Rappel cours précédents I

Définition

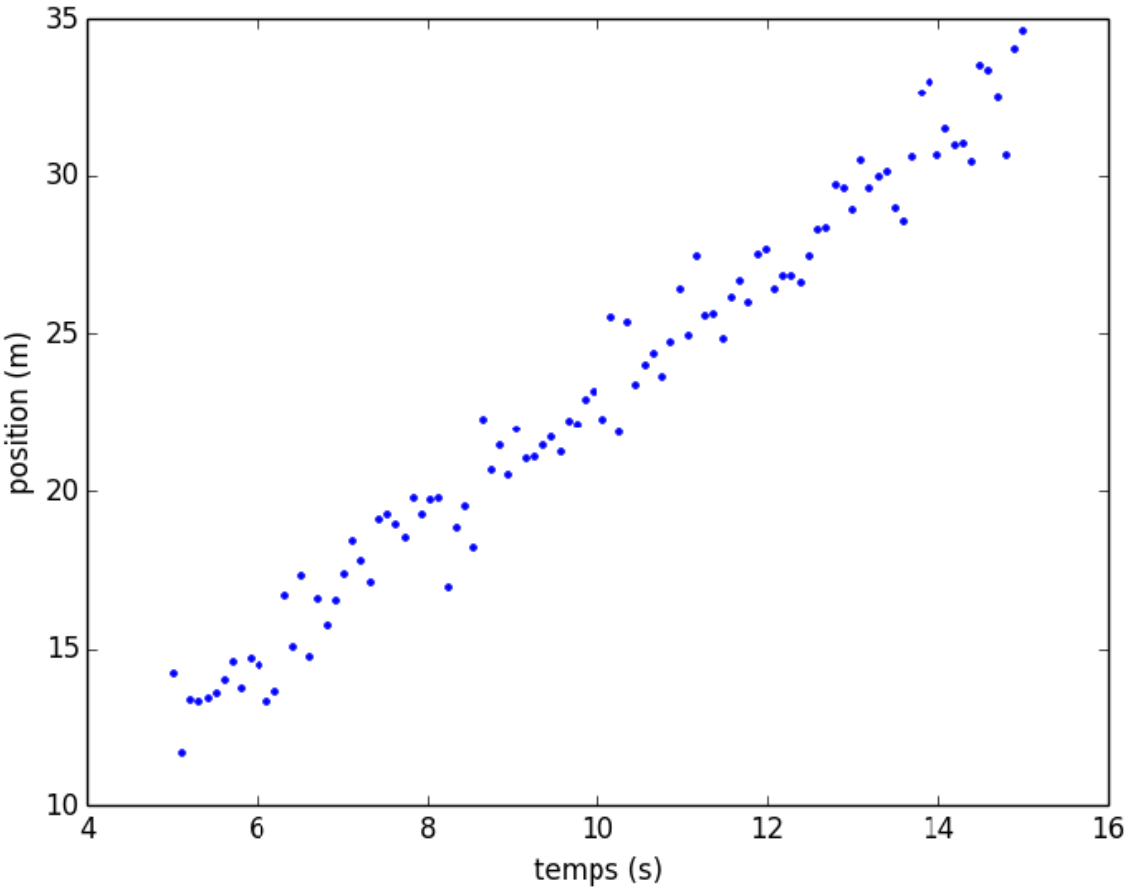
Etant donné un ensemble $\mathcal{D}$ de N données sous forme de couples $\{x_i, y_i = f(x_i)\}_{0<i<N+1}$ avec $x_i \in \mathcal{X}$ et $y_i \in \mathcal{Y}$ (souvent $\mathcal{X} = \mathbb{R}^m$ et $\mathcal{Y} = \mathbb{R}$), il s'agit de trouver une fonction $f_N(x) \in \mathcal{H}$ qui permette d'expliquer et de prédire la relation entre des entrées quelconques (x) et les sorties (y) correspondantes. $\mathcal{H}$ est appelé l'espace d'hypothèse.

Oracle

L'oracle connaît la relation, il a donné les couples de la base. Il agit selon une loi de probabilité $P(Y|X)$

Rappel cours précédents II

Si sortie continue : pas de problème



Rappel cours précédents III

Choix de la représentation

- On se choisit une représentation linéaire :

$$y = \sum_m \theta_m x_m = \theta^T \mathbf{x}$$

- On cherche les paramètres θ grâce aux données $\{\mathbf{x}^{(i)}, y^{(i)}\}$.

