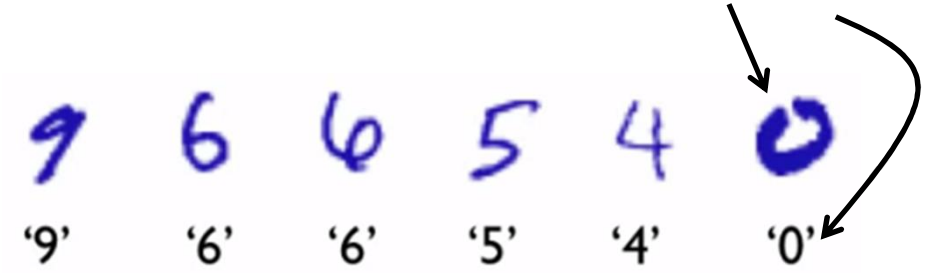
Apprentissage supervisé

Principe de l'apprentissage supervisé

Les algorithmes d'apprentissage supervisé procèdent comme suit:

- On fournit à l'algorithme des **données d'entrainement**:

$$\mathcal{D} = \left\{\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \cdots, \left(x^{(N)}, y^{(N)}\right)\right\}$$

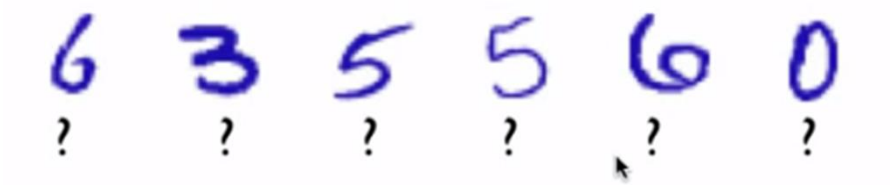


- On appelle $x^{(i)}$ **l'entrée** et $y^{(i)}$ **la cible** du *i*-ième exemple.

- Un élément de $\mathcal{D}$ est appelé **exemple d'apprentissage ou une instance de données**.

Principe de l'apprentissage supervisé

• L'algorithme retourne un «programme» capable de **se généraliser** à de nouvelles données:

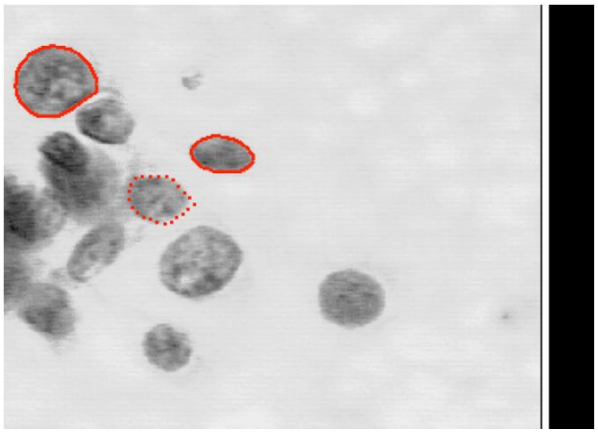


• On note le «programme» généré par l'algorithme d'apprentissage $f(x)$.

• On appelle $f(x)$ un **modèle** ou une **hypothèse**.

• On utilise souvent un ensemble de test $\mathcal{D}_{\text{test}}$ pour mesurer la performance du modèle $f(x)$.

Exemple motivation



• Des cellules cancéreuses sont prises de tumeurs de cancer du sein avant la chirurgie et elles sont photographiées.

• Les tumeurs sont excisées.

• Les patients sont suivis pour voir s'il y a récurrence du cancer. On mesure le temps avant que la récurrence du cancer ou que le patient est déclaré sans la maladie.

Exemple motivation

- On utilise 30 caractéristiques par tumeur.
- Deux variables sont prédites:

  ✓ **Résultat** (**R**: récurrence, **N**: non-récurrence).
  ✓ **Temps** (Jusqu'a récurrence, pour R, et en santé, pour N).

- L'ensemble de données est représenté par une matrice $\mathbf{X}$:

$\mathbf{X}$

tumor size	texture	perimeter	. . .	outcome	time
18.02	27.6	117.5		N	31
17.99	10.38	122.8		N	61
20.29	14.34	135.1		R	27
. . .					

Exemple motivation

- Les colonnes sont appelées **variables d'entrée**, **attributs** ou **caractéristiques**.

- Le résultat et le temps (que nous essayons de prédire) sont appelés les **variables résultats** ou **les cibles**.

- Une ligne du tableau est appelée un **exemple d'entrainement** ou **instance**.

- Le tableau en entier est appelé **l'ensemble d'entrainement**.

Types de prédiction

tumor size	texture	perimeter	. . .	outcome	time
18.02	27.6	117.5		N	31
17.99	10.38	122.8		N	61
20.29	14.34	135.1		R	27
. . .					

- Le problème de prédiction des **résultats** de la maladie est appelé une **classification**.

- Le problème de prédiction **du temps** est appelé **régression**.

Types de prédiction (suite)

- Un exemple d'entrainement à la forme $\left(x^{(i)}, y^{(i)}\right)$, où:

$$x^{(i)} = \left({x_1}^{(i)}, {x_2}^{(i)}, \cdots, {x_D}^{(i)}\right)$$

- D est le nombre d'attributs (dans notre cas 30).
- L'ensemble d'entrée $\mathcal{D}$ contient N exemples.
- On dénote par $\boldsymbol{\mathcal{X}}$ l'espace des variables d'entrées (ex. $\mathbb{R}^D$)
- On dénote par $\boldsymbol{\mathcal{Y}}$ l'espace des variables de sortie (ex. $\mathbb{R}$)

Apprentissage supervisé

- Ayant un ensemble de données $\mathcal{D} \subset \mathcal{X} \times \mathcal{Y}$, trouver une fonction :

$$f: \mathcal{X} \rightarrow \mathcal{Y}$$

- Telle que f est un bon prédicteur de la valeur de y. La fonction f est appelée **une hypothèse**.

