

# Crowd-Sourcing

Pralon, Thiriet

Décembre 2022



# Présentation générale

## Crowd-Sourcing

- *Problématique* : Attribution de labels à une image/patient
- *Approche intuitive* : vote majoritaire
- *Approche de cette présentation* : algorithme EM

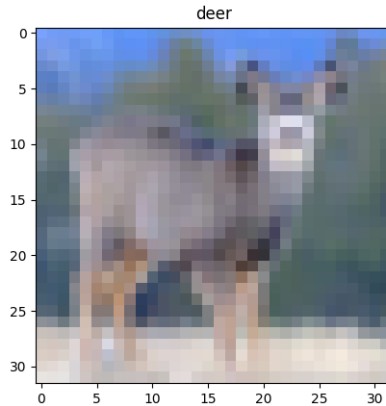
# Plan

- 1 Présentation générale
- 2 Modélisation du problème
- 3 Résolution du problème



[https://github.com/Aurelien2021/Projet\\_Crowd\\_Sourcing](https://github.com/Aurelien2021/Projet_Crowd_Sourcing)

# Enjeu



0	airplane
1	automobile
2	bird
3	cat
4	deer
5	dog
6	frog
7	horse
8	ship
9	truck

Table 1:  $J = 10$  labels

$K$  annotateurs  
/  
images  
 $J$  labels

# Modélisation du problème

- $(X_i)_{i \in \{1, \dots, I\}}$ , tel que  $\forall i$ ,  $X_i$  à valeur dans  $\{1, \dots, J\}$  correspond à l'objet illustré sur l'image  $i$
- $Y_i^k$  à valeur dans  $\{0, 1\}^J$ , la labélisation du participant  $k$  à l'image  $i$



Nous supposons que,  $\forall j, k \in \llbracket 1, J \rrbracket \times \llbracket 1, K \rrbracket$  :

$$Y_i^k | X_i = \alpha_i \sim \text{Multinomiale}(n, (\pi_{\alpha_i, j}^k)_{j \in \llbracket 1, J \rrbracket})$$

où  $(\pi_{\alpha_i, j}^k)_{j \in \llbracket 1, J \rrbracket}$  proba que  $k$  attribue  $j$  à l'image  $i$  sachant que le label correct est  $\alpha_i$ .

$$\mathbb{P}_{Y_i^k | X_i = \alpha_i} = \sum_{(n_{i,1}^k, \dots, n_{i,J}^k)} \underbrace{\frac{(\sum_{j=1}^J n_{i,j}^k)!}{\prod_{j=1}^J (n_{i,j}^k)!} \prod_{j=1}^J (\pi_{\alpha_i,j}^k)^{n_{i,j}^k}}_{\text{densité suivant}} \delta_{(n_{i,1}^k, \dots, n_{i,J}^k)}$$

$$\mathbb{P} \left( \bigcap_{k=1}^K Y_i^k | X_i = \alpha_i \right) = \prod_{k=1}^K \mathbb{P} \left( Y_i^k | X_i = \alpha_i \right), \text{ car } (Y_i^k)_k \text{ indep}$$
$$\propto \prod_{k=1}^K \prod_{j=1}^J \left( \pi_{\alpha_i, j}^k \right)^{n_{i, j}^k}$$

$$\begin{aligned}\mathbb{P}\left(\left(\bigcap_k Y_i^k\right) \cap (X_i = \alpha_i)\right) &= \mathbb{P}\left(\bigcap_{k=1}^K Y_i^k | X_i = \alpha_i\right) \times \mathbb{P}(X_i = \alpha_i) \\ &\propto p_{\alpha_i} \times \prod_{k=1}^K \prod_{j=1}^J \left(\pi_{\alpha_i,j}^k\right)^{n_{i,j}^k}\end{aligned}$$

Plus généralement, en définissant  $T_{i,l}$  la variable aléatoire telle que  $T_{i,l} = 1$ , si le label de l'image  $i$  est  $l$  et  $T_{i,l} = 0$  sinon, on a :

$$\mathbb{P} \left( \left( \bigcap_k Y_i^k \right) \cap (X_i = \alpha_i) \right) \propto \prod_{l=1}^J \left( p_l \prod_{k=1}^K \prod_{j=1}^J \left( \pi_{l,j}^k \right)^{n_{i,j}^k} \right)^{T_{i,l}}$$

$$\begin{aligned}
 \mathbb{P} \left( \left( \bigcap_i \bigcap_k Y_i^k \right) \cap \left( \bigcap_i X_i = \alpha_i \right) \right) &= \mathbb{P} \left( \bigcap_i \left[ \left( \bigcap_k Y_i^k \right) \cap (X_i = \alpha_i) \right] \right) \\
 &= \prod_i \mathbb{P} \left( \left( \bigcap_k Y_i^k \right) \cap (X_i = \alpha_i) \right) \\
 &\propto \prod_{i=1}^I \prod_{l=1}^J \left( \rho_l \prod_{k=1}^K \prod_{j=1}^J (\pi_{l,j}^k)^{n_{i,j}^k} \right)^{T_{i,l}}
 \end{aligned}$$

