IFSBM Module 11

GUSTAVE ROUSSY
CANCER CAMPUS GRAND PARIS

université PARIS-SACLAY

IFSBM
INSTITUT DE FORMATION SUPÉRIEURE BIOMÉDICALE

Méthodes de classification supervisée

Yoann Pradat

https://github.com/gustaveroussy/IFSBM-bigdata/tree/master/Presentations/Mardi

Sommaire

1.1 Le jeu de données

Notations

$\mathrm{X} : \Omega \mapsto \mathscr{X}$: variable aléatoire (scalaire ou vectorielle)

$x \in \mathscr{X}$: une observation de la variable aléatoire

Exemples:

1. On lance un dé, $\mathscr{X} = \{1, 2, \cdots, 6\}$

1.1 Le jeu de données

Notations

$\mathrm{X} : \Omega \mapsto \mathscr{X}$: variable aléatoire (scalaire ou vectorielle)

$x \in \mathscr{X}$: une observation de la variable aléatoire

Exemples:

1. On lance un dé, $\mathscr{X} = \{1, 2, \cdots, 6\}$
2. On mesure la longueur et la largeur de fleurs $\mathscr{X} = \mathbb{R}^+ \times \mathbb{R}^+$

1.1 Le jeu de données

Notations

$\mathrm{X} : \Omega \mapsto \mathscr{X}$: variable aléatoire (scalaire ou vectorielle)

$x \in \mathscr{X}$: une observation de la variable aléatoire

Exemples:

1. On lance un dé, $\mathscr{X} = \{1, 2, \cdots, 6\}$
2. On mesure la longueur et la largeur de fleurs $\mathscr{X} = \mathbb{R}^+ \times \mathbb{R}^+$
3. On mesure le niveau d'expression de gènes relatif à 1M (K gènes)

$$\mathscr{X} = [0, 1M]^K = [0, 1M] \times [0, 1M] \times \cdots \times [0, 1M]$$

1.1 Le jeu de données

Notations

$\mathrm{X} : \Omega \mapsto \mathscr{X}$: variable aléatoire (scalaire ou vectorielle)

$x \in \mathscr{X}$: une observation de la variable aléatoire

Exemples:

1. On lance un dé, $\mathscr{X} = \{1, 2, \cdots, 6\}$
2. On mesure la longueur et la largeur de fleurs $\mathscr{X} = \mathbb{R}^+ \times \mathbb{R}^+$
3. On mesure le niveau d'expression de gènes relatif à 1M (K gènes)

$$\mathscr{X} = [0, 1M]^K = [0, 1M] \times [0, 1M] \times \cdots \times [0, 1M]$$

$n \in \mathbb{N}^*$: nombre d'observations (=individus, échantillons)

$p \in \mathbb{N}^*$: nombre de variables (=covariables, prédicteurs, features)

$x_{1:n} = (x_1, \cdots, x_n)$: ensemble d'observations (=échantillon, =dataset)

1.1 Le jeu de données

Exemple:

1. On réalise un essai clinique de réponse à un traitement sur 100 patients

$$X = \begin{array}{l}(\text{Age, Poids, Nb Tx antérieurs,} \\ \text{Exp gene 1, Exp gene 2, Mutation gene 1,} \\ \text{Temps avant rechute, Meilleure réponse})\end{array}$$

1.1 Le jeu de données

Exemple:

1. On réalise un essai clinique de réponse à un traitement sur 100 patients

$$
\begin{aligned}
X = \quad & (\text{Age, Poids, Nb Tx antérieurs,} \\
& \text{Exp gene 1, Exp gene 2, Mutation gene 1,} \\
& \text{Temps avant rechute, Meilleure réponse}) \\
x_1 = \quad & (45.2, 78.2, 3, 1032, 258, 1, 85, \text{PR}) \\
x_2 = \quad & (81, 63, 6, 589, 903, 0, 390, \text{SD}) \\
\ldots &
\end{aligned}
$$

1.1 Le jeu de données

Exemple:

1. On réalise un essai clinique de réponse à un traitement sur 100 patients

$$
\begin{aligned}
X = \quad & (\text{Age, Poids, Nb Tx antérieurs,} \\
& \text{Exp gene 1, Exp gene 2, Mutation gene 1,} \\
& \text{Temps avant rechute, Meilleure réponse}) \\
x_1 = \quad & (45.2, 78.2, 3, 1032, 258, 1, 85, \text{PR}) \\
x_2 = \quad & (81, 63, 6, 589, 903, 0, 390, \text{SD}) \\
\cdots &
\end{aligned}
\Bigg\} \; (x_1, x_2, \cdots, x_{100}) = \text{jeu de données}
$$

Objectifs

1. **Proposer un modèle** mathématique pour modéliser une variable en fonction d'autres. Exemple $\text{Meilleure réponse vs (Exp gene 1, Exp gene 2, Age)}$
   $\text{Modèle} = \text{Régression logistique multinomiale}$

1.1 Le jeu de données

Exemple:

1. On réalise un essai clinique de réponse à un traitement sur 100 patients

$$
\begin{aligned}
X = \quad & (\text{Age, Poids, Nb Tx antérieurs,} \\
& \text{Exp gene 1, Exp gene 2, Mutation gene 1,} \\
& \text{Temps avant rechute, Meilleure réponse}) \\
x_1 = \quad & (45.2, 78.2, 3, 1032, 258, 1, 85, \text{PR}) \\
x_2 = \quad & (81, 63, 6, 589, 903, 0, 390, \text{SD}) \\
\cdots &
\end{aligned}
\qquad \Big\} \quad (x_1, x_2, \cdots, x_{100}) = \text{jeu de données}
$$

Objectifs

1. **Proposer un modèle** mathématique pour modéliser une variable en fonction d'autres. Exemple $\text{Meilleure réponse vs (Exp gene 1, Exp gene 2, Age)}$ $\text{Modèle} = \text{Régression logistique multinomiale}$
2. **Estimer les paramètres** du modèle à partir de $(x_1, x_2, \cdots, x_{100})$

1.2 L'estimation

Exemple:

1. On veut predire le volume de la tumeur à partir de l'expression de 5000 gènes

   -> $(\mathrm{G}^1, \cdots, \mathrm{G}^{5000})$ profil d'expression et $\mathrm{V} =$ volume

   $x_{1:n} = (g^1, \cdots, g^{5000}, v)_{1:n}$ observations de $\mathrm{X} = (\mathrm{G}^1, \cdots, \mathrm{G}^{5000}, \mathrm{V})$

1.2 L'estimation

Exemple:

1. On veut predire le volume de la tumeur à partir de l'expression de 5000 gènes

   -> $(\mathrm{G}^1, \cdots, \mathrm{G}^{5000})$ profil d'expression et $\mathrm{V} =$ volume

   $x_{1:n} = (g^1, \cdots, g^{5000}, v)_{1:n}$ observations de $\mathrm{X} = (\mathrm{G}^1, \cdots, \mathrm{G}^{5000}, \mathrm{V})$

   -> on prend comme modèle un ***modèle de regression linéaire***

$$\mathrm{V} = f_\theta(\mathrm{G}^1, \mathrm{G}^{5000})$$
$$= \theta_1 \mathrm{G}^1 + \cdots + \theta_{5000} \mathrm{G}^{5000}$$

Schématique, pas mathématiquement exact

1.2 L'estimation

Exemple:

1. On veut predire le volume de la tumeur à partir de l'expression de 5000 gènes

   -> $(\mathrm{G}^1, \cdots, \mathrm{G}^{5000})$ profil d'expression et $\mathrm{V} =$ volume

   $x_{1:n} = (g^1, \cdots, g^{5000}, v)_{1:n}$ observations de $\mathrm{X} = (\mathrm{G}^1, \cdots, \mathrm{G}^{5000}, \mathrm{V})$

   -> on prend comme modèle un ***modèle de regression linéaire***

$$\mathrm{V} = f_\theta(\mathrm{G}^1, \mathrm{G}^{5000})$$
$$= \theta_1 \mathrm{G}^1 + \cdots + \theta_{5000} \mathrm{G}^{5000}$$

Schématique, pas mathématiquement exact

Fonction objectif (=coût) Ca veut dire quoi "coller au mieux aux données" ?
Mesure de distance entre l'observation v_i et la prediction $\hat{v_i} = f_\theta(g_i^1, \cdots, g_i^{5000})$?

1.2 L'estimation

Exemple:

1. On veut predire le volume de la tumeur à partir de l'expression de 5000 gènes

   -> $(\mathrm{G}^1, \cdots, \mathrm{G}^{5000})$ profil d'expression et $\mathrm{V} =$ volume

   $x_{1:n} = (g^1, \cdots, g^{5000}, v)_{1:n}$ observations de $\mathrm{X} = (\mathrm{G}^1, \cdots, \mathrm{G}^{5000}, \mathrm{V})$

   -> on prend comme modèle un ***modèle de regression linéaire***

$$\mathrm{V} = f_\theta(\mathrm{G}^1, \mathrm{G}^{5000})$$
$$= \theta_1 \mathrm{G}^1 + \cdots + \theta_{5000} \mathrm{G}^{5000}$$

Schématique, pas mathématiquement exact

Fonction objectif (=coût) Ca veut dire quoi "coller au mieux aux données" ?

Mesure de distance entre l'observation v_i et la prediction $\hat{v}_i = f_\theta(g_i^1, \cdots, g_i^{5000})$?

$d(v_i, \hat{v}_i) = (v_i - \hat{v}_i)$?

1.2 L'estimation

Exemple:

1. On veut predire le volume de la tumeur à partir de l'expression de 5000 gènes

   -> $(\mathrm{G}^1, \cdots, \mathrm{G}^{5000})$ profil d'expression et V = volume

   $x_{1:n} = (g^1, \cdots, g^{5000}, v)_{1:n}$ observations de $\mathrm{X} = (\mathrm{G}^1, \cdots, \mathrm{G}^{5000}, \mathrm{V})$

   -> on prend comme modèle un ***modèle de regression linéaire***

$$\mathrm{V} = f_\theta(\mathrm{G}^1, \mathrm{G}^{5000})$$
$$= \theta_1 \mathrm{G}^1 + \cdots + \theta_{5000} \mathrm{G}^{5000}$$

Schématique, pas mathématiquement exact

Fonction objectif (=coût) Ca veut dire quoi "coller au mieux aux données" ?

Mesure de distance entre l'observation v_i et la prediction $\hat{v}_i = f_\theta(g_i^1, \cdots, g_i^{5000})$?

$d(v_i, \hat{v}_i) = (v_i - \hat{v}_i)$? $\quad (v_i - \hat{v}_i)^2$?

1.2 L'estimation

Exemple:

1. On veut predire le volume de la tumeur à partir de l'expression de 5000 gènes

   -> $(\mathrm{G}^1, \cdots, \mathrm{G}^{5000})$ profil d'expression et V = volume

   $x_{1:n} = (g^1, \cdots, g^{5000}, v)_{1:n}$ observations de $\mathrm{X} = (\mathrm{G}^1, \cdots, \mathrm{G}^{5000}, \mathrm{V})$

   -> on prend comme modèle un ***modèle de regression linéaire***

$$\mathrm{V} = f_\theta(\mathrm{G}^1, \mathrm{G}^{5000})$$
$$= \theta_1 \mathrm{G}^1 + \cdots + \theta_{5000} \mathrm{G}^{5000}$$

Schématique, pas mathématiquement exact

Fonction objectif (=coût) Ca veut dire quoi "coller au mieux aux données" ?

Mesure de distance entre l'observation v_i et la prediction $\hat{v}_i = f_\theta(g_i^1, \cdots, g_i^{5000})$?

$d(v_i, \hat{v}_i) = (v_i - \hat{v}_i)$? $\quad (v_i - \hat{v}_i)^2$? $\quad (v_i - \hat{v}_i)^4$?

1.2 L'estimation

Exemple:

1. On veut predire le volume de la tumeur à partir de l'expression de 5000 gènes

   -> $(\mathrm{G}^1, \cdots, \mathrm{G}^{5000})$ profil d'expression et V = volume

   $x_{1:n} = (g^1, \cdots, g^{5000}, v)_{1:n}$ observations de $\mathrm{X} = (\mathrm{G}^1, \cdots, \mathrm{G}^{5000}, \mathrm{V})$

   -> on prend comme modèle un ***modèle de regression linéaire***

$$\mathrm{V} = f_\theta(\mathrm{G}^1, \mathrm{G}^{5000})$$
$$= \theta_1 \mathrm{G}^1 + \cdots + \theta_{5000} \mathrm{G}^{5000}$$

Schématique, pas mathématiquement exact

Fonction objectif (=coût) Ca veut dire quoi "coller au mieux aux données" ?

Mesure de distance entre l'observation v_i et la prediction $\hat{v}_i = f_\theta(g_i^1, \cdots, g_i^{5000})$?

$d(v_i, \hat{v}_i) = (v_i - \hat{v}_i)$? $\quad (v_i - \hat{v}_i)^2$? $\quad (v_i - \hat{v}_i)^4$? $\quad e^{|v_i - \hat{v}_i|} - 1$?

1.2 L'estimation

Exemple:

1. On veut predire le volume de la tumeur à partir de l'expression de 5000 gènes

-> $(\mathrm{G}^1, \cdots, \mathrm{G}^{5000})$ profil d'expression et V = volume

$x_{1:n} = (g^1, \cdots, g^{5000}, v)_{1:n}$ observations de $\mathrm{X} = (\mathrm{G}^1, \cdots, \mathrm{G}^{5000}, \mathrm{V})$

-> on prend comme modèle un ***modèle de regression linéaire***

$$\mathrm{V} = f_\theta(\mathrm{G}^1, \mathrm{G}^{5000})$$
$$= \theta_1 \mathrm{G}^1 + \cdots + \theta_{5000} \mathrm{G}^{5000}$$

Schématique, pas mathématiquement exact

Fonction objectif (=coût) Ca veut dire quoi "coller au mieux aux données" ?

Mesure de distance entre l'observation v_i et la prediction $\hat{v}_i = f_\theta(g_i^1, \cdots, g_i^{5000})$?

$d(v_i, \hat{v}_i) = (v_i - \hat{v}_i)$? $\quad (v_i - \hat{v}_i)^2$? $\quad (v_i - \hat{v}_i)^4$? $\quad e^{|v_i - \hat{v}_i|} - 1$?

Estimation $\quad \hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^{n} d(v_i, \hat{v}_i)$

1.2 L'estimation

Quels modèles ? Le modèle peut avoir une interpretation probabiliste ou non.

-> si interpretation probabiliste

fonction objectif = **- maximum de vraisemblance**

Exemples: régression linéaire, régression logistique, classification bayésienne naïve

1.2 L'estimation

Quels modèles ? Le modèle peut avoir une interpretation probabiliste ou non.

-> si interpretation probabiliste

fonction objectif = **- maximum de vraisemblance**

Exemples: régression linéaire, régression logistique, classification bayésienne naïve

-> si pas d'interpretation probabiliste

fonction objectif = **a la main**

Exemples: machine à vecteur de support (SVM), forêt aléatoire, reseaux de neurones

1.2 L'estimation

Quels modèles ? Le modèle peut avoir une interpretation probabiliste ou non.

