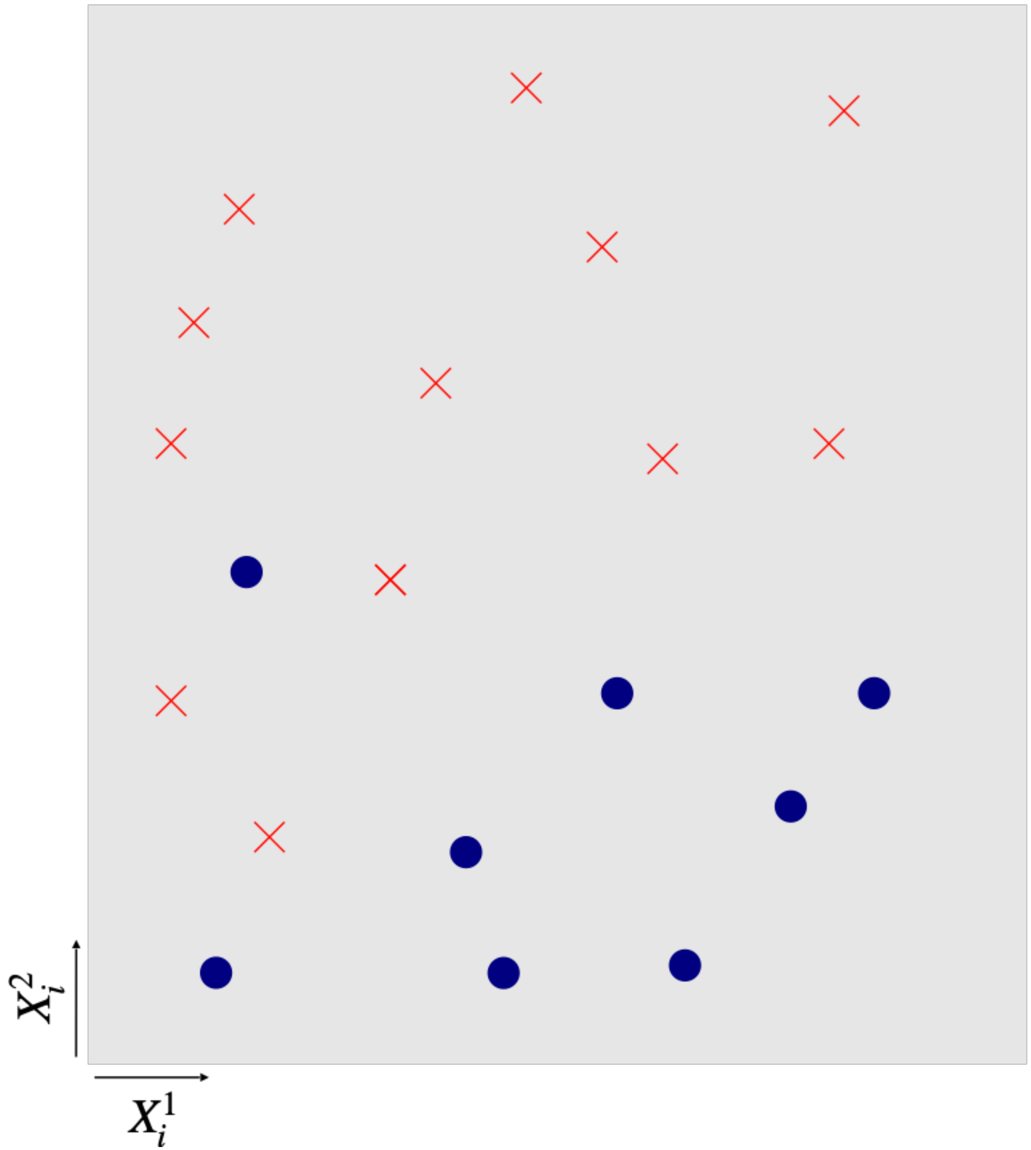
Aide au diagnostic

Base d'apprentissage	
Patient 1 : • Age = 40 • Globule Blancs/L = 6	Sain
Patient 2 : • Age = 28 • Globule Blancs/L = 12	Rhume
Patient N : • Age = 57 • Globule Blancs/L = 8	Sain

Nouveau Patient :
- Age = 34
- Globule Blancs/L = 5

→ Sain ou rhume ???

