


Classification

Master AIDN: Applications Interactives et Données Numériques

Sylvie Gibet

1

1



Régression Linéaire - rappels

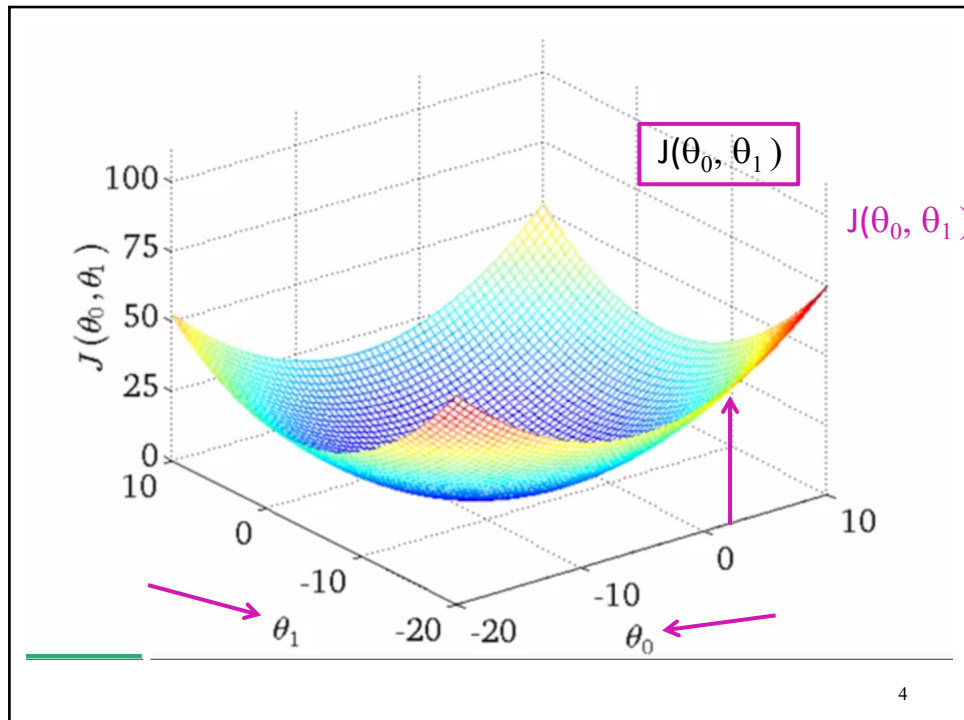
2

2

- Hypothèse : $h_{\theta}(x) = \theta_0 + \theta_1 x$
- Paramètres : θ_0, θ_1
- Fonction coût : $J(\theta_0, \theta_1) = \frac{1}{2m} \cdot \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$
- But : minimiser $J(\theta_0, \theta_1)$
 θ_0, θ_1

3

3



4

4

- Fonction coût : $J(\theta_0, \theta_1)$ $J(\theta_0, \theta_1, \theta_2, \dots \theta_n)$
- But : minimiser $J(\theta_0, \theta_1)$ minimiser $J(\theta_0, \theta_1, \theta_2, \dots \theta_n)$
 θ_0, θ_1 $\theta_0, \theta_1, \theta_2, \dots \theta_n$
- Algorithme :
 - Commencer avec θ_0, θ_1 (initial guesses, e.g., $\theta_0 = 0, \theta_1 = 0$)
 - Changer θ_0, θ_1 pour réduire $J(\theta_0, \theta_1)$ jusqu'à atteindre un minimum

5

5

Algorithme de minimisation du coût : descente de gradient

- Algorithme :
 - Répéter jusqu'à convergence :
 - Repeat {

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

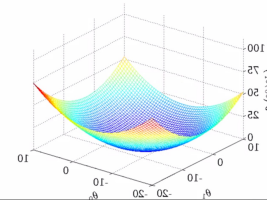
 until θ_0, θ_1 converge
 - }
 - Mises à jour simultanées de θ_0 et θ_1

6

6

Régression linéaire - Résumé

- Droite qui approxime les données – trouver les paramètres de la droite
 - Fonction coût quadratique entre la droite et les données
 - L'algorithme pour trouver les paramètres de la droite est un algorithme de minimisation du coût quadratique, de type descente de gradient : algorithme incrémental, qui met à jour simultanément les 2 paramètres de la droite
 - Cet algorithme converge toujours vers un minimum global (voir forme de la fonction coût)



7



Vers la classification Régression logistique

8

8

Classification

- Email: Spam / Not Spam?
- Transactions en ligne : frauduleuse (OUI/NON)?
- Tumeur : Maligne / Bénine ?

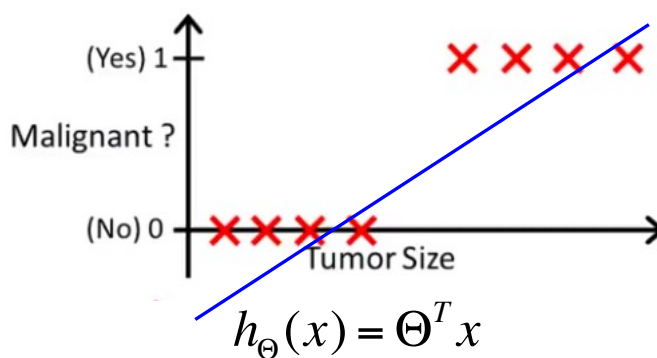
$y \in \{0,1\}$ 0: Classe négative (e.g., tumeur bénine)
 1: Classe positive (e.g., tumeur maligne)

$y \in \{0,1,2,3\}$ problème multi-classes

9

9

Classification

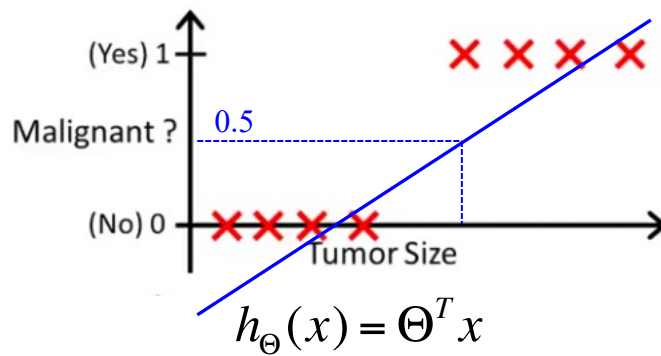


- Classifieur à seuil (threshold) : output $h_{\theta}(x)$ à 0.5:
 - si $h_{\theta}(x) \geq 0.5$, prédire $y = 1$
 - si $h_{\theta}(x) < 0.5$, prédire $y = 0$

10

10

Classification



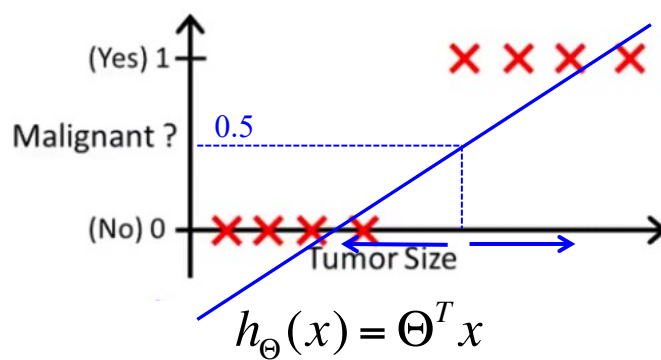
□ Classifieur à seuil (threshold) : output $h_{\theta}(x)$ à 0.5:

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

11

11

Classification



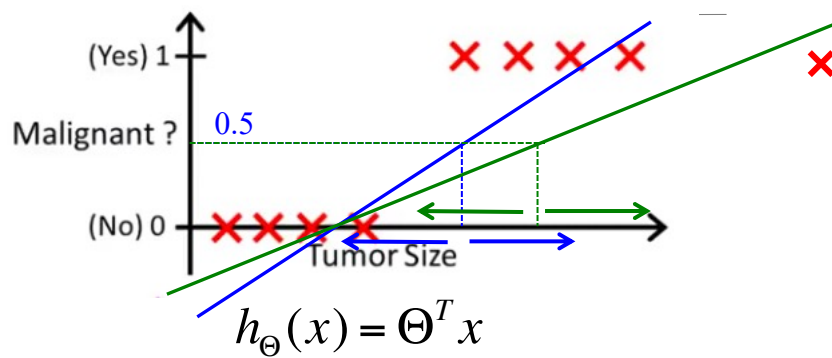
□ Classifieur à seuil (threshold) : output $h_{\theta}(x)$ à 0.5:

