Classification par Régression Logistique

Reconnaissance des Formes (Semaine VII)

Université de Lille, France

6 mars 2013

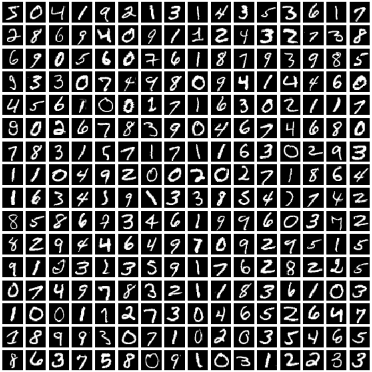
Rappel

Classification

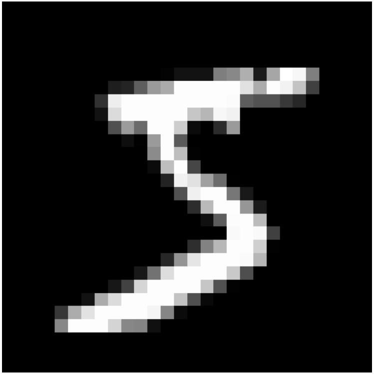
- L'ensemble des données est partagé en quelques classes différentes,
- On a quelques exemplaires pour qui on sait les classes correctes.
- On veut classifier correctement les données pour qui les classes correctes sont inconnus.

Exemple : Classification des caractères manuscrits

Données d'entrainement



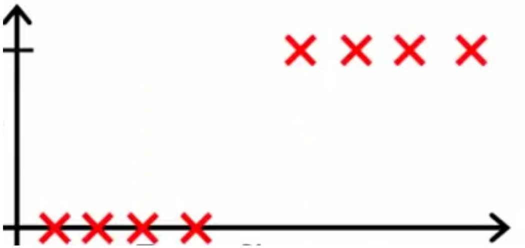
Ça corresponds à quel chiffre ?



Le problème de classification

le cas le plus simple

- Les données à classifier sont représenter en une dimension i.e. $x \in \mathbb{R}$
- Nous avons deux classes seulement, i.e. $y \in \{0, 1\}$.



Le Problème de classification

On cherche une fonction $h : \mathbb{R} \to \{0, 1\}$ telle que pour chaque x on a

- $h(x) = 0$ si x est dans classe 0 et
- $h(x) = 1$ si x est dans classe 1

Le Problème de classification

On cherche une fonction $h : \mathbb{R} \to \{0, 1\}$ telle que pour chaque x on a

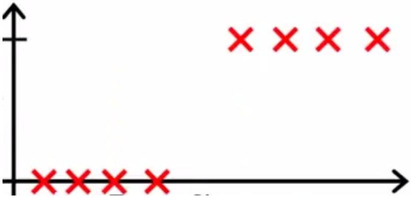
- $h(x) = 0$ si x est dans classe 0 et
- $h(x) = 1$ si x est dans classe 1

Si cette fonction existe, le problème de classification est déjà résolu.
Nous allons utiliser les données d'entrainement pour trouver h.

Une Fonction Linéaire ?

- On défini un séparateur linéaire $\theta_0 + \theta_1 x$
- On utilise les exemplaires disponible pour chercher les valeurs convenables pour θ_0 et θ_1. On prédit la classe y de la manière suivante

$$y = 1 \text{ si } \theta_0 + \theta_1 x \geq 0.5$$
$$y = 0 \text{ si } \theta_0 + \theta_1 x < 0.5$$



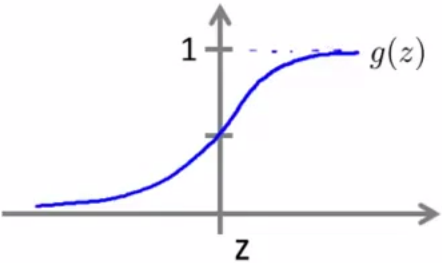
Question

Est-ce que cette approche va marcher ?

Le Modèle

La fonction Sigmoid où Logistique

$$g(z) := \frac{1}{1 + e^{-z}}$$



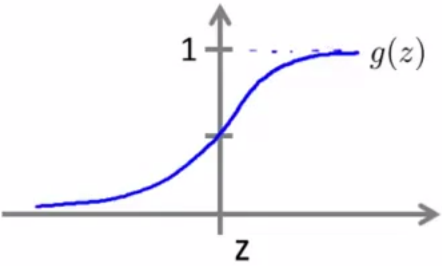
Remarque

- $0 \leq g(z) \leq 1$
- $g(z) \geq 0.5$ si $z \geq 0$
- $g(z) < 0.5$ si $z < 0$

Le Modèle

La fonction **Sigmoid** où **Logistique**

$$g(z) := \frac{1}{1 + e^{-z}}$$



Remarque

- $0 \leq g(z) \leq 1$
- $g(z) \geq 0.5$ si $z \geq 0$
- $g(z) < 0.5$ si $z < 0$

Le Modèle

$$h_\theta(x) := g(\theta_0 + \theta_1 x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x)}}$$

