(\beta) > 0$.

Existence de γ^*. Pour prouver que le paramètre $\gamma^* := \max_{\|\beta\|=1} \gamma(\beta)$ est bien défini, notons $\mathbb{S} := \{x \in \mathbb{R}^d : \|x\| = 1\}$ la sphère unité de $\mathbb{R}^d$ muni de la norme euclidienne. Elle est un compact de

$\mathbb{R}^d$. En outre, l'application $\gamma : \beta \in \mathbb{R}^d \mapsto \min_{1 \leq i \leq n} Y_i \, \beta^\intercal X_i \in \mathbb{R}$ est continue comme minimum de n fonctions linéaires. L'image $\gamma(\mathbb{S})$ de $\mathbb{S}$ par γ est donc un compact de $\mathbb{R}$. Le théorème des valeurs extrêmes nous garantit alors qu'elle possède un maximum, *i.e.* que $\gamma^* = \max \gamma(\mathbb{S})$ existe bien.

Signe de γ^*. Lorsque les données sont linéairement séparables, nous avons vu qu'il existe $\beta \in \mathbb{R}^d$ de norme unitaire tel que $\gamma(\beta) > 0$. Il vient immédiatement que $\gamma^* \geq \gamma(\beta) > 0$.
Ce paramètre donne la plus grande marge des séparateurs vectoriels.

6) Nous allons démontrer la convergence de l'algorithme du Perceptron dans un cadre plus général que celui suggéré dans l'exercice. On considère en particulier le cas où le taux d'apprentissage est une fonction non constante du nombre de mises à jour, notée $\rho : k \in \mathbb{N}^* \mapsto \rho_k \in \mathbb{R}_+^*$. On ne suppose plus non plus que $\beta^0 = 0$, pour bien voir l'influence du paramètre d'initialisation.

Supposons que le paramètre initial ne permet pas de séparer les deux classes de données, et soit une mise à jour $k \in \mathbb{N}^*$ précédant la convergence. D'après l'algorithme de la question 4, il existe bien au moins une donnée $i_k \in \{1, \dots, n\}$ mal classée par β^k, *i.e.* telle que $\operatorname{sgn}\left(\langle \beta^{k-1}, X_{i_k} \rangle\right) \neq Y_{i_k}$. Elle sert ensuite à adapter le paramètre pour donner β^k.

Soit en outre $\beta^* \in \mathbb{R}^d$ de norme 1 tel que $\gamma^* = \gamma(\beta^*)$ (on a vu qu'il en existait bien un). Nous allons encadrer la norme de $\beta^k - \beta^0$.

Minorant. Tout d'abord, on a $\left\|\beta^k - \beta^0\right\| = \left\|\beta^k - \beta^0\right\| \left\|\beta^*\right\| \geq \langle \beta^k - \beta^0, \beta^* \rangle$ d'après l'inégalité de Cauchy-Schwartz. Ensuite, on remarque que

$$\begin{aligned}\langle \beta^k - \beta^0, \beta^* \rangle &= \langle \beta^{k-1} + \rho_k \, Y_{i_k} X_{i_k} - \beta^0, \beta^* \rangle = \langle \beta^{k-1} - \beta^0, \beta^* \rangle + \rho_k \, Y_{i_k} \, \langle X_{i_k}, \beta^* \rangle \\ &\geq \langle \beta^{k-1} - \beta^0, \beta^* \rangle + \rho_k \min_{1 \leq i \leq n} Y_i \, \langle \beta^*, X_i \rangle = \langle \beta^{k-1} - \beta^0, \beta^* \rangle + \rho_k \, \gamma^*.\end{aligned}$$

Par récurrence cette dernière inégalité implique que $\langle \beta^k - \beta^0, \beta^* \rangle \geq \langle \beta^0 - \beta^0, \beta^* \rangle + \gamma^* \sum_{\ell=1}^{k} \rho_\ell$. On obtient donc un minorant de la norme de $\beta^k - \beta^0$:

$$\left\|\beta^k - \beta^0\right\| \geq \gamma^* \sum_{\ell=1}^{k} \rho_\ell. \tag{1}$$

