



Projet Master 2 SSD

Crowd-Sourcing

Rédigé par

PRALON Nicolas

THIRIET Aurelien

Encadrant : Joseph SALMON

IMAG
INSTITUT MONTPELLIERAIN
ALEXANDER GROTHENDIECK

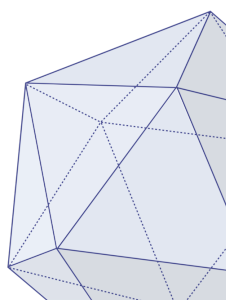


Table des matières

Introduction	2
1 Modélisation du problème	3
1.1 Modèle Dawid-Skene	3
1.2 Résolution du problème	5
1.3 Initialisation des $T_{i,l}$	7
2 Pseudo-Code algorithme EM	8
2.1 L'étape E (Expectation)	8
2.2 L'étape M (Maximization)	8
2.3 Pseudo-code	8

Introduction

En science appliquée et notamment en statistique inférentielle, le recueil de données constitue une étape primordiale, il nous permet d'élaborer des modèles et de construire des estimateurs. Il convient à ce titre d'être attentif à l'observation des données.

Dans un cadre idéal, les modèles construits nécessitent un faible nombre de données dont l'observation peut être réalisée par des professionnels du domaine concerné. Toujours est-il que ces situations sont peu courantes et l'observation, par ces experts, d'un grand nombre de données n'est pas envisageable. C'est à ce titre que le "Crowd-Sourcing" est couramment utilisé, puisqu'un grand nombre d'intervenants permet de construire la base de données nécessaire. Toutefois des erreurs d'observations quant à la nature des données peuvent être commises, il est alors essentiel de s'accorder sur une décision.

C'est à ce problème, que ce projet propose de présenter différentes solutions. Nous étudierons le modèle construit par DAWID et SKENE.

Chapitre 1

Modélisation du problème

1.1 Modèle Dawid-Skene

Dans ce chapitre, nous établissons le cadre nous permettant de traiter le problème de décision lorsque plusieurs observations d'une même donnée sont fournies.

Nous définissons alors l'exemple d'un ensemble d'images à labeliser, les experts du domaine décident alors de faire appel à plusieurs participants pour labéliser ces images.

On définit le cadre suivant :

- ✱ On dispose d'un ensemble d'images $\{1, \dots, I\}$ et on suppose le vecteur aléatoire $(X_i)_{i \in \{1, \dots, I\}}$, tel que $\forall i$, X_i à valeur dans $\{1, \dots, J\}$ correspond à l'objet illustré sur l'image i .
- ✱ On dispose également d'un ensemble de participant $\{1, \dots, K\}$, et on considère $\forall i \in \{1, \dots, I\}$ le vecteur aléatoire $(Y_i^k)_{k \in \{1, \dots, K\}}$ la labélisation du participant k à l'image i , Y_i^k à valeur dans $\{0, 1\}^J$.
- ✱ Chaque participant répond indépendamment des autres et chaque image est indépendante des autres, elle ne donne aucune indication sur les autres images. On note l'indépendance 2 à 2, $(Y_i^k)_k \perp\!\!\!\perp$ et $(X_i)_i \perp\!\!\!\perp$.

A partir des résultats donnés par chaque participant nous pouvons essayer d'attribuer un label à chaque image. Toutefois nous pouvons nous retrouver dans la situation où le label donné par un participant à une image est différent de celui d'un autre participant, pour diverses raisons ; erreur d'annotation, fatigue, spam. Il convient dans ce cas de trouver une solution à la question : quel label donner à chaque image ?

Afin de simplifier les démarches, nous allons envisager le cadre suivant ; nous allons considérer, en plus de l'observation des résultats des participants, observer les labels des images.

Ce cadre s'écarte légèrement de celui du problème puisque l'information sur les labels des images n'est pas accessible, cependant nous verrons comment palier à cela dans la partie suivante ; notamment par une garantie théorique d'un algorithme que l'on appelle algorithme EM.

Dans un premier temps nous considérons le cas d'un participant et une image.

