Pruning Nearest Neighbor Cluster Trees

Samory Kpotufe SAMORY@TUEBINGEN.MPG.DE
Ulrike von Luxburg ULRIKE.LUXBURG@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems, Tuebingen, Germany

Abstract

Nearest neighbor (k-NN) graphs are widely used in machine learning and data mining applications, and our aim is to better understand what they reveal about the cluster structure of the unknown underlying distribution of points. Moreover, is it possible to identify spurious structures that might arise due to sampling variability?

Our first contribution is a statistical analysis that reveals how certain subgraphs of a k-NN graph form a consistent estimator of the cluster tree of the underlying distribution of points. Our second and perhaps most important contribution is the following finite sample guarantee. We carefully work out the tradeoff between aggressive and conservative pruning and are able to guarantee the removal of all spurious cluster structures at all levels of the tree while at the same time guaranteeing the recovery of salient clusters. This is the first such finite sample result in the context of clustering.

1. Introduction

In this work, we consider the nearest neighbor (k-NN) graph where each sample point is linked to its nearest neighbors. These graphs are widely used in machine learning and data mining applications, and interestingly there is still much to understand about their expressiveness. In particular we would like to better understand what such a graph on a finite sample of points might reveal about the cluster structure of the underlying distribution of points. More importantly we are interested in whether one can identify spurious structures that are artifacts of sampling variability, i.e. spurious structures that are not representative of the true cluster structure of the distribution.

Our first contribution is in exposing more of the richness of k-NN graphs. Let G_n be a k-NN graph over an n-sample from a distribution $\mathcal{F}$ with density f. Previous work (Maier et al., 2009) has shown that the connected components (CC) of a given level set of f can be approximated by the CCs of some subgraph of G_n, provided the level set satisfies certain boundary conditions. However it remained unclear whether or when all level sets of f might satisfy these conditions, in other words, whether the CCs of any level set can be recovered. We show under mild assumptions on f that CCs of any level set can be recovered by subgraphs of G_n for n sufficiently large. Interestingly, these subgraphs are obtained in a rather simple way: just remove points from the graph in decreasing order of their k-NN radius (distance to the k'th nearest neighbor), and we obtain a nested hierarchy of subgraphs which approximates the *cluster tree* of $\mathcal{F}$, i.e. the nested hierarchy formed by the level sets of f (see Figure 1, also Section 2.1).

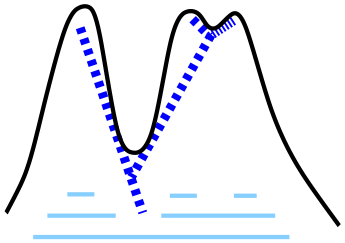


Figure 1. A density f (black line) and its cluster tree (dashed). The CCs of 3 level sets are shown in lighter color at the bottom.

Our second, and perhaps more important contribution is in providing the first concrete approach in the context of clustering that guarantees the pruning of all spurious cluster structures at any tree level. We carefully work out the tradeoff between pruning "aggressively" (and potentially removing important clusters) and pruning "conservatively" (with the risk of keeping spurious clusters) and derive tuning settings that require no knowledge of the underlying distribution beyond an upper bound on f. We can thus guarantee in a finite sample setting that (a) all clusters remaining at any level of the pruned tree correspond to CCs of some level set of f, i.e. all spurious clusters are pruned away, and (b) salient clusters are still discovered, where the

Appearing in *Proceedings of the 28^{th} International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

degree of *saliency* depends on the sample size n. We can show furthermore that the pruned tree remains a consistent estimator of the underlying cluster tree, i.e. the CCs of any level set of f are recovered for sufficiently large n. Interestingly, the pruning procedure is not tied to the k-NN method, but is based on a simple intuition that can be applied to other cluster tree methods (see Section 3).

Our results rely on a central "connectedness" lemma (Section 5.2) that identifies which CCs of f remain connected in the empirical tree. This is done by analizing the way in which k-NN radii vary along a path in a dense region.

1.1. Related work

Recovering the cluster tree of the underlying density is a clean formalism of hierarchical clustering proposed in 1981 by J. A. Hartigan (Hartigan, 1981). Hartigan showed in the same seminal paper that the single-linkage algorithm is a consistent estimator of the cluster tree for densities on $\mathbb{R}$. For $\mathbb{R}^d, d > 1$ it is known that the empirical cluster tree of a consistent density estimate is a consistent estimator of the underlying cluster tree (see e.g. (Wong & Lane, 1983)), unfortunately there is no known algorithm for computing this empirical tree. Nonetheless, the idea has led to the development of interesting heuristics based on first estimating density, then approximating the cluster tree of the density estimate in high dimension (Wong & Lane, 1983; Stueltze & Nugent, 2010).

Many other related work such as (Rigollet & Vert, 2009; Singh et al., 2009; Maier et al., 2009; Rinaldo & Wasserman, 2010) consider the task of recovering the CCs of a single level set, the closest to the present work being (Maier et al., 2009) which uses a k-NN graph for level set estimation. As previously discussed, level set estimation however never led to a consistent estimator of the cluster tree, since these results typically impose technical requirements on the level set being recovered but do not work out how or when these requirements might be satisfied by all level sets of a distribution.

A recent insightful paper of Chaudhuri & Dasgupta (2010) presents the first provably consistent algorithm for estimating the cluster tree. At each level of the empirical cluster tree, they retain only those samples whose k-NN radii are below a scale parameter r which indexes the level; CCs at this level are then discovered by building an r-neighborhood graph on the retained samples. This is similar to an earlier generalization of single-linkage by D. Wishart (Wishart, 1969) which however was given without a convergence analysis. The k-NN tree studied here differs in that, at an equivalent level r, points are connected to the subset of their k-nearest neighbors retained at that level. One practical appeal of our method is its simplicity: we need only remove points from an initial k-NN graph to obtain the various levels of the empirical cluster tree.

(Chaudhuri & Dasgupta, 2010) provides finite sample results for a particular setting of $k \approx \log n$. In contrast our finite sample results are given for a wide range of values of k, namely for $\log n \lesssim k \lesssim n^{1/O(d)}$. In both cases the finite sample results establish natural separation conditions under which the CCs of level sets are recovered (see Theorem 1). The result of (Chaudhuri & Dasgupta, 2010) however allows the possibility that some empirical clusters are just artifacts of sampling variability. We provide a simple pruning procedure that ensures that clusters discovered empirically at any level correspond to true clusters at some level or the underlying cluster tree. Note that this can be trivially guaranteed by returning a single cluster at all levels, so we additionally guarantee that the algorithm discovers salient modes of the density, where the saliency depends on empirical quantities (see Theorem 2).