-> si interpretation probabiliste (=modèle statistique)

fonction objectif = **- maximum de vraisemblance**

Exemples: régression linéaire, régression logistique, classification bayésienne naïve

-> si pas d'interpretation probabiliste (=modèle statistique?)

fonction objectif = **a la main**

Exemples: machine à vecteur de support (SVM), forêt aléatoire, reseaux de neurones

Les reseaux de neurones sont-ils des modèles statistiques ?
https://ai.stackexchange.com/questions/10289/are-neural-networks-statistical-models/18580#18580
Cf aussi sur la definition des modèles statistiques https://www.stat.uchicago.edu/~pmcc/pubs/AOS023.pdf

1.2 L'estimation

-> si interpretation probabiliste (=modèle statistique)

Modèle statistique = ensemble de lois (=mesures) de probabilité $\mathbb{P}$ sur l'espace des observations $\mathcal{X}$. Si proba paramétriques, alors on parle de modèle paramétrique. Enfin si les proba ont une densité p alors le modèle s'écrit

$$\mathcal{M}_\Theta = \{p_\theta | \theta \in \Theta\}$$

1.2 L'estimation

-> si interpretation probabiliste (=modèle statistique)

Modèle statistique = ensemble de lois (=mesures) de probabilité $\mathbb{P}$ sur l'espace des observations $\mathcal{X}$. Si proba paramétriques, alors on parle de modèle paramétrique. Enfin si les proba ont une densité p alors le modèle s'écrit

$$\mathcal{M}_\Theta = \{p_\theta | \theta \in \Theta\}$$

Quand on modélise, on fait l'hypothèse qu'il existe θ^* tel que $\mathbb{P}_X = \mathbb{P}_{\theta^*}$

Modéliser = calculer $\hat{\theta}$ en "espérant" que $\hat{\theta} \approx \theta^*$

1.2 L'estimation

Modèle statistique $\mathcal{M}_\Theta = \{p_\theta | \theta \in \Theta\}$ Il existe θ^* tel que $\mathbb{P}_X = \mathbb{P}_{\theta^*}$

Pour le modèle candidat $\mathbb{P}_\theta$

$$p_\theta(x_i) = \text{vraisemblance échantillon i}$$

$$\prod_{i=1}^{n} p_\theta(x_i) = \text{vraisemblance jeu de données}$$

1.2 L'estimation

Modèle statistique $\mathcal{M}_\Theta = \{p_\theta | \theta \in \Theta\}$ Il existe θ^* tel que $\mathbb{P}_X = \mathbb{P}_{\theta^*}$

Pour le modèle candidat $\mathbb{P}_\theta$

$$p_\theta(x_i) = \text{vraisemblance échantillon i}$$

$$\prod_{i=1}^{n} p_\theta(x_i) = \text{vraisemblance jeu de données}$$

$$L(\theta) = -\prod_{i=1}^{n} p_\theta(x_i)$$

Fonction de coût à minimiser

1.2 L'estimation

Modèle statistique $\mathcal{M}_\Theta = \{p_\theta | \theta \in \Theta\}$ Il existe θ^* tel que $\mathbb{P}_X = \mathbb{P}_{\theta^*}$

Pour le modèle candidat $\mathbb{P}_\theta$

$p_\theta(x_i) =$ vraisemblance échantillon i

$\prod_{i=1}^{n} p_\theta(x_i) =$ vraisemblance jeu de données

$$L(\theta) = -\prod_{i=1}^{n} p_\theta(x_i)$$ equivalent à $$\ell(\theta) = -\sum_{i=1}^{n} \log p_\theta(x_i)$$

Fonction de coût à minimiser = - log de la vraisemblance

Sommaire

2.1 Régression logistique binaire

Modèle statistique de regression Prédire Y à partir de X

Jeu de données $x_{1:n}, y_{1:n}$

Un modèle de regression est une famille de lois de probabilités *conditionnelles*

$$\mathcal{M}_\Theta = \{\mathbb{P}^\theta_{Y|X=x} | \theta \in \Theta, x \in \mathscr{X}\}$$

2.1 Régression logistique binaire

Modèle statistique de regression Prédire Y à partir de X

Jeu de données $x_{1:n}, y_{1:n}$

Un modèle de regression est une famille de lois de probabilités *conditionnelles*

$$\mathcal{M}_\Theta = \{\mathbb{P}^\theta_{Y|X=x} | \theta \in \Theta, x \in \mathscr{X}\}$$

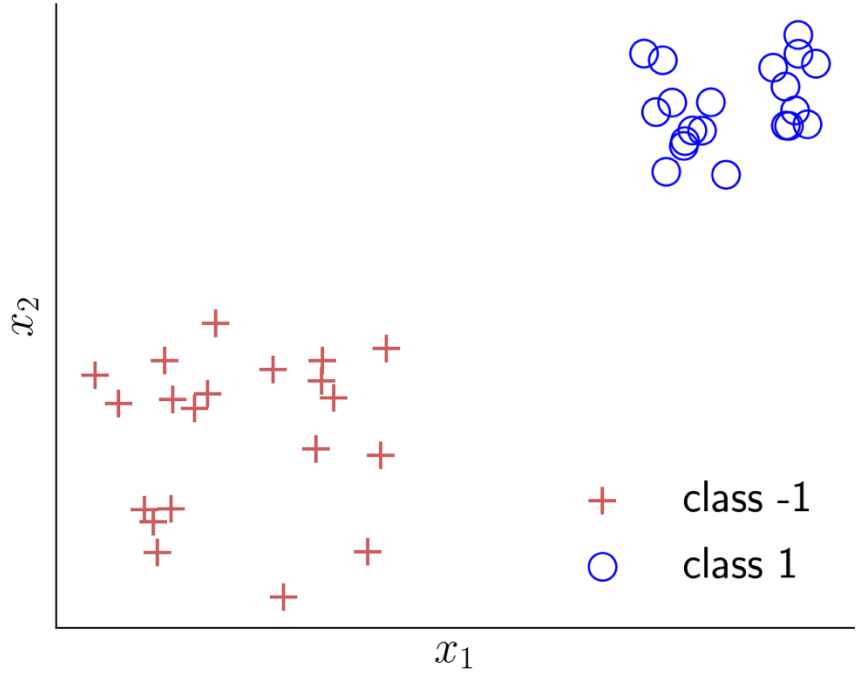
Regression logistique $Y \in \{0,1\}, \quad X \in \mathbb{R}^p$

$$\mathbb{P}^\theta_{Y|X=x} = \text{Binomial}(\sigma(x^\top \theta)) \qquad \sigma(u) = \frac{1}{1+e^{-u}}$$

2.1 Régression logistique binaire

Regression logistique $Y \in \{0,1\}, \quad X \in \mathbb{R}^p \quad \mathbb{P}^{\theta}_{Y|X=x} = \text{Binomial}(\sigma(x^\top \theta))$

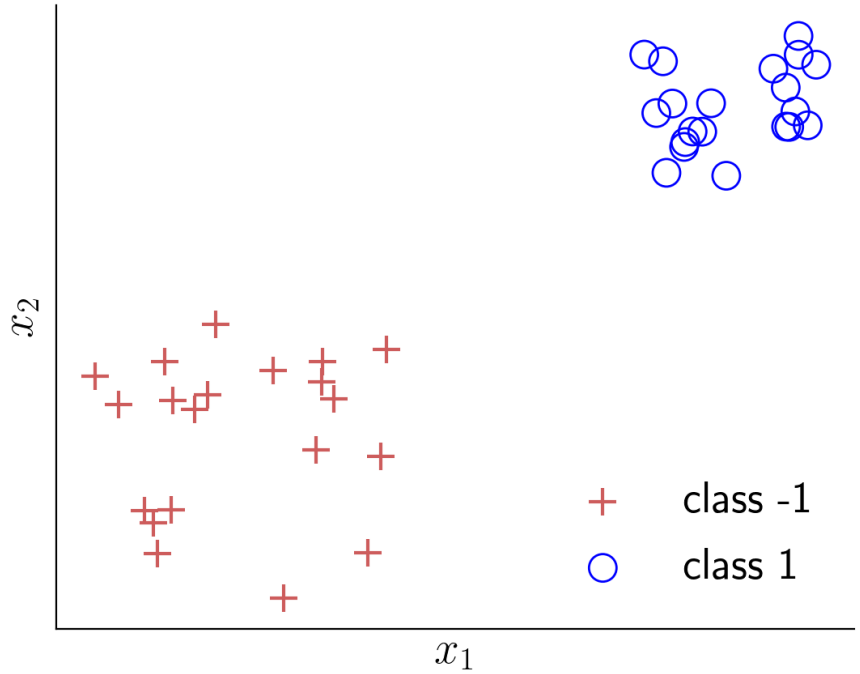
Exemple: p=2, $X = (X^1, X^2)$

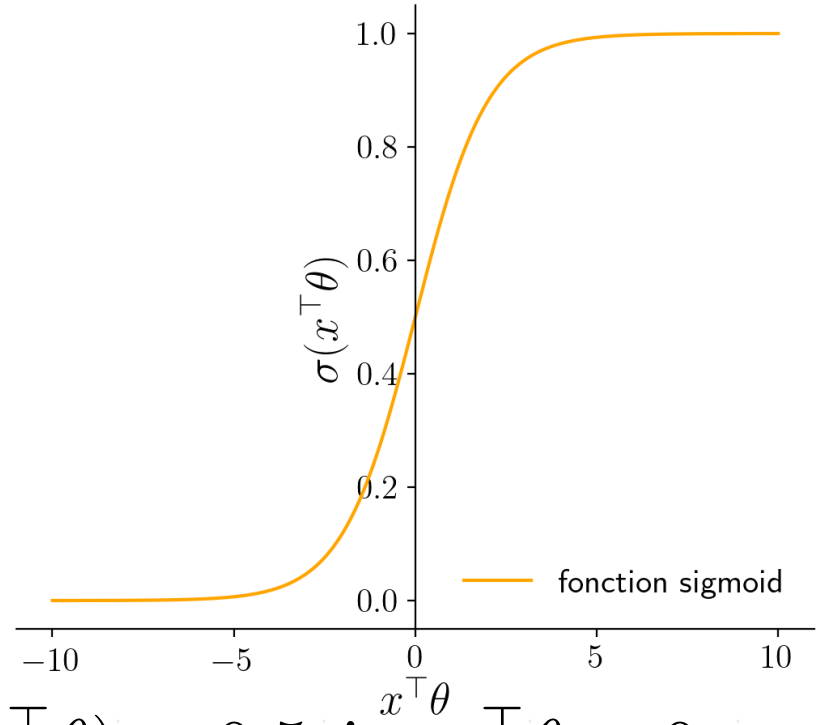


2.1 Régression logistique binaire

Regression logistique $Y \in \{0, 1\}, \quad X \in \mathbb{R}^p \quad \mathbb{P}^{\theta}_{Y|X=x} = \text{Binomial}(\sigma(x^\top \theta))$

Exemple: p=2, $X = (X^1, X^2)$



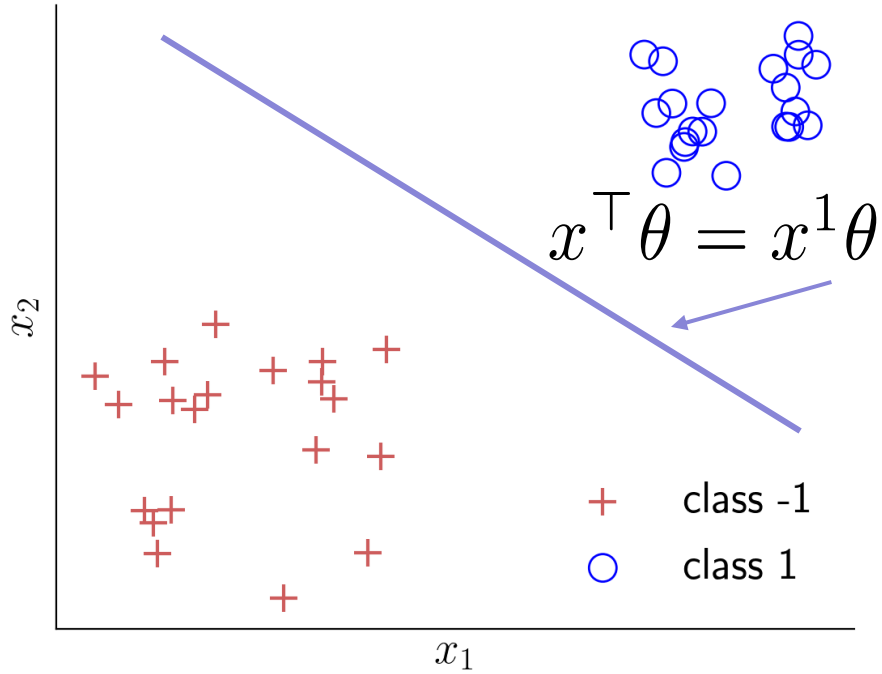


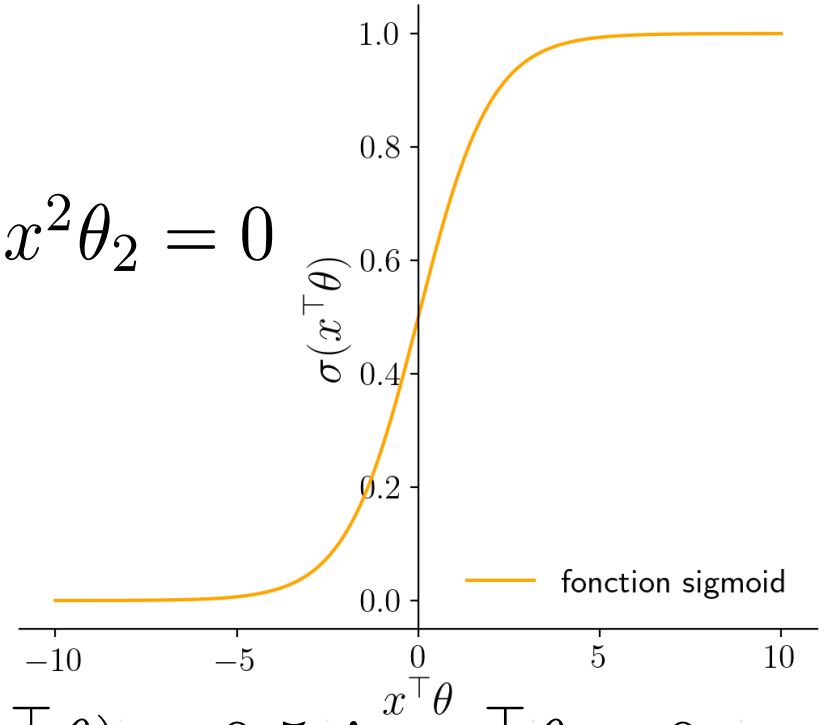
$$\text{Pour } X = x, \text{ prédiction} = \begin{cases} 1 & \text{si } \sigma(x^\top\theta) > 0.5, \text{i.e } x^\top\theta > 0 \\ 0 & \text{si } x^\top\theta < 0 \end{cases}$$

