Hicham Janati

hjanati@insea.ac.ma

INSEA

Chapitre 3. Applications et thématiques avancées

1. Modèles Bayésiens hiérarchiques
2. Bayesian Machine learning

INSEA

Modèles Bayésiens hiérarchiques

On souhaite modéliser la fréquence des sinistres d'un ensemble de conducteurs dans trois villes différentes A, B, C

Trois variables à estimer: $\lambda_A, \lambda_B, \lambda_C$

On peut considérer une approche indépendante:

$$\begin{array}{ccc} \boxed{\textbf{Données A}} & \boxed{\textbf{Données B}} & \boxed{\textbf{Données C}} \\ \uparrow & \uparrow & \uparrow \\ \lambda_A & \lambda_B & \lambda_C \\ \lambda_A \sim \text{prior}_A & \lambda_B \sim \text{prior}_B & \lambda_C \sim \text{prior}_C \end{array}$$

Quels sont les inconvénients de ce modèle ?

Aucun lien entre les régions: on n'exploite pas les similarités entre les régions

Et si on utilise la même prior ?

$$\begin{array}{ccc} \boxed{\textbf{Données A}} & \boxed{\textbf{Données B}} & \boxed{\textbf{Données C}} \\ \uparrow & \uparrow & \uparrow \\ \lambda_A & \lambda_B & \lambda_C \end{array}$$

$$\lambda_A, \lambda_B, \lambda_C \sim \text{Gamma}(\alpha, \beta)$$

α, β fixés (vaguement, ou données historiques)

Implicitement à quoi correspondent les quantités:

$$\frac{\alpha}{\beta} \text{ et } \frac{\alpha}{\beta^2}$$

Quels sont les inconvénients de ce modèle ?

INSEA

Modèles Bayésiens hiérarchiques

Données A

Données B

Données C

λ_A λ_B λ_C

Ne pas forcer les paramètres a priori, les considérer comme des variables aléatoires à estimer:

prior $\lambda_A, \lambda_B, \lambda_C \sim \text{Gamma}(\alpha, \beta)$

hyperprior $\alpha \sim \text{prior}(a)$ $\beta \sim \text{prior}(b)$ a, b fixés (vaguement, données historiques)

Un modèle bayésien hiérarchique modélise les similarités et les différences entre les groupes à partir des données

Données de mortalité dans des hôpitaux américains.

YEAR	HOSPITAL	Procedure/Condition	# of Deaths	# of Cases
2016	Highland Hospital	Acute Stroke	17	147
2016	Highland Hospital	Acute Stroke Hemorrhagic	10	36
2016	Highland Hospital	Acute Stroke Ischemic	6	106
2016	Highland Hospital	Acute Stroke Subarachnoid	1	5
2016	Highland Hospital	Carotid Endarterectomy	0	5
2016	Highland Hospital	Espophageal Resection	0	3
2016	Highland Hospital	GI Hemorrhage	4	147
2016	Highland Hospital	Heart Failure	1	317
2016	Highland Hospital	Hip Fracture	1	38
2016	Highland Hospital	PCI	10	132

1. Vous êtes data scientist.
2. Votre tâche est vague: “on veut un rapport sur les hôpitaux dans le pays”
3. Que faites-vous ?

INSEA

YEAR	HOSPITAL	Procedure/Condition	# of Deaths	# of Cases
2016	Highland Hospital	Acute Stroke	17	147
2016	Highland Hospital	Acute Stroke Hemorrhagic	10	36
2016	Highland Hospital	Acute Stroke Ischemic	6	106
2016	Highland Hospital	Acute Stroke Subarachnoid	1	5
2016	Highland Hospital	Carotid Endarterectomy	0	5
2016	Highland Hospital	Espophageal Resection	0	3

1. Problématiques simples :

1. Classement des hôpitaux par taux de mortalité
2. Classement des procédures par taux de mortalité
3. Classement des hôpitaux + procédures par taux de mortalité
4. Étudier l'évolution des taux de mortalité dans le temps

2. Statistiques descriptives:

1. Combien y a-t-il d'hôpitaux ? de procédures ? d'années ?
2. Données manquantes / dupliquées ?
3. Calculer un taux de mortalité fréquentiste.
4. Visualiser les hôpitaux / procédures avec une ACP.
5. Clusters évidents ? Outliers ?

3. Modélisation bayésienne

1. Pourquoi ne pas se contenter des taux fréquentistes ?
2. Définir les groupes et les lois a priori
3. Interpréter les taux de mortalité avec leur HDI

4. Expliquer ces données avec des données externes

1. Données géographiques (ville / quartier de l'hôpital)
2. Données par hôpital (effectif, technologies utilisées, reviews)
3. Données temporelles (événements rares: accidents, pandémies..)

Chapitre 3. Applications et thématiques avancées

1. Modèles Bayésiens hiérarchiques (Assurance / Biostats)
2. Classical Machine learning: zero to hero
3. Bayesian Machine learning

Un opérateur téléphonique a les données historiques sur ses clients.

Dependents	TechSupport	Contract	InternetService	Months	MonthlyCharges	Churn
0	1	0	1	12	75.65	0
1	0	0	0	24	89.50	0
0	0	0	1	6	65.25	1
0	1	1	0	48	35.30	?
1	0	0	1	48	85.81	?

Churn = 1: client a annulé son abonnement

L'entreprise souhaite anticiper le "churn" avec un algorithme de prédiction pour cibler les clients concernés

$$\mathbf{X} = (\mathbf{X}^1, \ldots, \mathbf{X}^6) \to y \in \{0, 1\}$$

On cherche une fonction f telle que: $f(\mathbf{X}) \approx y$ $\quad \min_f \sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2$ Erreur de prédiction

f doit donner 1 ou 0, on considère alors des fonctions de type: $f(\mathbf{x}) = \mathbb{1}_{g(\mathbf{x}) \geq 0}$

On ne peut pas chercher g dans la totalité de l'espace des fonctions (dimension infinie), il faut paramétriser g

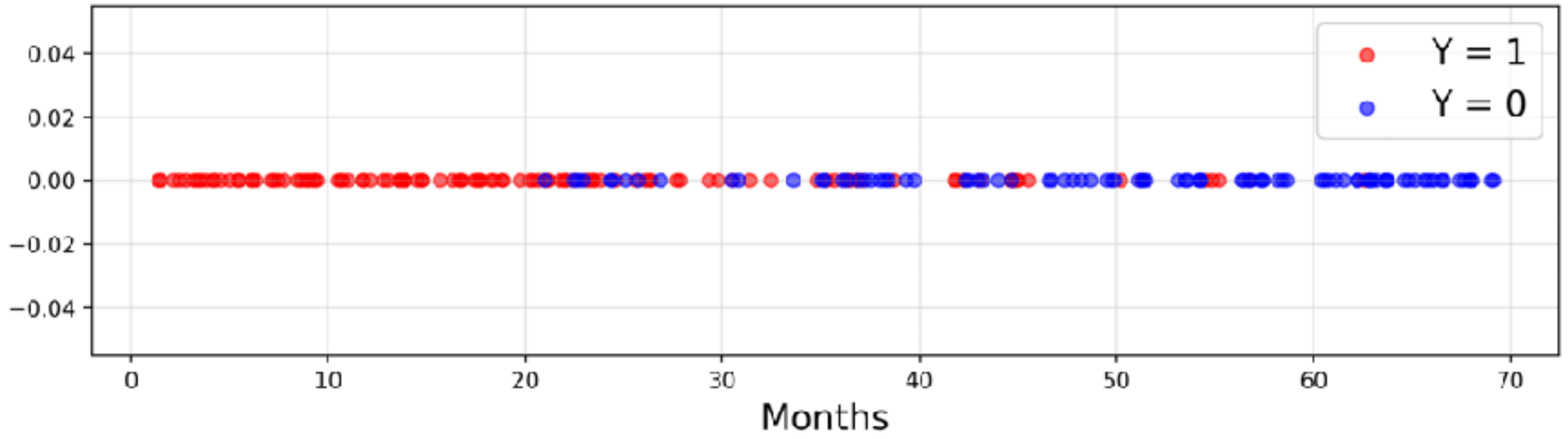
INSEA

f doit donner 1 ou 0, on considère alors des fonctions de type: $f(\mathbf{x}) = \mathbb{1}_{g(\mathbf{x})\geq 0}$

On ne peut pas chercher g dans la totalité de l'espace des fonctions (dimension infinie), il faut paramétriser g

On considère une seule variable "Months" qui donne la durée du contrat:

$$\mathbf{x} = \text{Months} \in \mathbb{R}$$



Quelle serait la fonction paramétrée g la plus simple ici ?

$$g(\mathbf{x}) = \beta_1 \mathbf{x} + \beta_0, \quad \beta_0, \beta_1 \in \mathbb{R}$$