Majorant. Commençons par remarquer que

$$\left\|\beta^k - \beta^0\right\|^2 = \left\|\beta^{k-1} + \rho_k \, Y_{i_k} X_{i_k} - \beta^0\right\|^2 = \left\|\beta^{k-1} - \beta^0\right\|^2 + \rho_k^2 \, Y_{i_k}^2 \left\|X_{i_k}\right\|^2 + 2 \, \rho_k \, Y_{i_k} \, \langle \beta^{k-1} - \beta^0, X_{i_k} \rangle.$$

Or par hypothèse $Y_{i_k}^2 = 1$ et $\left\|X_{i_k}\right\| \leq B^2$ presque-sûrement. En outre, $Y_{i_k} \, \langle \beta^{k-1}, X_{i_k} \rangle < 0$ par construction de l'algorithme et $Y_{i_k} \langle \beta^0, X_{i_k} \rangle \geq \min_{1 \leq i \leq n} Y_i \, \langle \beta^0, X_i \rangle = \gamma(\beta^0)$ par définition. Pour

simplifier les notations, posons $\gamma^0 := \gamma(\beta^0)$. Cette quantité est négative (ou nulle si β^0 est le vecteur nul), sinon β^0 serait directement solution du problème d'optimisation, contrairement à ce qui a été supposé. On obtient donc que $\left\|\beta^k - \beta^0\right\|^2 < \left\|\beta^{k-1} - \beta^0\right\|^2 + \rho_k^2 B^2 + 2\rho_k |\gamma^0|$, ce qui implique par récurrence que

$$\left\|\beta^k - \beta^0\right\|^2 < B^2 \sum_{\ell=1}^{k} \rho_\ell^2 + 2|\gamma^0| \sum_{\ell=1}^{k} \rho_\ell. \tag{2}$$

Conclusion. En combinant les équations (1) et (2) on obtient que si la convergence n'a pas été atteinte avant la k^{e} mise à jour, alors k satisfait l'inégalité suivante :

$$1 < \left(\frac{B}{\gamma^*}\right)^2 \frac{\sum_{\ell=1}^{k} \rho_\ell^2}{\left(\sum_{\ell=1}^{k} \rho_\ell\right)^2} + \frac{2|\gamma^0|}{(\gamma^*)^2} \frac{1}{\sum_{\ell=1}^{k} \rho_\ell}. \tag{3}$$

► Si ρ est une fonction constante égale à $\rho_0 > 0$ et β^0 est le vecteur nul, alors l'inégalité (3) se simplifie pour donner $k < \left(\frac{B}{\gamma^*}\right)^2$. Dans ce cas, le taux d'apprentissage n'a aucun impact sur la vitesse de convergence de l'algorithme.

► Si ρ est une fonction constante égale à $\rho_0 > 0$ et β^0 est de norme strictement positive, alors cette (3) devient $k < \left(\frac{B}{\gamma^*}\right)^2 + \frac{2|\gamma^0|}{\rho_0 (\gamma^*)^2}$.

► Si ρ est telle que $\lim_{k\to+\infty} \sum_{\ell=1}^{k} \rho_\ell = +\infty$ et $\lim_{k\to+\infty} \frac{\sum_{\ell=1}^{k} \rho_\ell^2}{\left(\sum_{\ell=1}^{k} \rho_\ell\right)^2} = 0$, alors la partie de droite de l'inégalité (3) tend vers 0 lorsque k grandit. En d'autres termes, il existe un $k_0 \in \mathbb{N}^*$ au delà duquel (3) n'est plus respectée. Cela garantit que l'algorithme converge en un nombre fini de mises à jour, et ce d'autant plus rapidement que la partie de droite de (3) tend vers 0.

Compléments sur la Question 3

Notons $\mathcal{P} := \left\{i \in [\![1, n]\!] : Y_i = +1\right\}$ et $\mathcal{N} := \left\{i \in [\![1, n]\!] : Y_i = -1\right\}$ les ensembles d'observations de chacune des deux classes. Les données sont linéairement séparables ssi il existe $\beta \in \mathbb{R}^d$ et $\beta_0 \in \mathbb{R}$ tels que

$$\min_{i\in\mathcal{P}} \langle \beta, X_i \rangle + \beta_0 > 0 \quad \text{et} \quad \max_{i\in\mathcal{N}} \langle \beta, X_i \rangle + \beta_0 \leq 0. \tag{4}$$

Prenons maintenant les enveloppes convexes des descripteurs des deux classes, notées

$$C_{\mathcal{P}} := \left\{ \sum_{i\in\mathcal{P}} \lambda_i X_i : \min_{i\in\mathcal{P}} \lambda_i \geq 0\,,\ \sum_{i\in\mathcal{P}} \lambda_i = 1 \right\} \quad \text{et} \quad C_{\mathcal{N}} := \left\{ \sum_{i\in\mathcal{N}} \lambda_i X_i : \min_{i\in\mathcal{N}} \lambda_i \geq 0\,,\ \sum_{i\in\mathcal{N}} \lambda_i = 1 \right\}.$$

Remarque. Un hyperplan sépare les deux classes de données ssi il sépare $C_{\mathcal{P}}$ et $C_{\mathcal{N}}$. La preuve est aisée et laissée en exercice.

