

Nearest Neighbor Clustering

Ulrike von Luxburg, Sébastien Bubeck¹, Stefanie Jegelka, Michael Kaufmann²



Clustering is often formulated as a discrete optimization problem: given a finite set of sample points, the objective is to find, among all partitions of the data set, the best one according to some quality measure. However, in the statistical setting where we assume that the finite data set has been sampled from some underlying space, the goal is not to find the best partition of the given sample, but to approximate the true partition of the underlying space.

The first insight gained in this project is that the discrete optimization approach is not well-suited to achieve this goal. To see this, we provide examples where this approach leads to overfitting: the clustering constructed on a finite sample does not converge to the optimal clustering of the underlying space, as the sample size n tends to infinity. Then we show that in order to avoid such overfitting effects, one has to apply the same cure as in supervised learning: one has to restrict the class of functions from which the clustering on the finite data space can be chosen to some “small” function space. For appropriate “small” function classes we can prove very general consistency theorems for clustering optimization schemes.

As one particular algorithm for clustering with a restricted function space we introduce “nearest neighbor clustering”. This algorithm can be seen as a general baseline algorithm to minimize arbitrary clustering objective functions. In essence, it works with a class of functions which are constant on local neighborhoods. This function class is “small”. On the other hand, we allow the function class to slowly grow with n , by allowing the neighborhoods to become smaller and smaller. In the limit for $n \rightarrow \infty$, we can then approximate any clustering on the underlying data space up to arbitrary precision. Formally, we prove that nearest neighbor clustering is statistically consistent for all commonly used clustering objective functions such as the

k -means objective function, graph cut objective functions, and many others. This consistency result is stronger than most of the existing consistency results in the literature such as for k -means [2] or spectral clustering [3].

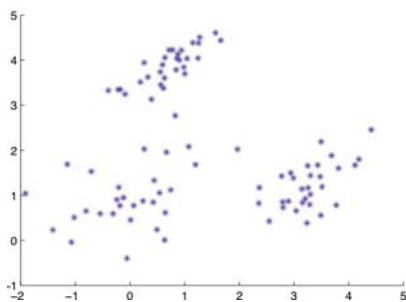
An interesting side effect of our approach is as follows. Most discrete optimization problems in clustering are NP hard and thus inherently difficult to solve. Consequently, in practice people usually apply heuristic approximations to “solve” the optimization problems, usually without any approximation guarantees to the true global optimum. In our approach, however, we provide a polynomial-time algorithm (due to the polynomially small function class) which has (stochastic) approximation guarantees to the global solution of the problem.

This work has been published in von Luxburg et al. [4] and Bubeck and von Luxburg [1].

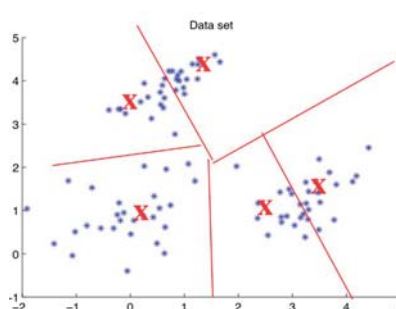
References

1. **Bubeck S., U. von Luxburg:** Nearest Neighbor Clustering: A Baseline Methode for Consistent Clustering with Arbitrary Objective Functions. *Journal of Machine Learning Research*, (accepted) (2009).
2. Pollard, D.: Strong Consistency of k -means Clustering. *Annals of Statistics* 9(1), 135–140 (1981).
3. **von Luxburg, U., M. Belkin, O. Bousquet:** Consistency of Spectral Clustering. *Annals of Statistics* 36(2), 555–586 (2008).
4. **von Luxburg, U., S. Bubeck, S. Jegelka, M. Kaufmann:** Consistent Minimization of Clustering Objective Functions. In *Advances in Neural Information Processing Systems 20: Proceedings of the 2007 Conference*, 961–968. (Eds.) Platt, J. C., D. Koller, Y. Singer, S. Roweis, MIT Press, Cambridge, MA, USA (2008).

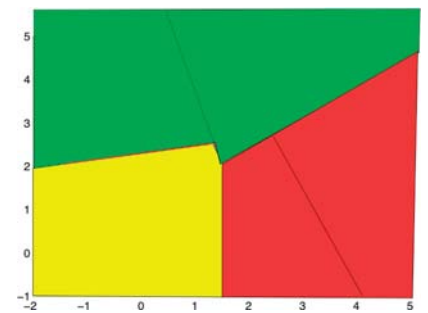
Data points



Seeds and their neighborhoods



Nearest neighbor clustering



Nearest neighbor clustering selects random seed points from the data and only considers candidate clusterings which are constant on the induced Voronoi cells. Among those clusterings, NNC chooses the “best one” according to a given clustering objective function.

