

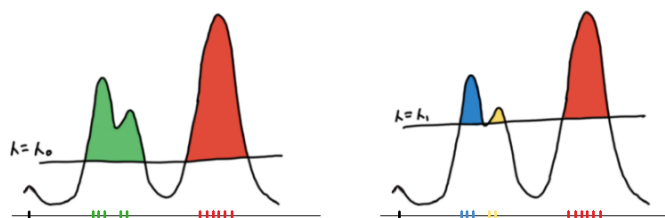
### Co bude v dnešní přednášce

- Shlukování pomocí hustoty
- DBSCAN
- Evaluace shlukování pomocí Silhouette skóre
- Asociační pravidla

## 45 Shlukování pomocí hustoty

### Úvod do shlukování pomocí hustoty

- V dnešní přednášce se zaměříme na shlukování, které využívá odhad hustoty rozdělení dat na prostoru  $\mathcal{X}$  možných hodnot příznaků.
- Z pravděpodobnostního pohledu chápeme pozorovaná data jako **realizace náhodného vektoru**  $\mathbf{X} = (X_1, \dots, X_p)^T$ .
- Pokud budeme mít nějakým způsobem odhadnutou hustotu pravděpodobnosti  $f_{\mathbf{X}}(\mathbf{x}) \equiv f_{\mathbf{X}}(x_1, \dots, x_p)$  (resp. něco, co jí je úměrné), můžeme ji využít k získání shlukování včetně identifikace bodů, které do žádných shluků nepatří.
- Shluky můžeme získat jako **souvislé oblasti**, ve kterých odhad hustoty překročí nějakou zvolenou hranici.
- Body mimo tyto oblasti pak označíme jako **šum**.

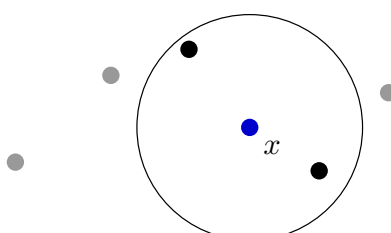


## 46 DBSCAN

### DBSCAN - úvodní pojmy

- Jedním s aktuálně nejpoužívanějších algoritmů shlukování, který je implicitně založen na principu odhadu hustoty, je algoritmus *DBSCAN*, což je zkratka angl. *density-based spatial clustering of applications with noise*, [Ester et al. (1996)].
- Připravme si nyní základní pojmy, na kterých DBSCAN stojí.
- Uvažujme metrický prostor  $\mathcal{X}$  s metrikou  $d(x, y)$  ze kterého pochází dataset  $\mathcal{D}$  a parametry  $\varepsilon > 0$  a  $\text{MinPts} \in \mathbb{N}^+$ .
- Definujme  $\varepsilon$  okolí bodu  $x$  v  $\mathcal{D}$  (angl.  $\varepsilon$ -neighborhood) jako množinu

$$N_\varepsilon(x) = \{y \in \mathcal{D} | d(x, y) \leq \varepsilon\}.$$



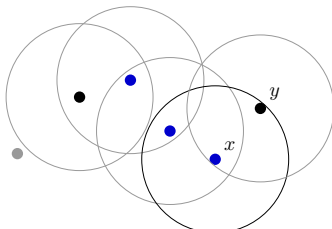
Znázornění  $\varepsilon$  okolí.

## DBSCAN - klíčové body, přímá dosažitelnost

- Bod  $x \in \mathcal{D}$  je *klíčový bod* (angl. *core point*), jestliže v jeho  $\varepsilon$  okolí v  $\mathcal{D}$  je alespoň  $\text{MinPts}$  bodů,

$$|N_\varepsilon(x)| \geq \text{MinPts}.$$

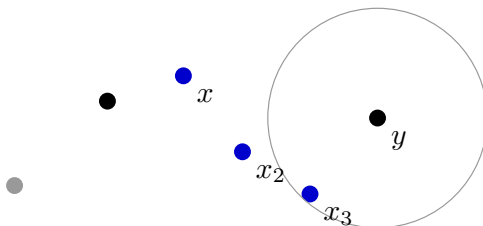
- Bod  $y \in \mathcal{D}$  je *přímo dosažitelný* (angl. *directly density-reachable*) z bodu  $x \in \mathcal{D}$ , jestliže  $x$  je klíčový bod a  $y \in N_\varepsilon(x)$ .
- Relace přímé dosažitelnosti je symetrická pro dvojici klíčových bodů, je ale nesymetrická pro tzv. *okrajový bod* (angl. *border point*), což je bod, který není klíčový, ale je přímo dosažitelný z klíčového bodu.