Le Modèle

$$h_\theta(x) := \frac{1}{1 + e^{-(\theta_0 + \theta_1 x)}}$$

Prédiction

$$y = 1 \text{ si } h_\theta(x) \geq 0.5$$
$$y = 0 \text{ si } h_\theta(x) < 0.5$$

Le Modèle

$$h_\theta(x) := \frac{1}{1 + e^{-(\theta_0 + \theta_1 x)}}$$

Prédiction

$$y = 1 \text{ si } h_\theta(x) \geq 0.5$$
$$y = 0 \text{ si } h_\theta(x) < 0.5$$

L'intuition

$h_\theta(x)$ est l'estimation de la probabilité de $y = 1$ sachant x, i.e. $P(y = 1|x)$. Comme on a $y = 0$ **où** $y = 1$, on peut dériver $P(y = 0|x)$ i.e.

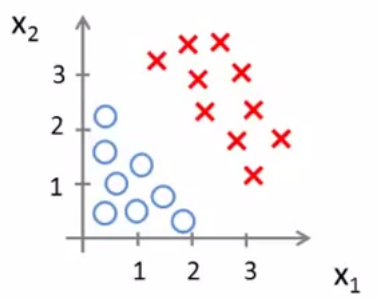
$$P(y = 0|x) = 1 - P(y = 1|x) = 1 - h_\theta(x).$$

Le Modèle

En général les données n'ont pas qu'une seule dimension !
Chaque point $\mathbf{x}$ est un vecteur de dimension k i.e.
$\mathbf{x} := (x_1, x_2, \ldots, x_k) \in \mathbb{R}^k$.

Exemple

Pour $k = 2$ on a $\mathbf{x} := (x_1, x_2)$ où $x_1 \in \mathbb{R}$ est $x_2 \in \mathbb{R}$, donc $\mathbf{x} \in \mathbb{R}^2$.



Le Modèle Général

Chaque point $(x_1, x_2, \ldots, x_k) \in \mathbb{R}^k$ est un vecteur de dimension k,.

$$h_\theta(\mathbf{x}) := \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k)}} = \frac{1}{1 + e^{-\sum_{i=0}^{k} \theta_i x_i}}$$

Le Modèle Général

Chaque point $(x_1, x_2, \ldots, x_k) \in \mathbb{R}^k$ est un vecteur de dimension k,.

$$h_\theta(\mathbf{x}) := \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k)}} = \frac{1}{1 + e^{-\sum_{i=0}^{k} \theta_i x_i}}$$

$$= \frac{1}{1 + e^{-(\boldsymbol{\theta}^T \mathbf{x})}} \text{ où}$$

$$\boldsymbol{\theta} := \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_k \end{bmatrix} \text{ et } \mathbf{x} := \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_k \end{bmatrix} \text{ donc } \boldsymbol{\theta}^T \mathbf{x} = \sum_{i=0}^{k} \theta_i x_i.$$

Le Modèle Général

Prédiction

$y = 1$ si $h_\theta(\mathbf{x}) \geq 0.5$
$y = 0$ si $h_\theta(\mathbf{x}) < 0.5$ où

$$h_\theta(\mathbf{x}) := \frac{1}{1 + e^{-(\boldsymbol{\theta}^T \mathbf{x})}}$$

Le Modèle Général

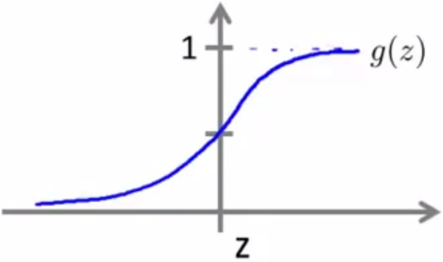
Prédiction

$y = 1$ si $h_\theta(\mathbf{x}) \geq 0.5$
$y = 0$ si $h_\theta(\mathbf{x}) < 0.5$ où

$$h_\theta(\mathbf{x}) := \frac{1}{1 + e^{-(\boldsymbol{\theta}^T \mathbf{x})}}$$

$$g(z) := \frac{1}{1 + e^{-z}}$$

$g(z) \geq 0.5$ si $z \geq 0$ et $g(z) < 0.5$ si $z < 0$



Le Modèle Général

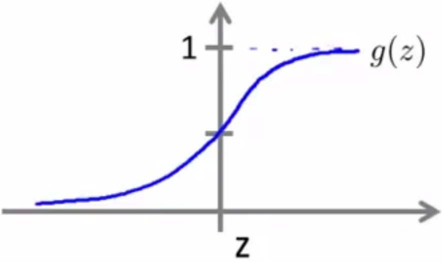
Prédiction

$y = 1$ si $h_\theta(\mathbf{x}) \geq 0.5$ i.e. si $\boldsymbol{\theta}^T\mathbf{x} \geq 0$
$y = 0$ si $h_\theta(\mathbf{x}) < 0.5$ i.e. si $\boldsymbol{\theta}^T\mathbf{x}$

