Matematika pro zpracování znalostí

Relace

(n-ární) **relace** R je podmnožina kartézského součinu množin $R \subseteq A_1 \times \ldots \times A_n = \{(a_1, \ldots, a_n) \mid (\forall i = 1, \ldots, n) : a_i \in A_i\}$.

Dělení podle arity:

- unární $R_1 \subseteq A_1$
- binární $R_2 \subseteq A_1 \times A_2$
  - homogenní $A_1 = A_2$
  - heterogenní $A_1 \neq A_2$
- ternární
- n-ární

Relační systém je uspořádaná dvojice nosičů a relací $(\{A_1, \ldots, A_n\}, \{R_1, \ldots, R_m\})$.

Triviální relace:

- prázdná $R = \emptyset$
- identická $\mathrm{id}_A = \{(a, a) \mid a \in A\}$
- úplná $R = A_1 \times \ldots A_n$

Binární relace

Binární relaci $R_1 \subseteq X \times Y$ nazveme **inverzní** k $R_2 \subseteq Y \times X \iff (\forall x \in X)(\forall y \in Y) : xR_1y = yR_2x$. Inverzní relace odpovídá otočení hran v orientovaném grafu (transpozice matice sousednosti).

Buď $R_1 \subseteq X \times Y$, $R_2 \subseteq Y \times Z$ a $R \subseteq X \times Z$ relace. **Složením relací** R_1 a R_2 rozumíme $(x, z) \in (R_1 \circ R_2) \Leftrightarrow (\exists y \in A_2) : xR_1y \wedge yR_2z$.

- není komutativní
- je asociativní

Vlastnosti relací $(\forall x, y, z)$:

- **RE**flexivita xRx
- **IR**eflexivita $(x, x) \notin R$
- **SY**metrie $xRy \Rightarrow yRx$
- **AS**ymetrie $(x, y) \in R \Rightarrow (y, x) \notin R$
- **AN**tisymetrie $xRy \wedge yRx \Rightarrow x = y$
- **TR**anzitivita $xRy \wedge yRz \Rightarrow xRz$
- **ÚP**lnost $xRy \vee yRx$ $(\mathbb{A} + \mathbb{A}^T = \mathbf{1}^{n,n})$
- **SO**uvislost $x \neq y \Rightarrow xRy \vee yRx$

ASymetrie $\Rightarrow$ **AN**tisymetrie.

TRanzitivita – každá cesta délky 2 má i zkratku.

Uzávěry binárních relací

- **reflexivní** $\mathrm{Re}(R) = R \cup \mathrm{id}_X$
- **symetrický** $\mathrm{Sy}(R) = R \cup R^{-1}$
- **tranzitivní** $\mathrm{Tr}(R) = \bigcup_{i=1}^{|X|} R^i$
- **tolerantní** $\mathrm{Sy}(\mathrm{Re}(R))$
- **ekvivalentní** $\mathrm{Tr}(\mathrm{Sy}(\mathrm{Re}(R)))$

Speciální binární relace

- **tolerance** - **ReSy**, prvek $a \in X$ je v toleranci $\tau \iff [a]_\tau = \{x \in X \mid (x, a) \in \tau\}$
- **ekvivalence** - **ReSyTr**, prvek $a \in X$ je v ekvivalenci $\varepsilon \iff [a]_\varepsilon = \{x \in X \mid (x, a) \in \varepsilon\}$. $\varepsilon_1 \cap \varepsilon_2$ je ekvivalence a $\varepsilon^n = \varepsilon$.
- **uspořádání** (částečné / úplné (=lineární, totální), ostré / neostré)
  - '$<$' $\iff$ IR, AS, TR (+SO)
  - '$\leq$' $\iff$ RE, AN, TR (+ÚP)

Pokrytí a rozklad

Systém množin $\{[a_i]_{\tau/\varepsilon} \mid a_i \in A\}$ tvoří **pokrytí/rozklad** $A \iff$

1. $[a_i]_\tau \neq \emptyset$
2. $\bigcup_i [a_i]_\tau = A$

1. $[a_i]_\varepsilon \neq \emptyset$
2. $\bigcup_i [a_i]_\varepsilon = A$
3. $\forall i \neq j \Rightarrow [a_i]_\varepsilon \cap [a_j]_\varepsilon = \emptyset$

Faktorová množina množiny X podle ekvivalence ε je $X/\varepsilon = \{[a]_\varepsilon \mid a \in A\}$.

Zobrazení $\kappa_\varepsilon \colon X \to X/\varepsilon$ nazýváme jako **přirozené / kanonické**.

Zobrazení

Binární relaci $f \subseteq X \times Y$ nazveme **zobrazením** $f \colon X \to Y \iff (\forall x \in X)(\forall y_1, y_2 \in Y) : [f(x) = y_1 \wedge f(x) = y_2] \Rightarrow y_1 = y_2$. Pokud navíc $(\exists y \in Y) : f(x) = y$, tak mluvíme o **totálním zobrazení**.

