Discriminative level and Similarities between variables

Description: discriminative level of variables The discriminative level of variable j, denoted by $\text{Discrim}(j) \in [0,1]$, is defined by

$$\text{Discrim}(j) = 1 - \frac{\sum_{k=1}^{K} E_{kj}}{n \ln K} \tag{1}$$

where $E_{kj} = -\sum_{i=1}^{n} P(Z_i = k|X_{ij} = x_{ij}) \ln P(Z_i = k|X_{ij} = x_{ij})$ is the marginal entropy of component k for variable j. A high value of $\text{Discrim}(j)$ (close to one) means that X_j is highly discriminating. A low value of $\text{Discrim}(j)$ (close to zero) means that X_j is poorly discriminating.

Description: discriminative level of variables in a cluster The discriminative level of variable j for a cluster k, denoted by $\text{Discrim}(j,k) \in [0,1]$, is defined by

$$\text{Discrim}(j,k) = 1 - \frac{E_{kj} + \bar{E}_{kj}}{n \ln 2} \tag{2}$$

where $\bar{E}_{kj} = -\sum_{i=1}^{n} \big(1 - P(Z_i = k|X_{ij} = x_{ij})\big) \ln \big(1 - P(Z_i = k|X_{ij} = x_{ij})\big)$ is the marginal entropy of component k for all variables except j. A high value of $\text{Discrim}(j,k)$ (close to one) means that X_j is highly discriminating in cluster k. A low value of $\text{Discrim}(j,k)$ (close to zero) means that X_j is poorly discriminating in cluster k.

Description: similarities between variables for the clustering task The similarity between variables j and h, denoted by $\Delta(j,h) \in [0,1]$, is defined by

$$\Delta(j,h) = 1 - \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}(P(Z_i = k|X_{ij} = x_{ij}) - P(Z_i = l|X_{ih} = x_{ih}))^2}. \tag{3}$$

A high value of $\Delta(j,h)$ (close to one) means that X_j and X_h provide the same information for the clustering task (i.e. similar partitions). A low value of $\Delta(j,h)$ (close to zero) means that X_j and X_h provide some different information for the clustering task (i.e. different partitions).

Notations Data $\mathbf{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ are composed of n i.i.d. observations $\boldsymbol{x}_i = (x_{i1}, \dots, x_{id})$ 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

$$f(\boldsymbol{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\alpha}_k), \text{ with } f_k(\boldsymbol{x}_i; \boldsymbol{\alpha}_k) = \prod_{j=1}^{d} f_{kj}(x_{ij}; \boldsymbol{\alpha}_{kj}), \tag{4}$$

where $\boldsymbol{\theta} = (\pi_k, \boldsymbol{\alpha}_k; k = 1, \ldots, K)$ 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 $z = (z_1, \ldots, z_n)$ where $z_i = k$ means that observation i arises from component k. Therefore,

$$P(Z_i = k | \boldsymbol{X}_i = \boldsymbol{x}_i) = \frac{\pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\alpha}_k)}{\sum_{l=1}^{K} \pi_l f_l(\boldsymbol{x}_i; \boldsymbol{\alpha}_l)}. \tag{5}$$

If only the realization of variable j is observed then

$$P(Z_i = k | X_{ij} = x_{ij}) = \frac{\pi_k f_{kj}(x_{ij}; \boldsymbol{\alpha}_{kj})}{\sum_{l=1}^{K} \pi_l f_{lj}(x_{ij}; \boldsymbol{\alpha}_{lj})}. \tag{6}$$