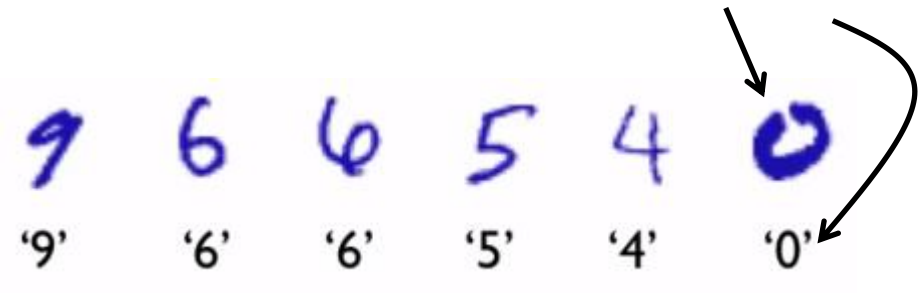
Apprentissage supervisé

Principe de l'apprentissage supervisé

Les algorithmes d'apprentissage supervisé procèdent comme suit:

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

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

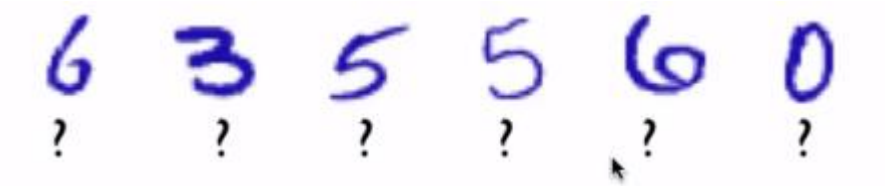


- On appelle $x^{(i)}$ **l'entrée** et $y^{(i)}$ **la cible** du *i*-ième exemple.

- 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:

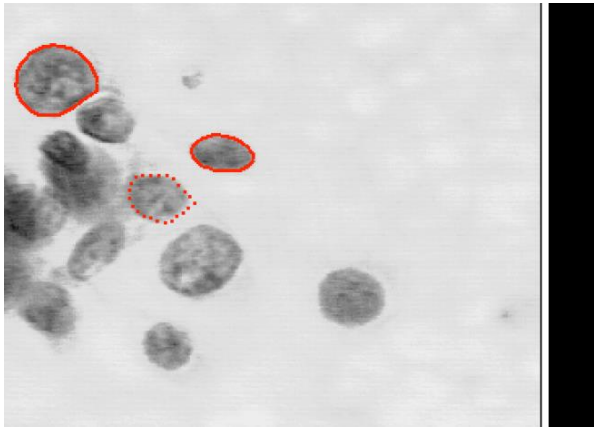


• On note le «programme» généré par l'algorithme d'apprentissage $f(x)$.

• 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 $\mathcal{D} \subset \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{Y}}$, trouver une fonction :

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

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

Apprentissage supervisé

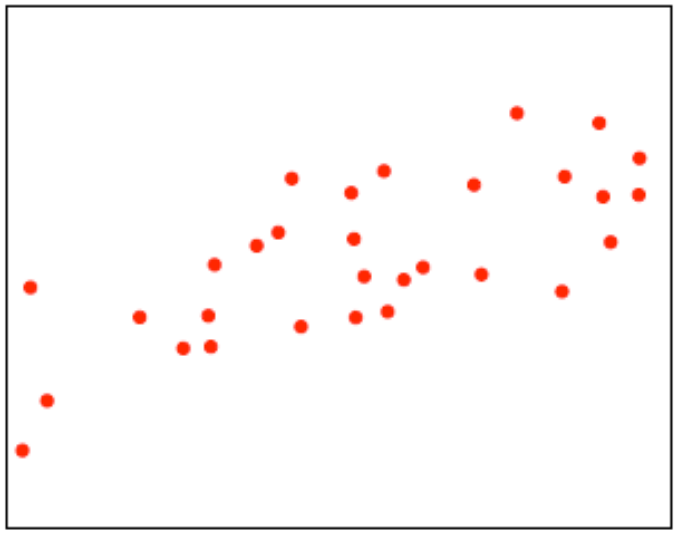
Les étapes pour résoudre un problème d'apprentissage supervisé.

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

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:

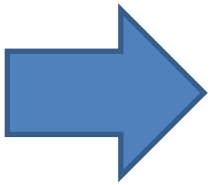
$$2\mathbf{X}^T\mathbf{X}\tilde{\mathbf{w}} - 2\mathbf{X}^T\boldsymbol{y} = 0 \quad \Rightarrow \quad \mathbf{X}^T\mathbf{X}\tilde{\mathbf{w}} = \mathbf{X}^T\boldsymbol{y}$$

