TD3 - Optimisation Stochastique - Maximum de vraisemblance.

Introduction

Au cours d'une enquète, on recherche un suspect dont la taille est 1m80. Afin d'orienter les recherches rapipdement, doit-on chercher un homme ou une femme ? Si m_h et m_f désignent la taille moyenne des hommes et des femmes, et si σ_h et σ_f désignent les différents écarts types, on suppose que la distribution de probabilité suit la loi gaussienne de densité

$$p(x, \theta) = \frac{1}{2\pi\sigma^2} e^{-(x-m)^2/2\sigma^2}$$

Pour décider de rechercher un suspect masculin, on a en réalité chercher le jeu de paramètres θ_h ou θ_f maximisant la fonction

$$\theta \longmapsto p(x, \theta)$$

On observe un échantillon X de loi de paramètre θ inconnu. Pour tout $\theta \in \Theta$, on définit la vraisemblance de l'échantillon $X_1, \dots X_n$ comme

$$p(X_1, \dots X_n, \theta) = p(X, \theta) = P(X|\theta)$$

De même, la fonction $L(X, \theta) = \log p(X, \theta)$ désigne la log vraisemblance de l'échantillon X pour le jeu de paramètres θ. L'estimateur du maximum de vraisemblance est le jeu de paramètre permettant de maximiser cette log vraisemblance (ou vraisemblance). C'est ainsi le *jeu de paramètres le plus probable.*

Application du maximum de vraisemblance

Exercice 1

On considère le modèle gaussien $\mathcal{P} = \left\{\mathcal{N}(\mu, \sigma_0^2)\right\}$ où la variance est connue mais pas la moyenne. On mesure une réalisation d'un n échantillon $X_1, \dots X_n$, donner l'esimateur $\hat{\mu}_n$ du maximum de vraisemblance de l'échantillon. Cet estimateur est-il convergent, estime-t-il correctement le bon paramètre μ ?

Exercice 2

On considère un modèle uniforme $\left\{\mathcal{U}_{[0;\theta]}, \theta > 0\right\}$. Calculer de la même façon que dans le premier exercice l'estimateur du maximum de vraisemblance d'un n échantillon $X_1, \dots X_n$. Démontrer en plus que cet estimateur est convergent vers le bon paramètre θ.

Exercice 3

Reprendre l'exercice 1 dans le cas où l'écart type est cette fois inconnu.

Algortithme EM

On suppose qu'une variable aléatoire X de densité $f(X|\theta)$ où $\theta \in \Theta$ est inconnu. On suppose de plus que l'observation est incomplète, c'est-à-dire qu'en réalité le vecteur de données complètes

$x \in \mathcal{X}$ n'est pas accessible directement et on ne connaît que $y(x)$. On cherche néanmoins à calculer le θ le plus probable associé au modèle. C'est possible grâce à l'algorithme EM (Expectation Maximization).

1. Si l'on pouvait mesurer x, quel serait la méthode naturelle pour trouver la valeur de θ ?
2. On définit
$$l_c(\theta|x) = \log f(x|\theta)$$
l'algorithme consiste alors à calculer une suite de paramètres $(\theta_k)_{k\geq 0}$ tel que
   - θ_0 est pris quelconque dans Θ.
   - Si l'on suppose connu θ_k, on calcule la fonction
$$q(\theta, \theta_k) = \mathbb{E}_{\theta_k}\left[l_c(\theta|X)|y\right]$$
C'est donc l'espérance conditionnelle à l'observation de y lorsqu'on suppose que X est donné par une loi de paramètre θ_k.
   - On pose θ_{k+1} qui maximise la fonction $q(\theta, \theta_k)$ sur Θ.

   On définit
$$L(\theta|y) = \int_{\mathcal{X}(y)} f(x|\theta)dx$$
où $\mathcal{X}(y)$ désigne l'ensemble des x tels que $y(x) = y$. Démontrer que la suite $L(\theta_k|y)$ est une suite croissante.
3. Expliciter les étapes E et M lorsque la densité de X s'écrit :
$$f(x|\theta) = b(x)e^{c^t(\theta)t(x)}/a(\theta)$$
Trouver un exemple de modèle paramétré d'une telle façon.