The HDBSCAN* clustering algorithm

Alexander Rolle

3 May 2018

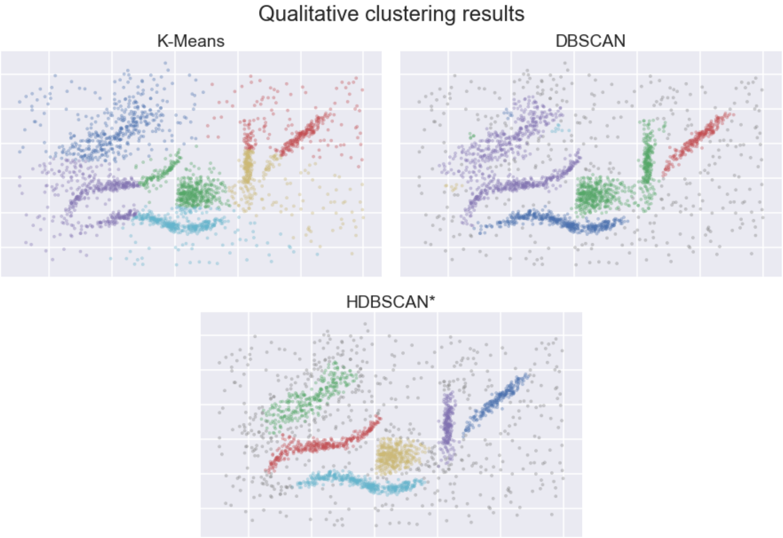
References

- Healy, John and Leland McInnes. Accelerated heirarchical density clustering. Preprint, arXiv: 1705.07321v2 [stat.ML], 2017.
- Jardine, J.F. Cluster graphs. Preprint, http://uwo.ca/math/faculty/jardine/preprints/preprints.html 2017

Introduction

HDBSCAN* is a clustering algorithm for exploratory data analysis.

Introduction



[1, p2]

Introduction

HDBSCAN* is a clustering algorithm for exploratory data analysis.

We have to make two choices to get started:

a minimum cluster size m,

and an integer k that defines a density threshold.

Vietoris-Rips complex

A data cloud X is a finite set of points in some metric space.

For $\epsilon \geq 0$, $V_\epsilon(X)$ is a semi-simplicial set, with

$$V_\epsilon(X)_n = \{(x_1, \ldots, x_n) \,:\, d(x_i, x_j) < \epsilon \text{ for all } i, j\}$$

Vietoris-Rips complex

A data cloud X is a finite set of points in some metric space.

For $\epsilon \geq 0$, $V_\epsilon(X)$ is a semi-simplicial set, with

$$V_\epsilon(X)_n = \{(x_1, \ldots, x_n) \, : \, d(x_i, x_j) < \epsilon \text{ for all } i, j\}$$

$$d_i : V_\epsilon(X)_n \to V_\epsilon(X)_{n-1}$$
$$(x_1, \ldots, x_n) \mapsto (x_1, \ldots, \hat{x}_i, \ldots, x_n)$$

Path components

$V_\epsilon(X)_0$ is just the set X.

Say $x \sim y$ if there are $x_1, \ldots, x_k \in X$ with

$$d(x, x_1) < \epsilon, \quad d(x_i, x_{i+1}) < \epsilon, \quad d(x_k, y) < \epsilon$$

Path components

$V_\epsilon(X)_0$ is just the set X.

Say $x \sim y$ if there are $x_1, \ldots, x_k \in X$ with

$$d(x, x_1) < \epsilon, \quad d(x_i, x_{i+1}) < \epsilon, \quad d(x_k, y) < \epsilon$$

For example

$$\bullet x \qquad \bullet x_1 \qquad \bullet y$$

Path components

$V_\epsilon(X)_0$ is just the set X.