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

- 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



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

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

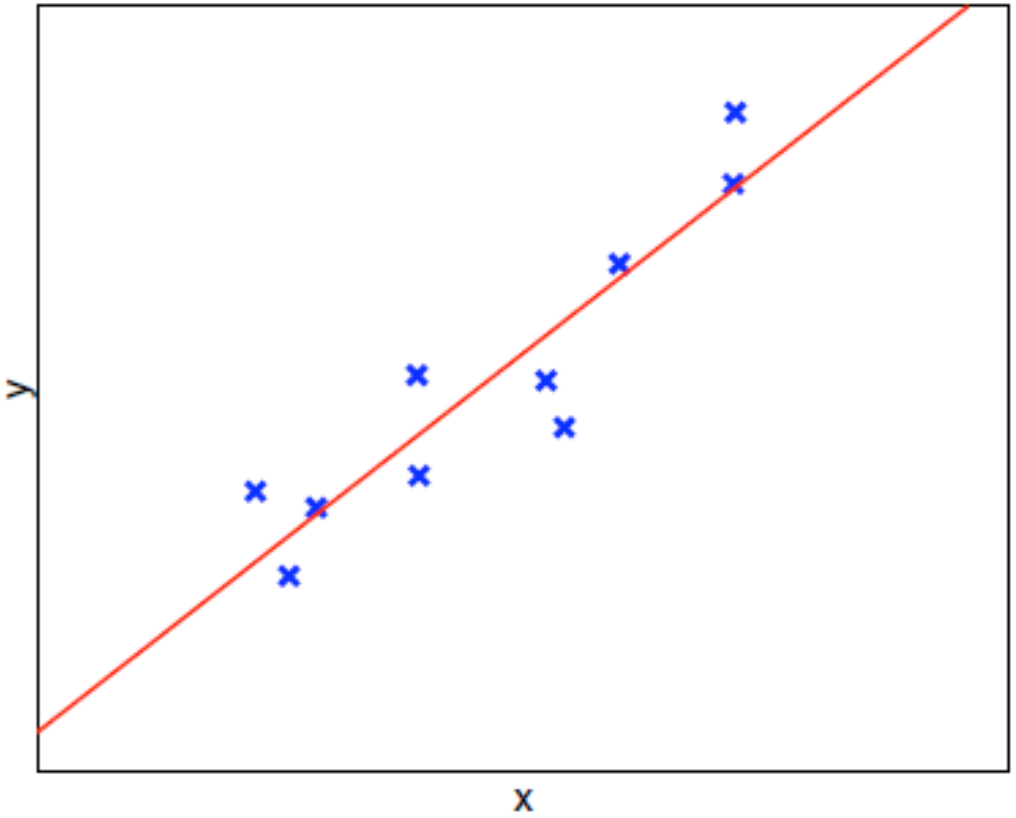
Exemple

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

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

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

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

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

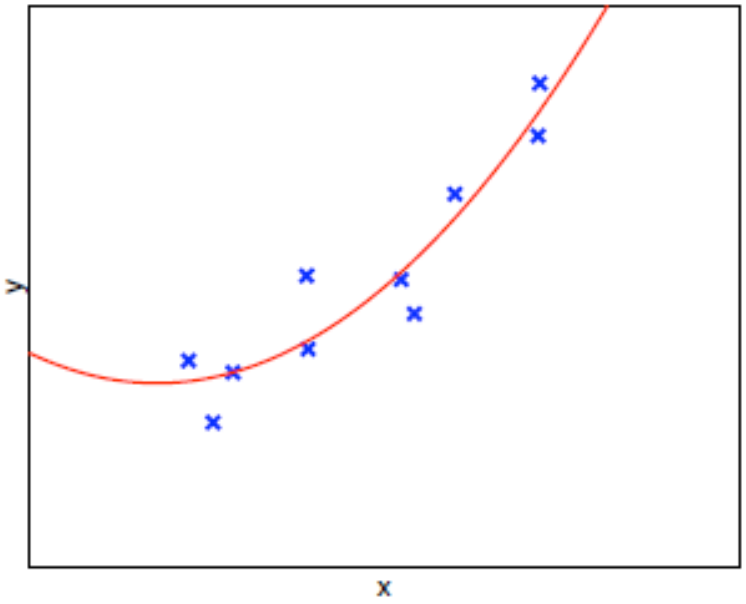
$$\mathbf{X}^T\mathbf{y} = \begin{bmatrix} 0.75 & 0.01 & 0.73 & 0.76 & 0.19 & 0.18 & 1.22 & 0.16 & 0.93 & 0.03 \\ 0.86 & 0.09 & -0.85 & 0.87 & -0.44 & -0.43 & -1.10 & 0.40 & -0.96 & 0.17 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} 2.49 \\ 0.83 \\ -0.25 \\ 3.10 \\ 0.87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

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

Fonctions d'ordre supérieur

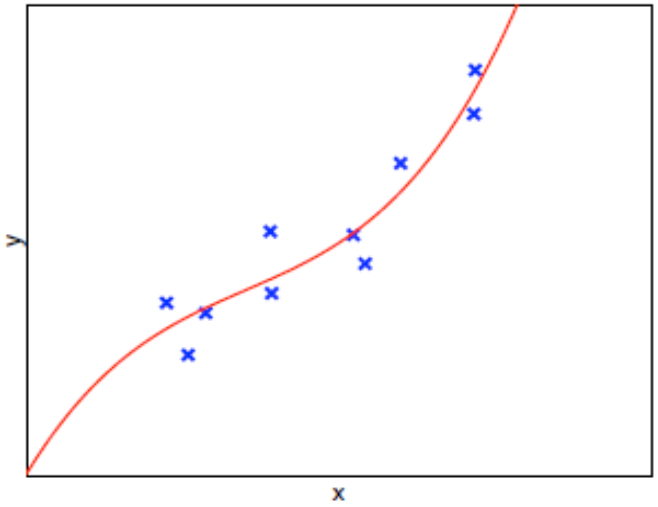
$$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y} = \begin{bmatrix} 4.11 & -1.64 & 4.95 \\ -1.64 & 4.95 & -1.39 \\ 4.95 & -1.39 & 10 \end{bmatrix}^{-1} \begin{bmatrix} 3.60 \\ 6.49 \\ 8.34 \end{bmatrix} = \begin{bmatrix} 0.68 \\ 1.74 \\ 0.73 \end{bmatrix}$$