Induction

On choisit de minimiser l'erreur quadratique moyenne (MSE) :

$$J(\theta) = \frac{1}{N} \sum_{i=1}^{N} (y^{(i)} - \theta^T \mathbf{x}^{(i)})^2$$

Rappel cours précédents IV

Solution

- Pour minimiser, on dérive et on annule :

$$\nabla_\theta \frac{1}{N} \sum_{i=1}^{N} (y^{(i)} - \theta^T \mathbf{x}^{(i)})^2 = 0$$

- La solution est :

$$\theta = (\sum_i \mathbf{x}^{(i)} \mathbf{x}^{(i)T})^{-1} \sum_i \mathbf{x}^{(i)} y^{(i)}$$

Rappel cours précédents V

Version matricielle

$$J(\theta) = \frac{1}{N}\|(Y - X^T\theta)\|^2$$

$$\theta = (XX^T)^{-1}(XY)$$

Descente de gradient

$$\theta^{(t+1)} = \theta^{(t)} + \alpha_t \frac{1}{N} X \left(Y - X^T\theta^{(t)}\right)$$

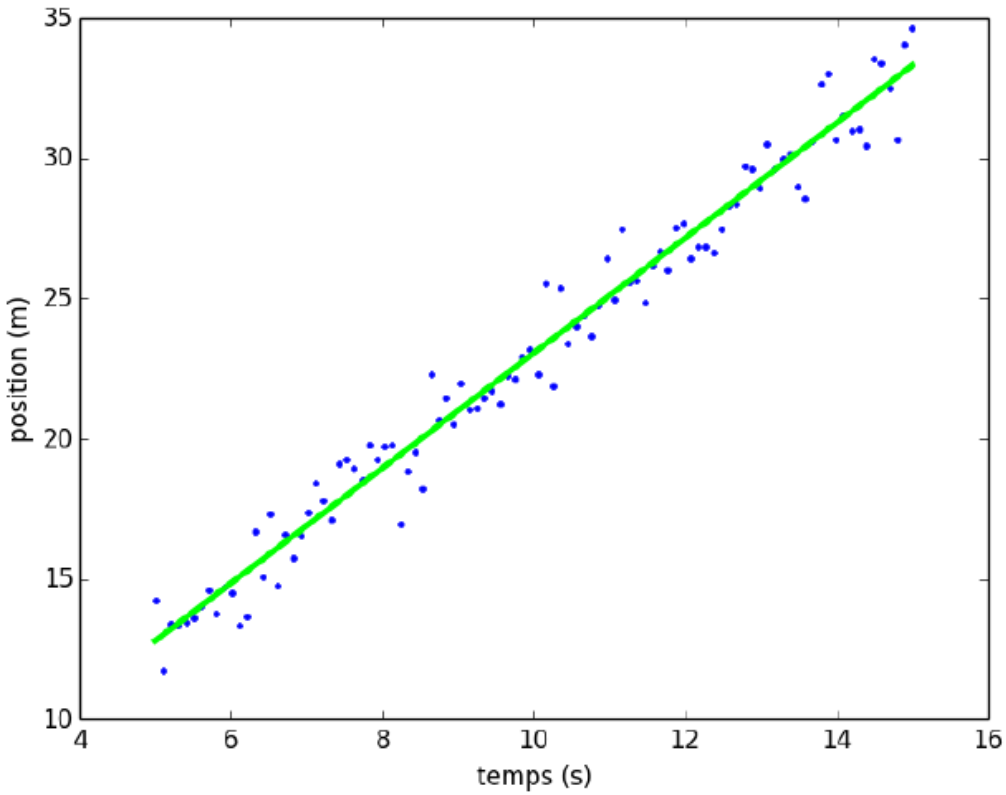
$$\theta^{(t+1)} = \theta^{(t)} + \alpha_t \left(x^{(i)} \left(y^{(i)} - \theta^{(t)^T}\mathbf{x}^{(i)}\right)\right)$$

Pourquoi ça marche ?

- Fonction de coût convexe et dérivable
- La dérivée donne le minimum

Rappel cours précédents VI

Résultat



Rappel cours précédents VII

Fonction de perte (*Loss function*)

$\mathcal{L}(f(x), y)$: coût de la décision associée à x par une fonction f quelconque, sachant que la bonne décision était y.
Par exemple :

- Coût quadratique : $(y_i - f(x_i))^2$

Risque Réel

Le risque associé à la fonction de perte est

$$R(f) = \int_{\mathcal{X}\times\mathcal{Y}} \mathcal{L}(f(x), y) dP(x, y)$$

C'est donc l'espérance de la fonction de coût.