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

12

12

Classification



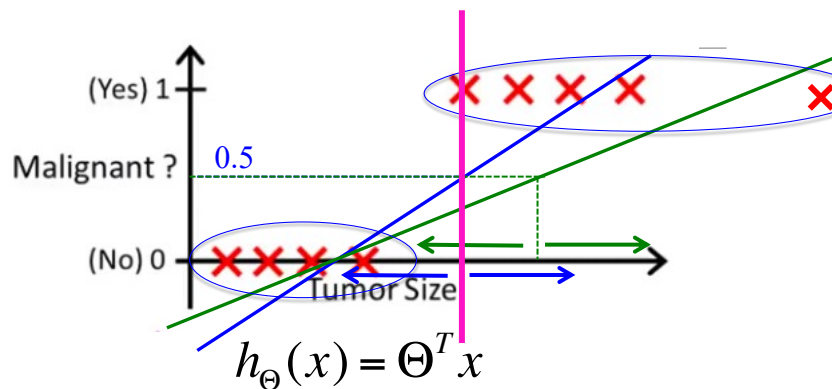
□ Classifieur à seuil (threshold) : output $h_{\Theta}(x)$ à 0.5:

- si $h_{\Theta}(x) \geq 0.5$, prédire $y = 1$
- si $h_{\Theta}(x) < 0.5$, prédire $y = 0$

13

13

Classification

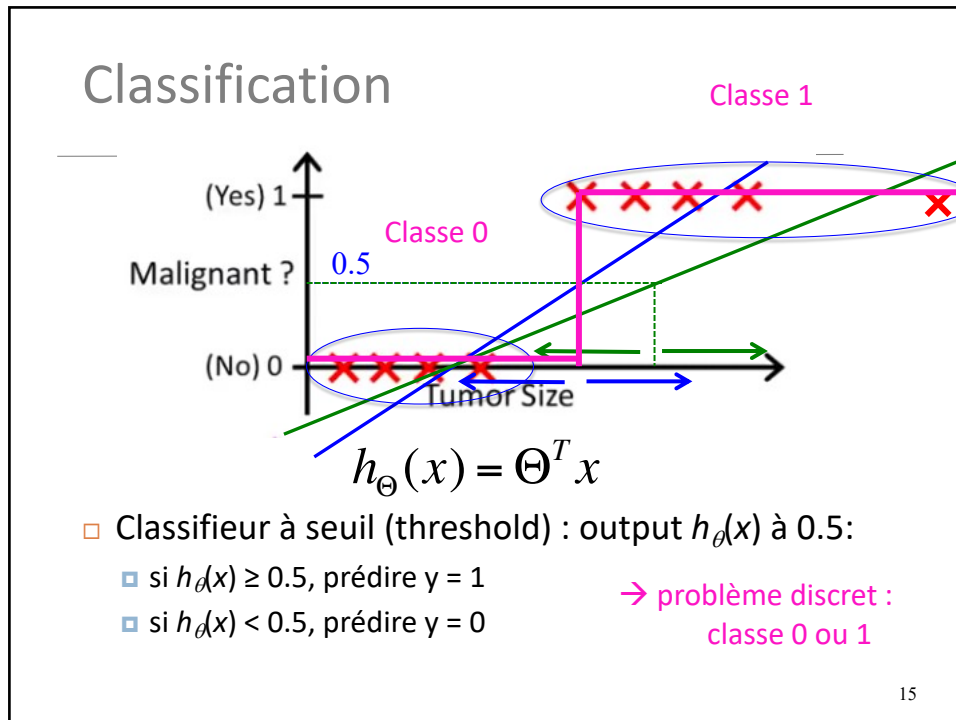


□ Classifieur à seuil (threshold) : output $h_{\Theta}(x)$ à 0.5:

- si $h_{\Theta}(x) \geq 0.5$, prédire $y = 1$
- si $h_{\Theta}(x) < 0.5$, prédire $y = 0$

14

14



15

Régression / Classification

Régression linéaire

$y = 0$ or 1
 mais $h_{\Theta}(x)$ peut être > 1 ou < 0 !

Régression logistique

On veut que $0 \leq h_{\Theta}(x) \leq 1$

16

16

Quelle proposition suivante est juste ?

- Si l'ensemble d'apprentissage (training set) satisfait $0 \leq y^{(i)} \leq 1$ pour tout exemple $(x^{(i)}, y^{(i)})$, alors la prédiction par régression linéaire satisfera aussi :
 $0 \leq h_{\theta}(x) \leq 1$ pour toute valeur de x .
- S'il y a un vecteur de feature x qui prédit parfaitement y , i.e.
 $y = 1$ quand $x \geq c$
et $y = 0$ quand $x < c$ (pour une constante c),
alors la régression linéaire conduira à une erreur de classification qui vaut zéro.
- Aucune des propositions précédentes n'est vraie

17

17

Quelle proposition suivante est juste ?

- Si l'ensemble d'apprentissage (training set) satisfait $0 \leq y^{(i)} \leq 1$ pour tout exemple $(x^{(i)}, y^{(i)})$, alors la prédiction par régression linéaire satisfera aussi :
 $0 \leq h_{\theta}(x) \leq 1$ pour toute valeur de x . False
- S'il y a un vecteur de feature x qui prédit parfaitement y , i.e.
 $\hat{y} = 1$ quand $x \geq c$
et $\hat{y} = 0$ quand $x < c$ (pour une constante c),
Alors la régression linéaire conduira à une erreur de classification qui vaut zéro. True
- Aucune des propositions précédentes n'est vraie

18

18

Classification

Régression linéaire

$y = 0$ ou 1

mais $h_{\theta}(x)$ (prédiction) peut-être > 1 ou < 0 , ce qu'on ne veut pas

Classification utilisant la **Régression Logistique**

$$0 \leq h_{\theta}(x) \leq 1$$

Les prédictions de la régression logistique sont toujours comprises entre 0 et 1

La régression logistique est l'algorithme de classification que l'on applique lorsque l'étiquette y a une valeur discrète.

19

19



Régression logistique

Hypothèse h

20

20

Problème

- Représentation de l'hypothèse : i.e., quelle fonction va-t-on utiliser pour représenter notre prédiction (hypothèse) dans le cadre d'un problème de classification ?

- Modèle de régression logistique

$$0 \leq h_{\theta}(x) \leq 1$$

Hypothèse : les prédictions devraient être entre 0 et 1

21

21

Modèle de régression logistique

- En régression linéaire, on avait : $h_{\theta}(x) = \Theta^T x$
avec

- Droite : $z = \theta_0 + \theta_1 x_1$

Une variable en entrée, une variable prédite en sortie

- Généralisation :

$$z = \Theta^T x = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$$

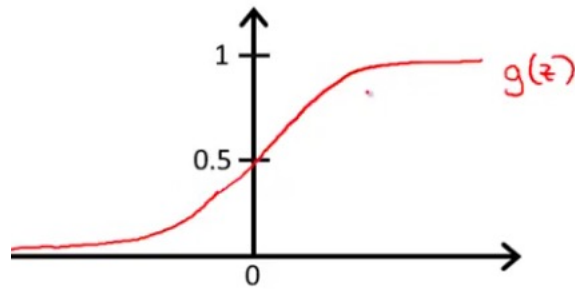
Plusieurs variables en entrée x_1, x_2, x_n , une variable prédite en sortie z

22

22

Modèle de régression logistique

- En régression linéaire, on avait : $h_{\Theta}(x) = \Theta^T x$
avec $z = \Theta^T x = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$