Say $x \sim y$ if there are $x_1, \ldots, x_k \in X$ with

$$d(x, x_1) < \epsilon, \quad d(x_i, x_{i+1}) < \epsilon, \quad d(x_k, y) < \epsilon$$

For example

$$\bullet x \qquad \bullet x_1 \qquad \bullet y$$

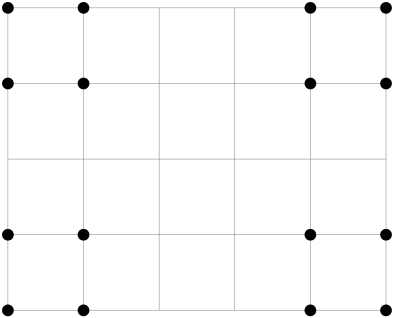
Write

$$\pi_0(V_\epsilon(X)) = X \,/ \sim$$

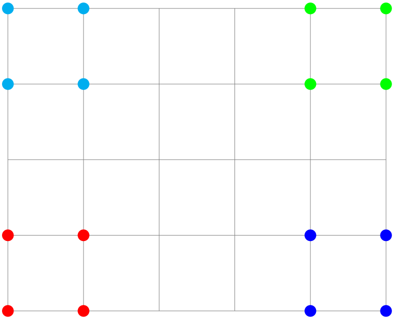
Note that π_0 is a functor.

Example

Say our data cloud X looks like this

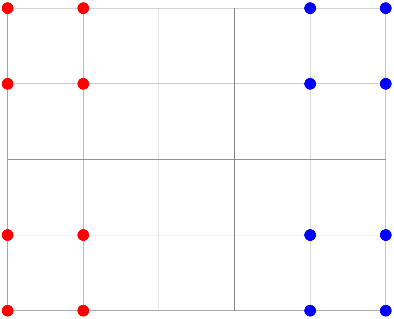


Example



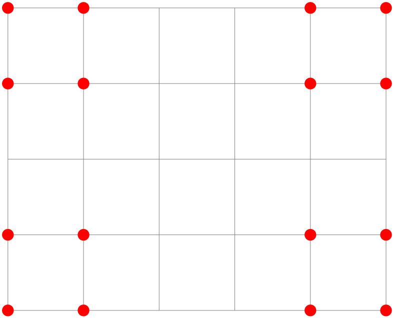
$$1 < \epsilon < 2$$

Example



$$2 < \epsilon < 3$$

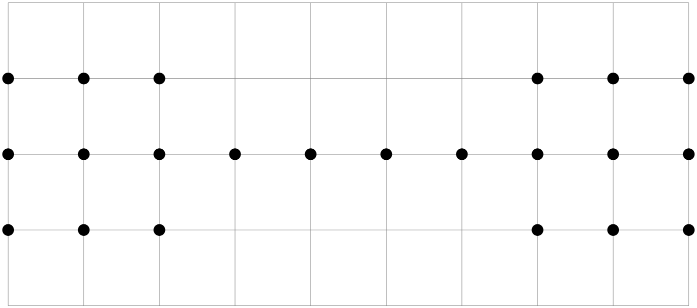
Example



$$3 < \epsilon < \infty$$

Less nice example

But, say our data cloud X looks like this

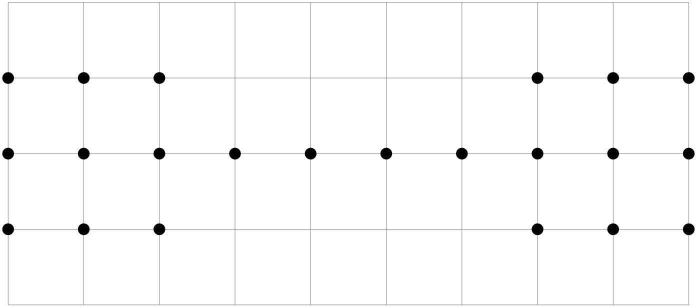


Lesnick filtration

Choose an integer $k \geq 0$. Define $L_{\epsilon,k} \subset V_\epsilon$ to be the full subcomplex on those vertices of degree at least k.

