Projet Master 2 SSD

Crowd-Sourcing

Rédigé par

PRALON Nicolas

THIRIET Aurelien

Encadrant : Joseph SALMON

Table des matières

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 construit 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'intervenant permet de construire la base de donnée nécessaire. Toute fois des erreurs d'obervations quant à la nature des données peut-être commisent, 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 obervations d'une même donnée sont fournit.

Nous définissons alors l'exemple d'un esemble d'images à labeliser, les experts du domaines décident alors de faire appel à plusieurs participants pour labéliser ces images.
On définit le cadre suivant :

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

A partir des résultats donner par chaque participant nous pouvons essayer d'attribuer un label à chaque image. Toute fois nous pouvons nous retrouver dans la situation où le label donner par un participant à une image est diffère 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 accèssible, 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,* $\forall j, k \in [\![1, J]\!] \times [\![1, K]\!]$, $Y_i^k | X_i = \alpha_i \sim Multinomiale((\pi^k_{\alpha_i,j})_{j\in[\![1,J]\!]}, 1)$

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 indépendant.

Les paramètres $(\pi^k_{\alpha_i,j})_{j\in[\![1,J]\!]}$ 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

$$\mathbb{P}_{Y_i^k|X_i=\alpha_i} = \sum_{(n_{i,1}^k,\cdots,n_{i,J}^k)} \underbrace{\frac{(\sum_{j=1}^{J} n_{i,j}^k)!}{\prod_{j=1}^{J}(n_{i,j}^k)!} \prod_{j=1}^{J} \left(\pi_{\alpha_i,j}^k\right)^{n_{i,j}^k}}_{densit\acute{e}\ suivant \sum \delta_{(n_{i,1}^k,\cdots,n_{i,J}^k)}} \quad \delta_{(n_{i,1}^k,\cdots,n_{i,J}^k)}$$

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

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 particpant 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 particpant 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 $(1, \cdots, K)$.

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

Pour tous les participants on a dans ce cas :

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

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 $\mathbb{P}\left(\left(\bigcap_k Y_i^k\right) \bigcap (X_i = \alpha_i)\right)$.

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

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

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 :

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

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

Il manque maintenant à se placer dans le cas où on observe les résutats 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 :

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

La vraisemblance est également équivalente à cette même expréssion.

1.2 Résolution du problème

Pour faire un choix quant à la maladie de chaque patient, la démarche que l'on adopte est la suivante :

Soit le patient i, on décide que ce patient est atteint de la maladie j si :

$$j = argmax_j(\mathbb{P}(X_i = j) = p_j = \mathbb{P}(T_{i,j} = 1))$$

Les paramètres $(p_j)_j$ n'étant pas connu on souhaite les estimer, et on les estime par maximum de vraisemblance, à partir du modèle défini dans la partie précédente.
Un problème tout fois 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 connu, en particulier il nous faut avoir accès aux valeurs des $T_{i,l}$ dans l'exprèssion de la vraisemblance.

Dans un premier temps nous pouvons considérer connu les valeurs de Ti, l afin de déterminer les expréssions des estimateurs du maximum de vraisemblance, on se concentrera dans un second temps à comment déterminer les $T_{i,l}$.

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

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

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

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 3. 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 $\pi_{l,j}^{k}$ (repectivement en p_l). On obtient alors les termes suivant $\prod_{i}^{I} \left(\pi_{l,j}^{k}\right)^{n_{i,j}^{k} T_{i,l}} a_i$ et $\prod_{i}^{I} (p_l)^{T_{i,l}} b_i$ 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.

A présent que l'on dispose des estimateurs du maximum de vraisemblance, il nous faut déterminer les $T_{i,l}$, nécessaire au calcul des estimateurs. Ces derniers 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 $\mathscr{L}^2$, à partir de ce que l'on connait $< (Y_i^k)_k >$ (les réponses des médecins).

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

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

En explicitant cette probabilité on a :

$$\begin{aligned}
\mathbb{P}\left(T_{i,l} = 1|(Y_i^k)_k\right) &= \frac{\mathbb{P}\left((T_{i,l} = 1) \cap \left(Y_i^k\right)_k\right)}{\mathbb{P}\left(\left(Y_i^k\right)_k\right)} \\
&= \frac{\mathbb{P}\left(\left(Y_i^k\right)_k | (T_{i,l} = 1)\right) \times \mathbb{P}(T_{i,l} = 1)}{\sum_{l}^{J} \mathbb{P}\left(\left(Y_i^k\right)_k | (X_i = l)\right) \times \mathbb{P}(X_i = l)} \\
&= \frac{\mathbb{P}\left(\left(Y_i^k\right)_k | (X_i = l)\right) \times \mathbb{P}(X_i = l)}{\sum_{l}^{J} \mathbb{P}\left(\left(Y_i^k\right)_k | (X_i = l)\right) \times \mathbb{P}(X_i = l)} \\
&= \frac{p_l \prod_{k}^{K} \prod_{j}^{J} (\pi_{l,j}^{k})^{n_{i,j}^{k}}}{\sum_{l}^{J} p_l \prod_{k}^{K} \prod_{j}^{J} (\pi_{l,j}^{k})^{n_{i,j}^{k}}}, \textit{ car } \forall i, l \ \{X_i = l\} = \{T_{i,l} = 1\}
\end{aligned}$$

On a donc réussi à obtenir une estimation des $T_{i,l}$, mais celle-ci dépend des paramètres recherchés. 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.
Toute fois, on voit alors apparaître une procédure permettant d'estimer les estimateurs du maximum de vraisemblance, 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'algoritme EM.

Il n'existe pas de preuvre de convergence de la suite de paramètres établie par l'algorithme EM et le chois de bons paramètres initiaux est de fait primordial.
Comme sus-mentionné en début de partie, nous disposons d'une garantie théorique de l'algoritme EM justifiant la démarche que nous avons introduit ; celle-ci étant que les estimateurs ainsi obtenus ont la particularités de faire croitre, à chaque itération, la vraisemblance des observations.
Toute fois, 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.