→ sigmoïde : non linéaire
Ramène les valeurs entre 0 et 1

23

23

Modèle de régression logistique

- En régression linéaire, on avait : $h_{\Theta}(x) = \Theta^T x$
avec $z = \Theta^T x = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$
- En régression logistique on transforme le problème ainsi :

$$h_{\Theta}(x) = g(\Theta^T x)$$

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

Où g est une fonction Sigmoid
Encore appelée la **fonction logistique**

24

24

Modèle de régression logistique

- En régression linéaire, on avait : $h_{\Theta}(x) = \Theta^T x$
- En régression logistique on transforme le problème ainsi :

$$h_{\Theta}(x) = g(\Theta^T x)$$

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

$$h_{\Theta}(x) = \frac{1}{1 + e^{-\Theta^T x}}$$

Où g est une fonction Sigmoid
encore appelée **fonction logistique**
 $z = \Theta^T x$

25

25

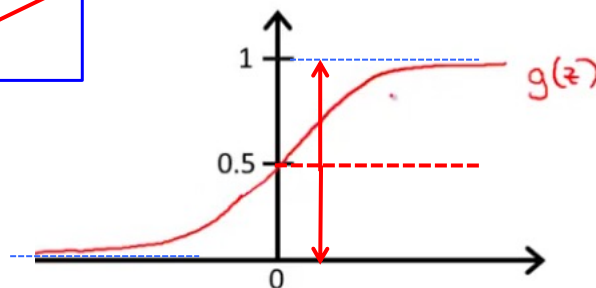
Modèle de régression logistique

$$h_{\Theta}(x) = g(\Theta^T x)$$

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

$$h_{\Theta}(x) = \frac{1}{1 + e^{-\Theta^T x}}$$

g : fonction Sigmoid
fonction logistique



Les valeurs de $g(z)$ sont entre 0 et 1, et celles de $h_{\Theta}(x)$ aussi

26

26

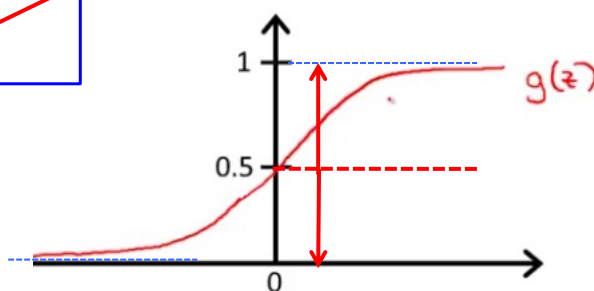
Modèle de régression logistique

$$h_{\Theta}(x) = g(\Theta^T x)$$

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

g : fonction Sigmoid
fonction logistique

$$h_{\Theta}(x) = \frac{1}{1 + e^{-\Theta^T x}}$$



Finalement, on veut trouver les **parameters θ** qui correspondent (fit) le mieux avec nos données, ce qui constituera nos prédictions

27

Interprétation de l'hypothèse de sortie

- $h_{\theta}(x)$ = probabilité estimée que $y = 1$ pour l'input x

28

28

Interprétation de l'hypothèse de sortie

- $h_{\theta}(x)$ = probabilité estimée que $y = 1$ pour l'input x

- Exemple : si
$$x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ tumorSize \end{bmatrix}$$

Si $h_{\theta}(x) = 0.7$ (étant donné les données x de mon patient)

Alors je peux dire à mon patient qu'il y a 70% de chance que sa tumeur soit maligne

29

29

Interprétation de l'hypothèse de sortie

- $h_{\theta}(x)$ = probabilité estimée que $y = 1$ pour l'input x

- Exemple : si
$$x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ tumorSize \end{bmatrix}$$

Si $h_{\theta}(x) = 0.7$ (étant donné les données x de mon patient)

Alors je peux dire à mon patient qu'il y a 70% de chance que sa tumeur soit maligne

- Formellement: $h_{\theta}(x) = P(y = 1 \mid x ; \theta)$ "probabilité que $y = 1$, étant donné x , paramétré par θ

30

30

Interprétation de l'hypothèse de sortie

- $h_{\theta}(x)$ = probabilité estimée que $y = 1$ pour l'input x

- Exemple : si
$$x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ tumorSize \end{bmatrix}$$

Si $h_{\theta}(x) = 0.7$ (étant donné les données x de mon patient)

Alors je peux dire à mon patient qu'il y a 70% de chance que sa tumeur soit maligne

- Formellement: $h_{\theta}(x) = P(y = 1 \mid x ; \theta)$ "probabilité que $y = 1$, étant donné x , paramétré par θ "

L'hypothèse (le modèle) estime la probabilité que la sortie y soit égale à 1.

31

31

Interprétation de l'hypothèse de sortie

- Puisque c'est une tâche de classification, on sait que la sortie doit être discrete : $y = 0$ ou 1

- soit dans l'ensemble d'entraînement (training set)
 - soit pour de nouveaux patients (test set)

32

32

Quelles propositions sont vraies ?

Supposez que l'on veut prédire, à partir d'une donnée x sur une tumeur, si cette tumeur est maligne ($y=1$) ou bénigne ($y=0$). Notre classifieur logistique est tel que, pour notre tumeur spécifique, la sortie est donnée par :

$h_\theta(x) = P(y = 1 \mid x; \theta) = 0.7$, ainsi on estime qu'il y a 70% de chance que cette tumeur soit maligne. Que devrait être notre estimation :

$P(y = 0 \mid x; \theta)$: probabilité pour que la tumeur soit bénigne ?

- ☐ $P(y = 0 \mid x; \theta) = 0.3$
- ☐ $P(y = 0 \mid x; \theta) = 0.7$
- ☐ $P(y = 0 \mid x; \theta) = 0.7^2$
- ☐ $P(y = 0 \mid x; \theta) = 0.3 \times 0.7$

33

33

Quelles propositions sont vraies ?

- ☐ $P(y = 0 \mid x; \theta) + P(y = 1 \mid x; \theta) = 1$

$$P(y = 0) + P(y = 1) = 1$$

- ☐ $\rightarrow P(y = 0 \mid x; \theta) = 1 - P(y = 1 \mid x; \theta) = 1 - 0.7 = 0.3$

34

34

Régression logistique

Frontière de décision
(boundary decision)

