Nathanael Fillmore, proposal for summer research, 2009

Manifold K-means clustering via algebraic topology

We consider the problem of clustering n points $X = \{x_1, \ldots, x_n\}$, $x_i \in \mathbb{R}^p$, when p is large. This problem—clustering in high-dimensional spaces—is of practical importance, because high-dimensional data abounds; for example, typical DNA microarray data has more than 10000 dimensions [20], and it is of scientific interest to find patterns in this data.

Learning in high-dimensional spaces is associated with many difficulties due to Bellman's "curse of dimensionality" [2] and related issues (e.g., [6]). Manifold learning is attractive because it offers a way to avoid these issues. If the points of interest are drawn from a q-dimensional manifold $\mathcal{M}$ embedded in $\mathbb{R}^p$, and $q \ll p$, then this offers the possibility of finding an algorithm whose theoretical properties depend on q rather than p.

There is good reason to believe that much data of interest does lie on low-dimensional manifolds: frequently the process used to generate the data has only a few degrees of freedom. For example, [10] uses physical properties of the human vocal tract to argue that speech sounds lie on a low-dimensional manifold embedded in L_2.

Some classical approaches, such as principal component analysis [11], are effective, but are limited to learning linear subspaces. In many cases, we have reason to believe that a dataset lies on a nonlinear manifold. For this reason, a large number of methods, such as locally linear embedding [17], Laplacian eigenmaps [1], kernel PCA [18], and others, have been proposed for learning nonlinear manifolds. See [13, 19] for an overview.

In recent work, Niyogi, Smale, and Weinberger (NSW) have formulated a new manifold learning framework based on insights from algebraic topology [15, 16]. A distinctive property of their method is that, under appropriate conditions, its sample complexity is independent of p and its computational complexity is only dependent on p insofar as the points must be examined initially. Recent related work includes [3, 4, 5, 7].

NSW's method is based on the following result, stated here informally: If the curvature of the manifold $\mathcal{M}$ is not too large, the variance of the distribution centered on $\mathcal{M}$ from which the points X are drawn is not too large, the codimension of the manifold is not too small, and other technical conditions hold, then with high probability the true homology of $\mathcal{M}$ can be found, using an algorithm that depends on p only linearly in computational complexity and not at all in sample complexity ([16], Thm. 2.1).

This result suggests an efficient way to cluster the points in X, in two steps [16]: (1) Use X to find the homology of $\mathcal{M}$. (2) Designate the connected components of the homology as clusters. The number of connected components, and thus clusters, is equal to the zeroth Bette number β_0 of the homology.

A weakness of this approach is that in many cases there is not a one-to-one relationship between con-

nected components and clusters, as NSW's method assumes. First, a single connected component can be comprised of more than one cluster. Consider for example the following three manifolds:

(a) (b) (c)

How many clusters are there? For many applications, we might want to say that there is one cluster in (a) and there are two clusters in both (b) and (c). However, (a) is homeomorphic to (b), while (b) is not homeomorphic to (c), so NSW's algorithm above will find just one cluster for (b).

On the other hand, a cluster can be comprised of multiple disconnected components. Consider, for example, a mixture of a large number of Gaussians, with means as shown below by a dot, and with variance comparatively large:

Each mean will have its own connected component, so NSW's algorithm will find a large number of clusters, but it might be better for some applications to find only two clusters in a dataset generated from this mixture.

K-means, a classical clustering algorithm (e.g., [14]), would give reasonable results on both of the above examples. The K-means algorithm assigns each point x_i to the cluster j, centered at μ_j ($j = 1, \ldots, K$), for which $\|x_i - \mu_j\|$ is smallest in some norm. An assignment of x_i to cluster j is recorded in an indicator matrix w, where

$$w_{ij} = \begin{cases} 1 & \text{if } x_i \text{ is in cluster } j \\ 0 & \text{otherwise} \end{cases}$$

A locally optimal choice of μ and w is classically found by iterating between the following two steps until convergence: (i) Update w so that each x_i is assigned to the nearest μ_j. (ii) Update each μ_j as the average $\sum_{i=1}^n w_{ij} x_i / \sum_{i=1}^n w_{ij}$.