- Les problèmes sont classés par type de domaine de sortie:

☞ Si $\mathcal{Y} = \mathbb{R}$, on parle alors de **régression**.
☞ Si $\mathcal{Y}$ est un ensemble discret fini, on parle de **classification**.
☞ Si $\mathcal{Y}$ a 2 éléments, on parle de **classification binaire.**

Apprentissage supervisé

Les étapes pour résoudre un problème d'apprentissage supervisé.

A. Définir les variables d'entrée et de sortie.
B. Définir le codage des variables d'entrée et de sortie (X et Y).
C. Choisir la classe d'hypothèse/représentations H.
D. Trouver l'hypothèse f **la plus optimale** pour la prédication.

Apprentissage supervisé pour la régression

Rappels d'algèbre

Matrices

- **Une matrice** est un tableau 2D composés d'éléments rangés en **lignes** et en **colonnes**.

- Pour une matrice **A** à m lignes et n colonnes, on dit qu'elle est **de dimension** m x n, et elle sera représentée comme suit:

$$\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

- Dans notre cours, les matrices sont dénotées par des lettres majuscules en gras (ex. **A**, **B**,…).

Rappels d'algèbre

- **Un vecteur** est une matrice à une colonne**.**

$$\mathbf{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix}$$

- **Une matrice carrée** possède le même nombre de lignes et de colonnes ($m=n$).

- Une **matrice identité** $\mathbf{I}$ est une matrice carrée qui a tous ses éléments de la diagonale égaux à 1 et le reste des éléments égaux à 0.

$$\mathbf{I} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}$$

Rappels d'algèbre

- **Le transposé** d'une matrice $\mathbf{A}$ est dénoté par $\mathbf{A}^T$.

$$\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \quad \Rightarrow \quad \mathbf{A}^T = \begin{bmatrix} a_{11} & \cdots & a_{m1} \\ \vdots & \ddots & \vdots \\ a_{1n} & \cdots & a_{mn} \end{bmatrix}$$

- On a alors les propriétés:

$$\mathbf{I}^T = \mathbf{I}$$

$$(\mathbf{A}^T)^T = \mathbf{A}$$

Rappels d'algèbre

- **La somme** (**soustraction**) de 2 matrices $\mathbf{A}$ et $\mathbf{B}$ ne peuvent se faire que si les 2 matrices ont le même nombre d'éléments.

$$\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{m1} & \cdots & b_{mn} \end{bmatrix}$$

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} + b_{11} & \cdots & a_{1n} + b_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & \cdots & a_{mn} + b_{mn} \end{bmatrix}$$

$$\mathbf{A} - \mathbf{B} = \begin{bmatrix} a_{11} - b_{11} & \cdots & a_{1n} - b_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} - b_{m1} & \cdots & a_{mn} - b_{mn} \end{bmatrix}$$

Rappels d'algèbre

- **La multiplication** de 2 matrices $\mathbf{A}$ et $\mathbf{B}$, $\mathbf{A} \cdot \mathbf{B}$, ne peut se faire que si la matrice $\mathbf{A}$ **a le même nombre de colonnes que le nombre de lignes** de la matrice $\mathbf{B}$.

$$\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} b_{11} & \cdots & b_{1r} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{nr} \end{bmatrix}$$

$$\mathbf{C} = \mathbf{A} \cdot \mathbf{B} = \begin{bmatrix} c_{11} & \cdots & c_{1r} \\ \vdots & \ddots & \vdots \\ c_{m1} & \cdots & a_{mr} \end{bmatrix} \quad \text{où} \quad c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

Rappels d'algèbre

Quelques Propriétés

$$\mathbf{A} \cdot \mathbf{I} = \mathbf{I} \cdot \mathbf{A} = \mathbf{A}$$ (**I**: matrice identité)

$$\mathbf{A} \cdot \mathbf{B} \neq \mathbf{B} \cdot \mathbf{A}$$ (Non commutativité)

$$\mathbf{A} \cdot (\mathbf{B} \cdot \mathbf{C}) = (\mathbf{A} \cdot \mathbf{B}) \cdot \mathbf{C}$$ (Associativité)

$$\mathbf{A} \cdot (\mathbf{B} + \mathbf{C}) = (\mathbf{A} \cdot \mathbf{B}) + (\mathbf{A} \cdot \mathbf{C})$$ (Distributivité)

$$(\mathbf{A} \cdot \mathbf{B})^T = \mathbf{B}^T \cdot \mathbf{A}^T$$

$$(\mathbf{A} \cdot \mathbf{B} \cdot \mathbf{C})^T = \mathbf{C}^T \cdot \mathbf{B}^T \cdot \mathbf{A}^T$$

Rappels d'algèbre

Déterminant

- Chaque matrice carrée $\mathbf{A}$ possède **un déterminant** dénoté par $\mathbf{det(A)}$ ou $\mathbf{|A|}$.

- Le déterminant d'une matrice 1x1 [a] est a.

- Le déterminant d'une matrice 2x2: $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ est:

$$|\mathbf{A}| = ad - cb$$

- Pour les matrices de plus hautes dimensions, on utilisera une procédure récursive pour calculer leurs déterminants.

Rappels d'algèbre

- **Un mineur** d'une matrice $\mathbf{A}$ est le déterminant d'une sous-matrice carrée de $\mathbf{A}$.

- **Le cofacteur** C_{ij} d'une matrice carrée $\mathbf{A}$ à une position $\mathbf{(i, j)}$ est le scalaire obtenu en multipliant $\mathbf{(-1)^{i+j}}$ par le mineur obtenu par la suppression de la ligne $\mathbf{i}$ et la colonne $\mathbf{j}$ de $\mathbf{A}$.

Ex.

