Group Sparsity with Probabilistic Feature Clustering

Lucas E. Morales {lucasem}

Abstract

High-dimensional datasets often have sets of variables which are co-dependent. Recognizing these dependence relations permits a natural degree of sparsity for interpretable reduction of dimensionality. In this paper, we demonstrate how to find these co-dependent sets of variables using a combination of probabilistic programming, information theory, and statistical learning. We also discuss the applicability of this method to sparsity-based learning problems.

1 Introduction

Datasets with hundreds of variables are not rare finds. These variables, though representing different features of a datum, are often related. This relation is often calculated as the correlation, a statistical metric. We abandon this method of relating variables in favor of a more probabilistic approach to better express the generative models behind the data. These relations are then used to cluster related variables, yielding a sparse set of grouped variables.

We treat the dataset as a sample from a population, described by a generative probabilistic model program. We use tools from probabilistic programming to infer a distribution over suitable programs, and use the resulting meta-model to estimate the dependence between different variables of the dataset. We then use clustering algorithms to group sets of probably co-dependent variables together, ultimately yielding a probably expressive set of grouped variables and a reduction in dimensionality. Finally, we discuss the applicability of this method to problems of sparsity, and the enhanced utility in semi-supervised learning scenarios.

2 Modeling the data

The dataset can be conceptualized as a table with a finite number of columns and rows. A *population*, from which the dataset is sampled, can analogously be thought of as a table with a finite number of columns but an infinite number of rows. A generative population model (GPM), then, characterizes the generating process for a population [Saad and Mansinghka, 2016b]. We use Cross-Categorization (CrossCat), a non-parametric Bayesian meta-GPM, to infer models appropriate for the dataset [Mansinghka et al., 2015]. In the words of the aforementioned paper, the CrossCat model "consists of a Dirichlet process mixture over the columns of a table where each mixture component is itself an independent Dirichlet process mixture over the rows of the table." We chose to use CrossCat for data analysis because this model structure inherently represents dependencies between variables.

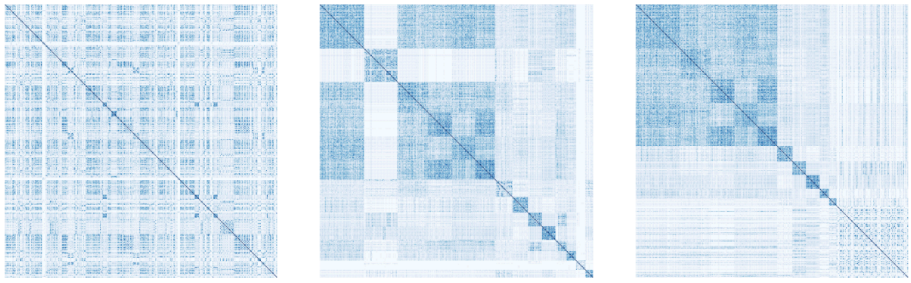

Figure 1: Heatmap of dependence probabilities in the Gapminder dataset of global statistics. Left: arbitrary order. Center: ordered according to cluster assignment after performing k-means ($k = 14$). Right: ordered according to cluster assignment after thresholding the mean inter-cluster dependence probabilities ($k = 10 + 84$).

3 Measuring dependence between features

After modeling the data using CrossCat, we detect pairwise predictive relationships between every two variables (x_i, x_j) using the dependence probability

$$\mathbb{P}[\mathcal{I}(x_i : x_j) > 0]$$

where $\mathcal{I}(x : y)$ is the mutual information of x and y. This value can be upper-bounded by the posterior probability that x_i and x_j are assigned to the same variable partition in CrossCat [Saad and Mansinghka, 2016a]. From this we construct a table of pairwise dependence probabilities, illustrated in the left-most diagram of Figure 1.

4 Learning clusters of dependent features

Many variables in the dataset may exhibit similar dependencies with other variables. We learn clusters to group these variables accordingly for a given cluster count k using k-means algorithm with k-means++ initialization [Lloyd, 1982][Arthur and Vassilvitskii, 2007]:

$$S^{*(k)} = \arg\min_{S} \sum_{i=1}^{k} \sum_{j \in S_i} ||x_j - u_i||^2$$

We determine k using the X-means method, minimizing the Bayesian information criterion

$$k^* = \arg\min_{k} -2\hat{L}_k + p_k \ln n$$

where $\hat{L}_k$ is the maximum likelihood for the data given the k-clustering, p_k is the number of free parameters, and n is the sample size [Pelleg et al., 2000]. These learned clusters are shown in the center diagram of Figure 1.