2.1 Régression logistique binaire

Regression logistique $Y \in \{0,1\}, \quad X \in \mathbb{R}^p \quad \mathbb{P}^{\theta}_{Y|X=x} = \text{Binomial}(\sigma(x^\top \theta))$

Exemple: p=2, $X = (X^1, X^2)$





$$\text{Pour } X = x, \text{ prédiction} = \begin{cases} 1 & \text{si } \sigma(x^\top\theta) > 0.5, \text{i.e } x^\top\theta > 0 \\ 0 & \text{si } x^\top\theta < 0 \end{cases}$$

2.1 Régression logistique binaire

Regression logistique $Y \in \{0,1\}, \quad X \in \mathbb{R}^p$

$$\mathbb{P}^{\theta}_{Y|X=x} = \text{Binomial}(\sigma(x^\top \theta)) \qquad \sigma(u) = \frac{1}{1+e^{-u}}$$

Comment s'écrit la vraisemblance de $y_i|x_i$?

$$p_\theta(.|x_i) = \begin{cases} 1 & \text{avec probabilité } \sigma(x_i^\top \theta) \\ 0 & \text{avec probabilité } 1-\sigma(x_i^\top \theta) \end{cases}$$

2.1 Régression logistique binaire

Regression logistique $Y \in \{0,1\}, \quad X \in \mathbb{R}^p$

$$\mathbb{P}^{\theta}_{Y|X=x} = \text{Binomial}(\sigma(x^\top \theta)) \qquad \sigma(u) = \frac{1}{1+e^{-u}}$$

Comment s'écrit la vraisemblance de $y_i|x_i$?

$$p_\theta(.|x_i) = \begin{cases} 1 & \text{avec probabilité } \sigma(x_i^\top \theta) \\ 0 & \text{avec probabilité } 1-\sigma(x_i^\top \theta) \end{cases}$$

$$p_\theta(y_i|x_i) = \sigma(\theta^\top x_i)^{y_i}(1-\sigma(\theta^\top x_i))^{1-y_i}$$

2.1 Régression logistique binaire

Regression logistique $Y \in \{0,1\}, \quad X \in \mathbb{R}^p$

$$\mathbb{P}^{\theta}_{Y|X=x} = \text{Binomial}(\sigma(x^\top \theta)) \qquad \sigma(u) = \frac{1}{1+e^{-u}}$$

Comment s'écrit la vraisemblance de $y_i|x_i$?

$$p_\theta(.|x_i) = \begin{cases} 1 & \text{avec probabilité } \sigma(x_i^\top \theta) \\ 0 & \text{avec probabilité } 1-\sigma(x_i^\top \theta) \end{cases}$$

$$p_\theta(y_i|x_i) = \sigma(\theta^\top x_i)^{y_i}(1-\sigma(\theta^\top x_i))^{1-y_i}$$

Fonction de coût

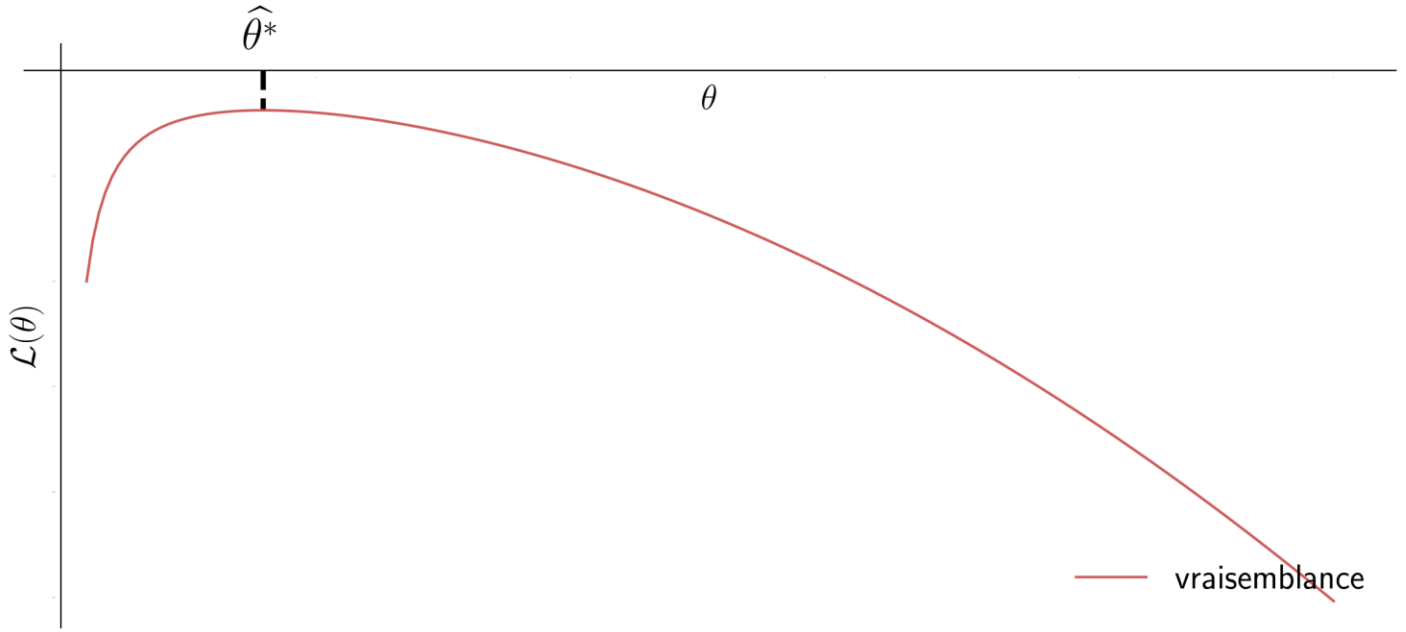
$$\ell(\theta; x_{1:n}, y_{1:n}) = -\sum_{i=1}^{n} y_i \log \sigma(\theta^\top x_i) + (1-y_i)\log(1-\sigma(\theta^\top x_i))$$

2.1 Régression logistique binaire

Regression logistique

Fonction de coût

$$\ell(\theta; x_{1:n}, y_{1:n}) = -\sum_{i=1}^{n} y_i \log \sigma(\theta^\top x_i) + (1 - y_i) \log(1 - \sigma(\theta^\top x_i))$$



Concave !

2.1 Régression logistique binaire

Regression logistique

Fonction de coût

$$\ell(\theta; x_{1:n}, y_{1:n}) = -\sum_{i=1}^{n} y_i \log \sigma(\theta^\top x_i) + (1 - y_i) \log(1 - \sigma(\theta^\top x_i))$$

Minimisation par descente de gradient

$$\ell(\theta^{t+1}) = \ell(\theta^t) + (\theta^{t+1} - \theta^t)\nabla\ell(\theta^t) + o(\theta^{t+1} - \theta^t)$$

(dvp de Taylor ordre 1)

2.1 Régression logistique binaire

Regression logistique

Fonction de coût

$$\ell(\theta; x_{1:n}, y_{1:n}) = -\sum_{i=1}^{n} y_i \log \sigma(\theta^\top x_i) + (1 - y_i) \log(1 - \sigma(\theta^\top x_i))$$

Minimisation par descente de gradient

$$\ell(\theta^{t+1}) = \ell(\theta^t) + (\theta^{t+1} - \theta^t)\nabla\ell(\theta^t) + o(\theta^{t+1} - \theta^t)$$

(dvp de Taylor ordre 1)

Idée. Choisir θ^{t+1} tel que $(\theta^{t+1} - \theta^t) = -\nabla\ell(\theta^t)$

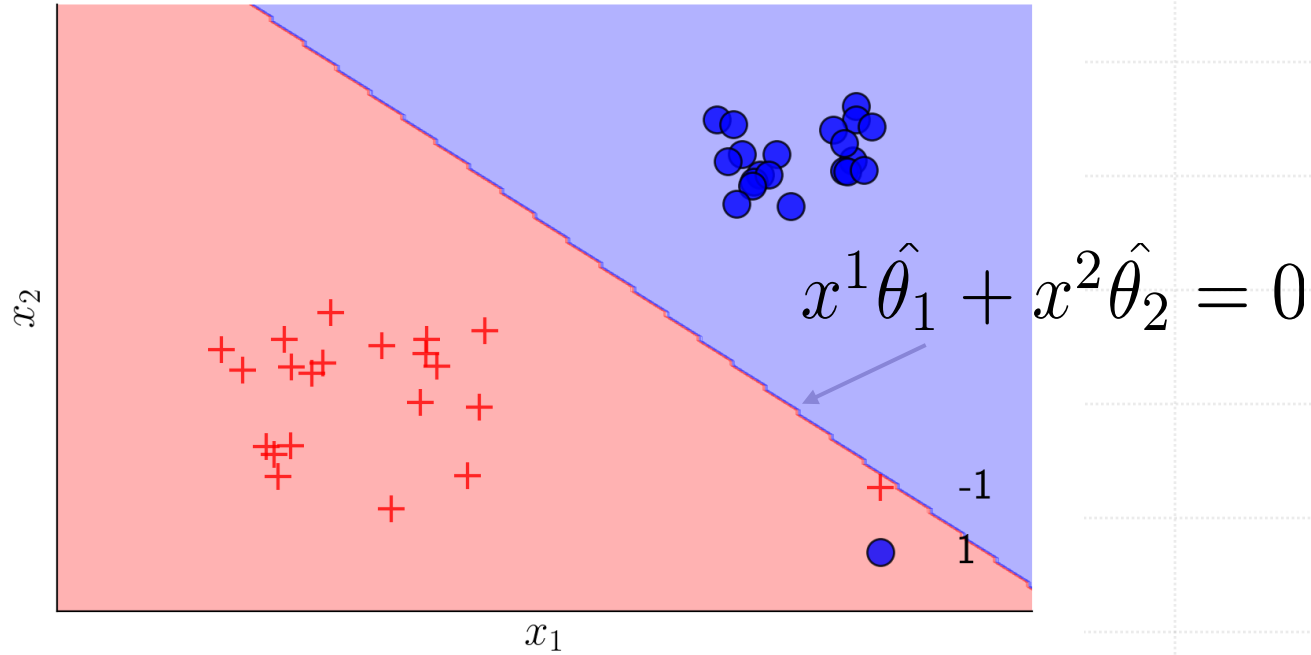
2.1 Régression logistique binaire

Regression logistique

Fonction de coût

$$\ell(\theta; x_{1:n}, y_{1:n}) = -\sum_{i=1}^{n} y_i \log \sigma(\theta^\top x_i) + (1 - y_i) \log(1 - \sigma(\theta^\top x_i))$$

Minimisation par descente de gradient



2.2 Régression logistique multinomiale

Regression logistique multinomiale $Y \in \{1, 2, \cdots, K\}, \quad X \in \mathbb{R}^p$

$$\mathbb{P}^{\Theta}_{Y|X=x} = \text{Multinomial}(\sigma(\Theta^\top x))$$

$$\sigma(\mathrm{z}) = \frac{1}{\sum_{k=1}^{K} e^{z_k}} \begin{bmatrix} e^{z_1} \\ \vdots \\ e^{z_K} \end{bmatrix}$$

$$\Theta = \begin{bmatrix} \theta_{1,1} & \cdots & \theta_{1,p} \\ \vdots & \ddots & \vdots \\ \theta_{K,1} & \cdots & \theta_{K,p} \end{bmatrix}$$

2.2 Régression logistique multinomiale

Regression logistique multinomiale $Y \in \{1, 2, \cdots, K\}, \quad X \in \mathbb{R}^p$

$$\mathbb{P}^{\Theta}_{Y|X=x} = \text{Multinomial}(\sigma(\Theta^\top x))$$

$$\sigma(\mathrm{z}) = \frac{1}{\sum_{k=1}^{K} e^{z_k}} \begin{bmatrix} e^{z_1} \\ \vdots \\ e^{z_K} \end{bmatrix}$$

$$\Theta = \begin{bmatrix} \theta_{1,1} & \cdots & \theta_{1,p} \\ \vdots & \ddots & \vdots \\ \theta_{K,1} & \cdots & \theta_{K,p} \end{bmatrix}$$

$$\begin{bmatrix} \mathbb{P}^{\Theta}_{Y|X=x}(Y=1) \\ \vdots \\ \mathbb{P}^{\Theta}_{Y|X=x}(Y=K) \end{bmatrix} = \frac{1}{\sum_{k=1}^{K} e^{(\Theta\top x)_k}} \begin{bmatrix} e^{(\Theta\top x)_1} \\ \vdots \\ e^{(\Theta\top x)_K} \end{bmatrix}$$

2.2 Régression logistique multinomiale

Regression logistique multinomiale $Y \in \{1, 2, \cdots, K\}, \quad X \in \mathbb{R}^p$

Comment s'écrit la vraisemblance de $y_i | x_i$?

$$p_\Theta(.|x_i) = k \quad \text{avec probabilité } \sigma(\Theta^\top x)_k$$

$$p_\Theta(y_i|x_i) = \prod_{k=1}^{K} \sigma(\Theta^\top x_i)_k^{\mathbb{1}_{y_i=k}}$$

2.2 Régression logistique multinomiale

Regression logistique multinomiale $Y \in \{1, 2, \cdots, K\}, \quad X \in \mathbb{R}^p$

Comment s'écrit la vraisemblance de $y_i | x_i$?

$$p_\Theta(.|x_i) = k \quad \text{avec probabilité } \sigma(\Theta^\top x)_k$$

$$p_\Theta(y_i|x_i) = \prod_{k=1}^{K} \sigma(\Theta^\top x_i)_k^{\mathbb{1}_{y_i=k}}$$

Fonction de coût

$$\ell(\Theta; x_{1:n}, y_{1:n}) = -\sum_{i=1}^{n}\sum_{k=1}^{K} \mathbb{1}_{y_i=k} \log \sigma(\Theta^\top x_i)_k$$

2.2 Régression logistique multinomiale

Estimation par descente de gradient $(\Theta^{t+1} - \Theta^t) = -\nabla \ell(\Theta^t)$

2.2 Régression logistique multinomiale

Estimation par descente de gradient $(\Theta^{t+1} - \Theta^t) = -\nabla \ell(\Theta^t)$

NOTE 1: modèle surparamétré, car

$$\text{si} \quad \Theta \leftarrow \Theta - \psi = \begin{bmatrix} \Theta_{1,:} - \psi \\ \vdots \\ \Theta_{K,:} - \psi \end{bmatrix}$$

2.2 Régression logistique multinomiale

Estimation par descente de gradient $(\Theta^{t+1} - \Theta^t) = -\nabla \ell(\Theta^t)$