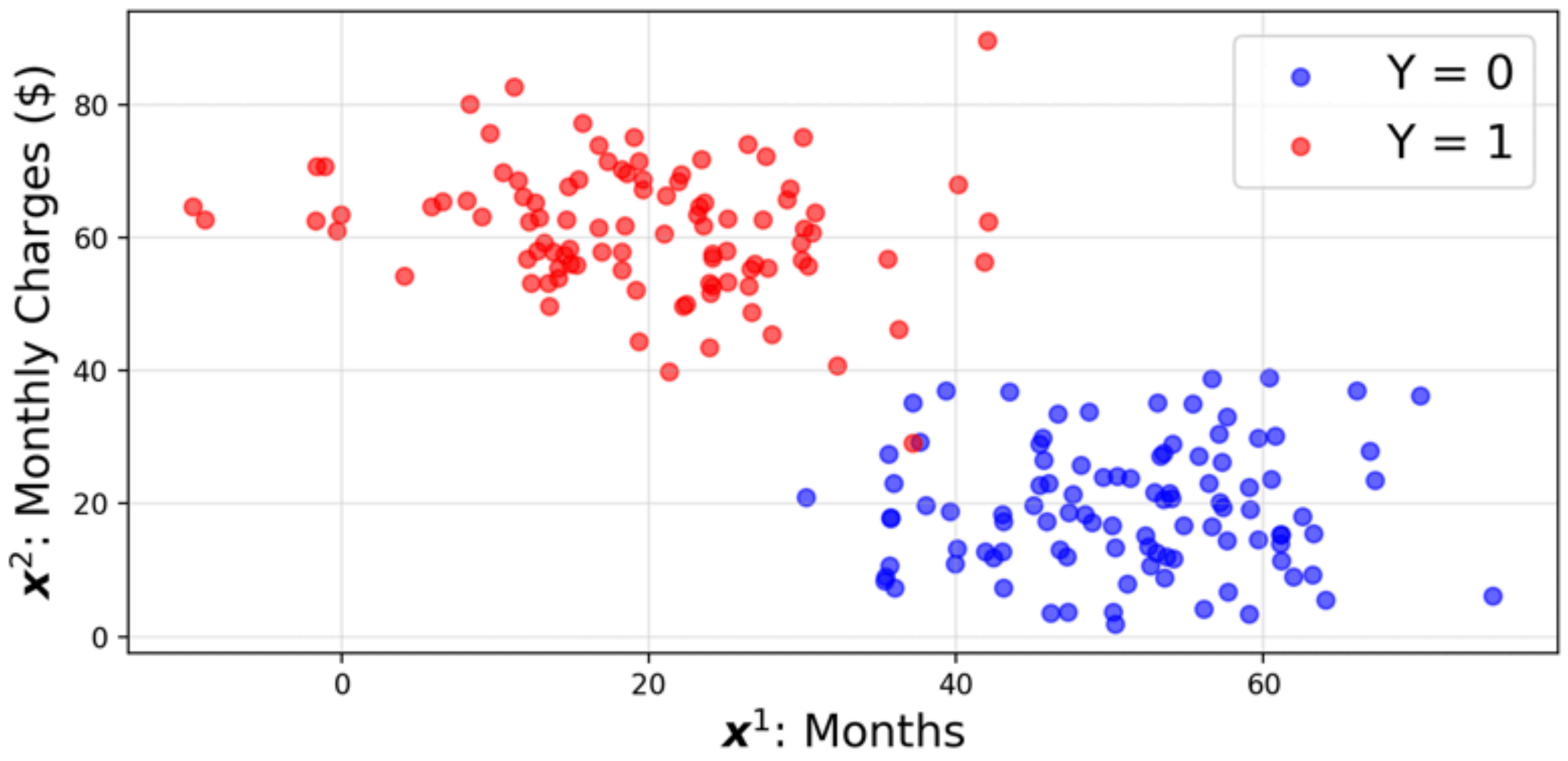
Chercher la meilleure f = chercher le meilleur β: $$\min_{\beta \in \mathbb{R}^2} \sum_{i=1}^{n} (\mathbb{1}_{\{\beta_1 \mathbf{x}_i + \beta_0 \geq 0\}} - y_i)^2$$

Pouvez-vous donner des estimations vagues de ces paramètres ?

séparateur linéaire en dimension 2

On considère une deux variables: “Months” et “MonthlyCharges”:

$$\mathbf{x} = (\mathbf{x}^1, \mathbf{x}^2) \qquad f(\mathbf{x}) = \mathbb{1}_{g(\mathbf{x})\geq 0}$$



Quelle serait la fonction paramétrée g la plus simple ici ?

$$g(\mathbf{x}) = \alpha + \beta_1 \mathbf{x}^1 + \beta_2 \mathbf{x}^2, \quad \alpha, \beta_1, \beta_2 \in \mathbb{R}$$

$$g(\mathbf{x}) = \alpha + \langle \beta, \mathbf{x} \rangle, \quad \alpha \in \mathbb{R}, \beta \in \mathbb{R}^2$$

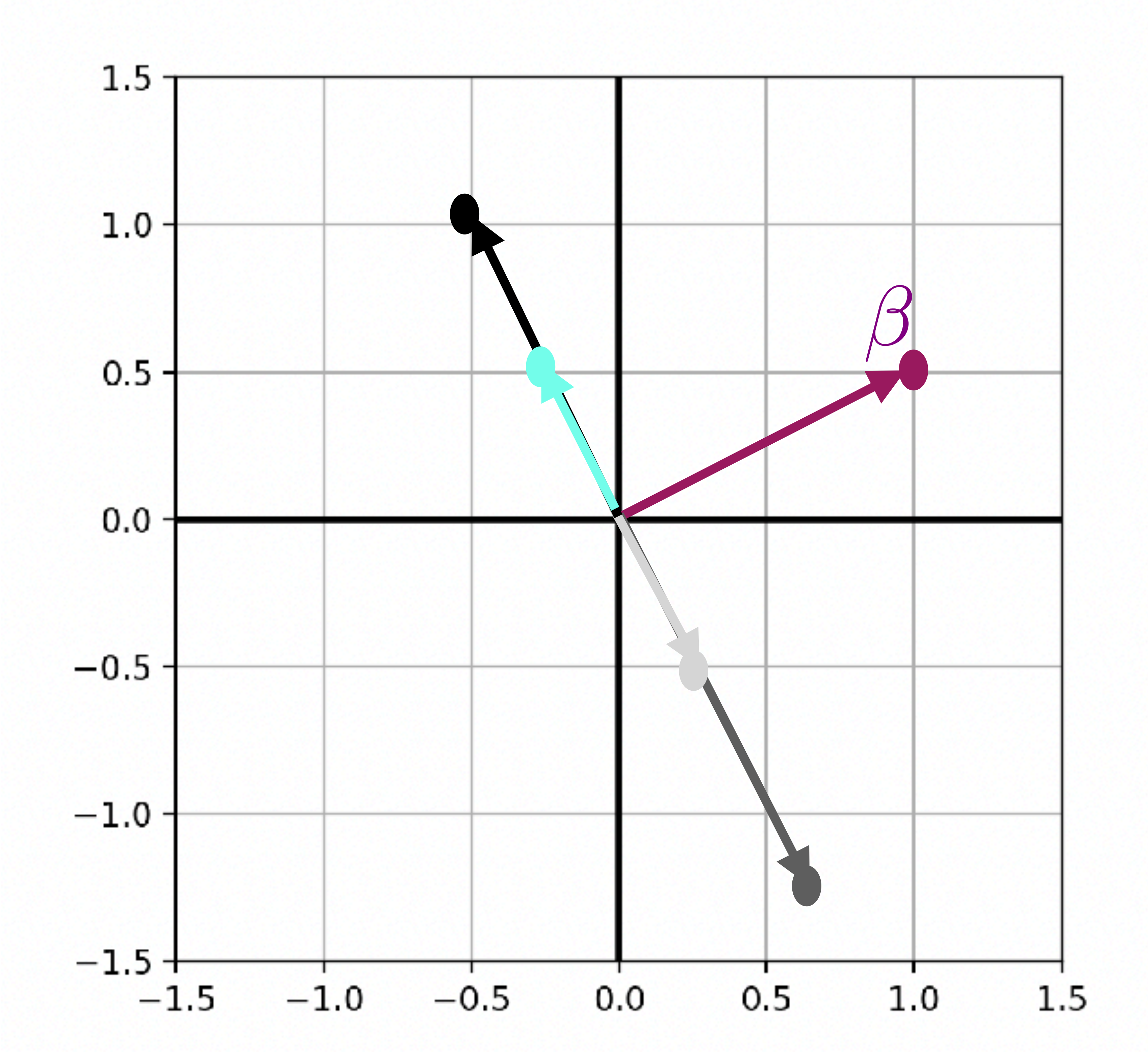
$$g(\mathbf{x}) = \alpha + \beta^\top \mathbf{x}, \quad \alpha \in \mathbb{R}, \beta \in \mathbb{R}^2$$

$$\min_{\alpha\in\mathbb{R}, \beta\in\mathbb{R}^2} \sum_{i=1}^{n} (\mathbb{1}_{\{\alpha+\beta^\top \mathbf{x}_i \geq 0\}} - y_i)^2$$

À quoi ressemble l’ensemble des fonctions g ?

On considère $g : \mathbf{x} \mapsto \beta^\top \mathbf{x}$. Étudions ses courbes de niveaux, c-à-d pour $c \in \mathbb{R}$ les ensembles: $\{\mathbf{x} | g(\mathbf{x}) = c\}$.

On considère $g : \mathbf{x} \mapsto \beta^\top \mathbf{x}$. Étudions ses courbes de niveaux, c-à-d pour $c \in \mathbb{R}$ les ensembles: $\{\mathbf{x} | g(\mathbf{x}) = c\}$.