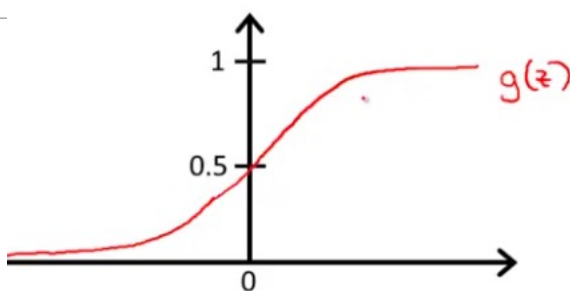
35

35

Régression logistique

$$h_{\Theta}(x) = g(\Theta^T x)$$

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



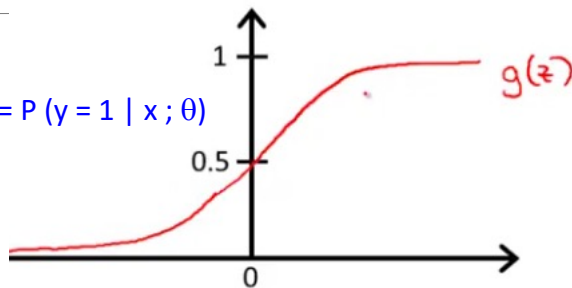
36

36

Régression logistique

$$h_{\theta}(x) = g(\Theta^T x) = P(y = 1 | x; \theta)$$

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



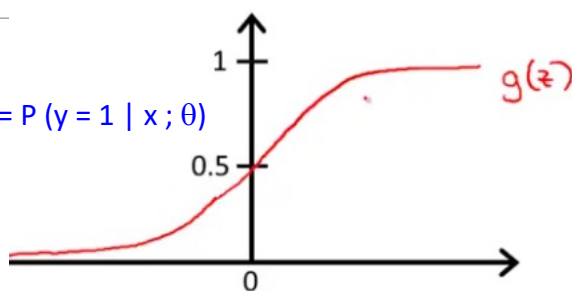
37

37

Régression logistique

$$h_{\theta}(x) = g(\Theta^T x) = P(y = 1 | x; \theta)$$

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



Supposons qu'on prédise $y = 1$ si $h_{\theta}(x) \geq 0.5$

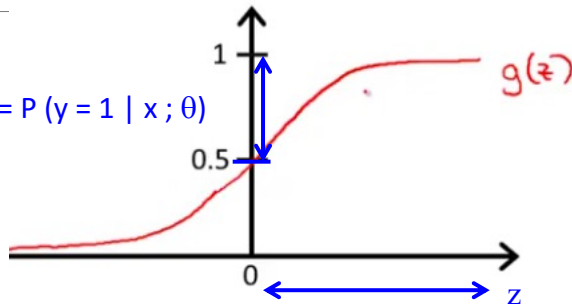
38

38

Régression logistique

$$h_{\theta}(x) = g(\Theta^T x) = P(y = 1 | x; \theta)$$

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



Supposons qu'on prédise $y = 1$ si $h_{\theta}(x) \geq 0.5$ $g(z) \geq 0.5$

quand $z \geq 0$

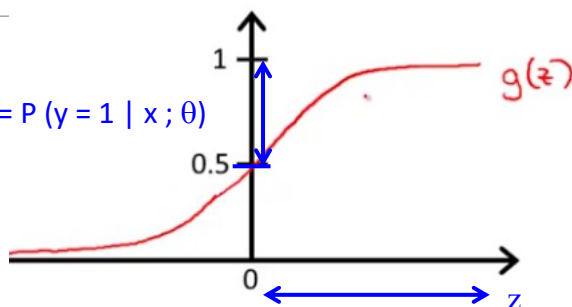
Hypothèse $h_{\theta}(x) = g(\Theta^T x) \geq 0.5$
quand $\theta^T x \geq 0$

39

Régression logistique

$$h_{\theta}(x) = g(\Theta^T x) = P(y = 1 | x; \theta)$$

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



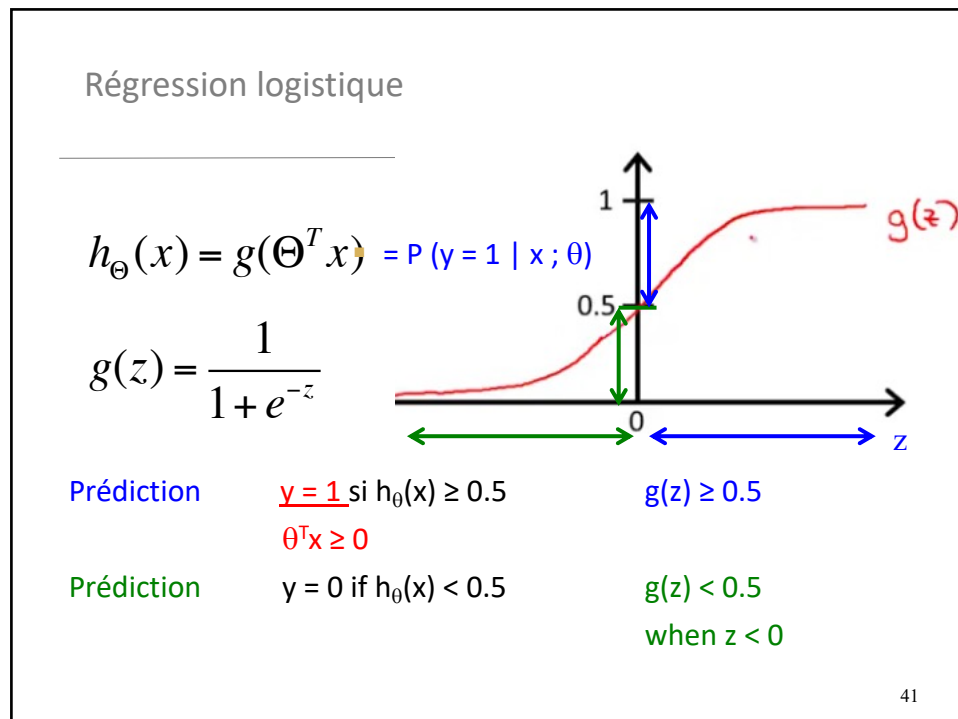
Supposons qu'on prédise $y = 1$ si $h_{\theta}(x) \geq 0.5$ $g(z) \geq 0.5$

$\theta^T x \geq 0$

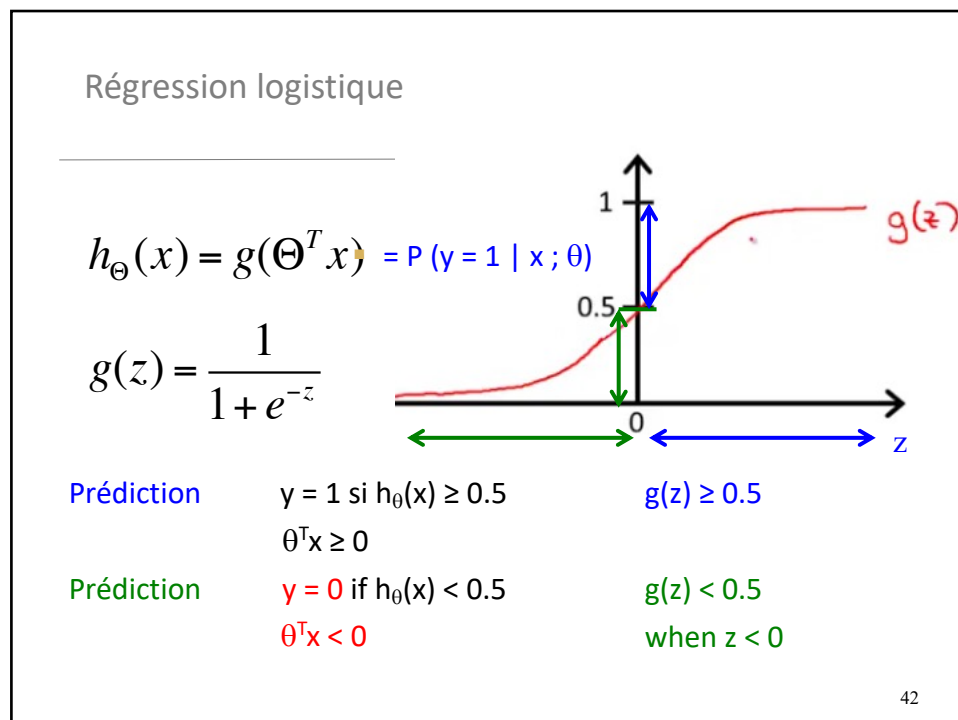
quand $z \geq 0$

Hypothèse $h_{\theta}(x) = g(\Theta^T x) \geq 0.5$
quand $\theta^T x \geq 0$

40

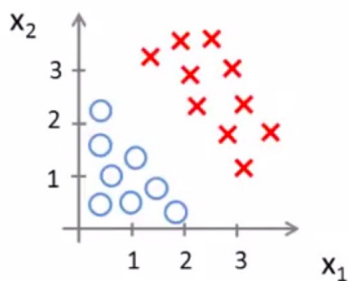


41



42

Frontière de décision

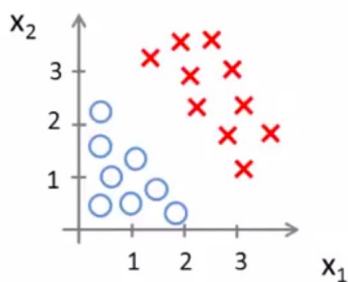


$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

43

43

Frontière de décision



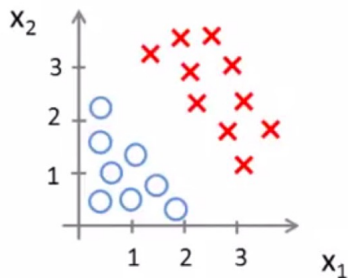
$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2) \quad \theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$$