NOTE 1: modèle surparamétré, car

$$\text{si} \quad \Theta \leftarrow \Theta - \psi = \begin{bmatrix} \Theta_{1,:} - \psi \\ \vdots \\ \Theta_{K,:} - \psi \end{bmatrix} \quad \text{alors} \quad \begin{bmatrix} \mathbb{P}^{\Theta}_{Y|X=x}(Y=1) \\ \vdots \\ \mathbb{P}^{\Theta}_{Y|X=x}(Y=K) \end{bmatrix} = \begin{bmatrix} \mathbb{P}^{\Theta-\psi}_{Y|X=x}(Y=1) \\ \vdots \\ \mathbb{P}^{\Theta-\psi}_{Y|X=x}(Y=K) \end{bmatrix}$$

2.2 Régression logistique multinomiale

Estimation par descente de gradient $(\Theta^{t+1} - \Theta^t) = -\nabla \ell(\Theta^t)$

NOTE 1: modèle surparamétré, car

$$\text{si} \quad \Theta \leftarrow \Theta - \psi = \begin{bmatrix} \Theta_{1,:} - \psi \\ \vdots \\ \Theta_{K,:} - \psi \end{bmatrix} \quad \text{alors} \quad \begin{bmatrix} \mathbb{P}^{\Theta}_{Y|X=x}(Y=1) \\ \vdots \\ \mathbb{P}^{\Theta}_{Y|X=x}(Y=K) \end{bmatrix} = \begin{bmatrix} \mathbb{P}^{\Theta-\psi}_{Y|X=x}(Y=1) \\ \vdots \\ \mathbb{P}^{\Theta-\psi}_{Y|X=x}(Y=K) \end{bmatrix}$$

On fixe donc $\Theta_{K,:} = \mathbf{1}$

2.2 Régression logistique multinomiale

Estimation par descente de gradient $(\Theta^{t+1} - \Theta^t) = -\nabla \ell(\Theta^t)$

NOTE 1: modèle surparamétré, car

NOTE 2: généralisation de la régression logistique binaire

2.2 Régression logistique multinomiale

Estimation par descente de gradient $(\Theta^{t+1} - \Theta^t) = -\nabla \ell(\Theta^t)$

NOTE 1: modèle surparamétré, car

NOTE 2: généralisation de la régression logistique binaire

NOTE 3: donne le même modèle de classification que K modèles de regression logistique binaire combinés en stratégie multiclasse one-vs-rest

$$f^1_{\theta^1} = \text{Reg. log. binaire } Y = 1 \text{ vs } Y \neq 1$$

...

$$f^K_{\theta^K} = \text{Reg. log. binaire } Y = K \text{ vs } Y \neq K$$

2.3 Régression linéaire

Modèle statistique de regression Prédire Y à partir de X

Jeu de données $x_{1:n}, y_{1:n}$

Un modèle de regression est une famille de lois de probabilités *conditionnelles*

$$\mathcal{M}_\Theta = \{\mathbb{P}^\theta_{Y|X=x} | \theta \in \Theta, x \in \mathscr{X}\}$$

Regression linéaire

$$Y \in \mathbb{R}, \quad X \in \mathbb{R}^p$$

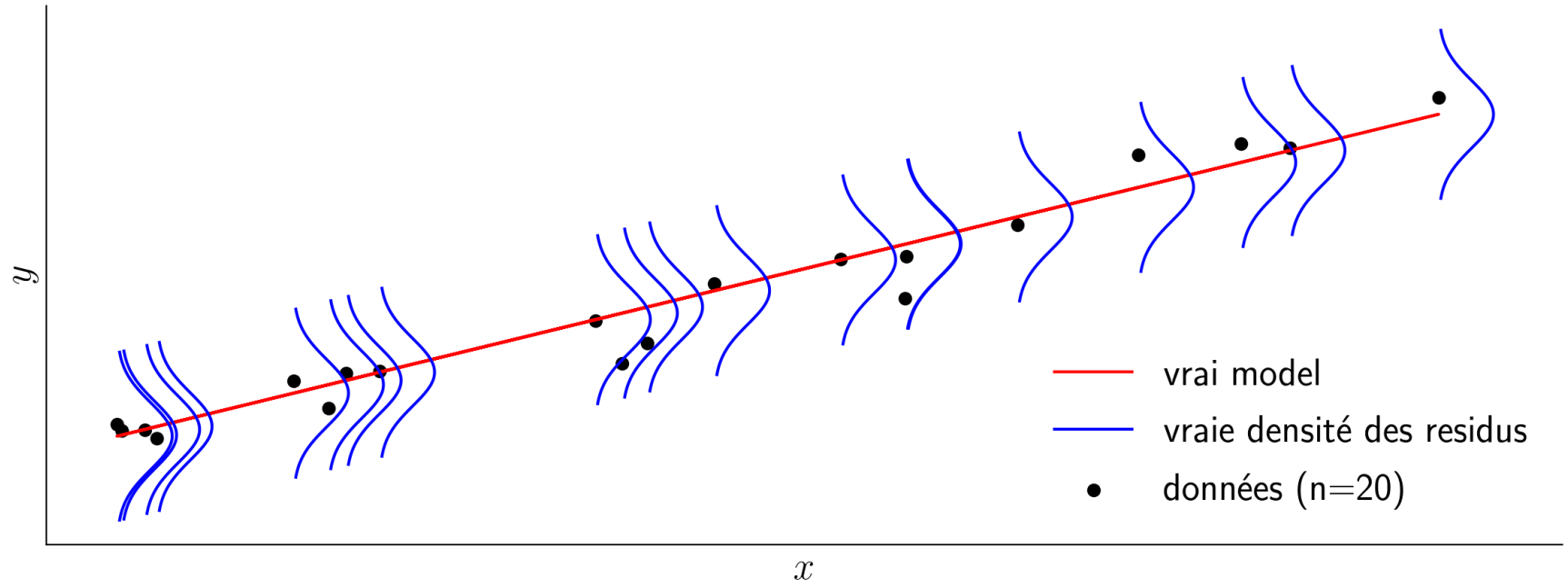
$$\mathbb{P}^\theta_{Y|X=x} = \mathcal{N}(\beta_0 + x^\top \beta, \sigma^2) \quad \theta = (\beta_0, \beta_1, \cdots, \beta_p, \sigma^2)$$

2.3 Régression linéaire

Regression linéaire

$$Y \in \mathbb{R}, \quad X \in \mathbb{R}^p$$

$$\mathbb{P}^{\theta}_{Y|X=x} = \mathcal{N}(\beta_0 + x^\top \beta, \sigma^2) \quad \theta = (\beta_0, \beta_1, \cdots, \beta_p, \sigma^2)$$



2.3 Régression linéaire

Estimation par maximum de vraisemblance

$$\mathcal{L}(\theta; \mathbf{x_{1:n}}, \mathbf{y_{1:n}}) = \prod_{i=1}^{n} p_{\mathrm{Y_i}|\mathrm{X_i}=\mathbf{x_i}}(y_i)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{\sigma^2 2\pi}} e^{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \mathbf{x_i}^\top \beta)^2}$$

2.3 Régression linéaire

Estimation par maximum de vraisemblance

$$\mathcal{L}(\theta; \mathbf{x_{1:n}}, \mathbf{y_{1:n}}) = \prod_{i=1}^{n} p_{\mathrm{Y_i}|\mathrm{X_i}=\mathbf{x_i}}(y_i)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{\sigma^2 2\pi}} e^{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \mathbf{x_i}^\top \beta)^2}$$

On peut montrer que $\theta \mapsto \mathcal{L}(\theta; \mathbf{x_{1:n}}, \mathbf{y_{1:n}})$ est concave.

$\theta \mapsto \ell(\theta) = -\log \mathcal{L}(\theta; \mathbf{x_{1:n}}, \mathbf{y_{1:n}})$ est convexe.

2.3 Régression linéaire

Estimation par maximum de vraisemblance

$$\mathcal{L}(\theta; \mathbf{x_{1:n}}, \mathbf{y_{1:n}}) = \prod_{i=1}^{n} p_{\mathrm{Y_i}|\mathrm{X_i}=\mathbf{x_i}}(y_i)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{\sigma^2 2\pi}} e^{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \mathbf{x_i}^\top \beta)^2}$$

On peut montrer que $\theta \mapsto \mathcal{L}(\theta; \mathbf{x_{1:n}}, \mathbf{y_{1:n}})$ est concave.

$\theta \mapsto \ell(\theta) = -\log \mathcal{L}(\theta; \mathbf{x_{1:n}}, \mathbf{y_{1:n}})$ est convexe.

$$\ell(\theta; \mathbf{x_{1:n}}, \mathbf{y_{1:n}}) = \frac{n}{2} 2\pi + \frac{n}{2} \log(\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \mathbf{x_i}^\top \beta)^2$$

2.3 Régression linéaire

Estimation par maximum de vraisemblance

$$\mathcal{L}(\theta; \mathbf{x_{1:n}}, \mathbf{y_{1:n}}) = \prod_{i=1}^{n} p_{\mathrm{Y_i}|\mathrm{X_i}=\mathbf{x_i}}(y_i)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{\sigma^2 2\pi}} e^{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \mathbf{x_i}^\top \beta)^2}$$

On peut montrer que $\theta \mapsto \mathcal{L}(\theta; \mathbf{x_{1:n}}, \mathbf{y_{1:n}})$ est concave.

$\theta \mapsto \ell(\theta) = -\log \mathcal{L}(\theta; \mathbf{x_{1:n}}, \mathbf{y_{1:n}})$ est convexe.

Moindres carrés !

$$\ell(\theta; \mathbf{x_{1:n}}, \mathbf{y_{1:n}}) = \frac{n}{2} 2\pi + \frac{n}{2} \log(\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \mathbf{x_i}^\top \beta)^2$$

2.3 Régression linéaire

Estimation par maximum de vraisemblance

Le minimum est réalisé aux dérivées nulles, i.e

$$\frac{\partial \ell}{\partial \beta_0} = 0,\ \frac{\partial \ell}{\partial \beta_1} = 0, \cdots,\ \frac{\partial \ell}{\partial \beta_p} = 0,\ \text{et}\ \frac{\partial \ell}{\partial \sigma^2} = 0$$

2.3 Régression linéaire

Estimation par maximum de vraisemblance

Le minimum est réalisé aux dérivées nulles, i.e

$$\frac{\partial \ell}{\partial \beta_0} = 0,\ \frac{\partial \ell}{\partial \beta_1} = 0, \cdots, \frac{\partial \ell}{\partial \beta_p} = 0, \text{ et } \frac{\partial \ell}{\partial \sigma^2} = 0$$

Tous calculs faits

$$\hat{\beta}(\mathbf{X_{1:n}}, \mathbf{Y_{1:n}}) = \left(\tilde{\mathbf{X}}_{\mathbf{1:n}}^{\top} \tilde{\mathbf{X}}_{\mathbf{1:n}}\right)^{-1} \tilde{\mathbf{X}}_{\mathbf{1:n}}^{\top} \mathbf{Y_{1:n}} \qquad (1)$$

Avec

$$\tilde{\mathbf{X}}_{\mathbf{1:n}} = \begin{bmatrix} 1 & \cdots & 1 \\ x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{bmatrix}$$

2.4 Pénalisation lasso

Observation: Les estimateurs précédents sans biais mais forte variance.
Idée: Diminuer la variance au prix de biais

$$\min_{\beta_0,\beta} = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x_i}^\top\beta)^2$$

$$\min_{\beta_0,\beta} = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x_i}^\top\beta)^2$$

$$\text{s.t} \sum_{i=0}^{p}|\beta_i| = \|\beta\|_1 \leq t$$

2.4 Pénalisation lasso

Observation: Les estimateurs précédents sans biais mais forte variance.
Idée: Diminuer la variance au prix de biais

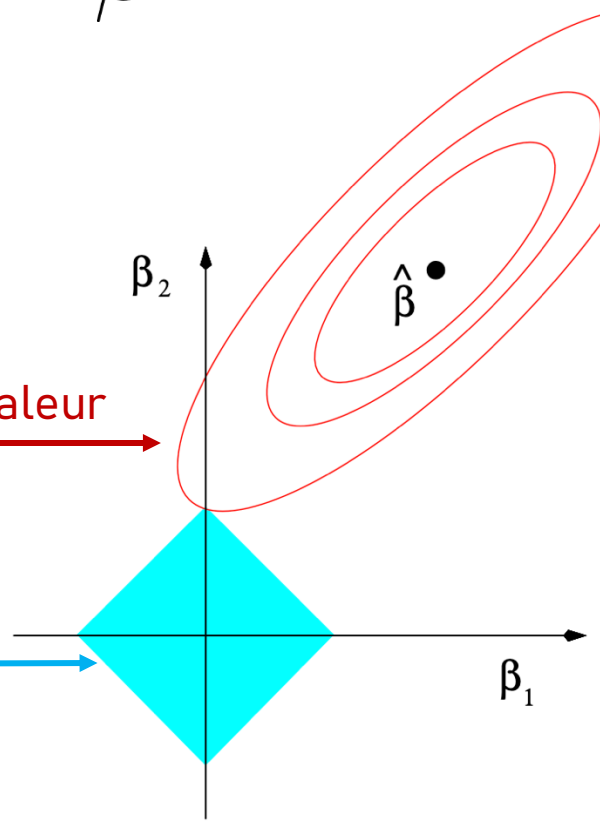
$\hat{\beta}$ Estimateur moindres carrés

$$\min_{\beta_0,\beta} = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x_i}^\top\beta)^2$$

$$\min_{\beta_0,\beta} = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x_i}^\top\beta)^2$$

$$\text{s.t} \sum_{i=0}^{p}|\beta_i| = \|\beta\|_1 \leq t$$

isovaleur



2.4 Pénalisation lasso

Observation: Les estimateurs précédents sans biais mais forte variance.
Idée: Diminuer la variance au prix de biais

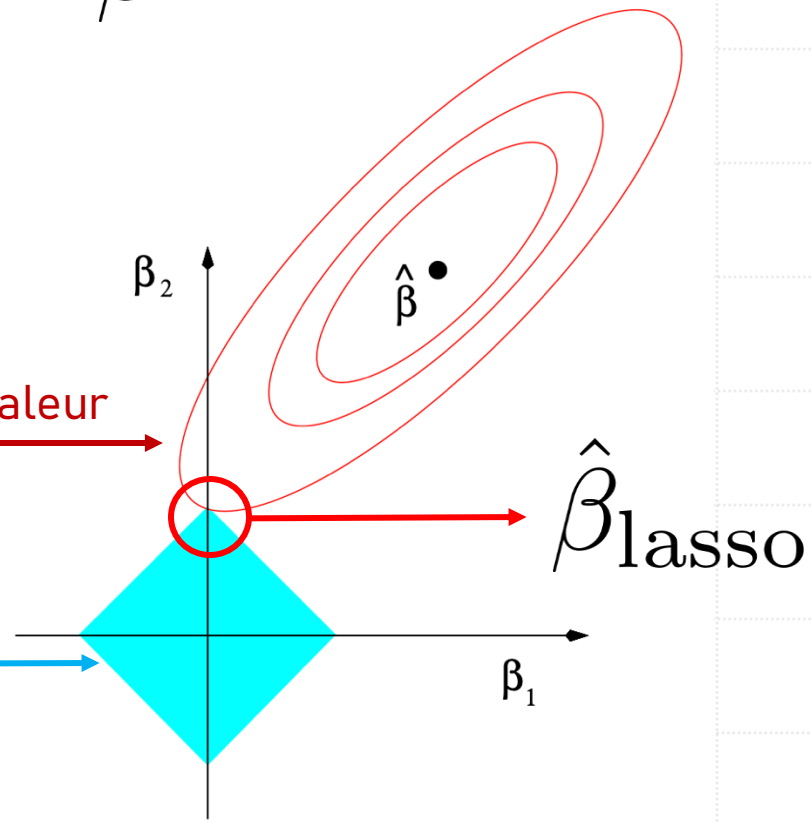
$$\min_{\beta_0,\beta} = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x_i}^\top\beta)^2$$

$$\min_{\beta_0,\beta} = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x_i}^\top\beta)^2$$

$$\text{s.t} \sum_{i=0}^{p}|\beta_i| = \|\beta\|_1 \leq t$$

$\hat{\beta}$ Estimateur moindres carrés

isovaleur



2.4 Pénalisation lasso

$$\min_{\beta_0,\beta} = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x_i}^\top\beta)^2$$

$$\text{s.t} \sum_{i=0}^{p}|\beta_i| = \|\beta\|_1 \leq t$$

Dualité de Lagrange $\Longleftrightarrow$

$$\min_{\beta_0,\beta} = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x_i}^\top\beta)^2 + \lambda\|\beta\|_1$$

1:1

Question: Comment choisir t/λ ?