Lesnick filtration

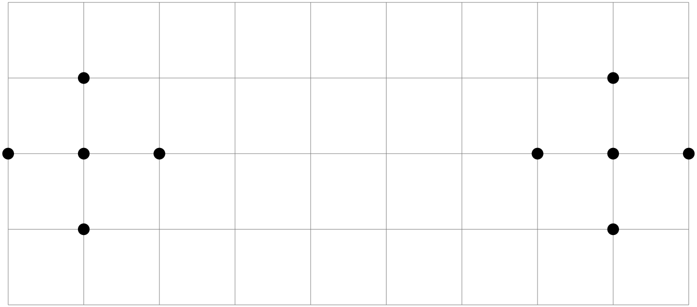
Choose an integer $k \geq 0$. Define $L_{\epsilon,k} \subset V_{\epsilon}$ to be the full subcomplex on those vertices of degree at least k.



$1 < \epsilon$ and $k = 0, 1, 2$

Lesnick filtration

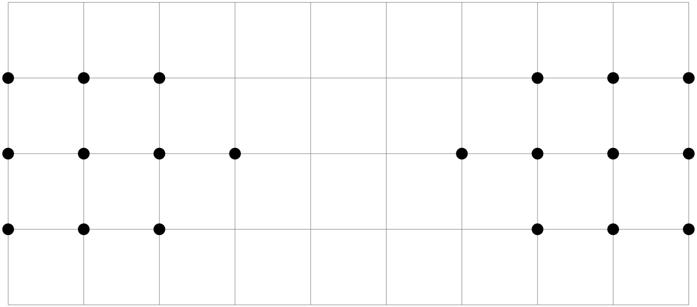
Choose an integer $k \geq 0$. Define $L_{\epsilon,k} \subset V_\epsilon$ to be the full subcomplex on those vertices of degree at least k.



$1 < \epsilon < \sqrt{2}$ and $k = 3$

Lesnick filtration

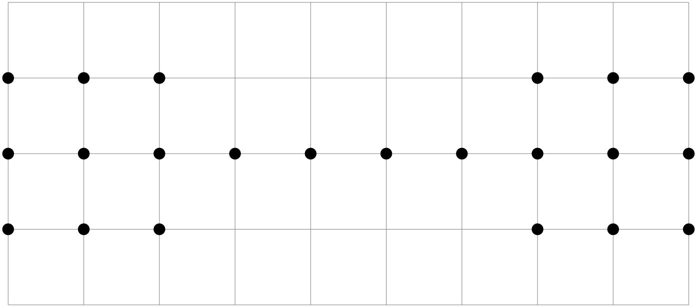
Choose an integer $k \geq 0$. Define $L_{\epsilon,k} \subset V_\epsilon$ to be the full subcomplex on those vertices of degree at least k.



$\sqrt{2} < \epsilon < \sqrt{5}$ and $k = 3$

Lesnick filtration

Choose an integer $k \geq 0$. Define $L_{\epsilon,k} \subset V_\epsilon$ to be the full subcomplex on those vertices of degree at least k.



$\sqrt{5} < \epsilon$ and $k = 3$

What is a cluster?

Provisional answer: a cluster is a path component of V_ϵ (or $L_{\epsilon,k}$) that survives for some range of ϵ.

HDBSCAN*

Recall that we've chosen non-negative integers k and m.

HDBSCAN*

Recall that we've chosen non-negative integers k and m.

Consider the poset

$$\mathbb{R}_{\geq 0} = \{x \in \mathbb{R} \ : \ x \geq 0\}$$

with $x \leftarrow y$ if $x \leq y$.

HDBSCAN*

Recall that we've chosen non-negative integers k and m.

Consider the poset

$$\mathbb{R}_{\geq 0} = \{x \in \mathbb{R} \;:\; x \geq 0\}$$

with $x \leftarrow y$ if $x \leq y$.

Define a functor $F : \mathbb{R}_{\geq 0}^{op} \to \mathsf{Set}$

$$F(x) = \{C_0 \;:\; C \in \pi_0(L_{x,k})\}$$

where C_0 is the set of points in the path component C.

HDBSCAN*

Recall that we've chosen non-negative integers k and m.

Consider the poset

$$\mathbb{R}_{\geq 0} = \{x \in \mathbb{R} \ : \ x \geq 0\}$$

with $x \leftarrow y$ if $x \leq y$.