-3 1 1

44

44

Frontière de décision



$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

$$\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$$

Prédire $y = 1$ quand $-3 + x_1 + x_2 \geq 0$ pour tout exemple x_1, x_2

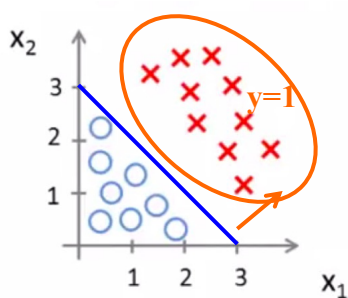
$$\theta^T x$$

ou $x_1 + x_2 \geq 3$

45

45

Frontière de décision



$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

$$\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$$

Prédire $y = 1$ quand $-3 + x_1 + x_2 \geq 0$ pour tout exemple x_1, x_2

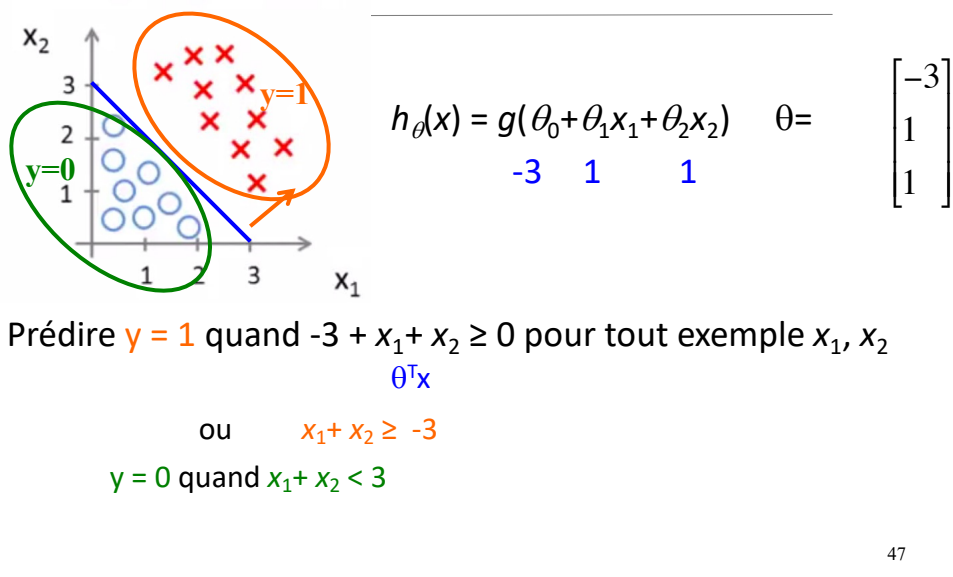
$$\theta^T x$$

ou $x_1 + x_2 \geq 3$

46

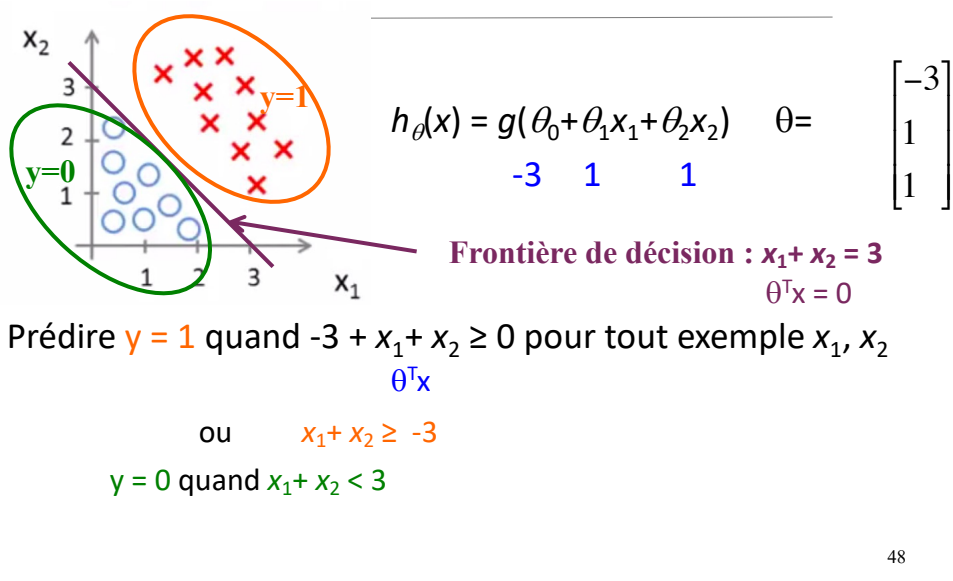
46

Frontière de décision



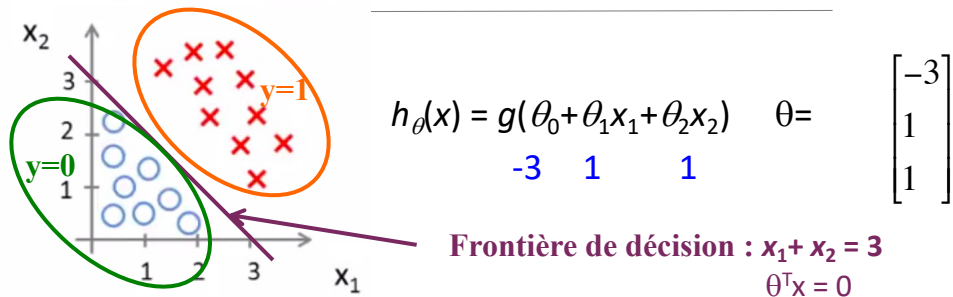
47

Frontière de décision



48

Frontière de décision



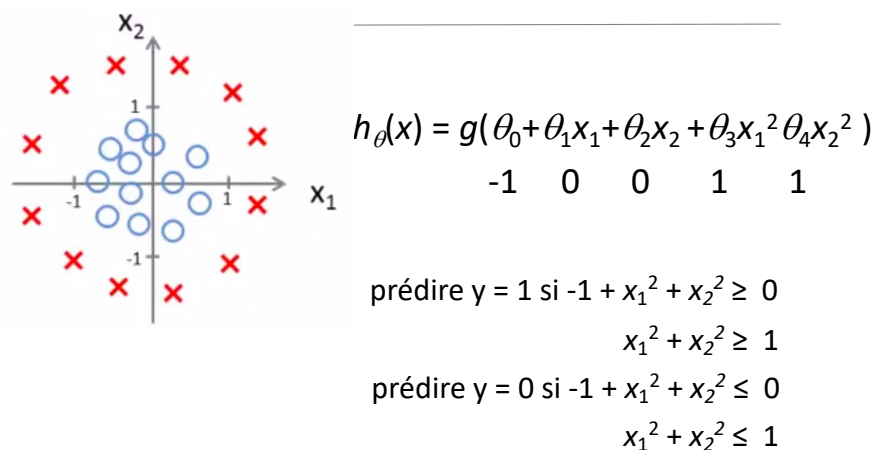
Frontière de décision : propriété de l'hypothèse incluant les paramètres $\theta_0, \theta_1, \theta_2$ et non propriété du dataset.

On verra plus tard comment faire “fitter” les paramètres aux données d’entraînement.

49

49

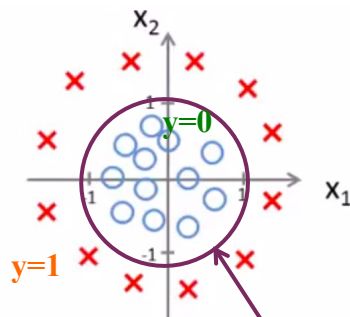
Frontière de décision non linéaire



50

50

Frontière de décision non linéaire



$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

$$\begin{matrix} & -1 & 0 & 0 & 1 & 1 \end{matrix}$$

prédire $y = 1$ si $-1 + x_1^2 + x_2^2 \geq 0$

$$x_1^2 + x_2^2 \geq 1$$

prédire $y = 0$ si $-1 + x_1^2 + x_2^2 \leq 0$

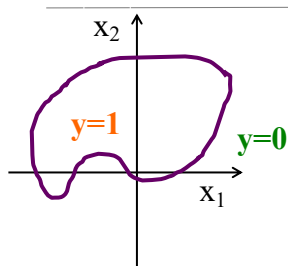
$$x_1^2 + x_2^2 \leq 1$$

frontière de décision : $x_1^2 + x_2^2 = 1$

51

51

Frontière de décision non linéaire



$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_4 x_1^2 x_2 + \theta_5 x_1^2 x_2^2 + \theta_6 x_1^3 x_2 + \dots)$$

52

52

□ Hypothèse : $h_{\theta}(x) = \theta_0 + \theta_1 x$

□ Paramètres : θ_0, θ_1

□ Fonction coût : $J(\theta_0, \theta_1) = \frac{1}{2m} \cdot \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$

□ But : minimiser $J(\theta_0, \theta_1)$
 θ_0, θ_1

53

53



Régression logistique

Fonction coût
pour aller plus loin

54

54

Fonction coût

- Training set : $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$
 m exemples

$$x \in \begin{bmatrix} x_0 \\ x_1 \\ \dots \\ x_n \end{bmatrix} \quad x_0=1, \quad y \in \{0,1\}$$

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

Comment choisir (fitter) les paramètres θ ?