Apprentissage supervisé — classification



Observations d'entrée (*X*) :

- *n* observations $X_i \in \mathbb{R}^p$
- Ici *n = 20* et *p = 2*

Observations de sortie (*Y*) :

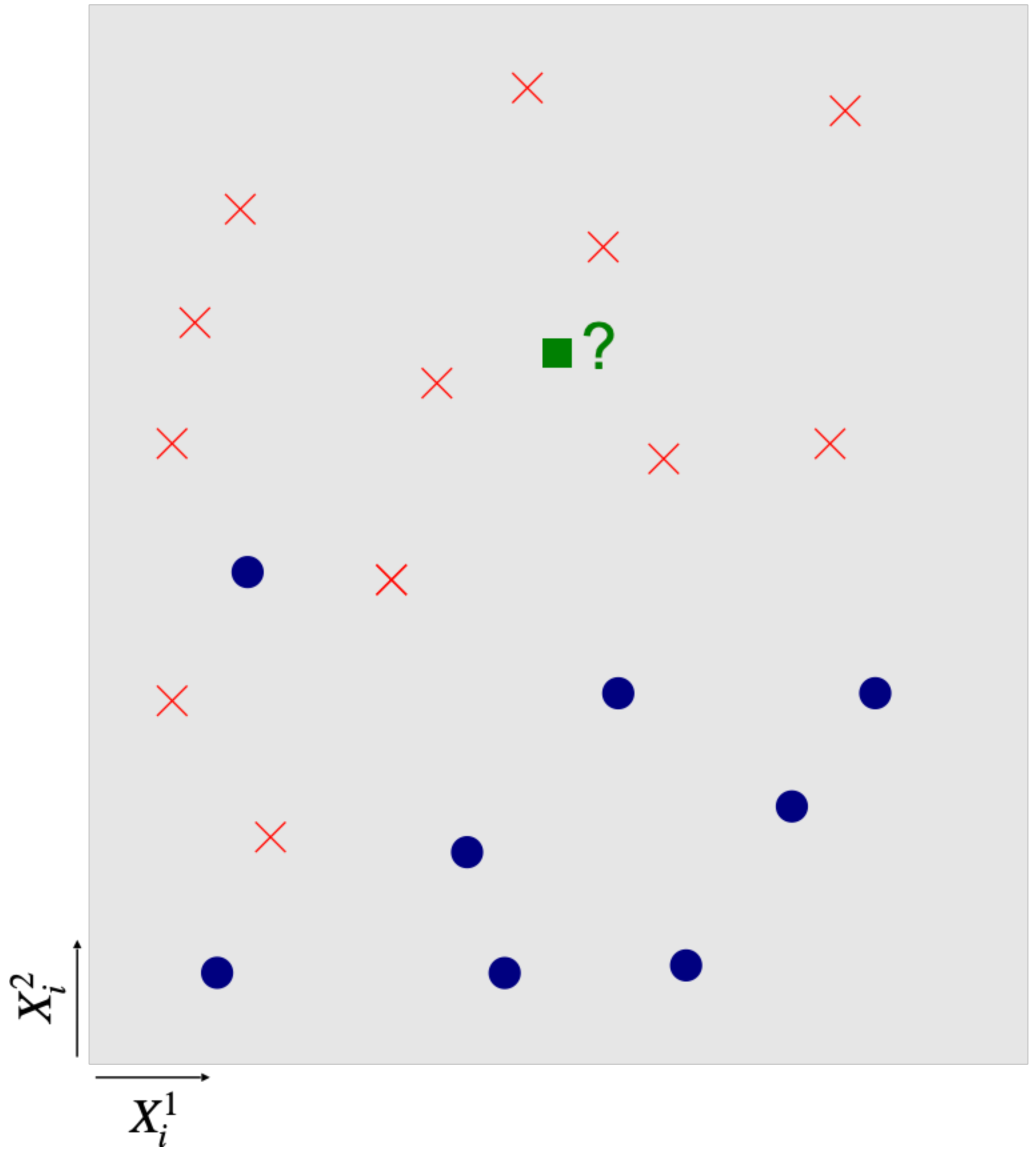
- *n* Labels $Y_i \in \{-1,1\}$
- × $Y_i = 1$
- ● $Y_i = -1$

<u>Dans notre exemple</u> :