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

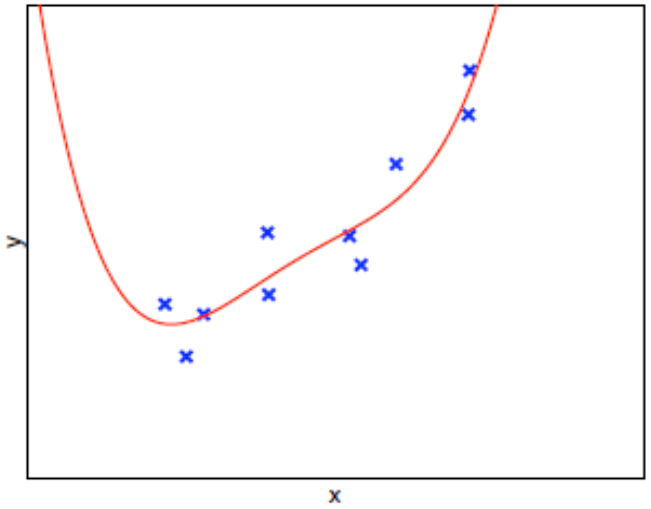


Fonctions d'ordre supérieur

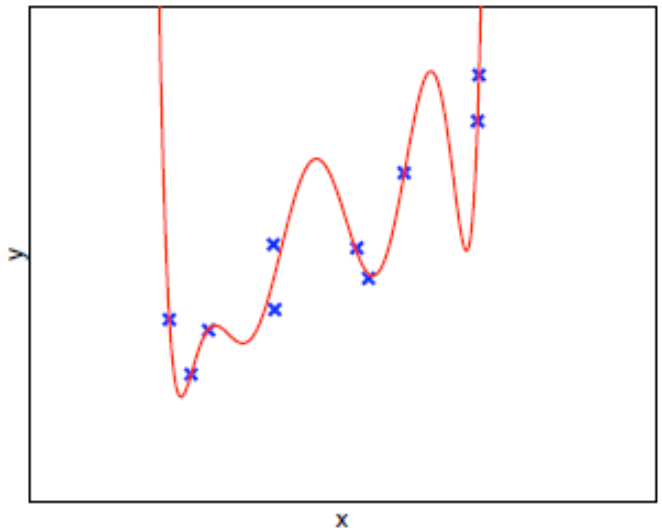
$m = 4$



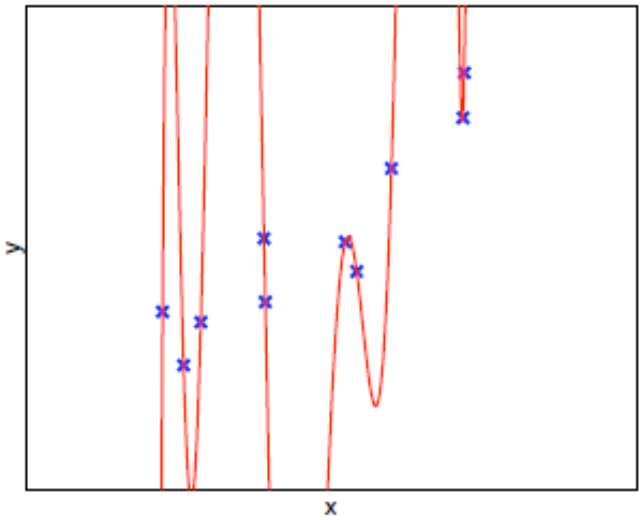
$m = 5$



$m = 8$



$m = 9$



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 $p(C_k \mid x)$. Cette probabilité est formulée par:

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

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 $\{R_1, R_2, \ldots, R_K\}$ 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:

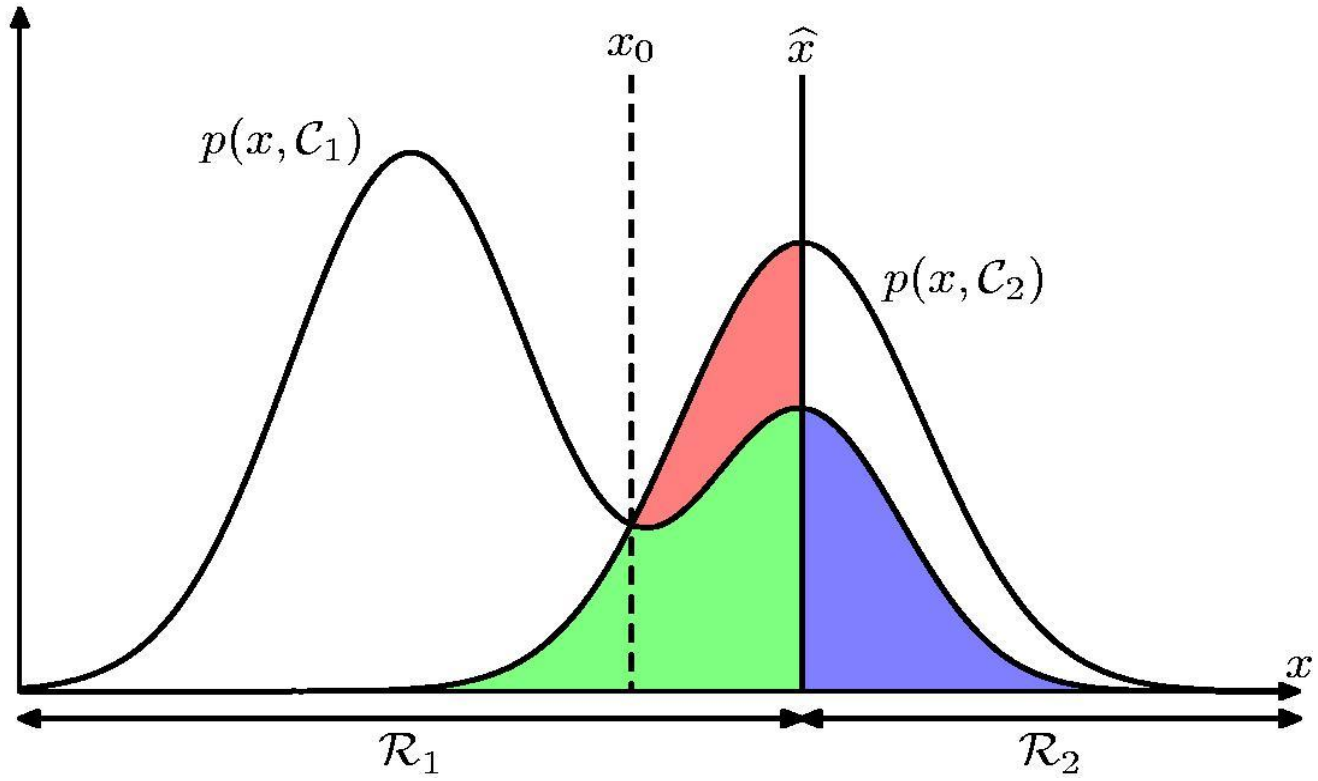
$$p(erreur) = p(x \in R_1|C_2) + p(x \in R_2|C_1)$$

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

- 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.

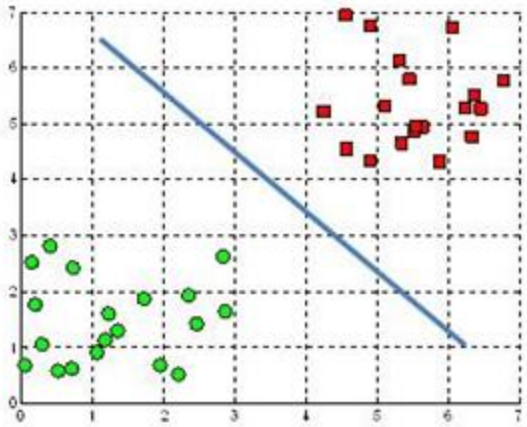
- **Un classificateur linéaire** utilise une **frontière de décision linaire** pour assigner les classes aux données.