55

55

Fonction coût

- Régression logistique

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

$$J(\theta) = \frac{1}{m} \cdot \sum_{i=1}^m \frac{1}{2} (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

$$\text{Cost}(h_{\theta}(x^{(i)}), y^{(i)})$$

56

56

Fonction coût

□ Régression logistique

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

$$J(\theta) = \frac{1}{m} \cdot \sum_{i=1}^m \frac{1}{2} (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

Or, $Cost(h_{\theta}(x), y) = \frac{1}{2} (h_{\theta}(x) - y)^2$

n'est pas une fonction convexe car h n'est pas linéaire (sigmoïde)

$$Cost(h_{\theta}(x^{(i)}), y^{(i)}) \quad ?$$

57

57

Fonction coût

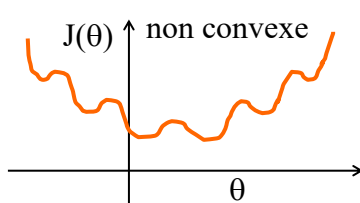
□ Régression logistique

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

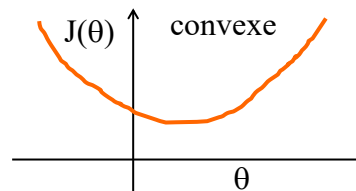
$$J(\theta) = \frac{1}{m} \cdot \sum_{i=1}^m \frac{1}{2} (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

$$Cost(h_{\theta}(x^{(i)}), y^{(i)})$$

$$Cost(h_{\theta}(x), y) = \frac{1}{2} (h_{\theta}(x) - y)^2 \quad \text{non convexe}$$



minimum global : non
(gradient descent)



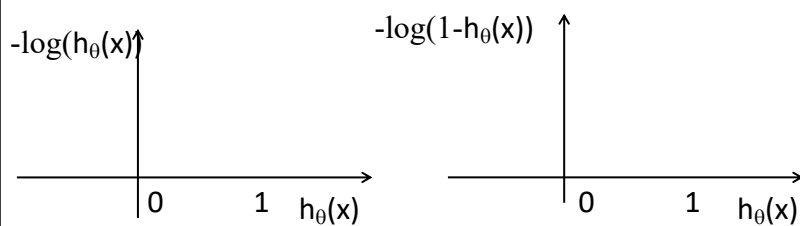
minimum global : oui

58

58

Régression logistique

$$Cost(h_{\theta}(x), y) = \begin{cases} -\log(h_{\theta}(x)) & \text{if } y = 1 \\ -\log(1 - h_{\theta}(x)) & \text{if } y = 0 \end{cases}$$

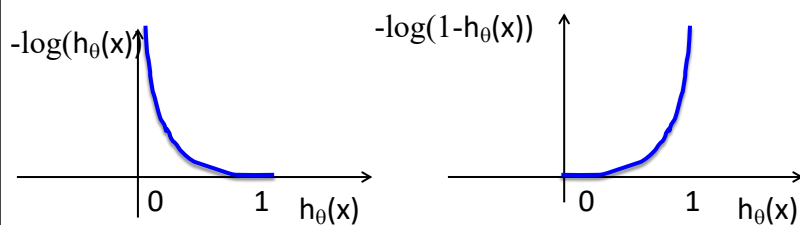


59

59

Régression logistique

$$Cost(h_{\theta}(x), y) = \begin{cases} -\log(h_{\theta}(x)) & \text{if } y = 1 \\ -\log(1 - h_{\theta}(x)) & \text{if } y = 0 \end{cases}$$

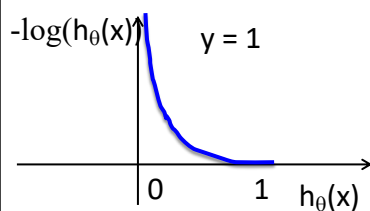


60

60

Régression logistique- Fonction coût

$$Cost(h_{\theta}(x), y) = \begin{cases} -\log(h_{\theta}(x)) & \text{if } y = 1 \\ -\log(1 - h_{\theta}(x)) & \text{if } y = 0 \end{cases}$$



Coût = 0 si $h_{\theta}(x) \rightarrow 1$ et $y = 1$
 coût minimum si bonne prédiction
 on prédit ($P(y=1|x; \theta) = 1$), et $y = 1$

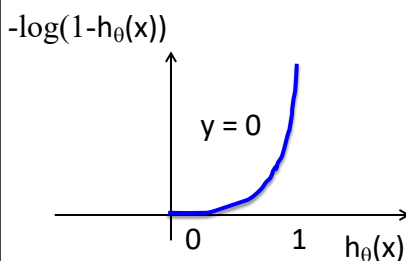
Coût $\rightarrow \infty$ si $h_{\theta}(x) \rightarrow 0$ et $y = 1$
 mauvaise prédiction : on pénalise en
 mettant un coût très important
 on prédit ($P(y=1|x; \theta) = 0$), et $y = 1$

61

61

Régression logistique- Fonction coût

$$Cost(h_{\theta}(x), y) = \begin{cases} -\log(h_{\theta}(x)) & \text{if } y = 1 \\ -\log(1 - h_{\theta}(x)) & \text{if } y = 0 \end{cases}$$



Coût = 0 si $h_{\theta}(x) \rightarrow 0$ et $y = 0$
 coût minimum si bonne prédiction
 on prédit ($P(y=0|x; \theta) = 0$), et $y = 0$

Coût $\rightarrow \infty$ si $h_{\theta}(x) \rightarrow 1$ et $y = 0$
 mauvaise prédiction : on pénalise en
 mettant un coût très important
 on prédit ($P(y=0|x; \theta) = 1$), et $y = 0$

62

62

Régression logistique

Fonction coût simplifiée et descente de gradient

63

63

Fonction coût

- Fonction coût de la régression logistique

$$J(\theta) = \frac{1}{m} \cdot \sum_{i=1}^m \text{Cost}(h_{\theta}(x^{(i)}), y^{(i)})$$
$$\text{Cost}(h_{\theta}(x), y) = \begin{cases} -\log(h_{\theta}(x)) & \text{if } y = 1 \\ -\log(1 - h_{\theta}(x)) & \text{if } y = 0 \end{cases}$$

- Note: on a TOUJOURS $y = 0$ ou 1

64

64

Fonction coût

$$Cost(h_{\theta}(x), y) = \begin{cases} -\log(h_{\theta}(x)) & \text{if } y = 1 \\ -\log(1 - h_{\theta}(x)) & \text{if } y = 0 \end{cases}$$

$$Cost(h_{\theta}(x), y) = -y \cdot \log(h_{\theta}(x)) - (1 - y) \cdot \log(1 - h_{\theta}(x))$$

- si $y = 1$: $Cost(h_{\theta}(x), y) = -\log(h_{\theta}(x))$
- si $y = 0$: $Cost(h_{\theta}(x), y) = -(1 - y) \cdot \log(h_{\theta}(x))$

65

65

Descente du gradient

$$Cost(h_{\theta}(x), y) = -y \cdot \log(h_{\theta}(x)) - (1 - y) \cdot \log(1 - h_{\theta}(x))$$

$$J(\theta) = -\frac{1}{m} \cdot \left[\sum_{i=1}^m y^{(i)} \cdot \log(h_{\theta}(x^{(i)})) + (1 - y^{(i)}) \cdot \log(1 - h_{\theta}(x^{(i)})) \right]$$