Réponse: Par validation croisée!

2.4 Pénalisation lasso

$$\min_{\beta_0,\beta} = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x_i}^\top\beta)^2$$

$$\text{s.t} \sum_{i=0}^{p}|\beta_i| = \|\beta\|_1 \leq t$$

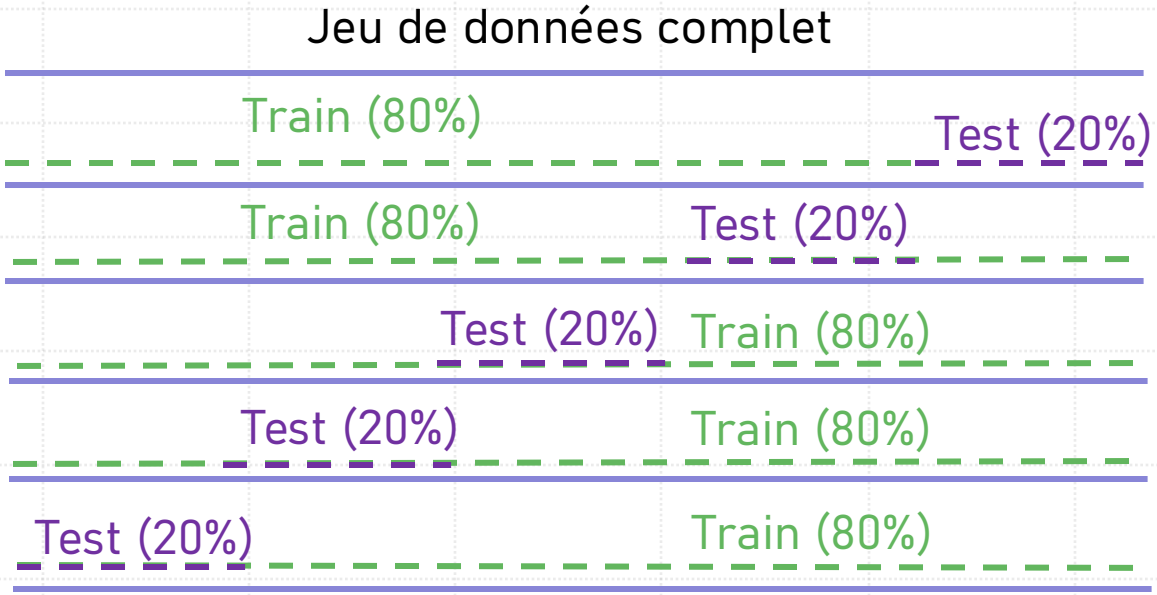
Dualité de Lagrange $\Longleftrightarrow$

$$\min_{\beta_0,\beta} = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x_i}^\top\beta)^2 + \lambda\|\beta\|_1$$

1:1

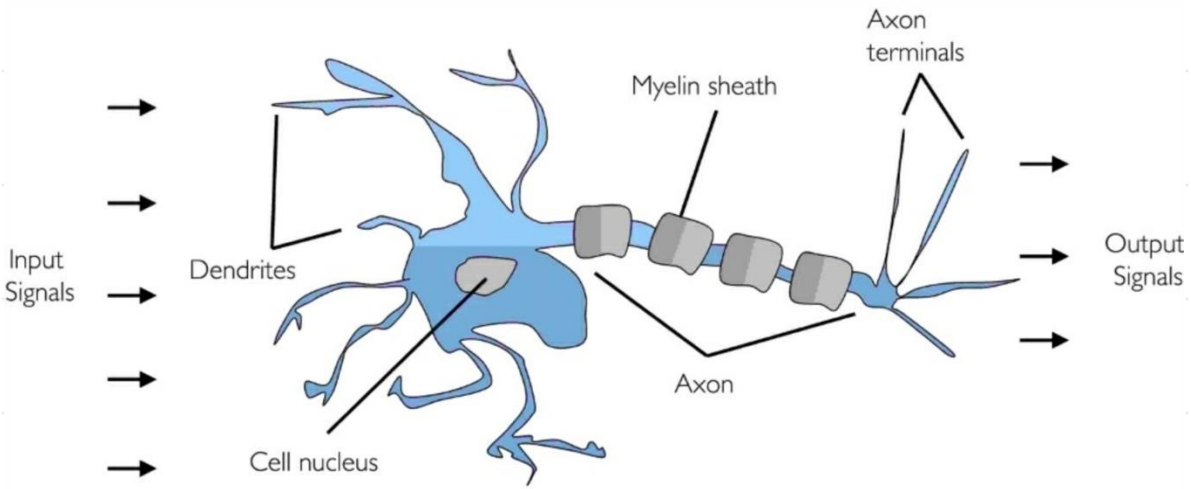
Question: Comment choisir t/λ ?

Réponse: Par validation croisée!



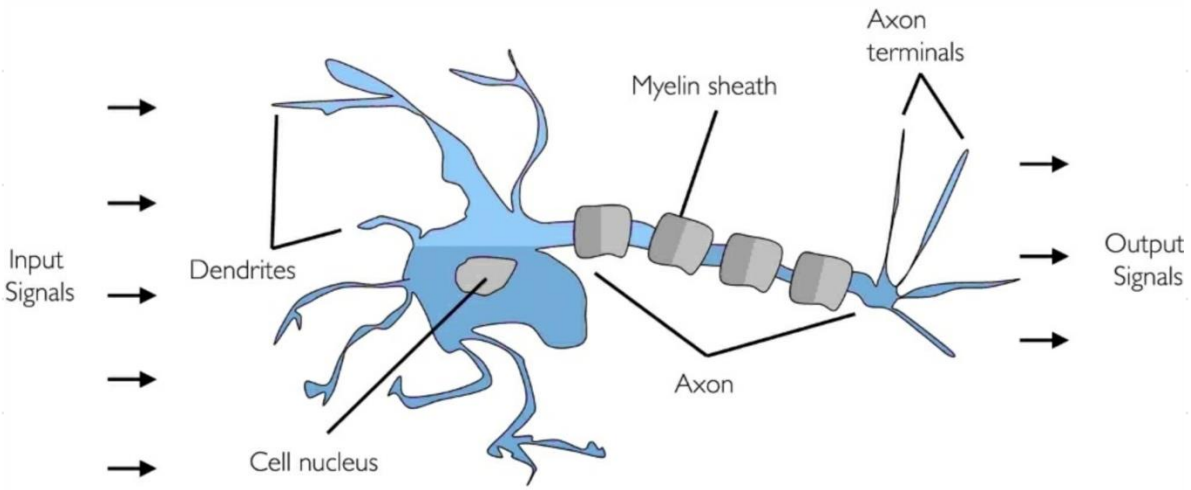
2.5 Réseau monocouche

Un neurone biologique:
si somme des signaux entrées > seuil
-> potential d'action généré
sinon inactif.



2.5 Réseau monocouche

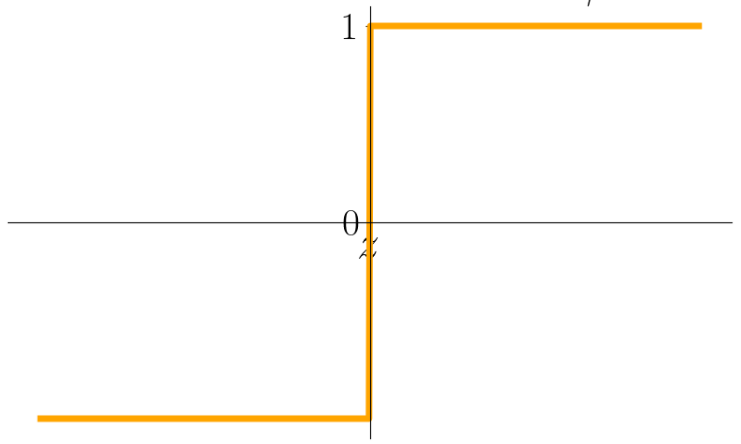
Un neurone biologique:
si somme des signaux entrées > seuil
-> potential d'action généré
sinon inactif.



Un neurone artificiel Rosenblatt 1957

- Entrées $x = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$, poids $w = \begin{bmatrix} w_1 \\ \vdots \\ w_p \end{bmatrix}$
- Entrées agrégées $z = x^\top w$
- Activation $\phi(z)$.

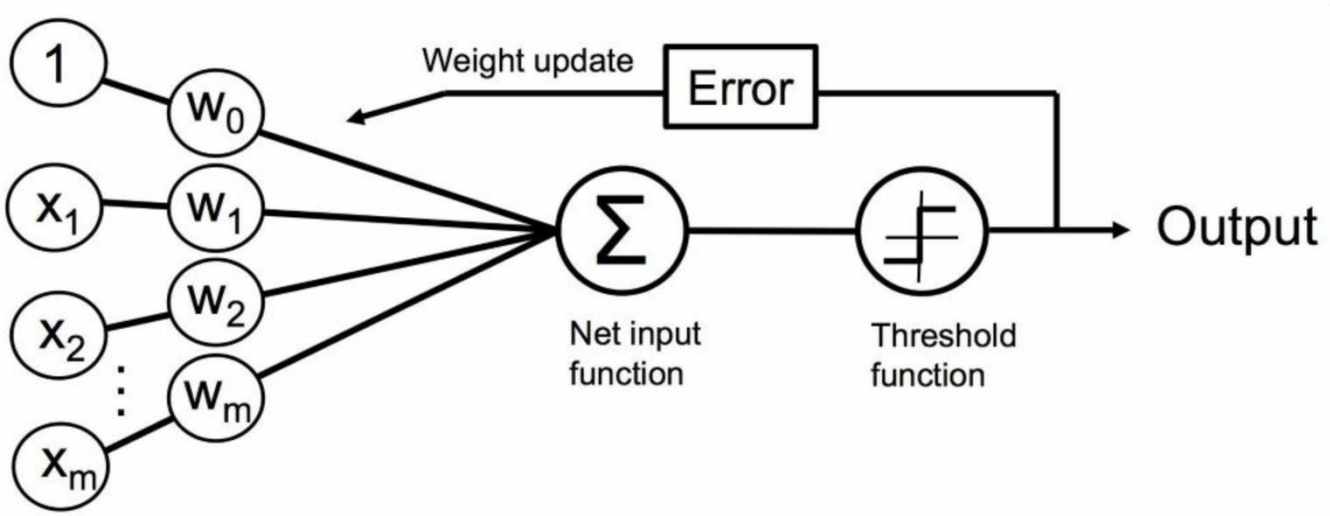
Fonction activation ϕ



2.5 Réseau monocouche

Règle du Perceptron: jeu de données $x_{1:n}, y_{1:n}$

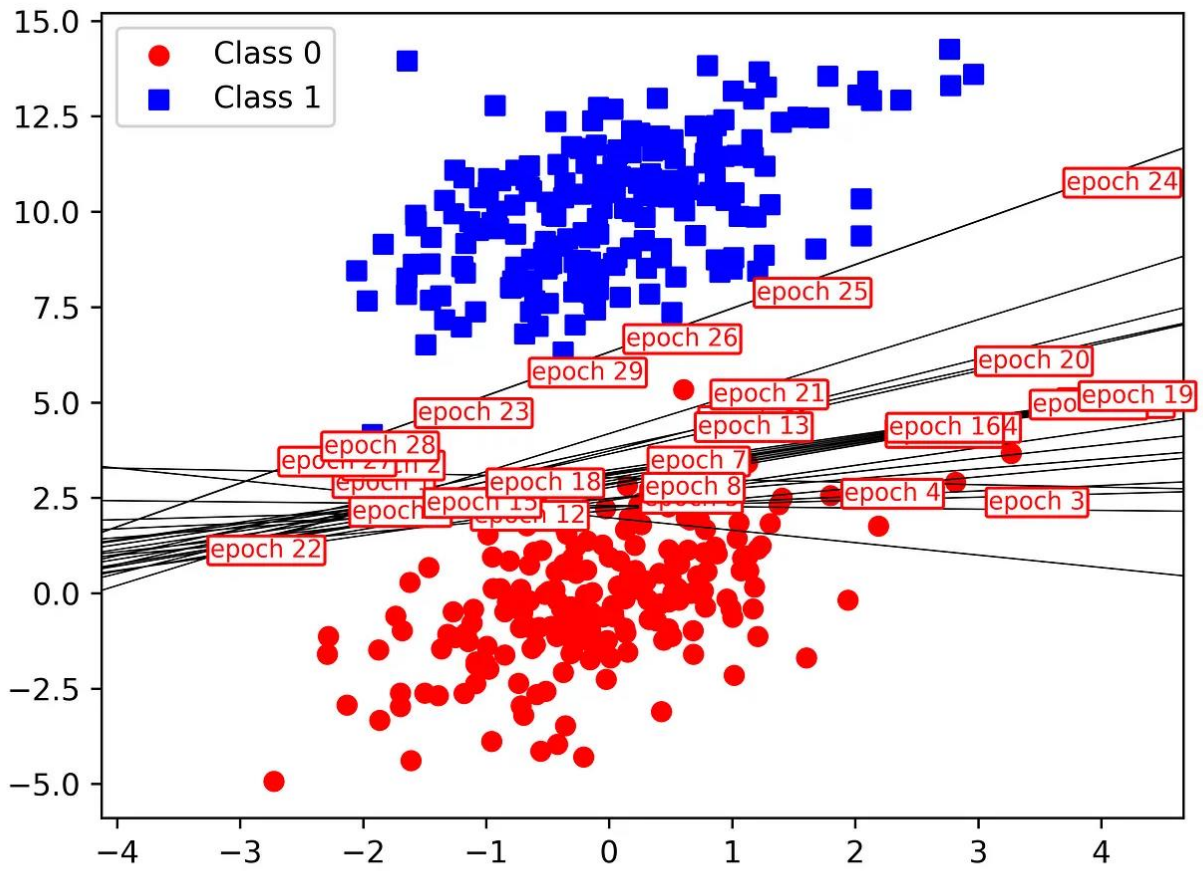
1. initialisation w aléatoire;
2. pour chaque $i = 1, \ldots, n$,
   (a) $\hat{y}_i = \phi(x_i^\top w)$
   (b) $\delta = \eta(y_i - \hat{y}_i)$.
   (c) $w = w + \delta x_i$