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

$$C_{11} = (-1)^{1+1}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} = a_{22}a_{33} - a_{32}a_{23}$$

$$C_{21} = (-1)^{2+1}\begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} = a_{32}a_{13} - a_{12}a_{33}$$

Rappels d'algèbre

Pour calculer le déterminant d'une matrice carrée $\mathbf{A}$ de taille n x n:

- Prendre une ligne ou une colonne de la matrice.
- Pour chaque élément a_{ij} de la ligne (ou de la colonne), calculer son **cofacteur** C_{ij}.
- Multiplier chaque élément a_{ij} avec son cofacteur C_{ij} et faire la somme des multiplications.

$$|\mathbf{A}| = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} C_{ij}$$

Rappels d'algèbre

- **L'inverse** d'une matrice carrée $\mathbf{A}$, s'il existe, est dénoté par $\mathbf{A}^{-1}$. il est défini comme suit:

$$\mathbf{A}^{-1} \cdot \mathbf{A} = \mathbf{A} \cdot \mathbf{A}^{-1} = \mathbf{I}$$

- La matrice carrée $\mathbf{A}$ possède un inverse si $|\mathbf{A}| = 0$.
- Lorsque $|\mathbf{A}| = 0$, on dit que la matrice est **singulière**.
- Une matrice est **singulière** si ses vecteurs sont **linéairement dépendants**.
- Si $\mathbf{A}$ est **non-singulière**, alors $|\mathbf{A}^{-1}| = 1/|\mathbf{A}|$.

Rappels d'algèbre

Résolution d'une système d'équations

Exemple: soit le système de 2 équations: $\begin{cases} x+3y=1 \\ 2x+5y=3 \end{cases}$

On peut réécrire ce système sous forme matricielle comme suit:

$$\mathbf{A}=\begin{pmatrix}1 & 3\\ 2 & 5\end{pmatrix} \quad \mathbf{b}=\begin{pmatrix}1\\ 3\end{pmatrix} \quad \mathbf{x}=\begin{pmatrix}x\\ y\end{pmatrix}$$

D'où: $\mathbf{Ax}=\mathbf{b}$

A est inversible, alors: $\mathbf{A^{-1}Ax}=\mathbf{A^{-1}b} \Rightarrow \boxed{\mathbf{x}=\mathbf{A^{-1}b}}$

Rappel sur les dérivées

- Soit une fonction $h(u_1, u_2, \dots, u_D)$: $\mathbb{R}^D \rightarrow \mathbb{R}$ (pour nous, c'est une fonction d'erreur à minimiser).

- La dérivée partielle de h par rapport à u_i est définie par:

$$\frac{\partial h}{\partial u_d} h(u_1, u_2, \dots, u_D): \mathbb{R}^D \rightarrow \mathbb{R}$$

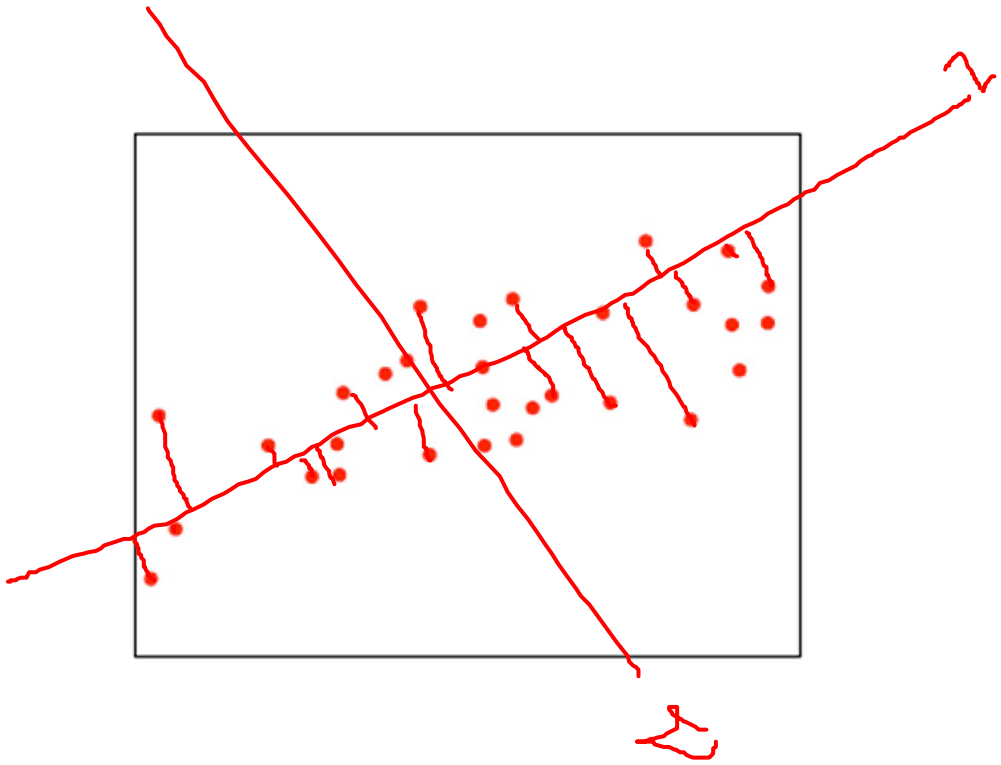
- Elle est la dérivée de h par rapport à l'axe u_i en gardant les autres variables fixes (comme des constantes).

- Le **gradient** de $\nabla h(u_1, u_2, \dots, u_D)$ est le vecteur défini par:

$$\nabla h(u_1, u_2, \dots, u_D) = \left(\frac{\partial h}{\partial u_1}, \frac{\partial h}{\partial u_2}, \cdots, \frac{\partial h}{\partial u_D} \right)$$

Principe de la régression

Exemple: (régression linéaire**)**



Espace de données

x	y
0.86	2.49
0.09	0.83
-0.85	-0.25
0.87	3.10
-0.44	0.87
-0.43	0.02
-1.10	-0.12
0.40	1.81
-0.96	-0.83
0.17	0.43

Échantillon

Régression linéaire

Exemple: supposons une hypothèse **de régression linéaire.**

$$y = f(x) = w_0 + w_1 x_1 + \cdots$$

où $x = (x_1, x_2, \ldots, x_D)$.

- Les w_d sont appelés des **paramètres** ou des **poids**.
- Pour simplifier la notation, ajouter un attribut $x_0 = 1$.

$$y = f(x) = \sum_{d=0}^{D} w_d x_d = \tilde{\mathbf{w}}^T \tilde{x}$$

où $\tilde{\mathbf{w}}$ et $\tilde{x}$ sont des vecteurs de dimension $D + 1$.