- Soit D la dimension de x :

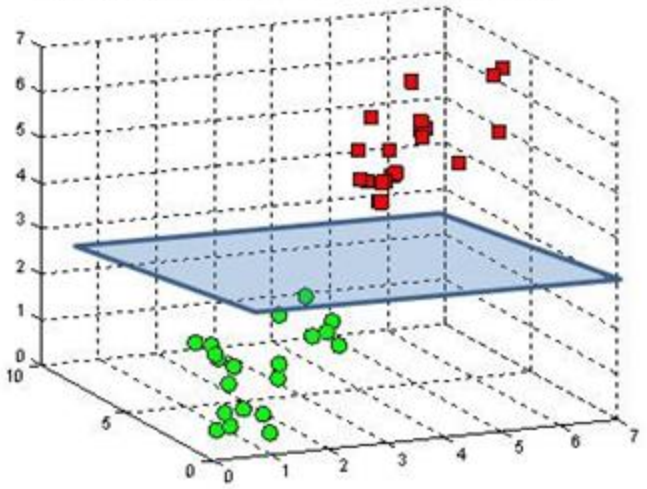
  - Pour D = 2, la frontière sera **une droite**.
  - Pour D = 3, la frontière sera **un plan**.
  - Pour D > 3, la frontière sera appelée **hyperplan**.

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

Frontière de décision linéaire

- Les classes qui peuvent être séparée par des frontières de décision linéaires sont dites **linéairement séparable**.

- Rappelons que pour le problème de régression, la variable cible y a une valeur réelle.

- Pour la classification, on peut utiliser plusieurs manière pour représenter la variable cible y. Soit $y=1$ pour la classe C_1 et $y=0$ pour la classe C_2.

- Pour $K > 2$, on peut utiliser un codage avec un vecteur $\mathbf{y} = (y_1, \ldots, y_K)$ où, si $\mathbf{x} \in C_k$, alors $y_k = 1$ et $y_j = 0, \forall j \neq k$.

Frontière de décision linéaire

Exemple de codage avec un vecteur:

- Pour $K > 2$, on peut utiliser un codage avec un vecteur $\mathbf{y} = (\boldsymbol{y_1}, \ldots, \boldsymbol{y_K})$ **où, si** $\mathbf{x} \in \boldsymbol{C_k}$**, alors** $\boldsymbol{y_k} = \mathbf{1}$ **et** $\boldsymbol{y_j} = \mathbf{0}, \forall \boldsymbol{j} \neq \boldsymbol{k}$.
- **Exemple :** si K = 5 alors

Donnée (x)	Classe	Codage (y)
	C_1	$(\mathbf{1},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0})$
	C_2	$(\mathbf{0},\mathbf{1},\mathbf{0},\mathbf{0},\mathbf{0})$
	C_3	$(\mathbf{0},\mathbf{0},\mathbf{1},\mathbf{0},\mathbf{0})$
	C_4	$(\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{1},\mathbf{0})$
	C_5	$(\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{0},\mathbf{1})$

Frontière de décision linéaire

- Dans le modèle de **régression linéaire**, la prédiction de y se fait par une fonction de la forme: $f(x) = \mathbf{w}^T x + w_0$.

- Pour la **classification**, on voudrait prédire **les étiquettes de classes** (class labels), ou plus généralement **les probabilités a posteriori des classes** dans l'intervalle [0,1].

- On peut généraliser le modèle linéaire de régression pour la classification en ayant une fonction linéaire de $\mathbf{w}$ qu'on transforme avec **une fonction non-linéaire**:

$$f(x) = g(\mathbf{w}^T x + w_0)$$

Frontière de décision linéaire

- On appelle $g(x)$ **une fonction d'activation**. Son inverse est appelé **une fonction de lien** en statistique.

- Pour ce modèle, les surfaces de décision correspondent aux valeurs $f(x) = \text{constante}$, de telle sorte que

$$\boldsymbol{w^T x + w_0 = \text{constante}.}$$

- Le modèles reposant sur l'équation précédentes s'appellent des **modèles linéaires généralisés**.

- Notons, qu'à cause de la fonction non-linéaire $g(x)$, ces modèles ne sont pas linéaires des paramètres **w**.

Cas de classification binaire (K = 2)

- La représentation la plus simple d'une fonction discriminante est obtenue en prenant une **fonction linéaire** du vecteur d'entrée x:

