

# SYSTÈME DE CLASSIFICATION BAYESIEN NAÏF

Vincent Guigue, Romain Thoreau  
[vincent.guigue@agroparistech.fr](mailto:vincent.guigue@agroparistech.fr)



# Exemple introductif : classification automatique de chiffres manuscrits

Un programme informatique peut-il reconnaître des chiffres ?

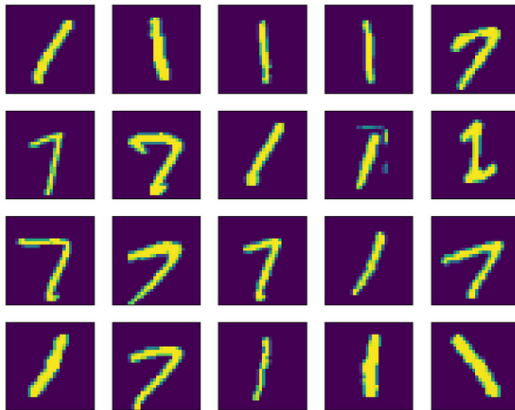


Figure 1: Exemples de chiffres manuscrits du jeu de données MNIST



# Plan du cours

1 LOIS DE PROBABILITÉS

2 NAIVE BAYES

# LOIS DE PROBABILITÉS



# Loi de Bernoulli

## Définition

**Épreuve de Bernoulli** = expérience aléatoire qui ne peut prendre que deux résultats (*succès* et *échec*)

$p$  = proba de succès, et  $q = 1 - p$  = proba d'échec.



# Loi de Bernoulli

## Définition

**Épreuve de Bernoulli** = expérience aléatoire qui ne peut prendre que deux résultats (*succès* et *échec*)

$p$  = proba de succès, et  $q = 1 - p$  = proba d'échec.

## Loi de Bernoulli

Variable  $X$  à support  $\mathcal{X} = \{0, 1\}$  telle que:

$$P(X = 1) = p \text{ et } P(X = 0) = 1 - p$$

$$E(X) = p \quad V(X) = p(1 - p)$$

$\implies X$  = le nombre de succès de l'épreuve de Bernoulli



# Loi binomiale

## Définition

**Épreuve binomiale** = expérience aléatoire telle que:

- 1 on répète  $n$  fois la même épreuve de Bernoulli,
- 2 les probas  $p$  et  $q$  restent inchangées pour chaque épreuve de Bernoulli,
- 3 les épreuves de Bernoulli sont toutes réalisées indépendamment les unes des autres.



# Loi binomiale

## Définition

**Épreuve binomiale** = expérience aléatoire telle que:

- 1 on répète  $n$  fois la même épreuve de Bernoulli,
- 2 les probas  $p$  et  $q$  restent inchangées pour chaque épreuve de Bernoulli,
- 3 les épreuves de Bernoulli sont toutes réalisées indépendamment les unes des autres.

## Loi binomiale de paramètres $n$ et $p$

- $X$  = nombre de succès de l'épreuve binomiale
- $X \sim \mathcal{B}(n, p)$
- $P(X = k) = C_n^k p^k (1 - p)^{n-k}, \forall k = 0, \dots, n$
- $E(X) = np \quad V(X) = np(1 - p)$



# Loi normale



Loi extrêmement importante : souvent une très bonne approximation de la loi réelle

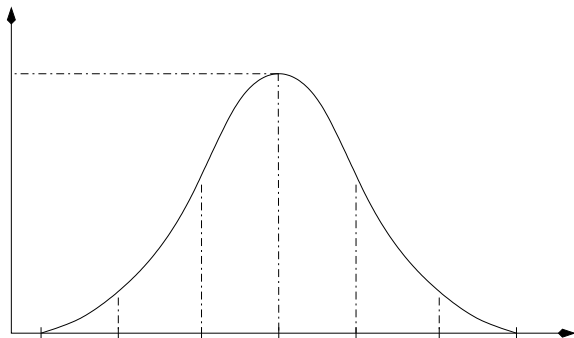
*Définition : loi normale de paramètres  $\mu$  et  $\sigma^2$*

