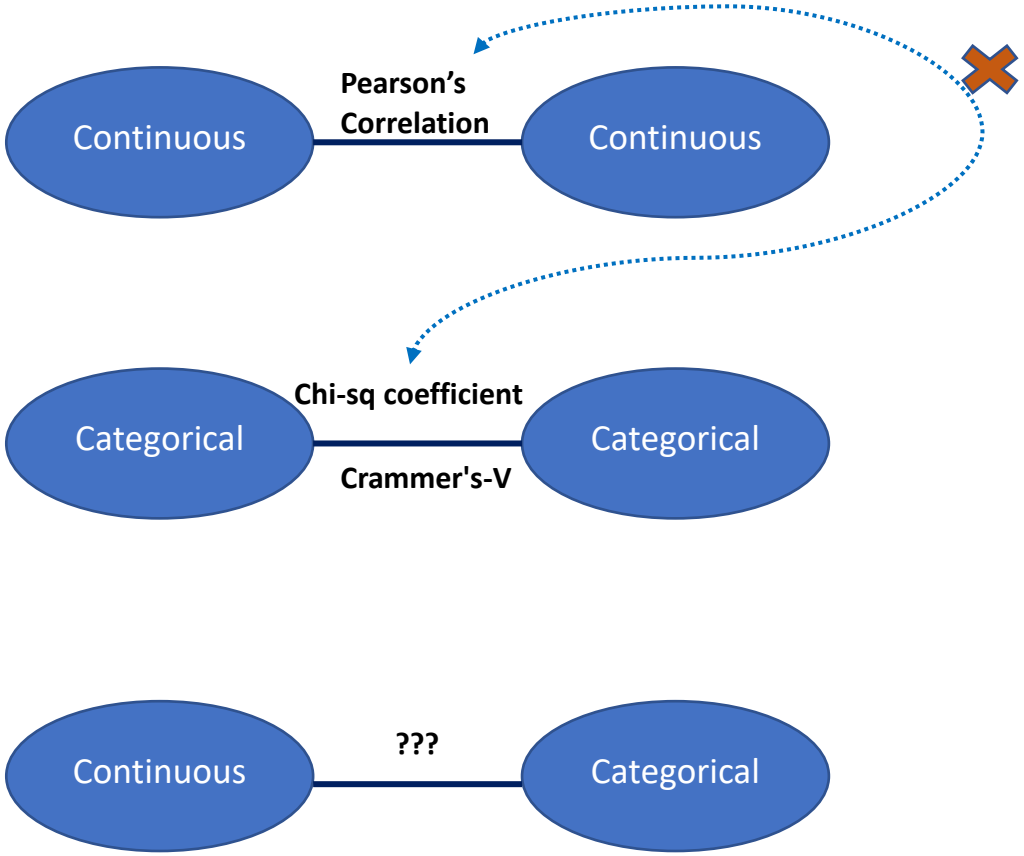
Graph Spectral Feature Learning for Mixed Data of Categorical and Numerical Type

Saswata Sahoo – Gartner
Souradip Chakraborty – Walmart Labs

Motivation & Complexity of the Problem

- Representation Learning in a heterogeneous space with mixed variables of numerical and categorical types has interesting challenges due to its complex feature manifold.
- Moreover, feature learning in an unsupervised setup, without class labels and a suitable learning loss function, adds to the problem complexity further.
- Hence the object is finding a suitable transformation which retains important information on the mixed variables without distorting the data geometry