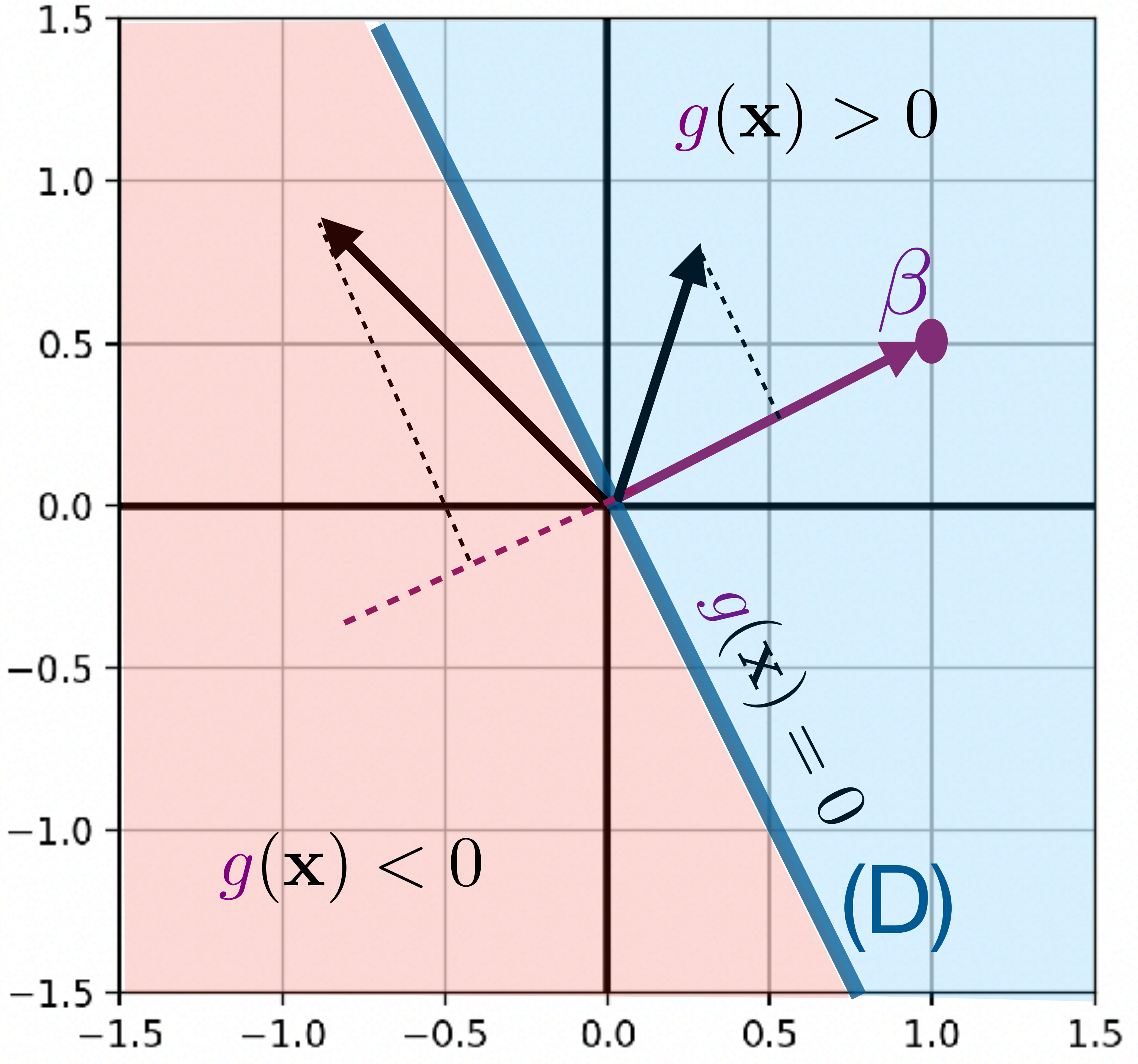
Exemple avec $\beta = (1, 0.5)^\top$ et $c = 0$.

Quels sont les $\mathbf{x}$ tels que $\beta^\top \mathbf{x} = 0$?

Tous les vecteurs orthogonaux à β.

$\{\mathbf{x} \in \mathbb{R}^2 | \beta^\top \mathbf{x} = 0\}$ est la droite perpendiculaire à β.

On considère $g : \mathbf{x} \mapsto \beta^\top \mathbf{x}$. Étudions ses courbes de niveaux, c-à-d pour $c \in \mathbb{R}$ les ensembles: $\{\mathbf{x} | g(\mathbf{x}) = c\}$.



Exemple avec $\beta = (1, 0.5)^\top$ et $c = 0$.

Quels sont les $\mathbf{x}$ tels que $\beta^\top \mathbf{x} = 0$?

Tous les vecteurs orthogonaux à β.

$\{\mathbf{x} \in \mathbb{R}^2 | \beta^\top \mathbf{x} = 0\}$ est la droite perpendiculaire à β.

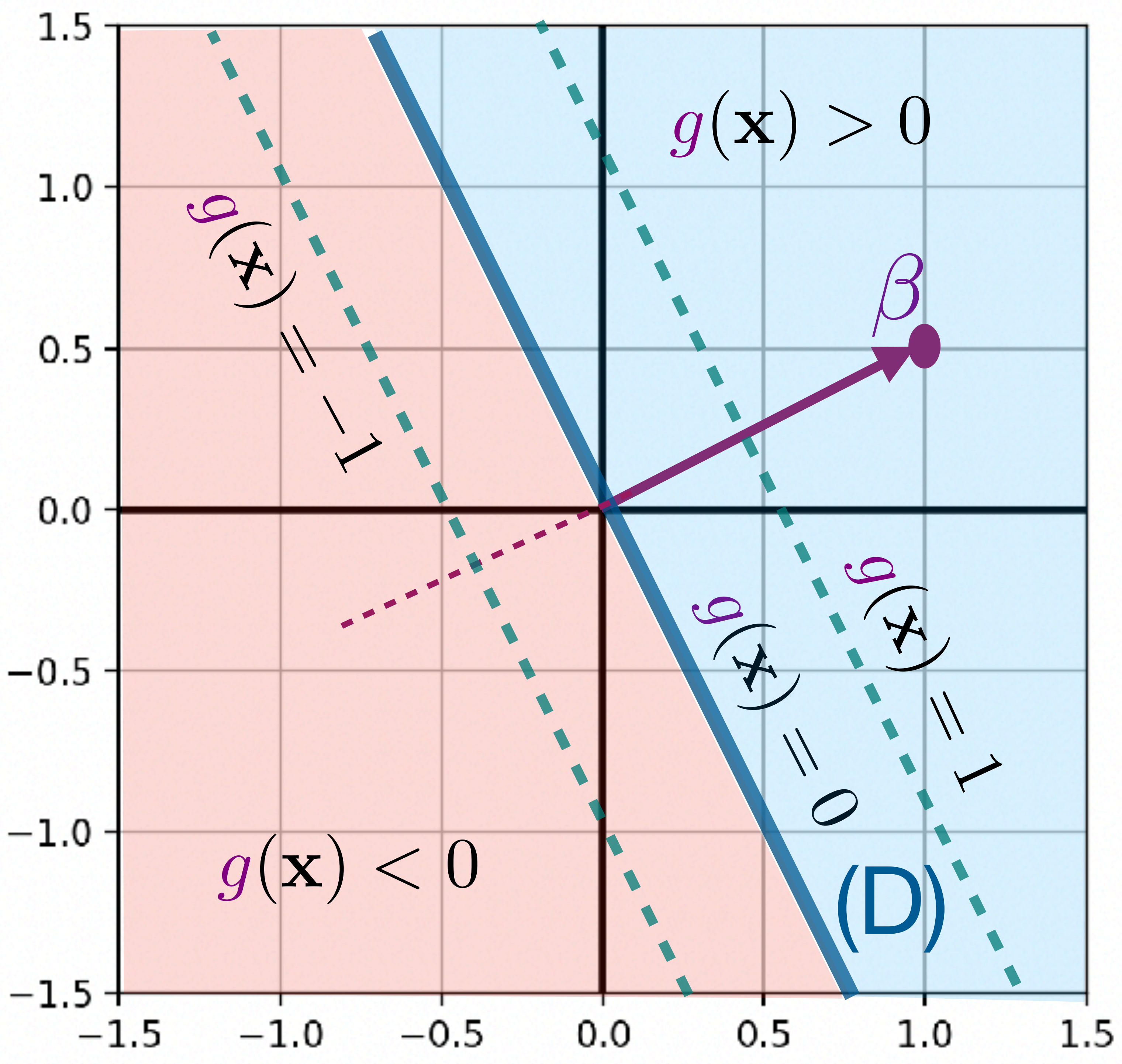
à droite de (D), $\beta^\top \mathbf{x} > 0$

à gauche de (D), $\beta^\top \mathbf{x} < 0$

et si $c = 1$? ou $c = -1$?

INSEA

On considère $g : \mathbf{x} \mapsto \beta^\top \mathbf{x}$. Étudions ses courbes de niveaux, c-à-d pour $c \in \mathbb{R}$ les ensembles: $\{\mathbf{x} | g(\mathbf{x}) = c\}$.



Exemple avec $\beta = (1, 0.5)^\top$ et $c = 0$.

Quels sont les $\mathbf{x}$ tels que $\beta^\top \mathbf{x} = 0$?

Tous les vecteurs orthogonaux à β.

