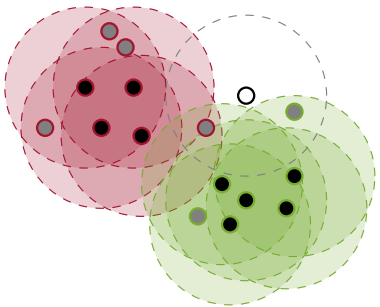
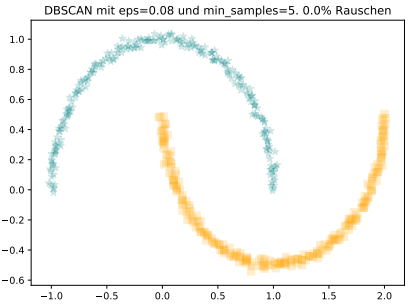
DBSCAN

Prof. Dr. Jörg Frochte

Maschinelles Lernen

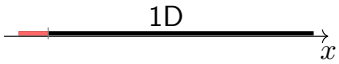


Hochschule Bochum
Bochum University
of Applied Sciences
Campus Velbert/Heiligenhaus



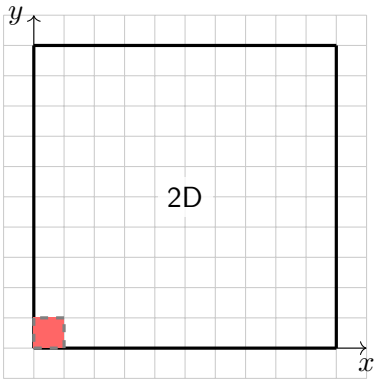
Fluch der Dimensionalität – Dichte

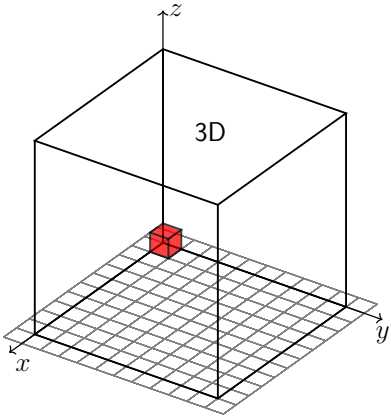
Wie groß ist der Anteil des Merkmalsraums, der 1/10 jedes Merkmals belegt? Und welcher Anteil an Daten liegt dort bei gleichverteilten Daten?



- In 1D: 10%
- In 2D: 1%
- In 3D: 0.1%
- usw.

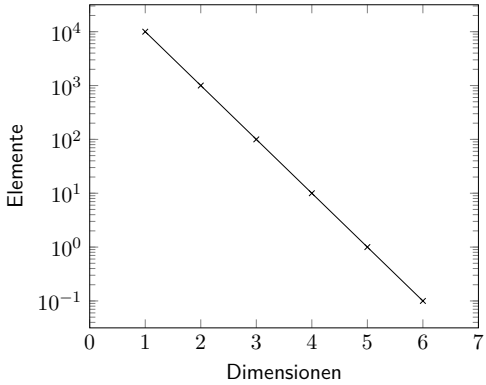
Mit steigender Dimension wird die Datenlage zunehmend dünner. Die Datendichte nimmt ab.





Fluch der Dimensionalität

- Stehen in 1D noch 10 000 Einträge in diesem Bereich zur Verfügung, um ein Regressionsmodell anzupassen, sind es in drei Raumdimensionen nur noch 100.
- Wie die Abbildung links zeigt kann man mit steigender Dimension schnell nicht mehr davon ausgehen z. B. 3 Nachbarn in einem Abstand von 0.1 zu finden.
- Wir haben bereits Verfahren kennengelernt um die Dimensionen zu reduzieren bzw. geeignete Merkmale auszuwählen.



Anteil des Teilgebietes mit der Kantenlänge 0.1 in zwei und drei Dimensionen

DBSCAN

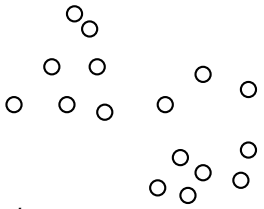
- **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise) gehört zur Klasse der dichte-basierten Clusteralgorithmen.
- Er wurde 1996 erstmals vorgestellt und weiterentwickelt.
- Wir besprechen die Grundform von DBSCAN, die man auch in Bibliotheken und Veröffentlichungen meistens vorfindet.
- DBSCAN kann mehrere Cluster erkennen, ohne dass wie bei den k-Means-Varianten zuvor die Anzahl der Cluster bekannt sein muss.
- Darüber hinaus werden Rauschpunkte im Laufe der Clusteranalyse erkannt, für das Clustering ignoriert und separat zurückgeliefert.
- Zu den Nachteilen im Vergleich zu k-Means gehört, dass er weniger gut skaliert bzgl. großer Datenmengen und weit stärker vom Fluch der Dimension betroffen ist.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5



Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

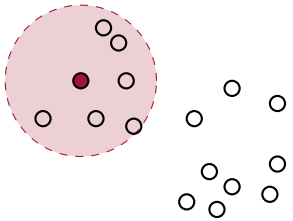
- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5



Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

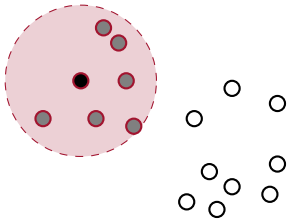
- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5



Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

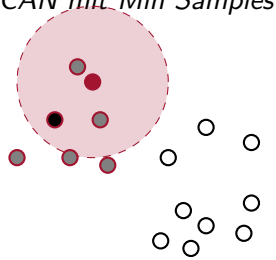
- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5



Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

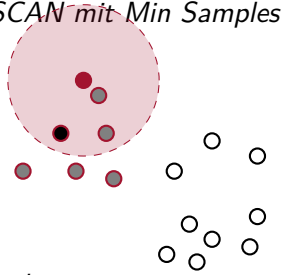
- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5



Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

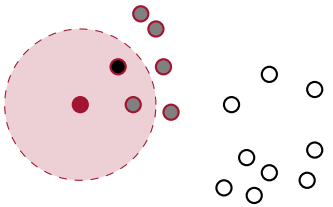
- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5



Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

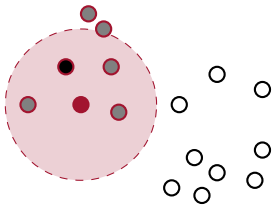
- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5



Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

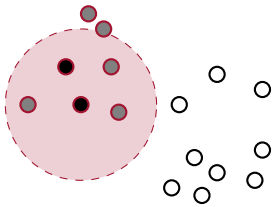
- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5



Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

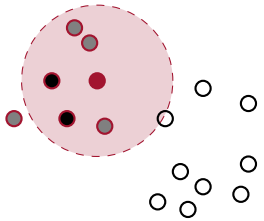
- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5



Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

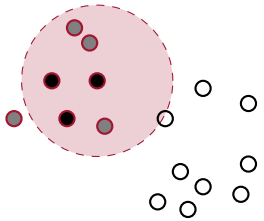
- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5



Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

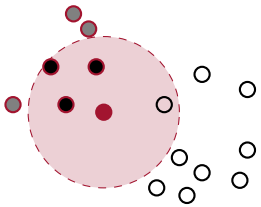
- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5



Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

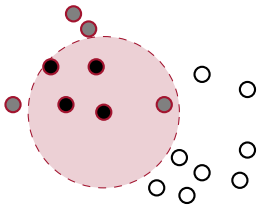
- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5



Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

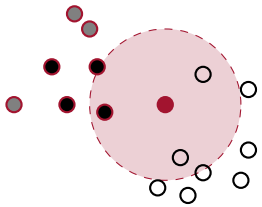
- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5



Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

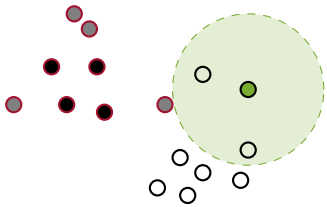
- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5



Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

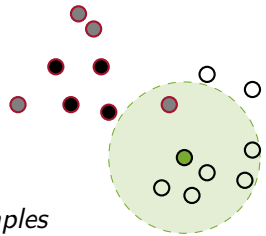
- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5



Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

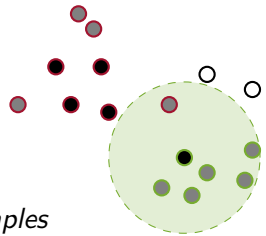
- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5



Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

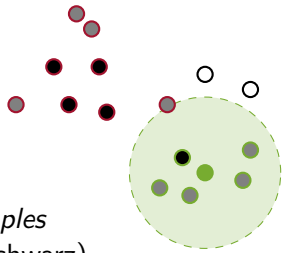
- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5



Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

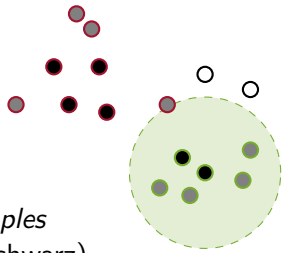
- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5



Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

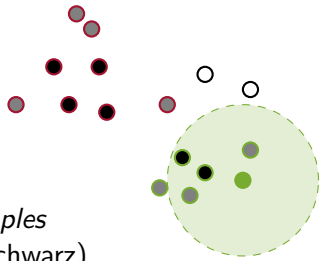
- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5



Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

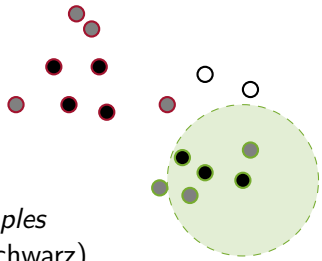
- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5



Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

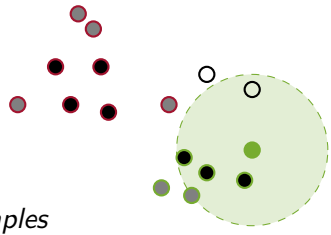
- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5



Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

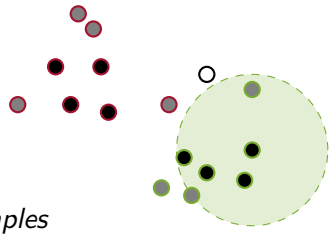
- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5



Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

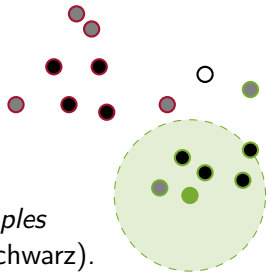
- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5



Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

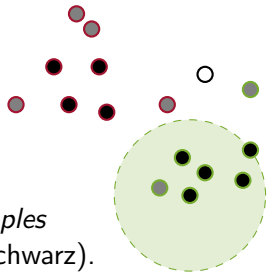
- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5



Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

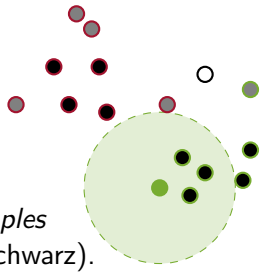
- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5



Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

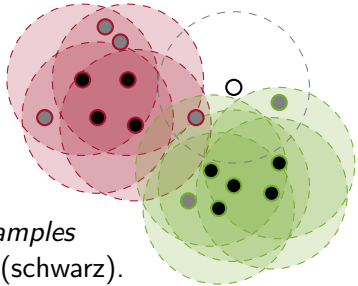
- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Algorithmus illustriert

Drei Arten von Punkten:

- Kernpunkte (Core Points)
- Randpunkte
- Rauschen bzw. Rauschpunkte (Noise)

Stationen im DBSCAN mit Min Samples = 5

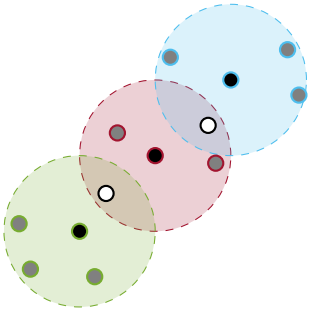


Welcher Art ein Punkt ist, hängt von zwei Parametern ab: ε und *Min Samples*.

- Liegen in der ε-Umgebung eines Punktes mindestens *Min Samples* Punkte – inklusive dem Punkt selbst – ist es ein Kernpunkt (schwarz).
- Die Punkte, die in der ε-Umgebung liegen, sind zumindest Randpunkte (grau).
- Im Beispiel liegen am Schluss zwei Cluster aus acht bzw. sieben Punkten vor, sowie ein Punkt, der als Rauschen eingeordnet wird. Die Zuordnung des Punktes zwischen den Clustern (hier: rot ●) hängt von der Reihenfolge der Untersuchung ab.

Verhalten der Randpunkte

- DBSCAN gilt als im Wesentlichen deterministisch und reihenfolgeunabhängig.
- Das bedeutet: Kernpunkte und die meisten Randpunkte werden unabhängig von der Reihenfolge der Datensätze gruppiert.
- Randpunkte, die in der Reichweite von Kernpunkten verschiedener Gruppen sind, werden reihenfolgenabhängig zugewiesen.
- Eine falsche Annahme ist, dass jeder Cluster mind. *Min Samples*-Objekte beinhaltet.



- In der Umgebung eines Kernpunktes liegen zwar *Min Samples*-Objekte, aber davon können einige einem anderen Cluster zugeordnet sein.