A recent archived paper (Rinaldo et al., 2010) also treats the problem of false clusters in cluster tree estimation, but the result is not algorithmic as they only consider the cluster tree of an empirical density estimate, and do not provide a way to compute this cluster tree.

There exist many pruning heuristics in the literature which typically consist of removing *small* clusters (Maier et al., 2009; Stueltze & Nugent, 2010) using some form of thresholding. The difficulty with these approaches is in how to define *small* without making strong assumptions on the unknown underlying distribution, or on the tree level being pruned (levels correspond to different resolutions or cluster sizes). Moreover, even the assumption that spurious clusters must be small does not necessarily hold. Consider for example a cluster made up of two large regions connected by a thin bridge of low mass; the two large regions can easily appear as two separate clusters in a finite sample. Some more sophisticated methods such as (Stueltze & Nugent, 2009) do not rely on cluster size for pruning, instead they return confidence values for the empirical clusters based on various notions of cluster stability; unfortunately they do not provide finite sample guarantees. Our pruning guarantees the removal of all spurious clusters, large and small (see Figure 2); we make no assumption on the shape of clusters beyond a smoothness assumption on the density; we provide a simple tuning parameter whose setting requires just an upper bound on the density.

2. Preliminaries

Assume the finite dataset $\mathbf{X} = \{X_i\}_{i=1}^n$ is drawn i.i.d. from a distribution $\mathcal{F}$ over $\mathbb{R}^d$ with density function f.

We start with some simple definitions related to k-NN operations. All balls, unless otherwise specified, denote closed balls in $\mathbb{R}^d$.

Definition 1 (k-NN radii)**.** *For $x \in \mathcal{X}$, let $r_{k,n}(x)$ denote the radius of the smallest ball centered at x containing k points from $\mathbf{X} \setminus \{x\}$. Also, let $r_k(x)$ denote the radius of the smallest ball centered at x of $\mathcal{F}$-mass k/n.*

Definition 2 (k-NN and mutual k-NN graphs)**.** *The k-NN graph is that whose vertices are the points in $\mathbf{X}$, and where X_i is connected to X_j iff $X_i \in B(X_j, \theta r_k(X_j))$ or $X_j \in B(X_i, \theta r_k(X_i))$ for some $\theta > 0$. The mutual k-NN graph is that where X_i is connected to X_j iff $X_i \in B(X_j, \theta r_k(X_j))$ and $X_j \in B(X_i, \theta r_k(X_i))$.*

2.1. Cluster tree

Definition 3 (Connectedness)**.** *We say $A \subset \mathbb{R}^d$ is connected if for every $x, x' \in A$ there exists a continuous $1-1$ function $P : [0,1] \mapsto A$ where $P(0) = x$ and $P(1) = x'$. P is called a path in A between x and x'.*

The cluster tree of f will be denoted $\{G(\lambda)\}_{\lambda>0}$, where $G(\lambda)$ are the CCs of the level set $\{x : f(x) \geq \lambda\}$. Notice that $\{G(\lambda)\}_{\lambda>0}$ forms a (infinite) tree hierarchy where for any two components A, A', either $A \cap A' = \emptyset$ or one is a descendant of the other, i.e $A \subset A'$ or $A' \subset A$.

3. Algorithm

Definition 4 (k-NN density estimate)**.** *Define the density estimate at $x \in \mathbb{R}^d$ as :*

$$f_n(x) \doteq \frac{k}{n \cdot \mathrm{vol}\left(B(x, r_{k,n}(x))\right)} = \frac{k}{n \cdot v_d r_{k,n}^d(x)},$$

where v_d is the volume of the unit ball in $\mathbb{R}^d$.

Let G_n be the k-NN or mutual k-NN graph. For $\lambda > 0$ define $G_n(\lambda)$ as the subgraph of G_n containing only vertices in $\{X_i : f_n(X_i) \geq \lambda\}$ and corresponding edges. The CCs of $\{G_n(\lambda)\}_{\lambda>0}$ form a tree: let A_n and A'_n be two such CCs, either $A_n \cap A'_n = \emptyset$ or one is a descendant of the other, i.e. A_n is a subgraph of A'_n or vice versa. To simplify notation, we let the set $\{G_n(\lambda)\}_{\lambda>0}$ denote the empirical cluster tree before pruning.

Pruning

The pruning procedure (Algorithm 1) consists of simple lookups: it reconnects CCs at level λ if they are part of the same CC at level $\lambda - \tilde{\epsilon}$ where the tuning parameter $\tilde{\epsilon} \geq 0$ controls how aggressively we prune. We show its behavior on a finite sample in Figure 2.

The intuition behind the procedure is the following. Suppose $A_n, A'_n \subset \mathbf{X}$ are disconnected at some level λ in the empirical tree before pruning. However, they ought to be connected, i.e. their vertices belong to the same CC A at the highest level where they are all contained in the underlying cluster tree. Then, key sample points from A that would have kept them connected are missing at level λ in the empirical tree. These key points have f_n values lower than λ, but probably not much lower. By looking down to a lower level near λ we find that A_n, A'_n are connected and thus detect the situation. Notice that this intuition is not tied to the k-NN cluster tree but can be applied to any other cluster tree procedure. All that is required is that all points from A (as discussed above) be connected at some level in the tree close to λ.

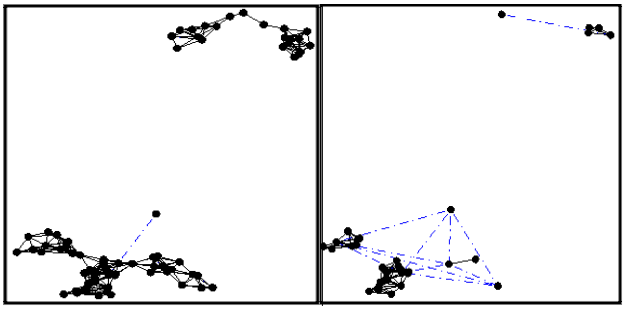


Figure 2. Pruning at work: it reconnects CCs independent of size. The dashed lines are reconnection edges from pruning. Shown are two levels of the k-NN tree of a 500-sample from the 2-modes mixture $0.5\mathcal{N}([0,0], I_2) + 0.5\mathcal{N}([1,4], I_2)$. Here $k = 12$, $\theta = 1$, $\tilde{\epsilon} = F/\sqrt{k}$ where $F = 2.73$ is the maximum f_n value. From left to right, level $\lambda = 0.9$ has 72 points, and level $\lambda = 1.3$ has 33.

