DENSITY TRAVERSAL CLUSTERING - Amos J Storkey and Tom G Griffiths, School of Informatics, University of Edinburgh.

In this work we present a generative procedure for clustered data. Clusters are understood to be a density dependent concept rather than a data dependent one. To encompass this, a Markov chain diffusion model is defined with respect to a specific density, and clusters are understood to be regions which mix well under the Markov chain. By interpreting the initialisation points of this diffusion model as cluster centres, we can obtain a likelihood from the probability, under the diffusion model, of arriving a particular point having started at the cluster centre. However, this likelihood is dependent on an underlying density, called the traversal density. We show that because of the form of the diffusion model we choose, we can approximate the traversal density with the empirical data distribution. The result is a fully computable model of the same form as a that of spectral clustering by Markov relaxation, but with different affinity matrices. The difference allows the affinity matrices (or powers of the affinity matrices) to be properly interpreted and used as (approximate) scaled likelihoods, and hence allows the clustering procedure to be incorporated into many different probabilistic modelling procedures.

We have shown that this model can be practically used for clustering using variational Bayesian methods and that the probabilistic formalism allows the estimation of the number of clusters by model comparison.

A number of different approaches for describing and implementing spectral clustering methods have been proposed in the past, all of which shed light on different aspects of the spectral clustering problem, and highlight the depth of the approach. The work of Shi and Malik [3] developed the approach for the benefit of the machine learning and vision communities, utilising work in spectral graph theory, and improving on a proposal of Wu and Leahy [5]. Since that, the paper of Ng, Jordan and Weiss [2] provided the most commonly used approach for spectral clustering, using k-means clustering in the principal eigenspace.

Tishby and Slonim [4] described the spectral clustering problem in terms of Markov relaxation, where clusters are found within the nth iteration of a Markov transition matrix. Meila and Shi [1] focus on edge flows within a random walk to establish the positions of clusters. The work of Brand and Huang establishes an interpretation of spectral clustering in terms of the dynamics of the angles between eigenvectors as the number of considered eigenvalues of the affinity matrix reduces. Spectral clustering can also be used as a precursor for k-means clustering [6]. Of the different understandings of spectral clustering, the Markov relaxation view will the the most useful in the context of this work.

Density Traversal Clustering

Our approach to understanding and modelling clustering has the following benefits. First, we take a starting point that clusters are properties of densities not of data. Any model of clusters should be meaningfully applicable to a density as well as a set of data points, and different datasets drawn from the same underlying density fundamentally must have the same clusters. The issue is inferring what those clusters are from given data. Second, we define the clustering model as a generative process. An implication of this is that the space in which the data lies is not ignored in the underlying framework. Third, the model is flexible; it is applicable to many different forms of density, including those where the data lies on submanifolds. Fourth, the method directly relates to spectral clustering and hence sheds light on the working of other spectral clustering methods.

In our formalism, we utilise a diffusing Markov chain to model the generation of clusters from some cluster centre. We define this process dependent on some global latent distribution which we call the traversal distribution.

Suppose we have a cluster centre at position $\mathbf{x}_0$. We define a Markov chain, using the traversal distribution P^*, by the transitions

$$P(\mathbf{x}_t|\mathbf{x}_{t-1}) = \frac{D(\mathbf{x}_t,\mathbf{x}_{t-1})P^*(\mathbf{x}_t)S^{-1}(\mathbf{x}_t)}{Z(\mathbf{x}_t)} \text{ where } S \text{ is given by the solution of } S(\mathbf{x}) = \int d\mathbf{y} \frac{P(\mathbf{y})D(\mathbf{y},\mathbf{x})}{S(\mathbf{y})} \tag{1}$$

and $D(\mathbf{y},\mathbf{x})$ is some locality function, typically a squared-exponential, $D(\mathbf{y},\mathbf{x}) = \exp((\mathbf{y}-\mathbf{x})^2)$. Z is a

normalisation function:

$$Z(\mathbf{x}) = \int d\mathbf{y} D(\mathbf{y}, \mathbf{x}) P^*(\mathbf{y}) S^{-1}(\mathbf{y}) \tag{2}$$

One can think of this transition in three parts. First the D term ensures the transitions are local. Transitions between nearby points have by far the highest probability. Second, the P^* term ensures that the transitions respect the traversal density P^*: the chain is unlikely to transition to regions of low probability for P^*. Third, the S term ensures global consistency - this makes sure that ultimately the Markov chain visits regions of the space in a way that is consistent with the traversal density. This can be seen from the fact that this chain satisfies detailed balance with respect to the traversal distribution, and hence the equilibrium distribution of this Markov chain is precisely P^*.