Vlastnosti zobrazení

Zobrazení $f \colon X \to Y$ nazveme:

- **injektivní** (prosté) $\iff (\forall x_1, x_2 \in X) : x_1 \neq x_2 \Rightarrow f(x_1) \neq f(x_2)$
- **surjektivní** (na) $\iff (\forall y \in Y)(\exists x \in X) : f(x) = y$
- **bijektivní** (vzájemně jednoznačné) $\iff$ f je zároveň injektivní i prosté

Operace

n-ární **operace na množinách** je $f \colon A_1 \times \ldots \times A_n \to B$.

Direktní součin

Direktní součin $A \otimes B$ relačních struktur (A, R), (B, S) s binárními homogenními relacemi je kartézský součin $A \times B$ s relací T takovou, že $(\forall (a_1, b_1), (a_2, b_2) \in A \times B) : (a_1, b_1)T(a_2, b_2) \iff a_1Ra_2 \wedge b_1Sb_2$.

Významné prvky v uspořádání

Buď relační systém $(A, \leq)$ (uspořádaná množina) a buď $M \subseteq A$. Pak prvek

- $a \in M$ je minimální prvek $M \iff (\nexists x \in M) : x < a$
- $a \in M$ je maximální prvek $M \iff (\nexists x \in M) : x > a$
- $a \in M$ je nejmenší (porovnatelný) prvek $M \iff (\forall x \in M) : a \leq x$
- $a \in M$ je největší (porovnatelný) prvek $M \iff (\forall x \in M) : a \geq x$
- $a \in A$ je dolní závora (LB) $M \iff (\forall x \in M) : a \leq x$
- $a \in A$ je horní závora (UB) $M \iff (\forall x \in M) : a \geq x$
- $\inf(M) = \max\{LB(M)\}$
- $\sup(M) = \min\{UB(M)\}$

Prvky závor musí být porovnatelné!

Hasseův diagram

1. Odstraním smyčky.
2. Odstraním tranzitivní hrany.
3. Kreslím šipky směrem vzhůru $\uparrow$.

Uzávěrový systém

Uzávěrový systém $\mathcal{C} \subseteq 2^A$ obsahuje A a je uzavřený vůči průniku svých prvků, tzn.

1. $A \in \mathcal{C}$,
2. $B_i \in \mathcal{C} \Rightarrow \bigcap_i B_i \in \mathcal{C}$

Uzávěrový operátor je zobrazení $\alpha \colon 2^A \to 2^A$, které splňuje:

1. extenzionalitu $(\forall B \subseteq A) : B \subseteq \alpha(B)$
2. idempotentnost $(\forall B \subseteq A) : \alpha(\alpha(B)) \subseteq \alpha(B)$
3. monoticitu $(\forall B \subseteq C \subseteq A) : \alpha(B) \subseteq \alpha(C)$

Algebry

Algebraický systém je uspořádaná dvojice nosičů a operací $(\{A_i\}, \{f_j\})$

Vlastnosti binárních operací. Buď $\circ$ binární homogenní operace na mže A.

- Uzavřenost (**UZ**)
  $(\forall a, b \in A)(\exists c \in A) : a \circ b = c$
- Asociativita (**AS**)
  $(\forall a, b, c \in A) : a \circ (b \circ c) = (a \circ b) \circ c$
- Existence jednotkového/neutrálního prvku (**EJ**)
  $(\forall x \in A)(\exists e \in A) : a \circ e = e \circ a = a$
- Existence nulového (agresivního) prvku (**EN**)
  $(\forall x \in A)(\exists n \in A) : a \circ n = n \circ a = n$
- Existence inverzního prvku (**IN**)
  $(\forall a \in A)(\exists a^{-1} \in A) :$
  $a \circ a^{-1} = a^{-1} \circ a = e$
- Komutativita (**KO**)
  $(\forall a, b \in A) : a \circ b = b \circ a$
- Idempotentnost (**ID**)
  $(\forall a \in A) : a \circ a = a$

Binární operací na množině A nazveme každé zobrazení $\circ : A \times A \to A$. (Výsledek zůstane v A – uzavřenost.)

Grupoidy

Grupoid je uspořádaná dvojice $(A, \circ)$, kde A je neprázdná množina a $\circ$ je binární operace na A (**UZ**).

Pologrupa je asociativní grupoid (+**AS**).

Monoid je pologrupa s jednotkovým prvkem(+**EJ**).

Grupa je monoid s inverzními prvky ke každému prvku (+**IN**).

Abelova grupa je komutativní grupa (+**KO**).

Nechť $(A, \circ)$. Pokud existuje jednotkový/nulový prvek, pak je právě jeden.