Régression linéaire

Comment prendre les paramètres $\widetilde{\mathbf{w}}$?

- $\widetilde{\mathbf{w}}$ doit rendre $f(x)$ très proche des valeurs des y.

- On doit alors définir une **fonction d'erreur** ou **de perte** pour mesurer combien notre prédiction est loin des «vraies» valeurs.

- On prend $\widetilde{\mathbf{w}}$ qui minimise la fonction d'erreur.

Exemple: erreur des moindres carrés.

Erreur des moindres carrés

Principe

- Essayer de rendre $f(\mathbf{x})$ très proche des valeurs des y dans tous les exemples d'apprentissage dans $\mathcal{D}$.

- Définir une fonction d'erreur par la somme:

$$E(\tilde{\mathbf{w}}) = \frac{1}{2}\sum_{i=1}^{N}\left(f\left(x^{(i)}\right) - y^{(i)}\right)^2$$

- Choisir $\tilde{\mathbf{w}}$ qui minimise la fonction d'erreur $E(\tilde{\mathbf{w}})$?.

Erreur des moindres carrés

Sur notre exemple (**Un peu d'algèbre!)**

$$\frac{\partial E(\tilde{\mathbf{w}})}{\partial w_d} = \frac{\partial}{\partial w_d}\left(\frac{1}{2}\sum_{i=1}^{N}\left(f\left(x^{(i)}\right) - y^{(i)}\right)^2\right)$$

$$= 2\left(\frac{1}{2}\sum_{i=1}^{N}\left(f\left(x^{(i)}\right) - y^{(i)}\right)\frac{\partial}{\partial w_d}\left(f\left(x^{(i)}\right) - y^{(i)}\right)\right)$$

$$= \sum_{i=1}^{N}\left(f\left(x^{(i)}\right) - y^{(i)}\right)\frac{\partial}{\partial w_d}\left(\sum_{d=0}^{D} w_d x_d^{(i)} - y^{(i)}\right)$$

$$= \sum_{i=1}^{N}\left(f\left(x^{(i)}\right) - y^{(i)}\right)x_d^{(i)}$$

Notation matricielle

- On a:

$$\nabla E(\tilde{\mathbf{w}}) = \nabla E((\mathbf{X}\tilde{\mathbf{w}} - \mathbf{y})^T(\mathbf{X}\tilde{\mathbf{w}} - \mathbf{y}))$$

$$= 2\mathbf{X}^T(\mathbf{X}\tilde{\mathbf{w}} - \boldsymbol{y}) \text{ (formule de Taylor)}$$

$$= 2\mathbf{X}^T\mathbf{X}\tilde{\mathbf{w}} - 2\mathbf{X}^T\boldsymbol{y}$$

- En posant la dérivée égale à zéro, on obtient:

$$2\mathbf{X}^T\mathbf{X}\tilde{\mathbf{w}} - 2\mathbf{X}^T\boldsymbol{y} = 0 \quad \Rightarrow \quad \mathbf{X}^T\mathbf{X}\tilde{\mathbf{w}} = \mathbf{X}^T\boldsymbol{y}$$

$$\Rightarrow \quad \tilde{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$$

- L'inverse de $\mathbf{X}^T\mathbf{X}$ existe si les colonnes de $\mathbf{X}$ sont linéairement indépendantes.

Exemple

x	y
0.86	2.49
0.09	0.83
-0.85	-0.25
0.87	3.10
-0.44	0.87
-0.43	0.02
-1.10	-0.12
0.40	1.81
-0.96	-0.83
0.17	0.43

$$\mathbf{X} = \begin{bmatrix} 0.86 & 1 \\ 0.09 & 1 \\ -0.85 & 1 \\ 0.87 & 1 \\ -0.44 & 1 \\ -0.43 & 1 \\ -1.10 & 1 \\ 0.40 & 1 \\ -0.96 & 1 \\ 0.17 & 1 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 2.49 \\ 0.83 \\ -0.25 \\ 3.10 \\ 0.87 \\ 0.02 \\ -0.12 \\ 1.81 \\ -0.83 \\ 0.43 \end{bmatrix}$$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 0.86 & 0.09 & -0.85 & 0.87 & -0.44 & -0.43 & -1.10 & 0.40 & -0.96 & 0.17 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} 0.86 & 1 \\ 0.09 & 1 \\ -0.85 & 1 \\ 0.87 & 1 \\ -0.44 & 1 \\ -0.43 & 1 \\ -1.10 & 1 \\ 0.40 & 1 \\ -0.96 & 1 \\ 0.17 & 1 \end{bmatrix} = \begin{bmatrix} 4.95 & -1.39 \\ -1.39 & 10 \end{bmatrix}$$