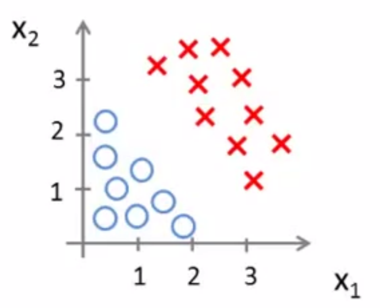
$$h_\theta(\mathbf{x}) := \frac{1}{1 + e^{-(\boldsymbol{\theta}^T\mathbf{x})}}$$

$$g(z) := \frac{1}{1 + e^{-z}}$$

$g(z) \geq 0.5$ si $z \geq 0$ et $g(z) < 0.5$ si $z < 0$



Exemple



poids

Prédiction

$y = 1$ si $\theta_0 + \theta_1 x_1 + \theta_2 x_2 \geq 0$

$y = 0$ si $\theta_0 + \theta_1 x_1 + \theta_2 x_2 < 0$

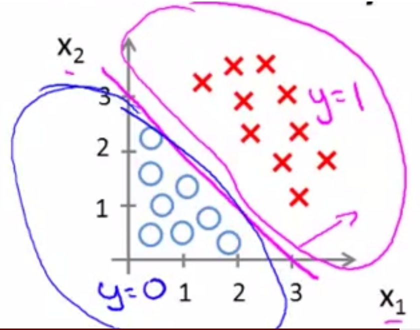
Remarquez que $\theta_0,\ \theta_1$ et θ_2 sont inconnus. On doit les chercher !

Exemple

Supposons que $\boldsymbol{\theta} := \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$ est donné donc,

$$\theta_0 = -3, \ \theta_1 = 1, \ \theta_2 = 1$$

X = (x1, x2)



Prédiction

$y = 1$ si $x_1 + x_2 \geq 3$
$y = 0$ si $x_1 + x_2 < 3$

$x_1 + x_2 = 3$ est le **frontière de décision**.

Comment trouver $\boldsymbol{\theta}$?

Coût d'une prédiction

$$\textbf{coût}(h_\theta(\mathbf{x}), y) := \begin{cases} -\log(h_\theta(\mathbf{x})) & \text{si } y = 1 \\ -\log(1 - h_\theta(\mathbf{x})) & \text{si } y = 0 \end{cases}$$

L'intuition

Imaginons que $y = 1$

Si $h_\theta(\mathbf{x}) = 1$ on a coût $= 0$ mais

coût $\to \infty$ lorsque $h_\theta(\mathbf{x}) \to 0$

Ç'est une façon de pénaliser l'algorithme pour la mauvaise prédiction !

Il faut chercher les θ qui donnent le coût le plus petit possible.

Régression Logistique : Comment trouver $\boldsymbol{\theta}$?

Coût d'une prédiction

$$\begin{aligned}\mathbf{coût}(h_\theta(\mathsf{x}), y) &:= \begin{cases} -\log(h_\theta(\mathsf{x})) & \text{si } y = 1 \\ -\log(1 - h_\theta(\mathsf{x})) & \text{si } y = 0 \end{cases} \\ &= -y \log(h_\theta(\mathsf{x})) - (1 - y) \log(1 - h_\theta(\mathsf{x})) \quad \textbf{pourquoi ?}\end{aligned}$$

Régression Logistique : Comment trouver $\boldsymbol{\theta}$?

Coût d'une prédiction

$$\textbf{coût}(h_\theta(\mathbf{x}), y) := \begin{cases} -\log(h_\theta(\mathbf{x})) & \text{si } y = 1 \\ -\log(1 - h_\theta(\mathbf{x})) & \text{si } y = 0 \end{cases}$$

$$= -y\log(h_\theta(\mathbf{x})) - (1-y)\log(1 - h_\theta(\mathbf{x})) \quad \textbf{pourquoi ?}$$

La fonction de coût $J(\theta)$

On a m exemplaires d'entrainement $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(m)}\}$. Soit $y^{(i)}$ la classe de $\mathbf{x}^{(i)}$, $i = 1..m$.
On défini la fonction de coût par

$$J(\theta) := -\frac{1}{m}\sum_{i=1}^{m} y^{(i)}\log(h_\theta(\mathbf{x}^{(i)})) - (1 - y^{(i)})\log(1 - h_\theta(\mathbf{x}^{(i)}))$$

Comment trouver $\boldsymbol{\theta}$?