Nechť $(A, \circ)$, $\circ$ je AS, a platí EJ na A. Pak $(\forall a \in A) : [\exists!\, a^{-1} \vee \nexists\, a^{-1}]$.

Nechť $(A, \circ)$ je grupa, pak $(\forall a \in A) : \exists!\, a^{-1}$.

Pokud v Cayleyho tabulce existuje řádek nebo sloupec s neunikátními hodnotami, pak se nemůže jednat o grupu.

Neutrální prvek je tam, kde se zkopíruje záhlaví Cayleyho tabulky.

Podgrupy

Nechť $(G, \circ)$ je grupa a platí:

1. $H \subseteq G$
2. $H \neq \emptyset$
3. $(\forall a, b \in G) : \boxed{a \circ b^{-1} \in H}$

Pak $(H, \circ)$ je podgrupou $(G, \circ)$.

Nechť $(G, \circ)$ je grupa a platí:

1. $H \subseteq G$
2. $H \neq \emptyset$
3. H je konečná množina
4. $(\forall a, b \in G) : \boxed{a \circ b \in H}$ (UZ)

Pak $(H, \circ)$ je podgrupou $(G, \circ)$.

Průnik podgrup je podgrupa.

Lagrangeova věta

Buď $(G, \circ)$ grupa, $a \in G$. Nejmenší $n \in \mathbb{N}$ splňující $\boxed{a^n = e}$ nazveme **řádem prvku** a. **Řád grupy** je $|G|$.

Nechť $(G, \circ)$ je *konečná grupa*, $(H, \circ)$ její podgrupa. Pak

- Řád prvku dělí řád grupy $\boxed{n \,\big|\, |G|}$
- Řád podgrupy dělí řád grupy $\boxed{|H| \,\big|\, |G|}$

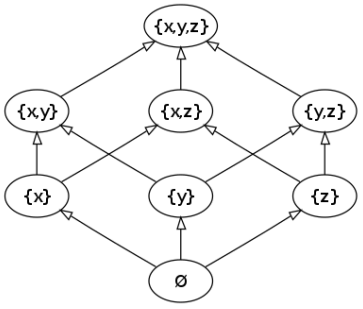
Cyklická grupa

Cyklická grupa je grupa tvořená mocninami jediného prvku. Řád tohoto prvku, tzv. generátoru grupy, je roven řádu grupy.

Svazy

Svazem *(lattice)* nazveme relační strukturu $(X, \subseteq) \iff (\forall (x, y) \in X) :$
$\boxed{\inf\{x, y\} \in X \ \wedge \ \sup\{x, y\} \in X.}$

Svaz nazveme **úplným** $\iff$
$\left(\forall M \subseteq \left\{2^X \setminus \emptyset\right\}\right) :$
$\boxed{\inf(M) \in X \ \wedge \ \sup(M) \in X.}$



Vlastnosti svazů

- Každý úplný svaz je svaz, protože inf a sup existuje i pro dvouprvkové podmnožiny.
- Každý *konečný* svaz je *úplný*.
- Každý úplný svaz $(X, \leq)$ obsahuje nejmenší prvek, **svazovou nulu** ($\inf X$), a největší prvek, **svazovou jedničku** ($\sup X$).
- „Motýlek“ nebo „kuří nožky“ znamenají, že daná struktura není svaz, protože mezi takovými prvky nenajdeme inf/sup.

Izotonní a antitonní zobrazení

Buď $(G, \leq_G)$ a $(H, \leq_H)$ uspořádané množiny. Řekneme, že $f\colon G \to H$ je

- **izotonní** zobrazení, které zachovává uspořádání, $\iff (\forall x, y \in G) :$
  $\boxed{x \leq_G y \quad \Rightarrow \quad f(x) \leq_H f(y)}$
- **antitonní** zobrazení, které „otáčí“ uspořádání, $\iff (\forall x, y \in G) :$
  $\boxed{x \leq_G y \quad \Rightarrow \quad f(x) \geq_H f(y)}$

Nechť $(X, \leq)$ je úplný svaz a zobrazení $\phi\colon X \to X$ je izotonní. Pak existuje pevný bod $\phi(x) = x$.

Kategorizace svazů

Podsvaz není obecně podgrafem. Musí zachovávat vlastnosti původního svazu (inf, sup).

Svaz je **distributivní** $\Leftrightarrow$ neobsahuje podsvaz izomorfní se svazem typu **diamant** nebo **pentagon**.

Svaz je **modulární** $\Leftrightarrow$ neobsahuje podsvaz izomorfní se svazem typu **pentagon**.

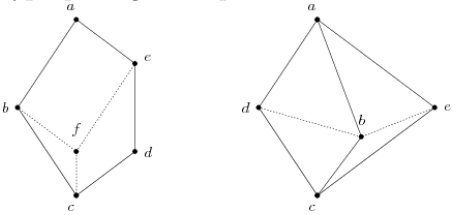
Každý distributivní svaz je modulární.