$$f(\mathbf{x}) = \mathbf{w}^T x + w_0$$

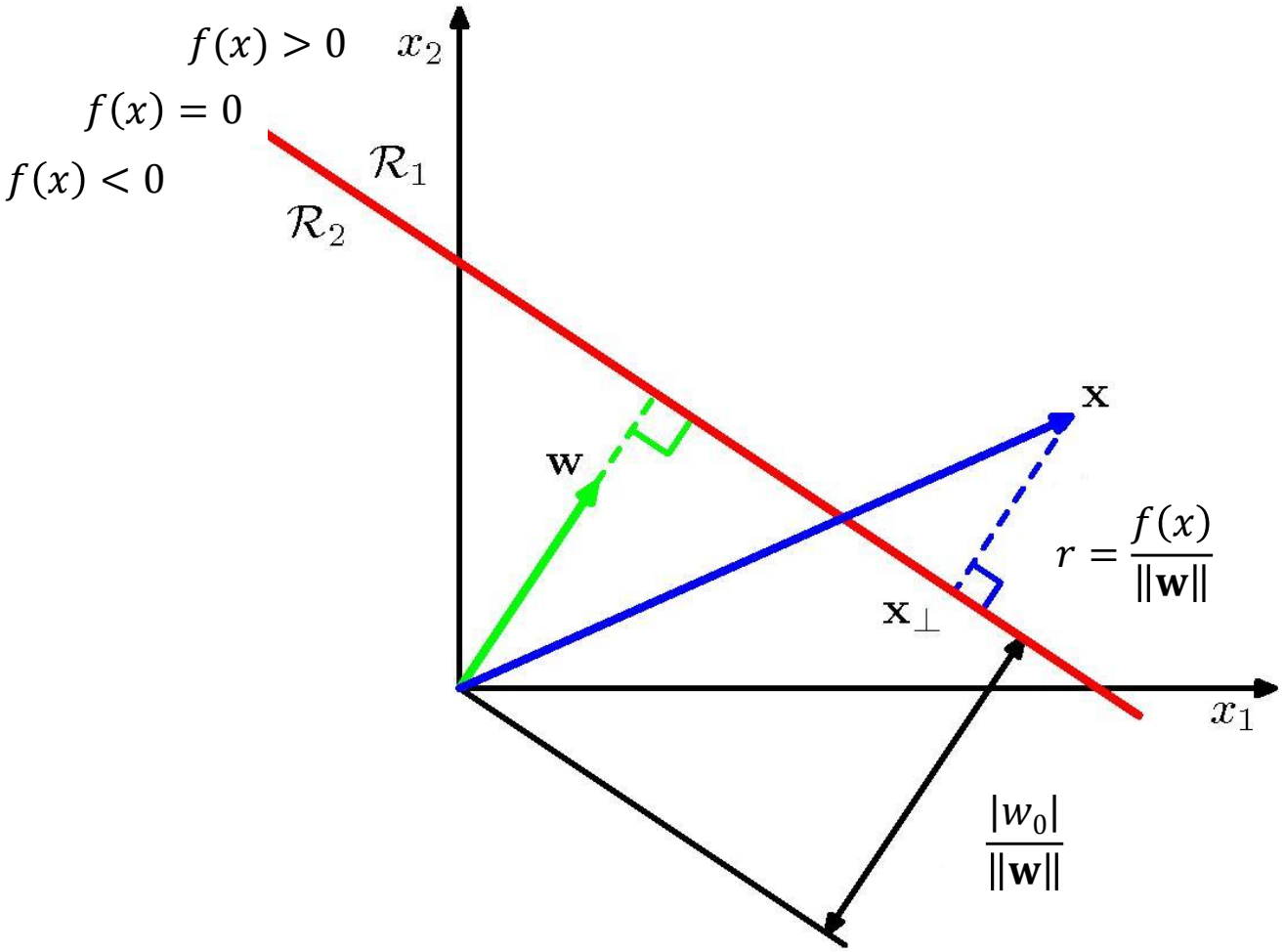
- Le vecteur **x** sera assigné à la classe C_1 si $f(x) \geq 0$ et à la classe C_2 dans le cas contraire.

- Si deux points $x^{(i)}$ et $x^{(j)}$ appartiennent à la frontière de décision, alors: $\mathbf{w}^T(x^{(i)} - x^{(j)}) = 0$.

- Donc, le vecteur **w** **est orthogonale à la frontière**.

Cas de classification binaire (K = 2)

Soit un problème de classification à 2 classes (**classification binaire**). Dans le cas de d = 2, on a la géométrie suivante:



Cas de classification binaire (K = 2)

- Si un point $\mathbf{x}$ appartient à la frontière de décision alors on aura $f(\mathrm{x}) = 0$.

- **La distance normalisée** entre l'origine des coordonnées et la frontière de décision est calculée comme suit:

$$\frac{\mathbf{w}^T x}{\|\mathbf{w}\|} = -\frac{w_0}{\|\mathbf{w}\|}$$

- Le paramètre w_0 détermine **la position de la frontière de décision** par rapport à l'origine des coordonnées.

Cas de classification binaire (K = 2)

- Si un point x est un point quelconque, la valeur de $f(x)$ donne une **mesure signée** de **la distance perpendiculaire** r entre le point x et la frontière de décision (voir la figure précédente).

- Soit $x_\perp$ **la projection** de x sur la frontière de décision, alors:

$$x = x_\perp + r\frac{\mathrm{w}}{\|\mathbf{w}\|}$$

$$\Rightarrow \quad \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 = \mathbf{w}^{\mathrm{T}}x_\perp + w_0 + r\frac{\mathbf{w}^{\mathrm{T}}\mathbf{w}}{\|\mathbf{w}\|}$$

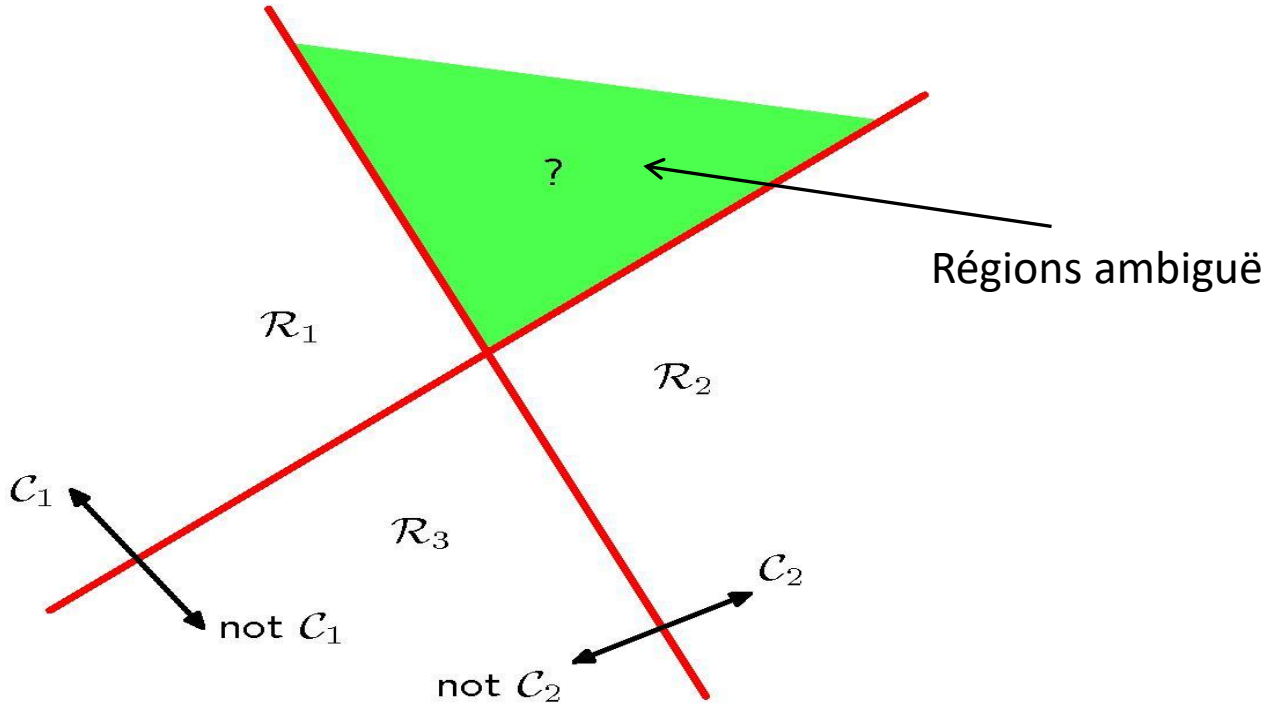
On a $\mathbf{w}^{\mathrm{T}}x_\perp + w_0 = 0$, car $f(x_\perp) = 0$