- notée  $\mathcal{N}(\mu, \sigma^2)$
- s'applique pour des variables aléatoires continues
- densité positive sur tout  $\mathbb{R}$  :

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \exp \left\{ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right\}$$

- $E(X) = \mu \quad V(X) = \sigma^2$

# Fonction de densité de la loi normale



Quelques reflexes:

- 2/3 de la masse entre  $+\sigma$  et  $-\sigma$
- Support infini...

Mais empiriquement  $\sim$  toutes les observations entre  $+3\sigma$  et  $-3\sigma$

- Facile à dériver, à tronquer, ...



# Loi normale en pratique

## *Théorème*

$$X \sim \mathcal{N}(\mu; \sigma^2)$$

Alors la variable  $Y = aX + b$  obéit à la loi  $\mathcal{N}(a\mu + b; a^2\sigma^2)$ .

⇒ toute transformée affine d'une variable aléatoire suivant  
une loi normale suit aussi une loi normale



# Loi normale en pratique

## Théorème

$$X \sim \mathcal{N}(\mu; \sigma^2)$$

Alors la variable  $Y = aX + b$  obéit à la loi  $\mathcal{N}(a\mu + b; a^2\sigma^2)$ .

$\implies$  toute transformée affine d'une variable aléatoire suivant une loi normale suit aussi une loi normale

## Corollaire

- $X$  une variable aléatoire obéissant à une loi  $\mathcal{N}(\mu; \sigma^2)$   
 $\implies Z = \frac{X - \mu}{\sigma}$  suit la loi  $\mathcal{N}(0; 1)$
- $Z$  suit une loi normale centrée (à cause de la moyenne en 0) réduite (à cause du  $\sigma^2$  égal à 1)



# Loi normale en pratique (2)

## *Théorème*

$$X_1 \sim \mathcal{N}(\mu_1; \sigma_1^2), \quad X_2 \sim \mathcal{N}(\mu_2; \sigma_2^2)$$

Si les variables sont indépendantes, alors la variable  $Y = X_1 + X_2$  obéit à la loi  $\mathcal{N}(\mu_1 + \mu_2; \sigma_1^2 + \sigma_2^2)$ .

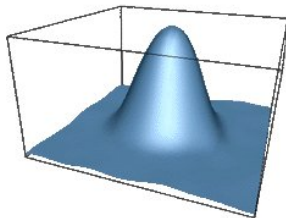
# Loi normale bi-dimensionnelle

## Définition : loi normale bi-dimensionnelle

- couple de variables  $(X, Y)$
- densité dans  $\mathbb{R}^2$  :

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times \\ \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[ \left( \frac{x-\mu_x}{\sigma_x} \right)^2 - 2\rho \frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \left( \frac{y-\mu_y}{\sigma_y} \right)^2 \right] \right\}$$

où  $\rho = \frac{\text{cov}(X, Y)}{\sigma_x\sigma_y} =$  **coefficient de corrélation linéaire**





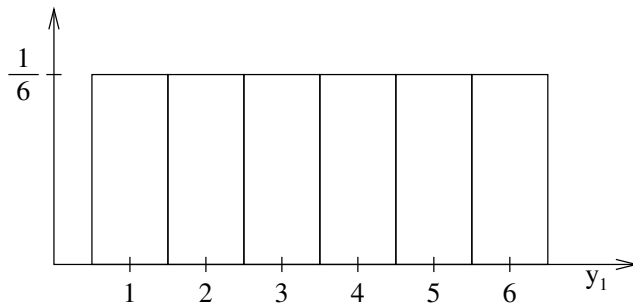
# Loi normale = limite d'autres lois (1/4)