$i \longrightarrow$ Patient de la base d'apprentissage
$X_i^1 \longrightarrow$ Age
$X_i^2 \longrightarrow$ Globule Blancs/L
$Y_i \longrightarrow$ Sain ou rhume

5.1 : Préambule – Classification

Apprentissage supervisé — classification



Observations d’entrée (*X*) :

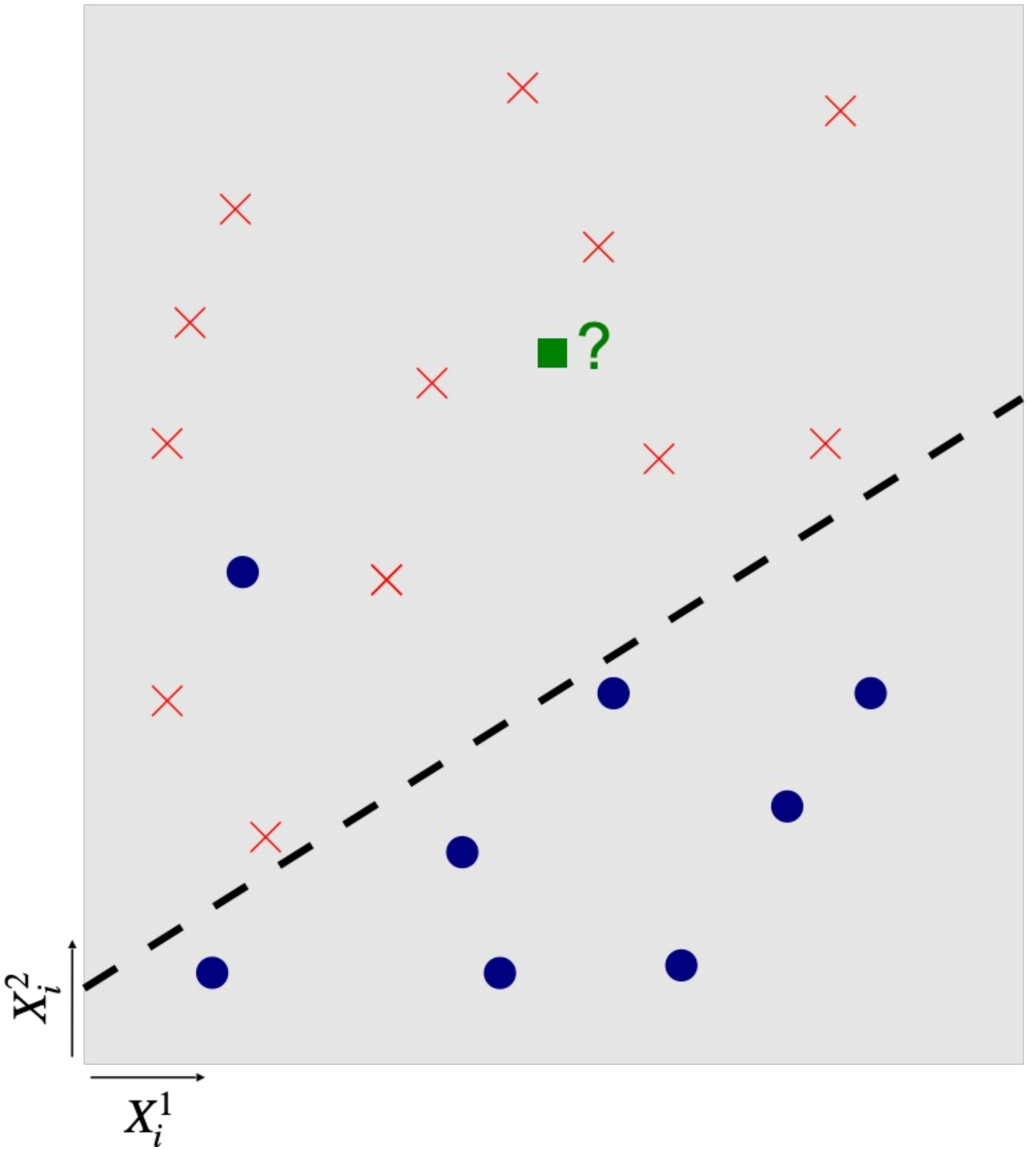
- *n* observations $X_i \in \mathbb{R}^p$
- Ici *n = 20* et *p = 2*

Observations de sortie (*Y*) :

- *n* Labels $Y_i \in \{-1,1\}$
- $Y_i = 1$
- $Y_i = -1$

Label le plus probable de ?