Pro  $\text{MinPts} = 3$  je bod  $y$  přímo dosažitelný z  $x$ . Všechny klíčové body jsou modré, všechny okrajové body jsou černé.

## DBSCAN - dosažitelnost

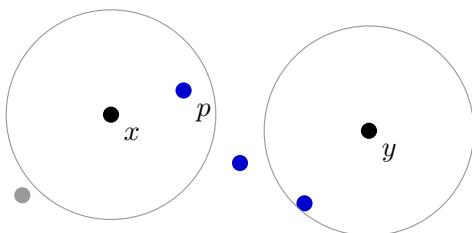
- Bod  $y \in \mathcal{D}$  je *dosažitelný* (angl. *density-reachable*) z bodu  $x \in \mathcal{D}$ , pokud existuje posloupnost  $x_1, x_2, \dots, x_n \in \mathcal{D}$  bodů tak, že  $x_1 = x$ ,  $x_n = y$  a pro každé  $i = 1, \dots, n - 1$  je  $x_{i+1}$  přímo dosažitelný z bodu  $x_i$ .
- Z toho plyne, že všechny body po cestě **kromě posledního** musí být klíčové body.



Pro  $\text{MinPts} = 3$  je bod  $y$  dosažitelný z bodu  $x$ .

## DBSCAN - spojenost

- Bod  $y \in \mathcal{D}$  je *spojený* (angl. *density-connected*) s bodem  $x \in \mathcal{D}$ , jestliže existuje (klíčový bod)  $p \in \mathcal{D}$ , tak, že  $x$  i  $y$  jsou dosažitelné z bodu  $p$ .
- Relace spojenosti je zjevně symetrická. Pro klíčové body je také tranzitivní.
- Jestliže jsou dva body spojené a jeden z nich je klíčový bod, tak ten druhý je z toho prvního dosažitelný.



Pro  $\text{MinPts} = 3$  je bod  $y$  spojený s bodem  $x$ .

## DBSCAN - shluky, šum

Shluk nyní definujeme jako **maximální množinu spojených** bodů.

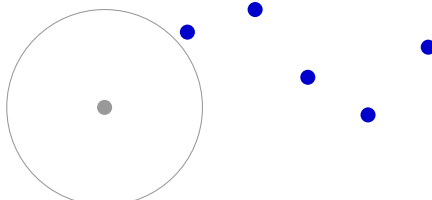
### Definice 46.1

Shluk (angl. *cluster*)  $C$  je podmnožina  $\mathcal{D}$  obsahující alespoň jeden klíčový bod tak, že:

- Pro každé  $x, y \in \mathcal{D}$  platí, že když  $x \in C$  a  $y$  je dosažitelný z  $x$ , pak  $y \in C$  (**maximalita**).
- Pro každé  $x, y \in C$  je  $x$  spojený s  $y$  (**souvislost**).

Označme  $C_1, \dots, C_k$  množinu všech shluků v  $\mathcal{D}$  (vzhledem k  $\varepsilon$  a  $k$ ). Množinu  $N$  bodů z  $\mathcal{D}$ , které nejsou v žádném ze shluků nazýváme *šumem* (angl. *noise*),

$$N = \mathcal{D} \setminus \bigcup_{i=1}^k C_i$$



Body ve shluku jsou modré, bod šumu je šedivý.

### Poznámky k definici shluků

- Shluk  $C$  vždy obsahuje nějaký klíčový bod  $p \in C$ , ze kterého jsou dosažitelné všechny body v jeho  $\varepsilon$  okolí, kterých je alespoň MinPts.

Právě jsme tedy ukázali, že každý shluk obsahuje alespoň MinPts bodů.