- Pourquoi cette fonction coût ?
 - ▣ Fonction **convexe** !
 - ▣ Se ramène à une estimation du maximum de vraisemblance (statistique)

66

66

Descente du gradient

$$Cost(h_{\theta}(x), y) = -y \cdot \log(h_{\theta}(x)) - (1 - y) \cdot \log(1 - h_{\theta}(x))$$

$$J(\theta) = -\frac{1}{m} \cdot \left[\sum_{i=1}^m y^{(i)} \cdot \log(h_{\theta}(x^{(i)})) + (1 - y^{(i)}) \cdot \log(1 - h_{\theta}(x^{(i)})) \right]$$

67

67

Descente du gradient

$$Cost(h_{\theta}(x), y) = -y \cdot \log(h_{\theta}(x)) - (1 - y) \cdot \log(1 - h_{\theta}(x))$$

$$J(\theta) = -\frac{1}{m} \cdot \left[\sum_{i=1}^m y^{(i)} \cdot \log(h_{\theta}(x^{(i)})) + (1 - y^{(i)}) \cdot \log(1 - h_{\theta}(x^{(i)})) \right]$$

- Pour trouver les paramètres θ automatiquement, il faut minimiser $J(\theta)$ selon θ :

$$\min_{\theta} J(\theta)$$

68

68

Descente du gradient

$$\text{Cost}(h_{\theta}(x), y) = -y \cdot \log(h_{\theta}(x)) - (1 - y) \cdot \log(1 - h_{\theta}(x))$$

$$J(\theta) = -\frac{1}{m} \cdot \left[\sum_{i=1}^m y^{(i)} \cdot \log(h_{\theta}(x^{(i)})) + (1 - y^{(i)}) \cdot \log(1 - h_{\theta}(x^{(i)})) \right]$$

- Pour trouver les paramètres θ automatiquement, il faut minimiser $J(\theta)$ selon θ :

$$\min_{\theta} J(\theta)$$

- Prédiction : étant donné x (nouvelle entrée), on calcule la prédiction :

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

69

69

Descente du gradient

$$J(\theta) = -\frac{1}{m} \cdot \left[\sum_{i=1}^m y^{(i)} \cdot \log(h_{\theta}(x^{(i)})) + (1 - y^{(i)}) \cdot \log(1 - h_{\theta}(x^{(i)})) \right]$$

- On veut minimiser $J(\theta)$ suivant θ

Algorithme

Repeat {

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

}
Mise à jour des paramètres θ_j
simultanément

70

70

Descente du gradient

$$J(\theta) = -\frac{1}{m} \cdot \left[\sum_{i=1}^m y^{(i)} \cdot \log(h_{\theta}(x^{(i)})) + (1 - y^{(i)}) \cdot \log(1 - h_{\theta}(x^{(i)})) \right]$$

□ On veut minimiser $J(\theta)$ suivant θ

Repeat {

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^m (h_{\theta} x^{(i)} - y^{(i)}) x_j^{(i)}$$

$$\theta_j = \theta_j - \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) \cdot x^{(i)}$$

}

Algorithme : identique à celui de la régression linéaire

71

Descente du gradient

MAIS la fonction hypothèse $h_{\theta}(x)$ a changé !

Repeat {

$$\theta_j = \theta_j - \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) \cdot x^{(i)} \quad \text{for } j = 0, 1, \dots, n$$

□ Régression linéaire : $h_{\theta}(x) = \theta^T x$

□ Régression logistique :

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

72

72



Logistic Regression

Optimization avancée

73

73

Algorithme d'optimization

- Fonction coût $J(\theta)$; on veut $\min_{\theta} J(\theta)$
- Etant donné θ , on a du code qui calcule :

$$J(\theta)$$

$$\frac{\partial}{\partial \theta_j} J(\theta)$$

pour $j = 0, 1, \dots, n$

- Descente de gradient

Repeat {

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

}

74

74

Optimization algorithm

- Etant donné θ , on a du code qui calcule :

$$J(\theta)$$

$$\frac{\partial}{\partial \theta_j} J(\theta)$$

pour $j = 0, 1, \dots, n$

Algorithmes d'optimization

- Gradient descent

- Conjugate gradient

- BFGS

- L-BFGS

...

Avantages

- plus rapide que GD

Inconvénients

- plus complexe

75

75



Régression logistique

Classification multiclasse :
One-versus all

76

76

Classification multiclasse

Exemples

- Tag Email : Travail, Amis, Famille

$y=1$ $y=2$ $y=3$

- Diagnostic médical : Sain, Rhume, Grippe, Autre

$y=1$ $y=2$ $y=3$ $y=4$

- Météo : Ensoleillé, nuageux, pluvieux, neigeux, venté

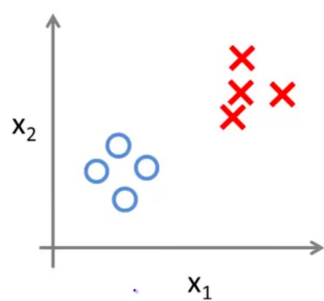
$y=1$ $y=2$ $y=3$ $y=4$ $y=5$

77

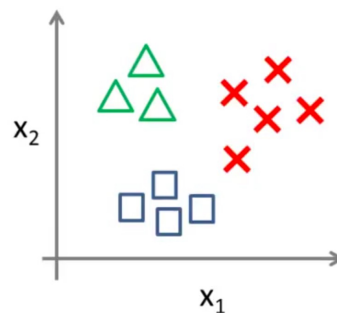
77

Classification multiclasse

- Classification binaire



- Classification multi-classe

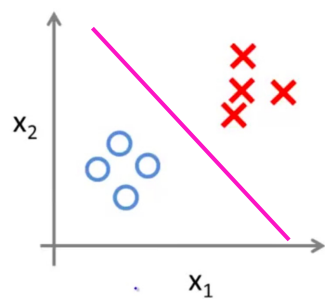


78

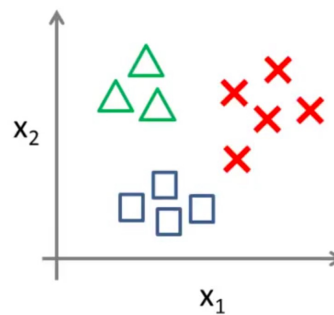
78

Classification multiclasse

Classification binaire



Classification multi-classe

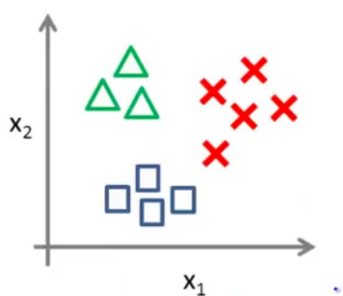





79

79

Classification multiclasse

One-vs-all



Class 1: 
 Class 2: 
 Class 3: 

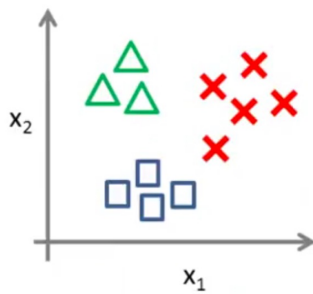
- On transforme le problème de classification en 3 problèmes de séparation binaire à 2 classes.

80

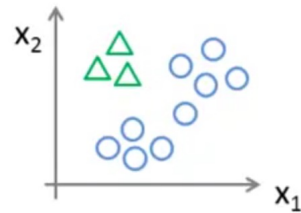
80

Classification multiclasse

One-vs-all



Class 1: \triangle
 Class 2: \square
 Class 3: \times



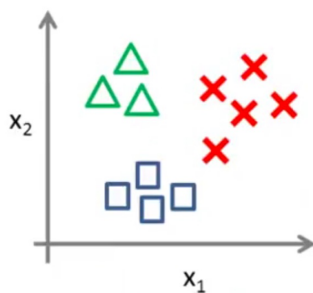
- On crée un nouveau "fake" training set, où les classes 2 et 3 sont affectées à la classe négative, et la classe 1 est affectée à la classe positive.