Lancés de dés à 6 faces



⇒ on compte la somme des résultats des dés

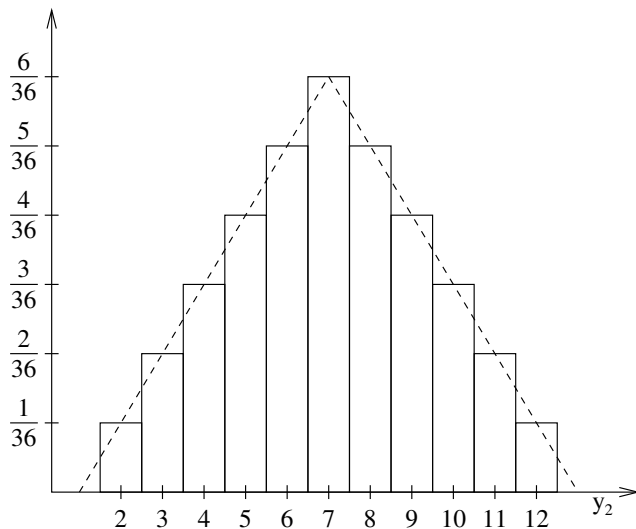
Somme pour 1 jet de dé





# Loi normale = limite d'autres lois (2/4)

Somme pour 2 jets de dés

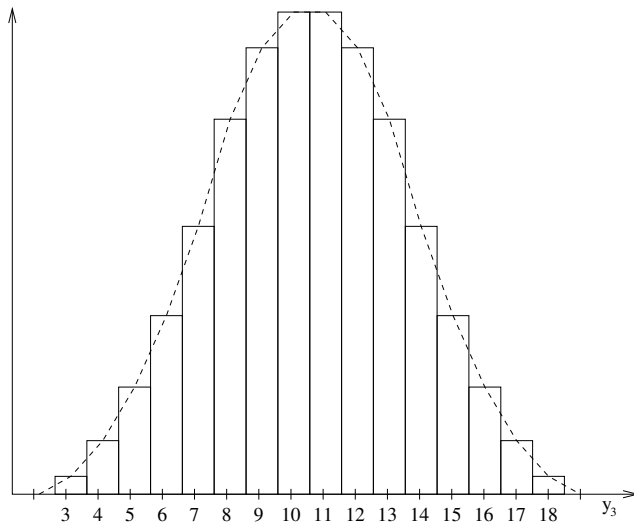






# Loi normale = limite d'autres lois (3/4)

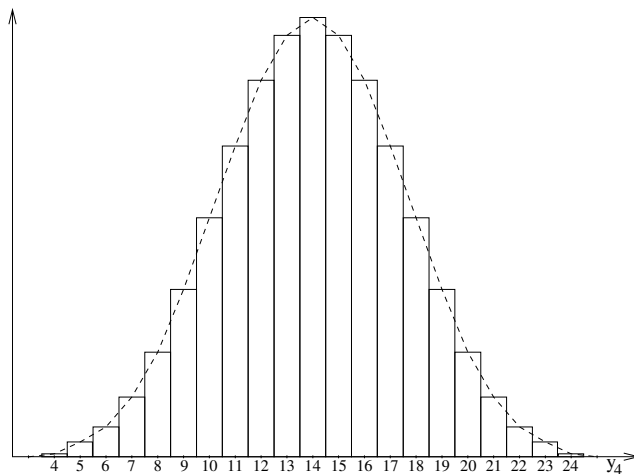
## Somme pour 3 jets de dés



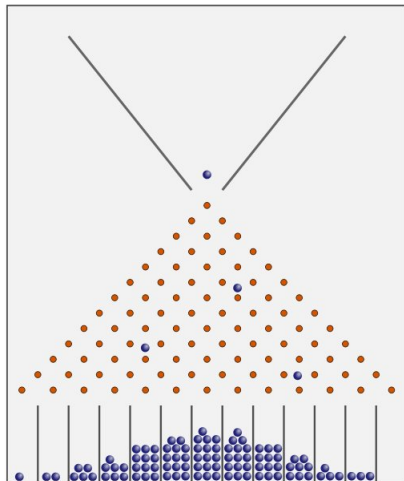


# Loi normale = limite d'autres lois (4/4)

Somme pour 4 jets de dés



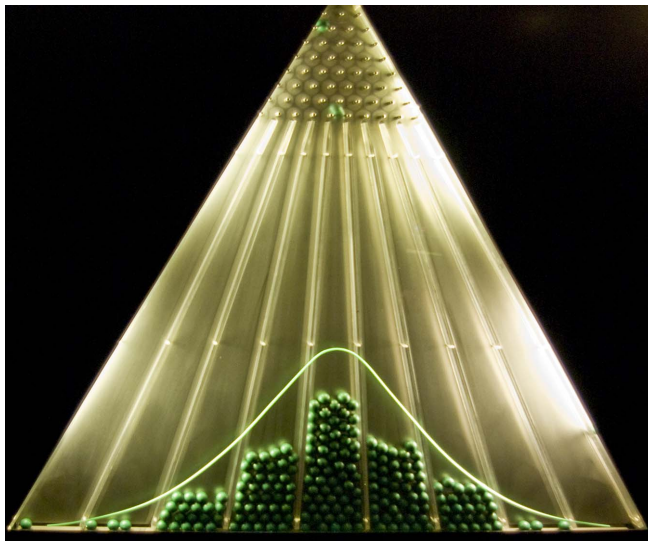
# La planche de Galton



- chaque niveau  $\Rightarrow$  expérience de Bernoulli
- $\Rightarrow X \sim$  loi binomiale



# La planche de Galton





# Théorème central-limite

## *Théorème central-limite*

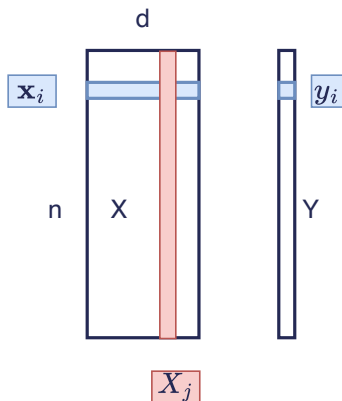
- $(X_n)_{n \in \mathbb{N}}$  : suite de variables
  - de même loi
  - d'espérance  $\mu$
  - de variance  $\sigma^2$
  - **mutuellement** indépendantes
- alors la suite des moyennes empiriques centrées réduites

