

Similarities between classes for the clustering task

Description The similarity between classes k and ℓ , denoted by $\Sigma(k, \ell) \in [0, 1]$, is defined by

$$\Sigma(k, \ell) = 1 - \sqrt{\frac{1}{n} \sum_{i=1}^n (P(Z_i = k | \mathbf{X}_i = \mathbf{x}_i) - P(Z_i = \ell | \mathbf{X}_i = \mathbf{x}_i))^2}. \quad (1)$$

A high value of $\Sigma(k, \ell)$ (close to one) means that components k and ℓ are strongly different (*i.e.*, low overlapping). A low value of $\Sigma(k, \ell)$ (close to zero) means that components k and ℓ are similar for the clustering task (*i.e.*, high overlapping).

Notations Data $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ are composed of n i.i.d observations $\mathbf{x}_i = (x_{i1}, \dots, x_{id})$ described by d variables and defined on space \mathcal{X} . Clustering is achieved with a mixture model of K components assuming independence within components between variables. Therefore the probability distribution function (pdf) of the mixture model is

$$f(\mathbf{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k f_k(\mathbf{x}_i; \boldsymbol{\alpha}_k), \text{ with } f_k(\mathbf{x}_i; \boldsymbol{\alpha}_k) = \prod_{j=1}^d f_{kj}(x_{ij}; \boldsymbol{\alpha}_{kj}), \quad (2)$$

where $\boldsymbol{\theta} = (\pi_k, \boldsymbol{\alpha}_k; k = 1, \dots, K)$ groups the model parameters, π_k is the proportion of component k , f_k is the pdf of component k whose parameters are denoted by $\boldsymbol{\alpha}_k$ and f_{kj} is the pdf of variable j for component k whose parameters are denoted by $\boldsymbol{\alpha}_{kj}$.

The partition is denoted by $\mathbf{z} = (z_1, \dots, z_n)$ where $z_i = k$ means that observation i arises from component k . Therefore,

$$P(Z_i = k | \mathbf{X}_i = \mathbf{x}_i) = \frac{\pi_k f_k(\mathbf{x}_i; \boldsymbol{\alpha}_k)}{\sum_{\ell=1}^K \pi_\ell f_\ell(\mathbf{x}_i; \boldsymbol{\alpha}_\ell)}. \quad (3)$$