$$donc \quad r = \frac{f(x)}{\|\mathbf{w}\|}$$

Cas de classification multiple (K > 2)

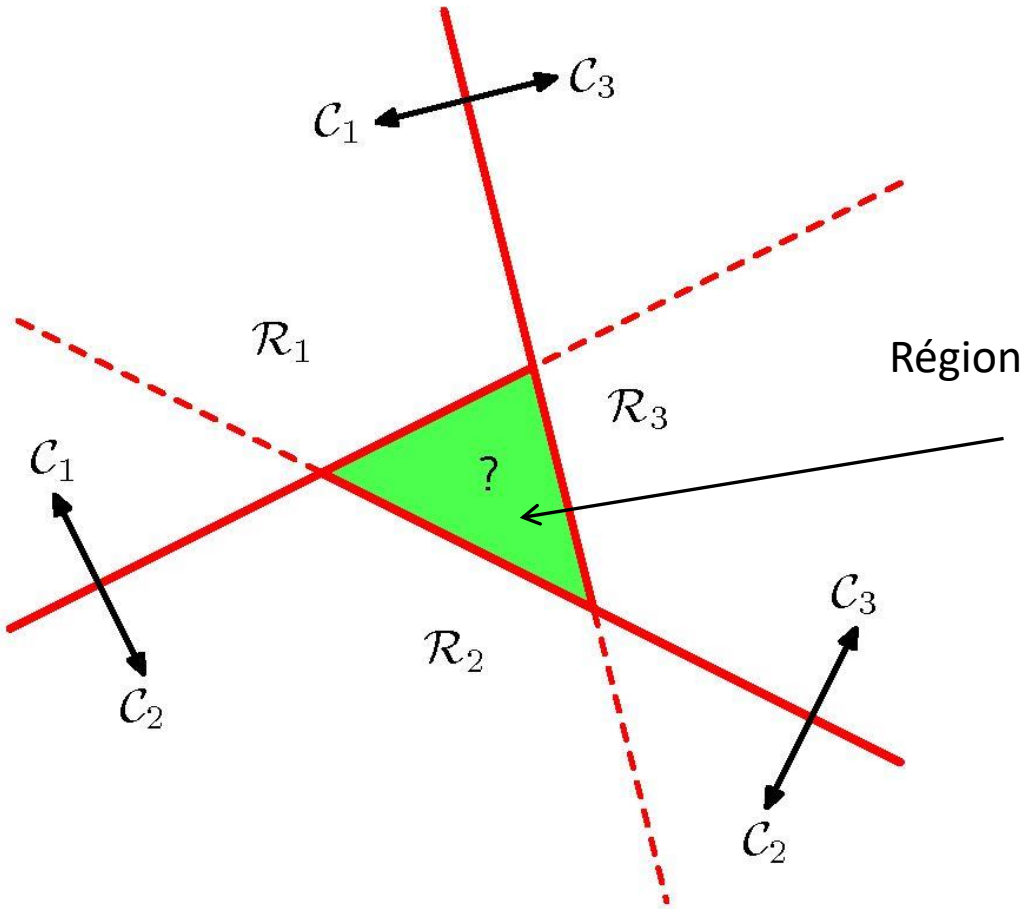
- Une des façons de construire **un classificateur à K classes** est de **combiner plusieurs classificateurs binaires**.

- Soit (K-1) classificateurs binaires chacun séparant une des classes C_k, k = 1…, K-1 (**classification un-versus-tous**).



Cas de classification multiple (K > 2)

- Une autre alternative est de construire K(K-1)/2 classificateurs binaires, un pour chaque deux classes (**classification un-versus-un**).



Cas de classification multiple (K > 2)

- On peut éviter ces difficultés en considérant un classificateur comprenant K **fonctions linéaires discriminantes** de la forme:

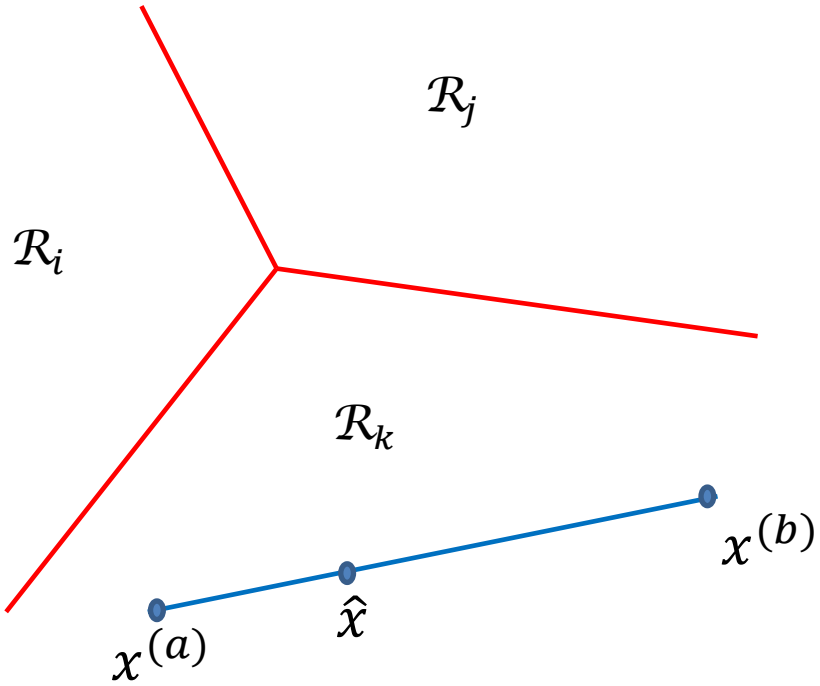
$$f_k(x) = \mathbf{w}^T x + w_{k0}$$

- On assigne x pour une classe C_k si $f_k(x) > f_j(x), \ \forall j \neq k$.

