Régression logistique

Cours de machine learning avancé

AI For You

Régression logistique

1. Définition

- La classification consiste à attribuer une classe ou une catégorie aux observations, en fonction des données statistiques de ces observations. Après une phase d'apprentissage où une fonction est paramétrée avec des données dont la classe est connue. Cette fonction permet de classer les nouvelles données qui ne sont pas étiquetées.

- Exemples:
  - Prédire la classe de l’objet en photo ;
  - Prédire si un patient est atteint d’une maladie ;
  - Prédire si le client ouvrira l’e-mail ;
  - Reconnaître des chiffres écrit à la main.

2. Jeu de données

	Nb d'e-mails ouvert (x_1)	Nb de produits achetés(x_2)	Chiffre d'affaires (x_3)	Ouverture de l'e-mail (y)
1	12	3	120	1
2	0	1	40	0
3	30	10	1800	1
4	14	5	799	1
…	…	…	…	…
m	25	2	260	0

Jeu de données d'entraînement pour la prédiction de l'ouverture d'un email

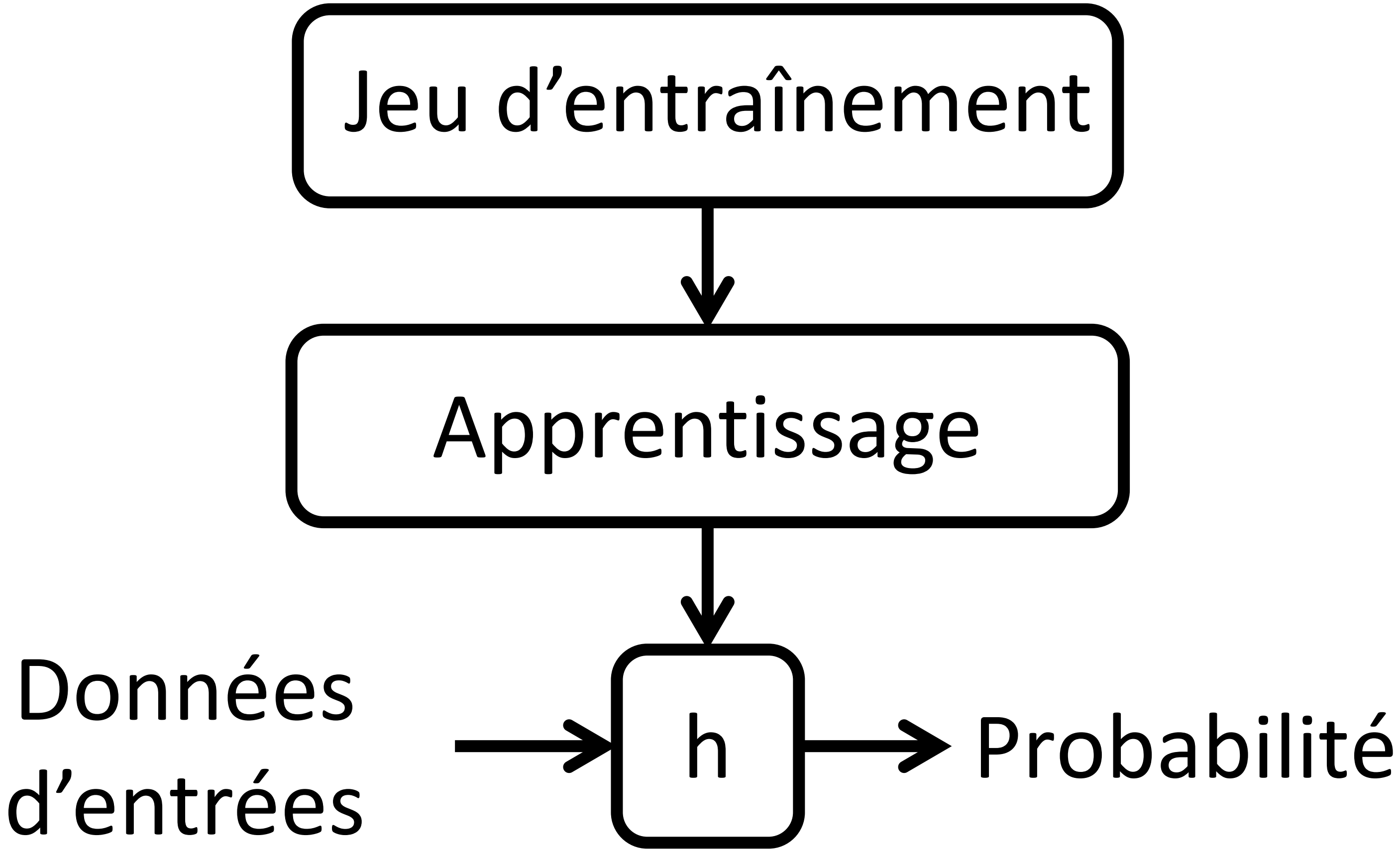
$x^{(1)}$ = (12, 3, 120)