Apprentissage supervisé — classification



Observations d'entrée (*X*) :
- *n* observations $X_i \in \mathbb{R}^p$
- Ici *n = 20* et *p = 2*

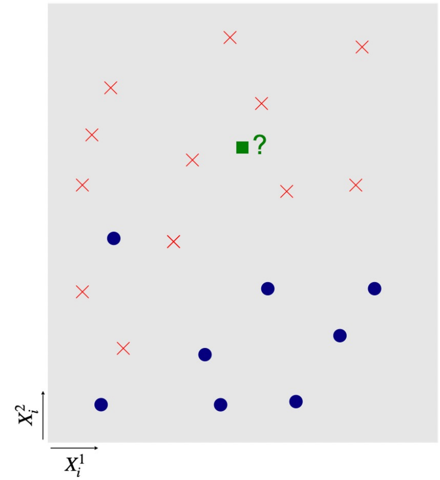
Observations de sortie (*Y*) :
- *n* Labels $Y_i \in \{-1,1\}$
- × $Y_i = 1$
- ● $Y_i = -1$

Label le plus probable de ■ ?

1. **Choix d'un modèle** pour séparer les données d'apprentissage, i.e. les ● et les × .
2. **Apprentissage des paramètres** optimaux
3. Une fois les paramètres du modèle appris, **prédiction** extrêmement simple et rapide de ■ .

n observations d'apprentissage

- Entrée : $x_i = (x_i^1, \dots, x_i^p) \in \mathbb{R}^p$
- Sortie : $y_i \in \{-1,1\}$

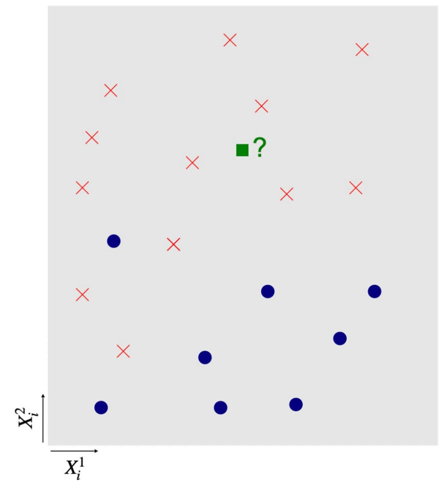


On note $\mathbb{P}(Y=1|X)$ la loi conditionnelle que Y soit égal à 1 sachant X.

5.2 : Régression logistique simple

n observations d'apprentissage

- Entrée : $x_i = (x_i^1, \dots, x_i^p) \in \mathbb{R}^p$
- Sortie : $y_i \in \{-1,1\}$



On note $\mathbb{P}(Y=1|X)$ la loi conditionnelle que Y soit égal à 1 sachant X.

<u>Remarque importante</u> : Nous pourrions utiliser le modèle linéaire directement …

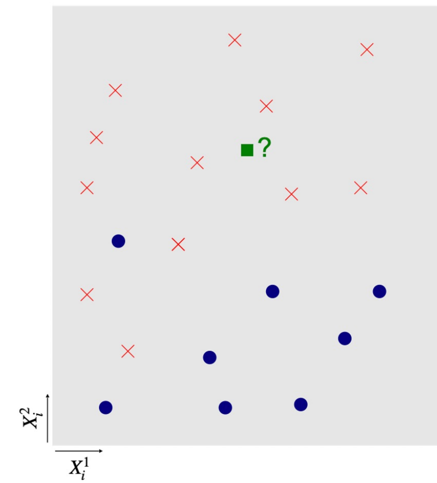
$$\mathbb{P}(Y=1|X) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

A apprendre Connu

… mais il serait très difficile de garantir que les probabilités aient des valeurs dans [0,1]

n observations d'apprentissage

- Entrée : $x_i = (x_i^1, \dots, x_i^p) \in \mathbb{R}^p$
- Sortie : $y_i \in \{-1,1\}$