- Každý bod ve shluku  $C$  je spojený se všemi klíčovými body v  $C$  a tedy je z libovolného klíčového bodu dosažitelný. Shluk je tedy tvořen všemi body, které jsou dosažitelné z libovolného klíčového bodu v  $C$ .
- To nám dává návod, jak může algoritmus tvorby shluků fungovat. Najdeme klíčový bod a vytvoříme k němu shluk jako množinu všech z něho dosažitelných bodů (které ještě nejsou v jiném shluku).

### Abstraktní popis algoritmu

Nyní si abstraktním způsobem popíšeme běh algoritmu DBSCAN.

#### Algoritmus DBSCAN

*Nalezení klíčových bodů:* Spočítáme  $\varepsilon$  okolí každého bodu a identifikujeme klíčové body.

*Vytvoření zárodků shluků:* Spojíme sousední (přímo dosažitelné) klíčové body do shluků.

*Pro každý bod, který není klíčový:*

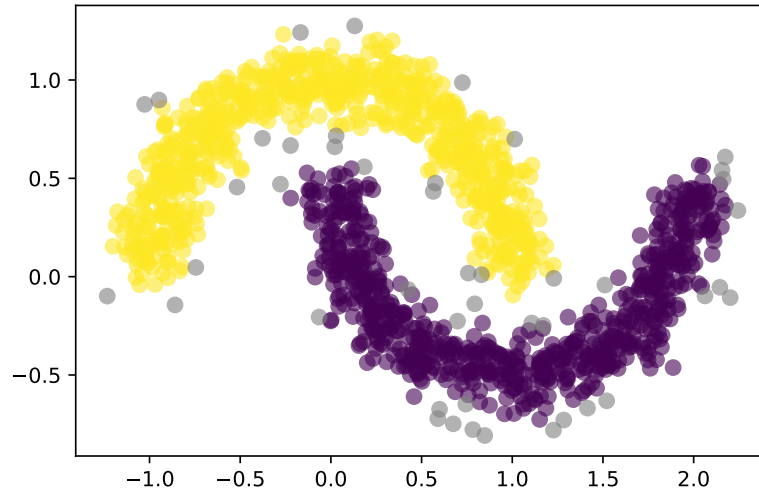
- Přidáme do shluku podle klíčového bodu v jeho okolí.
- Pokud takový neexistuje, přidáme mezi šum.
- Okrajový bod, který má ve svém  $\varepsilon$  okolí klíčové body z různých (zárodků) shluků spadne do prvního z těchto shluků, ke kterému se algoritmus dostane.
- Finálním výsledkem v takovém případě budou shluky, které **nesplňují podmínku maximality** v předchozí definici, protože okrajový bod bude pouze v jednom shluku a nikoliv ve všech, kde by dle maximality měl být.
- Lze ukázat ([Schubert et al. (2017)]), že složitost v nejhorším případě je  $O(n^2)$ . V mnoha reálných situacích se ale lze dostat na  $O(n \log n)$ .

### DBSCAN - volba parametrů

- Zaměříme se nyní krátce na volbu parametrů algoritmu DBSCAN.
- Ukazuje se, že parametr MinPts je mnohem méně důležitý než parametr  $\varepsilon$ . Obvykle dobrou volbou jsou hodnoty okolo 4 – 6 (někdy bývá doporučováno  $2 \dots p$ , kde  $p$  je počet příznaků).
- Pro parametr  $\varepsilon$  bývá doporučováno volit co nejmenší hodnotu s tím, že lze například brát průměrnou vzdálenost bodů k jejich nejbližšímu  $(2 \cdot p - 1)$ tému sousedovi.

- Také můžeme sledovat velikost šumu. Uvádí se, že obvykle by poměr šumu měl být mezi 1% a 30%.
- Dále je dobré sledovat velikost největšího shluku. Pokud jeho velikost překračuje 50% velikosti datasetu, bývá vhodné zmenšit  $\varepsilon$ , případně použít nějaký pokročilejší algoritmus (např. HDBSCAN).
- Více detailů k volbě parametrů najdete v [Schubert et al. (2017)] nebo v [Sander et al. (1998)].

### DBSCAN - ukázka a poznámky



Znázornění shluků a šumu.