$y^{(4)}$ = 1

$x_2^{(3)}$ = 10

$(x^{(3)}, y^{(3)})$ = (30, 10, 1800, 1)

3. Schéma d’un modèle de classification



Modèle d'un algorithme de classification

Régression logistique

- I/ Classification
- **II/ Définir le problème**
- III/ Hypothèse
- IV/ Fonction de coût
- V/ Descente de gradient
- VI/ Interprétation

1. Problème de Classification

- Imaginez, vous travaillez dans un laboratoire médical, votre but est de distinguer les tumeurs bénignes des tumeurs malignes. Le but est de savoir avec la plus grande certitude de quel type de tumeur il s'agit.

- **Comment différencier les tumeurs malignes des tumeurs bénignes ?**

- La différenciation n'est pas simple, les humains même les experts ont un biais selon leurs expériences et chaque tumeur maligne non détectée est lourde de conséquences. Vous voulez une solution neutre et précise pour votre laboratoire, vous voulez donc utiliser un algorithme d'apprentissage automatique qui classera automatiquement vos tumeurs en fonction de ses données.

2. Jeu de données

- Un modèle de régression logistique est un modèle de classification qui cherche à établir une relation linéaire entre une ou plusieurs variables quantitatives, appelées variables explicatives, et une autre variable qualitative, appelée variable cible.

	Taille (x_1)	Age (x_2)	Malignes (y)
1	20	50	0
2	10	45	0
3	50	55	1
4	70	80	1
…	…	…	…
m	15	20	0

- **m** nombre de lignes (observations).
- **X** est le vecteur des variables d'entrée appelées variables explicatives.
- **y** est la variable de sortie appelée variable cible.

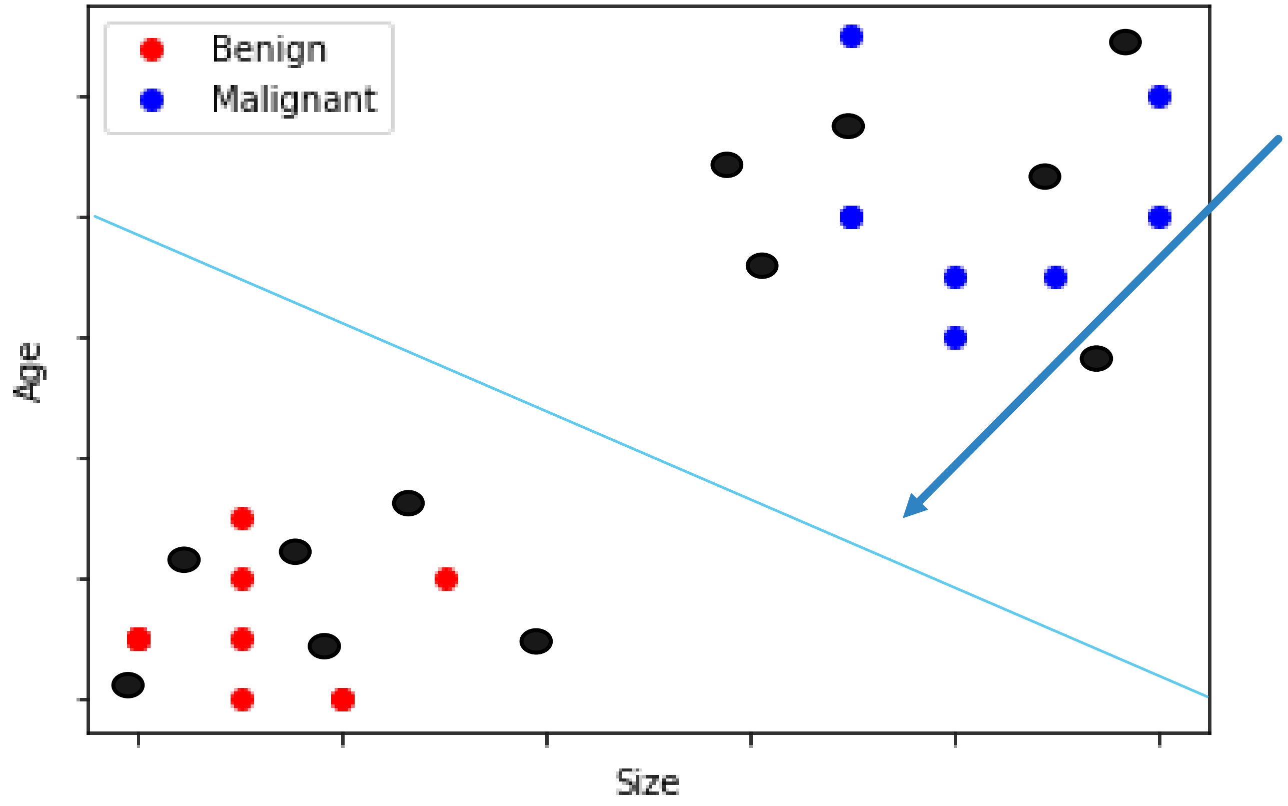
Exemples de lecture de données :

