The Importance of Data Clustering Stability

Ezekiel Cochran
Advised by Dr. Blevins

LeTourneau University

February 8, 2024

figures/letu.png

- HDBSCAN stands for “Hierarchical Density-Based Spatial Clustering of Applications with Noise”
- It was introduced in 2015, by Campello, Moulavi, and Zimek [2], for automatic clustering of n-dimensional data sets.
- It uses a “mutual reachability distance”, and identifies clusters by iteratively merging or splitting groups of data points.

Mutual Reachability Distance

Definition

The **mutual reachability distance** of two data points x and y, written $\mathrm{mrd}(x, y)$, with a paramater $k \in \mathbb{Z}^+$, is the maximum of

$$\begin{cases} \text{Distance between } x \text{ and } y \\ \text{Distance to } k\text{-nearest neighbor of } x \\ \text{Distance to } k\text{-nearest neighbor of } y \end{cases}$$

where “distance” refers to the standard ℓ^2 norm of $x - y$, or euclidian distance.

Note: The k-nearest neighbor is the kth closest point *not counting the point itself*. Undefined if there are k or less total data points.

HDBSCAN

figures/from_tutorial/scatterplot.png

figures/from_tutorial/mst.png

figures/from_tutorial/clusters.png

“Unreasonable Stability”

In [1], Dr. Blevins and her advisor Dr. Bridges of Oak Ridge National Labratory showed that distinct minimum spanning trees on the same data set give the same clusters.

figures/stability.png

Goals

1. Verify that it is possible to get different minimum spanning trees over the same data
2. Test whether this can occur in real-world data (and get a rough sense of how often)
3. Classify exactly when we get non-unique minimum spanning trees

Base Case

Example

Different orderings of the simple data set $[[0,0],[0,1],[1,1],[1,0]]$ give different minimum spanning trees

figures/base_case.png

Iris

Example

the Iris data set [6] records information about various types of iris flowers.

figures/iris2d.png

Iris Closeup

figures/iris2d_closeup.png

Note: All graphs (except the first) are a two-dimensional projection of the higher-dimensional data sets, but the python script verifies that there are truly multiple trees in the native dimension.

MNIST Closeup

Example

MNIST is a massive data set of handwritten digits. We plotted the first 50.

figures/mnist2d_closeup.png

core$_k$

Definition

Given some $k \in \mathbb{Z}^+$, for a data point x, the core$_k(x)$ is the distance to the k-nearest neighbor of x.

Note that this definition depends on which distance we are using. Usually we will use the ℓ^2 norm.
If the language seems strange, it is inspired by the HDBSCAN algorithm itself.

NECKSc

Definition

Let X, along with some distance definition dist, be a metric space. And, with a paramater $k \in \mathbb{Z}^{+}$, define

$$\epsilon = \min_{x \in X} \mathrm{core}_k(x) = \mathrm{core}_k(x_0) \text{ for some } x_0 \in X.$$

Then X is a 'not-everywhere core_k sparse (closed)' (**NECKSc**) metric space, if and only if

$\exists y, z \in \{x \in X : 0 < \mathrm{dist}(x, x_0) \leq \epsilon\}$ with $y \neq z$ and $\mathrm{dist}(y, z) \leq \epsilon$.

In other words, a metric space is NECKSc if there are two points in the k-nearest neighbors of x_0 that are also within ϵ of each other.

Original NECKS Definition

Note: This is a tweaked version of a NECKS metric space.

figures/necks_def.png

NECKSc $\implies$ Multiple Minimum Spanning Trees

Theorem

(From [1]) Let (X, dist) *be a NECKS metric space for some* $k \in \mathbb{N}$*. Let* $G(V, E)$ *be the complete graph such that* $V = X$ *and for* $(a, b) \in E$*,* $w(a, b) = \mathrm{mrd}(a, b)$*. Then* G *has multiple minimum spanning trees.*

Dr. Blevins and Dr. Bridges showed this for NECKS spaces, but it also holds for NECKSc spaces. So we need only to show that a data set (along with a definition of distance) is a NECKSc metric space to show that it has multiple minimum spanning trees.

Sufficient k value

Theorem

For $k \geq 6$, any two dimensional data set, with the ℓ^2 norm as distance, is a NECKSc space, and thus has multiple minimum spanning trees.

Proof Sketch

Increasing k cannot decrease ϵ, so proving the theorem for $k = 6$ is sufficient to show it's truth for all $k \geq 6$.

- Assume BWOC there exists some two-dimensional data set X that is not a NECKSc metric space for $k = 6$ (with standard euclidian distance).
- Order the six closest points to x_0 in terms of their angle relative to x_0.
- By the definition of $\text{core}_6(x_0)$, these are all within ϵ of x_0.
- No pair of these points is within ϵ of each other, becuase X is not a NECKSc space.