Algorithm 1 Prune $G_n(\lambda)$

Given: tuning parameter $\tilde{\epsilon} \geq 0$, same for all levels.
$\widetilde{G}_n(\lambda) \leftarrow G_n(\lambda)$.
if $\lambda > \tilde{\epsilon}$ **then**
  Connect components A_n, A'_n of $\widetilde{G}_n(\lambda)$ if they are part of the same component of $G_n(\lambda - \tilde{\epsilon})$.
else
  Connect all $\widetilde{G}_n(\lambda)$.
end if

It is not hard to see that the CCs of the pruned subgraphs $\left\{\widetilde{G}_n(\lambda)\right\}_{\lambda>0}$ still form a tree. We will hence denote the pruned empirical tree by $\left\{\widetilde{G}_n(\lambda)\right\}_{\lambda>0}$.

4. Results Overview

We make the following assumptions on the density f.

(A.1) $\exists F > 0, \sup_{x \in \mathbb{R}^d} f(x) \leq F$.

(A.2) f is Hoelder-continuous, i.e. there exists $L, \alpha > 0$ such that for all $x, x' \in \mathbb{R}^d$,

$$|f(x) - f(x')| \leq L \left\|x - x'\right\|^{\alpha}.$$

Theorem 1 below is a finite sample result that establishes conditions under which samples from a connected subset

of $\mathbb{R}^d$ remain connected in the empirical cluster tree, and samples from two disconnected subsets of $\mathbb{R}^d$ remain disconnected even after pruning. Essentially, for k sufficiently large, points from connected subsets A remain connected below some level. Also, provided k is not too large, disjoint subsets A and A' which are separated by a large enough region of low density (relative to n, k and $\tilde{\epsilon}$), remain disconnected above some level.

We require the following two definitions.

Definition 5 (Envelope of $A \subset \mathbb{R}^d$)**.** *Let $A \subset \mathbb{R}^d$ and for $r > 0$, define: $A_{+r} \doteq \{y : \exists x \in A, y \in B(x, r)\}$.*

Definition 6 ((ϵ, r)-separated sets)**.** *$A, A' \subset \mathbb{R}^d$ are (ϵ, r)-separated if there exists a separating set S such that every path in $\mathbb{R}^d$ between A and A' intersects S, and*

$$\sup_{x \in S_{+r}} f(x) \leq \inf_{x \in A \cup A'} f(x) - \epsilon.$$

Theorem 1. *Suppose f satisfies (A.1) and (A.2). Let G_n be the k-NN or mutual k-NN graph. Let $\delta > 0$ and define $\epsilon_k \doteq 11F\sqrt{\ln(2n/\delta)/k}$. There exist C and $C' = C'(\mathcal{F})$ such that, for*

$$C\left(\max\left\{1, \sqrt{2}/\theta\right\}\right)^d d\ln(n/\delta) \leq k \leq C'\left(F\sqrt{\ln(n/\delta)}\right)^{2(\alpha+d)/(3\alpha+d)} n^{2\alpha/(3\alpha+d)} \quad (1)$$

the following holds with probability at least $1 - 3\delta$ simultaneously for subsets A of $\mathbb{R}^d$.

(a) Let A be a connected subset of $\mathbb{R}^d$, and let $\lambda \doteq \inf_{x \in A} f(x) > 2\epsilon_k$. All points in $A \cap \mathbf{X}$ belong to the same CC of $\widetilde{G}_n(\lambda - 2\epsilon_k)$.

(b) Let A and A' be two disjoints subsets of $\mathbb{R}^d$, and define $\lambda = \inf_{x \in A \cup A'} f(x)$. Recall that $\tilde{\epsilon} \geq 0$ is the tuning parameter. Suppose A and A' are (ϵ, r)-separated for $\epsilon = 6\epsilon_k + 2\tilde{\epsilon}$ and $r = \frac{\theta}{2}\left(4k/v_d n\lambda\right)^{1/d}$. Then $A \cap \mathbf{X}$ and $A' \cap \mathbf{X}$ are disconnected in $\widetilde{G}_n(\lambda - 2\epsilon_k)$.

Theorem 1 above, although written in terms of $\widetilde{G}_n$, applies also to G_n by just setting $\tilde{\epsilon} = 0$. The theorem implies consistency of both pruned and unpruned k-NN trees under mild additional conditions. Some such conditions are illustrated in the corollary below. A nice practical aspect of the pruning procedure is that consistency is obtained for a wide range of settings of $\tilde{\epsilon}$ and k as functions of n.

Corollary 1 (Consistency)**.** *Suppose that f satisfies (A.1) and (A.2) and that, in addition, $\mathcal{F}$ is supported on a compact set, and for any $\lambda > 0$, there are finitely many components in $G(\lambda)$. Assume that, as $n \to \infty$, $\tilde{\epsilon} = \tilde{\epsilon}(n) \to 0$ and $k/\log n \to 0$ while $k = k(n)$ satisfies (1). For any $A \subset \mathbb{R}^d$, let A_n denote the smallest component of $\left\{\widetilde{G}_n(\lambda)\right\}_{\lambda > 0}$ containing $A \cap \mathbf{X}$. Fix $\lambda > 0$. We have $\lim_{n \to \infty} \mathbb{P}\left(\forall A, A' \in G(\lambda),\ A_n \text{ is disjoint from } A'_n\right) = 1$.*

Proof. Let A and A' be separate components of $G(\lambda)$. The assumptions ensure that all paths between A and A' traverse a compact set S satisfying $\lambda - \max_{x \in S} f(x) \doteq \epsilon_S > 0$ (see Lemma 14 of (Chaudhuri & Dasgupta, 2010)). Let $\epsilon = 6\epsilon_k + 2\tilde{\epsilon}$ and $r = \frac{\theta}{2}\left(4k/v_d n\lambda\right)^{1/d}$. By uniform continuity of f, there exists N_1 such that for $n > N_1$, r is small enough so that $\lambda - \max_{x \in S_{+r}} f(x) > \epsilon_S/2$. Also, there exists $N_2 > N_1$ such that for $n > N_2$, $\epsilon < \epsilon_S/2$, in other words $\sup_{x \in S_{+r}} f(x) \leq \lambda - \epsilon$.