$x^{(1)}$ = 70

$y^{(4)}$ = 178

$(x^{(3)}, y^{(3)})$ = (40, 232)

3. Frontière de décision



$$W^T X = w_0 . x_0 + \ldots + w_n . x_n$$

Si y=1 :

$$w_0 . x_0 + \ldots + w_n . x_n \geq 0$$

Si y=0 :

$$w_0 . x_0 + \ldots + w_n . x_n < 0$$

Régression logistique

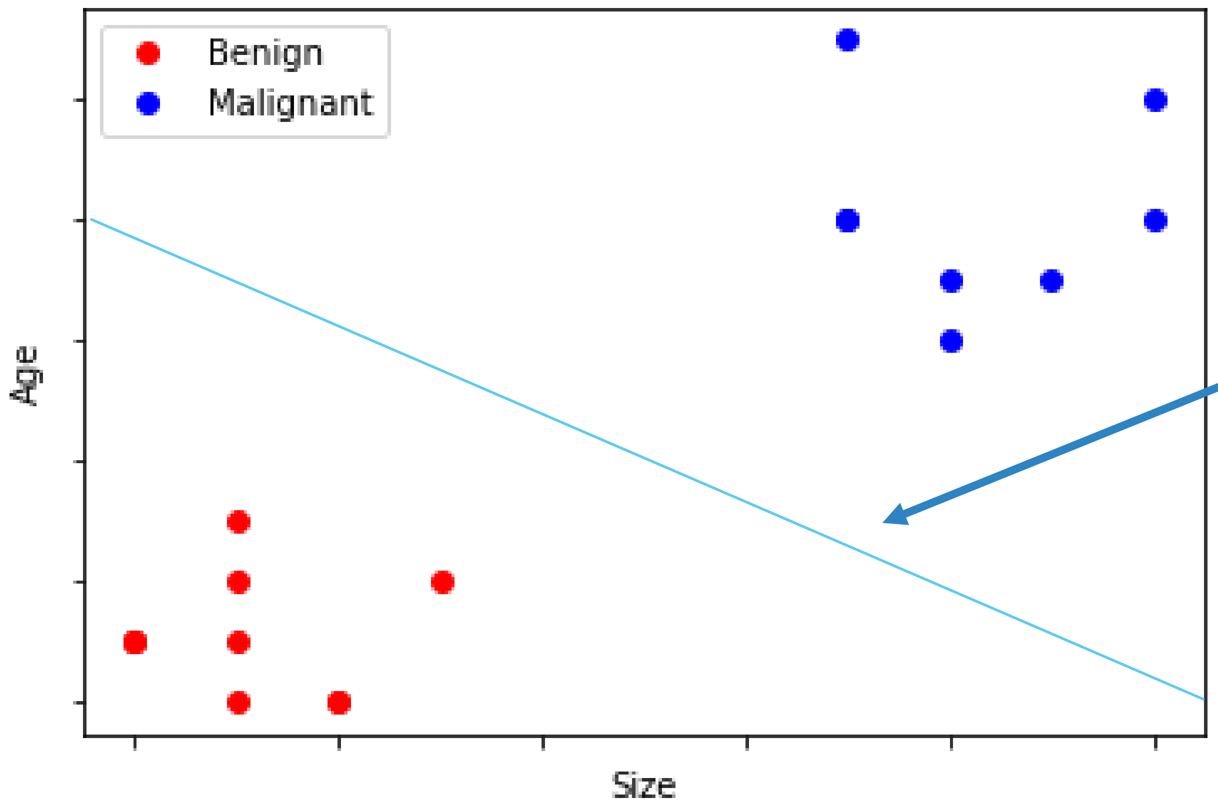
1. Similitude avec la régression linéaire