Nous émettons l'hypothèse suivante :

Hypothèse 1. Nous supposons que, $\forall j, k \in \llbracket 1, J \rrbracket \times \llbracket 1, K \rrbracket$, $Y_i^k | X_i = \alpha_i \sim \text{Multinomiale}((\pi_{\alpha_i, j}^k)_{j \in \llbracket 1, J \rrbracket}, 1)$

Sachant que l'image i a pour label α_i , la réponse du participant k suit une loi multinomiale. Dans le cas où le label de l'image i est connu, le participant k peut aussi y avoir accès, et nous avons Y_i^k, X_i non indépendants.

Les paramètres $(\pi_{\alpha_i, j}^k)_{j \in \llbracket 1, J \rrbracket}$ correspondent à la probabilité que le participant k annote le label j à l'image i sachant que le label correct est le label α_i . On suppose que chaque participant ne labelise qu'une seule fois chaque image.

On ainsi

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

Avec $n_{i,j}^k$ le nombre de fois que le participant k à annoter le label j à l'image i

On a alors que la vraisemblance de $Y_i^k|X_i = \alpha_i$ est équivalent à $\prod_{j=1}^J (\pi_{\alpha_i,j}^k)^{n_{i,j}^k}$

Remarque 1. Ici on a une dépendance des paramètres $(\pi_{\alpha_i,j}^k)$ en k, j et α_i (pour le conditionnement), mais l'indice i sert simplement à indiquer qu'on considère l'image i et ne consitue pas un paramètre.

Cela signifie que, sachant que le label de l'image i est α_i et que le label de l'image i' est $\alpha_{i'}$ mais $\alpha_i = \alpha_{i'}$, la probabilité que le participant k annote le label j à l'image i est la même de celle où il annote le label j pour l'image i' . En tout état de connaissance, le participant sachant que les deux images ont le même label, la probabilité qu'il annote le label j à l'une, et j à l'autre n'a pas lieu d'être différente. Il s'agit pour le participant d'une même situation et le choix ne dépend que de lui.

Dans le but de se ramener au cas de plusieurs images et plusieurs participants, on considère ici une image i et tous les participants $(1, \dots, K)$.

Remarque 2. Par soucis de lisibilité on notera $\mathbb{P}(Y_i^k = (n_{i,1}^k, \dots, n_{i,J}^k)|X_i = \alpha_i)$ par $\mathbb{P}(Y_i^k|X_i = \alpha_i)$.

Pour tous les participants on a dans ce cas :

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

La vraisemblance du vecteur $(Y_i^k|X_i = \alpha_i)_k$ est équivalente à cette même valeur.

Ramenons nous à présent au cadre de notre problème et intéressons nous au vecteur $((Y_i^k)_k, X_i)$.

Nous pouvons de ce qui précède déterminer la probabilité $\mathbb{P}\left(\left(\bigcap_k Y_i^k\right) \cap (X_i = \alpha_i)\right)$.

On note $\forall i, j \in \{1, \dots, J\}^2, p_j = \mathbb{P}(X_i = j)$.

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

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 :

$$p_{\alpha_i} \prod_{k=1}^K \prod_{j=1}^J (\pi_{\alpha_i,j}^k)^{n_{i,j}^k} = \prod_{l=1}^J \left(p_l \prod_{k=1}^K \prod_{j=1}^J (\pi_{l,j}^k)^{n_{i,j}^k} \right)^{T_{i,l}}$$

De plus l'événement $\{X_i = \alpha_i\} = \{T_{i,\alpha_i} = 1\}$, $\forall l \in \{1, \dots, J\}$ $T_{i,l}$ est alors connu.

Il manque maintenant à se placer dans le cas où on observe les résultats des participants pour toutes les images et les labels de chaque image.

$\forall k, (Y_i^k) \perp\!\!\!\perp X_{i'}, i' \neq i$ et $(Y_i^k)_i \perp\!\!\!\perp$, on a alors $Y_i^k \perp\!\!\!\perp (Y_{i'}^k, X_{i'})$ puis $(Y_i^k, X_i) \perp\!\!\!\perp (Y_{i'}^k, X_{i'})$
On a ainsi la probabilité :

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

La vraisemblance est également équivalente à cette même expression.

1.2 Résolution du problème

Pour faire un choix quant au label de chaque image, la démarche que l'on adopte est la suivante :

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

$$j = \operatorname{argmax}_j (T_{i,j})$$

Les paramètres $T_{i,j}$ n'étant pas connu on souhaite les estimer.