81

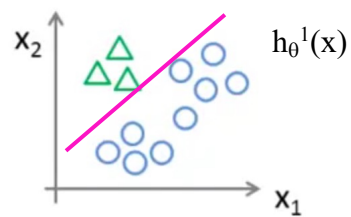
81

Classification multiclasse

One-vs-all



Class 1: \triangle
 Class 2: \square
 Class 3: \times



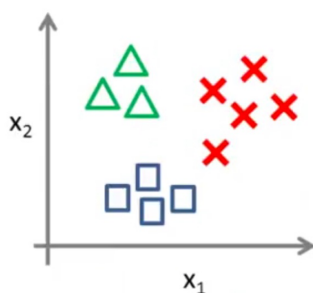
- On entraîne un classifieur de régression logistique standard sur ce problème à 2 classes : $h_0^1(x)$

82

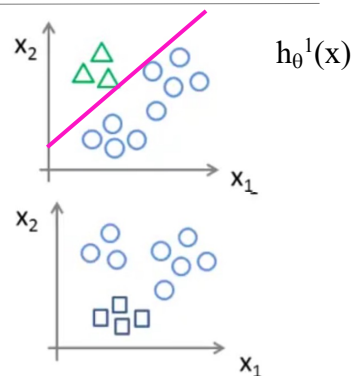
82

Classification multiclasse

One-vs-all



Class 1: \triangle
 Class 2: \square
 Class 3: \times

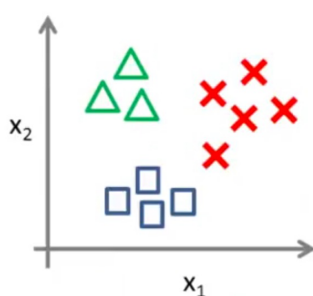


83

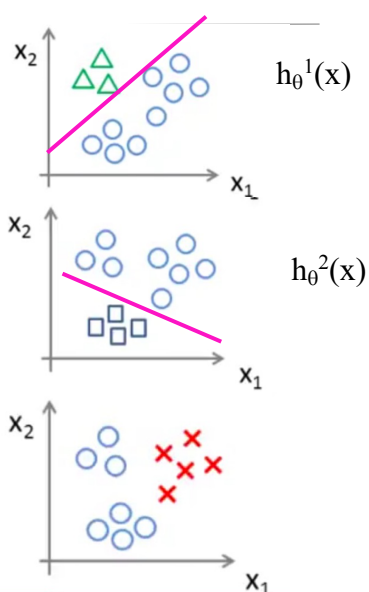
83

Classification multiclasse

One-vs-all

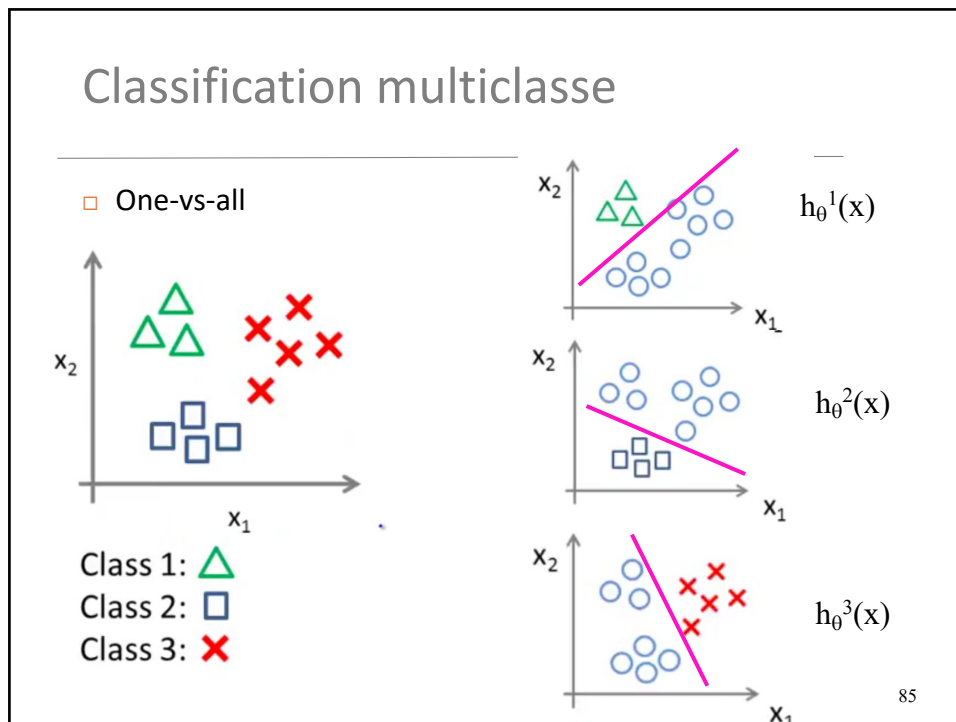


Class 1: \triangle
 Class 2: \square
 Class 3: \times

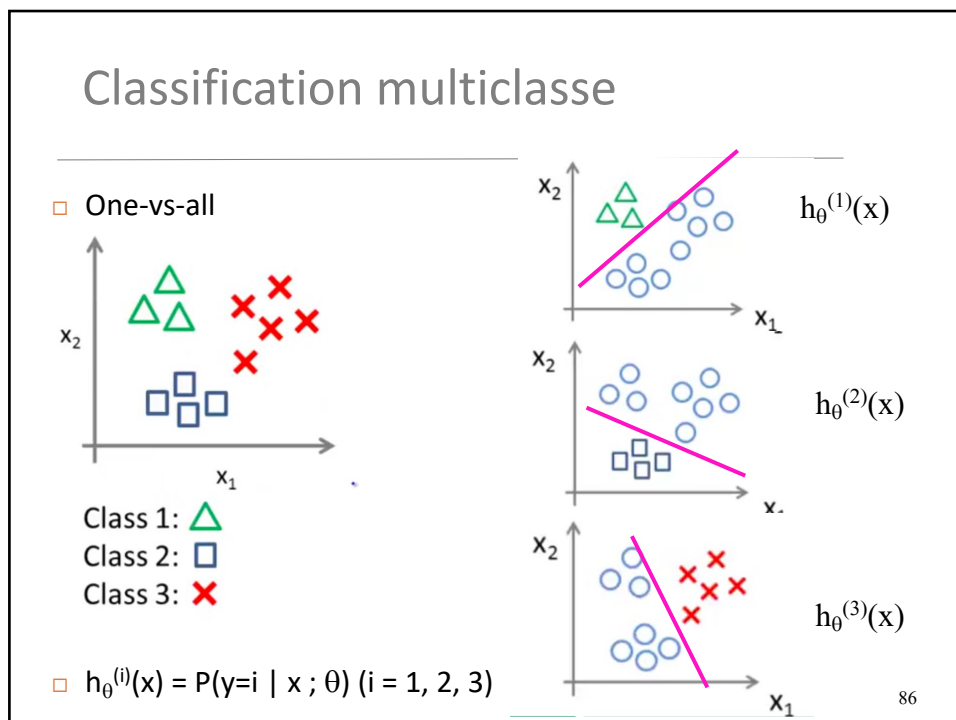


84

84



85



86

Classification multiclasse : One-vs-all

- On entraîne un classifieur de régression logistique $h_{\theta}^{(i)}(x)$ pour chaque classe i afin de prédire la probabilité pour que $y = i$
- Sur une nouvelle entrée x , on effectue une prédiction, et on choisit la classe i qui maximise $h_{\theta}^{(i)}(x)$:

$$\max_i h_{\theta}^{(i)}(x)$$

87

87

Classification multiclasse : One-vs-all

- Supposez que vous avez un problème de classification multiclasse à k classes (y appartient à $\{1, 2, \dots, k\}$). En utilisant la méthode one-vs-all, combien de classifieurs de régression logistique aurez vous besoin d'entraîner ?
 - $k-1$
 - k
 - $k+1$
 - Approximativement $\log_2(k)$

88

88