- Nous voulons une ligne séparant nos deux ensembles de données, nous utiliserons donc l'équation de la régression linéaire multivariée.

$$h(X) = w_0 . x_0 + w_1 . x_1 + w_2 . x_2 + \dots + w_n . x_n$$

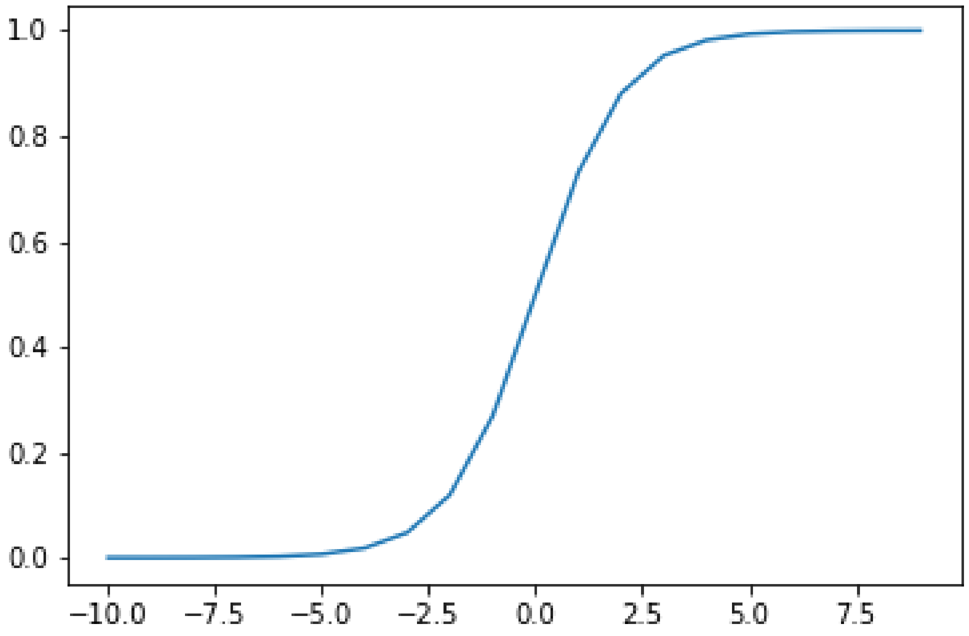
- Pour la régression linéaire $h(X) \in R$, pour la régression logistique $0 \leq h(X) \leq 1$.

2. Frontière de décision



$$W^T X = w_0 . x_0 + \dots + w_n . x_n$$

3. La fonction Sigmoïd



$$g(z) = \frac{1}{1+e^{-z}} \quad ; \quad 0 \leq g(z) \leq 1$$

3. Hypothèse

- Cette fonction prend en entrée les paramètres et les variables afin de réaliser une combinaison linéaire limitée entre 0 et 1.

$$h(X) = g(W^T X) = \frac{1}{1+e^{-W^T X}}$$

- Plus h(X) est proche de 1 plus l'observation à de chance d'appartenir à la classe 1, plus h(X) est proche de 0 plus l'observation à de chance d'appartenir à la classe 0.