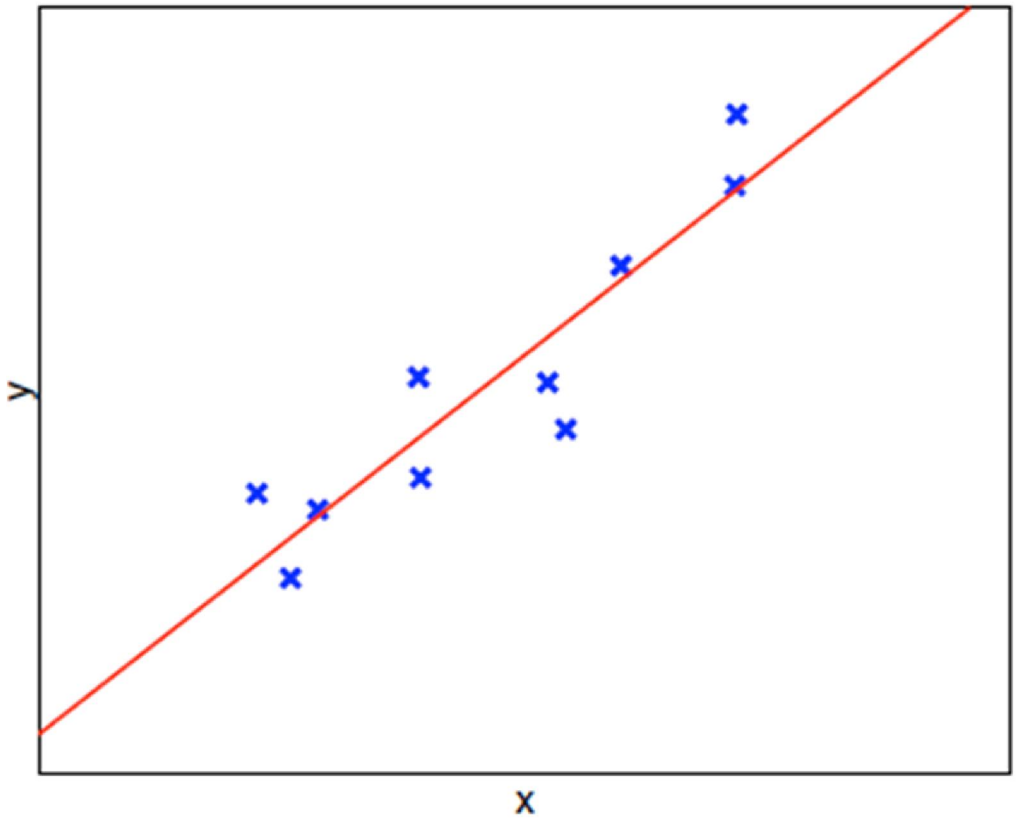
Exemple

$$\mathbf{X}^T\mathbf{y} = \begin{bmatrix} 0.86 & 0.09 & -0.85 & 0.87 & -0.44 & -0.43 & -1.10 & 0.40 & -0.96 & 0.17 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} 2.49 \\ 0.83 \\ -0.25 \\ 3.10 \\ 0.87 \\ 0.02 \\ -0.12 \\ 1.81 \\ -0.83 \\ 0.43 \end{bmatrix} = \begin{bmatrix} 6.49 \\ 8.34 \end{bmatrix}$$

$$\widetilde{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y} = \begin{bmatrix} 4.95 & -1.39 \\ -1.39 & 10 \end{bmatrix}^{-1} \begin{bmatrix} 6.49 \\ 8.34 \end{bmatrix} = \begin{bmatrix} 1.60 \\ 1.05 \end{bmatrix}$$

Exemple

La meilleure ligne est égale alors à: $y = 1.6x + 1.05$



Fonctions d'ordre supérieur

- La régression linéaire est **trop simple** pour les problèmes les plus réalistes, mais elle devrait être **la première chose** que vous essayer pour les sorties à valeur réelles.

- Si $\mathbf{X}^T\mathbf{X}$ n'Est pas inversible:

  ☞ **Transformer les données:** ajouter des termes d'ordre supérieur. Plus généralement, appliquer une transformation des entrées de $\mathcal{X}$ dans une autre espace $\mathcal{X}^*$, puis faire la régression linéaire dans le nouveau espace.

  ☞ **Changer de classe d'hypothèses** $\mathcal{H}$.

Fonctions d'ordre supérieur

- Soit x une variable d'entrée unidimensionnelle $(D = 1)$. Si nous voulons appliquer un polynôme d'ordre supérieur aux données d'apprentissage, on aura:

Exemple: $f(x) = w_0 + w_1 x + w_2 x^2$

- Pour un polynôme d'ordre m, on aura:

$$\mathbf{X} = \begin{bmatrix} x^{(1)^m} & \cdots & x^{(1)^2} & x^{(1)} & 1 \\ x^{(2)^m} & & x^{(2)^2} & x^{(2)} & 1 \\ \vdots & \ddots & & \vdots & \\ x^{(N)^m} & \cdots & x^{(N)^2} & x^{(N)} & 1 \end{bmatrix}$$

- Résoudre le problème: $\mathbf{X}^T\mathbf{w} \approx \boldsymbol{y}$

Fonctions d'ordre supérieur

Pour notre exemple, **une régression quadratique** (m=2) aura la forme:

$$\mathbf{X} = \begin{bmatrix} 0.75 & 0.86 & 1 \\ 0.01 & 0.09 & 1 \\ 0.73 & -0.85 & 1 \\ 0.76 & 0.87 & 1 \\ 0.19 & -0.44 & 1 \\ 0.18 & -0.43 & 1 \\ 1.22 & -1.10 & 1 \\ 0.16 & 0.40 & 1 \\ 0.93 & -0.96 & 1 \\ 0.03 & 0.17 & 1 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 2.49 \\ 0.83 \\ -0.25 \\ 3.10 \\ 0.87 \\ 0.02 \\ -0.12 \\ 1.81 \\ -0.83 \\ 0.43 \end{bmatrix}$$

Fonctions d'ordre supérieur

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 0.75 & 0.01 & 0.73 & 0.76 & 0.19 & 0.18 & 1.22 & 0.16 & 0.93 & 0.03 \\ 0.86 & 0.09 & -0.85 & 0.87 & -0.44 & -0.43 & -1.10 & 0.40 & -0.96 & 0.17 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} 0.75 & 0.86 & 1 \\ 0.01 & 0.09 & 1 \\ 0.73 & -0.85 & 1 \\ 0.76 & 0.87 & 1 \\ 0.19 & -0.44 & 1 \\ 0.18 & -0.43 & 1 \\ 1.22 & -1.10 & 1 \\ 0.16 & 0.40 & 1 \\ 0.93 & -0.96 & 1 \\ 0.03 & 0.17 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 4.11 & -1.64 & 4.95 \\ -1.64 & 4.95 & -1.39 \\ 4.95 & -1.39 & 10 \end{bmatrix}$$