Soient $x^+ \in C_{\mathcal{P}}$ et $x^- \in C_{\mathcal{N}}$ tels que $\|x^+ - x^-\| = \min\left\{\|x - z\| : x \in C_{\mathcal{P}}, z \in C_{\mathcal{N}}\right\}$; ce minimum existe bien car $(x, z) \in C_{\mathcal{P}} \times C_{\mathcal{N}} \mapsto \|x - z\| \in \mathbb{R}_+$ est un fonction continue sur un compact. On note $\mathcal{S} := \left\{u\, x^+ + (1-u)\, x^- : u \in [0, 1[\right\}$ le segment d'extrémités x^- et x^+ privé de ce dernier.

► Tout hyperplan séparant les données coupe $\mathcal{S}$.

▷ *Preuve.* Soient $\beta \in \mathbb{R}^d$ et $\beta_0 \in \mathbb{R}$ satisfaisant (4). L'hyperplan $\mathrm{H}_{\beta,\beta_0}$ coupe $\mathcal{S}$ ssi il existe $u \in [0, 1[$ tel que $\langle \beta, u\, x^+ + (1-u)\, x^- \rangle + \beta_0 = 0$, *i.e.* vérifiant $u \langle \beta, x^+ - x^- \rangle = -\langle \beta, x^- \rangle - \beta_0$. Or d'après la remarque précédente on a d'une part $\langle \beta, x^+ \rangle + \beta_0 > 0$ et d'autre part $\langle \beta, x^- \rangle + \beta_0 \leq 0$. Par conséquent, $\langle \beta, x^+ - x^- \rangle = \langle \beta, x^+ \rangle + \beta_0 - \langle \beta, x^- \rangle - \beta_0 > -\langle \beta, x^- \rangle - \beta_0 \geq 0$. On en déduit que $u = -\dfrac{\langle \beta, x^- \rangle + \beta_0}{\langle \beta, x^+ - x^- \rangle} \in [0, 1[$ satisfait bien la contrainte désirée.

► Réciproquement, pour tout point de $\mathcal{S}$ on peut trouver un hyperplan séparant les données qui le rencontre.

▷ *Preuve.* Soient $m := \dfrac{x^+ + x^-}{2}$ le milieu du segment $\mathcal{S}$ et $r := \dfrac{\|x^+ - x^-\|}{2}$ sa demi-longueur. On note $\mathcal{B} := \left\{x \in \mathbb{R}^d : \|x - m\| < r\right\}$ la boule ouverte de centre m et de rayon r.

On remarque d'abord que $\mathcal{B} \cap C_{\mathcal{P}} = \mathcal{B} \cap C_{\mathcal{N}} = \emptyset$. En effet, soit $x_0 \in C_{\mathcal{P}} \cup C_{\mathcal{N}}$. Par l'absurde, si $x_0 \in \mathcal{B}$, alors $\|x_0 - x^-\| \vee \|x_0 - x^+\| \leq 2\, r \vee 2\, r = 2\, r = \|x^+ - x^-\| := \min\left\{\|x - z\| : x \in C_{\mathcal{P}}, z \in C_{\mathcal{N}}\right\}$, ce qui est contradictoire.