Performed in the ambient space with a Euclidean norm, K-means can only find linearly separable clusters. In light of this and the examples above, it is of interest to consider performing K-means on the manifold $\mathcal{M}$. This can be done, using existing methods that rely on a geodesic distance function, e.g. [8, 12].

An interesting question, and the principal question we propose to consider this summer, is this: how can it be done efficiently? NSW's clustering algorithm is particularly efficient because the clustering falls out as a natural byproduct of the learning of the manifold. Can we modify NSW's manifold learning algorithm so that a K-means clustering, on the manifold, falls out as a byproduct of the construction? As an example of what is possible, we mention Ding and He [9], who showed a close connection between K-means and PCA and used this connection to construct an alternative K-means algorithm.

For ease of analysis, we will first consider a noiseless setting, where each point in X is drawn uniformly at random from the manifold. After we find a suitable algorithm in the noiseless setting and determine its sample and computational complexity, we will consider the more realistic noisy setting where each point is drawn from a distribution centered on $\mathcal{M}$ but with support on all of $\mathbb{R}^p$.

Finally, we will validate our results empirically, using both intrinsic and extrinsic measures, and we will consider connections and differences with other results in the literature. An important longer-term goal of this project is to develop connections between topology and machine learning.

References

[1] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6), 2003.

[2] R. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.

[3] J.-D. Boissonnat, L. J. Guibas, and S. Y. Oudot. Manifold reconstruction in arbitrary dimensions using witness complexes. *Proc. 23rd ACM Sympos. on Comput. Geom.*, 2007.

[4] F. Chazal and A. Lieutier. Smooth manifold reconstruction from noisy and non uniform approximation with guarantees. *Computational Geometry: Theory and Applications*, 40, 2008.

[5] S.W. Cheng, T.K. Dey, and E.A. Ramos. Manifold reconstruction from point samples. *SODA*, 2005.

[6] Robert Clarke, Habtom W. Ressom, Antai Wang, Jianhua Xuan, Minetta C. Liu, Edmund A. Gehan, and Yue Wang. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer*, 2008.

[7] Tamal K. Dey and Kuiyu Li. Topology from data via geodesic complexes. 2009. Manuscript.

[8] Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: spectral clustering and normalized cuts. In *KDD*, 2004.

[9] Chris Ding and Xiaofeng He. k-means clustering via principal component analysis. In *ICML*, 2004.

[10] Aren Jansen and Partha Niyogi. Intrinsic fourier analysis on the manifold of speech sounds. *ICASSP*, 2006.

[11] I.T. Jolliffe. *Principal Component Analysis*. Springer, 2002.

[12] Jaehwan Kim, Kwang-Hyun Shim, and Seungjin Choi. Soft geodesic kernel k-means. *ICASSP*, 2007.

[13] John A. Lee and Michel Verleysen. *Nonlinear Dimensionality Reduction*. Springer, 2007.

[14] D. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge Univ. Press, 2003.

[15] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of the submanifolds with high confidence from random samples. In *Discrete and Computational Geometry*, 2006.

[16] P. Niyogi, S. Smale, and S. Weinberger. A topological view of unsupervised learning from noisy data. 2008. Manuscript.

[17] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. In *Science*, 2000.

[18] Bernhard Schlkopf, Alexander Smola, and Klaus-Robert Mller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, (5).

[19] L.J.P. van der Maaten, E.O. Postma, and H.J. van den Herik. Dimensionality reduction: A comparative review. 2008.

[20] Eric P. Xing, Michael I. Jordan, and Richard M. Karp. Feature selection for high-dimensional genomic microarray data. In *ICML*, 2001.