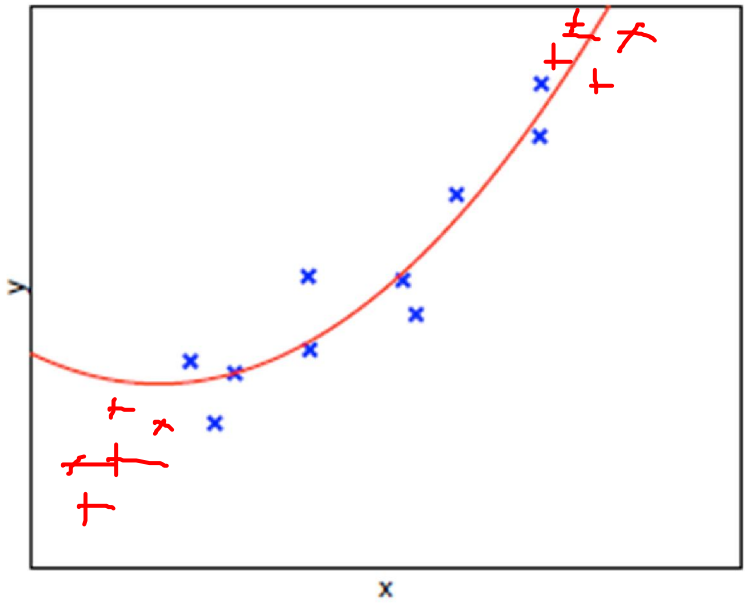
$$\mathbf{X}^T\mathbf{y} = \begin{bmatrix} 0.75 & 0.01 & 0.73 & 0.76 & 0.19 & 0.18 & 1.22 & 0.16 & 0.93 & 0.03 \\ 0.86 & 0.09 & -0.85 & 0.87 & -0.44 & -0.43 & -1.10 & 0.40 & -0.96 & 0.17 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} 2.49 \\ 0.83 \\ -0.25 \\ 3.10 \\ 0.87 \\ 0.02 \\ -0.12 \\ 1.81 \\ -0.83 \\ 0.43 \end{bmatrix}$$

$$= \begin{bmatrix} 3.60 \\ 6.49 \\ 8.34 \end{bmatrix}$$

Fonctions d'ordre supérieur

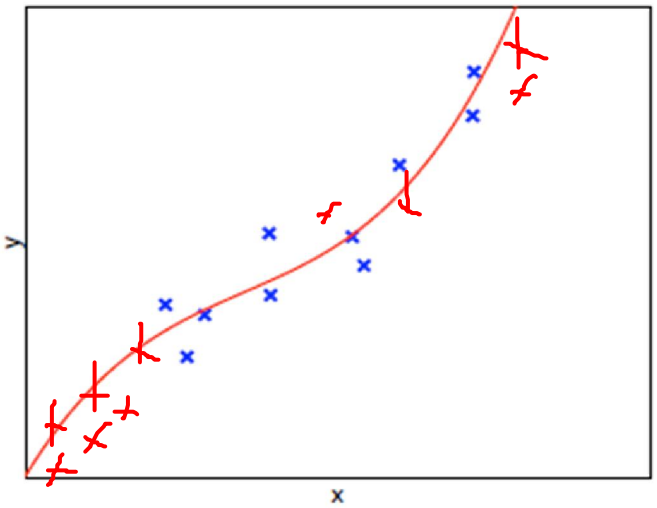
$$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y} = \begin{bmatrix} 4.11 & -1.64 & 4.95 \\ -1.64 & 4.95 & -1.39 \\ 4.95 & -1.39 & 10 \end{bmatrix}^{-1} \begin{bmatrix} 3.60 \\ 6.49 \\ 8.34 \end{bmatrix} = \begin{bmatrix} 0.68 \\ 1.74 \\ 0.73 \end{bmatrix}$$