Régression logistique

1. Rappel de la régression linéaire

$$Minimize\ J(W) = \frac{1}{2m}\sum_{i=1}^{m}(h(x^{(i)}) - y^{(i)})^2$$

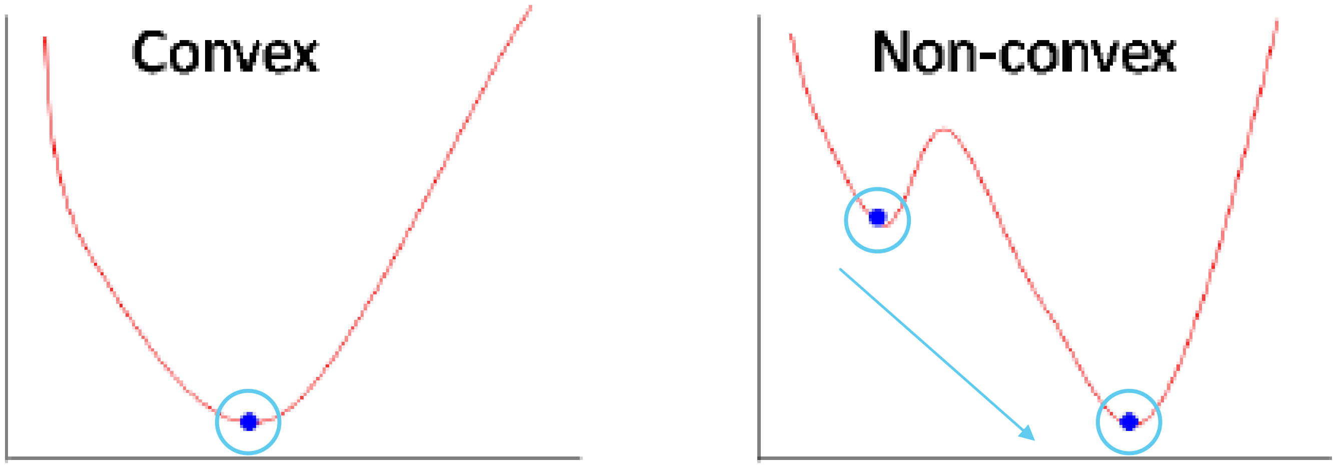
$$Minimize\ J(W) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}(h(x^{(i)}) - y^{(i)})^2$$

$$Minimize\ J(W) = \frac{1}{m}\sum_{i=1}^{m}Cost(h(x^{(i)}), y)$$

2. Non-convexe vs. convexe

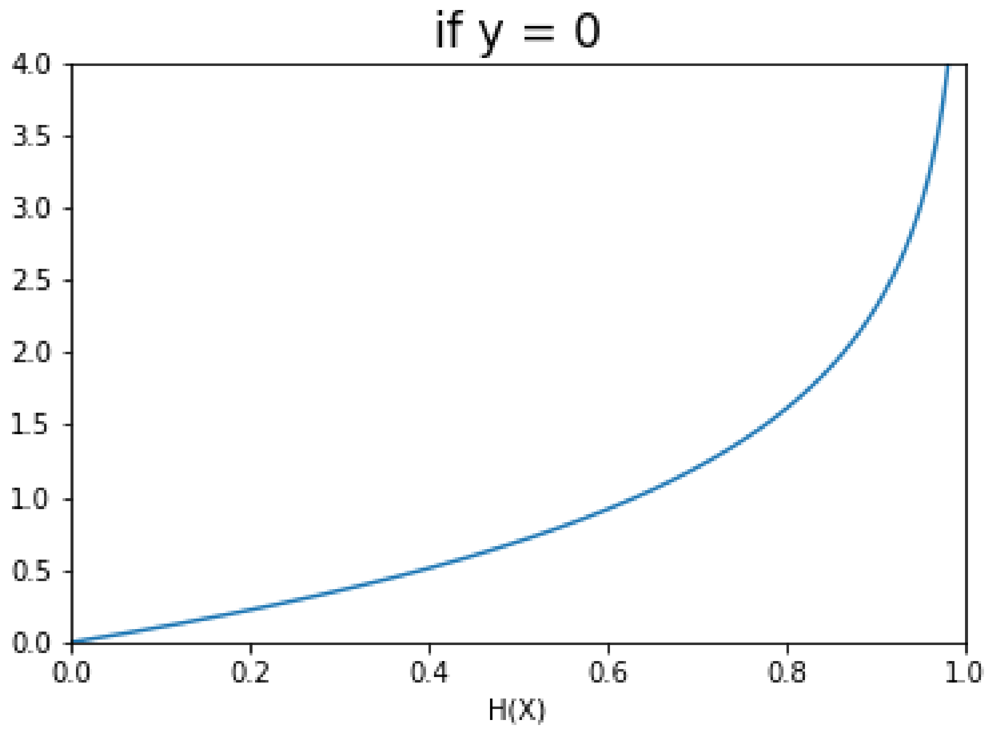
$$Cost(h(x^{(i)}), y) = \frac{1}{2}\left(\boxed{h(x^{(i)})} - y^{(i)}\right)^2$$