¹ INRIA Futurs, Lille, France; ² Department of Computer Science, Universität Tübingen

Understanding Clustering Stability

Ulrike von Luxburg, Shai Ben-David¹

Clustering stability is a popular tool to select the number K of clusters. The rationale is that the “best” K is the one for which the clustering algorithm leads to the “most stable” clustering results. Even though this approach is very popular in practice, its theoretical foundations are far from being understood. In [2] we already pointed out that in many cases, clustering stability is *not* related to the “correct” number of clusters, at least in the large sample size regime. Instead, clustering stability just detects whether the objective function has one or several global optima, which often is not related to the correct number of clusters. Several follow-up papers took up the discussion [3], Shamir and Tishby [5], Shamir and Tishby [4].

Subsequently, we tried to relate the stability of clustering algorithms (on finite sample sizes) to properties of the optimal data clustering itself. In the small/moderate sample region, a conjecture was formulated during a PASCAL workshop organized in summer 2007 in Tübingen [6]: *stable clusterings tend to have their boundaries in regions of low density*. If this conjecture were true, it would have a very large impact on understanding the mechanism of stability based model selection. For example, in the case of K-means, this would explain the success of stability-based methods by demonstrating that stability adds the “the missing piece” to the algorithm. As the K-means clustering criterion is only concerned by within-cluster similarity, but not with between-cluster dissimilarity, a model selection criterion based on low density areas would add a valuable aspect to the algorithm. Our recent paper [1] studies this conjecture in depth. On the one hand we show that stability can be upper bounded by certain properties of the optimal clustering, namely by the mass in a small tube around the cluster boundaries. On the other hand, we provide counterexamples which show that a reverse statement is not true in general: stability cannot be lower bounded by the mass around the cluster boundary. Unfortunately, the “missing direction” would have been the important one, as it would have explained the nature of the clusterings selected by the stability criterion. As it is now, we can only provide statements about the clusterings which are *not*

selected by the stability criterion: instable clusterings have undesirable properties in that they cut in high density areas. All in all, this paper provides a negative result again, so that theoretical justifications why clustering stability works successfully in practice remain an important topic for future research. Currently, I am working on an idea how algorithmic effects such as randomization are exploited by clustering stability. Preliminary results show that this might finally lead to a positive result on clustering stability, but this work is not finished yet.



References

1. Ben-David, S., **U. von Luxburg**: Relating clustering stability to properties of cluster boundaries. In Proceedings of the 21st Annual Conference on Learning Theory (COLT 2008), 379–390. (Eds.) Bshouty N., C. Gentile, Springer, Berlin, Germany (2008).
2. Ben-David, S., **U. von Luxburg**, D. Pál: A Sober Look at Clustering Stability. Learning Theory: In Proceedings of the 19th Annual Conference on Learning Theory (COLT 2006), 5–19. (Eds.) Lugosi, G., H.-U. Simon, Springer, Berlin, Germany (2006).
3. Ben-David, S., D. Pál, H.-U. Simon: Stability of k-means Clustering. In Proceedings of the 20th Annual Conference on Learning Theory (COLT 2007), 20–34. (Eds.) Bshouty, N., C. Gentile, Springer, Berlin, Germany (2007).
4. Shamir O., N. Tishby: Model Selection and Stability in k-means Clustering. In Proceedings of the 21st Annual Conference on Learning Theory (COLT 2008). (Eds.) Bshouty, N., C. Gentile, Springer, Berlin, Germany (2008).
5. Shamir O., T. Tishby: Cluster Stability for Finite Samples. In Advances in Neural Information Processing Systems 20: Proceedings of the 2007 Conference. (Eds.) Platt, J. C., D. Koller, Y. Singer, S. Roweis, MIT Press, Cambridge, MA, USA (2008).
6. **von Luxburg, U.**, T. Lange (Organizers): PASCAL Workshop on Stability and Resampling Methods for Clustering. http://www.kyb.tuebingen.mpg.de/bs/people/ule/stability_workshop_07/stability_workshop.html.

¹ David R. Cheriton School of Computer Science, University of Waterloo, Canada

Statistical Analysis of Manifold Methods Based on Random Geometric Graphs

Matthias Hein¹, Jean-Yves Audibert², Ulrike von Luxburg, Markus Maier



It is often observed in practice that real data in a high-dimensional Euclidean space has a rather small number of intrinsic parameters, e.g. a sequence of images of a rotating object. In this case it is reasonable to assume that the data effectively lies on a low-dimensional submanifold. Since this submanifold is unknown a priori, it can be approximated by building a certain neighborhood graph based on the Euclidean distance with the data points as vertices.