Since $G_n(\lambda)$ is finite, there exists N such that for $n > N$, all pairs A, A' have a suitable (ϵ, r)-separating set S. Thus by Theorem 1, for $n > N$, with probability at least $1 - 3\delta$, $\forall A, A' \in G(\lambda)$, $A \cap \mathbf{X}$ and $A' \cap \mathbf{X}$ are fully contained in $\widetilde{G}_n(\lambda - 2\epsilon_k)$ and are disjoint. They are thus disjoint at any higher level, so A_n and A'_n are also disjoint.

The above holds for all $\delta > 0$, so the statement follows. □

While Theorem 1 establishes that a connected set A remains connected below some level, it does not guarantee against parts of A becoming disconnected at higher levels, creating spurious clusters. Note that the removal of spurious clusters can be trivially guaranteed by just letting the parameter $\tilde{\epsilon}$ very large, but the ability of the algorithm to discover true clusters is necessarily affected. We are interested in how to set $\tilde{\epsilon}$ in order to guarantee the removal of spurious clusters while still recovering important ones.

Theorem 2 guarantees that, by setting $\tilde{\epsilon}$ as $\Omega(\epsilon_k)$ (recall ϵ_k from Theorem 1), separate CCs of the empirical cluster tree correspond to actual clusters of the (unknown) underlying distribution, i.e. all spurious clusters are removed. The setting of $\tilde{\epsilon}$ only requires an upper-bound F on the density f [1]. Note that, under such a setting, consistency is maintained per Corollary 1, and in light of Theorem 1 (b), we can expect that interesting clusters are discovered. In particular the following salient modes of f are discovered.

Definition 7 ((ϵ, r)-salient mode)**.** *An (ϵ, r)-salient mode is a leaf node A of the cluster tree $\{G(\lambda)\}_{\lambda > 0}$ which has an ancestor $A_k \supset A$ (possibly A itself) satisfying:*

(i) A_k is the ancestor of a single leaf of $\{G(\lambda)\}_{\lambda > 0}$, namely A.

(ii) A_k is large: $\exists x \in A_k, B(x, r_k(x)) \subset A_k$.

[1] We might just use $\max_{i \in [n]} f_n(X_i)$ in practice, which in light of Lemma 1 can be a good surrogate for F (see Figure 3).

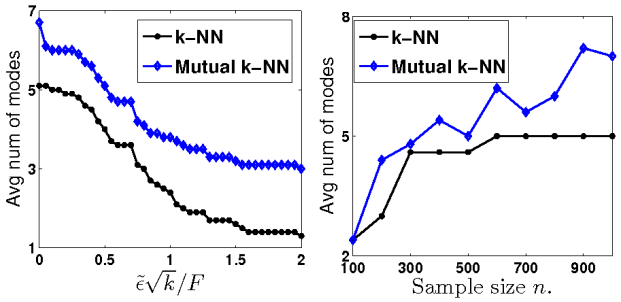


Figure 3. (LEFT). Number of modes (leaves of the empirical tree) as we increase $\tilde{\epsilon}$ from 0. The trees are built on 500-samples (results are averaged over ten such 500-samples) from the 5-modes mixture $\sum_{i=1}^{5} 0.2\mathcal{N}(2\sqrt{d}e_i, I_d)$, $d = 7$. Here $k = (\log n)^{1.5}$, $\theta = 1$, and F is the maximum f_n value over the 10 samples. The mutual k-NN tree being more sparse is rather brittle and requires more pruning. (RIGHT) We fix $\tilde{\epsilon} = F/4\sqrt{k}$, $k = (\log n)^{1.5}$, as we increase n. Results are averaged over 10 n-samples for each n, and F is again the max f_n value over the 10 samples for each n. The k-NN tree quickly asymptotes at 5 modes. The mutual k-NN being more brittle, we're underpruning for $n > 500$, i.e. $\tilde{\epsilon}$ is too small; thus for these settings we would require larger n to obtain the correct number of modes.

(iii) *A_k is sufficiently separated from other components at its level: let $\lambda \doteq \inf_{x\in A_k} f(x)$; A_k and $(\{x : f(x) \geq \lambda\} \setminus A_k)$ are (ϵ, r)-separated.*

Notice that, under the assumptions of Corollary 1, every mode of f is (ϵ, r)-salient for sufficiently large k and $1/\tilde{\epsilon}$.

Theorem 2 (Pruning guarantees)**.** *Let $\delta > 0$. Under the assumptions of Theorem 1, the following holds with probability at least $1 - 3\delta$.*

(a) Suppose the tuning parameter $\tilde{\epsilon} \geq 3\epsilon_k$. Consider two disjoint CCs A_n and A'_n at the same level in $\left\{\widetilde{G}_n(\lambda)\right\}_{\lambda>0}$. Let V be the union of vertices of A_n and A'_n, and define $\lambda \doteq \inf_{x\in V} f(x)$. The vertices of A_n and those of A'_n are in separate CCs of $G(\lambda)$.

(b) Let $\epsilon = 6\epsilon_k + 2\tilde{\epsilon}$ and $r = \frac{\theta}{2}(4k/v_d n\lambda)^{1/d}$. There exists a $1-1$ map from the set of (ϵ, r)-salient modes to the leaves of the empirical tree $\left\{\widetilde{G}_n(\lambda)\right\}_{\lambda>0}$.

The behavior of both the k-NN and mutual k-NN tree, as guaranteed in Theorem 2, is illustrated in Figure 3.

5. Analysis

Theorem 1 follows from lemmas 3 and 6 below. These two lemmas depend on the events described by lemmas 1, 2 and 4 which happen with a combined probability of at least $1 - 3\delta$ for a confidence parameter $\delta > 0$.

Theorem 2 follows from lemmas 5 and 7 below. These two lemmas also depend on the events described by lemmas 1, 2 and 4 which happen with a combined probability of at least $1 - 3\delta$.