$\frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}}$  tend en loi vers la loi normale centrée réduite :

$$\frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} \xrightarrow{\text{loi}} \mathcal{N}(0, 1)$$

# NAIVE BAYES

## Notations et représentation des données



$X$  matrice des données

- composée de  $n$  individus  $\mathbf{x}_i \in \mathcal{X}$
- presque toujours,  $\mathcal{X} = \mathbb{R}^d$

$Y$  étiquettes des données,  $y_i \in \mathcal{Y}$

- $y_i \in \mathbb{R}$   $\Rightarrow$  régression
- $y_i \in \{1, \dots, C\}$   $\Rightarrow$  classification en  $C$  catégories

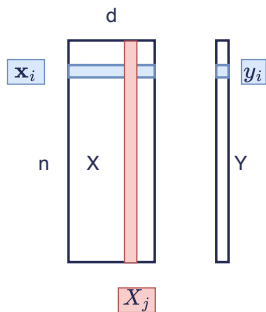
# Apprentissage automatique

A partir des données, construire une fonction  $f$  telle que:

$$\forall (\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}, \quad f(\mathbf{x}) \approx y$$



# Algorithme bayésien naïf



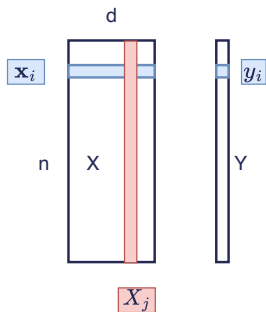
Hypothèse d'indépendance des variables descriptives  $X_j$

Pourquoi c'est très naïf?