Proof Sketch (continued)

- This means that blue lines are strictly greater than ϵ, and red/green lines are less than or equal to ϵ.
- The angles formed around x_0 sum to more than 2π.

figures/2d_proof_fig.png

The Kissing Number

Definition

Given an n-dimensional sphere S with radius r, the n-**dimensional kissing number** $\text{kiss}(n)$ is the maximum number of spheres of radius r that can be tangent to S without the interiors of any two spheres overlapping*.

*Two spheres' interiors overlap if their intersection is neither the empty set nor a single point.

Background on the Kissing Number

Already hard in three dimensions: Newton disagreed with David Gregory about kiss(3). "Extra" space in 3D unlike in 2D.

figures/1dkn.png

figures/kissing_picture.jpg

Kissing Numbers (from [3])

figures/kissing_nums_list.png

$k > \mathrm{kiss}(n) \implies$ NECKSc

Theorem

Let X be n-dimensional data set, and let dist *be euclidian distance. Then (X, dist) is a NECKSc space if*

1. $k > \mathrm{kiss}(n)$ *and*
2. $|X| > k$*

*This second point is formally necessary but practically meaningless: $|X| \leq k$ would mean the data is all in one cluster.

Proof Sketch

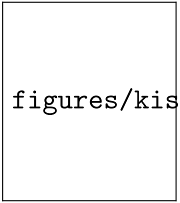
Let X be an n-dimensional data set, and let (X, dist) **not** be a NECKSc space. We show that this implies $k \leq \text{kiss}(n)$.

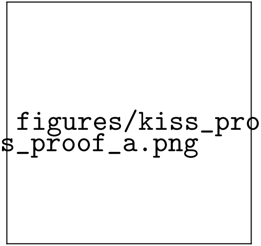
- Consider the k closest points to x_0, call this set A.
- Notice that $|A| = k$.
- Let each $p \in A$ correspond to a new p', on $\overrightarrow{x_0 p}$, with $\text{dist}(x_0, p') = \epsilon \geq \text{dist}(x_0, p)$. (Call this set of new points A').
- $|A'| = |A| = k$.
- For all $p, q \in A$ with $p \neq q$, $\text{dist}(p, q) > \epsilon$, becuase X is not a NECKSc space.
- Because new points travel outward along diverging lines, we also know $\text{dist}(p', q') > \epsilon$ for all $p', q' \in A'$ with $p' \neq q'$.

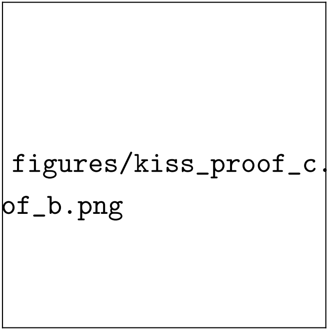
Proof Sketch (continued)

- We place a hypersphere of radius $\frac{\epsilon}{2}$ centered at each $p' \in A'$, as well as one centered at x_0.
- These hyperspheres are all tangent to the hypersphere centered at x_0, because for each $p' \in A'$, $\mathrm{dist}(p', x_0) = \epsilon$.
- But no hypersphere is tangent to any other, because $\mathrm{dist}(p', q') > \epsilon$ for all $p', q' \in A'$ with $p' \neq q'$.
- These all have radius $\frac{\epsilon}{2}$, so $\mathrm{kiss}(n) \geq k$ by definition.

Proof Illustration







Conclusion

- We have verified that it is possible to get different minimum spanning trees from HDBSCAN over the same data set.
- We have tested that this can occur in real-world data, and have empirically shown that this is often.
- We have shown conditions that gurantee a metric space is NECKSc, which implies non-unique minimum spanning trees.

References

Deborah Blevins.
The unreasonable stability of hdbscan.
Presentation, 2017.

Ricardo Campello, Davoud Moulavi, and Arthur Zimek.
Hierarchical density estimates for data clustering, visualization, and outlier detection.
ACM Transactions on Knowledge Discovery from Data, 10(5), 2015.

Henry Cohn.
Kissing numbers.
https://cohn.mit.edu/kissing-numbers, 2023.

Dariel Dato-On.
Mnist in csv.
https://www.kaggle.com/datasets/oddrationale/mnist-in-csv, 2018.

Leland McInnes, John Healy, and Steve Astels.
How hdbscan works.
https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html, 2016.

Antony Unwin and Kim Kleinman.
The iris data set: In search of the source of virginica.
Significance, 18, 2021.

Mary Wootters.
Compressed sensing and the restricted isometry property.