2.5 Réseau monocouche

Règle du Perceptron: jeu de données $x_{1:n}, y_{1:n}$

1. initialisation w aléatoire;
2. pour chaque $i = 1, \ldots, n$,
   (a) $\hat{y}_i = \phi(x_i^\top w)$
   (b) $\delta = \eta(y_i - \hat{y}_i)$.
   (c) $w = w + \delta x_i$

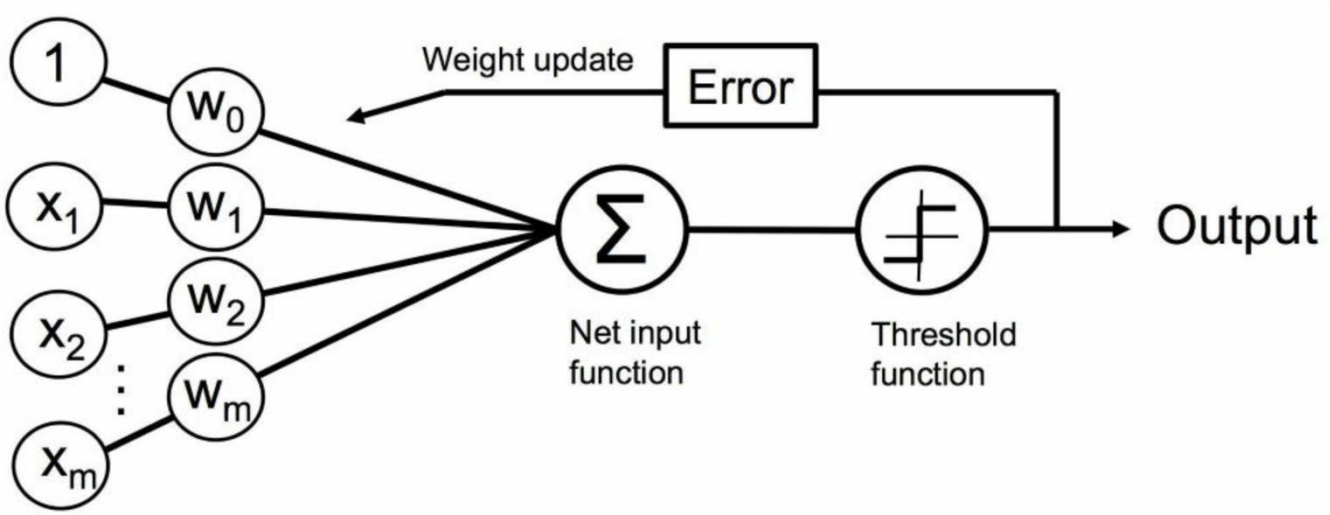


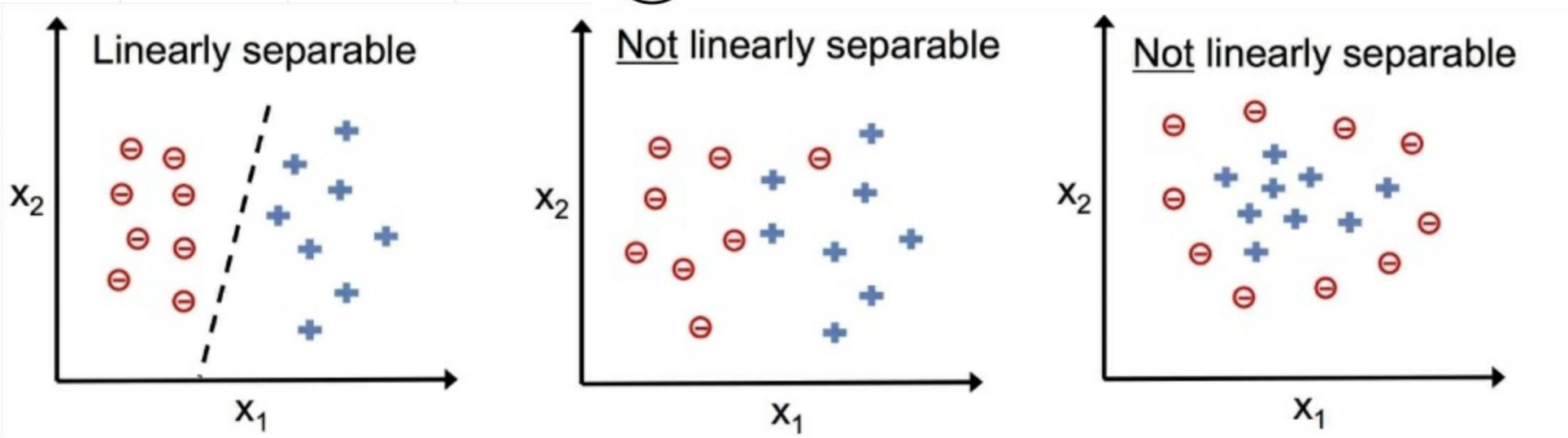
https://towardsdatascience.com/classification-with-rosenblatts-perceptron-e7f49e3af562

2.5 Réseau monocouche

Règle du Perceptron: jeu de données $x_{1:n}, y_{1:n}$

1. initialisation w aléatoire;
2. pour chaque $i = 1, \ldots, n$,
   (a) $\hat{y}_i = \phi(x_i^\top w)$
   (b) $\delta = \eta(y_i - \hat{y}_i)$.
   (c) $w = w + \delta x_i$

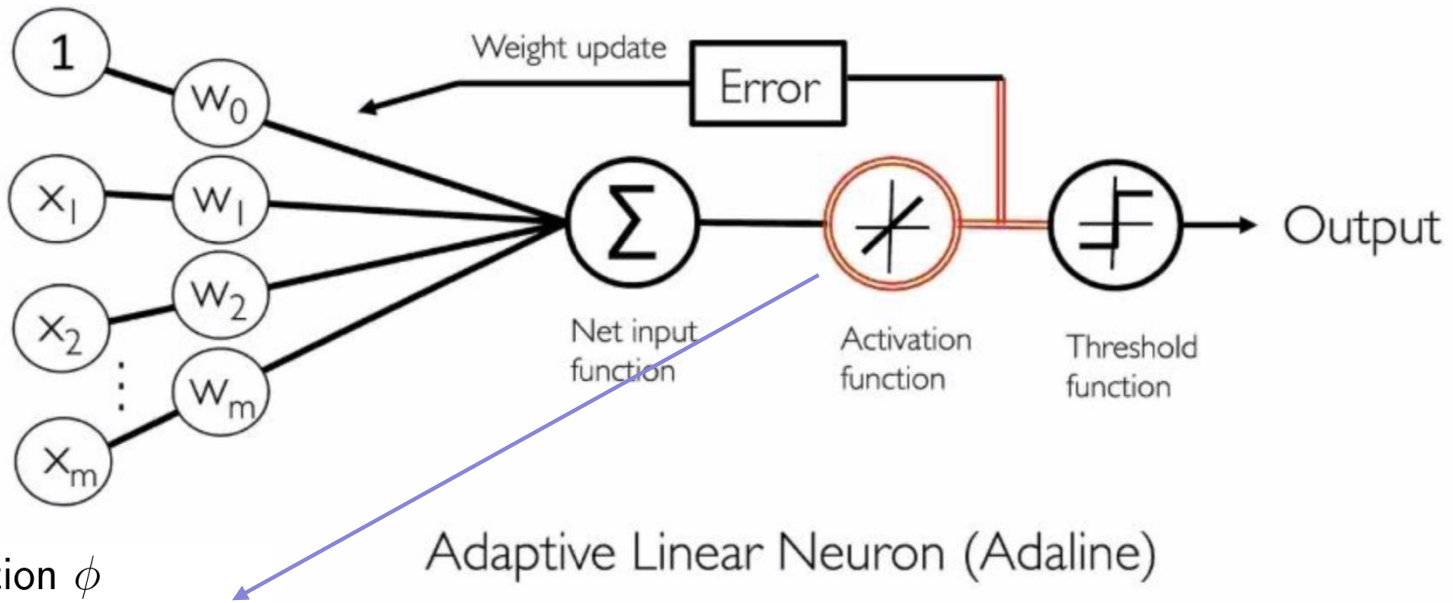




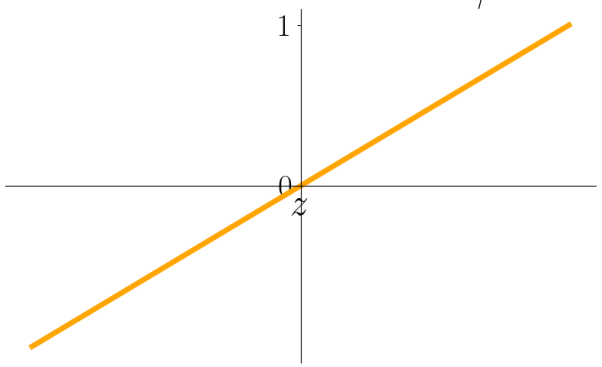
2.5 Réseau monocouche

Reseau Adaline: jeu de données $x_{1:n}, y_{1:n}$

1. initialisation w aléatoire;
2. pour chaque $i = 1, \ldots, n$,
   (a) $\hat{y}_i = \phi(x_i^\top w)$
   (b) $\delta = \eta(y_i - x_i^\top w)$.
   (c) $w = w + \delta x_i$



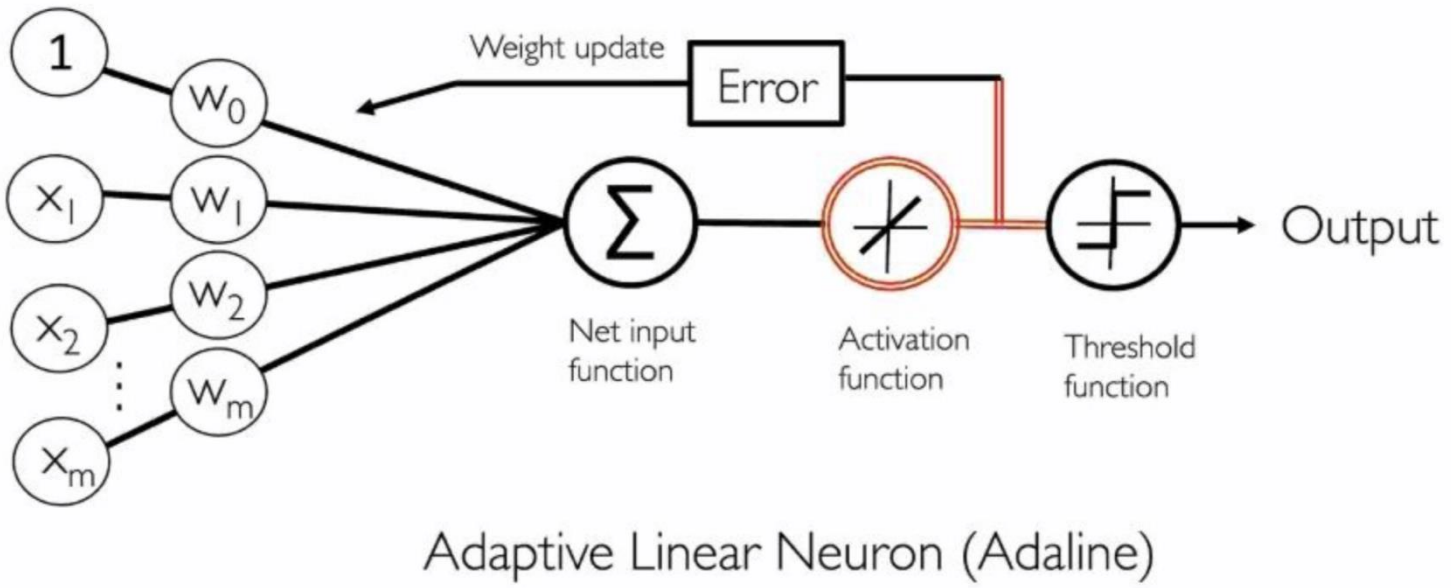
Fonction activation ϕ



2.5 Réseau monocouche

Reseau Adaline: jeu de données $x_{1:n}, y_{1:n}$

1. initialisation w aléatoire;
2. pour chaque $i = 1, \ldots, n$,
   (a) $\hat{y}_i = \phi(x_i^\top w)$
   (b) $\delta = \eta(y_i - x_i^\top w)$.
   (c) $w = w + \delta x_i$



Adaptive Linear Neuron (Adaline)

Pourquoi: $y_i - x_i^\top w$

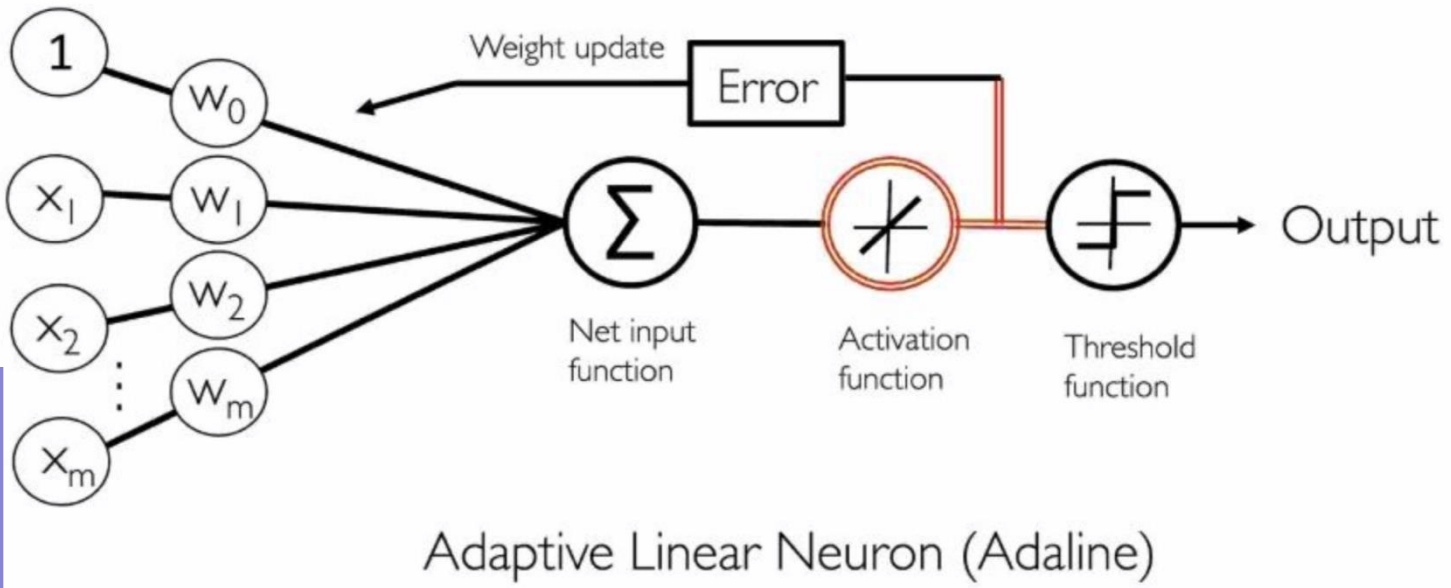
$$\hat{w}(x_{1:n}, y_{1:n}) \in \operatorname*{argmin}_{w \in \mathbb{R}^p} J(w) = \frac{1}{2} \sum_{i=1}^{n} (y_i - x_i^\top w)^2$$

descente de gradient sur un échantillon $\quad$ pas $= -\nabla J_i(w) = x_i(y_i - x_i^\top w)$