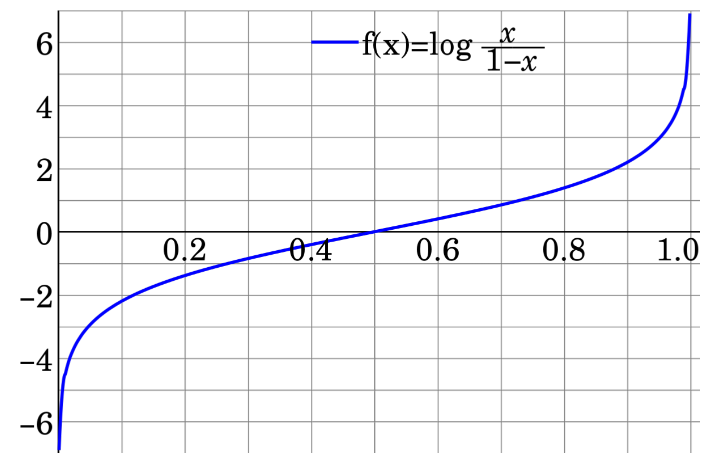


On note $\mathbb{P}(Y = 1|X)$ la loi conditionnelle que Y soit égal à 1 sachant X.

On suppose alors que : $\ln \frac{\mathbb{P}(Y=1|X)}{1-\mathbb{P}(Y=1|X)} = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$

A apprendre

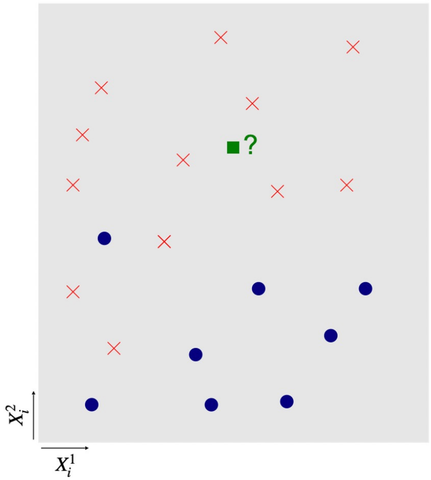
Connu



Fonction logit

n observations d'apprentissage

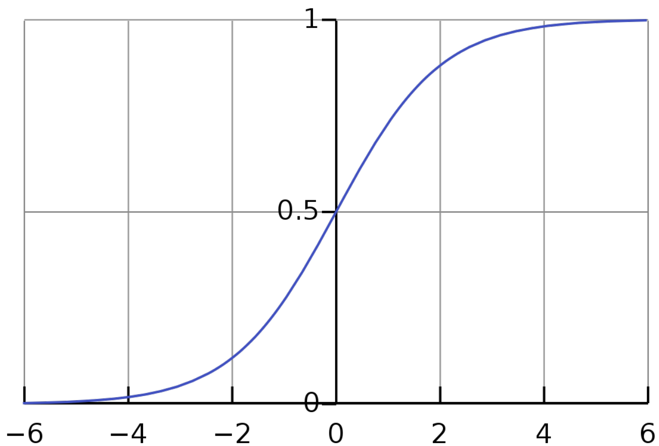
- Entrée : $x_i = (x_i^1, \ldots, x_i^p) \in \mathbb{R}^p$
- Sortie : $y_i \in \{-1,1\}$



On note $\mathbb{P}(Y = 1|X)$ la loi conditionnelle que Y soit égal à 1 sachant X.

On suppose alors que : $$\ln \frac{\mathbb{P}(Y=1|X)}{1-\mathbb{P}(Y=1|X)} = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

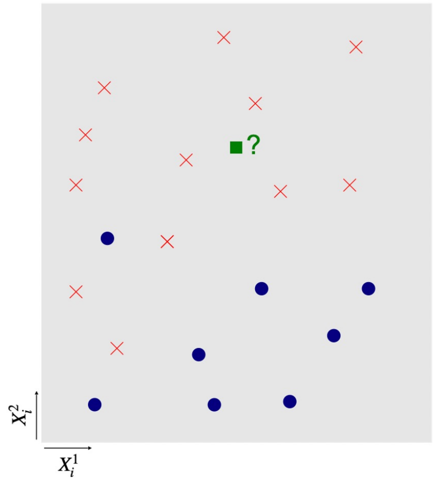
$$\mathbb{P}(Y=1|X) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}}$$



