DBSCAN - Derivations & Proofs

Paul F. Roysdon, Ph.D.

Contents

1 Mathematical Derivations & Proofs

1.1 Introduction

DBSCAN is a *density-based* clustering algorithm that discovers clusters of arbitrary shape by expanding connected regions of sufficiently high point density and labeling isolated points as noise. From first principles, DBSCAN formalizes density via an ε-neighborhood and a minimum point count (`minPts`); clusters are then the maximal sets of points that are mutually reachable through chains of dense neighborhoods. We present the exact mathematical definitions, prove key properties (reachability and connectivity), show that the algorithm recovers the axiomatic notion of a cluster, and provide the complete procedure with complexity and parameter guidelines.

1.2 Data and Notation

Let the dataset be

$$\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{n}, \qquad \mathbf{x}_i \in \mathbb{R}^d.$$

Let $d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ be a metric (e.g., Euclidean). Fix parameters

$$\varepsilon > 0 \quad \text{(neighborhood radius)}, \qquad m \in \mathbb{N},\ m \geq 1 \quad (\texttt{minPts}).$$

For any $\mathbf{x} \in \mathbb{R}^d$ define the (closed) ε-neighborhood and its cardinality

$$\mathcal{N}_\varepsilon(\mathbf{x}) \;=\; \{\mathbf{z} \in \mathcal{D} :\; d(\mathbf{z}, \mathbf{x}) \leq \varepsilon\}, \qquad \deg_\varepsilon(\mathbf{x}) \;=\; |\mathcal{N}_\varepsilon(\mathbf{x})|. \tag{1}$$

(We adopt the standard convention that $\mathbf{x} \in \mathcal{N}_\varepsilon(\mathbf{x})$.)

1.3 Density-Based Notions (Core, Border, Noise)

Core point. $\mathbf{x}$ is a *core point* iff $\deg_\varepsilon(\mathbf{x}) \geq m$.

Direct density-reachability. A point $\mathbf{y}$ is *directly density-reachable* from $\mathbf{x}$ iff

$$\mathbf{y} \in \mathcal{N}_\varepsilon(\mathbf{x}) \quad \text{and} \quad \deg_\varepsilon(\mathbf{x}) \geq m \quad (\mathbf{x} \text{ is core}). \tag{2}$$

Density-reachability. A point $\mathbf{y}$ is *density-reachable* from $\mathbf{x}$ (w.r.t. ε, m) iff there exists a finite sequence

$$\mathbf{x} = \mathbf{x}^{(0)},\ \mathbf{x}^{(1)},\ \ldots,\ \mathbf{x}^{(L)} = \mathbf{y}$$

such that $\mathbf{x}^{(\ell)}$ is directly density-reachable from $\mathbf{x}^{(\ell-1)}$ for all $\ell = 1, \ldots, L$. We write $\mathbf{x} \xrightarrow{\varepsilon,m} \mathbf{y}$.

Density-connectedness. Points $\mathbf{x}$ and $\mathbf{y}$ are *density-connected* iff there exists a point $\mathbf{o}$ such that both $\mathbf{x}$ and $\mathbf{y}$ are density-reachable from $\mathbf{o}$:

$$\exists\, \mathbf{o} \in \mathcal{D} \text{ with } \mathbf{o} \xrightarrow{\varepsilon,m} \mathbf{x} \text{ and } \mathbf{o} \xrightarrow{\varepsilon,m} \mathbf{y}. \tag{3}$$

Border and noise points. A non-core point that lies in the ε-neighborhood of some core point is called a *border* (or *boundary*) point. A point that is neither core nor border is labeled *noise* (outlier).

1.4 Basic Properties (Proofs)

We collect the fundamental properties used by DBSCAN.

Lemma 1 (Asymmetry of direct density-reachability)**.** *Direct density-reachability need not be symmetric: it is possible that* $\mathbf{y}$ *is directly density-reachable from* $\mathbf{x}$ *but* $\mathbf{x}$ *is not directly density-reachable from* $\mathbf{y}$*.*

Proof. Take $\mathbf{x}$ to be core, $\deg_\varepsilon(\mathbf{x}) \geq m$, and $\mathbf{y} \in \mathcal{N}_\varepsilon(\mathbf{x})$ with $\deg_\varepsilon(\mathbf{y}) < m$ (non-core). Then $\mathbf{y}$ is directly density-reachable from $\mathbf{x}$ by Eqn. (2). Conversely, direct density-reachability from $\mathbf{y}$ to $\mathbf{x}$ would require $\mathbf{y}$ to be core, which it is not. ■

Lemma 2 (Transitivity of density-reachability)**.** *If* $\mathbf{x} \xrightarrow{\varepsilon,m} \mathbf{y}$ *and* $\mathbf{y} \xrightarrow{\varepsilon,m} \mathbf{z}$*, then* $\mathbf{x} \xrightarrow{\varepsilon,m} \mathbf{z}$*.*