5.1. Maintaining Separation

In this section we establish conditions under which points from two disconnected subsets of $\mathbb{R}^d$ remain disconnected in the empirical tree, even after pruning.

The following is an important lemma which establishes the estimation error of f_n relative to f on the sample $\mathbf{X}$. Although of independent interest, we could not find this sort of finite sample statement in the literature on k-NN[2], at least not under our assumptions. The proof combines intuition from an asymptotic analysis of (Devroye & Wagner, 1977) with concentration bounds from (Angluin & Valiant, 1979). The proof for this lemma and the next one are given in (Kpotufe & von Luxburg, 2011).

Lemma 1. *Suppose f satisfies (A.1) and (A.2). There exists $C = C(\mathcal{F})$ such that for $\delta > 0$, for $\epsilon = 11F\sqrt{\ln(2n/\delta)/k}$ and*

$$121\ln(2n/\delta) \leq k \leq C\left(F\sqrt{\ln(2n/\delta)}\right)^{2(\alpha+d)/(3\alpha+d)} n^{2\alpha/(3\alpha+d)},$$

we have with probability at least $1 - \delta$ that $\sup_{X_i\in\mathbf{X}} |f_n(X_i) - f(X_i)| \leq \epsilon$.

The next lemma bounds $r_{k,n}(X_i)$ in terms of $r_k(X_i)$, and hence, in terms of the density at X_i.

Lemma 2. *Suppose f satisfies (A.1) and (A.2). Fix $\lambda > 0$ and let $\mathcal{L}_\lambda \doteq \{x : f(x) \geq \lambda\}$.*

(a) Let $r \doteq \frac{1}{2}(\lambda/2L)^{1/\alpha}$. We have $\forall x, x' \in \mathbb{R}^d$, $\|x - x'\| \leq 2r \implies |f(x) - f(x')| \leq \lambda/2$. If in addition $x \in \mathcal{L}_\lambda$, it follows that $f(x)/2 \leq f(x') \leq 2f(x)$.

(b) Suppose $k \leq 2^{-(d+3)} v_d (2L)^{-d/\alpha} \lambda^{(d+\alpha)/\alpha} n$. We have

$$\forall x \in \mathcal{L}_\lambda,\ r_k(x) \leq \min\left\{2^{-3/d}r, \left(\frac{2k}{v_d n f(x)}\right)^{1/d}\right\}.$$

For $\delta > 0$, if in addition $k \geq 192\ln(2n/\delta)$, we have with probability at least $1-\delta$ that for all $X_i \in \mathbf{X} \cap \mathcal{L}_\lambda$

$$2^{-3/d} r_k(X_i) \leq r_{k,n}(X_i) \leq 2^{3/d} r_k(X_i).$$

The main separation lemma is next. It says that if A and A' are separated by a sufficiently large low density region, then they remain separated in the empirical tree.

[2]There are however many asymptotic analyses of k-NN methods such as (Devroye & Wagner, 1977).

Lemma 3 (Separation)**.** *Suppose f satisfies (A.1) and (A.2). Let G_n be the k-NN or mutual k-NN graph. Define $\epsilon_k \doteq 11F\sqrt{\ln(2n/\delta)/k}$, and let $\delta > 0$. There exists $C = C(\mathcal{F})$ such that, for*

$$192\ln(2n/\delta) \le k \le C\left(F\sqrt{\ln(n/\delta)}\right)^{2(\alpha+d)/(3\alpha+d)} n^{2\alpha/(3\alpha+d)},$$

the following holds with probability at least $1-2\delta$ simultaneously for any two disjoint subsets A, A' of $\mathbb{R}^d$.

Let $\lambda = \inf_{x\in A\cup A'} f(x)$. If A and A' are (ϵ, r)-separated for $\epsilon = 6\epsilon_k + 2\tilde{\epsilon}$ and $r = \frac{\theta}{2}(4k/v_d n\lambda)^{1/d}$, then $A\cap\mathbf{X}$ and $A'\cap\mathbf{X}$ are disconnected in $G_n(\lambda - 2\epsilon_k - \tilde{\epsilon})$ and therefore in $\widetilde{G}_n(\lambda - 2\epsilon_k)$.

Proof. Applying Lemma 1, it's immediate that, with probability at least $1-\delta$, all points of any $A\cup A'\cap\mathbf{X}$ are in $G_n(\lambda-\epsilon_k)$ and lower levels, and no point from $S_{+r}\cap\mathbf{X}$ is in $G_n(\lambda - 5\epsilon_k - 2\tilde{\epsilon})$ or higher levels. Thus any path between A and A' in $G_n(\lambda-2\epsilon_k-\tilde{\epsilon})$ must have an edge through the center $x\in S$ of a ball $B(x,r)\subset S_{+r}$. This edge must therefore have length greater than $2r$. We just need to show that no such edge exists in $G_n(\lambda-2\epsilon_k-\tilde{\epsilon})$.

Let V be the set of points (vertices) in $G_n(\lambda-2\epsilon_k-\tilde{\epsilon})$. By Lemma 1, $\min_{X_i\in V} f(X_i) \ge \lambda - 3\epsilon_k - \tilde{\epsilon}$. Given the density assumption on S, $\lambda \ge 6\epsilon_k + 2\tilde{\epsilon}$ so $\min_{X_i\in V} f(X_i) \ge \lambda/2$ and $V\subset\mathcal{L}_{\epsilon_k}$. Now, given the range of k, Lemma 2 holds for the level set $\mathcal{L}_{\epsilon_k}$. It follows that with probability at least $1-\delta$ (uniform over any such choice of A, A' since the event is a function of $\mathcal{L}_{\epsilon_k}$),

$$\max_{X_i\in V} r_{k,n}(X_i) \le 2^{3/d}\max_{X_i\in V} r_k(X_i) \le \frac{2r}{\theta}.$$

Thus, edge lengths in $G_n(\lambda-2\epsilon_k-\tilde{\epsilon})$ are at most $2r$. □

5.1.1. Identifying Modes

As a corollary to Lemma 3, we can guarantee in Lemma 5 that certain salient modes are recovered by the empirical cluster tree. For this to happen, we require in Definition 7 (ii) that an (ϵ, r)-salient mode A is contained in a sufficiently large set A_k so that we sample points near the mode.

We start with the following VC lemma establishing conditions under which subsets of $\mathbb{R}^d$ contain samples from $\mathbf{X}$.