2.5 Réseau monocouche

Reseau Adaline: jeu de données $x_{1:n}, y_{1:n}$

1. initialisation w aléatoire;
2. pour chaque $i = 1, \ldots, n$,
   (a) $\hat{y}_i = \phi(x_i^\top w)$
   (b) $\delta = \eta(y_i - x_i^\top w)$
   (c) $w = w + \delta x_i$



Adaptive Linear Neuron (Adaline)

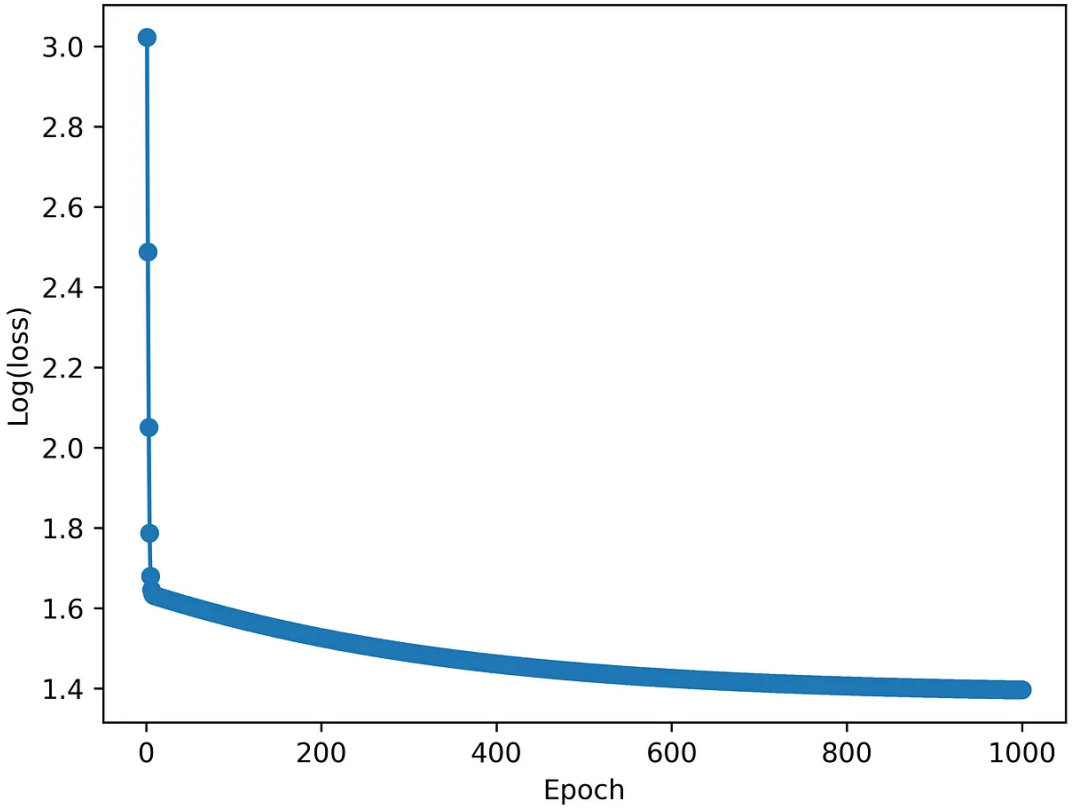
Pourquoi: $y_i - x_i^\top w$

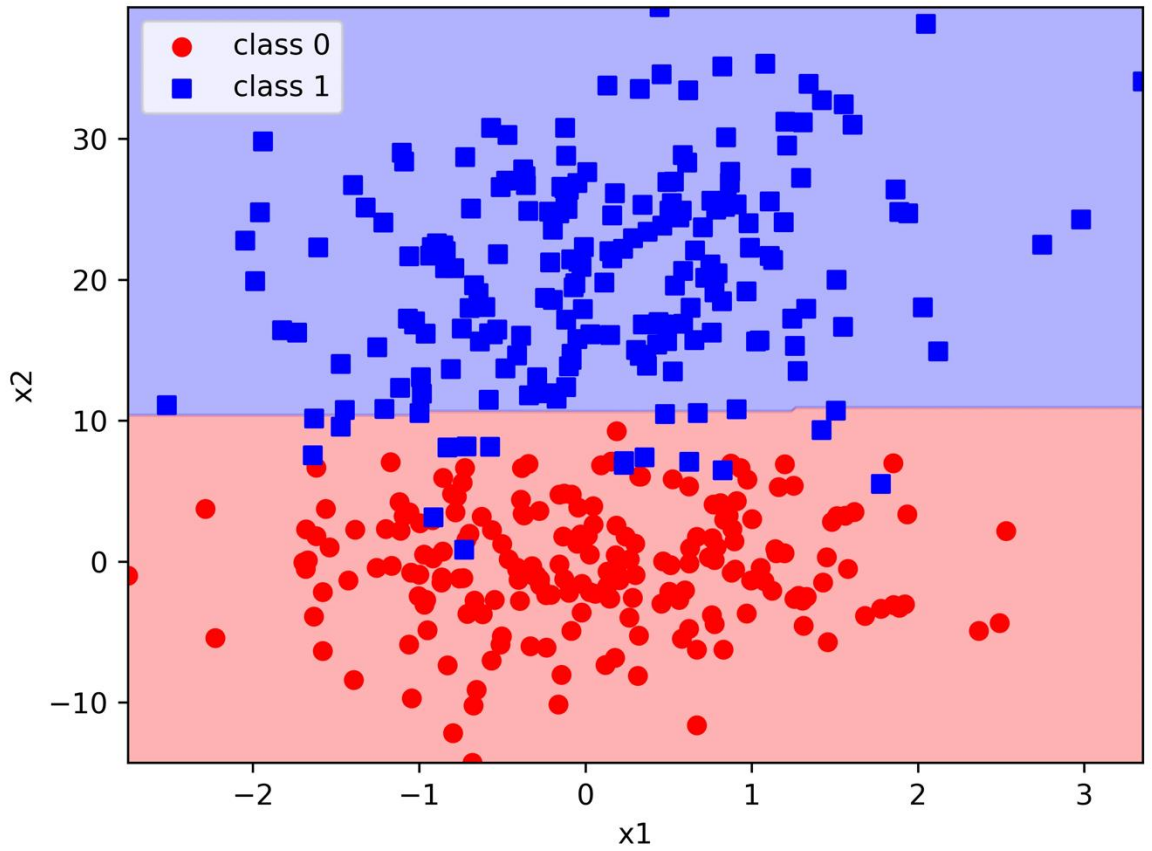
$$\hat{w}(x_{1:n}, y_{1:n}) \in \operatorname*{argmin}_{w \in \mathbb{R}^p} J(w) = \frac{1}{2} \sum_{i=1}^{n} (y_i - x_i^\top w)^2$$

descente de gradient sur un échantillon $\quad \text{pas} = -\nabla J_i(w) = x_i(y_i - x_i^\top w)$

2.5 Réseau monocouche

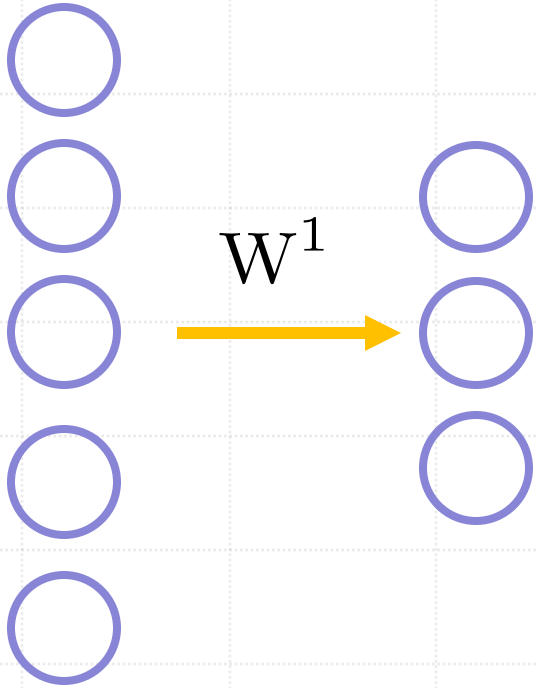
Reseau Adaline: jeu de données $x_{1:n}, y_{1:n}$





https://towardsdatascience.com/from-the-perceptron-to-adaline-1730e33d41c5

2.6 Réseau multicouches

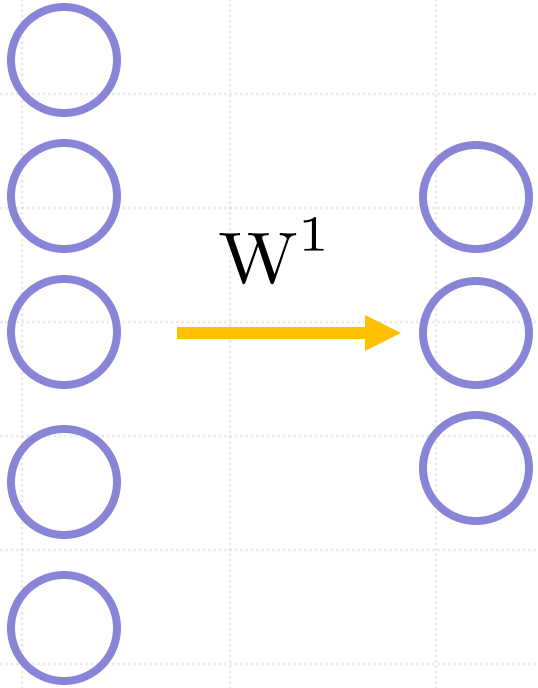


$a^1 = x$

$z^2 = W^1 a^1 + b^1$

$a^2 = \phi(z^2)$

2.6 Réseau multicouches



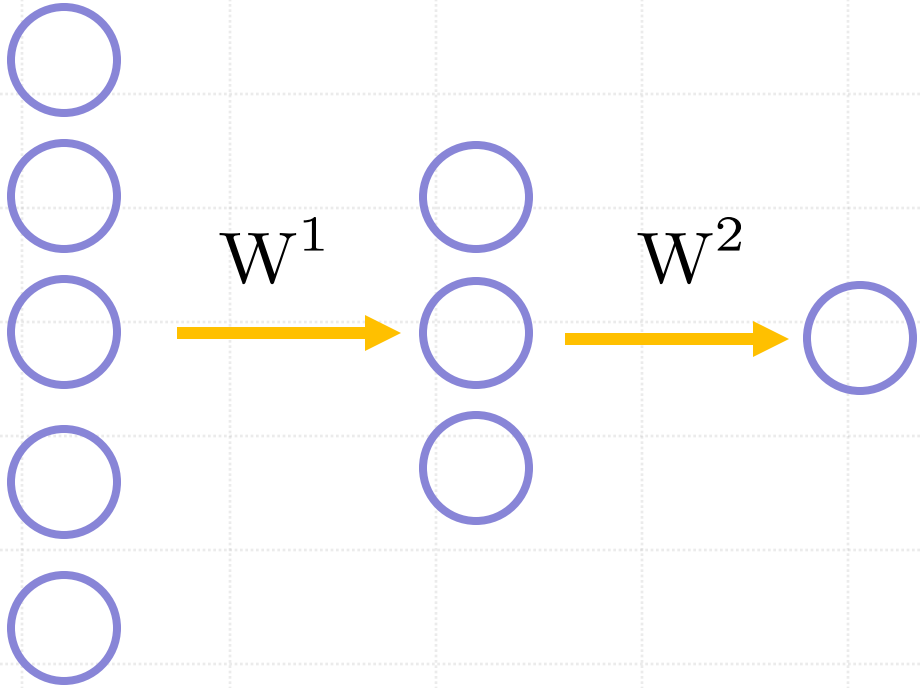
$$\mathrm{W}^1 = \begin{bmatrix} w_{11}^1 & \cdots & w_{15}^1 \\ \vdots & \ddots & \vdots \\ w_{31}^1 & \cdots & w_{35}^1 \end{bmatrix} \quad \mathrm{b}^1 = \begin{bmatrix} b_1^1 \\ \vdots \\ b_3^1 \end{bmatrix}$$

$$\mathrm{a}^1 = \mathrm{x}$$

$$\mathrm{z}^2 = \mathrm{W}^1\mathrm{a}^1 + \mathrm{b}^1$$

$$\mathrm{a}^2 = \phi(\mathrm{z}^2)$$

2.6 Réseau multicouches



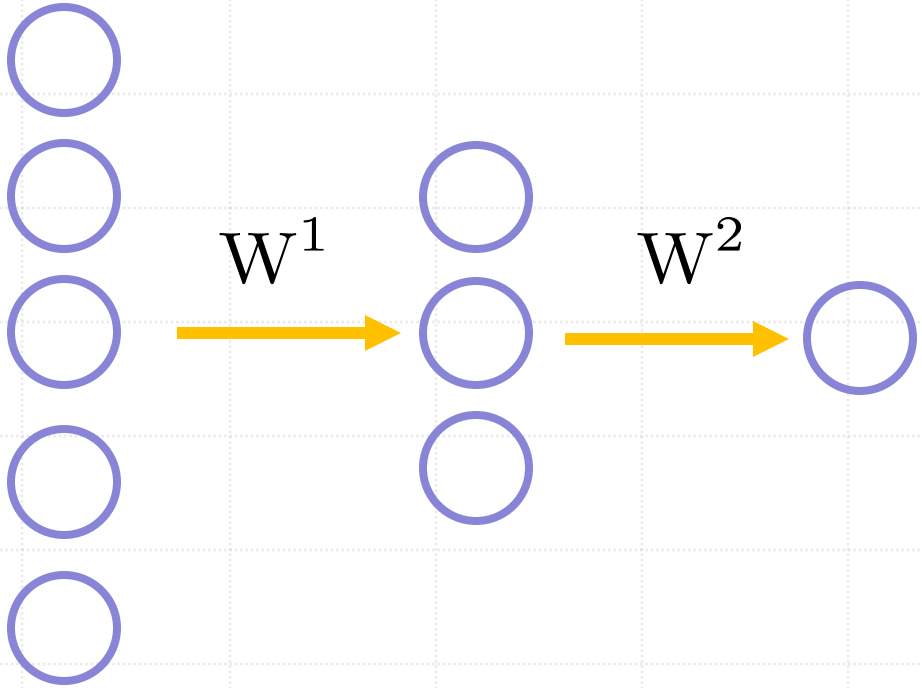
$$\mathrm{W}^1 = \begin{bmatrix} w_{11}^1 & \cdots & w_{15}^1 \\ \vdots & \ddots & \vdots \\ w_{31}^1 & \cdots & w_{35}^1 \end{bmatrix} \quad \mathrm{b}^1 = \begin{bmatrix} b_1^1 \\ \vdots \\ b_3^1 \end{bmatrix}$$

$$\mathrm{W}^2 = \begin{bmatrix} w_{11}^2 & \cdots & w_{13}^2 \end{bmatrix} \quad \mathrm{b}^2 = \begin{bmatrix} b_1^2 \end{bmatrix}$$

$$\mathrm{a}^1 = \mathrm{x} \qquad \mathrm{z}^2 = \mathrm{W}^1\mathrm{a}^1 + \mathrm{b}^1 \qquad \mathrm{z}^3 = \mathrm{W}^2\mathrm{a}^2 + \mathrm{b}^2$$

$$\mathrm{a}^2 = \phi(\mathrm{z}^2) \qquad \mathrm{a}^3 = \phi(\mathrm{z}^3)$$

2.6 Réseau multicouches



$$\mathrm{W}^1 = \begin{bmatrix} w_{11}^1 & \cdots & w_{15}^1 \\ \vdots & \ddots & \vdots \\ w_{31}^1 & \cdots & w_{35}^1 \end{bmatrix} \quad \mathrm{b}^1 = \begin{bmatrix} b_1^1 \\ \vdots \\ b_3^1 \end{bmatrix}$$

$$\mathrm{W}^2 = \begin{bmatrix} w_{11}^2 & \cdots & w_{13}^2 \end{bmatrix} \quad \mathrm{b}^2 = \begin{bmatrix} b_1^2 \end{bmatrix}$$