- Mezi hlavní výhody DBSCAN (a obecně metod založených na hustotě) patří možnost nalezení nekonevných shluků.
- Další výhodou je snížená citlivost na odlehlé hodnoty - které jsou označeny jako šum.

## 47 Silhouette skóre

### Evaluace pomocí Silhouette skóre (1/3)

- Vraťme se nyní k obecné úloze shlukování a ukažme si jednoduchou a používanou metodu pro evaluaci shlukování pomocí metody *Silhouette* [Rousseeuw, P. (1987)], kterou lze využít i k určení optimálního počtu shluků.
- Uvažujme shlukování  $\mathcal{D} = C_1 \cup \dots \cup C_k$  na metrickém prostoru  $\mathcal{X}$  s metrikou  $d(x, y)$  a pro libovolný bod  $x \in \mathcal{D}$  označme  $j(x)$  index shluku do kterého  $x$  patří, tj.  $x \in C_{j(x)}$ .
- Pro bod  $x \in \mathcal{D}$  nyní:

- Spočteme  $a(x)$  jako průměrnou vzdálenost bodu  $x$  od všech ostatních bodů ve stejném shluku (vnitřní rozdílnost, angl. within dissimilarity)

$$a(x) = \frac{1}{|C_{j(x)}| - 1} \sum_{y \in C_{j(x)}, y \neq x} d(x, y).$$

- Pro každý další shluk  $C_i$ ,  $i \neq j(x)$  spočteme průměrnou vzdálenost bodu  $x$  od všech bodů v tomto shluku

$$d(x, C_i) = \frac{1}{|C_i|} \sum_{y \in C_i} d(x, y)$$

- Spočteme  $b(x)$  jako minimum z těchto průměrných vzdáleností od ostatních shluků (sousední rozdílnost, angl. between dissimilarity)

$$b(x) = \min_{i \neq j(x)} d(x, C_i)$$

### Evaluace pomocí Silhouette skóre (2/3)

Finální silhouette skóre bodu  $x \in \mathcal{D}$  získáme vztahem

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}.$$

- Jestliže máme pouze jeden shluk, položíme  $s(x) = 0$ .
- Z definice vidíme, že vždy platí

$$-1 \leq s(x) \leq 1.$$

- Hodnota  $s(x)$  blízka 1 znamená, že vnitřní rozdílnost  $a(x)$  je mnohem menší než sousední rozdílnost  $b(x)$ , což značí, že ten bod je dobře zatříděn.
- Hodnota  $s(x)$  okolo 0 znamená, že  $a(x)$  je podobně velké jako  $b(x)$  a tedy bod  $x$  je někde na okraji svého shluku a sousední shluk je blízko. Čili by ten bod klidně mohl patřit i do toho druhého shluku.
- Hodnota  $s(x)$  blízka  $-1$  znamená, že vnitřní rozdílnost  $a(x)$  je mnohem větší než sousední rozdílnost  $b(x)$  a tedy bod  $x$  by měl spíše patřit do toho sousedního shluku než do toho svého. Je tedy špatně přiřazen.

### Evaluace pomocí Silhouette skóre (3/3)

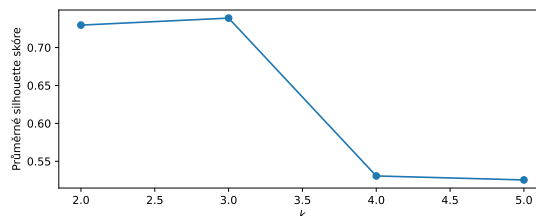
- Z ohodnocení  $s(x)$  každého z bodů  $x$  můžeme získat jednak průměrné skóre  $s_i$  pro každý shluk  $C_i$  zvlášť