$$R(f) = E[\mathcal{L}(f(x), y)] = \mu_{\mathcal{L}(f(x),y)}$$

Rappel cours précédents VIII

Risque empirique

$$R_N(f) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f(x_i), y_i)).$$

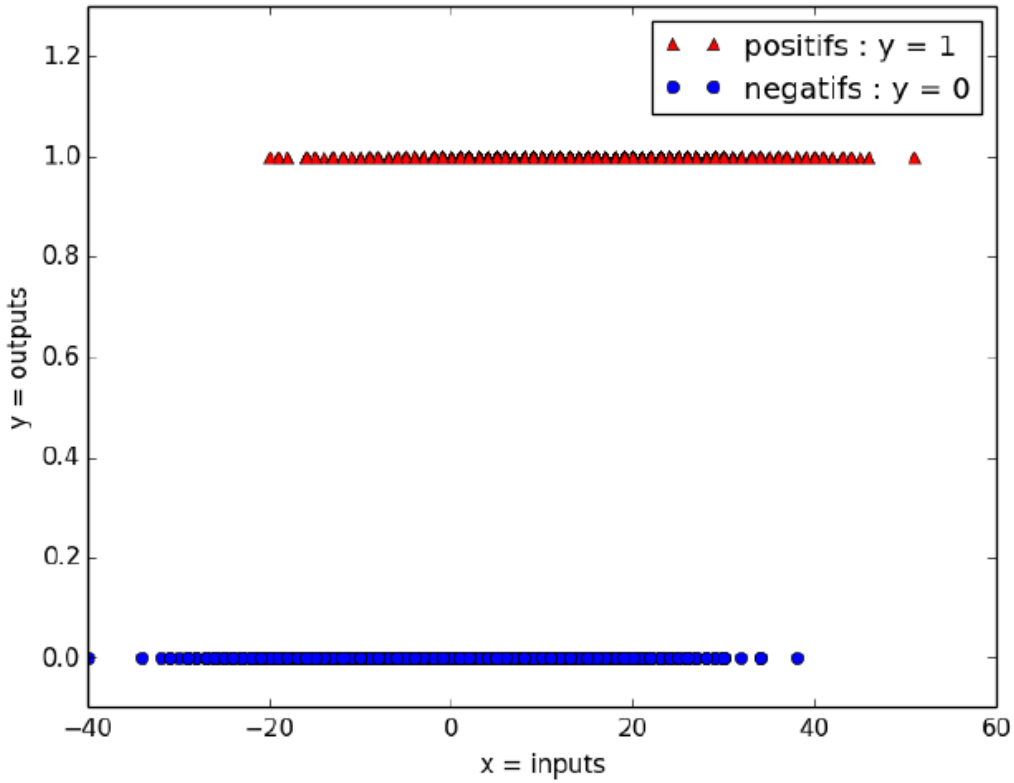
Par exemple pour la fonction de perte quadratique :

$$R_N(f) = \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i))^2 .$$

On a donc minimisé le risque empirique pour la fonction de perte quadratique !

Classification $\neq$ Régression

La sortie est discrète (ici, binaire) : $\mathcal{Y} = \{0, 1\}$



La régression ne peut pas fonctionner !

Idée : régresser $P(Y|X)$ I

Probabilité conditionnelle

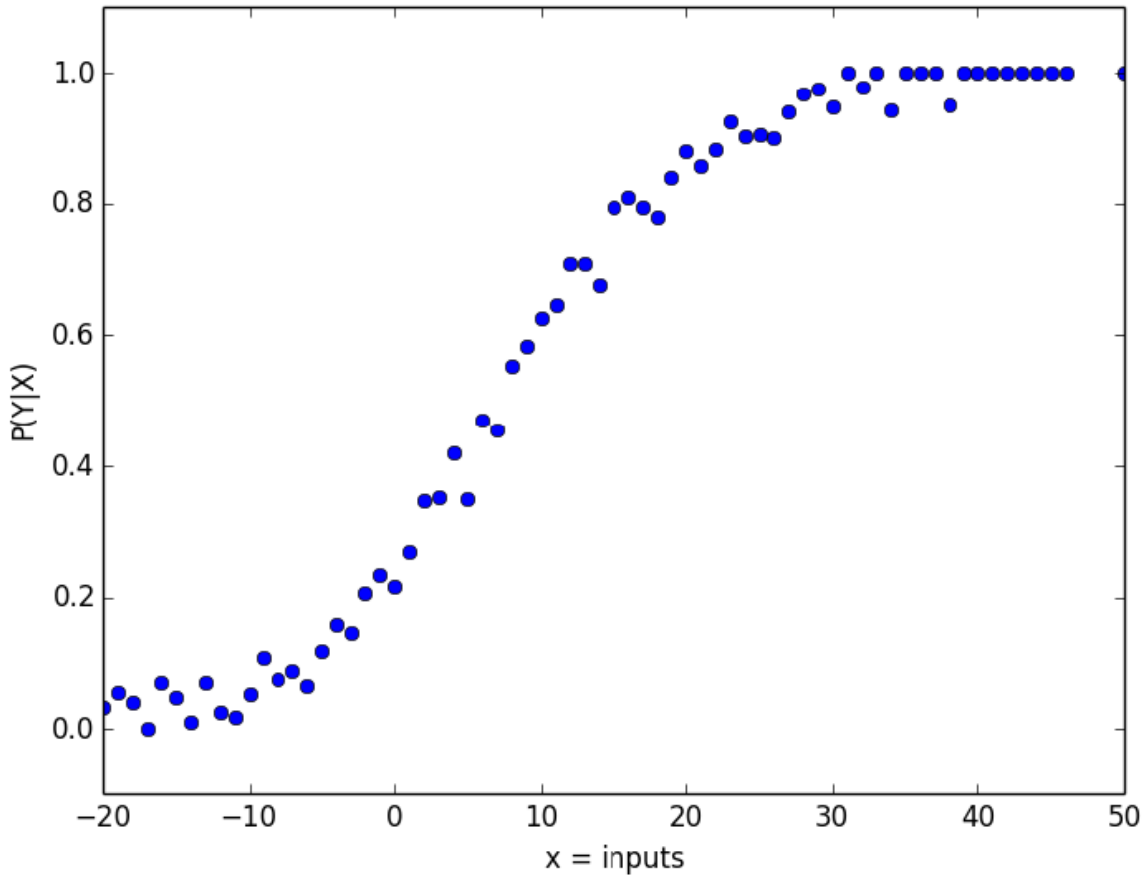
Par définition :