Lemma 4 (Lemma 5.1 of (Bousquet et al., 2004))**.** *Suppose $\mathcal{C}$ is a class of subsets of $\mathbb{R}^d$. Let $\mathcal{S}_\mathcal{C}(2n)$ denote the $2n$-shatter coefficient of $\mathcal{C}$. Let $\mathcal{F}_n$ denote the empirical distribution over n samples drawn i.i.d from $\mathcal{F}$. For $\delta > 0$, with probability at least $1-\delta$,*

$$\sup_{A\in\mathcal{C}} \frac{\mathcal{F}(A) - \mathcal{F}_n(A)}{\sqrt{\mathcal{F}(A)}} \le 2\sqrt{\frac{\log\mathcal{S}_\mathcal{C}(2n) + \log 4/\delta}{n}}.$$

Lemma 5 (Modes)**.** *Suppose f satisfies (A.1) and (A.2). Let G_n be the k-NN or mutual k-NN graph. Let $\delta > 0$. There exist C and $C' = C'(\mathcal{F})$ such that, for*

$$Cd\ln(n/\delta) \le k \le C'\left(F\sqrt{\ln(n/\delta)}\right)^{2(\alpha+d)/(3\alpha+d)} n^{2\alpha/(3\alpha+d)}$$

the following holds with probability at least $1-3\delta$. Let $\epsilon = 6\epsilon_k + 2\tilde{\epsilon}$ and $r = \frac{\theta}{2}(4k/v_d n\lambda)^{1/d}$. There exists a $1-1$ map from the set of (ϵ, r)-salient modes to the leaves of the empirical tree $\left\{\widetilde{G}_n(\lambda)\right\}_{\lambda>0}$.

Proof. First, with probability at least $1-\delta$, for any (ϵ, r)-salient mode A, there are samples in $\mathbf{X}$ from the containing set A_k (as defined in Definition 7). To arrive at this we apply Lemma 4 for the class $\mathcal{C}$ of all possible balls $B\in\mathbb{R}^d$, (for this class $\mathcal{S}_\mathcal{C}(2n)\le(2n)^{d+1}$). We have with probability at least $1-\delta$ that for all B, $\mathcal{F}_n(B) > 0$ whenever

$$\mathcal{F}(B) \ge \frac{Cd\ln(n/\delta)}{n} > 4\frac{(d+1)\log(2n) + \log(4/\delta)}{n},$$

where C is appropriately chosen to satisfy the last inequality. Now, from the definition of A_k, there exists x such that $B(x, r_k(x)) \subset A_k$, while we have $\mathcal{F}(B(x, r_k(x))) = k/n \ge Cd\ln(n/\delta)/n$, implying that $\mathcal{F}_n(A_k) \ge \mathcal{F}_n(B(x, r_k(x))) \ge 1/n$.

As a consequence of the above argument, there is a finite number m of (ϵ, r)-salient modes since each contributes some points to the final sample $\mathbf{X}$. We can therefore arrange them as $\{A^i\}_{i=1}^m$ so that for $i<j$, we have $\lambda_i \le \lambda_j$ where $\lambda_i = \inf_{x\in A_k^i} f(x)$. An injective map can now be constructed iteratively as follows.

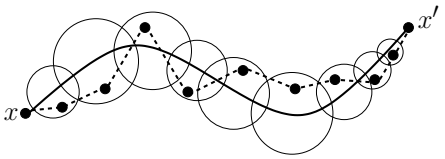
Starting with $i=1$, we have by Lemma 3 that, with probability at least $1-2\delta$, $A_k^i\cap\mathbf{X}$ is disconnected in $\widetilde{G}_n(\lambda_i - 2\epsilon_k)$ from all $A_k^j, j>i$. Let U be the union of those CCs of $\widetilde{G}_n(\lambda_i - 2\epsilon_k)$ containing points from $A_k^i\cap\mathbf{X}$. We've already established that U contains no point from any $A_k^j, j>i$. For $i>1$, U also contains no point from any $A_k^j, j<i$. This is because, again by Lemma 3, $A_k^j\cap\mathbf{X}$ is disconnected in $\widetilde{G}_n(\lambda_j - 2\epsilon_k)$ from $A_k^i\cap\mathbf{X}$, therefore disconnected from U since all CCs in U remain connected at lower levels. Now, since U is disconnected from all $A_k^j, j\ne i$, we can just map A^i to any leaf rooted in U, A^i being the unique image of such a leaf. □

5.2. Maintaining Connectedness

In this section we show that sample points from a connected subset A of $\mathbb{R}^d$ remain connected in the empirical cluster tree before pruning (therefore also after pruning).

Similar to (Chaudhuri & Dasgupta, 2010), for any two points $x, x'\in A\cap\mathbf{X}$ we uncover a path in G_n near

a path P in A that connects the two. The path in G_n (the dashed path depicted below) consists of a sequence $x_1 = x, x_2, \ldots, x_i = x'$ of sample points from balls centered on the path P in A (the solid path depicted below). The intuition is that P is a high density route near which we can find enough sample points to connect x and x'.



The balls centered on P must be chosen sufficiently small and consecutively close so that consecutive terms x_i, x_{i+1} are adjacent in G_n. In (Chaudhuri & Dasgupta, 2010), points are adjacent (at any particular level) whenever they are less than some scale r apart; one can therefore choose balls of the same radius $o(r)$ and consecutively $o(r)$ close. In our particular case, no single scale determines adjacency. Adjacency is determined by the various nearest-neighbor radii and this creates a multiscale effect that complicates the analysis. One way to handle (and effectively get rid of) this multiscale effect is to choose balls on P of the same radius r corresponding to the smallest possible nearest-neighbor radius in G_n (restricted to $A \cap \mathbf{X}$). However, in order to get samples in such small balls one would need rather large sample size n, so the idea results in weak bounds. We instead use an inductive argument which keeps track of the various scales, the intuition being that nearest-neighbor-radii have to change slowly along the path P from x to x'.