Proof. Concatenate the witnessing chains from $\mathbf{x}$ to $\mathbf{y}$ and from $\mathbf{y}$ to $\mathbf{z}$ to obtain a valid chain from $\mathbf{x}$ to $\mathbf{z}$ in which each hop is direct density-reachability from a core point. ■

Lemma 3 (Symmetry of density-connectedness)**.** *Density-connectedness is symmetric and reflexive.*

Proof. Reflexivity holds by choosing $\mathbf{o} = \mathbf{x}$ in Eqn. (3). For symmetry, if $\mathbf{x}$ and $\mathbf{y}$ are density-connected via $\mathbf{o}$, then the same $\mathbf{o}$ witnesses that $\mathbf{y}$ and $\mathbf{x}$ are density-connected. ■

Proposition 1 (Clusters as maximal density-connected sets)**.** A *DBSCAN cluster* is any nonempty set $C \subseteq \mathcal{D}$ satisfying:

(C1) Maximality: If $\mathbf{x} \in C$ and $\mathbf{y}$ is density-reachable from $\mathbf{x}$, then $\mathbf{y} \in C$.

(C2) Connectivity: For any $\mathbf{x}, \mathbf{y} \in C$, $\mathbf{x}$ and $\mathbf{y}$ are density-connected.

Equivalently, for any core point $\mathbf{o}$, the set

$$\mathcal{R}(\mathbf{o}) \;=\; \{\mathbf{y} \in \mathcal{D} :\ \mathbf{o} \xrightarrow{\varepsilon,m} \mathbf{y}\}$$

is a cluster, and every cluster can be written as $\mathcal{R}(\mathbf{o})$ for some core $\mathbf{o}$.

Proof. (*First direction*) Fix a core $\mathbf{o}$ and set $C = \mathcal{R}(\mathbf{o})$. By definition of $\mathcal{R}(\mathbf{o})$ and transitivity of density-reachability, (C1) holds. For (C2), for any $\mathbf{x}, \mathbf{y} \in C$, both are density-reachable from $\mathbf{o}$, hence density-connected via $\mathbf{o}$. ■

Proof. (*Second direction*) Let C satisfy (C1)–(C2). Pick any $\mathbf{x} \in C$. If C contains no core point, then (C1) and (C2) force C to be empty (no border point can directly density-reach another without a core predecessor), contradiction. Thus C contains a core $\mathbf{o}$. For any $\mathbf{y} \in C$, (C2) gives that $\mathbf{x}$ and $\mathbf{y}$ are density-connected; in particular they are density-connected via some $\mathbf{o}' \in C$. Because C contains a core, we can choose a core witness (either $\mathbf{o}$ or another). Then $\mathbf{y}$ is density-reachable from that core, so $\mathbf{y} \in \mathcal{R}(\mathbf{o})$. Hence $C \subseteq \mathcal{R}(\mathbf{o})$. Maximality (C1) implies $\mathcal{R}(\mathbf{o}) \subseteq C$. Thus $C = \mathcal{R}(\mathbf{o})$. ■

1.5 Algorithmic Formulation (DBSCAN)

Given (ε, m), DBSCAN iteratively grows clusters by *region expansion* from yet-unvisited core points.

Region expansion. For a seed core point $\mathbf{o}$, define

$$\text{Seed} \leftarrow \{\mathbf{o}\}, \qquad C \leftarrow \emptyset.$$

While Seed $\neq \emptyset$:

1. Pop any $\mathbf{x}$ from Seed; add $\mathbf{x}$ to C if not already present.
2. Compute $\mathcal{N}_\varepsilon(\mathbf{x})$; if $\deg_\varepsilon(\mathbf{x}) \geq m$ (i.e., $\mathbf{x}$ is core), then

$$\text{Seed} \leftarrow \text{Seed} \ \cup \ \big(\mathcal{N}_\varepsilon(\mathbf{x}) \setminus C\big).$$

   (Neighbors of a core point are directly density-reachable and hence density-reachable from $\mathbf{o}$.)

When the queue empties, output the cluster C and mark all points in C as assigned.

Complete algorithm.

1. Initialize all points as *unvisited* and *unassigned*.
2. For each unvisited point $\mathbf{x}$:
   (a) Mark $\mathbf{x}$ visited; compute $\mathcal{N}_\varepsilon(\mathbf{x})$.
   (b) If $\deg_\varepsilon(\mathbf{x}) < m$, temporarily label $\mathbf{x}$ as *noise* and continue.
   (c) Else (core), start a new cluster C and perform region expansion from $\mathbf{x}$; mark all points discovered as members of C.
3. After all points processed, any point labeled noise that ended up within the expansion of some later cluster becomes a *border* point of that cluster; remaining noise labels are true outliers.

1.6 Correctness (Soundness and Completeness)

We sketch that DBSCAN enumerates exactly the clusters of Proposition 1.

Proposition 2 (Soundness)**.** Each set C returned by region expansion from a core $\mathbf{o}$ equals $\mathcal{R}(\mathbf{o})$ and thus satisfies (C1)–(C2).

