

Modèles de mélange de densités

Introduction

Véronique Tremblay

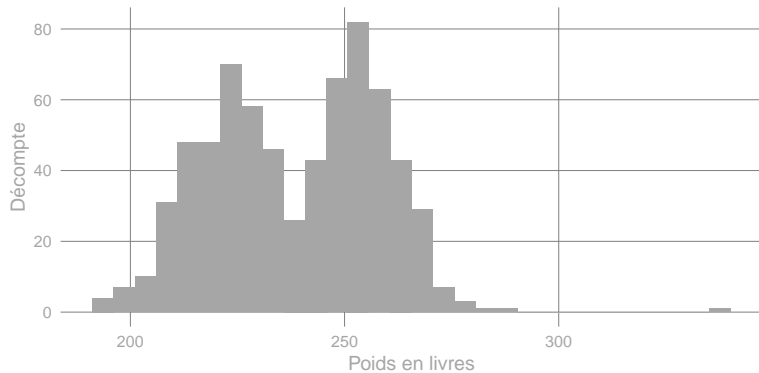
- Mélange de densités
- Model-based Clustering
- Mixture Modeling / modèles de mélange
- Gaussian mixture
- Latent Class Analysis (LCA) pour des variables nominales ou ordinales
- Latent Profile Analysis (LPA) pour des variables continues

Avantages de cette approche

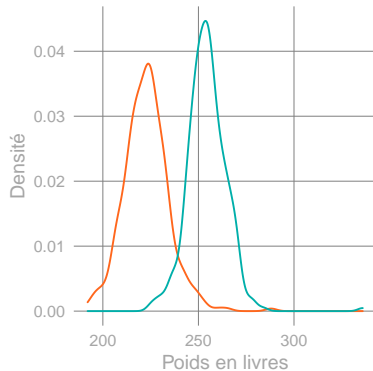
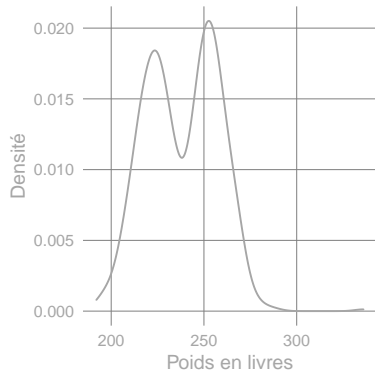
- Plusieurs types de variables
- Pas besoin de standardiser, ni de choisir une mesure de distance
- Elle se base sur la **vraisemblance**
- Permet d'obtenir une probabilité d'appartenance à un groupe

Le concept

Exemple classique



Exemple classique



Posons Y , un jeu de données de n observations et P variables.

$$f(Y, \theta) = \sum_{i=1}^K \pi_i f_i(Y, \theta_i)$$

Y : Les données

f_i : la loi du groupe i

π_i : le poids du groupe i ($\sum^K \pi_i = 1$)

θ_i : les paramètres de la loi du groupe i

$f(x)$ dépend du type de variable

Type d'attribut	Densité (paramètre-s)
Binaire	Binomiale ($\theta = p$)
Nominale + ordinale	Multinominale($\theta = p_1, \dots, p_{k-1}$)
Dénombrement	Poisson($\theta = \mu$)
Continue	Normale($\theta = \mu, \sigma^2$)
Continue positive	Gamma($\theta = \alpha, \beta$)

Estimation des paramètres

La vraisemblance aura la forme suivante:

$$\mathcal{L}(\theta|Y) = \prod_{j=1}^n \sum_{i=1}^K \pi_i f(Y_j, \theta_i)$$

La log-vraisemblance sera :

$$\ell\ell(\theta|Y) = \sum_{j=1}^n \log \left[\sum_{i=1}^K \pi_i f(Y_j, \theta_i) \right]$$

- Mélange de densités
- Basé sur le maximum de vraisemblance