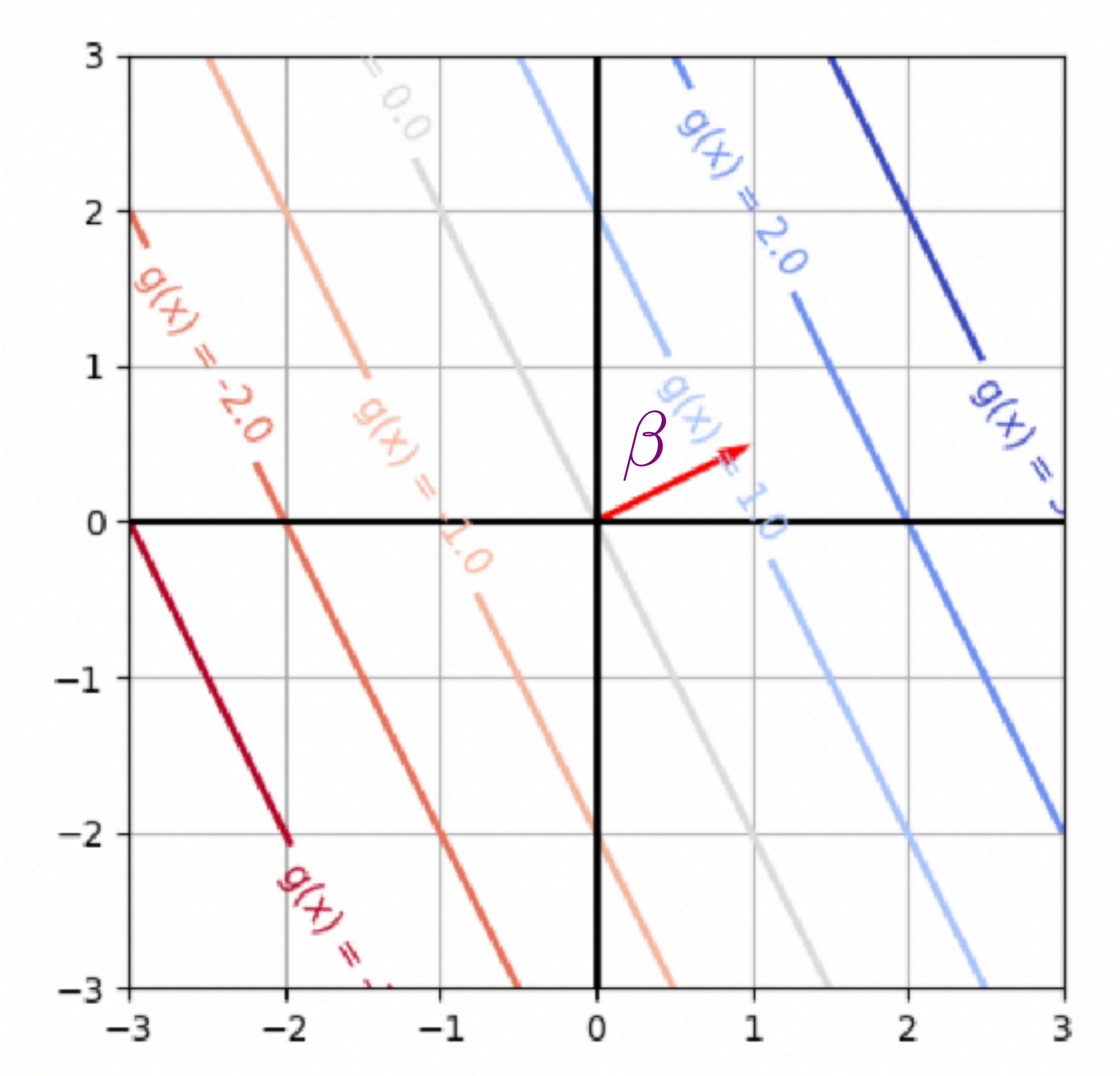
$\{\mathbf{x} \in \mathbb{R}^2 | \beta^\top \mathbf{x} = 0\}$ est la droite perpendiculaire à β.

à droite de (D), $\beta^\top \mathbf{x} > 0$

à gauche de (D), $\beta^\top \mathbf{x} < 0$

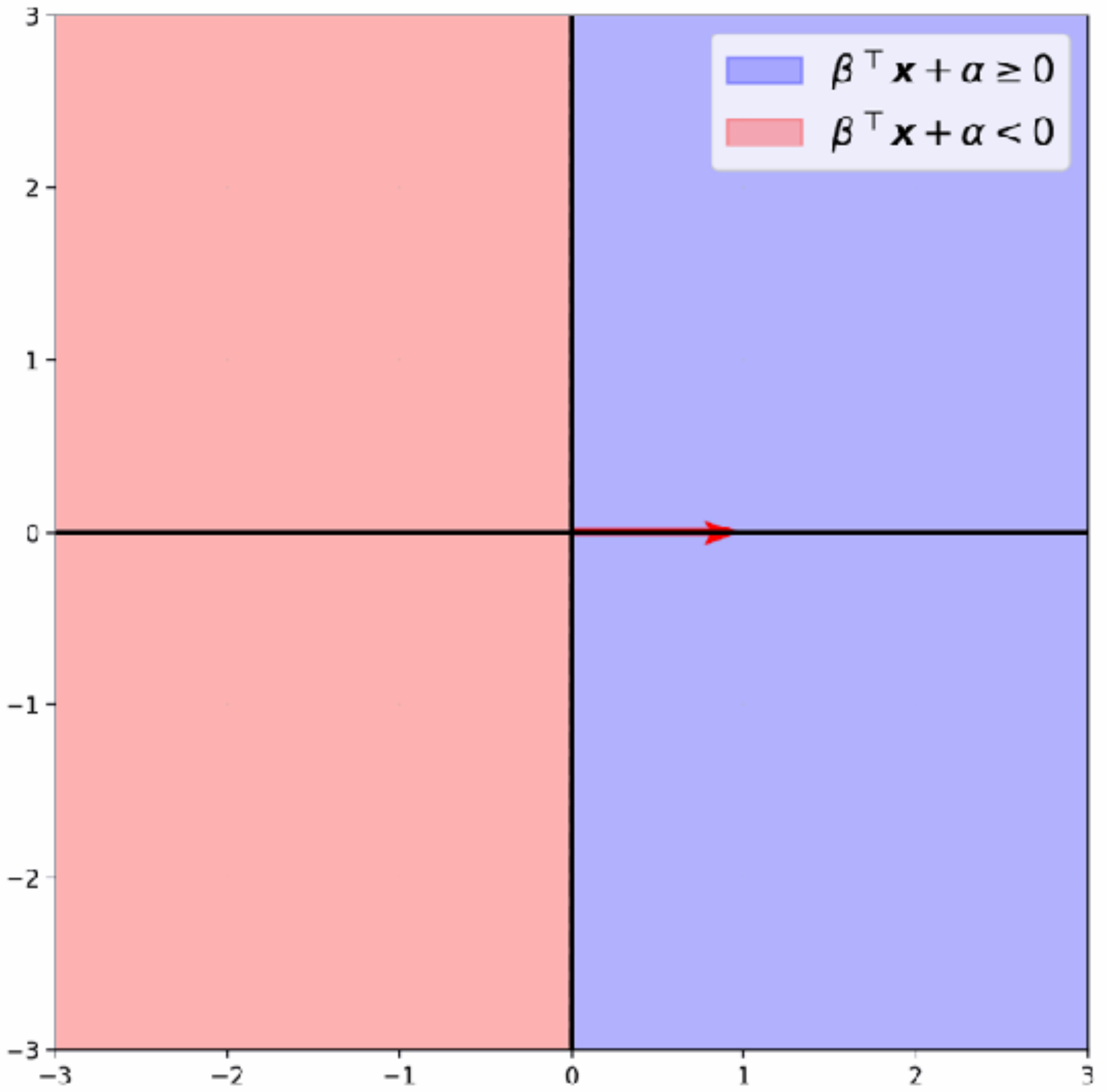
et si $c = 1$? ou $c = -1$?

INSEA

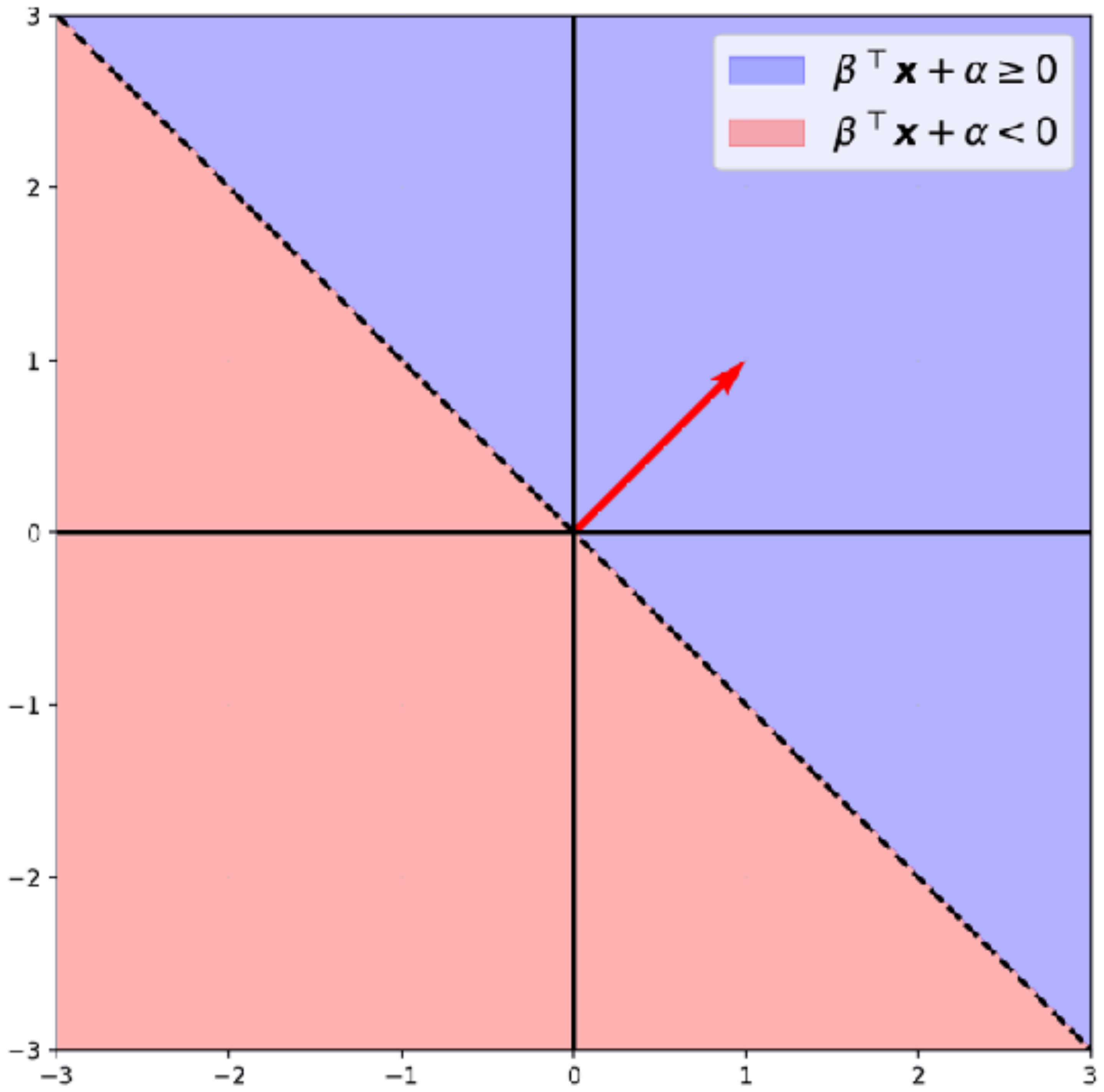


Comment change la fonction de prédiction $f : \mathbb{1}_{\{\alpha + \beta^\top \mathbf{x} \geq 0\}}$ en fonction de α et β ?