Úplný svaz je **komplementární** $\iff$ ke každému prvku existuje komplement x' takový že,

- $x \cup x' = 1$ (spojení ... $\sup(x)$)
- $x \cap x' = 0$ (průsek ... $\inf(x)$)

Svaz je **Booleův** $\iff$ je `distributivní` a `komplementární`.

Příklad. Svazy izomorfní se svazem typu pentagon, resp. diamant.



Okruh

Uspořádanou trojici $(R, +, \cdot)$, kde $R \neq \emptyset$ a $\{+, \cdot\}$ jsou zobrazení definované na $R \times R$, nazveme **okruhem** $\iff$

1. $(R, +)$ je `Abelova grupa`,
2. $(R, \cdot)$ je `pologrupa` a
3. platí `distributivní zákony`.

Pokud $(R, \cdot)$ je monoid (+EJ), mluvíme o **unitárním okruhu**.

Jednotkový prvek aditivní grupy okruhu je nulovým prvkem jeho multiplikativního monoidu. Tento prvek nazýváme nulovým prvkem okruhu.

Obor

Dělitelé nuly jsou $(a \neq 0 \wedge b \neq 0) : a \cdot b = 0$.

Obor je unitární okruh bez dělitelů nuly.

Obor integrity je komutativní obor.

Těleso

Těleso *(division ring)* je takový unitární okruh, že jeho nenulové prvky tvoří grupu vůči násobení. Tzn. $(A, +)$ je Abelova grupa a $(A \setminus \{0\}, \cdot)$ je grupa.

Pole *(field)* je komutativní těleso.

Těleso neobsahuje dělitele nuly.

Konečná tělesa nazýváme **Galoisovými tělesy**.

Těleso $\Rightarrow$ obor integrity.

Konečný obor integrity $\Rightarrow$
$\Rightarrow$ konečné pole.

Příklady

$(\mathbb{Z}_6, +, \cdot)$ je okruh, ale není to obor integrity, protože ex. dělitelé nuly

$$\overline{2_6} \cdot \overline{3_6} = \overline{6_6} = \overline{0_6}.$$

$(\mathbb{Z}, +, \cdot)$ je obor integrity, ale není to těleso, protože neex. některé inverze

$$2 \in \mathbb{Z} \quad \wedge \quad 2^{-1} = \frac{1}{2} \notin \mathbb{Z}.$$

$(\mathbb{R}[x]/_{\langle x^2+1 \rangle}, +, \cdot)$ je Galoisovo těleso.

$(\mathbb{Z}_p, +, \cdot)$, kde p je prvočíslo je konečné (Galoisovo) těleso.

Vektorový prostor (VP)

Reálným vektorovým prostorem nad tělesem T nazveme množinu $V = T^n$ s algebraickou operací sčítání

$$+ : (u, v) \in V \times V \to u + v \in V$$

(`sčítání vektorů`) a zobrazením

$$\cdot : (\alpha, v) \in T \times V \to \alpha \cdot v \in V$$

(`násobení skalárem`), přičemž platí axiomy:

1. UZ: $u + v \in V$
2. AS: $u + (v + w) = (u + v) + w$
3. KO: $u + v = v + u$
4. EJ: $u + o = u$
5. IN: $u + (-u) = o$
6. UZ: $\alpha \cdot u \in V$
7. EJ: $1 \cdot u = u$
8. AS: $\alpha \cdot (\beta \cdot u) = (\alpha \cdot \beta) \cdot u$
9. DZ: $\alpha \cdot (u + v) = \alpha \cdot u + \alpha \cdot v$
10. DZ: $u \cdot (\alpha + \beta) = \alpha \cdot u + \beta \cdot u$

Vektorový podprostor

Neprázdnou množinu $U \subset V$ nazveme podprostorem VP V, jestliže U je VP vzhledem ke sčítání vektorů a násobení skalárem, které je definované na V.

Množina $U \subset V$ je podprostor VP V $\iff$ $(\forall u, v \in V)(\forall \alpha \in T)$:

1. $U \neq \emptyset$
2. $u + v \in U$
3. $\alpha \cdot u \in U$

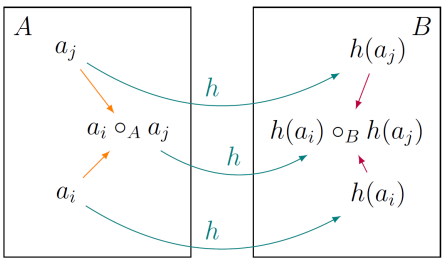
Morfismy

Morfismus je zobrazení $h\colon A \to A'$ mezi dvěma algebrami se stejnou signaturou, které *zachovává operaci.*

Nechť $(A, \circ_A)$ a $(B, \circ_B)$ jsou algebry s binární operací $\circ_A$, resp. $\circ_B$. Pak pro morfismus $h\colon A \to B$ platí

$$\boxed{h(a_i \circ_A a_j) = h(a_i) \circ_B h(a_j)}$$

Jádro morfismu je množina vzorů $1_B \in B$ (pokud $(B, \circ_B)$ je grupa).