$$f(x) = 0.73 + 0.68x^2 + 1.74x$$

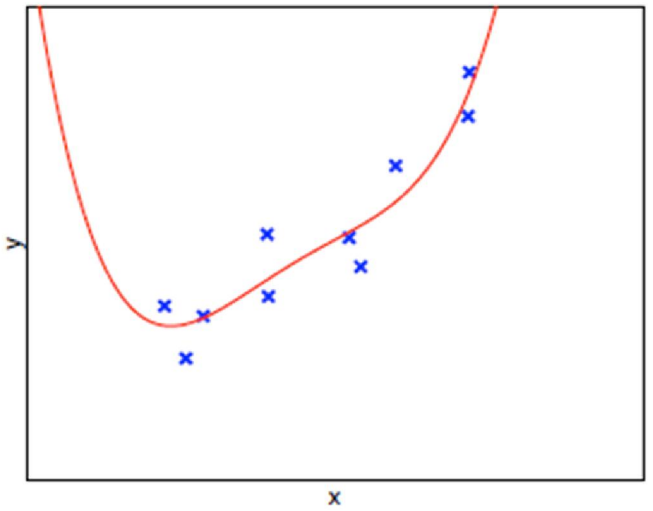


Fonctions d'ordre supérieur

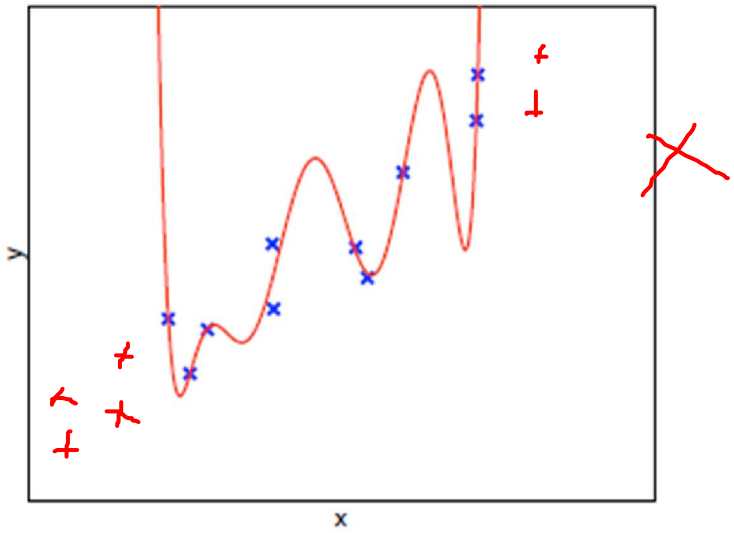
$m = 4$



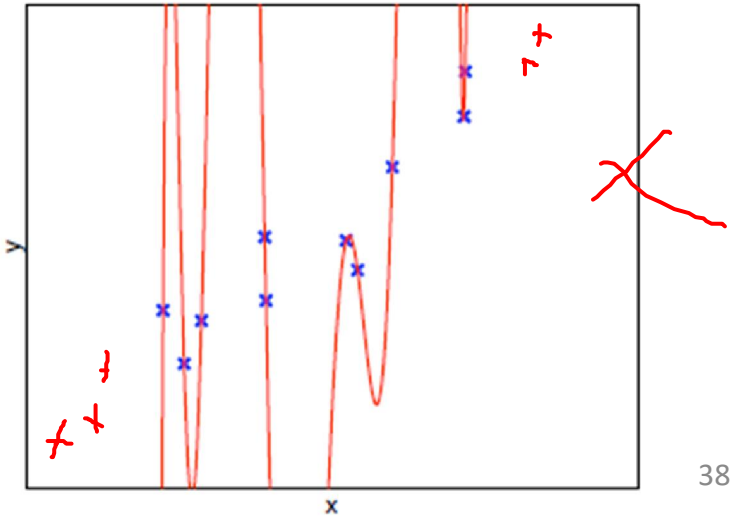
$m = 5$



$m = 8$



$m = 9$



Overfitting

- **Est un important problème pour les techniques d'apprentissage!**

- On peut trouver une hypothèse qui fait une bonne prédiction pour des données d'entrainement, mais qui **ne se généralise** pas bien pour le reste des données.

- Dans le reste du cours, nous verrons des méthodes pour atténuer le problème d'overfitting.

Théorie de la décision pour la classification

Principe

- Soit un problème de classification à K classes $\{C_1,\ldots,C_K\}$.

- Une **fonction discriminante** a le rôle de prendre une entrée x et de lui assigner une classe parmi K classes existantes.

- Soit x un vecteur d'entrée ayant une valeur cible y, et notre but est de prédire y ayant la donnée d'entrée x.

Exemple:

x : image de rayon-X.

y : présence/non-présence d'une certaine maladie (ex. cancer, sclérose, etc.) qui forment les classes C_1 et C_2.

Théorie de la décision

- **La théorie des probabilités** permet de quantifier et manipuler **l'incertitude** dans les expériences aléatoires.

- Elle peut aider aussi à la **prise des décisions** dans des situations impliquant **l'incertitude sur les résultats**.

- On peut choisir par exemple le codage suivant:

$$y^{(i)} = \begin{cases} 1 \quad si \quad x^{(i)} \in C_1 \\ 0 \quad si \quad x^{(i)} \in C_2 \end{cases}$$

Théorie de la décision

- Le problème revient alors à déterminer la probabilité $p(x,C_k)$ qui donnera **une description complète** de la situation.

- Lorsqu'on obtient l'image rayon-X x pour un nouveau malade, on doit décider à quelle classe il appartient.

- La probabilité d'une classe C_k est donnée par $p(C_k|\,x)$. Cette probabilité est formulée par:

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{p(x)}$$

C_1

C_2

$P(C_1|x)$ $\quad$ $P(C_2|x)$

Théorie de la décision

- On peut interpréter $p(C_k)$ comme étant la probabilité **a priori** pour observer la classe C_k.

- Le terme $p(C_k|x)$ correspond à la probabilité **a posteriori.**

- Intuitivement, **l'erreur de classification** est minimisée en assignant x à la classe ayant **la plus grande probabilité a posteriori**.

- La règle minimisant l'erreur de classification va diviser l'espace des données $\mathcal{D}$ en K régions $\{R_1, R_2, \ldots, R_{\mathrm{K}}\}$ appelées **régions de décisions**.

Minimiser l'erreur de classification

- Une **erreur** est produite lorsqu'une donné appartenant à C_1 est assignée à C_2 et vice-versa.

- **La probabilité d'occurrence d'erreur** est donné par:

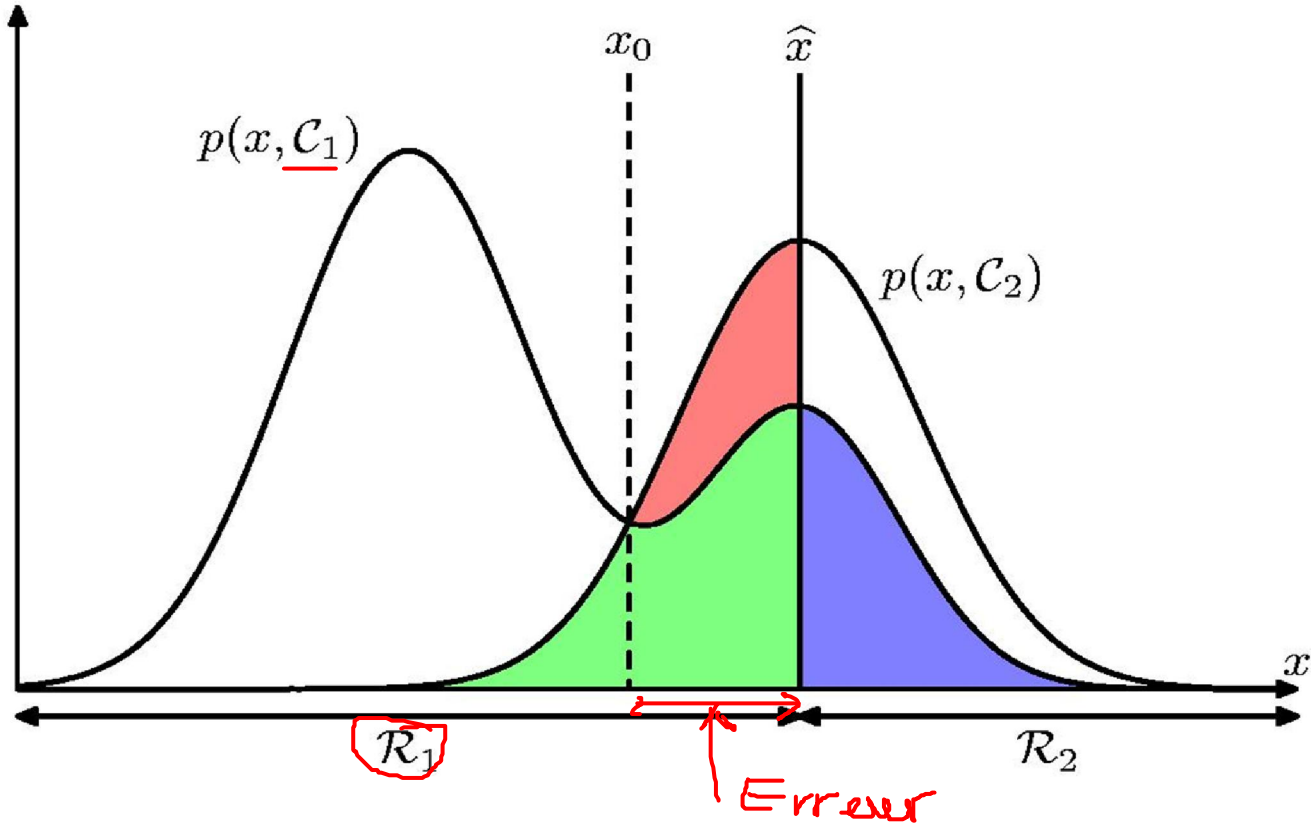
$$p(erreur) = p(x \in R_1|C_2) + p(x \in R_2|C_1)$$