$\alpha = 0$, β varie:

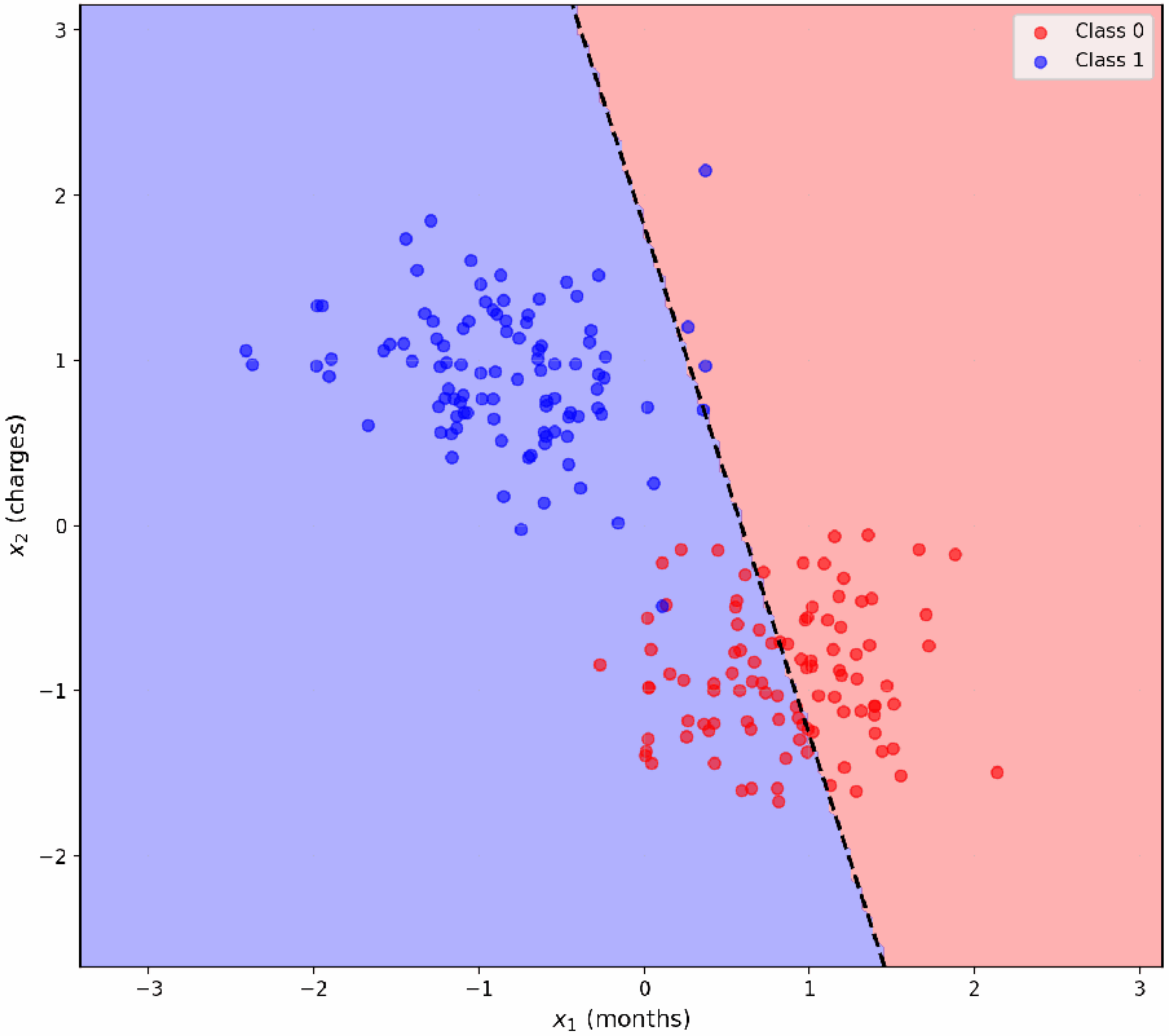


α varie, $\beta = [1, 1]$:

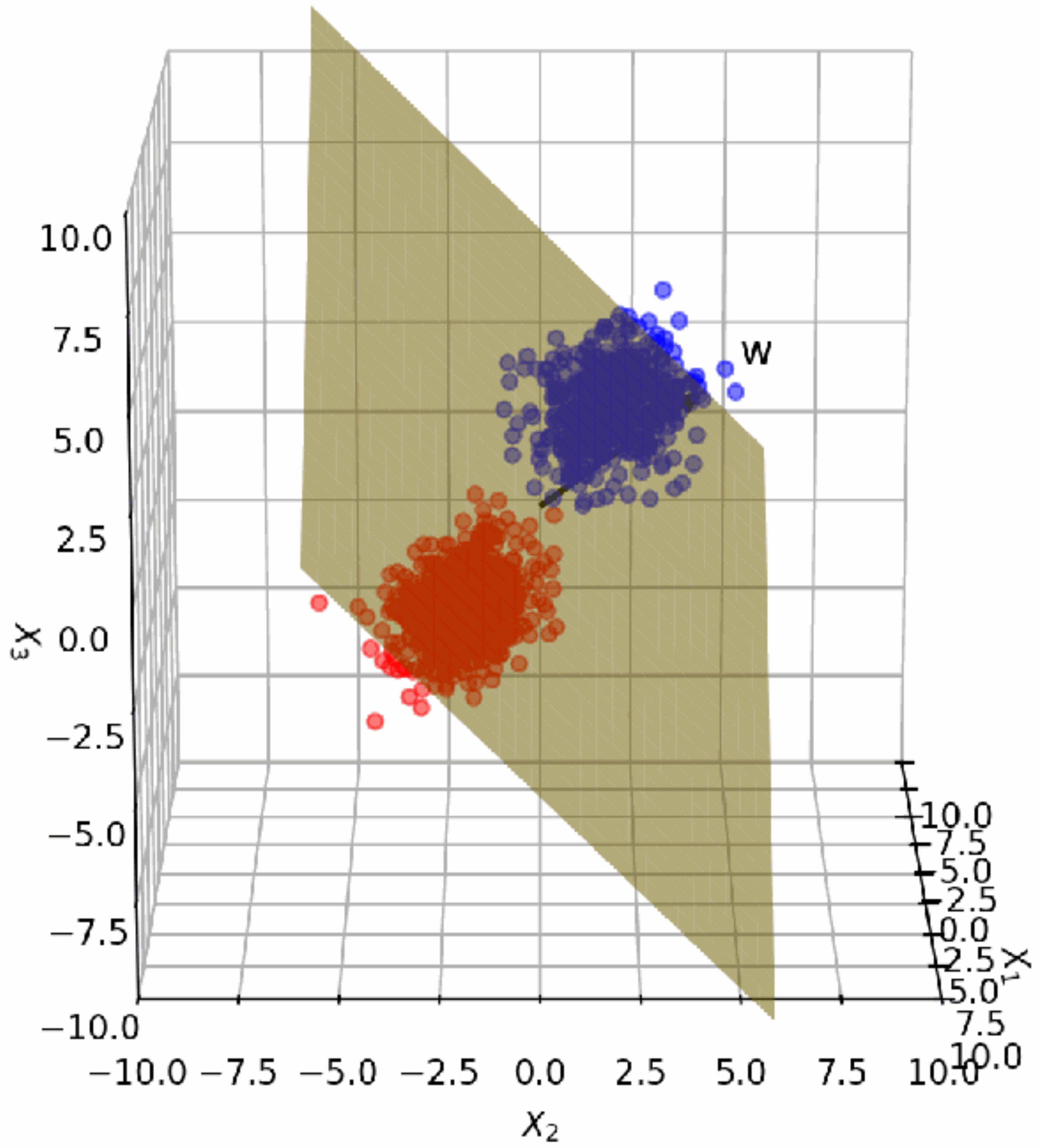


Comment change la fonction de prédiction $f : \mathbb{1}_{\{\alpha + \beta^\top \mathbf{x} \geq 0\}}$ en fonction de α et β ?

$$\min_{\alpha\in\mathbb{R},\beta\in\mathbb{R}^2} \sum_{i=1}^{n} (\mathbb{1}_{\{\alpha+\beta^\top \mathbf{x}_i \geq 0\}} - y_i)^2$$



Et si on utilise trois variables:



$$g(\mathbf{x}) = \alpha + \beta_1 \mathbf{x}^1 + \beta_2 \mathbf{x}^2 + \beta_3 \mathbf{x}^3$$

$$g(\mathbf{x}) = \alpha + \beta^\top \mathbf{x}$$

Que forment les $\mathbf{x}$ tels que $\{g(\mathbf{x}) = 0\}$?

En dimension d: $g(\mathbf{x}) = \alpha + \beta^\top \mathbf{x}, \quad \beta \in \mathbb{R}^d$

Que forment les $\mathbf{x}$ tels que $\{g(\mathbf{x}) = 0\}$?