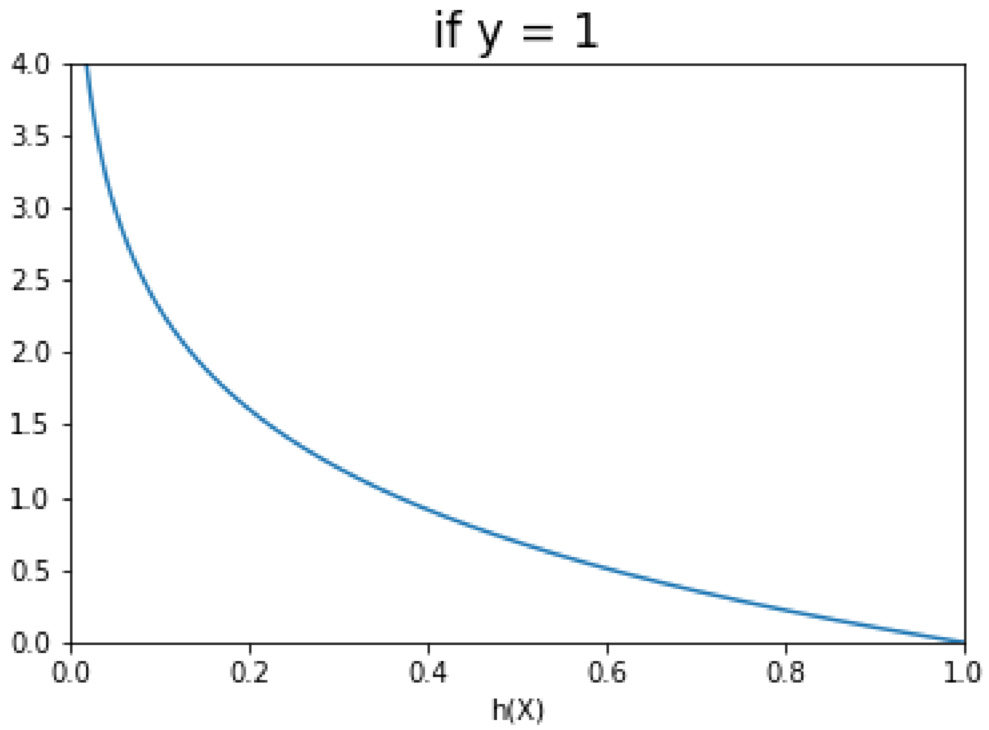
$$g(z) = \frac{1}{1 + e^{-z}}$$



3. Non-convexe à convexe

$$Cost(h(x^{(i)}), y) = \begin{cases} -\log\left(h\left(x^{(i)}\right)\right) & if\ y = 1 \\ -\log\left(1 - h\left(x^{(i)}\right)\right) & if\ y = 0 \end{cases}$$





4. Fonction de coût

$$Cost(h(x^{(i)}), y) = \begin{cases} -\log\left(h\left(x^{(i)}\right)\right) & if\ y = 1 \\ -\log\left(1 - h\left(x^{(i)}\right)\right) & if\ y = 0 \end{cases}$$

$$Cost(h(x^{(i)}), y) = -\boxed{y \log(h(x))} - \boxed{(1-y)\log(1-h(x))}$$

if y=1 $$Cost(h(x^{(i)}), y) = -1 \log(h(x))$$

if y=0 $$Cost(h(x^{(i)}), y) = -1 \log(1 - h(x))$$

$$Minimize\ J(W) = \frac{1}{m}\sum_{i=1}^{m} -y^i \log(h(x^{(i)})) - (1-y^i)\log(1-h(x^{(i)}))$$

Régression logistique

1. La descente de gradient

Répéter jusqu'à la convergence {

$w_j := w_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h(x^{(i)}) - y^{(i)}) . x_j^{(i)}$ (simultanément j = (0, …, n)

}

Régression logistique

1. Hypothèse

$$\mathrm{h}(X) = \frac{1}{1+e^{-W^T X}}$$

- Le résultat de cette fonction est limité entre 0 et 1, il peut être interprété comme la probabilité que l'observation soit classée dans la classe 1 :

$$h(x) = P(y = 1|x, W)$$

- h(x) = 0.7; Il y a 70% de chance que x appartiennent à la classe 1.

- Pour avoir la probabilité que les observations appartiennent à la classe 0 :

$$P(y = 0|x, W) + P(y = 1|x, W) = 1$$

$$P(y = 0|x, W) = 1 - P(y = 1|x, W)$$

$$P(y = 0|x, W) = 1 - \mathrm{h(x)}$$