Fonction logistique

n observations d'apprentissage

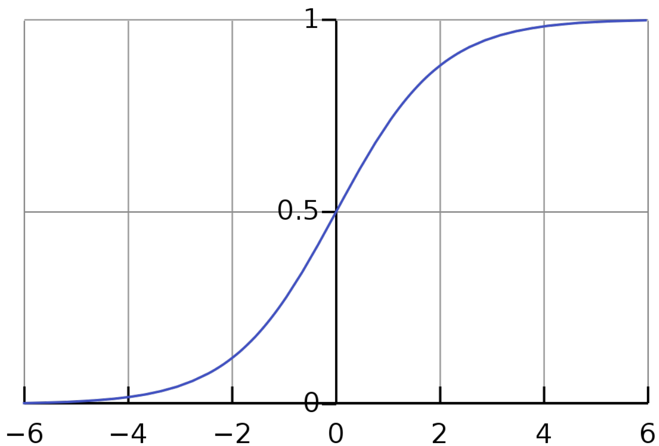
- Entrée : $x_i = (x_i^1, \ldots, x_i^p) \in \mathbb{R}^p$
- Sortie : $y_i \in \{-1,1\}$



On note $\mathbb{P}(Y = 1|X)$ la loi conditionnelle que Y soit égal à 1 sachant X.

On suppose alors que : $$\ln \frac{\mathbb{P}(Y=1|X)}{1-\mathbb{P}(Y=1|X)} = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

$$\mathbb{P}(Y=1|X) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}}$$



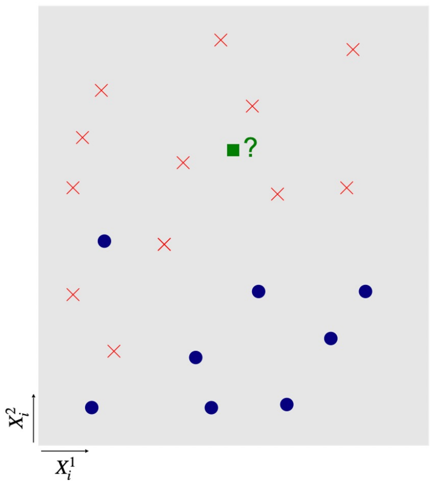
Fonction logistique

Ainsi : $$\beta_0 + \sum_{j=1}^{p} \beta_j X_j \in]-\infty, 0] \iff \mathbb{P}(Y=1\,|\,X) \in [0, 0.5]$$

$$\beta_0 + \sum_{j=1}^{p} \beta_j X_j \in [0, +\infty[\iff \mathbb{P}(Y=1\,|\,X) \in [0.5, 1]$$

n observations d'apprentissage

- Entrée : $x_i = (x_i^1, \dots, x_i^p) \in \mathbb{R}^p$
- Sortie : $y_i \in \{-1,1\}$



On note $\mathbb{P}(Y=1|X)$ la loi conditionnelle que Y soit égal à 1 sachant X.

On suppose alors que : $\ln \frac{\mathbb{P}(Y=1|X)}{1-\mathbb{P}(Y=1|X)} = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$

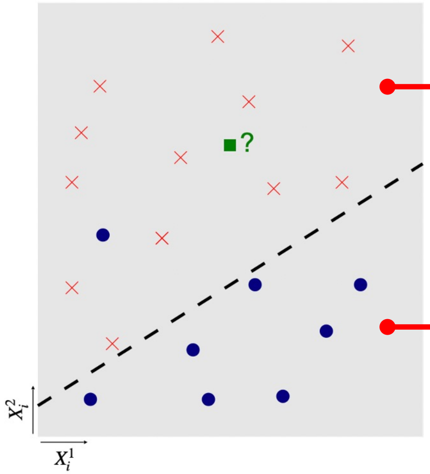
$$\mathbb{P}(Y=1|X) = \frac{e^{\beta_0 + \sum_{j=1}^p \beta_j X_j}}{1 + e^{\beta_0 + \sum_{j=1}^p \beta_j X_j}}$$

pour une observation i, $i = 1, \dots, n$: $p(y_i = 1|x_i^1, \dots, x_i^p) = \frac{e^{\beta_0 + \sum_{j=1}^p \beta_j x_i^j}}{1 + e^{\beta_0 + \sum_{j=1}^p \beta_j x_i^j}}$

Maximisation de la vraisemblance : $L(\beta) = \prod_{i=1}^{n} \left[\left(p(y_i = 1|x_i^1, \dots, x_i^p)\right)^{y_i} . \left(1 - p(y_i = 1|x_i^1, \dots, x_i^p)\right)^{1-y_i} \right]$

n observations d'apprentissage

- Entrée : $x_i = (x_i^1, \dots, x_i^p) \in \mathbb{R}^p$
- Sortie : $y_i \in \{-1,1\}$



$$\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{test}^j > 0$$

$$\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{test}^j < 0$$

On note $\mathbb{P}(Y=1|X)$ la loi conditionnelle que Y soit égal à 1 sachant X.