Define a functor $F : \mathbb{R}_{\geq 0}^{op} \to \mathrm{Set}$

$$F(x) = \{C_0 \ : \ C \in \pi_0(L_{x,k})\}$$

where C_0 is the set of points in the path component C.

Define $G \subset F$

$$G(x) = \{s \in F(x) \ : \ |s| \geq m\} \ .$$

HDBSCAN*

Write $S = \bigsqcup_{x \in \mathbb{R}_{\geq 0}} G(x)$.

HDBSCAN*

Write $S = \bigsqcup_{x \in \mathbb{R}_{\geq 0}} G(x)$.

If $x \leq y$ there is a structure map $G_{x,y} : G(x) \to G(y)$.

HDBSCAN*

Write $S = \bigsqcup_{x \in \mathbb{R}_{\geq 0}} G(x)$.

If $x \leq y$ there is a structure map $G_{x,y} : G(x) \to G(y)$.

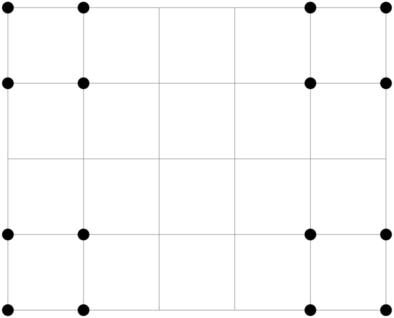
Let $s \in G(x)$ and $t \in G(y)$. We say $s \sim t$ if $G_{x,y}^{-1}(t) = \{s\}$, and for all z with $x \leq z \leq y$ we have $|G_{z,y}^{-1}(t)| = 1$.

HDBSCAN*

Write $S = \bigsqcup_{x \in \mathbb{R}_{\geq 0}} G(x)$.

If $x \leq y$ there is a structure map $G_{x,y} : G(x) \to G(y)$.

Let $s \in G(x)$ and $t \in G(y)$. We say $s \sim t$ if $G_{x,y}^{-1}(t) = \{s\}$, and for all z with $x \leq z \leq y$ we have $|G_{z,y}^{-1}(t)| = 1$.



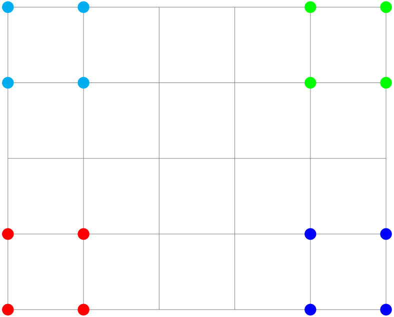
$m = 2, k = 0$

HDBSCAN*

Write $S = \bigsqcup_{x \in \mathbb{R}_{\geq 0}} G(x)$.

If $x \leq y$ there is a structure map $G_{x,y} : G(x) \to G(y)$.

Let $s \in G(x)$ and $t \in G(y)$. We say $s \sim t$ if $G_{x,y}^{-1}(t) = \{s\}$, and for all z with $x \leq z \leq y$ we have $|G_{z,y}^{-1}(t)| = 1$.



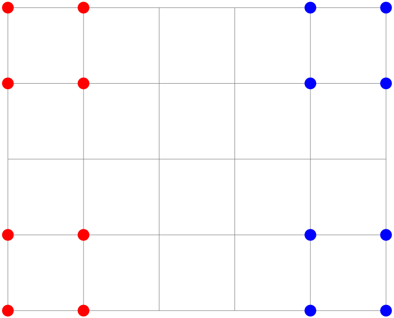
$1 < \epsilon < 2$

HDBSCAN*

Write $S = \bigsqcup_{x \in \mathbb{R}_{\geq 0}} G(x)$.

If $x \leq y$ there is a structure map $G_{x,y} : G(x) \to G(y)$.

Let $s \in G(x)$ and $t \in G(y)$. We say $s \sim t$ if $G_{x,y}^{-1}(t) = \{s\}$, and for all z with $x \leq z \leq y$ we have $|G_{z,y}^{-1}(t)| = 1$.