Un espace de dimension d-1: un hyperplan

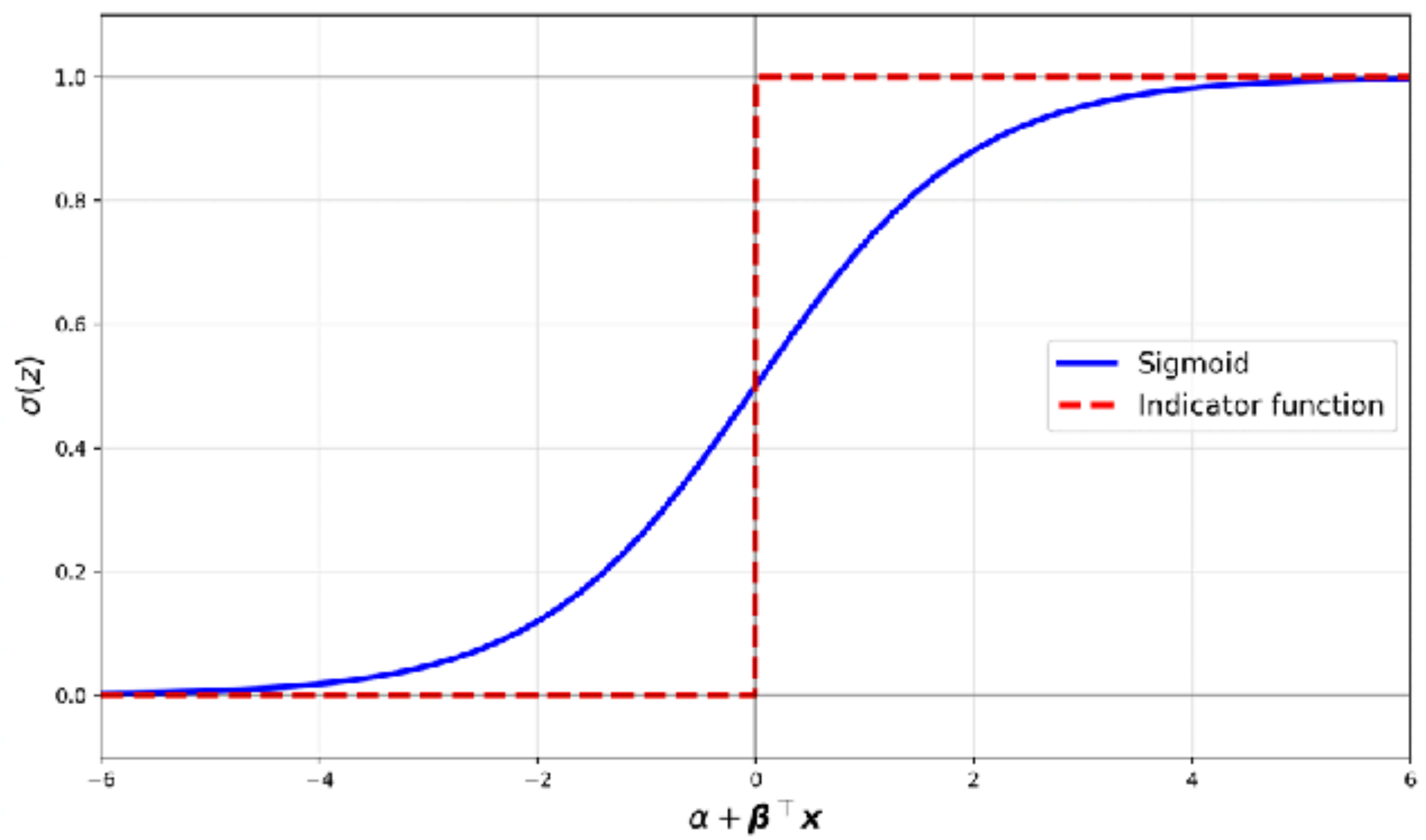
$$\min_{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^d} \sum_{i=1}^{n} (\mathbb{1}_{\{\alpha + \beta^\top \mathbf{x}_i \geq 0\}} - y_i)^2$$

Fonction non différentiable (discontinue même) difficile à optimiser

Au lieu de prendre le signe, transformer les scores $\alpha + \beta^\top \mathbf{x}_i$ vers $[0, 1]$ et modéliser des probabilités

sigmoid: $t \mapsto \frac{1}{1+e^{-t}}$ (logistique)

$$p_i \stackrel{\text{def}}{=} \mathbb{P}(y_i = 1 | \mathbf{x}_i) = \text{sigmoid}(\alpha + \beta^\top \mathbf{x}_i)$$



On peut comparer les p_i avec les y_i avec la *cross-entropy*:

$$\min_{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^d} - \sum_{i=1}^{n} y_i \log(p_i) + (1 - y_i) \log(1 - p_i)$$

On a donc une fonction de prédiction: $f^\star(\mathbf{x}_i) = 1 \Leftrightarrow \text{sigmoid}(\alpha^\star + \beta^{\star\top} \mathbf{x}_i) \geq \frac{1}{2}$

Modèle de régression logistique

$$p_i \stackrel{\text{def}}{=} \mathbb{P}(y_i = 1|\mathbf{x}_i) = \text{sigmoid}(\alpha + \beta^\top \mathbf{x}_i)$$

$$\min_{\alpha\in\mathbb{R},\beta\in\mathbb{R}^d} -\sum_{i=1}^{n} y_i \log(p_i) + (1-y_i)\log(1-p_i)$$

Optimisation faite sur $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$

"Training" data

$\mathbf{x}_1$	y_1
$\vdots$	$\vdots$
$\mathbf{x}_n$	y_n

⟶ "Training" ⟶ "Learned" f* ⟶

predictions	true labels
$f^\star(\mathbf{x}_1)$	y_1
$\vdots$	$\vdots$
$f^\star(\mathbf{x}_n)$	y_n

⟶ "Train" error

Est-ce une bonne manière d'évaluation la performance du modèle ?

L'erreur de prédiction sur ces données est optimisée: elle est forcément petite.

Il faut évaluer la performance du modèle sur des données nouvelles non vues à l'entraînement: "Test data"

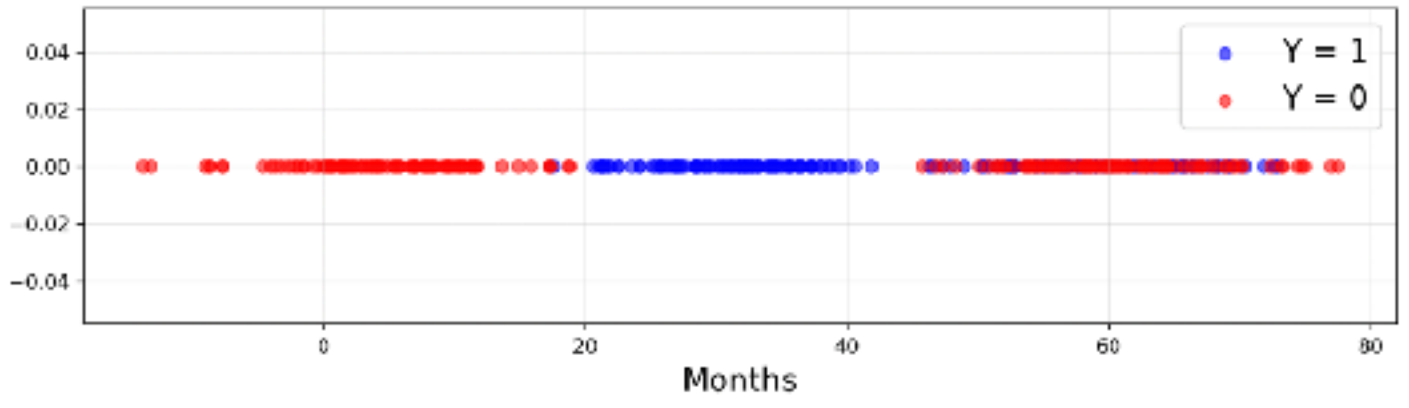
predictions	true labels
$f^\star(\mathbf{x}'_1)$	y'_1
$\vdots$	$\vdots$
$f^\star(\mathbf{x}'_m)$	y'_m

⟶ "Test" error

Train error goes to zero

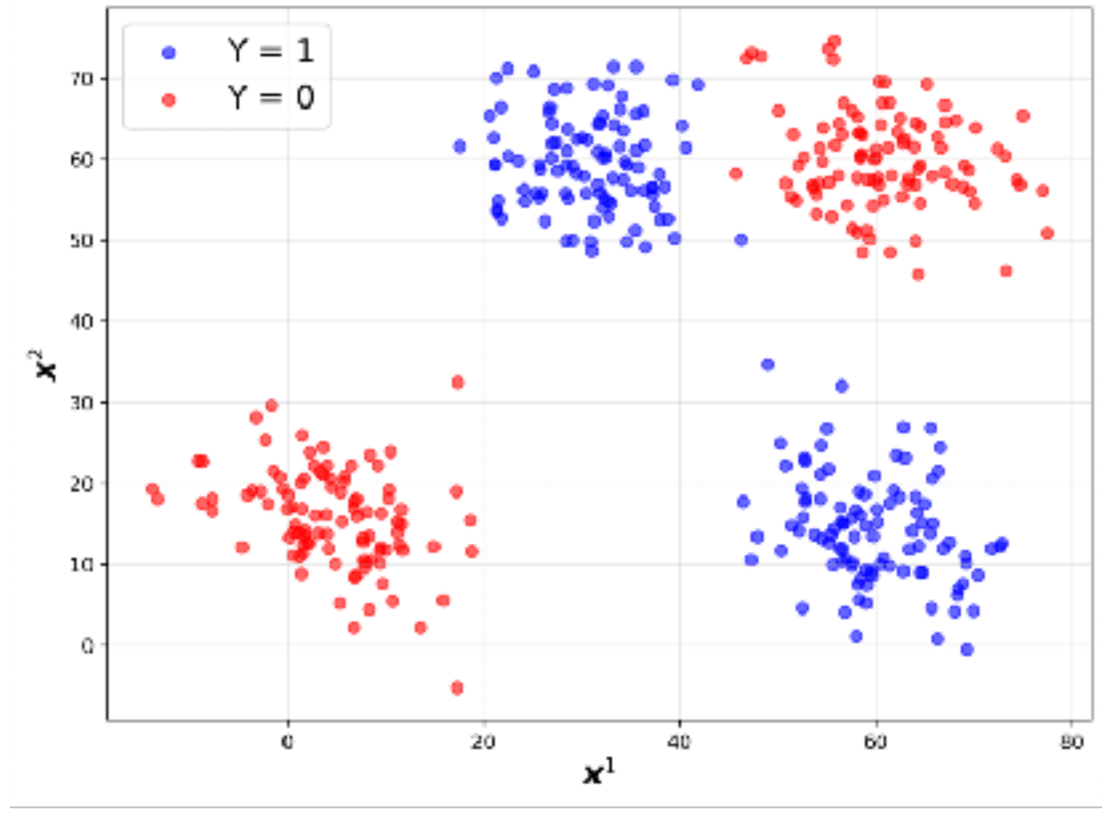
Peut-on séparer les classes avec une séparation linéaire dans ces cas ?