# Algorithme de Dawid et Skene (Algo EM)

Soit l'image  $i$ , on décide que le label de cette image est  $l$  si :

$$l = \operatorname{argmax}_l(T_{i,l})$$

$$\begin{aligned}\mathbb{E}[T_{i,l}|(Y_i^k)_k] &= \mathbb{E}[0 \times \mathbb{1}_{\{T_{i,l}=0\}} + 1 \times \mathbb{1}_{\{T_{i,l}=1\}}|(Y_i^k)_k] \\ &= \mathbb{E}[\mathbb{1}_{\{T_{i,l}=1\}}|(Y_i^k)_k] \\ &= \mathbb{P}(T_{i,l} = 1|(Y_i^k)_k)\end{aligned}$$



$$\begin{aligned}
 \mathbb{P}\left(T_{i,l} = 1 | (Y_i^k)_k\right) &= \frac{\mathbb{P}\left((T_{i,l} = 1) \cap (Y_i^k)_k\right)}{\mathbb{P}\left((Y_i^k)_k\right)} \\
 &= \frac{\mathbb{P}\left((Y_i^k)_k | (T_{i,l} = 1)\right) \times \mathbb{P}(T_{i,l} = 1)}{\sum_l^J \mathbb{P}\left((Y_i^k)_k | (X_i = l)\right) \times \mathbb{P}(X_i = l)} \\
 &= \frac{\mathbb{P}\left((Y_i^k)_k | (X_i = l)\right) \times \mathbb{P}(X_i = l)}{\sum_l^J \mathbb{P}\left((Y_i^k)_k | (X_i = l)\right) \times \mathbb{P}(X_i = l)}
 \end{aligned}$$

$$\mathbb{P}\left(T_{i,l} = 1 | (Y_i^k)_k\right) = \frac{\prod_k^K \prod_j^J (\pi_{l,j}^k)^{n_{i,j}^k}}{\sum_l p_l \prod_k^K \prod_j^J (\pi_{l,j}^k)^{n_{i,j}^k}}$$

Or  $(\pi_{l,j}^k)_{k,l,j}$  et  $p_l$  inconnus.  
 $\implies$  Algo EM

En effectuant la maximisation du Lagrangien de la fonction de vraisemblance:

$$\left( \left( \pi_{l,j}^k \right)_{l,j,k}, (p_j)_j \right) \mapsto \prod_i^I \prod_{l=1}^J \left( p_l \prod_{k=1}^K \prod_{j=1}^J \left( \pi_{l,j}^k \right)^{n_{i,j}^k} \right)^{T_{i,l}}$$

Sous la contrainte  $\forall l, k \in \llbracket 1, J \rrbracket \times \llbracket 1, K \rrbracket, \left( \sum_{j=1}^J \pi_{l,j}^k = 1 \right)$

On obtient les estimateurs du maximum de vraisemblance :

$$\widehat{\pi}_{l,j}^k = \frac{\sum_i T_{i,l} n_{i,j}^k}{\sum_j \sum_i T_{i,l} n_{i,j}^k}$$

$$\widehat{p}_l = \frac{\sum_i T_{i,l}}{I}$$

# Algo EM

La phase *expectation* consiste à calculer l'estimation des  $T_{i,l}$  par  $\mathbb{E}[T_{i,l} | (Y_i^k)_k]$  dont nous avons une forme analytique.

La phase *maximization* consiste à maximiser la vraisemblance du modèle. Ce maximum est atteint en les paramètres  $(\widehat{\pi_{l,j}^k}, \widehat{p_l})$  établis précédemment.

Initialisation des  $T_{i,l}$  :

$$T_{i,l} = \frac{\sum_k n_{i,l}^k}{\sum_k \sum_j n_{i,j}^k}$$

Initialisation des  $(T_{i,l})_{i,l} = (T_{i,l})_0$ , les observations  $(Y_i^k)_{k,i}$ ,  $(n_{i,l}^k)_{k,i,l}$  et  $N \in \mathbb{N}$  le nombre d'itération;  $n$  allant de 1 à  $N$

**ETAPE M :** Calculer les 
$$\left(\widehat{\pi}_{l,j}^k\right)_n = \frac{\sum_i (T_{i,l})_{n-1} n_{i,j}^k}{\sum_j \sum_i (T_{i,l})_{n-1} n_{i,j}^k},$$

$$(\widehat{\rho}_l)_n = \frac{\sum_i (T_{i,l})_{n-1}}{I}$$

**ETAPE E :** Calculer 
$$\left(\widehat{T}_{i,l}\right)_n = \frac{(\widehat{\rho}_l)_n \prod_k \prod_j \left(\widehat{\pi}_{l,j}^k\right)_n^{n_{i,j}^k}}{\sum_l (\widehat{\rho}_l)_n \prod_k \prod_j \left(\widehat{\pi}_{l,j}^k\right)_n^{n_{i,j}^k}};$$

**RETURN**  $\max_l \left(\widehat{T}_{i,l}\right)_N$