What we have defined is a process of diffusing from some point $\mathbf{x}_0$, with respect to some distribution P^*. If, in fact, $\mathbf{x}_0$ is sampled from the distribution P^* in the first place then the points after any τ iterations of the Markov chain will also be sampled from P^*. Suppose, then, we have the following generative model for the data. Points are generated from P^* and are then diffused using the Markov chain defined above. The resulting data points we see would also be sampled from P^*. However, given some data, inferring the origin $\mathbf{x}_0$ of a given data point $\mathbf{x}_\tau$ under this generative procedure is hard because a) we do not know the traversal distribution P^*, and b) we cannot calculate the necessary integrals in (1) and (2) and in calculating the required marginalisation over the variables involved in the intermediate steps of the Markov chain, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{\tau-1}$.

In order to produce a feasible approach we use the fact that, for this model, the empirical data is a sample from P^*, and revert to a sampling approximation. This follows, because the Markov chain satisfies detailed balance with respect to P^*. As all the integrals we need to calculate are integrals with respect to P^*, we can replace them by Monte Carlo sample approximations using the observed data as the samples. This enables us to obtain an estimate for the Markov chain which now only depends on the positions of the data points, but is computed in such a way as to be consistent with the underlying true distribution, at least in the limit of infinite data. The result is a process very similar to spectral clustering by Markov relaxation, except that the resulting transitions can properly be interpreted as approximate scaled likelihoods:

$$\frac{P(\mathbf{x}_i|\mathbf{x}_j)}{P(\mathbf{x}_j)} = n(\mathbf{A}^\tau)_{ij} \quad \text{where} \quad (\mathbf{A})_{ij} = \frac{D(\mathbf{x}_i, \mathbf{x}_j)/S(\mathbf{x}_j)}{\sum_i D(\mathbf{x}_i, \mathbf{x}_j)/S(\mathbf{x}_j)}, \tag{3}$$

where n is the number of data points, and $S(\mathbf{x}_i) = \sum_j D(\mathbf{x}_i, \mathbf{x}_j)/S(\mathbf{x}_j)$. These transitions are different from the standard affinity matrices used in spectral clustering due to the consistency term S, which allows us to properly and consistently utilise the data points as a proxy for the traversal distribution in order to gain the probabilistic interpretation as a scaled likelihood.

We have been able to utilise this interpretation to obtain a variational Bayes procedure for a clustering model utilising a mixture model for cluster centres. Because of the probabilistic framework, we can obtain estimates for the number of clusters by standard approaches to model comparison. It has also enabled an application to an improvement in Gaussian process models for clustered data in cases of missing targets. By also incorporating a Metropolis-Hastings rejection step, many other related models can be generated, resulting in slightly different affinity matrices. The approach has been tested on toy and real problems with at least comparable results with standard spectral clustering.

Conclusion

The formalism we describe here provides a generative understanding for affinity between points, based on the distribution from which points are generated, and which provides a basis for clustering models. By using an empirical approximation for the underlying distribution we recover a form for affinity matrices and the standard Markov relaxation view of spectral clustering. However, the exact procedure for deriving affinity matrices is slightly different from the usual procedure, but results in a form that can be interpreted in terms of a scaled likelihood. This probabilistic formalism for the generation of data allows spectral clustering methods to be slotted in to many probabilistic frameworks. As a simple example it allows standard model comparison methods for determining the number of clusters.

References

[1] M. Meila and J. Shi. Learning segmentation by random walks. In *Advances in Neural Information Processing Systems 13 (NIPS 2000)*, 2001.

[2] A.Y. Ng, M.I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14 (NIPS 2001)*, 2002.

[3] J. Shi and J. Malik. Normalised cuts and image segmentation. *IEEE PAMI*, 22:888–905, 2000.

[4] N. Tishby and N. Slonim. Data clustering by markovian relaxation and the information bottleneck method. In *Advances in Neural Information Processing 13*, pages 640–646, 2001.

[5] Z. Wu and R. Leahy. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15:1103–1113, 1993.

[6] H. Zha, X. He, C. Ding, M. Gu, and H. Simon. Spectral relaxation for k-means clustering. In *Advances in Neural Information Processing Systems 16 (NIPS 2003)*, 2004.