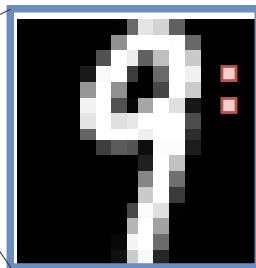
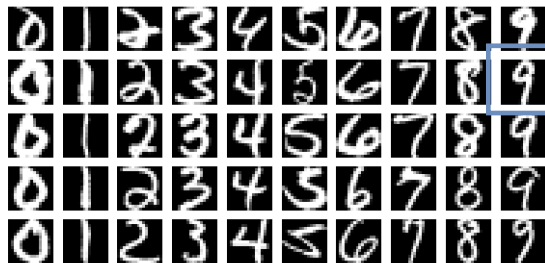


# Algorithme bayésien naïf



Hypothèse d'indépendance des variables descriptives  $X_j$

Pourquoi c'est très naïf?



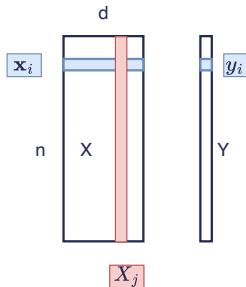
$$x_{ij} \sim X_j$$

$$x_{ik} \sim X_k$$

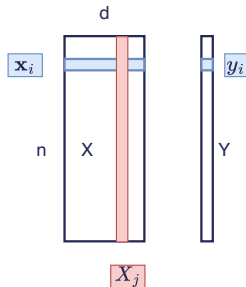


# Hypothèse variable par variable

- 1 Choix loi de probabilité pour (une ou toutes les)  $X_j$   
e.g Bernoulli pour une image binaire:  $X_j \sim \text{Ber}(p_j)$



# Hypothèse variable par variable



- 1 Choix loi de probabilité pour (une ou toutes les)  $X_j$   
e.g Bernoulli pour une image binaire:  $X_j \sim \text{Ber}(p_j)$

$$P(X_j = 1) = p_j \quad P(X_j = 0) = 1 - p_j$$

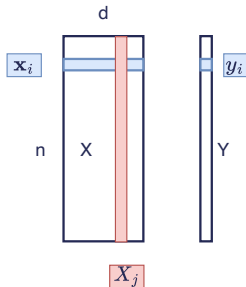
Vraisemblance de l'observation  $x_{ij}$ :

$$P(X_j = x_{ij}) = p_j^{x_{ij}} (1 - p_j)^{(1-x_{ij})}$$

- 2 Une variable descriptive  $X_j \Rightarrow 1$  paramètre  $p_j$   
On regroupe les paramètres :  $\Theta = \{p_1, \dots, p_d\}$
- 3 Optimisation des paramètres par max de vraisemblance



# Hypothèse variable par variable



- 1 Choix loi de probabilité pour (une ou toutes les)  $X_j$   
e.g Bernoulli pour une image binaire:  $X_j \sim \text{Ber}(p_j)$

$$P(X_j = 1) = p_j \quad P(X_j = 0) = 1 - p_j$$

Vraisemblance de l'observation  $x_{ij}$ :

$$P(X_j = x_{ij}) = p_j^{x_{ij}} (1 - p_j)^{(1-x_{ij})}$$

- 2 Une variable descriptive  $X_j \Rightarrow 1$  paramètre  $p_j$   
On regroupe les paramètres :  $\Theta = \{p_1, \dots, p_d\}$
- 3 Optimisation des paramètres par max de vraisemblance

$$\text{Echantillon i.i.d} + \text{NB} \Rightarrow \mathcal{L}(X) = \prod_{i=1}^n \prod_{j=1}^d P(x_{ij} | \Theta)$$

$$\text{Optimisation: } p_j^* = \arg \max_{p_j} \mathcal{L}(X)$$

## Apprentissage statistique

Identification des paramètres optimaux correspondant aux observations



# Calcul de la vraisemblance

- Pour une **valeur descriptive** , sous l'hypothèse de Bernoulli:

$$P(X_j = x_{ij}) = P(X_j = x_{ij} | p_j) = p_j^{x_{ij}} (1 - p_j)^{(1-x_{ij})}$$