$$P(Y|X) = \frac{P(X,Y)}{P(X)}$$

On peut estimer $P(Y = 0|X)$ par un processus de comptage :

$$\hat{P}(Y = 1|X = x) = \frac{count(X = x, Y = 1)}{count(X = x)}$$

Idée : régresser $P(Y|X)$ II



Loi de Bayes I

Par défintion

$$P(X, Y) = P(X|Y)P(Y)$$
$$P(X, Y) = P(Y|X)P(X)$$

Loi de Bayes

$$\underbrace{P(Y|X)}_{a\ posteriori} = \overbrace{P(X|Y)}^{vraisemblance} \cdot \frac{\overbrace{P(Y)}^{a\ priori}}{\underbrace{P(X)}_{a\ priori}}$$

Remarque

Aussi appelée Règle de Bayes ou Théorème de Bayes.

Loi de Bayes II

$$\text{Loi de Bayes : } P(h|e) = \frac{P(e|h)P(h)}{P(e)}$$

Interprétations

- **Diagnostic** : à partir de connaissances sur la causalité et sur la probabilité a priori d'événements, on peut inférer une cause à partir d'effets et donc raisonner dans l'incertitude.
- **Mise à jour** de l'état de croyance sur le monde : à partir de l'observation d'un événement, on peut changer l'état croyance en conséquence.

Exemple : Diagnostic I

Problème du diagnostic

En général on cherche la probabilité d'une cause après observation d'un effet ou un symptôme. Or, on connaît plus souvent les effets ou les symptômes provoqués par les causes.

Exemple

On sait qu'une méningite provoque un mal de nuque. Si on observe un mal de nuque, doit on diagnostiquer une méningite ?

Exemple : Diagnostic II

Données

- Probabilité d'observer une méningite si pas d'épidémie : 1/50000
- Probabilité d'observer une douleur à la nuque : 0,05
- Probabilité qu'une méningite provoque une douleur à la nuque : 0,5

Solution

$$P(meningite|douleur) = \frac{P(douleur|meningite)P(meningite)}{P(douleur)} = \frac{0.5*1/50000}{0.05} = 0.0002$$

rem : il faut tenir compte des proba *a priori* !

Marginalisation

Marginalisation

La probabilité marginale ou *a priori* d'un événement est la somme des probabilités d'occurrence de cet événement conjointement avec tous les autres :

$$P(X) = \sum_{y_i} P(X, y_i) = \sum_{y_i} P(X|y_i)P(y_i)$$

Autres formes de la loi de Bayes

$$P(Y|X) = \frac{P(X|Y)P(Y)}{\sum_i P(X|y_i)P(y_i)}$$

Fonction logistique ou sigmoïde I

Marginalisation pour la classification binaire

$$P(X) = P(X|Y=1) \cdot P(Y=1) + P(X|Y=0) \cdot P(Y=0)$$

Règle de Bayes

$$P(Y=1|X) = \frac{P(X|Y=1) \cdot P(Y=1)}{P(X|Y=1) \cdot P(Y=1) + P(X|Y=0) \cdot P(Y=0)}$$

Fonction logistique ou sigmoïde II

$$P(Y=1|X) = \frac{1}{1 + \frac{P(X|Y=0)}{P(X|Y=1)} \cdot \frac{P(Y=0)}{P(Y=1)}}$$