Motivation & Complexity of the Problem



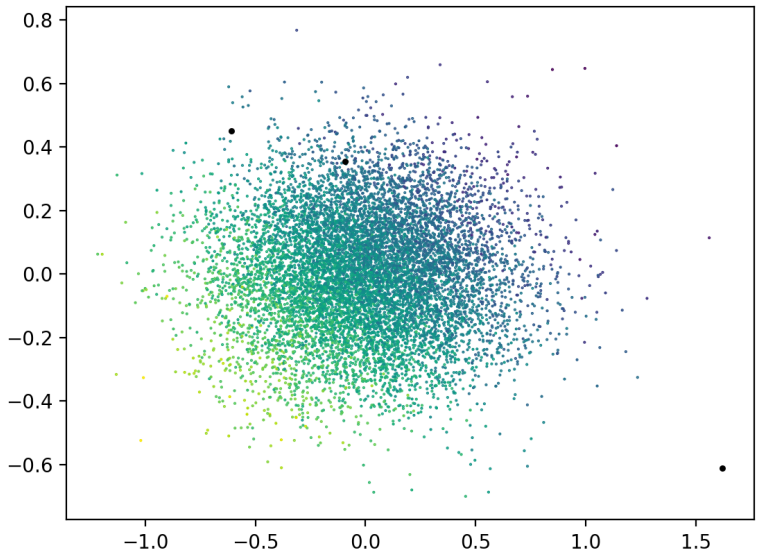
- Modelling the joint dependence structure of the mixed type of variables by a suitable representation learning model.
- The similarities between pairs of numerical, categorical or mixed variables should be directly comparable
- The entire framework should be in an unsupervised setting as in majority of the real-life scenarios limited amount of supervised information is available
- Supervised learning causes the representation to learn specifics on the dataset and not generic or robust representations which is a concern

Literature Survey & Prior Research - Mixed Space Representation Learning

- Past research works are based on linear projection on the Principal component space introduced by [*Pearson, 1901*] or based on projection on a suitable Kernel space given by [*Scholkopf et al., 1998*]
- Unlike such global approaches, there is a large class of methods involving locality-based manifold learning such as Local linear embedding [*Roweis and Saul, 2000*] or Isometric feature mapping [*Tenenbaum et al., 2000*]
- Singular value decomposition-based approach is used to get low rank representation of the data matrix which also gives feature representation of the data points preserving the pairwise dissimilarities

However, such approaches do not explicitly consider the inter-dependence structure among the categorical and continuous variables in a heterogeneous data.

Overview : Graph Spectral Feature Learning of Mixed Type Variables



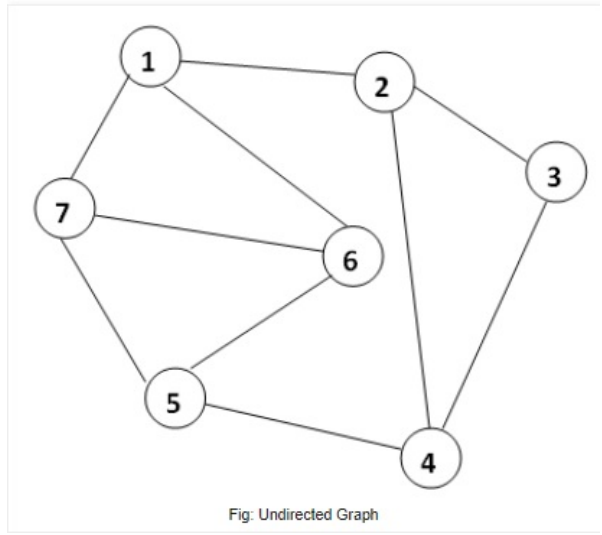


Fig: Undirected Graph

We propose a Graph Spectral Feature Learning based approach by explicitly modelling the probabilistic dependence structure among the mixed type of variables with a pairwise Markov model. Spectral decomposition of the Laplacian matrix provides the desired feature transformation

Problem Formulation & Definition

- Let us consider the mixed type of variables listed in the random vector $X = (X^{(num)'}, X^{(cat)'})'$ of dimension p, where

$$p = p1 + p2$$
$$X^{(num)} \rightarrow p1$$
$$X^{(cat)} \rightarrow p2$$

- There are n independent p dimensional observations on the random vector X and the data matrix D is partitioned into two submatrices giving observations on the numerical and categorical variables for all the data points as $D = [D^{(num)}, D^{(cat)}] \rightarrow p = p1 \times p2,\ p = p1 + p2$

- The main problem is to produce a feature transformation $\phi(.)$ which transforms the p dimensional mixed observation x(i) to $\phi(x(i))$.

Methodology : Undirected Graphical Model

- We assume the components of X as vertices of an undirected graph and the edges of the graph indicate the conditional dependence structure among the random variables. Let's say, the graph *G = (V,E)* with vertices V = {1,2,...,p} and edge set E.

- The joint distribution of the random vector X is assumed to factorize as $f(x) \propto exp[\sum_{(s,t)\in\mathcal{E}} \psi(x_s, x_t)]$

- For each edge *(s, t)* $\in E$, the function $\psi(x_s, x_t)$ is a mapping of the edge to the real line. We consider a special case:

$$\psi(x_s, x_t) = \theta_{st} h(x_s, x_t)$$

where, $\theta_{st} \in R$ and $h(x_s, x_t) \in [-1, 1]$ is a known function, indicating similarity between x_s and x_t.

