Group Sparsity with Probabilistic Feature Clustering

Lucas E. Morales {`lucasem`}

Workflow

CrossCat

$$\alpha_D \sim \text{Gamma}(k=1, \theta=1)$$
$$\vec{\lambda}_d \sim V_d(\cdot) \qquad \forall d \in \{1, \ldots, D\}$$
$$z_d \sim \text{CRP}(\{z_i | i \neq d\}; \alpha_D) \qquad \forall d \in \{1, \ldots, D\}$$
$$\alpha_v \sim \text{Gamma}(k=1, \theta=1) \qquad \forall v \in \hat{z}$$
$$y_r^v \sim \text{CRP}(\{y_i^v | i \neq r\}; \alpha_v) \qquad \forall v \in \hat{z},\ \forall r \in \{1, \ldots, N\}$$
$$\hat{\theta}_c^d \sim M_d(\cdot; \hat{\lambda}_d) \qquad \forall v \in \vec{z},\ \forall c \in \vec{y}^v$$
$$\forall d \in \{z_{(\cdot)} = v\}$$
$$\vec{x}_{(\cdot,d)}^c = \{x_{(r,d)} | y_r^{z_d} = c\} \sim \Pi_r L_d(\vec{\theta}_c^d) \qquad \forall v \in \vec{z},\ \forall c \in \vec{y}^v$$

Dependence

$$\mathbb{P}[\mathcal{I}_{\mathcal{G}}(x_i : x_j) > 0]$$
$$< \mathbb{P}[\{\mathcal{G} : z_i = z_j\}]$$
$$\approx \frac{1}{H} \sum_{h=1}^{H} \mathbb{I}[\hat{\mathcal{G}}_h : z_i^h = z_j^h]$$

K-Means Clustering

$$S^* = \arg\min_S \sum_{i=1}^{k} \sum_{j \in S_i} ||x_j - \mu_i||^2$$
$$\text{BIC} = -2 \cdot \ln \hat{L} + k \cdot \ln(n)$$

Cluster Thresholding

$$\frac{1}{||S_k||^2} \sum_{i,j \in S_k} x_i^j \geq p_t$$

Overview

- Dataset as subset of *population*
- Population associated with *generative population model* (GPM)
- Cross-Categorization (CrossCat) Bayesian meta-GPM for population
- Feature dependence as probability of non-zero mutual information
- Clustering using k-means with k-means++ initialization
- X-means using *Bayesian information criterion* (BIC) to determine *k*
- Threshold mean inter-cluster dependence prob. for cluster destruction

Application

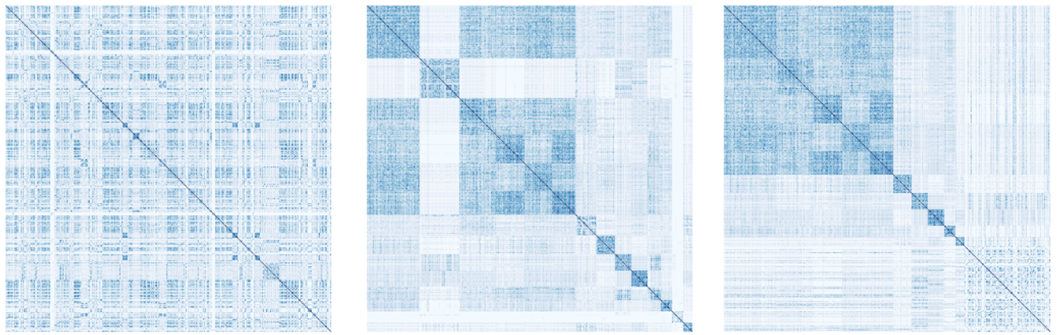
- Group sparsity for high-dimensional datasets
- Works with supervised, semi-supervised, and unsupervised learning

Selected Citations

- (2015) Mansinghka et al. Crosscat: A fully bayesian nonparametric method … arXiv preprint arXiv:1512.01272
- (2016) Saad and Mansinghka. Detecting dependencies in high-dimensional … arXiv preprint arXiv:1611.01708
- (2000) Pelleg et al. X-means: Extending k-means with efficient estimation of the number of … In ICML, (Vol. 1)

Experiment

Used on the Gapminder dataset:



Heatmap of dependence probabilities.
Left: arbitrary order.
Center: ordered by cluster assignment after k-means (k = 14).
Right: ordered by cluster assignment after cluster thresholding (k = 10 + 84).