Un estimateur naturel pouvant être prit pour chaque $T_{i,l}$ est l'espérance conditionnelle sachant $(Y_i^k)_k$, $\mathbb{E}[T_{i,l} | (Y_i^k)_k]$. Il s'agit de la meilleure approximation de $T_{i,l}$ au sens de la norme \mathcal{L}^2 , à partir de ce que l'on connaît $< (Y_i^k)_k >$ (les réponses des médecins).

$\mathbb{E}[T_{i,l} | (Y_i^k)_k]$ peut se réécrire sous la forme suivante :

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

En explicitant cette probabilité on a :

$$\begin{aligned}
\mathbb{P}(T_{i,l} = 1 | (Y_i^k)_k) &= \frac{\mathbb{P}((T_{i,l} = 1) \cap (Y_i^k)_k)}{\mathbb{P}((Y_i^k)_k)} \\
&= \frac{\mathbb{P}((Y_i^k)_k | (T_{i,l} = 1)) \times \mathbb{P}(T_{i,l} = 1)}{\sum_l^J \mathbb{P}((Y_i^k)_k | (X_i = l)) \times \mathbb{P}(X_i = l)} \\
&= \frac{\mathbb{P}((Y_i^k)_k | (X_i = l)) \times \mathbb{P}(X_i = l)}{\sum_l^J \mathbb{P}((Y_i^k)_k | (X_i = l)) \times \mathbb{P}(X_i = l)} \\
&= \frac{p_l \prod_k^K \prod_j^J (\pi_{l,j}^k)^{n_{i,j}^k}}{\sum_l^J p_l \prod_k^K \prod_j^J (\pi_{l,j}^k)^{n_{i,j}^k}}, \text{ car } \forall i, l \{X_i = l\} = \{T_{i,l} = 1\}
\end{aligned}$$

Remarque 3. On a donc réussi à obtenir une estimation des $T_{i,l}$ dépendant des paramètres $(\pi_{l,j}^k)_{k,l,j}$ et p_l . Cependant ces paramètres n'étant pas connus, il nous faut les estimer et on les estime par maximum de vraisemblance, à partir du modèle défini dans la partie précédente.

Un problème est présent, ceci étant dû au modèle que l'on a défini car, la détermination des estimateurs du maximum de vraisemblance nécessitent de bénéficier des observations des labels des images, qui ne sont pas connus, en particulier il nous faut avoir accès aux valeurs des $T_{i,l}$ dans l'expression de la vraisemblance.

Ceci ne nous permet pas, dans ce cas, de donner les estimateurs du maximum de vraisemblance des $(\pi_{l,j}^k)_{k,l,j}$ et p_l .

Toutefois, on voit apparaître une procédure permettant d'estimer les estimateurs du maximum de vraisemblance et l'espérance conditionnelle des $T_{i,l}$, il s'agit dans un premier temps, d'estimer arbitrairement les $T_{i,l}$ afin d'obtenir une estimation des estimateurs du maximum de vraisemblance $(\pi_{l,j}^k)_{k,l,j}$ et p_l .

On peut alors donner une estimation des $\mathbb{E}[T_{i,l} | (Y_i^k)_k]$, et de nouveau obtenir une estimation des estimateurs du maximum de vraisemblance des $(\pi_{l,j}^k)_{k,l,j}$ et p_l . Cette procédure itérative coorespond à ce que l'on appelle l'algorithme EM.

Il n'existe pas de preuve de convergence de la suite de paramètres établie par l'algorithme EM vers les estimateurs définie, et le choix de bons paramètres initiaux est de fait primordial.

Comme sus-mentionné en début de partie, nous disposons d'une garantie théorique de l'algorithme EM justifiant la démarche que nous avons introduit; celle-ci étant que les estimateurs ainsi obtenus ont la particularités de faire croître, à chaque itération, la vraisemblance des observations.

Toutefois, ce résultat ne prétend pas que l'algorithme EM maximisera la vraisemblance des observations. Ce dernier peut tout à fait produire une suite de paramètres correspondant à un maximum local; et une augmentation du nombre d'itérations n'apportera pas de solution à ce problème.

Considérons connus les valeurs de $T_{i,l}$, et déterminons à présent les expressions des estimateurs du maximum de vraisemblance.