- To define the similarity function h for mixed type of variables first it is important to describe the collective factorization strategy which maps observations on categorical variables to a continuous dense space, so that the mixed type of observations are directly comparable.

Collective Factorization of Numerical and Categorical Space

- The main idea behind the collective factorization is that both the categorical and numerical variables can be represented in a common hidden space.

- The common latent space representation of the data is done by considering a common matrix W for factorization of $D^{(num)}$ and $D^{(cat)}$ both, by the following optimization problem

$$\min_{W,H_1,H_2} : ||\mathcal{D}^{(num)} - WH_1|| + ||\mathcal{D}^{(cat)} - WH_2||$$

- The common latent space representation is achieved by restricting $W \in R^{n \times k}$ where k is the common latent dimension and W consists of factor loadings of latent features of both numerical and categorical variables.

- The similarity between the numerical observation $x^{(num)}$ and categorical observation $x^{(cat)}$ is measured by

$$\hat{x}^{(num)} = \hat{w}\hat{H}_1$$
$$\hat{x}^{(cat)} = \hat{w}\hat{H}_2$$

Similarity Function for the Edge potential

- The similarity function between observations on a pair of variables is defined as follows :

$$
\begin{aligned}
h(x_s, x_t) &= x_s x_t && \forall s,t \in \mathcal{V}^{(cat)} \\
&= g(x_s, x_t) && \forall s,t \in \mathcal{V}^{(num)} \\
&= g(\hat{x}_s, \hat{x}_t) && \forall s \in \mathcal{V}^{(num)}, \forall t \in \mathcal{V}^{(cat)}.
\end{aligned}
$$

- Note that the similarity function h(.,.) is bounded in [−1,1], where -1 and +1 respectively indicate extreme dissimilarity and extreme similarity of observations on a pair of variable

Model Estimation and Pseudo Likelihood

- The joint Pseudo log-likelihood function of the model parameters $\theta = \{\theta_{st}, s, t \in V\}$ based on the observed data points $D = [x_{(1)}, x_{(2)}, \ldots, x_{(n)}]'$ is given by

$$\log L(\theta, \mathcal{D}) = \sum_{i=1}^{n} \sum_{\forall (s,t) \in \mathcal{E}} \theta_{st} h(x_{(i)s}, x_{(i)t}) \; - n \log(Z(\theta))$$

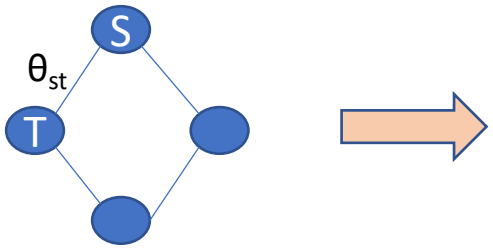
where, $Z(\theta) = \sum_{i=1}^{n} \exp(\sum_{\forall (s,t) \in \mathcal{E}} \theta_{st} h(x_{(i)s}, x_{(i)t}))$

- The parameters are estimated maximizing the Pseudo log-likelihood function based on the following gradient update : $\hat{\theta}_{st} = \theta_{st} - \eta \frac{\partial \log L(\theta, \tilde{\mathcal{D}})}{\partial \theta_{st}}$. The gradient direction is given by :

$$\frac{\partial \log L(\theta, \mathcal{D})}{\partial \theta_{st}} = \sum_{i=1}^{n} h(x_{(i)s}, x_{(i)t}) \; - \frac{n}{Z(\theta)} \sum_{i=1}^{n} \exp(\sum_{\forall s,t \in E} \theta_{st} h(x_{(i)s}, x_{(i)t})) h(x_{(i)s}, x_{(i)t})$$

Feature Transformation and Representation Learning

- The p dimensional square matrix of the model parameters are estimated maximizing the Pseudo log-likelihood function as $\hat{\Theta} = (\hat{\theta}_{st}, s, t \in \mathcal{V})$