Since the last SAB report we continued the study of finite sample bounds as well as of the limiting behavior of several objects defined on such graphs, as the sample size goes to infinity and the neighborhood shrinks to zero. One of these objects is the graph Laplacian, which is used in clustering, semi-supervised learning and dimensionality reduction. Its use is often motivated by properties of the (continuous) Laplacian of the submanifold. In [1] we established the pointwise limit of the so called “random walk” graph Laplacian and could show that it converges to the weighted Laplace-Beltrami operator. Moreover by the use of data-dependent edge weights we showed that one can control the influence of the density of the probability measure generating the data on the limit operator. This is of particular interest in clustering as well as in semi-supervised learning. Recently, we extended these results and provided the limits of all three graph Laplacians used in the literature [6]. The limit operators agree only as long as the probability measure on the manifold is uniform but are quite different as soon as the probability measure is non-uniform which is the regular case in machine learning applications.

Of particular interest in semi-supervised learning is the limit of the regularization functional induced by the graph Laplacian. We showed in [3] uniform convergence over the space of Hölder functions of the regularizer associated to the unnormalized graph Laplacian to the smoothness functional induced by the weighted Laplace-Beltrami operator. Particularly interesting is that using data-dependent weights on the graph one can adjust the influence of the density of the data on the limit smoothness functional. The assumption that the data lies on a submanifold is usually not true since due to noise the data is scattered around it. In high dimensions such noise can distort the distances severely so that points which are originally close on the submanifold can be far away in the ambient Euclidean space. In that case the concept of graph-based methods to build global from local structure breaks down. We tackle this problem by proposing a denoising method for manifold data. The key idea of the denoising algorithm we propose in [4] and [5] is that data on a manifold disturbed by Gaussian noise can be seen as the result of a diffusion process in Euclidean space generated by the Laplacian, where the length of the diffusion is proportional to the variance of the noise. Thus

the basic principle of the denoising method is to reverse this diffusion process. We do this directly on the samples using the graph Laplacian as diffusion operator. The key trick is that the graph is updated in every step and thus the inherent information about the underlying low-dimensional structure is enhanced in every step. The method works well for low-dimensional toy data even when it is disturbed by very high dimensional noise. But even more interesting is that the method used as a pre-processing step for semi-supervised learning can significantly improve the classification results. In [5] we show for the MNIST dataset that in particular for a small number of labeled points the method reduces the error by about 50%.

The dimension of the submanifold is an important qualitative description of the data, since it basically tells us how many free parameters the data generating process possesses. We proposed a new estimator for this quantity with rigorous theoretical underpinnings, see [2], and studied theoretically and empirically how high extrinsic curvature and non-smooth probability measures cause the results to deteriorate. The dimension of low-dimensional submanifolds with high curvature can nevertheless be estimated correctly with relatively small sample sizes.

References

1. **Hein, M., J.-Y. Audibert, U. von Luxburg:** From Graphs to Manifolds – Weak and Strong Pointwise Consistency of Graph Laplacians. In *Proceedings of the 18th Conference on Learning Theory (COLT)*, 470–485. (Eds.) P. Auer, R. Meir, Springer, Berlin, Germany (2005).
2. **Hein, M., J.-Y. Audibert:** Intrinsic Dimensionality Estimation of Submanifolds in Euclidean space. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, 289 – 296. (Eds.), L. De Raedt, S. Wrobel, ACM Press (2005).
3. **Hein, M.:** Uniform Convergence of Adaptive Graph-based Regularization. In *Proceedings of the 19th Annual Conference on Learning Theory (COLT)*, 50–64. (Eds.) G. Lugosi, H. U. Simon, Springer, New York, NY, USA (2006).
4. **Hein, M., M. Maier:** Manifold Denoising. In *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*. (Eds.) Schölkopf, B., J. Platt, T. Hofmann, MIT Press, Cambridge, MA, USA (2007).
5. **Hein, M., M. Maier:** Manifold Denoising as Preprocessing for Finding Natural Representations of Data. In *Proceedings of the 22nd AAAI*, The AAAI Press, Menlo Park, California, USA (2007).
6. **Hein, M., J.-Y. Audibert, U. von Luxburg:** Convergence of Graph Laplacians on Random Neighborhood Graphs. *Journal of Machine Learning Research* 8, 1325–1370 (2007).