Lemma 6 (Connectedness)**.** *Suppose f satisfies (A.1) and (A.2). Let G_n be the k-NN or mutual k-NN graph. Define $\epsilon_k \doteq 11F\sqrt{\ln(2n/\delta)/k}$ and let $\delta > 0$. There exist C and $C' = C'(\mathcal{F})$ such that, for*

$$C\left(\max\left\{1, \sqrt{2}/\theta\right\}\right)^d d\ln(n/\delta) \leq k \leq C'\left(F\sqrt{\ln(n/\delta)}\right)^{2(\alpha+d)/(3\alpha+d)} n^{2\alpha/(3\alpha+d)},$$

the following holds with probability at least $1 - 3\delta$ simultaneously for all connected subsets A of $\mathbb{R}^d$.

Let $\lambda \doteq \inf_{x\in A} f(x) > 2\epsilon_k$. All points in $A \cap \mathbf{X}$ belong to the same CC of $G_n(\lambda - 2\epsilon_k)$, therefore of $\widetilde{G}_n(\lambda - 2\epsilon_k)$.

Proof. First, let C and C' be large enough for lemmas 1 and 2 to hold. Define $r \doteq \frac{1}{2}\left(\epsilon_k/2L\right)^{1/\alpha}$. By Lemma 2 (a), we have that $f(x) \geq \lambda - \epsilon_k/2$ for any $x \in A_{+r}$. Applying Lemma 1, it follows that with probability at least $1 - \delta$ (uniform over choices of A), all points of $A_{+r} \cap \mathbf{X}$ are in $G_n(\lambda - 2\epsilon_k)$. We will show that $A \cap \mathbf{X}$ is connected in $G_n(\lambda - 2\epsilon_k)$ possibly through points in $A_{+r} \setminus A$.

In particular, any $x, x' \in A \cap \mathbf{X}$ are connected through a sequence $\{x_i\}_{i>1}, x_i \in A_{+r} \cap \mathbf{X}$ built according to the following procedure. Let P be a path in A between x and x'. Define $\tau \doteq \min\left\{1, \theta/\sqrt{2}\right\}$.

> Starting at $i = 1$ ($x_1 = x$), set $x_{i+1} = x'$ if $\|x_i - x'\| \leq \theta \min\{r_{k,n}(x_i), r_{k,n}(x')\}$, and we're done, otherwise:
> Let y_i be the point in $P \cap B\left(x_i, \tau 2^{-9/d} r_{k,n}(x_i)\right)$ farthest along the path P from x, i.e. $P^{-1}(y_i)$ is highest in the set. Define the half-ball
>
> $$H(y_i) \doteq \{z : \|z - y\| < \tau 2^{-18/d} r_{k,n}(x_i), (z - y_i)\cdot(x_i - y_i) \geq 0\}.$$
>
> Pick x_{i+1} in $H(y_i) \cap \mathbf{X}$, and continue.

The rest of the argument will proceed inductively as follows. First, assume that $x_i \in A_{+r}$ and that y_i exists. This is necessarily the case for x_1, y_1. Assume $x_{i+1} \neq x'$. We will show that x_{i+1} exists, is also in A_{+r}, and is adjacent to x_i in G_n. It will follow that y_{i+1} must exist (if the process does not end) and is distinct from $y_1, \ldots, y_i$. We'll then argue that the process must also end.

To see that x_{i+1} exists (under the aforementioned assumptions), we apply Lemma 4 for the class $\mathcal{C}$ of all possible half-balls $H(y)$ centered at $y \in \mathbb{R}^d$ (for this class $\mathcal{S}_{\mathcal{C}}(2n) \leq (2n)^{2d+1}$). We have with probability at least $1 - \delta$ that for all $H(y)$, $\mathcal{F}_n(H(y)) > 0$ whenever

$$\mathcal{F}(H(y_i)) \geq \frac{C_0 d\ln(\frac{n}{\delta})}{n} > \frac{(8d+4)\log(2n) + 4\log(\frac{4}{\delta})}{n},$$

where C_0 is appropriately chosen to satisfy the last inequality. We next show $\mathcal{F}(H(y_i))$ satisfies the first inequality.

We first apply Lemma 2 on $\mathcal{L}_{\epsilon_k} \supset A_{+r}$ (this inclusion was established earlier). We have with probability at least $1 - \delta$ (uniform over all A) that for $x_i \in A_{+r}$, $r_{k,n}(x_i) \leq 2^{3/d} r_k(x_i) \leq r$. Thus, for all $z \in H(y_i)$,

$$\begin{aligned}\|z - x_i\| &\leq 2 \cdot \tau 2^{-9/d} r_{k,n}(x_i) \\ &\leq 2 \cdot \tau 2^{-9/d} r \leq 2r, \end{aligned} \tag{2}$$

implying by the same Lemma 2 that $f(z) \geq f(x_i)/2$. Now, from Lemma 1, $f_n(x_i) \leq f(x_i) + \epsilon_k \leq 2f(x_i)$. We can thus write

$$\begin{aligned}\mathcal{F}(H(y_i)) &\geq \frac{1}{4}\operatorname{vol}\left(B(y_i, \tau 2^{-18/d} r_{k,n}(x_i))\right) f(x_i) \\ &= \tau^d 2^{-20} \operatorname{vol}\left(B(x_i, r_{k,n}(x_i))\right) f(x_i) \\ &\geq \tau^d 2^{-21} \operatorname{vol}\left(B(x_i, r_{k,n}(x_i))\right) f_n(x_i) \\ &= \tau^d 2^{-21} \frac{k}{n} \geq \frac{C_0 d\ln(n/\delta)}{n}, \text{ for } C \geq 2^{21} C_0.\end{aligned}$$

Therefore there is a point x_{i+1} in $H(y_i) \cap \mathbf{X}$. In addition $x_{i+1} \in A_{+r}$ since it is within r of $y_i \in A$.