Kategorizace morfismů

Epimorfismus – zobrazení h je `surjektivní` (na)

Monomorfismus – zobrazení h je `injektivní` (prosté)

Izomorfismus – zobrazení h je `bijekce`

Endomorfismus – zobrazení $h\colon A \to A$ je `zúžení`, tzn. $A' \subseteq A$

Automorfismus – pro zobrazení $h\colon A \to A$ platí $A' = A$

Algebry nazýváme **izomorfní** pokud mezi nimi existuje izomorfismus. Relace být izomorfní je relace ekvivalence.

Kongruence

Kongruencí na algebře $(A, \circ)$, kde $\circ$ je binární relace, rozumíme relaci ekvivalence $\mathsf{R} \subseteq A \times A$, která splňuje podmínku $(\forall a, b, c, d \in A)$:

$$\boxed{a\mathsf{R}b \wedge c\mathsf{R}d \Rightarrow (a \circ c)\mathsf{R}(b \circ d)}$$

Jádro kongruence je třída prvků ekvivalentních s jednotkovým prvkem.

FCA

Formální konceptuální analýza (FCA) je metoda, která pracuje s binárními tabulkovými daty (formální kontext), které popisují relaci mezi objekty a atributy.

Formální kontext je uspořádaná trojice (X, Y, I), kde

- X je množina `objektů`,
- Y je množina `atributů`,
- $I \subseteq X \times Y$ je `binární relace`.

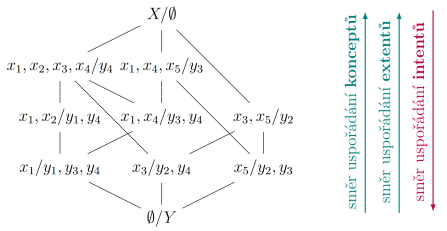
Skutečnost, že `objekt` $x \in X$ `má atribut` $y \in Y$, značíme $(x, y) \in I$.

Každý kontext indukuje **šipkové zobrazení** $^{\uparrow}\colon 2^X \to 2^Y$ a $^{\downarrow}\colon 2^Y \to 2^X$ definované:

- $A \subseteq X, A^{\uparrow} = \{y \in Y \mid (\forall x \in A) : (x, y) \in I\}$,
- $B \subseteq Y, B^{\downarrow} = \{x \in X \mid (\forall y \in B) : (x, y) \in I\}$.

Formální koncept kontextu (X, Y, I) definujeme jako dvojici (A, B), kde $X \supseteq A = B^{\downarrow}$ nazýváme **extent** a $Y \supseteq B = A^{\uparrow}$ nazýváme **intent**.

Konceptuální svaz je kontext (X, Y, I) spolu s relací '$\leq$' s.t. $(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 \Leftrightarrow B_1 \supseteq B_2$.



Asociační pravidla (AP)

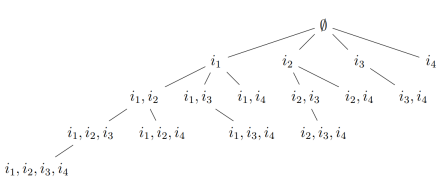
Podpora atributu m v matici transakcí $T \in \{0,1\}^{N,M}$ je jeho relativní četnost $\text{supp}(m) = \sum T_{:,m}/N$. Obdobně podpora více atributů $\text{supp}(m_1, m_2) = \sum(T_{:,m_1} \wedge T_{:,m_2})/N$.

Spolehlivost asociačního pravidla je

$$\boxed{\text{conf}(A \Rightarrow B) = \frac{\text{supp}(A, B)}{\text{supp}(A)}.}$$

(jako podmíněná pravděpodobnost)

Rymon Tree



Metrické prostory

Metrický prostor je dvojice $(\mathcal{M}, \rho)$, kde $\mathcal{M} \neq \emptyset$ a $\rho\colon \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ je metrika, která splňuje $\forall x, y, z \in \mathcal{M}$ axiomy:

- $\rho(x, y) = 0 \iff x = y$ ...totožnost
- $\rho(x, y) = \rho(y, x)$ ...symetrie
- $\rho(x, z) < \rho(x, y) + \rho(y, z)$ ...$\triangle \neq$

Metrika je nezáporná. $2\rho(x, y) = \rho(x, y) + \rho(y, x) \geq \rho(x, x) = 0$ □.

Podobnosti a nepodobnosti

Pseudometrika umožňuje, aby dva různé body měly mezi sebou nulovou vzdálenost (axiom totožnosti je zredukován na $\rho(x, x) = 0$).

Nepodobnost definujeme pomocí axiomů totožnosti a symetrie (neplatí $\triangle \neq$).