On cherche les paramètres qui donnent le coût le plus petit possible

$$\boldsymbol{\theta}^* := \min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \text{ où}$$

$$J(\boldsymbol{\theta}) := -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log(h_\theta(\mathsf{x}^{(i)})) - (1 - y^{(i)}) \log(1 - h_\theta(\mathsf{x}^{(i)}))$$

$\rightarrow$ minimiser la fonction de coût $J(\boldsymbol{\theta})$ et
utiliser le $\boldsymbol{\theta}$ qui correspond à $J(\boldsymbol{\theta})$ minimum

La minimisation de $J(\boldsymbol{\theta})$ est possible pas les méthodes de la théorie d'optimisation, mais cela dépasse le cadre de ce cours

$\rightarrow$ Nous allons utiliser les fonctions d'optimisations existantes en scilab.

Comment trouver $\boldsymbol{\theta}$?

On cherche les paramètres qui donnent le coût le plus petit possible

$$\boldsymbol{\theta}^* := \min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \text{ où}$$

$$J(\boldsymbol{\theta}) := -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log(h_\theta(\mathsf{x}^{(i)})) - (1 - y^{(i)}) \log(1 - h_\theta(\mathsf{x}^{(i)}))$$

$\rightarrow$ minimiser la fonction de coût $J(\boldsymbol{\theta})$ et
utiliser le $\boldsymbol{\theta}$ qui correspond à $J(\boldsymbol{\theta})$ de minimum

L'algorithme

Entrainement

Utiliser les données d'entrainement, i.e. les exemplaires $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(m)}\}$ et leur lables $\{y^{(1)}, y^{(2)}, \ldots, y^{(m)}\}$ pour trouver le vecteur $\boldsymbol{\theta}$. Donc,

m = nb d'entrainement

$$\boldsymbol{\theta}^* := \min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \text{ où}$$

une image

$$J(\boldsymbol{\theta}) := -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log(h_\theta(\mathbf{x}^{(i)})) - (1 - y^{(i)}) \log(1 - h_\theta(\mathbf{x}^{(i)}))$$

Prédiction de la classe d'un nouvel exemple

Pour chaque nouveau $\mathbf{x}$ $h_\theta(\mathbf{x}) := \frac{1}{1+e^{\theta^T\mathbf{x}}}$ classifier le de la manière suivante,

$$y = 1 \text{ si } h_\theta(\mathbf{x}) \geq 0.5$$

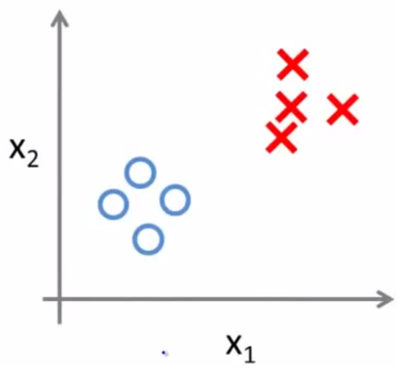
$$y = 0 \text{ si } h_\theta(\mathbf{x}) < 0.5$$

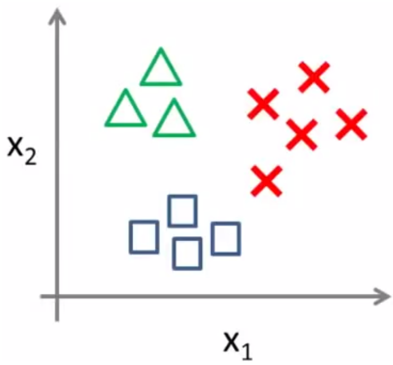
L'algorithme Multi-Classes

Pour l'instant cette approche marche que pour le cas de deux classe $y \in \{0, 1\}$ seulement.

Question

Comment peut on le généraliser pour le cas de plusieurs classes ?





L'algorithme Multi-Classes

Supposons que l'on a $l > 2$ classes, i.e. $y \in \{1, 2, \ldots, l\}$. On peut créer un classifier de la manière suivante.

1. **Entrainer** un classifier de régression logistique $h_\theta^{(i)}(\mathbf{x})$ pour chaque classe $i = 1, 2, \ldots, l$.

Rappel : $h_\theta^{(i)}(\mathbf{x})$ estime $P(y = i|\mathbf{x}),\ i = 1, 2, \ldots, l$

2. Pour **prédire** la classe d'un nouveau $\mathbf{x}$, choisir $y \in \{1, 2, \ldots, l\}$ tel que

$$y = \max_{i=1..l} h_\theta^{(i)}(\mathbf{x})$$

Ça veut dire que l'on suit la décision du classifier plus confiants.

L'algorithme Multi-Classes : Exemple