On suppose alors que : $\ln \frac{\mathbb{P}(Y=1|X)}{1-\mathbb{P}(Y=1|X)} = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$

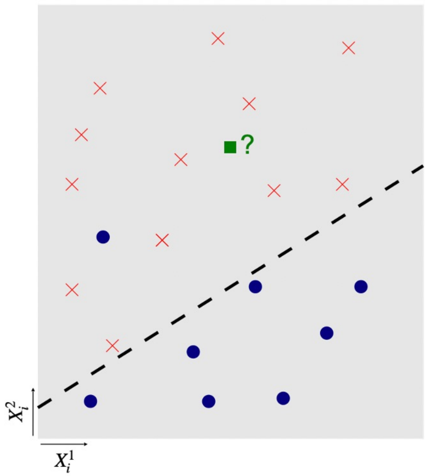
$$\mathbb{P}(Y=1|X) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}}{1+e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}}$$

pour une observation i, $i = 1, \dots, n$: $p(y_i = 1|x_i^1, \dots, x_i^p) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_i^j}}{1+e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_i^j}}$

Maximisation de la vraisemblance : $L(\beta) = \prod_{i=1}^{n} \left[\left(p(y_i = 1|x_i^1, \dots, x_i^p)\right)^{y_i} . \left(1 - p(y_i = 1|x_i^1, \dots, x_i^p)\right)^{1-y_i} \right]$

n observations d'apprentissage

- Entrée : $x_i = (x_i^1, \ldots, x_i^p) \in \mathbb{R}^p$
- Sortie : $y_i \in \{-1,1\}$



On note $\mathbb{P}(Y = 1|X)$ la loi conditionnelle que Y soit égal à 1 sachant X.

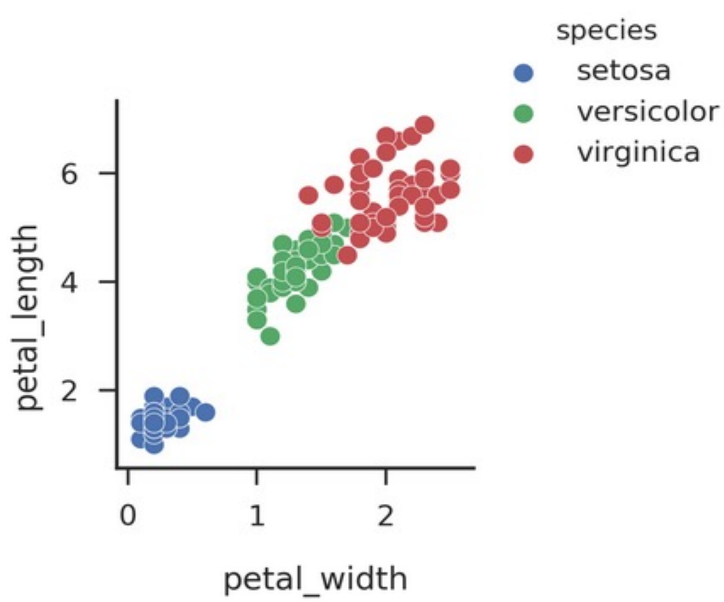
On suppose alors que : $$\ln \frac{\mathbb{P}(Y = 1|X)}{1 - \mathbb{P}(Y = 1|X)} = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

$$\mathbb{P}(Y = 1|X) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}}$$

pour une observation i, $i = 1, \ldots, n$: $$p(y_i = 1|x_i^1, \ldots, x_i^p) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_i^j}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_i^j}}$$

Maximisation de la vraisemblance : $$L(\beta) = \prod_{i=1}^{n} \left[\left(p(y_i = 1|x_i^1, \ldots, x_i^p) \right)^{y_i} . \left(1 - p(y_i = 1|x_i^1, \ldots, x_i^p) \right)^{1-y_i} \right]$$

Dans de nombreuses applications, il existe plus de 2 classes à distinguer, par exemple :



Jeu IRIS : Classification de **3** categories d'IRIS à partir de 4 variables décrivant la forme de l'IRIS.