- Pour un **individu**, avec **indépendance des variables descriptives**:

$$P(\mathbf{x}_i) = P(\mathbf{x}_i | \Theta) = \prod_{j=1}^d P(X_j = x_{ij})$$

- Pour l'**échantillon** entier, **sous hypothèse i.i.d**:

$$\mathcal{L}(X) = \prod_{i=1}^n \prod_{j=1}^d P(x_{ij} | \Theta) = \prod_{i=1}^n \prod_{j=1}^d p_j^{x_{ij}} (1 - p_j)^{(1-x_{ij})}$$



# Vraisemblance vs log-Vraisemblance

$$\mathcal{L}(X) \Rightarrow \log \mathcal{L}(X)$$

La vraisemblance a en générale vocation à être dérivée pour trouver les paramètres optimaux... Comme le log est une fonction croissante:

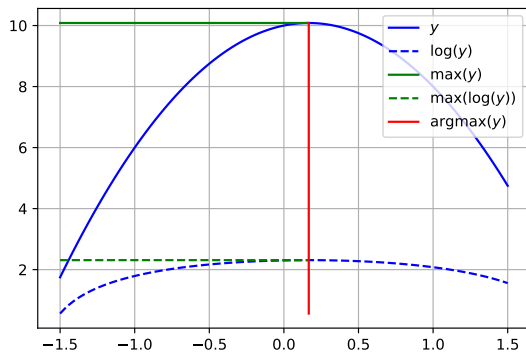
$$\arg \max_{\Theta} \mathcal{L}(X) = \arg \max_{\Theta} \log \mathcal{L}(X)$$

⇒ On travaille donc sur la log-vraisemblance, bien plus facile à dériver

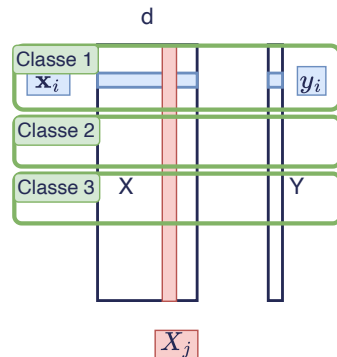
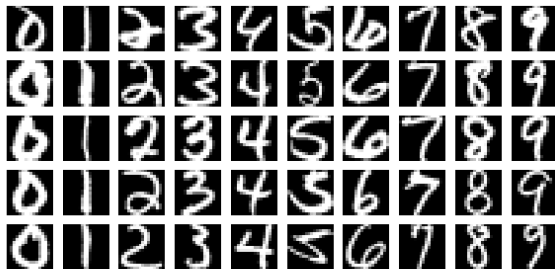
$$\mathcal{L}(X) = \prod_{i=1}^n \prod_{j=1}^d p_j^{x_{ij}} (1 - p_j)^{(1-x_{ij})}$$