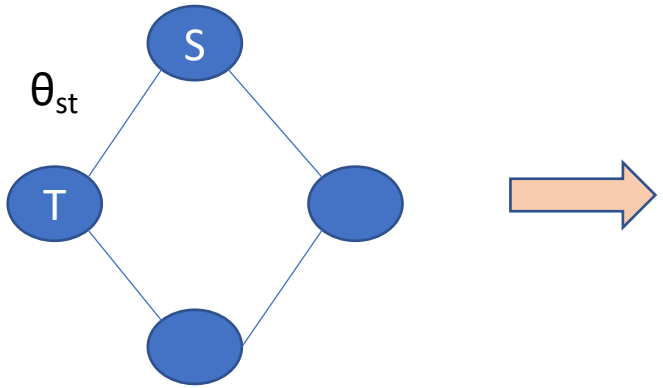
The conditional dependence structure among the random variables is estimated using a symmetric version of Θ, which is given by

$$\tilde{\Theta} = \frac{1}{2}(\hat{\Theta} + \hat{\Theta}^T)$$

- We evaluate the graph Laplacian given by Δ = D − Θ, where $D = Diag\left(\sum_t |\tilde{\theta}_{st}|, \forall s\right)$ is the degree matrix.

- The graph Laplacian has eigen decomposition given by $\Delta = \Phi\Lambda\Phi^T$, where $\Phi = (\phi_1, \phi_2 \ldots, \phi_p)$ are the orthonormal eigen vectors and Λ is the diagonal matrix of the eigen values

Feature Transformation and Representation Learning



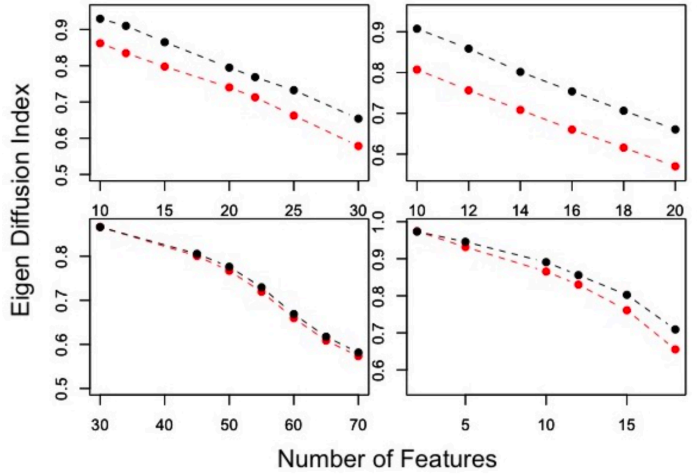
An observation on the mixed random vector X, denoted by x, is treated as a signal on the vertices V of the graph G and the desired feature transformation is given by

$$\phi(x) = \Phi^T x$$

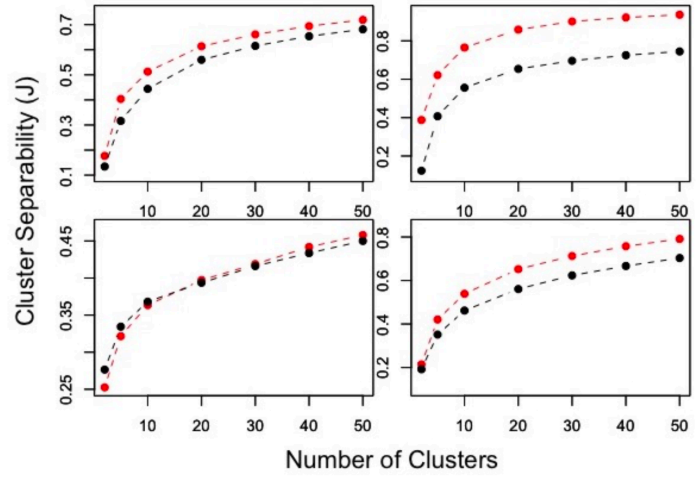
This feature map can be used to represent the observations on the mixed variables in a dense space.