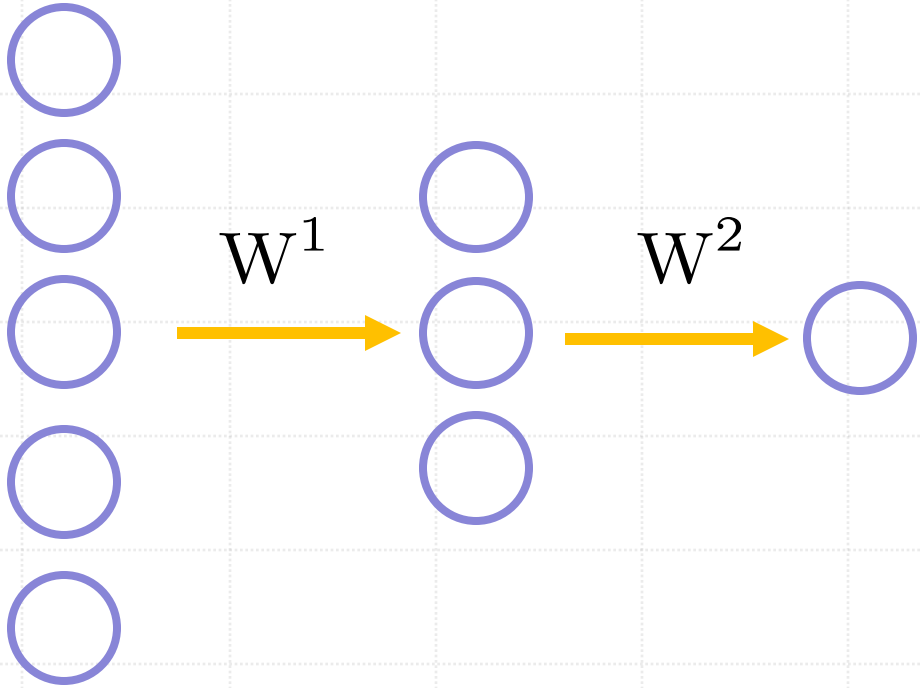
$$\mathrm{a}^1 = \mathrm{x} \qquad \mathrm{z}^2 = \mathrm{W}^1\mathrm{a}^1 + \mathrm{b}^1 \qquad \mathrm{z}^3 = \mathrm{W}^2\mathrm{a}^2 + \mathrm{b}^2$$

$$\mathrm{a}^2 = \phi(\mathrm{z}^2) \qquad \mathrm{a}^3 = \phi(\mathrm{z}^3)$$

Q1: combien de paramètres ?

Q2: que se passe-t-il si $\phi = \mathrm{Id}$?

2.6 Réseau multicouches



$$\mathrm{W}^1 = \begin{bmatrix} w_{11}^1 & \cdots & w_{15}^1 \\ \vdots & \ddots & \vdots \\ w_{31}^1 & \cdots & w_{35}^1 \end{bmatrix} \quad \mathrm{b}^1 = \begin{bmatrix} b_1^1 \\ \vdots \\ b_3^1 \end{bmatrix}$$

$$\mathrm{W}^2 = \begin{bmatrix} w_{11}^2 & \cdots & w_{13}^2 \end{bmatrix} \quad \mathrm{b}^2 = \begin{bmatrix} b_1^2 \end{bmatrix}$$

$$\mathrm{a}^1 = \mathrm{x} \qquad \mathrm{z}^2 = \mathrm{W}^1\mathrm{a}^1 + \mathrm{b}^1 \qquad \mathrm{z}^3 = \mathrm{W}^2\mathrm{a}^2 + \mathrm{b}^2$$

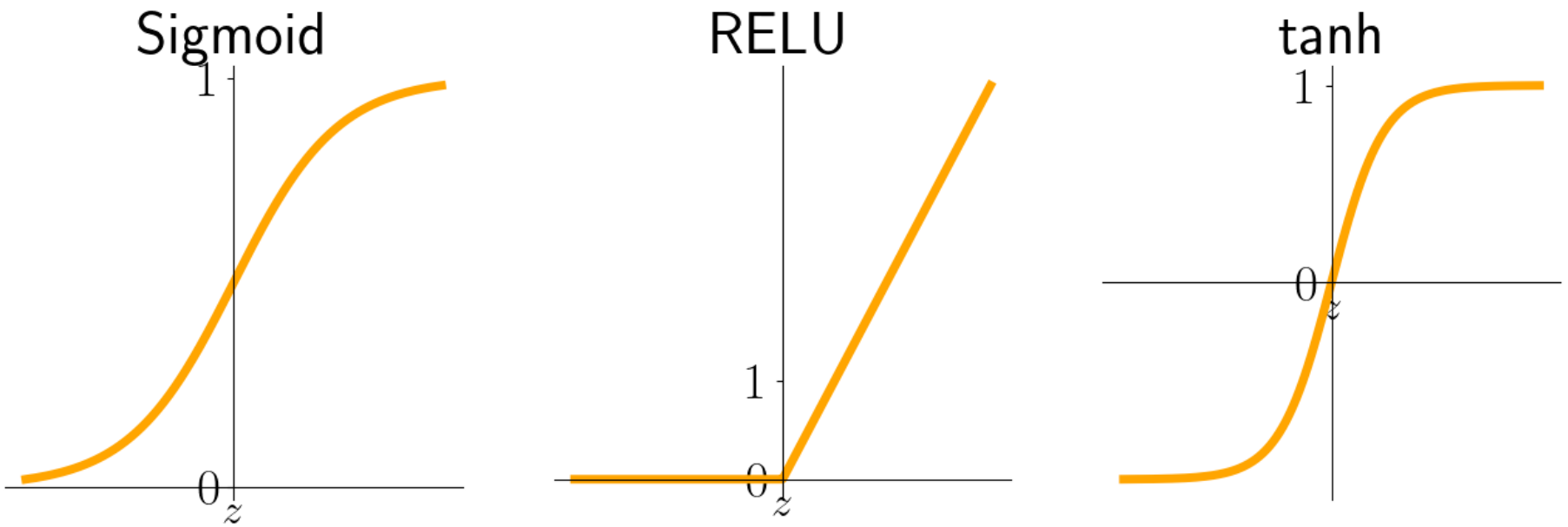
$$\mathrm{a}^2 = \phi(\mathrm{z}^2) \qquad \mathrm{a}^3 = \phi(\mathrm{z}^3)$$

Q1: combien de paramètres ? $3 \times 5 + 3 + 1 \times 3 + 1 = 22$

Q2: que se passe-t-il si $\phi = \mathrm{Id}$? $\mathrm{a}^3 = \mathrm{W}^2\mathrm{W}^1\mathrm{x} + \mathrm{W}^2\mathrm{b}^1 + \mathrm{b}^2$

$= \mathrm{mx} + \mathrm{p}$

2.6 Réseau multicouches



Exemples de fonctions d'activation non linéaires

2.6 Réseau multicouches

Comment apprendre les paramètres $\Theta = [\mathrm{W}^1, \mathrm{b}^1, \mathrm{W}^2, \mathrm{b}^2]$?

Jeu de données $x_{1:n}, y_{1:n}$ Modèle $f_{\Theta}(x) = a^3(x)$

2.6 Réseau multicouches

Comment apprendre les paramètres $\Theta = [\mathrm{W}^1, \mathrm{b}^1, \mathrm{W}^2, \mathrm{b}^2]$?

Jeu de données $x_{1:n}, y_{1:n}$ Modèle $f_{\Theta}(x) = a^3(x)$

-> **Choix** d'une function de coût

Exemples: $y_i \in \mathbb{R}$ $J(\Theta) = \frac{1}{2n} \sum_{i=1}^{n} \|y_i - f_{\Theta}(x)\|^2$ *tf.keras.losses.MeanSquaredError*

2.6 Réseau multicouches

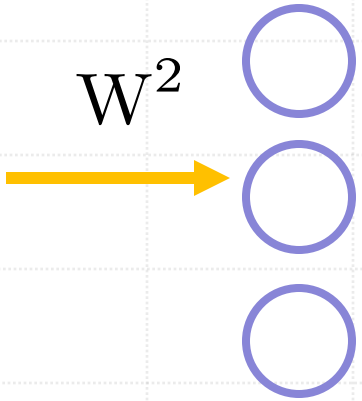
Comment apprendre les paramètres $\Theta = [\mathrm{W}^1, \mathrm{b}^1, \mathrm{W}^2, \mathrm{b}^2]$?

Jeu de données $x_{1:n}, y_{1:n}$ Modèle $f_\Theta(x) = a^3(x)$

-> **Choix** d'une function de coût

Exemples: $y_i \in \mathbb{R}$ $J(\Theta) = \frac{1}{2n} \sum_{i=1}^{n} \|y_i - f_\Theta(x)\|^2$ *tf.keras.losses.MeanSquaredError*

$y_i \in \{1, \cdots, K\}$ $J(\Theta) = \frac{-1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}_{y_i=k} \log\left((f_\Theta(x_i))_k\right)$ *tf.keras.losses.CategoricalCrossEntropy*



$$\mathrm{z}^3 = \mathrm{W}^2\mathrm{a}^2 + \mathrm{b}^2$$
$$\mathrm{a}^3 = \phi(\mathrm{z}^3)$$

$$\phi(\mathrm{z}) = \begin{bmatrix} \frac{e^{z1}}{\sum_{k=1}^{K} e^{z_k}} \\ \vdots \\ \frac{e^{zK}}{\sum_{k=1}^{K} e^{z_k}} \end{bmatrix}$$

"probabilities" (!)

Softmax activation

2.6 Réseau multicouches

Comment apprendre les paramètres $\Theta = [\mathrm{W}^1, \mathrm{b}^1, \mathrm{W}^2, \mathrm{b}^2]$?

Jeu de données $x_{1:n}, y_{1:n}$ Modèle $f_{\Theta}(x) = a^3(x)$

-> **Choix** d'une function de coût

-> Apprentissage par descente de gradient

$$\Theta^{t+1} = \Theta^t - \eta_t \nabla J(\Theta^t)$$

2.6 Réseau multicouches

Comment apprendre les paramètres $\Theta = [\mathrm{W}^1, \mathrm{b}^1, \mathrm{W}^2, \mathrm{b}^2]$?

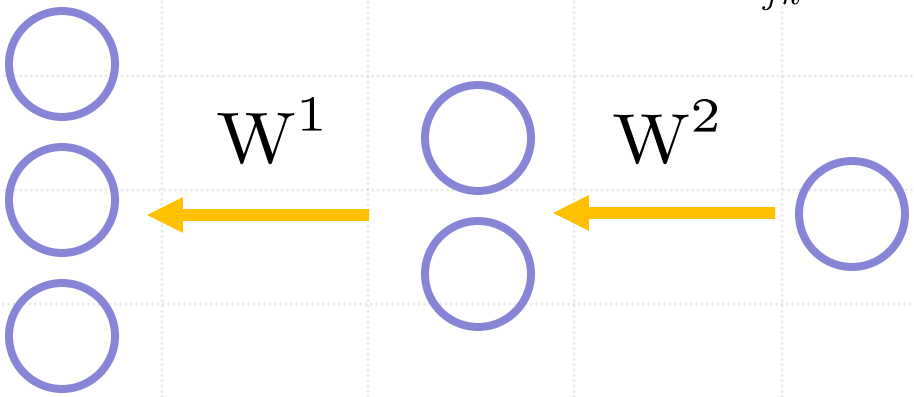
Jeu de données $x_{1:n}, y_{1:n}$ Modèle $f_{\Theta}(x) = a^3(x)$

-> **Choix** d'une function de coût

-> Apprentissage par descente de gradient

$$\Theta^{t+1} = \Theta^t - \eta_t \nabla J(\Theta^t)$$

Difficulté 1: comment calculer $\frac{\partial J}{\partial w^l_{jk}}(\Theta^t), \frac{\partial J}{\partial b^l_j}(\Theta^t)$ pour tout j, k, l ?



$\epsilon_i = y_i - f_{\Theta^t}(x_i)$ → Stochastic descent

$\epsilon_1 = y_i - f_{\Theta^t}(x_i)$

$\dots$

$\epsilon_m = y_i - f_{\Theta^t}(x_n)$

Batch descent

2.6 Réseau multicouches

Comment apprendre les paramètres $\Theta = [\mathrm{W}^1, \mathrm{b}^1, \mathrm{W}^2, \mathrm{b}^2]$?

Jeu de données $x_{1:n}, y_{1:n}$ Modèle $f_{\Theta}(x) = a^3(x)$

-> **Choix** d'une function de coût

-> Apprentissage par descente de gradient

$$\Theta^{t+1} = \Theta^t - \eta_t \nabla J(\Theta^t)$$

Difficulté 1: comment calculer $\frac{\partial J}{\partial w^l_{jk}}(\Theta^t), \frac{\partial J}{\partial b^l_j}(\Theta^t)$ pour tout j, k, l ?

Difficulté 2: comment choisir η_t ?

Constant, RMSProp, Adam, Adamax, Adadelta, Nadam, Adafactor, ...