Soient maintenant un point z de $\mathcal{S}$ puis β^z et β_0^z les paramètres de l'hyperplan (noté H^z) perpendiculaire à $\mathcal{S}$ passant par z, choisis tels que $\langle \beta^z, x^+ \rangle + \beta_0^z > 0$. Comme H^z sépare x^+ et x^- par construction, on a par ailleurs $\langle \beta^z, x^- \rangle + \beta_0^z \leq 0$. On montre par l'absurde que H^z sépare bien les deux classes de données (*cf.* Figure 1 pour une illustration). En effet, supposons qu'il existe $i \in \mathcal{P}$ tel que X_i est du même côté de H^z que x^-, *i.e.* tel que $\langle \beta^z, \mathrm{X}_i \rangle + \beta_0^z \leq 0$. Alors le segment d'extrémités X_i et x^+ rencontre forcément $\mathcal{B}$. Or par définition, ce segment est aussi contenu dans l'enveloppe convexe $C_{\mathcal{P}}$, d'où $C_{\mathcal{P}} \cap \mathcal{B} \neq \emptyset$, ce que l'on a montré être impossible. De la même manière si l'on suppose qu'il existe $i \in \mathcal{N}$ du même côté de H^z que x^+, alors on aboutit à la contradiction que $\mathcal{B} \cap C_{\mathcal{N}} \neq \emptyset$.

Interprétation. On déduit de ces propriétés les deux résultats suivants :

— Les données sont linéairement séparables ssi elles sont *strictement* linéairement séparables, *i.e.* ssi elles peuvent être séparées par un hyperplan qui ne contient aucune d'entre elles. Cela justifie la réponse à la question 4.

— Si l'on choisit un point z dans $\mathcal{S}$, alors l'ensemble des descripteurs translatés $\{\mathrm{X}_i - z\}_{1 \leq i \leq n}$ est séparable par un hyperplan vectoriel (non plus affine). S'il explique le bien-fondé de la question 3,

ce résultat n'a pas grand intérêt algorithmique : déterminer les enveloppes convexes de chacune des classes de données et trouver x^+ et x^- peut s'avérer très lourd si n et d sont assez grands.

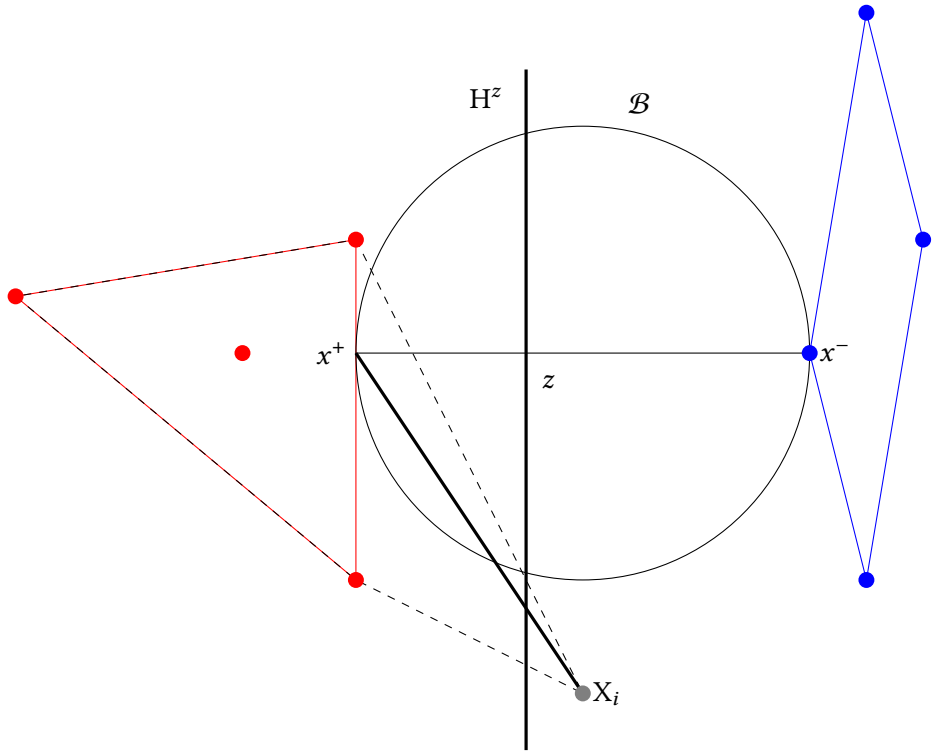

FIGURE 1 – Illustration de la situation considérée dans la preuve. Les labels positifs sont en rouge et les négatifs en bleu. Si X_i était une donnée de label positif, on voit que l'enveloppe convexe qui en résulterait (délimitée par les tirets noirs) intersecterait $\mathcal{B}$, créant une contradiction.

Références

[Ripley, 1996] Ripley, B. (1996). *Pattern recognition via neural networks.* Cambridge University Press. 2