- La frontiere de decision entre la classe C_k et C_j est donnée par la relation $f_k(x) = f_j(x)$, qui correspond à:

$$(\mathbf{w}_k - \mathbf{w}_j)^T x = w_{k0} - w_{j0}$$

Cas de classification multiple (K > 2)



On peut démontrer dans ce cas que les **régions de décisions** sont **connexes** et **convexes.**

Classification par les moindres carrées

- Comme on a fait pour la régression, on augmente d'abord la matrice de données $\mathbf{X}$ d'une colonne égale à **1**.

- On définit pour chaque donnée $x^{(i)}$ un vecteur colonne $y^{(i)} = \left({y_1}^{(i)}, {y_2}^{(i)}, \cdots, {y_K}^{(i)}\right)^{\mathrm{T}}$ et une matrice $\mathbf{Y}$ dont la i-ième ligne contient ${y^{(i)}}^T$.

- Soit la matrice $\widetilde{\mathbf{W}}$ dont la k-ième colonne contient $\widetilde{\mathbf{w}}_k$.

- **L'erreur des moindres carrées** est alors donnée par:

$$E(\widetilde{\mathbf{W}}) = \frac{1}{2} Tr\left\{\left(\mathbf{X}\widetilde{\mathbf{W}} - \mathbf{Y}\right)^T \left(\mathbf{X}\widetilde{\mathbf{W}} - \mathbf{Y}\right)\right\}$$

Classification par les moindres carrées

- En calculant la dérivée de E par rapport à $\widetilde{\mathbf{W}}$, on obtient:

$$\widetilde{\mathbf{W}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{X}'\mathbf{Y}$$

- On appelle $\mathbf{X}'$ **le pseudo-inverse** de la matrice $\mathbf{X}$.

- La fonction discriminante $\mathbf{f}$ est alors donnée par:

$$\mathbf{f}(\mathbf{x}) = \widetilde{\mathbf{W}}^T\mathbf{x} = \mathbf{Y}^T(\mathbf{X}')^T\mathbf{x}$$

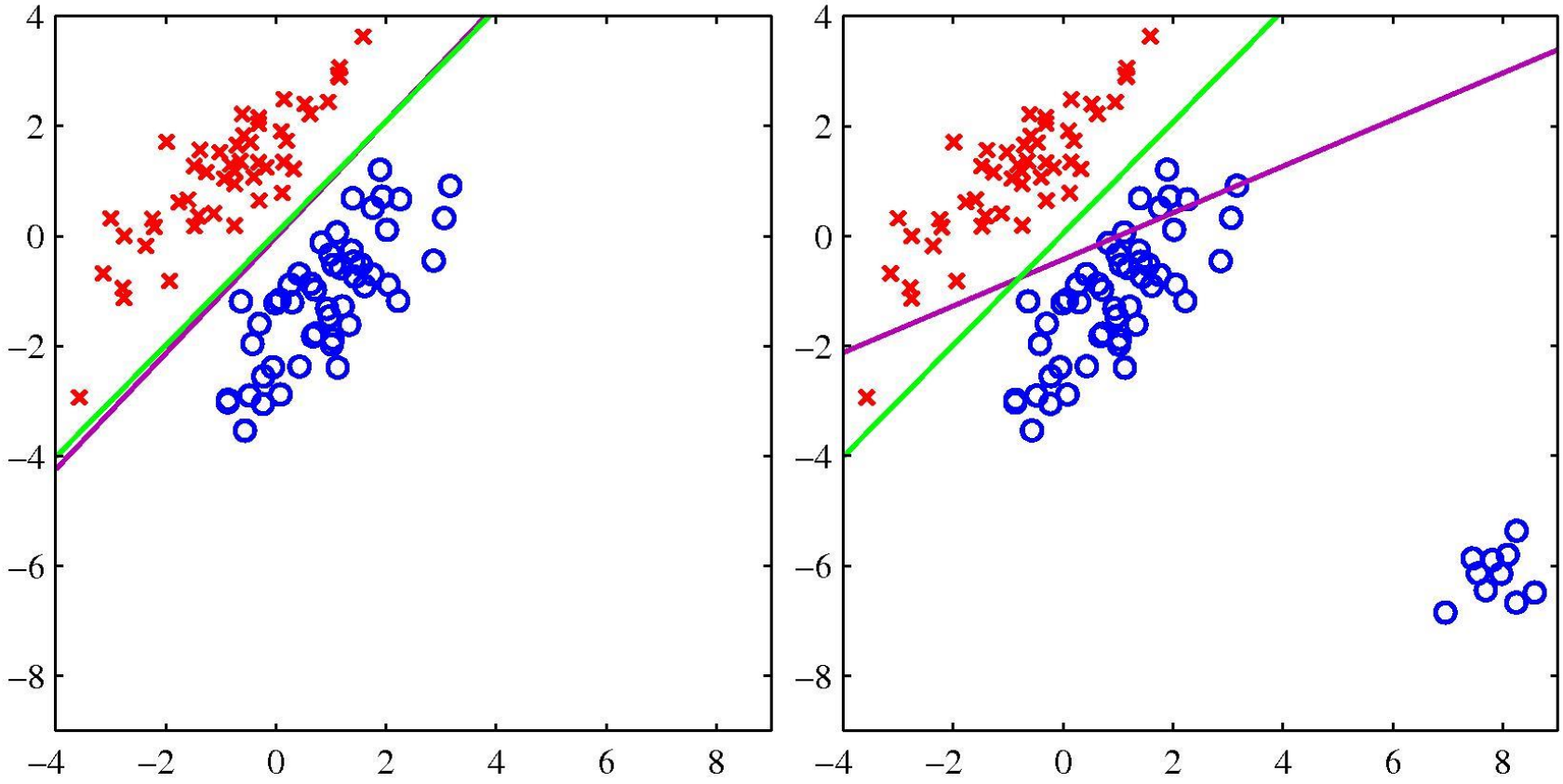
- Notons que les éléments de $\mathbf{f}$ auront la somme égale à 1. Mais les entrées de $\mathbf{f}$ ne sont pas pour autant des probabilités.