d = 1



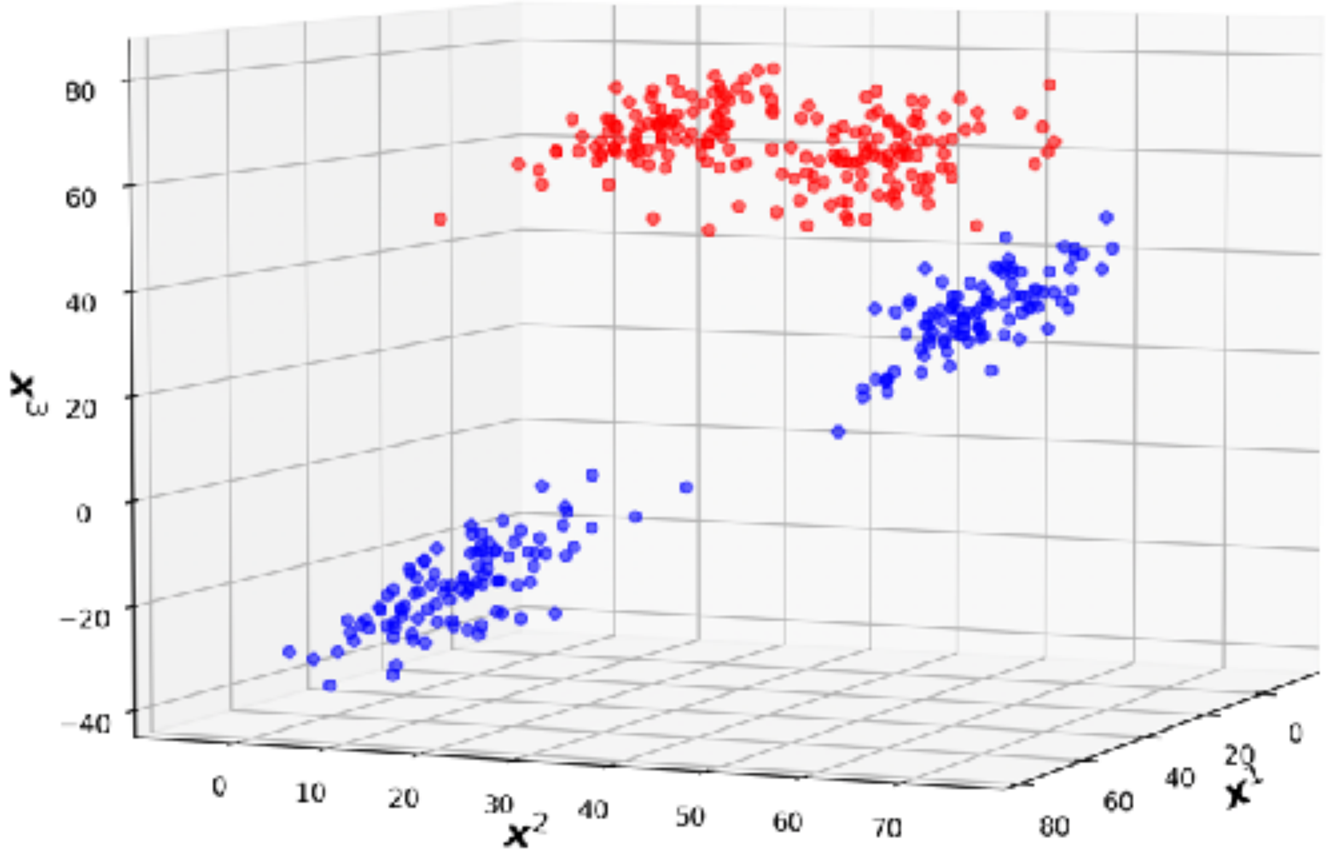
Non !

d = 2



Non !

d = 3



Oui !

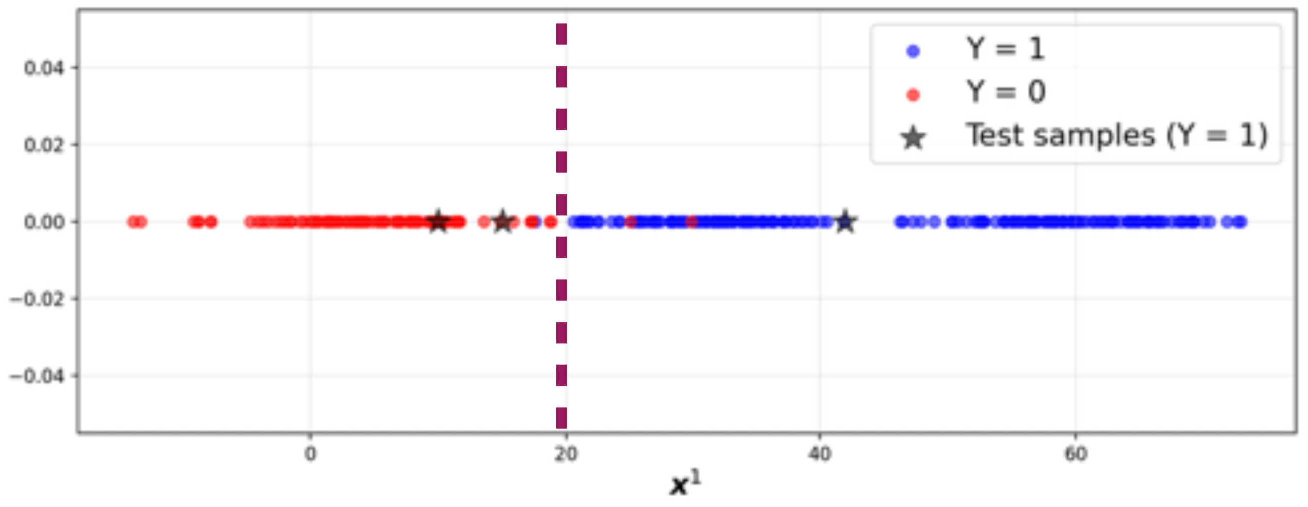
d + 1 représente le nombre de paramètres à estimer: plus d est grand, plus le modèle est riche, complexe.

Comment évolue l'erreur sur le train au fur-et-à mesure que la dimension d augmente ?