Proof. Every point enqueued during expansion is either a core point in $\mathcal{N}_\varepsilon(\mathbf{x})$ or a neighbor of such a core; inductively, each added point is density-reachable from $\mathbf{o}$. Conversely, any point density-reachable from $\mathbf{o}$ lies along a chain of direct density-reachability starting at $\mathbf{o}$; the BFS/DFS-style expansion visits that chain by construction. Hence the output cluster is exactly $\mathcal{R}(\mathbf{o})$. ■

Proposition 3 (Completeness)**.** Every cluster C (satisfying (C1)–(C2)) is discovered by starting region expansion at any core $\mathbf{o} \in C$.

Proof. By Proposition 1, $C = \mathcal{R}(\mathbf{o})$ for some core $\mathbf{o} \in C$. Starting expansion at $\mathbf{o}$ generates exactly $\mathcal{R}(\mathbf{o})$, hence C. ■

1.7 Geometry, Metrics, and Parameter Choices

Metric. DBSCAN requires any metric d. Common choices: Minkowski $d_p(\mathbf{x}, \mathbf{z}) = (\sum_{j=1}^{d} |x_j - z_j|^p)^{1/p}$ with $p \in \{1, 2\}$; cosine or Mahalanobis may be used if transformed to a metric on the embedded space.

Choosing m and ε. A practical guideline is $m \in [d+1, 2d]$ for moderate d (to ensure local dimensional support). The *k-distance plot* (with $k = m$) orders points by their distance to the kth nearest neighbor; ε is chosen at the *elbow* where distances begin to rise sharply, separating dense cores from sparse regions.

1.8 Complexity and Data Structures

Naively, evaluating all neighborhoods costs $O(n^2)$ distance computations. With spatial indexing (kd-tree/ball-tree) for d moderate, range queries $\mathcal{N}_\varepsilon(\cdot)$ cost $O(\log n + q)$ on average (q neighbors returned), yielding $O(n \log n)$ expected time. In high d, exact range searching deteriorates (curse of dimensionality); approximate neighbors or dimensionality reduction can help.

1.9 Edge Cases and Robustness

- **Border points and cluster assignment.** If a border point lies within ε of core points from different clusters, its assignment depends on the visitation order; DBSCAN allows such ties to break arbitrarily without affecting cluster cores.
- **Varying density.** A single (ε, m) pair may fail when clusters have markedly different densities; extensions (e.g., HDBSCAN) address this via hierarchical persistence of density.
- **Degenerate $m = 1$.** Every point is core (since $\deg_\varepsilon(\mathbf{x}) \geq 1$), and clusters reduce to connected components of the ε-graph.

1.10 Graph-Theoretic View

Let $G_\varepsilon = (V, E)$ with $V = \mathcal{D}$ and undirected edges between pairs at distance $\leq \varepsilon$. Let $V_{\text{core}} = \{\mathbf{x} : \deg_\varepsilon(\mathbf{x}) \geq m\}$. Define a directed graph $\vec{G}$ with arcs $\mathbf{x} \to \mathbf{y}$ iff $\mathbf{x} \in V_{\text{core}}$ and $\{\mathbf{x}, \mathbf{y}\} \in E$. Then

$$\mathbf{x} \xrightarrow{\varepsilon, m} \mathbf{y} \iff \text{there exists a directed path } \mathbf{x} \rightsquigarrow \mathbf{y} \text{ in } \vec{G},$$

and DBSCAN clusters are exactly the undirected connected components of the subgraph induced by V_{core}, each *augmented* with their immediate non-core neighbors (border points). This yields the same sets as $\mathcal{R}(\mathbf{o})$.

1.11 Summary of Variables and Their Dimensions

- $\mathbf{x}_i \in \mathbb{R}^d$: ith data vector (column vector, dimension $d \times 1$). n: number of samples. d: feature dimension.
- $d(\cdot, \cdot)$: metric on $\mathbb{R}^d$ (scalar distance).
- $\varepsilon > 0$: neighborhood radius (scalar). $m \in \mathbb{N}$: `minPts`.
- $\mathcal{N}_\varepsilon(\mathbf{x}) \subseteq \mathcal{D}$: ε-neighbors of $\mathbf{x}$; $\deg_\varepsilon(\mathbf{x}) = |\mathcal{N}_\varepsilon(\mathbf{x})|$.

- Core point: $\deg_\varepsilon(\mathbf{x}) \geq m$. Border: non-core within ε of a core. Noise: neither core nor border.
- $\xrightarrow{\varepsilon,m}$: density-reachability (via chains of direct reachability from cores).
- Density-connectedness: existence of $\mathbf{o}$ from which both endpoints are density-reachable.
- Cluster $C = \mathcal{R}(\mathbf{o})$: all points density-reachable from a core $\mathbf{o}$; satisfies maximality and connectivity.

1.12 Summary

From first principles, DBSCAN formalizes clusters as maximal sets of points connected through chains of dense ε-neighborhoods anchored at core points. The algorithmic expansion from any core exactly enumerates such sets while classifying residual points as borders or noise. This density-based viewpoint confers robustness to cluster shape and outliers, with computational performance governed chiefly by efficient neighborhood queries and sensible (ε, m) selection.