$$\log \mathcal{L}(X) =$$

$$\sum_{i=1}^n \sum_{j=1}^d x_{ij} \log(p_j) + (1 - x_{ij}) \log(1 - p_j)$$

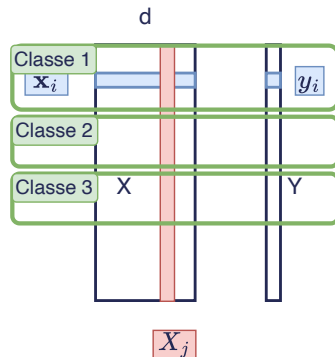
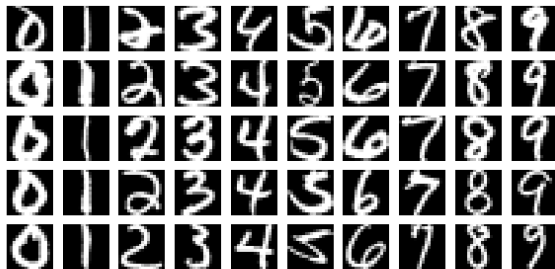


# Cas de la classification bayésienne naïve



- 1 classe  $C = 1$  sous-ensemble de données = 1 modèle optimisé (= un ensemble de paramètre  $\Theta_c$ )
- $C (\times d)$  problèmes d'optimisation distincts
- Combien de paramètres avec une modélisation de Bernoulli sur  $d = 256$  pixels?

# Cas de la classification bayésienne naïve



- 1 classe  $C = 1$  sous-ensemble de données = 1 modèle optimisé (= un ensemble de paramètre  $\Theta_c$ )
- $C (\times d)$  problèmes d'optimisation distincts
- Combien de paramètres avec une modélisation de Bernoulli sur  $d = 256$  pixels?
- $\Theta_c = \{p_{c,1}^*, \dots, p_{c,d}^*\}$  et  $\Theta = \{\Theta_1, \dots, \Theta_c, \dots, \Theta_C\} \Rightarrow 2560$  paramètres





# Apprentissage du modèle

Comment résoudre :

$$p_j^* = \arg \max_{p_j} \mathcal{L}(X) = \arg \max_{p_j} \sum_{i=1}^n \sum_{j=1}^d x_{ij} \log(p_j) + (1 - x_{ij}) \log(1 - p_j) ?$$

**Solution 1**

**Solution 2**

$$\frac{\partial \mathcal{L}_j(X)}{\partial p_j} = 0 \Leftrightarrow \dots$$

$$p_j^* = \dots$$



# Apprentissage du modèle

Comment résoudre :

$$p_j^* = \arg \max_{p_j} \mathcal{L}(X) = \arg \max_{p_j} \sum_{i=1}^n \sum_{j=1}^d x_{ij} \log(p_j) + (1 - x_{ij}) \log(1 - p_j) ?$$

## Solution 1

$$\frac{\partial \mathcal{L}_j(X)}{\partial p_j} = 0 \Leftrightarrow \dots$$

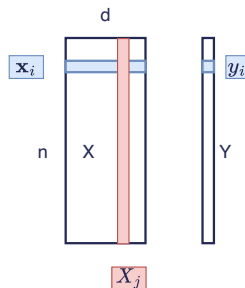
$$p_j^* = \dots$$

## Solution 2

Je connais la loi de Bernoulli

(ou j'ai accès à wikipedia)

$$p_j^* = \frac{\sum_i x_{ij}}{n}$$

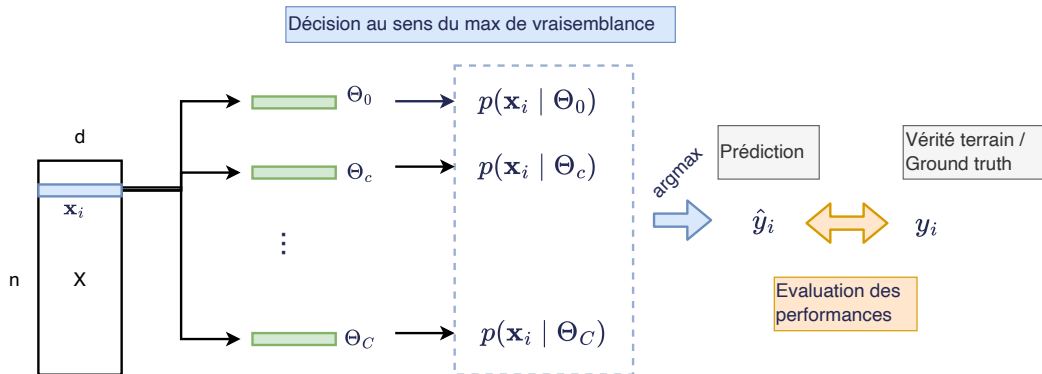


# Inférence

On utilise le modèle optimisé pour prédire la classe :