On pose : $f(X) = \log \frac{P(X|Y=1)}{P(X|Y=0)} + \log \frac{P(Y=1)}{P(Y=0)}$

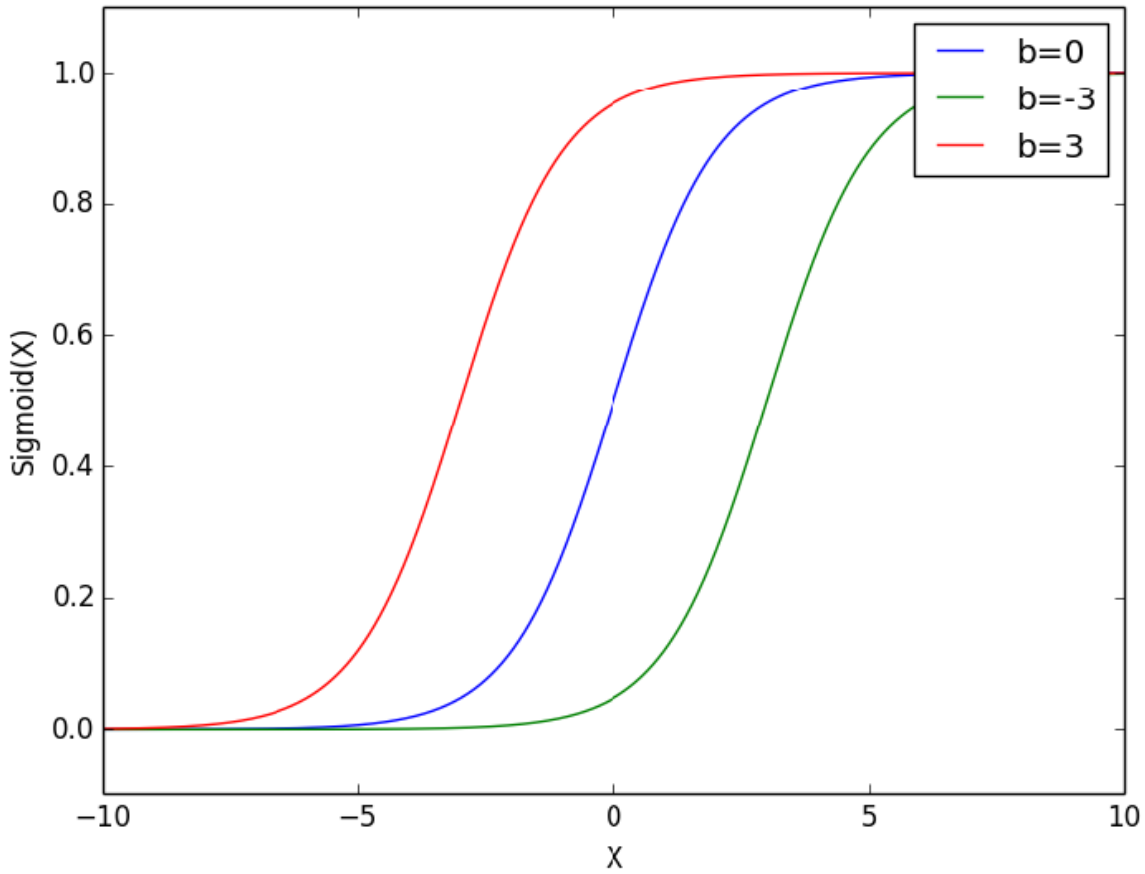
Fonction logistique ou sigmoïde σz

$$P(Y=1|X) = \sigma(f(X)) = \frac{1}{1+e^{-f(X)}}$$

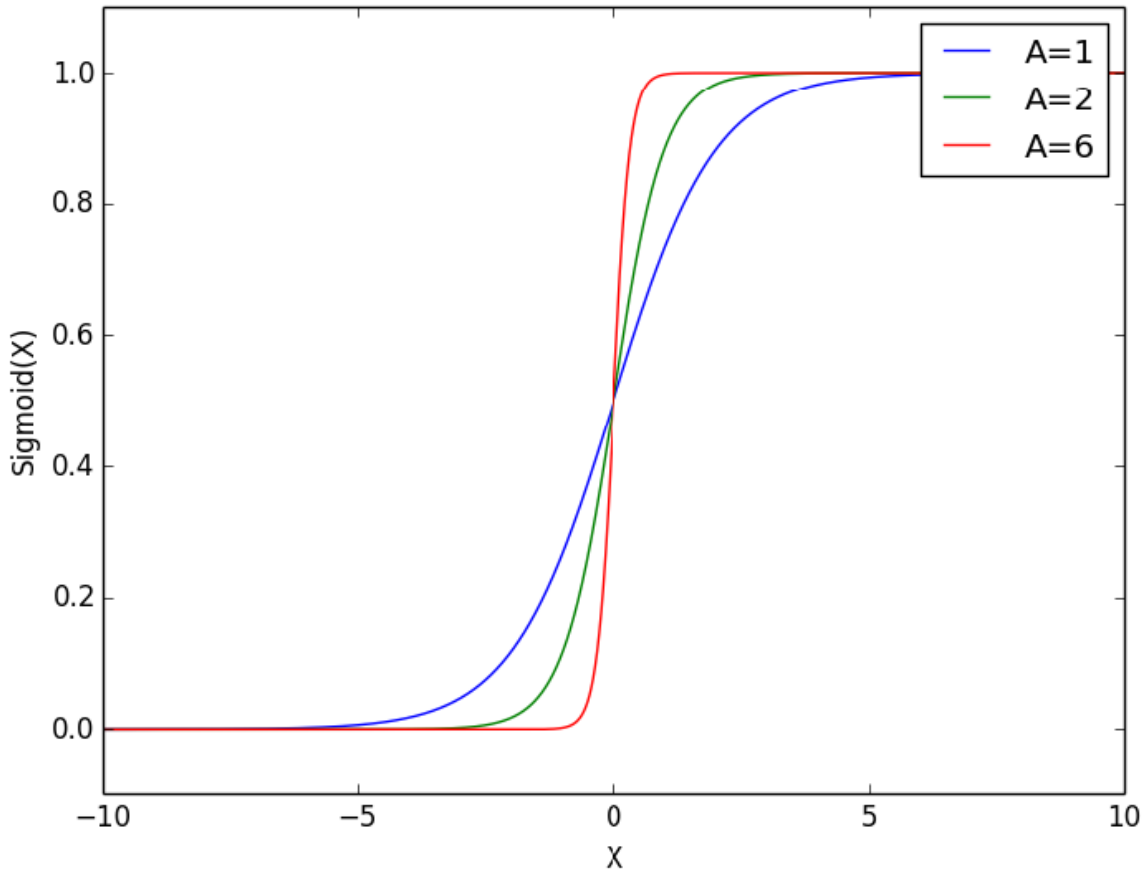
Cas particulier : $f(X)$ linéaire

$$P(Y=1|X) = \sigma(AX+b) = \frac{1}{1+e^{-(AX+b)}}$$