Next we establish that there is an edge between x_i and x_{i+1} in G_n. To this end we relate $r_{k,n}(x_{i+1})$ to $r_{k,n}(x_i)$ by first relating $r_k(x_{i+1})$ to $r_k(x_i)$. Remember that for $z \in A_{+r}$ we have $r_k(z) < r$ so that for any $z' \in B(z, r_k(z))$ we have $f(z)/2 \leq f(z') \leq 2f(z)$. Also recall that we always have $\|x_i - x_{i+1}\| \leq 2r$ (see (2)), implying $f(x_{i+1}) < 2f(x_i)$. We then have

$$\begin{aligned} v_d r_k^d(x_i) \cdot \frac{1}{2} f(x_i) \leq \frac{k}{n} &\leq v_d r_k^d(x_{i+1}) \cdot 2f(x_{i+1}) \\ &\leq v_d r_k^d(x_{i+1}) \cdot 4f(x_i), \end{aligned}$$

where for the first two inequalities we used the fact that both balls $B(x_i, r_k(x_i))$ and $B(x_{i+1}, r_k(x_{i+1}))$ have the same mass k/n. It follows that

$$\begin{aligned} r_{k,n}(x_{i+1}) &\geq 2^{-3/d} r_k(x_{i+1}) \geq 2^{-6/d} r_k(x_i) \\ &\geq 2^{-9/d} r_{k,n}(x_i), \end{aligned} \tag{3}$$

implying $2^{-9/d} r_{k,n}(x_i) \leq \min\{r_{k,n}(x_i), r_{k,n}(x_{i+1})\}$. We then get

$$\begin{aligned} \|x_i - x_{i+1}\|^2 &= \|x_i - y_i\|^2 + \|x_{i+1} - y_i\|^2 \\ &\quad - (x_i - y_i) \cdot (x_{i+1} - y_i) \\ &\leq \|x_i - y_i\|^2 + \|x_{i+1} - y_i\|^2 \\ &\leq 2\tau^2 \cdot \min\left\{r_{k,n}^2(x_i), r_{k,n}^2(x_{i+1})\right\} \\ &\leq \theta^2 \min\left\{r_{k,n}^2(x_i), r_{k,n}^2(x_{i+1})\right\}, \end{aligned}$$

meaning x_i and x_{i+1} are adjacent in G_n.

Finally we argue that y_{i+1} must exist. By (3) above we have

$$\|x_{i+1} - y_i\| < \tau 2^{-18/d} r_{k,n}(x_i) \leq \tau 2^{-9/d} r_{k,n}(x_{i+1}),$$

in other words the ball $B\left(x_{i+1}, \tau 2^{-9/d} r_{k,n}(x_{i+1})\right)$ contains $y_i \in P$ in its interior. It follows by continuity of P that there is a point y_{i+1} in this ball further along the path from x_i than y_i. Thus, recursively all y_i's must be distinct, implying that all x_i's must be distinct. Since all x_i's belong to the finite sample $\mathbf{X}$ the process must eventually terminate. □

5.2.1. Pruning of Spurious Branches

As a corollary to Lemma 6 we can guarantee in Lemma 7 that the pruning procedure will remove all spurious branchings, and hence, all spurious clusters.

Lemma 7 (Pruning)**.** *Let $\delta > 0$. Under the assumptions of Lemma 6, the following holds with probability at least $1 - 3\delta$, provided $\tilde{\epsilon} \geq 3\epsilon_k$.*

Consider two disjoint CCs A_n and A'_n at the same level in $\left\{\widetilde{G}_n(\lambda)\right\}_{\lambda>0}$. Let V be the union of vertices of A_n and A'_n, and define $\lambda \doteq \inf_{x \in V} f(x)$. The vertices of A_n and those of A'_n are in separate CCs of $G(\lambda)$.

Proof. Let $\lambda_n = \min_{x \in V} f_n(x)$ be the level in the empirical tree containing A_n, A'_n. By Lemma 1, $\sup_{x \in \mathbf{X}} |f_n(x) - f(x)| \leq \epsilon_k$ so $\lambda_n \leq \lambda + \epsilon_k$. Thus, we must have $\lambda > 2\epsilon_k$, since otherwise $\lambda_n \leq \tilde{\epsilon}$ implying $\widetilde{G}_n(\lambda_n)$ must have a single connected component.

Now suppose points in V were in the same component A of $G(\lambda)$. By Lemma 6, all of $A \cap \mathbf{X}$ is connected in $G_n(\lambda - 2\epsilon_k)$ and at lower levels. By the last argument $\lambda_n - \tilde{\epsilon} \leq \lambda - 2\epsilon_k$ so the pruning procedure reconnects A_n and A'_n. □

Acknowledgements

We thank Sanjoy Dasgupta for interesting discussions which helped improve presentation.

References

Angluin, D. and Valiant, L.G. Fast probabilistic algorithms for Hamiltonian circuits and matchings. *Journal of Computer and System Sciences*, 19:155–193, 1979.

Bousquet, O., Boucheron, S., and Lugosi, G. Introduction to statistical learning theory. *Lecture Notes in Artificial Intelligence*, (3176):169–207, 2004.

Chaudhuri, K. and Dasgupta, S. Rates of convergence for the cluster tree. *Neural Information Processing Systems*, 2010.

Devroye, L.P. and Wagner, T.J. The strong uniform consistency of nearest neighbor density estimates. *The Annals of Statistics*, 5:536–540, 1977.

Hartigan, J.A. Consistency of single linkage for high-density clusters. *Journal of the American Statistical Association*, 76(374): 388–394, 1981.

Kpotufe, S. and von Luxburg, U. Pruning nearest neighbor cluster trees. *arXiv:1105.0540*, 2011.

Maier, M., Hein, M., and von Luxburg, U. Optimal construction of k-nearest neighbor graphs for identifying noisy clusters. *Theoretical Computer Science*, 410:1749–1764, 2009.

Rigollet, P. and Vert, R. Fast rates for plug-in estimators of density level sets. *Bernouilli*, 15(4):1154–1178, 2009.

Rinaldo, A. and Wasserman, L. Generalized density clustering. *Annals of Statistics*, 38(5):2678–2722, 2010.

Rinaldo, A., Singh, A., Nugent, R., and Wasserman, L. Stability of density based clustering. *arXiv:1011.2771v1*, pp. 2760–2782, 2010.

Singh, A., Scott, C., and Nowak, R. Adaptive hausdorff estimation of density level sets. *Annals of Statistics*, 37(5), 2009.

Stueltze, W. and Nugent, R. Clustering with confidence: A binning approach. *International Federation Classification Societies Conference*, 2009.

Stueltze, W. and Nugent, R. A generalized single linkage method for estimating the cluster tree of a density. *Journal of Computational and Graphical Statistics*, 19(2):397–418, 2010.

Wishart, D. Mode analysis: A generalization of nearest neighbor which reduces chaining effects. *Numerical Taxonomy*, pp. 282–311, 1969.

Wong, M. and Lane, T. A kth nearest neighbor clustering procedure. *Journal of the Royal Statistical Society Series B*, 45(3): 362–368, 1983.