Some clusters correspond to similar high-dependency relations, while others have similar low-dependency relations. Because low-dependency relations shouldn't ultimately be clustered, we use a selection mechanism to choose clusters to either maintain or destruct. We compute the mean inter-cluster dependence probability for each cluster C as

$$\frac{1}{||C||^2} \sum_{i,j \in C} \mathbb{P}[\mathcal{I}(i : j) > 0]$$

and apply each to a threshold. A cluster is maintained iff the mean inter-cluster dependence probability meets the threshold. The resulting clusters are shown in the right-most diagram of Figure 1.

5 Comparison with correlation matrix

Pairwise correlations may be used instead of dependence probabilities for clustering. One of the more notable shortcomings of using correlations between two variables is that data must be dropped if they don't have values for either of the dimensions. Alternatively, CrossCat can impute undefined values based on a models inferred from defined data. Because the Gapminder dataset used in Figure 1 lacks several fields for many data samples, we use the Dartmouth Atlas of Healthcare for comparison [Rosling, 2008][Fisher et al., 2011]. An illustration of the clusters using correlation and dependence probability is shown in Figure 2 on the next page.

6 Application for group sparsity learning problems

The feature clusters represents co-dependent groups of variables in the dataset. As such, they make excellent candidates for use in group sparsity learning problems, where they would either be included in the learned model, or excluded entirely. This application, however, is not clearly useful for unsupervised or supervised problems, because the population modeled by CrossCat may be used directly. In a semi-supervised learning problem, supervised group sparsity methods can be used where the groups are learned, as described in this paper, in an entirely unsupervised manner. Models that fall within this scope include group LARS and group LASSO [Yuan and Lin, 2006], though any learning approach where non-overlapping groups are supplied (where the groups are indicative of a correspondence between sets of variables) are applicable.

7 Conclusion

We demonstrate a method for finding groups of probably co-dependent features in a dataset. We use CrossCat, a probabilistic meta-modeler, to infer sets of appropriate models for the data. These models are used to estimate an upper bound for the dependence probability between every pair of variables. We then used k-means to estimate sets of similar features according to their probable dependence relations, and destruct clusters which have low inter-cluster dependence. These new groups are then fit for application to groups sparsity learning problems. This method still needs to be tested on such problems and compared to alternative means of determining groups for it to be determined effective.

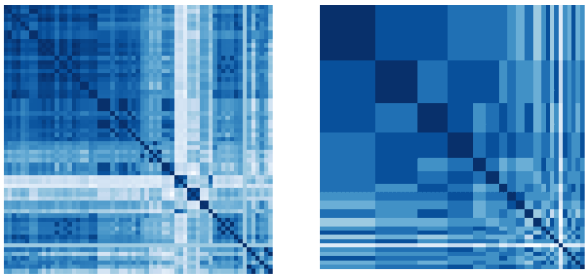

Figure 2: Clustered heatmaps of correlation (left, $k = 14$) and dependence probabilities (right, $k = 21$) in the Dartmouth Atlas of Healthcare.

References

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

Elliott S. Fisher, David C. Goodman, John E. Wennberg, and Kristen K. Bronner. The dartmouth atlas of health care. *URL http://www.dartmouthatlas.org/*, 2011.

Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

Vikash Mansinghka, Patrick Shafto, Eric Jonas, Cap Petschulat, Max Gasner, and Joshua B Tenenbaum. Crosscat: A fully bayesian nonparametric method for analyzing heterogeneous, high dimensional data. *arXiv preprint arXiv:1512.01272*, 2015.

Dan Pelleg, Andrew W Moore, et al. X-means: Extending k-means with efficient estimation of the number of clusters. In *ICML*, volume 1, 2000.

Hans Rosling. Gapminder: World. *URL http://www.gapminder.org/world*, 2008.

Feras Saad and Vikash Mansinghka. Detecting dependencies in high-dimensional, sparse databases using probabilistic programming and non-parametric bayes. *arXiv preprint arXiv:1611.01708*, 2016a.

Feras Saad and Vikash Mansinghka. Probabilistic data analysis with probabilistic programming. *arXiv preprint arXiv:1608.05347*, 2016b.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.