Ultrametrika je definovaná silnější $\triangle \neq$, $\rho(x, z) \leq \max\{\rho(x, y), \rho(y, z)\}$ (např. diskrétní prostor 0/1).

Podobnost je zobrazení $\text{sim}\colon \mathcal{V} \times \mathcal{V} \to \mathbb{R}$, kde $\mathcal{V} \neq \emptyset$, které splňuje $\forall x, y \in \mathcal{V}$ axiomy:

- $\text{sim}(x, y) \geq 0$ ...nezápornost
- $\text{sim}(x, y) \leq \text{sim}(x, x) \wedge \text{sim}(x, y) = \text{sim}(x, x) \iff x = y$ ...totožnost (bod je nejvíce podobný sám sobě)
- $\text{sim}(x, y) = \text{sim}(y, x)$ ...symetrie

Metriky v $\mathbb{R}^n$

Minkowského metrika (p-norma)

$$\rho_p(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}, p \in \langle 1, +\text{inf}\rangle$$

Manhattanská metrika (1-norma)

$$\rho_1(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

Eukleidova metrika (2-norma)

$$\rho_2(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Čebyševova metrika (∞-norma)

$$\rho_\infty(x, y) = \max_{i=1\ldots n} \{|x_i - y_i|\}$$

Mahalanobisova metrika ($\|x - y\|_{S^{-1}}$)

$$\rho_{\text{maha}}(x, y) = \sqrt{(x - y)^T S^{-1} (x - y)}$$

Pozn. platí $\boxed{\rho_1 \geq \rho_2 \geq \ldots \geq \rho_\infty}$

Metriky v $\{0,1\}^n$

Hammingova vzdálenost (počet rozdílných bitů - XOR)

$$\rho_H(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

Tanimotova vzdálenost

$$\rho_T(x, y) = 1 - \text{sim}_T(x, y)$$

Metriky nad množinami

Jaccardova vzdálenost množin

$$\rho_J(A, B) = 1 - \text{sim}_J(A, B)$$

Tanimotova vzdálenost množin

$$\rho_T(A, B) = 1 - \text{sim}_T(A, B)$$

Metriky nad abecedou $\sum^*$

Levenshteinova (editační) vzdálenost je definována jako nejmenší počet operací `vkládání`, `mazaní` a `substituce`.

Longest Common Subsequence (LCS) je definována jako nejmenší počet operací `vkládání` a `mazaní`.

Hammingova vzdálenost je definována pro stejně dlouhá slova jako `počet rozdílných pozic`.

Podobnosti nad $\mathbb{R}^n$

Kosinová podobnost

$$\text{sim}_C(x, y) = \frac{\langle x, y\rangle}{\|x\| \cdot \|y\|}$$

Růžičkova podobnost

$$\text{sim}_R(x, y) = \frac{\sum \min\{x_i, y_i\}}{\sum \max\{x_i, y_i\}}$$

Podobnosti nad množinami

Jaccardova podobnost množin (IoU)

$$\text{sim}_J(x, y) = \frac{|A \cap B|}{|A \cup B|}$$

Diceova podobnost množin

$$\text{sim}_D(x, y) = \frac{2 \cdot |A \cap B|}{|A| + |B|}$$

Tanimotova podobnost množin

$$\text{sim}_T(x, y) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Podobnosti nad $\{0,1\}^n$

Tanimotova podobnost binárních vektorů

$$\text{sim}_T(x, y) = \frac{\sum \min\{x_i, y_i\}}{\sum \max\{x_i, y_i\}}$$

Vlastnosti množin v $(\mathcal{M}, \rho)$

Průměr množiny $A \subseteq (\mathcal{M}, \rho)$ definujeme $diam(A) = \sup\{\rho(x,y) \mid x, y \in A\}$, resp. 0, pokud $A = \emptyset$.

Buď $x \in (\mathcal{M}, \rho)$ a $r \in (0, +\infty)$, pak následující množiny nazveme po řadě otevřená koule, uzavřená koule a sféra o středu x a poloměru r:

- $B(x,r) = \{y \in \mathcal{M} \mid \rho(x,y) < r\}$
- $\overline{B}(x,r) = \{y \in \mathcal{M} \mid \rho(x,y) \leq r\}$
- $S(x,r) = \{y \in \mathcal{M} \mid \rho(x,y) = r\}$

Pozn. otevřená koule je epsilon okolí, když zvolíme $r = \varepsilon$, tj. $\mathcal{O}(x) = B(x,r)$.