Fonction logistique ou sigmoïde III



Fonction logistique ou sigmoïde IV



Apprendre A et b

Problème

- $f(X)$ est linéaire
- $P(Y = 1|X)$ n'est pas linéaire

On ne peut pas utiliser directement les méthodes connues

Fonction logit

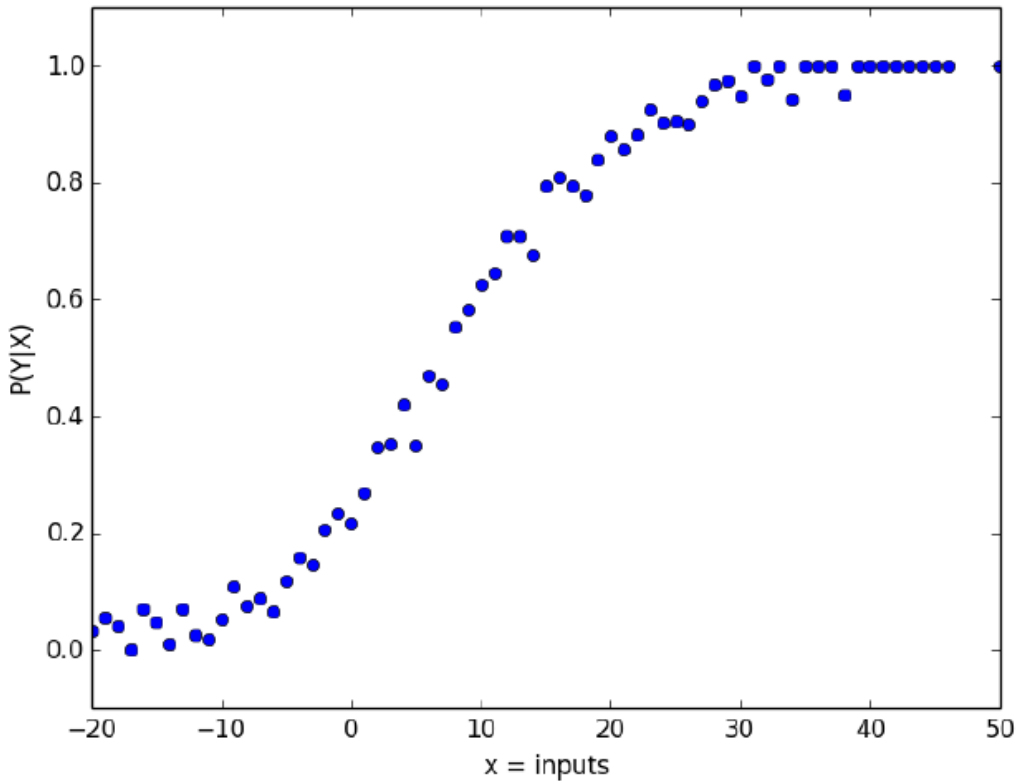
$$logit(x) = \log\left(\frac{x}{1-x}\right)$$