$$= \int_{R_1} p(x, C_2)\, dx + \int_{R_2} p(x, C_1)\, dx$$

- Pour minimiser l'erreur de classification, la règle de décision doit assigner chaque donnée x en minimisant $p(erreur)$.

Minimiser l'erreur de classification

- En ayant $p(x, C_1) = p(C_1|x)\ p(x)$, et $p(x)$ est un facteur commun entre les deux classes, le minimum sera obtenu en assignant la donnée x à la classe ayant le plus grand $p(C_k|x)$.



Modèles linéaires pour la classification

Introduction

- Pour la classification, on possède K classes $\{C_1, C_2, \ldots, C_K\}$. Dans la plupart des scenarios, **les classes sont disjointes**.

- L'espace d'entré est alors divisé en **régions** séparées par des **frontières (ou surface) de décision**.

- Dans cette partie du cours, nous considérons **les modèles linéaire** pour la classification, c.-à-d., la frontière de décision est une fonction linéaire de la variable d'entrée x.

- La frontière linéaire est définie dans l'espace à $(D-1)$ dimensions dans l'espace de données à D dimensions.

Modèle linéaire de classification

- Une **fonction discriminante** a le rôle de prendre une entrée x et de lui assigner une classe parmi K classes existantes.

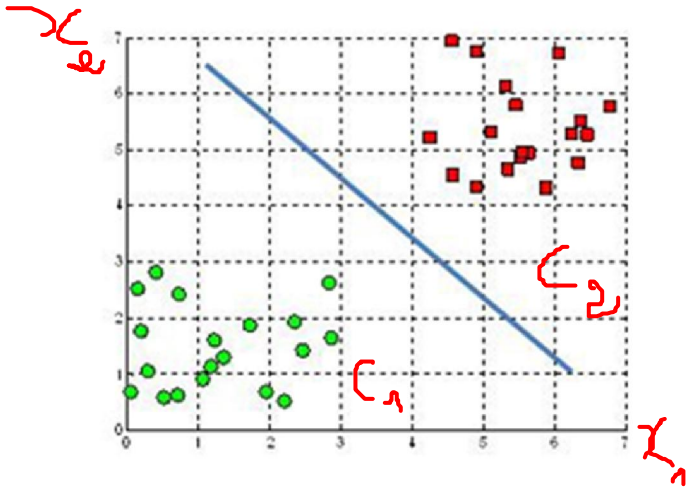
- **Un classificateur linéaire** utilise une **frontière de décision linaire** pour assigner les classes aux données.

- Soit D la dimension de x :

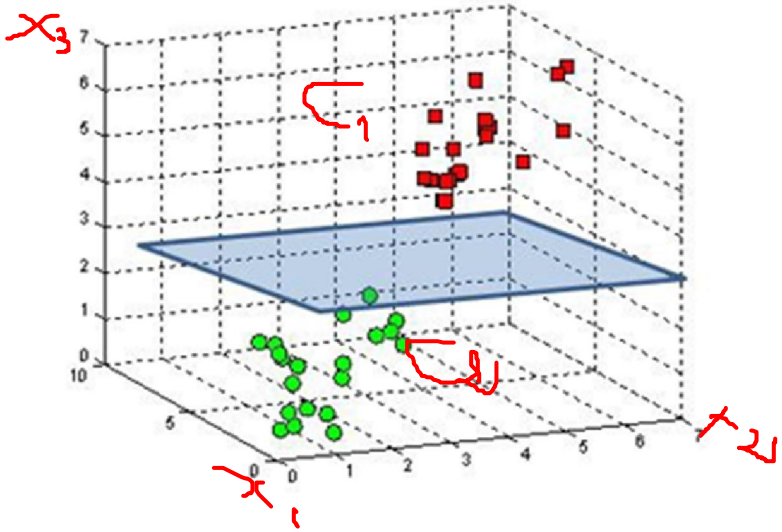
  - Pour $D = 2$, la frontière sera **une droite**.
  - Pour $D = 3$, la frontière sera **un plan**.
  - Pour $D > 3$, la frontière sera appelée **hyperplan**.

Exemple de frontières de décision linéaires
Pour D = 2, D=3

A hyperplane in $\mathbb{R}^2$ is a line



A hyperplane in $\mathbb{R}^3$ is a plane



A hyperplane in $\mathbb{R}^n$ is an *n*-1 dimensional subspace