¹ Now at: Universität des Saarlandes, Saarbrücken, Germany; ² CERTIS, Ecole Nationale des Ponts et Chaussées, Paris, France

Non-parametric Regression between Riemannian Manifolds

Florian Steinke, Matthias Hein¹, Bernhard Schölkopf

We investigate the problem of non-parametric regression between Riemannian manifolds. That is, unlike manifold learning, where either only the input space is a manifold or the manifold itself is learnt, we assume here that we are given training examples from *known* input and output manifolds and want to learn a prediction map from the input to the output manifold. This problem arises in many application areas ranging from signal processing, computer vision, over robotics to computer graphics.

Learning when input and output domain are Riemannian manifolds is quite distinct from standard multivariate regression. One fundamental problem of using traditional regression methods for manifold-valued regression is that standard techniques use the linear structure of the output space. It thus makes sense to linearly combine simple basis functions, since the addition of function values is still an element of the target space. While this approach still works for manifold-valued input, it is no longer feasible if the output space is a manifold, as general Riemannian manifolds do not allow an addition operation.

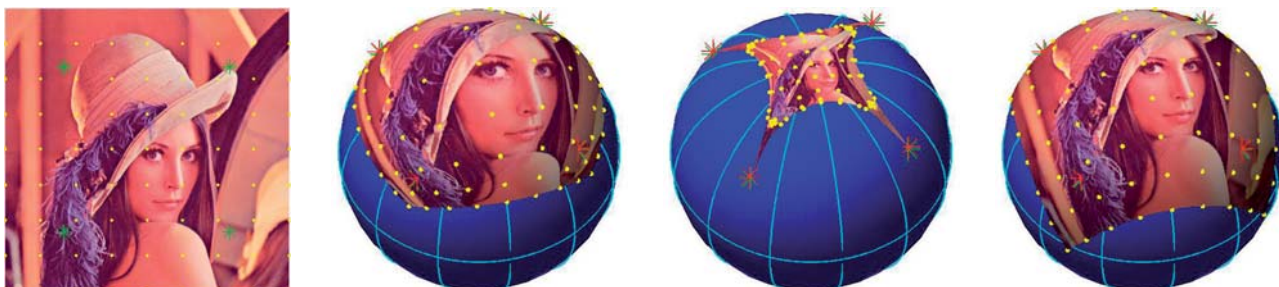
Instead, we focus on a variational formulation, minimizing a regularized empirical risk functional. On the one hand, the objective contains a data fidelity term formulated in terms of the geodesic distances between the predictions and the training examples. On the other hand, it includes a regularisation term that penalizes second derivatives of the prediction map. Note that computing the second derivatives for mappings between Riemannian manifolds is non-trivial since the curvature of the underlying spaces has to be accounted for properly.

Our proposed framework is independent of the parametrization of the manifolds, and depends only on their invariant geometric properties. It can be shown to be the generalisation of thin-plate splines to the case where the input and output spaces are Riemannian manifolds, the minimizers of our regularisation functional being valid generalizations of linear functions [1]. This property may explain the good performance of our method in applications in graphics, vision, and robotics [1, 2].

In the future we plan to investigate how manifold-valued regression is related to structured output learning. While the two problems look very different at first sight, one being a generalization of multi-class classification and the other an extension of regression, both types of learning are related in that they exploit dependencies between the different components of their predictions.

References

1. **Steinke, F., M. Hein:** Non-parametric Regression between Riemannian Manifolds. In *Advances in Neural Information Processing Systems 21: Proceedings of the 2008 Conference* (2009).
2. **Steinke, F., M. Hein, J. Peters, B. Schölkopf:** Manifold-valued Thin-Plate Splines with Applications in Computer Graphics. In *Computer Graphics Forum (Proc. Eurographics 2008) 27(2)*, 437–448 (2008).



a. Original in \mathbb{R}^2

b. Thin-plate energy to \mathbb{R}^3
+ Projection onto \mathbb{S}^2

c. Harmonic energy
to \mathbb{S}^2

d. Proposed Eells energy
to \mathbb{S}^2

Figure 1: The Lena image (a) is used to visualize a prediction map ϕ from the unit square in \mathbb{R}^2 to the unit sphere \mathbb{S}^2 in \mathbb{R}^3 . Green markers show the given training data point pairs, red stars on \mathbb{S}^2 denote positions of the training inputs in \mathbb{R}^2 mapped to the sphere via ϕ .