Classification par les moindres carrées

Exemples:

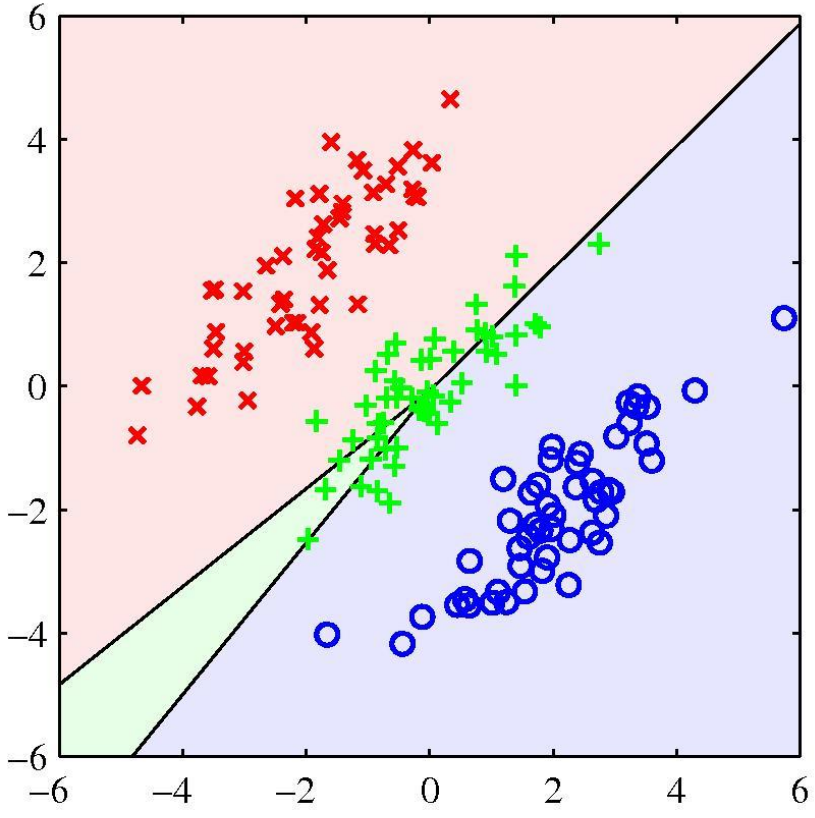
Moindres carrés.

Régression logistique.

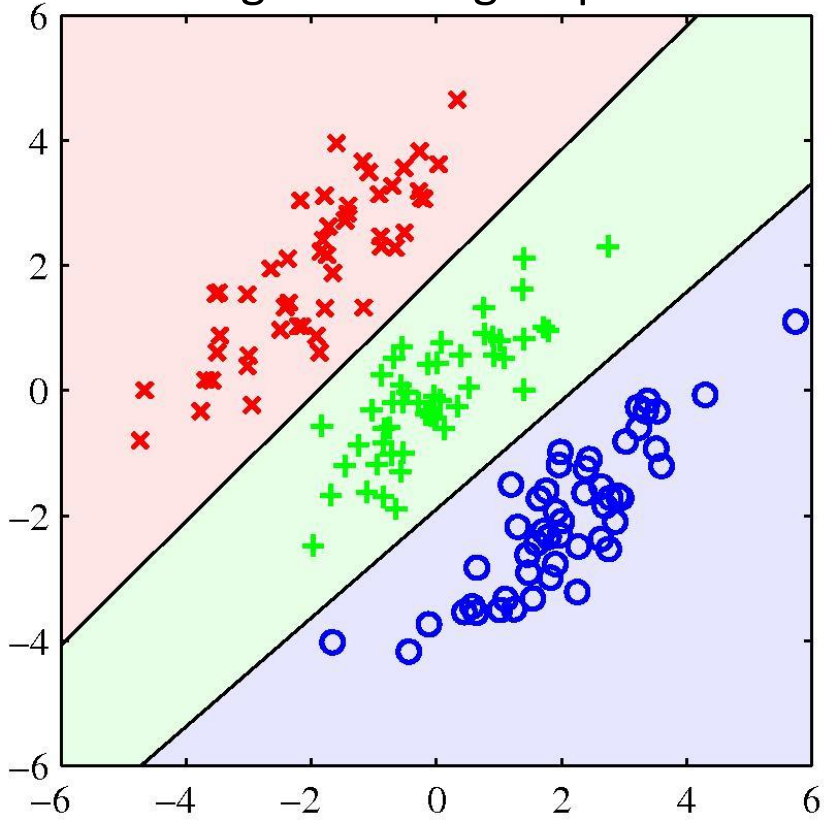


Classification par les moindres carrées

Moindres carrés



Régression logistique



Classification par les moindres carrées

- La méthode des **moindre carrés** donne **une solution exacte** pour la classification. Cependant, elle a des limitations:

  ➢ Elle fonctionne bien quand les classes sont **séparables** et les données des classes suivent des **lois Gaussiennes**.

  ➢ Elle n'est pas robuste **aux données aberrantes** (**outliers**) qui perturbe la frontière de décision.

  ➢ Par la suite, on verra une meilleure classification avec **la régression logistique**.

Références

1. M. S. Allili. Techniques d'apprentissage automatique (Cours de 2e cycle). Université du Québec en Outaouais (UQO), Québec, Canada. Hivers 2015.
2. S. Rogers et M Girolami. A first Course in machine learning, CRC press, 2012.
3. C. Bishop. Pattern Recognition and Machine learning. Springer 2006.
4. R. Duda, P. Storck et D. Hart. Pattern Classification. Prentice Hall, 2002.