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

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 suivant :

- On dispose d'un esemble de patient $\{1, \cdots, I\}$ tous atteint d'une maladie, et on suppose le vecteur aléatoire $(X_i)_{i \in \{1,\cdots,I\}}$, tel que $\forall i$, X_i à valeur dans $\{1, \cdots, J\}$ représente la maladie du partient i.
- On dispose egalement d'un esemble de médecin $\{1, \cdots, K\}$, et on considère $\forall i \in \{1, \cdots, I\}$ le vecteur aléatoire $(Y_i^k)_{k \in \{1,\cdots,K\}}$ le diagnostique du médecin k au patient i, Y_i^k à valeur dans $\{0, 1\}^J$.
- Chaque médecin répond indépendament des autres et la maladie de chaque patient est indépendante des autres. On note l'independance 2 à 2, $(Y_i^k)_k \perp\!\!\!\perp$ et $(X_i)_i \perp\!\!\!\perp$

Afin de diagnostiquer le meilleur traitement à chaque patient, on doit leur diagnostiquer une maladie. Toute fois nous pouvons nous retrouver dans la situation où le diagnostique différent d'un médecin à un autre, pour diverses raisons ; erreur de mesure, erreur d'annotation, fatigue. Il convient dans ce cas de trouver une solution à la question : quelle maladie est atteint chaque patient ?

Dans le but de simplifier l'approche, nous considérons dans un premier temps le cas d'un medecin et d'un patient, nous pouvons également émettre l'hypothèse vraisemblable 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 le partient i est atteint de la maladie α_i, la réponse du médecin k suit une loi multinomiale. Le médecin peut, ou non, commettre une érreur de mesure même en connaissant la vraie maladie du patient.

Les paramètres $(\pi^k_{\alpha_i,j})_{j \in [\![1,J]\!]}$ correspondent à la probabilité que le médecin k diagnostique la maladie j au patient i sachant qu'il est atteint de la vraie maladie α_i. On suppose que chaque médecin ne voit qu'une seule fois chaque patient.

On ainsi

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

Avec $n^k_{i,j}$ le nombre de fois que le médecin k à diagnostiqué la maladie j au patient i

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

Dans le but de se ramener au cas de plusieurs patients et plusieurs médecins, on considère ici un patient i et tous les médecins $(1, \cdots, K)$.