Kategorizace bodů v množině

Buď $A \subseteq (\mathcal{M}, \rho)$. Bod $x \in \mathcal{M}$ nazveme:

- **vnitřním** bodem $A \iff$ $\boxed{\mathcal{O}(x) \subset A \Leftrightarrow A \cap \mathcal{O}(x) = \mathcal{O}(x)}$,
- **vnějším** bodem $A \iff$ $\boxed{\mathcal{O}(x) \subset \mathcal{M} \setminus A \Leftrightarrow A \cap \mathcal{O}(x) = \emptyset}$,
- **hraničním** bodem $A \iff$ $\boxed{A \cap \mathcal{O}(x) = \emptyset \wedge (\mathcal{M} \setminus A) \cap \mathcal{O}(x) = \emptyset}$,
- **hromadným** bodem $A \iff$ $\boxed{\mathcal{O}(x) \setminus \{x\} = \emptyset}$ (tj. existuje posloupnost, která má limitu v x; konečné množiny hromadné body nemají),
- **izolovaným** bodem $A \iff$ $\boxed{\mathcal{O}(x) \cap A = \{x\}}$ (každý bod konečné množiny je izolovaný).

Vnitřek, **vnějšek** a **hranice** jsou postupně množiny všech vnitřních, vnějších a hraničních bodů (hranici A značíme ∂A nebo $bd(a)$).

Uzávěr mžy A je $\boxed{cl(A) = A \cup \partial A.}$

Topologické prostory (TP)

Na množině X je dána topologie pomocí **otevřených** množin, je-li dán systém podmnožin $\tau \subseteq 2^X$ takový, že:

1. Prázdná a celá množina, $\boxed{\emptyset \in \tau \wedge X \in \tau}$
2. Uzavřenost vůči `průniku dvojic`, $\boxed{A, B \in \tau \Rightarrow A \cap B \in \tau}$
3. Uzavřenost vůči `sjednocení`, $\boxed{(\forall A_i \in \tau) : \bigcup_i U_i \in \tau}$

Pak (X, τ) tvoří TP, kde prvky topologie τ jsou `otevřené` množiny.

Duálně, na množině X je dána topologie pomocí **uzavřených** množin, je-li dán systém podmnožin $\tau' \subseteq 2^X$ takový, že:

1. Prázdná a celá množina, $\boxed{\emptyset \in \tau' \wedge X \in \tau'}$
2. Uzavřenost vůči `sjednocení dvojic`, $\boxed{A, B \in \tau' \Rightarrow A \cup B \in \tau'}$
3. Uzavřenost vůči `průniku`, $\boxed{(\forall A_i \in \tau') : \bigcap_i U_i \in \tau'}$

Pak (X, τ') tvoří TP, kde prvky topologie τ' jsou `uzavřené` množiny.

Duální topologii vytvořím $\boxed{\tau' = X \setminus \tau.}$

Vlastnosti TP

Buď (X, τ') TP uzavřených množin. **Uzávěr** $A \subseteq X$ je nejmenší (uzavřená) množina τ', která A obsahuje, tj. $\boxed{cl(A) = \bigcap \{B \in \tau' \mid A \subseteq B\}.}$

Buď (X, τ) TP otevřených množin. **Okolí** bodu x v (X, τ) definujeme jako $O(x) \subseteq X \iff (\exists A \in \tau) : x \in A \wedge A \subseteq O(x)$.

Vnitřek $A \subseteq X$ v (X, τ) je největší (otevřená) mža v τ, která je celá v A, tj. $\boxed{\text{int}(A) = \bigcup \{B \in \tau \mid B \subseteq A\}.}$

Bod $x \in A, A \subseteq X$, nazveme **vnitřním** bodem $A \iff$
$(\exists$ (otevřená) mža $B \in \tau)$: $x \in B \subseteq A$.

Hranice $A \subseteq X$ v TP je $\boxed{\partial A = cl(A) \setminus \text{int}(A).}$

TP nazveme **souvislý** $\iff \emptyset$ a X jsou jediné podmnožiny X, které jsou `obojetné` (tzn. neex. jiná podmnožina X, která by byla zároveň otevřená i uzavřená).

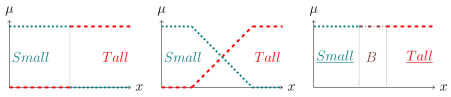
Komponenta TP je maximální souvislá podmnožina (tzn. podmnožina taková, aby TP byl souvislý).

Fuzzy a rough množiny

Fuzzy množiny neurčují přesné hranice množiny. Příslušnost daného prvku je popsána **charakteristickou funkcí**.

Rough množiny – prvky, pro které s jistotou známe přiřazení, přiřadíme do **dolní aproximace** ($\underline{Small}$, $\underline{Tall}$); pokud si nejsme jisti přiřazením, tak prvky umístíme do **horní aproximace** ($B = \overline{Small} \cup \overline{Tall}$, kde B je hraniční oblast).

Obr. booleovská/ostrá, fuzzy a rough (hrubá) množina:



Fuzzy množiny

Fuzzy podmnožina A univerza X je definovaná **funkcí příslušnosti** $\boxed{\mu_A \colon X \to \langle 0,1\rangle}$ Def. $A(x) := \mu_A(x)$.