Jeu MNIST : Classification des **10** chiffres à partir d'images representant des chiffres manuscrits.

Posons les notations :

$$X_i = (x_i^1, x_i^2, \dots, x_i^p)$$
$$y_i \in \{1,2,\dots,K\}$$

avec $i \in \{1,2,\dots,n\}$

On ne pourra plus avoir un classifieur *f* tel que : $f : \mathbb{R}^p \rightarrow \mathbb{R}$

car il n'y généralement pas de relation de rang entre les différentes classes

Posons les notations :

$$X_i = (x_i^1, x_i^2, \ldots, x_i^p)$$
$$y_i \in \{1,2,\ldots,K\}$$

avec $i \in \{1,2,\ldots,n\}$

On ne pourra plus avoir un classifieur *f* tel que : $f : \mathbb{R}^p \rightarrow \mathbb{R}$

car il n'y généralement pas de relation de rang entre les différentes classes

On utilisera une representation de type *one-hot-encoding* :

$$f(X) = \big(\mathbb{P}(y = 1 \,|\, X), \mathbb{P}(y = 2 \,|\, X), \ldots, \mathbb{P}(y = K \,|\, X)\big)$$

On utilise le modèle :

$$p(y_i = k \,|\, X_i) = \frac{e^{\beta_0^k + \sum_{j=1}^{p} \beta_j^k x_i^j}}{Z_i} = \frac{e^{\beta_0^k + \sum_{j=1}^{p} \beta_j^k x_i^j}}{\sum_{\tilde{k}=1}^{K} e^{\beta_0^{\tilde{k}} + \sum_{j=1}^{p} \beta_j^{\tilde{k}} x_i^j}}$$

où Z_i, la function de partition, garantit que l'on a des distributions de probabilités pour chaque observation *i*. Notons que cette stratégie est aussi très populaire pour la classification à classes multiples avec des réseaux de neurones. On parle de function softmax.

On maximise ainsi :

$$\hat{\beta} = \arg\max_{\beta} L(\beta)$$ avec $$\beta = (\beta_1^1, \beta_1^2, \ldots, \beta_1^K, \beta_2^1, \ldots, \beta_p^K)$$

et $$L(\beta) = \prod_{k=1}^{K} \prod_{i=1}^{n} p(y_i = k \,|\, x_i^1, x_i^2, \ldots, x_i^p) \mathbb{1}_{y_i = k}$$

Il est plus avantageux de maximiser la log-vraisemblance :

- Maximum identique à celui de la vraisemblance
- Pas de produits multiples qui conduisent la vraisemblance calculée sous le zero numérique

En pratique, on maximise ainsi :

$$
\begin{aligned}
\hat{\beta} &= \arg\max_{\beta} \log\left(L(\beta)\right) \\
&= \arg\max_{\beta} \log\left(\prod_{k=1}^{K}\prod_{i=1}^{n} p(y_i = k \,|\, x_i^1, x_i^2, \ldots, x_i^p)\mathbb{1}_{y_i=k}\right) \\
&= \arg\max_{\beta} \sum_{k=1}^{K}\sum_{i=1}^{n} \mathbb{1}_{y_i=k} \log\left(p(y_i = k \,|\, x_i^1, x_i^2, \ldots, x_i^p)\right) \\
&= \arg\max_{\beta} \sum_{k=1}^{K}\sum_{i=1}^{n} \mathbb{1}_{y_i=k} \log\left(\frac{e^{\beta_0^k + \sum_{j=1}^{p} \beta_j^k x_i^j}}{\sum_{\tilde{k}=1}^{K} e^{\beta_0^{\tilde{k}} + \sum_{j=1}^{p} \beta_j^{\tilde{k}} x_i^j}}\right) \\
&\boxed{= \arg\max_{\beta} \sum_{i=1}^{n} \log\left(\frac{e^{\beta_0^{y_i} + \sum_{j=1}^{p} \beta_j^{y_i} x_i^j}}{\sum_{\tilde{k}=1}^{K} e^{\beta_0^{\tilde{k}} + \sum_{j=1}^{p} \beta_j^{\tilde{k}} x_i^j}}\right)}
\end{aligned}
$$