Application à la sigmoïde

$$logit(\sigma(AX + b)) = Ax + b$$

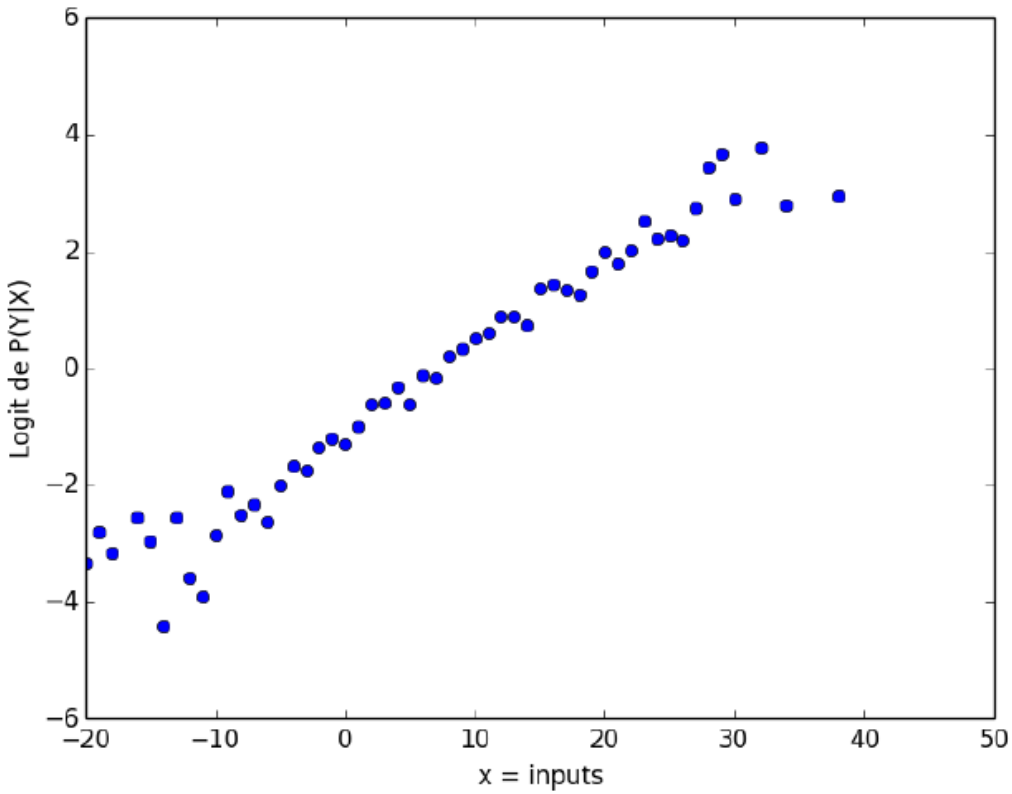
Régression de $logit(P(Y|X))$ I

Calcul des comptes $\hat{P}(Y = 1|X)$



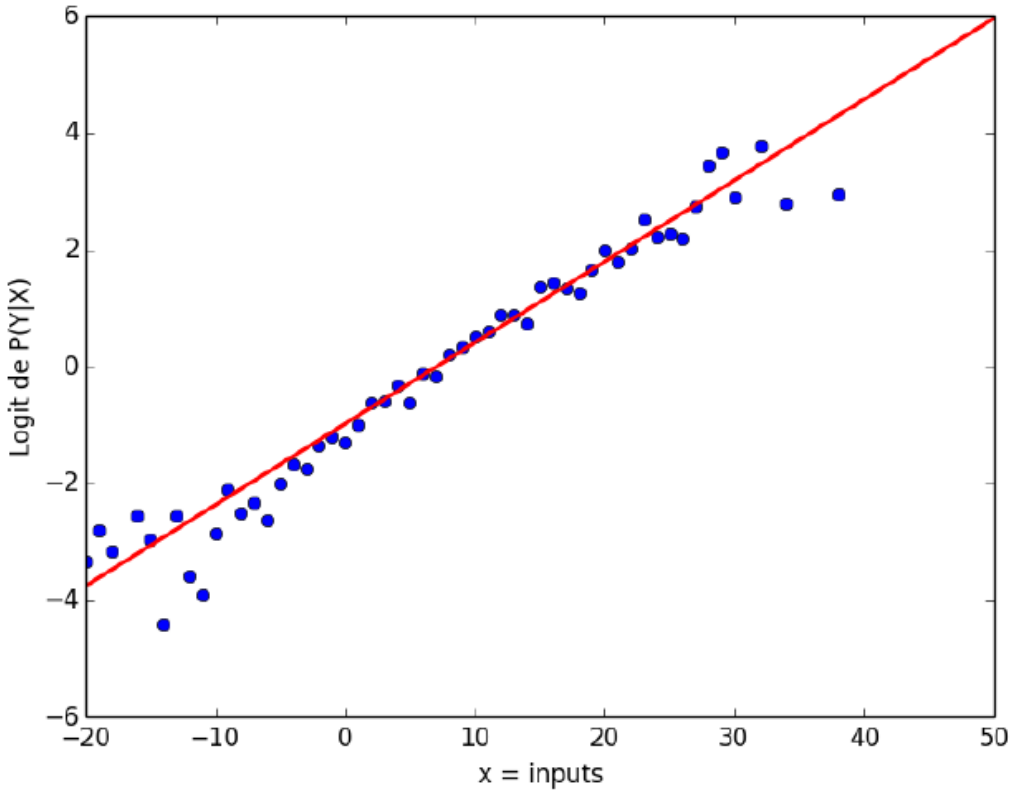
Régression de $logit(P(Y|X))$ II

Transformation des comptes $logit(\hat{P}(Y = 1|X))$



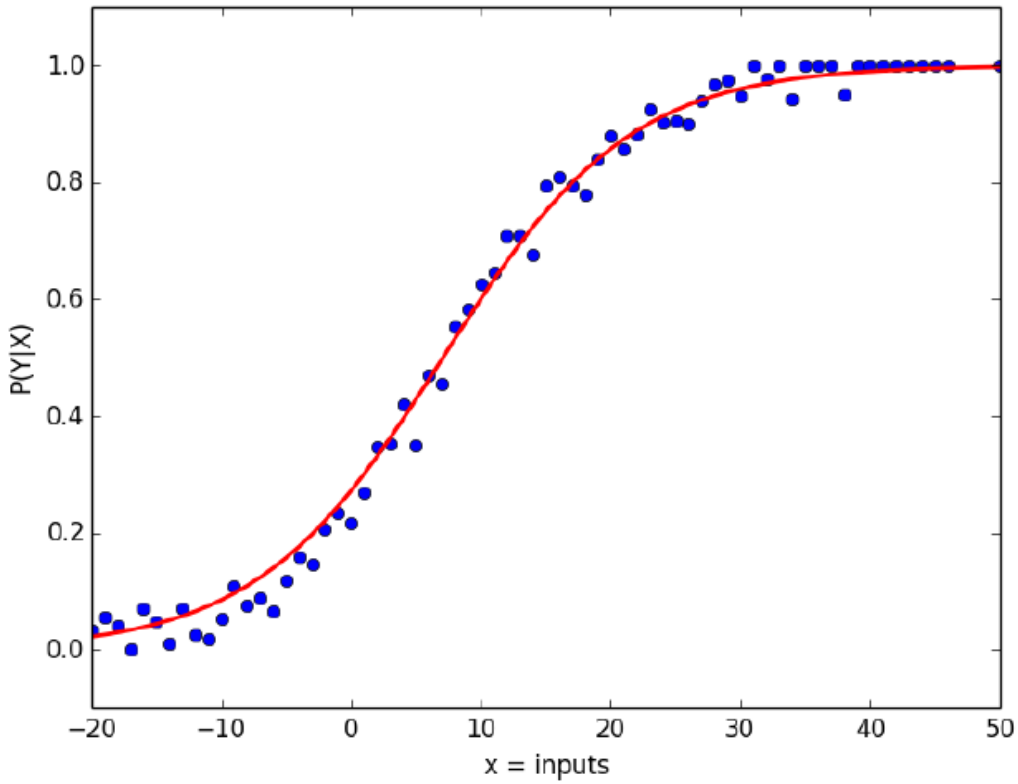
Régression de $logit(P(Y|X))$ III

Régression linéaire (par ex. moindres carrés) - on trouve A et b



Régression de $logit(P(Y|X))$ IV

Calcul de $\frac{1}{1+e^{-(AX+b)}}$



Pourquoi ce n'est pas une bonne idée ?

Plusieurs raisons

- Il faut forcément des comptes $\Rightarrow$ plusieurs valeurs de X avec le même Y
- Il faut passer par un logarithme (et $\log(0)$ n'est pas défini)
- On n'utilise qu'une partie des données $\Rightarrow$ on perd de l'information

Idée de solution

- Revenir à la méthodologie initiale
- Définir une fonction de perte
- Minimiser le risque empirique

Reposition du problème

Vraisemblance $P(Y|X;\theta)$

- $P(Y=1|X;\theta) = \sigma(\theta^T X)$
- $P(Y=0|X;\theta) = 1 - P(Y=1|X) = 1 - \sigma(\theta^T X)$
- $P(Y|X;\theta) = \sigma(\theta^T X)^y \cdot (1 - \sigma(\theta^T X))^{1-y}$

Maximiser la vraisemblance

$$\mathcal{L}(Y, f(X)) = -P(Y|X;\theta)$$

Perte logistique (log-vraisemblance)

$$\mathcal{L}(Y, f(X)) = -\log P(Y|X;\theta)$$