$$s_i = \frac{1}{|C_i|} \sum_{x \in C_i} s(x),$$

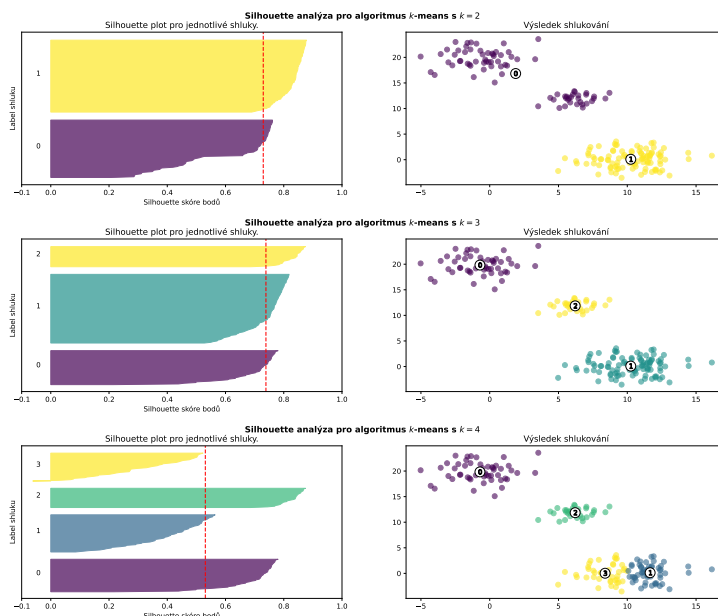
- a pak také průměrné skóre  $s$  pro celé shlukování

$$s = \frac{1}{|D|} \sum_{x \in D} s(x).$$

- Čím vyšší dostaneme hodnotu, tím jsou body ve shlucích správněji umístěné a tedy je celé shlukování lepší.
- Porovnání skóre pro různé počty shluků můžeme použít k nalezení vhodného počtu shluků jako hodnoty, při které je skóre  $s$  maximální.



### Znázornění silhouette skóre



## 48 Asociační pravidla

### Analýza asociačních pravidel

- *Analýza asociačních pravidel* je jedna ze standardních a oblíbených metod pro dobývání znalostí z komerčních databází.
- Obecným cílem je nalézt společné hodnoty příznaků  $\mathbf{X} = (X_1, \dots, X_p)^T$ , které se v databázi nejčastěji vyskytují.
- V zásadě jde přesně o jeden z hlavních cílů nesupervizovaného učení, kdy chceme nalézt oblasti prostoru, kde se data vyskytují s velkou pravděpodobností.
- V nejčastěji používaném zjednodušení se zabýváme pouze oblastmi, které jsou ve tvaru kartézského součinu pro jednotlivé příznaky, tj. chceme, aby pravděpodobnost

$$P\left(\bigcap_{j=1}^p (X_j \in s_j)\right),$$

kde  $s_j$  je podmnožina hodnot příznaku  $X_j$ , byla **relativně velká**.

- Průnik  $\bigcap_{j=1}^p (X_j \in s_j)$  se v takovém případě nazývá *konjunktivní pravidlo* (angl. *conjunctive rule*).

### Analýza nákupního košíku

- Analýza asociačních pravidel je nejčastěji aplikována v případě binárních příznaků,  $X_j \in \{0, 1\}$ , kdy se pak nazývá *analýza nákupního košíku* (angl. *market basket analysis*).
- Pro oblast, kterou hledáme ve tvaru kartézského součinu se v tomto případě používá ještě další omezení, a to, že  $s_j$  je buď jednoprvková množina  $\{1\}$  nebo všechny možnosti daného příznaku  $X_j$  (v tu chvíli příznak vypadne).
- Ekvivalentně tedy hledáme množinu indexů  $\mathcal{K} \subset \{1, \dots, p\}$  tak, že

$$P(\bigcap_{j \in \mathcal{K}} (X_j = 1)) = P\left(\prod_{j \in \mathcal{K}} X_j = 1\right)$$

je **relativně velká**.

- Množina  $\mathcal{K}$  se pak nazývá *množina položek* (angl. *item set*).
- Relativní velikost položek v datasetu, které danou množinu položek obsahují se značí  $T(\mathcal{K})$  a nazývá *podpora* (angl. *support*) množiny položek  $\mathcal{K}$  a odpovídá odhadu výše uvedené pravděpodobnosti:

$$T(\mathcal{K}) = \hat{P}(\bigcap_{j \in \mathcal{K}} (X_j = 1)) = \frac{1}{N} \sum_{i=1}^N \prod_{j \in \mathcal{K}} x_{i,j}.$$