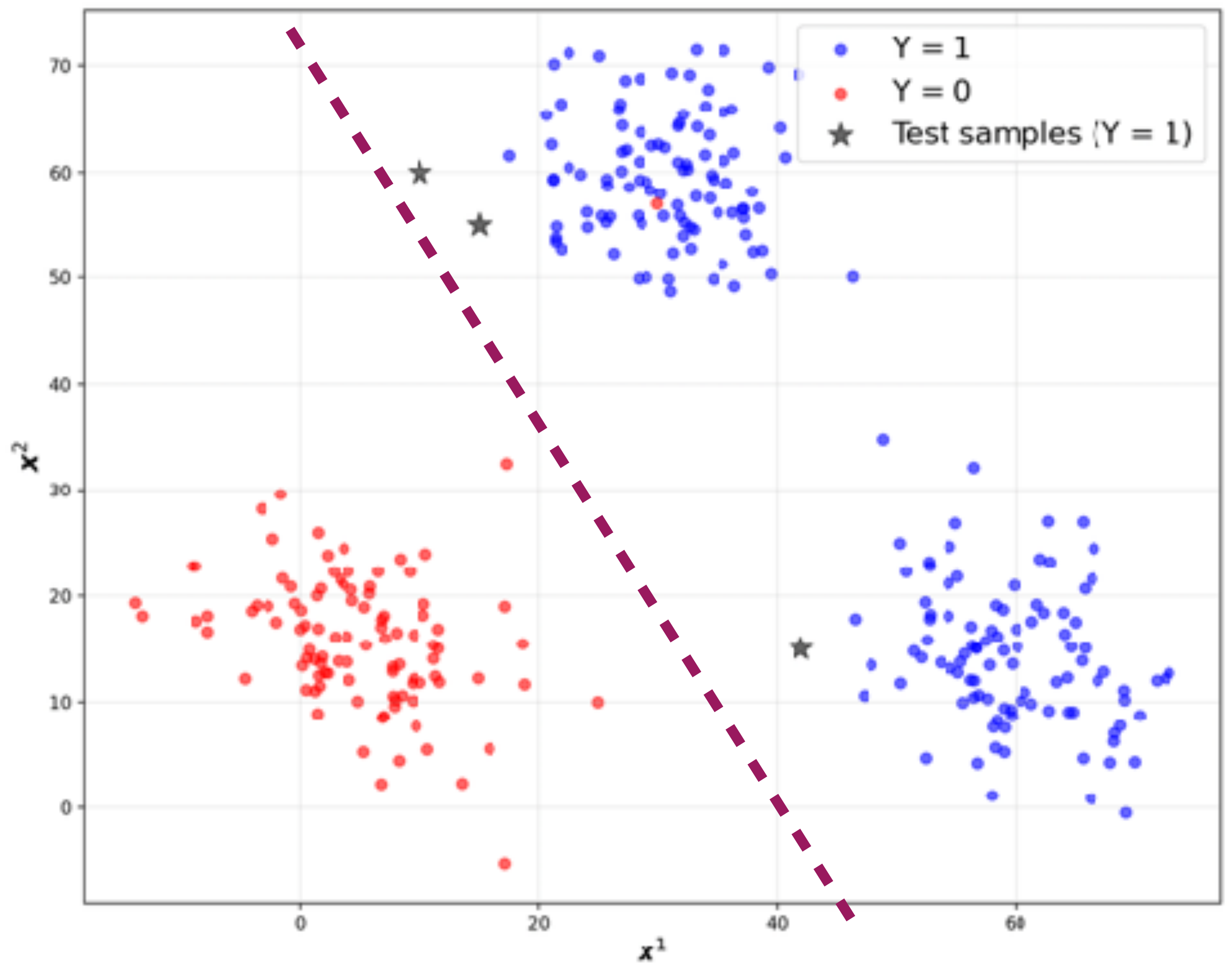
Test error makes a U shape

Quelle est la meilleure séparation linéaire sur ces données ?

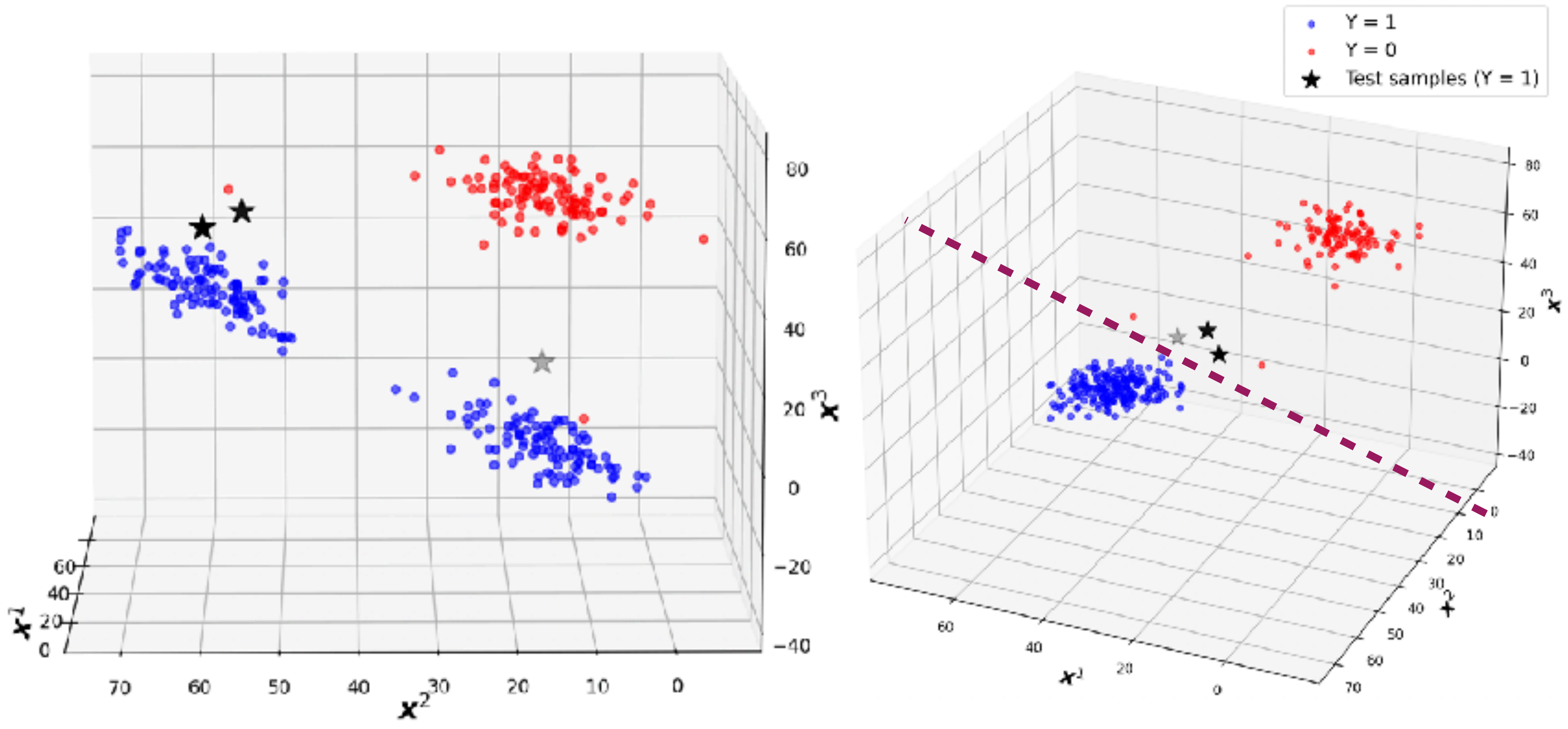
d = 1



d = 2



d = 3



Calculer l'erreur de train et de test.

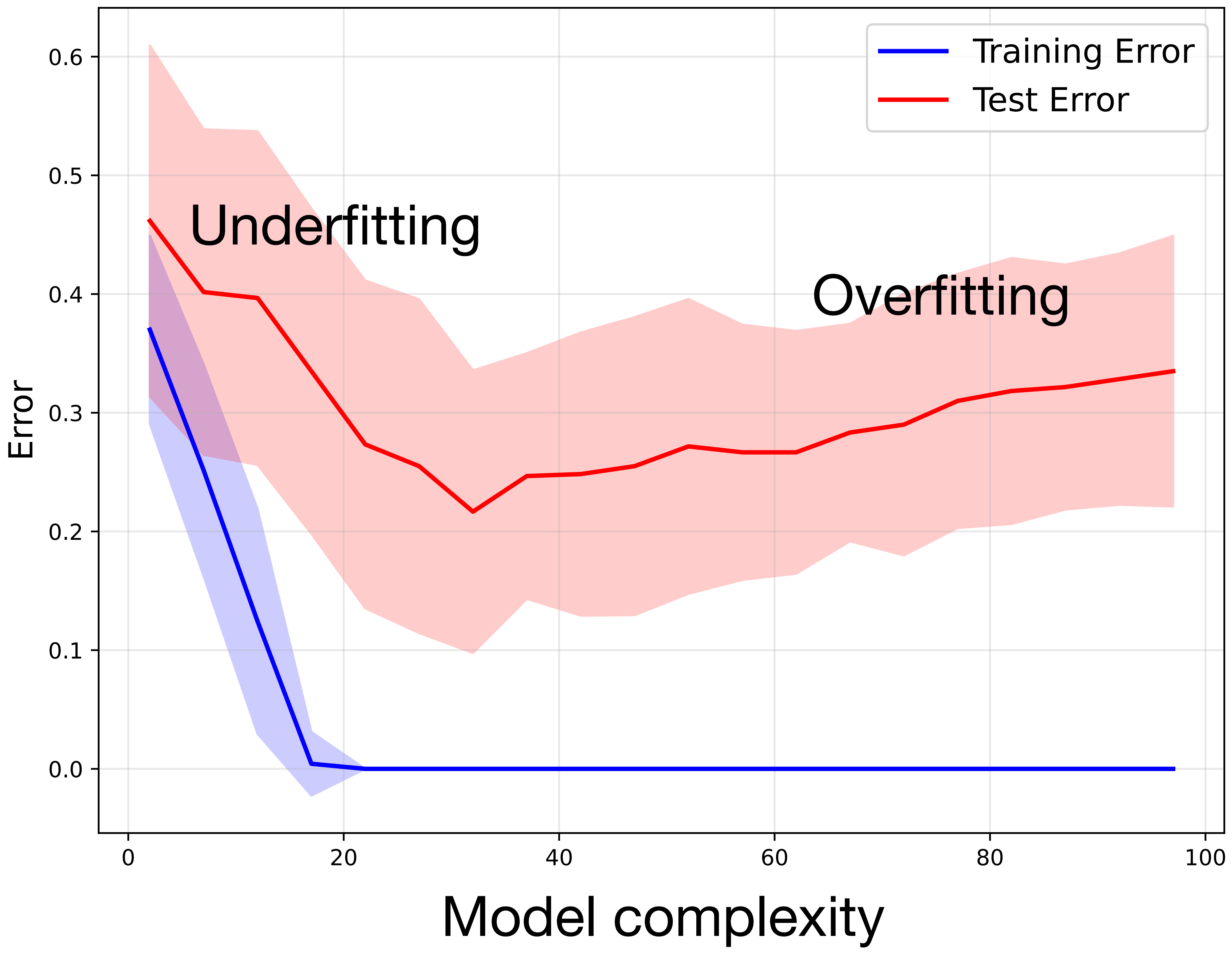
d = 3 donne la meilleure erreur de train = 0

d = 2 donne la meilleure erreur de test = 0

"La meilleure" séparation linéaire sur le train n'est pas la meilleure sur le test: elle est biaisée par les outliers

Une grande dimension peut causer **l'overfitting**

“Bias-Variance” tradeoff



Underfitting correspond à:

Grand biais ou grande variance ?

Variance nulle = prédiction constante
= underfitting