$$\mathcal{L}(Y, f(x)) = -y \cdot \log \sigma(\theta^T \mathbf{x}) - (1-y) \cdot \log\left(1 - \sigma(\theta^T \mathbf{x})\right)$$

Risque empirique et solution I

Risque empirique

$$R_N(f(x)) = \frac{1}{N}\sum_{i=1}^{N} -y^{(i)} \cdot \log \sigma(\theta^T \mathbf{x}^{(i)}) - (1-y) \cdot \log\left(1 - \sigma(\theta^T \mathbf{x}^{(i)})\right)$$

Propriétés

- $1 - \sigma(x) = \sigma(-x)$, $\frac{d\sigma(x)}{dx} = \sigma(x)\sigma(-x)$
- $\frac{d(\log f(x))}{dx} = \frac{f'(x)}{f(x)}$, $\frac{d\frac{1}{f(x)}}{dx} = \frac{f'(x)}{f^2(x)}$

Risque empirique et solution II

Gradient de $R_N(f(x))$

$$\nabla_\theta R_N(f(x)) = -\frac{1}{N}\sum_{i=1}^{N} \mathbf{x}^{(i)} \left(y^{(i)} - \sigma(\theta^T \mathbf{x}^{(i)})\right)$$

On ne peut pas l'annuler

Descente de Gradient de $R_N(f(x))$

$$\theta^{(t+1)} = \theta^{(t)} + \alpha_t \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}^{(i)} \left(y^{(i)} - \sigma(\theta^{(t)^T} \mathbf{x}^{(i)})\right)$$

$$\theta^{t+1} = \theta^{(t)} + \alpha_t \frac{1}{N} X \left(Y - \Sigma(X^T \theta^{(t)})\right)$$

Utilisation

Classification

- Calculer θ
- Fixer un seuil τ
- si $\sigma(\theta^T \mathbf{x}) \geqq \tau$ alors $\hat{y} = 1$
- si $\sigma(\theta^T \mathbf{x}) < \tau$ alors $\hat{y} = 0$

Vérification

$$Perf = \frac{1}{N} \sum_i |y_i - \hat{y}_i|$$