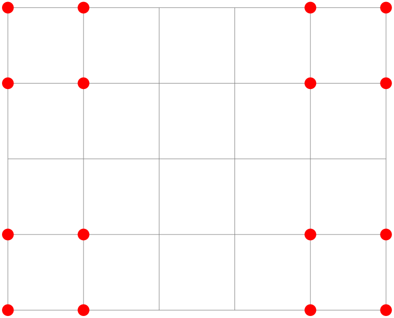
$2 < \epsilon < 3$

HDBSCAN*

Write $S = \bigsqcup_{x \in \mathbb{R}_{\geq 0}} G(x)$.

If $x \leq y$ there is a structure map $G_{x,y} : G(x) \to G(y)$.

Let $s \in G(x)$ and $t \in G(y)$. We say $s \sim t$ if $G_{x,y}^{-1}(t) = \{s\}$, and for all z with $x \leq z \leq y$ we have $|G_{z,y}^{-1}(t)| = 1$.



$3 < \epsilon < \infty$

HDBSCAN*

In this example, $S/\sim$ has seven elements:

$$\begin{aligned} &4 \text{ clusters with } 1 < \epsilon < 2 \\ +\ &2 \text{ clusters with } 2 < \epsilon < 3 \\ +\ &1 \text{ cluster with } 3 < \epsilon < \infty \end{aligned}$$

HDBSCAN*

In this example, $S/\sim$ has seven elements:

$$\begin{aligned} &4 \text{ clusters with } 1 < \epsilon < 2 \\ +\, &2 \text{ clusters with } 2 < \epsilon < 3 \\ +\, &1 \text{ cluster with } 3 < \epsilon < \infty \end{aligned}$$

There are three obvious ways to cluster this example; to choose the best one, the algorithm assigns each cluster a score.

Scoring in HDBSCAN*

Let $[s]$ be a cluster. For $x \geq 0$, let s_x be the member of $[s]$ that lies in $G(x)$, or the empty set if $[s]$ has no member in $G(x)$.

Define a step function $\hat{s}(x) = |s_x|$.

The persistence score σ of $[s]$ is

$$\sigma([s]) = \int_0^\infty \frac{\hat{s}(x)}{x^2}\, dx \ .$$

Scoring in HDBSCAN*

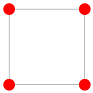
Let $[s]$ be a cluster. For $x \geq 0$, let s_x be the member of $[s]$ that lies in $G(x)$, or the empty set if $[s]$ has no member in $G(x)$.

Define a step function $\hat{s}(x) = |s_x|$.

The persistence score σ of $[s]$ is

$$\sigma([s]) = \int_0^\infty \frac{\hat{s}(x)}{x^2}\, dx \; .$$

For example, let $[s]$ be the cluster



It has persistence score

$$\sigma([s]) = \int_1^2 \frac{4}{x^2}\, dx \; = 2 \; .$$

Scoring in HDBSCAN*

Let $[s]$ be a cluster. For $x \geq 0$, let s_x be the member of $[s]$ that lies in $G(x)$, or the empty set if $[s]$ has no member in $G(x)$.

Define a step function $\hat{s}(x) = |s_x|$.

The persistence score σ of $[s]$ is

$$\sigma([s]) = \int_0^\infty \frac{\hat{s}(x)}{x^2}\, dx \ .$$

The term $\frac{1}{x^2}$ means that a cluster gets a higher score for existing at smaller distance scales.

Scoring in HDBSCAN*

Let $[s]$ be a cluster. For $x \geq 0$, let s_x be the member of $[s]$ that lies in $G(x)$, or the empty set if $[s]$ has no member in $G(x)$.

Define a step function $\hat{s}(x) = |s_x|$.

The persistence score σ of $[s]$ is

$$\sigma([s]) = \int_0^\infty \frac{\hat{s}(x)}{x^2}\, dx \; .$$

The term $\frac{1}{x^2}$ means that a cluster gets a higher score for existing at smaller distance scales.

Note that the integral

$$\int_0^\epsilon \frac{1}{x^2}\, dx$$

doesn't converge for any $\epsilon > 0$.

