Understanding your Neighbors: practical perspectives from modern analysis

Sanjoy Dasgupta, and Samory Kpotufe

Topics: k-NN prediction and similar methods, structured data and intrinsic dimension, efficient search, choice of metric, deep representation of data, modern uses and practical tradeoffs.

Abstract. Tutorial to be given at ICML 2018.

1 Overview

Description. Nearest-neighbor methods are among the most ubiquitous and oldest approaches in Machine Learning and other areas of data analysis. They are often used directly as predictive tools, or indirectly as integral parts of more sophisticated modern approaches (e.g. recent uses that exploit deep representations, uses in geometric graphs for clustering, integrations into time-series classification, or uses in ensemble methods for matrix completion). Furthermore, they have strong connections to other tools such as classification and regression trees, or even kernel machines, which are all (more sophisticated) forms of *local* prediction. Interestingly, our understanding of these methods is still evolving, with many recent results shedding new insights on performance under various settings describing the range of modern uses and application domains [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. Our aim is to cover such new perspectives on k-NN, and in particular, translate new theoretical insights (with potential practical implications) to a broader audience.

A tentative schedule of topics is outlined in Section 2 below. The presentation will consist of a mix of theoretical and empirical insights that addresses modern uses of k-NN and also aim to yield insights into related nonparametric methods.

Objectives. A first objective is to share new perspectives into proper practical uses of k-NN. We aim to address questions such as when (or on what type of data) might we expect good performance from k-NN methods; how to properly configure the method and which tradeoffs to expect (under choices of k, of metric, subsampling or other techniques to improve time and space efficiency). While much of such insights are present in the ML literature, and related areas, they are often of a technical nature that can make it inaccessible to the wide audience of practitioners. Our first objective is therefore to translate these insights to a wider audience.

Our second objective is to shed some light on the recent realization that *overparametrized procedures might still generalize properly*. Nearest neighbor methods, as quintessential *nonparametric* approaches, are very much overparametrized, yet they can be shown to *generalize* well, despite the fact that the usual insights from ML theory do not apply; therefore, modern insights on generalization of k-NN are interesting beyond the scope of this particular method.

Target audience. While much of the presentation will be rooted in recent theoretical insights, we will assume a non-theory audience, yet with the expected ML background in calculus (or basic analysis) and basic probability or statistics. However, we will expect that the more theoretical audience will also benefit from high-level overviews of new theoretical insights into generalization and practical tradeoffs.

Presenters. We plan to split the presentation, with the introduction and basic technical insights (Part 1) being covered by Samory Kpotufe, while more recent advances and modern uses (Part 2) will be covered by Sanjoy Dasgupta. Short bios are given below.

Sanjoy Dasgupta. (email: `dasgupta@cs.ucsd.edu`)
Sanjoy Dasgupta is a Professor in the Department of Computer Science and Engineering at UC San Diego. His area of research is algorithmic statistics, with a focus on interactive learning. He is the author of a textbook,

”Algorithms” (with Christos Papadimitriou and Umesh Vazirani), that appeared in 2006. He was program co-chair for COLT in 2009 and for ICML in 2013. Of relevance to this tutorial, Sanjoy Dasgupta has given a number of tutorials at various leading venues (ICML 2009; Machine Learning Summer School (MLSS), Chicago 2005, 2009; Series of three lectures at Institut Henri Poincare, Paris, May 2011; Microsoft Research Summer School for Machine Learning, Bangalore, June 2015; Symposium on Computational Geometry and Topology in the Sciences, College de France, June 2017).

Samory Kpotufe. (email: `samory@princeton.edu`)
Samory Kpotufe is an Assistant Professor at ORFE, Princeton University, and works at the intersection of Machine Leaning and Nonparametric Statistics. In particular, his work on local predictive methods (k-NN, tree-based regression and classification) has won honors at leading Machine Learning venues (best student paper at COLT, and plenary presentations at NIPS, and AISTATS). Of relevance to this tutorial, Samory was an invited lecturer at the Machine Learning Summer School (MLSS), Cadiz 2016, where he covered topics on modern Nonparametrics.

2 Detailed outline of topics *(tentative)*

PART I: Basic Insights for Structured Data

- Universality and Related Methods.
  (basic relations to trees, kernel machines, gaussian processes)
- Behavior of k-NN Distances (a denominator across modern analyses).
  (bias-variance, manifold, sparsity, choice of k, choice of metric)
- Why Classification is easier than Regression.
  (key insights about noise margin)
- Improving Performance, and some Tradeoffs.
  (Weighted NN methods and relations to Subsampling and Bagging)

PART II: Refined Analysis and Recent Uses

- Recent Formalisms that better Capture Modern Data.
  (volume-based smoothness, mixed costs regimes)
- Active and Semi-supervised learning using NNs.
  (key strategies to reduce labeling costs)
- Exploiting Deep Representations
  (some modern uses, and effects of better representation)
- Modern methods for Efficient NN search.
  (exact or approximate k-NN search, and statistical tradeoffs)

References

[1] R. J. Samworth *et al.*, “Optimal weighted nearest neighbour classifiers,” *The Annals of Statistics*, vol. 40, no. 5, pp. 2733–2763, 2012.

[2] G. Biau and L. Devroye, “On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification,” *Journal of Multivariate Analysis*, vol. 101, no. 10, pp. 2499–2518, 2010.

[3] R. Urner, S. Ben-David, and S. Shalev-Shwartz, “Access to unlabeled data can speed up prediction time,” 2011.

[4] S. Gadat, T. Klein, C. Marteau, *et al.*, “Classification in general finite dimensional spaces with the k-nearest neighbor rule,” *The Annals of Statistics*, vol. 44, no. 3, pp. 982–1009, 2016.

[5] G. H. Chen, S. Nikolov, and D. Shah, "A latent source model for nonparametric time series classification," in *Advances in Neural Information Processing Systems*, pp. 1088–1096, 2013.

[6] J. A. Costa and A. O. Hero, "Manifold learning using euclidean k-nearest neighbor graphs [image processing examples]," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 3, pp. iii–988, IEEE, 2004.

[7] C. Lee, Y. Li, D. Shah, and D. Song, "Blind regression via nearest neighbors under latent variable models," *arXiv preprint arXiv:1705.04867*, 2017.

[8] S. Singh and B. Póczos, "Analysis of k-nearest neighbor distances with application to entropy estimation," *arXiv preprint arXiv:1603.08578*, 2016.

[9] C. Berlind and R. Urner, "Active nearest neighbors in changing environments," in *International Conference on Machine Learning*, pp. 1870–1879, 2015.

[10] A. Kontorovich and R. Weiss, "A bayes consistent 1-nn classifier," in *Artificial Intelligence and Statistics*, pp. 480–488, 2015.

[11] L. Györfi, M. Döring, and H. Walk, "Exact rate of convergence of k-nearest-neighbor classification rule," 2017.

[12] S. Kpotufe, "k-NN Regression Adapts to Local Intrinsic Dimension," *NIPS*, 2011.

[13] K. Chaudhuri and S. Dasgupta, "Rates of convergence for nearest neighbor classification," in *Advances in Neural Information Processing Systems*, pp. 3437–3445, 2014.