Numerical Investigation of the Representation Learned

Eigen Diffusion index (α) for varying feature dimensions of the proposed spectral embedding feature map (in red) and naive principal component features (in black):
Clockwise starting from top left: Credit Approval Dataset, Adult Dataset, German Credit dataset, Heart Disease Dataset



Cluster Separability index (J) for varying number of clusters with naive *K-means* on the proposed spectral embedding features (in red) and actual mixed data points (in black): Clockwise starting from top left: Credit Approval Dataset, Adult Salary Dataset, German Credit Dataset, Heart Disease Dataset



Numerical Investigation of the Representation : Comparative Performance

Methods	SE-KMeans			SE-KMediods			SE-HAgglo		
Dataset	Number of Features	RI	NMI	Number of Features	RI	NMI	Number of Features	RI	NMI
Credit Approval Dataset	28	0.624	0.243	28	0.662	0.248	28	**0.671**	**0.306**
	37	**0.625**	**0.256**	30	0.673	0.295	30	0.629	0.215
	40	0.621	0.255	40	**0.699**	**0.328**	35	0.602	0.235
	45	0.573	0.188	45	0.667	0.292	40	0.565	0.174
Adult Salary Dataset	10	0.590	0.193	17	0.587	0.198	10	0.588	0.191
	15	0.588	0.191	20	0.586	0.189	15	0.632	**0.233**
	20	**0.639**	**0.226**	22	**0.588**	**0.208**	18	**0.639**	0.223
	25	0.561	0.183	25	0.556	0.010	22	0.638	0.226
German Credit Dataset	40	0.550	**0.028**	60	0.542	**0.040**	35	0.567	0.002
	50	0.586	0.002	62	**0.562**	0.036	50	0.591	0.003
	60	**0.590**	0.004	67	0.531	0.027	67	**0.678**	**0.017**
	70	0.574	0.004	70	0.529	0.001	70	0.567	0.004
Heart Disease Dataset	12	0.629	0.193	10	0.509	0.017	12	0.569	0.113
	15	0.633	0.199	12	0.653	0.267	15	0.566	0.109
	18	0.657	0.238	13	0.640	0.250	18	0.629	**0.245**
	19	**0.669**	**0.271**	19	**0.674**	**0.323**	19	**0.636**	0.211

Fig : Cluster quality obtained with varying dimension of the proposed feature map. The optimal values of RI and NMI are given in bold font.

Numerical Investigation of the Representation : Comparative Performance

	Datasets	Credit Approval Dataset		Adult Salary Dataset		German Credit Dataset		Heart Disease Dataset	
	Methods	RI	NMI	RI	NMI	RI	NMI	RI	NMI
Clustering on Proposed Features	SE-KMeans	0.625	0.257	**0.639**	**0.227**	**0.599**	**0.028**	**0.669**	**0.271**
	SE-KMediods	**0.699**	**0.328**	0.582	**0.208**	0.562	**0.036**	**0.674**	**0.323**
	SE-HAgglo	**0.672**	**0.306**	**0.639**	**0.226**	**0.678**	**0.017**	0.629	0.245
Competitor Clustering Methods for Mixed Data	K-Mediods	0.670	0.281	0.550	0.126	0.537	0.008	0.669	0.260
	K-Prototype	0.571	0.172	0.540	0.162	0.587	0.001	0.669	0.262
	WKM	0.662	0.281	0.603	0.092	0.504	0.003	0.611	0.188
	EWKM	0.654	0.243	0.626	0.012	0.511	0.013	0.643	0.221
	OCIL	0.601	0.182	0.624	0.004	0.559	0.003	0.641	0.231
	AE-KMeans	0.580	0.142	0.581	0.006	0.552	0.019	0.584	0.143

Fig : Cluster quality obtained using different clustering techniques. Clusters based on the proposed feature map outperforming all the competitors are shown in bold font

References

[1] H. Suresh and J. V. Guttag, “A framework for understanding unintended consequences of machine learning,” *arXiv preprint arXiv:1901.10002*, 2019.

[2] K. Pearson, “On lines and planes of closest fit to systems of points in space.” *Philos Mag*, 1901.

[3] H. Zou, T. Hastie, and R. Tibshirani, “Sparse principal component analysis,” *Journal of Computational and Graphical Statistics*, vol. 15, no. 2, pp. 265–286, 2006. [Online]. Available: http://www.jstor.org/stable/27594179

[4] B. Scholkopf, A. Smola, and K. R. Muller, “Nonlinear component analysis as a kernel eigenvalue problem.” *Neural Comput.*, vol. 10:1299e1319, 1998.

[5] S. Roweis and L. Saul, “Nonlinear dimensionality reduction by locally linear embedding.” *Science.*, vol. 290:2323e2326, pp. 405–416, 2000.

[6] J.Tenenbaum, V.Silva, and J.Langford, “A global geometric framework for nonlinear dimensionality reduction.” *Science.*, vol. 290:2319e2323, pp. 405–416, 2000.

Thank You !!!