Comparing scores

Define the "points" function p

$$p([s]) = \bigcup_{x=0}^{\infty} s_x \subset X \ .$$

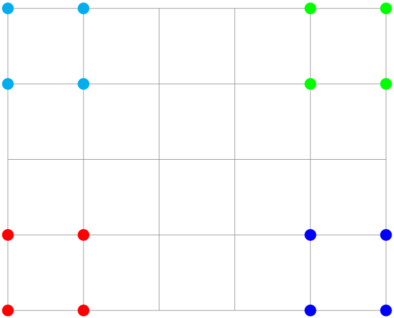
Say we have clusters $\{[s_1], \ldots, [s_n]\}$. We choose $I \subset \{1, \ldots, n\}$ to maximize

$$\sum_{i \in I} \sigma([s_i])$$

subject to the constraint that for all $i, j \in I$ with $i \neq j$,

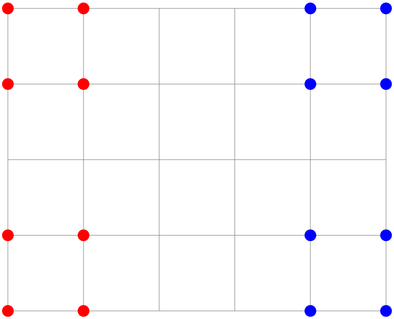
$$p([s_i]) \cap p([s_j]) = \emptyset \ .$$

Example



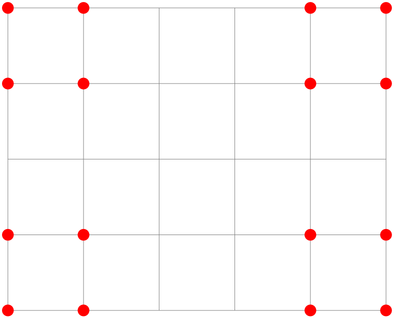
$$\sum_{i \in I} \sigma([s_i]) = 4 \cdot \int_1^2 \frac{4}{x^2}\, dx = 8$$

Example



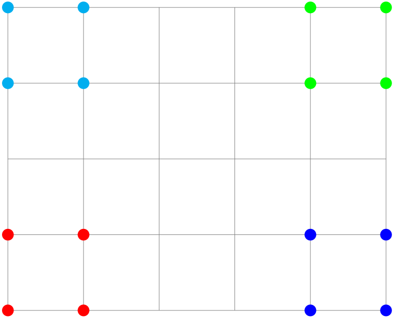
$$\sum_{i \in I} \sigma([s_i]) = 2 \cdot \int_2^3 \frac{8}{x^2}\, dx = \frac{8}{3}$$

Example



$$\sum_{i \in I} \sigma([s_i]) = 1 \cdot \int_3^\infty \frac{16}{x^2}\, dx = \frac{16}{3}$$

Example



HDBSCAN* clusters this example like this

2-dimensional persistence

Consider the poset $\mathbb{N} = \{0, 1, 2, \dots\}$ with $m \to n$ if $m \leq n$.

Then $\mathbb{R}_{\geq 0} \times \mathbb{N}$ is a poset with $(y, m) \to (x, n)$ if $x \leq y$ and $m \leq n$.

Define $F : \mathbb{R}_{\geq 0}^{op} \times \mathbb{N}^{op} \to \mathsf{Set}$

$$F(x, n) = \{ C_0 \ : \ C \in \pi_0(L_{x,n}) \} \ .$$

2-dimensional persistence

Consider the poset $\mathbb{N} = \{0, 1, 2, \dots\}$ with $m \to n$ if $m \leq n$.

Then $\mathbb{R}_{\geq 0} \times \mathbb{N}$ is a poset with $(y, m) \to (x, n)$ if $x \leq y$ and $m \leq n$.

Define $F : \mathbb{R}_{\geq 0}^{op} \times \mathbb{N}^{op} \to \mathsf{Set}$

$$F(x, n) = \{ C_0 \ : \ C \in \pi_0(L_{x,n}) \} \ .$$

This could be the starting point for a 2-dimensional clustering algorithm, which wouldn't require a density threshold k.

Healy and McInnes' approach to this problem is forthcoming.

Another approach is given in [2].