Minimax-optimal semi-supervised regression on unknown manifolds

Amit Moscovich, Ariel Jaffe, Boaz Nadler, Weizmann Institute of Science, Israel. amit@moscovich.org; ariel.jaffe@weizmann.ac.il; boaz.nadler@weizmann.ac.il

Abstract

We present a simple semi-supervised regression method for data from an unknown manifold. We prove that given a large *unlabeled* training set, it achieves the optimal finite-sample bound on the MSE, as if the manifold geometry were known. Furthermore, it demonstrates good empirical performance on manifold-structured data sets and can be implemented efficiently.

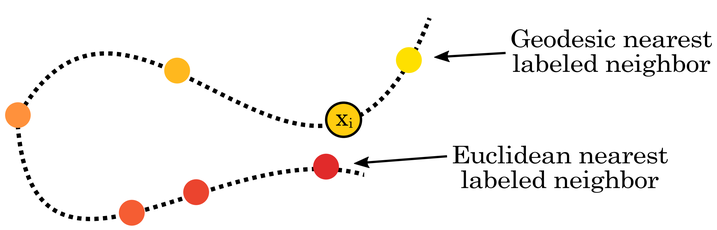
Introduction

Input: labeled set of n pairs $\mathcal{L} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ and an unlabeled set of m points $\mathcal{U} = \{\mathbf{x}_{n+1}, \ldots \mathbf{x}_{n+m}\}$. The points $\mathbf{x}_i$ are assumed to be sampled i.i.d. from some measure μ over $\mathbb{R}^D$ and $y_i = f(\mathbf{x}_i) + \text{noise}$.

Output (transductive): estimates $\hat{f}(\mathbf{x}_{n+1}), \ldots, \hat{f}(\mathbf{x}_{n+m})$
Output (inductive): regression *estimator* $\hat{f} : \mathbb{R}^D \to \mathbb{R}$.

We make two assumptions on the data:
(i) $\mathbf{x}_1, \ldots, \mathbf{x}_{n+m}$ lie on a d-dimensional manifold $\mathcal{M} \subset \mathbb{R}^{\mathbb{D}}$.
(ii) The regression function f is Lipschitz in the manifold geodesic distances $|f(\mathbf{x}_i) - f(\mathbf{x}_j)| \leq L d_{\mathcal{M}}(\mathbf{x}_i, \mathbf{x}_j)$.

Under these assumptions Kpotufe (2011) showed that as $n \to \infty$, standard knn regression achieves the minimax bound on the MSE $n^{-\frac{2}{2+d}}$ up to log factors. We show that it is possible to obtain the *finite-sample* minimax bound using a variant of knn regression which is based on estimates of manifold geodesic distances.



Geodesic knn regression

Step 1: Connect every pair of close points by an edge whose weight is their Euclidean distance, e.g.

$$w(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} \|\mathbf{x}_i - \mathbf{x}_j\| & \text{if } \|\mathbf{x}_i - \mathbf{x}_j\| < r \\ \infty & \text{otherwise.} \end{cases}$$

Step 2: Compute d_G, the graph shortest-path distances for every pair $\{(\mathbf{x}_i, \mathbf{x}_j) : \mathbf{x}_i \in \mathcal{L}, \mathbf{x}_j \in \mathcal{L} \cup \mathcal{U}\}$. The distances d_G uniformly approximate the manifold geodesic distances (Tenenbaum et al., 2000).

Step 3: Let $\text{knn}_G(\mathbf{x}_i) \subseteq \mathcal{L}$ denote the set of k nearest *labeled* neighbors to $\mathbf{x}_i$, as determined by d_G. Then for every $\mathbf{x}_i \in \mathcal{L} \cup \mathcal{U}$ the *geodesic knn regression* estimate is

$$\hat{f}(\mathbf{x}_i) := \frac{1}{|\text{knn}_G(\mathbf{x}_i)|} \sum_{(\mathbf{x}_j, y_j) \in \text{knn}_G(\mathbf{x}_i)} y_j. \qquad (1)$$

Inductive output: For a new instance $\mathbf{x} \notin \mathcal{L} \cup \mathcal{U}$ we first find its *Euclidean* nearest neighbor $\mathbf{x}^*$ from $\mathcal{L} \cup \mathcal{U}$. Then the geodesic knn regression estimate at $\mathbf{x}$ is

$$\hat{f}(\mathbf{x}) := \hat{f}(\mathbf{x}^*) = \hat{f}\left(\underset{\mathbf{x}' \in \mathcal{L} \cup \mathcal{U}}{\text{argmin}} \|\mathbf{x} - \mathbf{x}'\|\right).$$

Main result

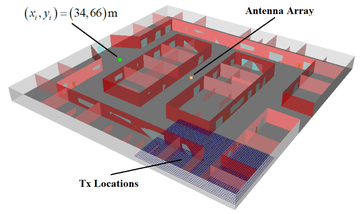
We assume that the manifold $\mathcal{M}$ satisfies several conditions and that the sampling measure μ satisfies, for every $\mathbf{x} \in \mathcal{M}$ and radius $r \leq R$ that $\mu(B_{\mathbf{x}}(r)) \geq Q r^d$.

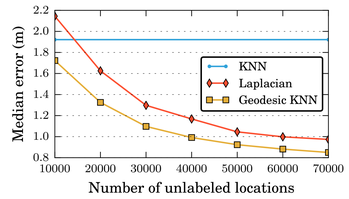
Theorem Let $\mathbf{x} \in \mathcal{M}$ be a point and let $f_D := f_{\max} - f_{\min}$. The MSE of the geodesic kNN regressor at x satisfies

$$\mathrm{E}(\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2 \leq c n^{-\frac{2}{2+d}} + c' e^{-c'' \cdot (n+m)} f_D^2. \qquad (2)$$

Conclusion: The estimator $\hat{f}$ obtains the finite-sample minimax bound on the mean squared error.

Application: indoor localization using WiFi fingerprints





Efficient computation

For the transductive regression estimates (1) we need to compute $\text{knn}_G(\mathbf{x}_i)$ for all $x_i \in \mathcal{L} \cup \mathcal{U}$. This can be done using all-pairs shortest-path algorithms or using Dijkstra's algorithm from each labeled point.

We describe a novel variant of Dijkstra's algorithm which simultaneously explores shortest paths from all labeled vertices.

Algorithm	Dense graph	Sparse graph
Floyd-Warshall	$O(N^3)$	$O(N^3)$
$n\times$Dijkstra	$O(nN^2)$	$O(nN \log N)$
Simultaneous Dijkstra	$O(kN^2)$	$O(kN \log N)$

The following table compares the empirical running time of geodesic knn to that of Belkin and Niyogi (2004).

N	Laplacian eigenbasis	Geodesic 7NN
1000	7.6 sec	2.3 sec
10000	195 sec	7 sec
100000	114 min	56 sec

References

Belkin, M. and Niyogi, P. (2004). Semi-Supervised Learning on Riemannian Manifolds. *Machine Learning*.

Kpotufe, S. (2011). k-NN regression adapts to local intrinsic dimension. In *NIPS*.

Tenenbaum, J. B., De Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*.