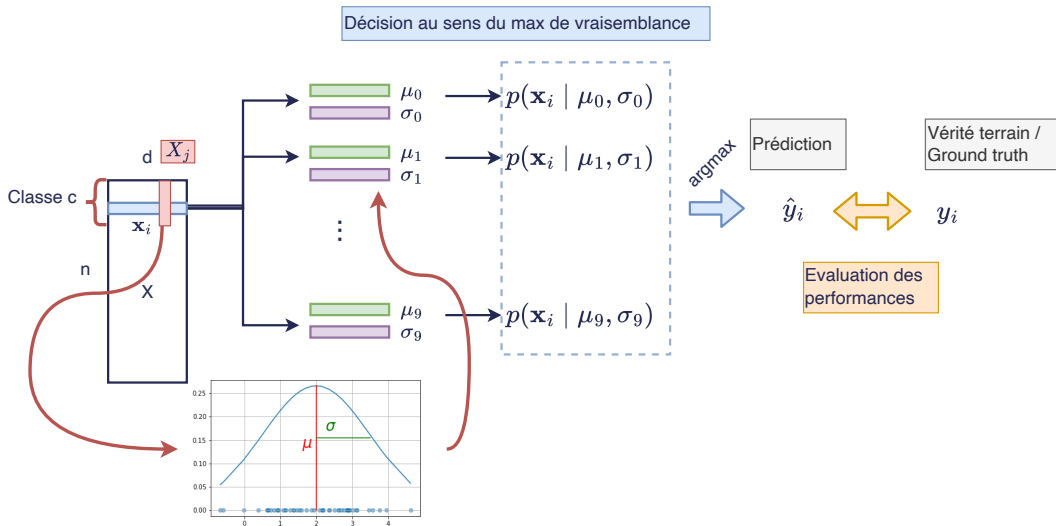
$$\hat{y}_i = \arg \max_k p(y_i = k | \mathbf{x}_i, \Theta).$$

De manière équivalente, est-ce que la donnée  $\mathbf{x}_i$  est plus vraisemblable sous le modèle de la classe 0, 1, ... ou  $C$  ?





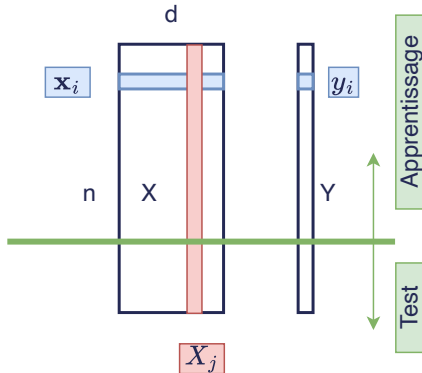
# Passage à la gaussienne



# Evaluation du modèle / Sélection de modèle

!! L'évaluation est aussi importante que l'apprentissage !!

- Evaluer sur les données d'apprentissage (=qui ont servi à régler les paramètres)  
⇒ **Tricherie, surestimation des performances**
- Evaluer sur des données vierges = OK



Problème de la répartition entre apprentissage et test

- La validation croisée

# Evaluation du modèle / Sélection de modèle

!! L'évaluation est aussi importante que l'apprentissage !!

- Evaluer sur les données d'apprentissage (=qui ont servi à régler les paramètres)  
⇒ **Tricherie, surestimation des performances**
- Evaluer sur des données vierges = OK
- La validation croisée