Pseudocode

```
1: function DBSCAN(D, ε, minSamples)
2:     C = 0
3:     for all unbesuchte P ∈ D do
4:         Markiere P als besucht
5:         N = {x ∈ D | ‖x − P‖ < ε}
6:         if #N < minSamples then
7:             Markiere P als Rauschen
8:         else
9:             C = C + 1
10:            Füge P dem Cluster Nr. C hinzu
11:            expandCluster(N, C, ε, minSamples)
12:        end if
13:    end for
14: end function
```

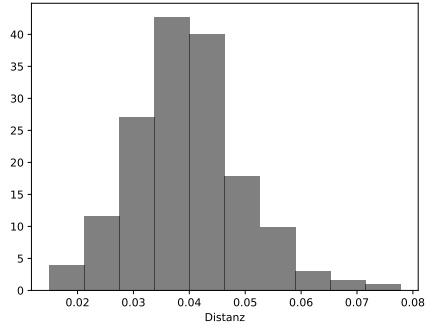
```
15: function expandCluster(N, C, ε, minSamples)
16:     for all P′ ∈ N do
17:         if P′ ist noch nicht besucht then
18:             Markiere P′ als besucht
19:             N′ = {x ∈ D | ‖x − P′‖ < ε}
20:             if #N′ >= minSamples then
21:                 N = N ∪ N′
22:             end if
23:         end if
24:         if P′ ist noch keinem Cluster zugewiesen then
25:             Füge P′ dem Cluster Nr. C hinzu
26:                 ▷ ggf. Zuordnung als Rauschen aufheben
27:         end if
28:     end for
29: end function
```

Laufzeitverhalten

- Wie man am Pseudocode schon erkennen kann, ist DBSCAN auf der obersten Ebene nur von linearer Komplexität.
- Das bedeutet, jeder Punkt wird im Wesentlichen nur einmal besucht.
- Das Problem liegt in den Zeilen 5 und 19, da die Berechnung der ε-Umgebung nicht von linearer Komplexität ist, sondern im Allgemeinen quadratisch.
- Für die Umsetzung kann man den kd-Baum verwenden, der für geringe Dimensionen gute Dienste leistet und den Algorithmus bzgl. dem Auffinden der Nachbarn deutlich beschleunigt.
- Neben den aus dem Pseudocode bekannten Parametern `eps` und *Min Samples* ist die Norm noch ein Design-Aspekt.
- Dazu kommen Parameter für den kd-Tree wie z. B. `leafSize`.

Wahl des Parameters ε

- Für den praktischen Einsatz von DBSCAN gibt es ein Problem: Die Wahl der beiden Parameter.
- Hier ist erneut der `kdTree` in der praktischen Umsetzung hilfreich.
- Es kann eine Funktion geschrieben werden, die für jeden Datenbankeintrag die minimale Distanz, in der sich *Min Samples*-Punkte befinden, berechnet.
- Dies gibt z. B. in einem Histogramm ein Gefühl für die Dichte in der Datenwolke; es ist nicht automatisch eine Empfehlung für ein spezielles ε.

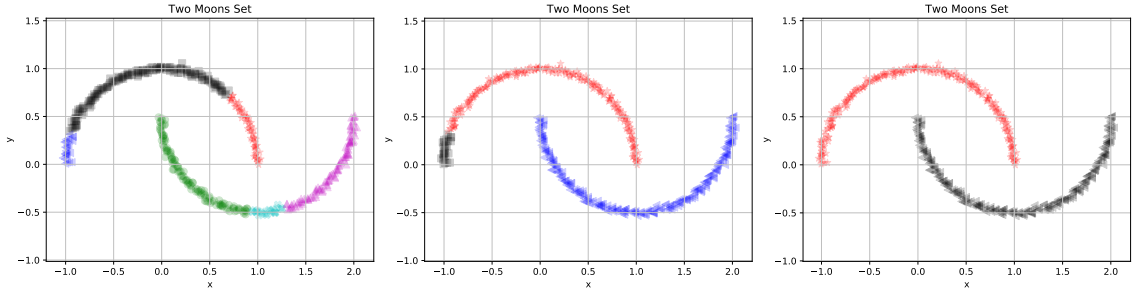


Beispielhistogram für das Two Moons Set

- Durch das Histogramm ist klar, dass der Wertebereich von 0.02 bis 0.08 sinnvoll ist.
- Innerhalb des Bereichs ist die Wahl nicht klar.

- Die Rauschpunkte geben ein Gefühl für die Wahl von ε, falls die Messungenauigkeit bekannt ist.
- Wir erwarten Rauschen unter 1%, daher probieren wir die letzten Fälle.
- Für 0.08 entspricht die Gruppierung den Erwartungen.

ε	Clusterzahl	Rauschen
0.02	5	95 %
0.03	24	66 %
0.04	38	16 %
0.05	11	1.4 %
0.06	6	0.0 %
0.07	3	0.0 %
0.08	2	0.0 %



Clusterbildung mit DBSCAN für $\varepsilon = 0.06$ (links) und 0.07 (rechts)

Clusterbildung mit DBSCAN auf den Testproblemen