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

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

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 suivant :

$$\widehat{\pi_{l,j}^k} = \frac{\sum_i T_{i,l} n_{i,j}^k}{\sum_l \sum_i T_{i,l} n_{i,j}^k}$$

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

Remarque 4. La preuve alourdissant la lecture, on se contente de donner l'idée générale permettant de retrouver les estimateurs.

Il convient de faire ressortir de la fonction ci-dessus, pour l'estimation des $\pi_{l,j}^k$ (respectivement des p_l), le terme en $\pi_{l,j}^k$ (respectivement en p_l). On obtient alors les termes suivant $\prod_i (\pi_{l,j}^k)^{n_{i,j}^k T_{i,l}} a_i$ et $\prod_i (p_l)^{T_{i,l}} b_i$ avec a_i et b_i correspondant au reste du produit.

Les contraintes de maximisation permettent ensuite d'obtenir les dénominateurs dans l'expression des estimateurs.

1.3 Initialisation des $T_{i,l}$

On peut proposer une initialisation simple des $T_{i,l}$ sans la base d'une quelconque information des données, en les estimant par :

$$T_{i,l} = \frac{1}{J}$$

Mais cette initialisation n'est pas optimale pour espérer estimer converger vers les estimateurs du maximum de vraisemblance.

Si les participants annotent suffisamment bien les images, on peut donner une meilleure initialisation des $T_{i,l}$ en se basant sur les résultats obtenus des participants, dans ce cas on estime $T_{i,l}$ par :

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

Il s'agit d'estimer $T_{i,l}$ par le nombre de fois que les participants ont annoté l à l'image i sur l'ensemble des annotations qu'ils ont donné pour l'image i , ici les $n_{i,l}^k$ valent soit 1 soit 0. On estime $T_{i,l}$ par vote majoritaire.

Chapitre 2

Pseudo-Code algorithme EM

Dans cette partie nous allons présenter l'algorithme EM, il s'agit d'une méthode itérative, constituée de deux étapes, à savoir une étape "Expectation" et une étape "Maximization". Cette algorithme nous permet de répondre à notre problématique initiale, par la garantie théorique de la croissance de la vraisemblance du modèle défini dans la partie Dawid-Skene.

2.1 L'étape E (Expectation)

La phase E consiste à calculer l'estimation des $T_{i,l}$ par $\mathbb{E}[T_{i,l} | (Y_i^k)_k]$.

Puisque seule son expression est nécessaire pour obtenir l'estimation des estimateurs du maximum de vraisemblance des $(\pi_{l,j}^k)_{k,l,j}$ et p_l . Son calcul est rendu possible par la forme analytique que l'on a pu déterminer précédemment à la page 6, il convient dans un premier temps d'estimer les $(\pi_{l,j}^k)_{k,l,j}$ et p_l en ayant utilisé l'initialisation des $T_{i,l}$.

2.2 L'étape M (Maximization)

La phase M consiste à maximiser la vraisemblance du modèle, ce maximum est atteint en les paramètres $(\widehat{\pi_{l,j}^k}, \widehat{p_l})$ et ces trois paramètres seront calculés à l'aide des maximas établis à la page 6. Ce sont ces estimateurs qui seront utilisés dans l'itération suivante pour mettre à jour estimation à l'étape E.

2.3 Pseudo-code

Algorithm 1 L'algorithme EM (Dempster et al., 1977).

Entrée(s) : 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;

1: **pour** n allant de 1 à N **faire**

2: **ETAPE M** : Calculer les $(\widehat{\pi_{l,j}^k})_n = \frac{\sum_i (T_{i,l})_{n-1} n_{i,j}^k}{\sum_l \sum_i (T_{i,l})_{n-1} n_{i,j}^k}$, $(\widehat{p_l})_n = \frac{\sum_i (T_{i,l})_{n-1}}{I}$

3: **ETAPE E** : Calculer $(\widehat{T_{i,l}})_n = \frac{(\widehat{p_l})_n \prod_k \prod_j (\widehat{\pi_{l,j}^k})_n^{n_{i,j}^k}}{\sum_l (\widehat{p_l})_n \prod_k \prod_j (\widehat{\pi_{l,j}^k})_n^{n_{i,j}^k}}$;

4: **fin du pour**

5: **retourner** $\max_l (\widehat{T_{i,l}})_N$;