### Asociační pravidla (1/2)

- Při provádění analýzy nákupního košíku hledáme všechny množiny položek, pro které je **podpora větší** než nějaká zvolená mez  $t$ :

$$\{\mathcal{K}_\ell | T(\mathcal{K}_\ell) > t\}.$$

- K nalezení řešení se používá efektivní algoritmus (případně jeho novější vylepšení), který se nazývá **Apriori algoritmus** ([Agrawal, Srikant (1994)]).
- Pro každou množinu položek  $\mathcal{K}$ , kterou takto získáme, dále hledáme vhodné rozložení na dvě disjunktní podmnožiny  $A$  a  $B$ ,  $A \cup B = \mathcal{K}$ , které budeme nazývat *asociační pravidlo* (angl. *association rule*) a značit

$$A \Rightarrow B.$$

- První položka  $A$  asociačního pravidla se nazývá *předpoklad* (angl. *antecedent*) a druhá položka  $B$  se nazývá *závěr* (angl. *consequent*).
- Podpora  $T(A \Rightarrow B)$  pravidla  $A \Rightarrow B$  je definována jako podpora sjednocení  $\mathcal{K} = A \cup B$ .
- *Spolehlivost* (angl. *confidence*)  $C(A \Rightarrow B)$  pravidla  $A \Rightarrow B$  je definována jako podpora pravidla podělená podporou předpokladu  $A$ :

$$C(A \Rightarrow B) = \frac{T(A \Rightarrow B)}{T(A)},$$

což odpovídá odhadu podmíněné pravděpodobnosti  $P(B|A)$ .

## Asociační pravidla (2/2)

- Asociační pravidla jsou volena tak, aby **spolehlivost byla větší** než nějaká zvolená mez  $c$ .
- **Finálním výstupem asociační analýzy** pravidel je množina asociačních pravidel, které splňují

$$T(A \Rightarrow B) > t \quad \text{a} \quad C(A \Rightarrow B) > c.$$

- U nalezených asociačních pravidel se dále často měří *zdvih* (angl. *lift*) definovaný jako

$$L(A \Rightarrow B) = \frac{C(A \Rightarrow B)}{T(B)},$$

který odpovídá odhadu podílu  $P(B|A)/P(B)$  což znamená, kolikanásobně je větší  $P(B|A)$  oproti  $P(B)$ .

- Někdy se také měří *pokrytí* (angl. *coverage*) definované jako

$$\text{Coverage}(A \Rightarrow B) = \frac{T(A \Rightarrow B)}{T(B)},$$

což odpovídá odhadu podmíněné pravděpodobnosti  $P(A|B)$ .

- Pokrytí tedy indikuje, jak často lze závěr vyložit coby důsledek předpokladu.

### Asociační pravidla - příklady

Klasickým příkladem asociačního pravidla je

$$\{\text{párky}\} \Rightarrow \{\text{hořčice, chléb}\}.$$

- **Podpora** 0.06 znamená, že v celém datasetu se trojice položek  $\{\text{párky, hořčice, chléb}\}$  vyskytuje v 6% případech.
- Pokud se **párky** vyskytují v datasetu v 8% případech, bude **spolehlivost**  $0.06/0.08 = 0.75$ , což znamená, že když si zákazník koupil **párky**, v 75% případech si koupil také **hořčici** a **chléb**.
- Pokud je dvojice **hořčice** a **chléb** v datasetu v 15% případech, **zdvih** bude  $0.75/0.15 = 5$ .
- **Pokrytí** v takovém případě bude  $0.06/0.15 = 0.4$ . Čili ve 40% případech lze **hořčici** a **chléb** uvažovat jako důsledek koupě **párků**.

Nevýhoda omezení na podporu je, že pravidla s velkou hodnotou spolehlivosti i zdvihu ale s nízkou podporou, jako např.

$$\{\text{doutník}\} \Rightarrow \{\text{rum}\},$$

nebudou nalezeny.

## ChangeLog

Verze	Datum	Autor	Log
1.0	4.12.2024	DV	Výchozí verze pro rok 2024/2025.
1.0	7.12.2023	DV	Výchozí verze pro rok 2023/2024.
1.0	15.12.2022	DV	Výchozí verze pro rok 2022/2023.