¹ Now at: Universität des Saarlandes, Saarbrücken, Germany

Neighborhood Graphs in Clustering

Markus Maier, Ulrike von Luxburg, Matthias Hein¹



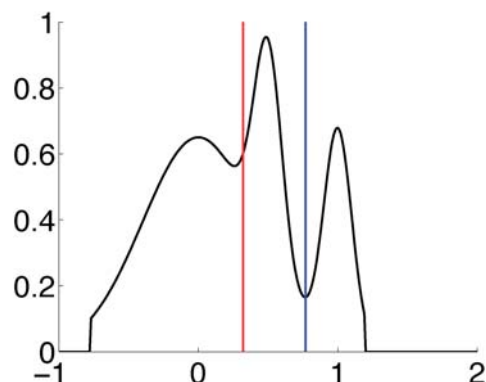
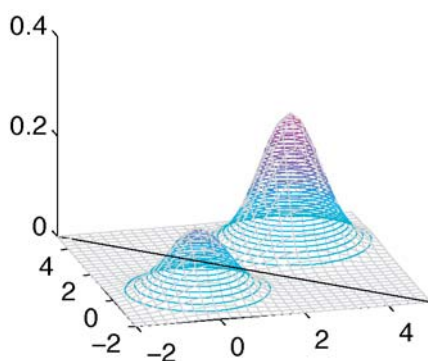
Graph clustering methods such as spectral clustering are defined for general weighted graphs. In machine learning, however, data often is not given in form of a graph, but in terms of similarity (or distance) values between points. In this case, first a neighborhood graph is constructed using the similarities between the points and then a graph clustering algorithm is applied to this graph. There are various types of neighborhood graphs, for example the r -neighborhood graph or different kinds of k -nearest neighbor graphs, and for each type of graph there are parameters that have to be chosen. The goal of this project is to investigate the influence of the construction of the similarity graph on the clustering results from a theoretical point of view.

In a first project we studied the neighborhood graph construction in a simplified setting, which is called “cluster identification”. We assumed that the true clusters are given as the level sets of the density our points are sampled from. A cluster is said to be identified in a neighborhood graph if its points are exactly one connected component of the graph. In [1] we compared two different kinds of k -nearest neighbor graphs, the mutual and the symmetric k -nearest neighbor graph, with respect to cluster identification. We found that they behave differently if our goal is to detect only the most significant clusters. In particular, the mutual k -nearest neighbor graph is better suited if one is interested in finding the “most significant clusters” only, whereas our results show no difference between the two graph types when one is interested in identifying all clusters. Furthermore, we gave the range of the parameter k for which clusters are identified in the graphs and derived the optimal choice of k . The surprising result was that to maximize the probability of cluster identification k has to be chosen in the order linear with n (and not in the order of $\log n$). In [2] we investigated the problem of cluster identification in a setting with considerably weaker assumptions. After a slight modification of the construction of the k -nearest neighbor graph we could show similar results. In future work we are going to compare our findings to algorithms in data mining that work according to the same principles.

In [3] we studied the convergence of graph clustering criteria such as the normalized cut (Ncut), which is used in spectral clustering, as the sample size tends to infinity. We assumed that the sample points are drawn from some density and fixed a hypersurface that separates the density into two parts. Then we studied the limit of graph clustering criteria for the cuts induced by a hyperplane for different types of unweighted neighborhood graphs and depending on the choice of parameters (see left Figure). We found that the limit expressions are different for different types of graphs. This means, for example, that Ncut on a k -nearest neighbor graph does something systematically different than Ncut on an r -neighborhood graph (see right Figure)! These differences could also be experimentally observed for toy and real data sets already for rather small sample sizes. This finding shows that graph clustering criteria cannot be studied independently of the kind of graph they will be applied to. In future work we are going to consider weighted neighborhood graphs which are closer to the graphs used in practice.

References

1. Maier, M., M. Hein, U. von Luxburg: Cluster Identification in Nearest-neighbor Graphs. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory (ALT 2007)*, 196–210. (Eds.) Hutter, M., R. A. Servedio, E. Takimoto, Springer, Berlin, Germany (2007).
2. Maier, M., M. Hein, U. von Luxburg: Optimal Construction of k -nearest Neighbor Graphs for Identifying Noisy Clusters. *Theoretical Computer Science*, (accepted).
3. Maier, M., U. von Luxburg, M. Hein: Influence of Graph Construction on Graph-based Clustering Measures. In *Advances in Neural Information Processing Systems 21: Proceedings of the 2008 Conference*, (2009).



Left: We study the limit of graph clustering criteria for neighborhood graphs whose vertices are drawn from a density and for a cut that is induced by a hyperplane. Right: The place of the optimal cuts are different for the k -nearest neighbor graph (red) and the r -neighborhood graph (blue).

¹ Now at: Universität des Saarlandes, Saarbrücken, Germany