Buď A, B fuzzy množiny. $\boxed{A \subseteq B \iff \forall x \in X \colon \mu_A(x) < \mu_B(x)}$

Standardní fuzzy operace $\alpha, \beta \in \langle 0,1\rangle$:

- negace $\boxed{\neg_s \alpha = 1 - \alpha}$
- konjunkce $\boxed{\alpha \wedge_s \beta = \min\{\alpha, \beta\}}$
- disjunkce $\boxed{\alpha \vee_s \beta = \max\{\alpha, \beta\}}$
- doplněk $\boxed{\mu_{\overline{A}}(X) = \neg \mu_A(X)}$
- průnik $\boxed{\mu_{A\cap B}(x) = \mu_A(x) \wedge \mu_B(x)}$
- sjednocení $\boxed{\mu_{A\cup B}(x) = \mu_A(x) \vee \mu_B(x)}$

Rough množiny

Informační systém (IS, aproximační prostor) je čtveřice $\boxed{\mathcal{A} = (U, A, V, f)}$, kde $U \neq \emptyset$ je konečné **univerzum objektů**, $A \neq \emptyset$ je konečná **množina atributů** a funkce f je popisující funkce (**informace**) $f \colon U \times A \to V$, kde V je množina hodnot, které nabývají atributy z A.

Buď $\mathcal{A} = (U, A, V, f)$ IS. Buď $B \subseteq A$ množina atributů, která indukuje rozklad univerza U s relací B-**nerozlišitelnosti** (B-indiscernability)

$$\mathrm{Ind}_B = \{(x,y) \in U \times U \mid \forall a \in B\big(f(x,a) = f(y,a)\big)\}$$

Ind_B je relace ekvivalence (ReSyTr), která rozkládá množinu U na konečný počet disjunktních tříd rozkladu $[x]_B = \big\{y \in U \mid \forall a \in B\big(f(x,a) = f(y,a)\big)\big\}$. Množina tříd rozkladu tvoří faktorovou množinu univerza $U/_{\mathrm{Ind}_B} = \{[x]_{\mathrm{Ind}_B} \mid x \in U\}$

Buď $\mathcal{A} = (U, A, V, f)$ IS a buď Ind_B na U pro atributy $B \subseteq A$. Pro $X \subseteq U$ def. B-**dolní aproximaci** X jako

$$\boxed{\underline{B}(X) = \{x \mid [x]_B \subseteq X\}}$$

a B-**horní aproximaci** X jako

$$\boxed{\overline{B}(X) = \{x \mid [x]_B \cap X \neq \emptyset\}}.$$

O cílové může $X \subseteq U$ a může atributů $B \subseteq A$ řekneme, že X je **definovatelná vzhledem** k B $\iff \boxed{\overline{B}(X) = \underline{B}(X).}$

Řekneme, že X je **hrubá množina** vzhledem k B $\iff \boxed{\overline{B}(X) \neq \underline{B}(X).}$

Hrubost množiny X vzhledem k B definujeme jako $\boxed{r_B(X) = 1 - \dfrac{\underline{B}(X)}{\overline{B}(X)}}$.

Rough set příklad

Sample Information System

Object	P_1	P_2	P_3	P_4	P_5
O_1	1	2	0	1	1
O_2	1	2	0	1	1
O_3	2	0	0	1	0
O_4	0	0	1	2	1
O_5	2	1	0	2	1
O_6	0	0	1	2	2
O_7	2	0	0	1	0
O_8	0	1	2	2	1
O_9	2	1	0	2	2
O_{10}	2	0	0	1	0

When the full set of attributes $P = \{P_1, P_2, P_3, P_4, P_5\}$ is considered, we see that we have the following seven equivalence classes:

$$\begin{cases} \{O_1, O_2\} \\ \{O_3, O_7, O_{10}\} \\ \{O_4\} \\ \{O_5\} \\ \{O_6\} \\ \{O_8\} \\ \{O_9\} \end{cases}$$

Thus, the two objects within the first equivalence class, $\{O_1, O_2\}$, cannot be distinguished from each other based on the available attributes, and the three objects within the second equivalence class, $\{O_3, O_7, O_{10}\}$, cannot be distinguished from one another based on the available attributes. The remaining five objects are each discernible from all other objects.

It is apparent that different attribute subset selections will in general lead to different indiscernibility classes. For example, if attribute $P = \{P_1\}$ alone is selected, we obtain the following, much coarser, equivalence-class structure:

$$\begin{cases} \{O_1, O_2\} \\ \{O_3, O_5, O_7, O_9, O_{10}\} \\ \{O_4, O_6, O_8\} \end{cases}$$

Shlukování

Podle hierarchizace shluků:

1. **hierarchické** shlukování ($\rightarrow$ dendrogramy)
   (a) `bottom-up` – **aglomerativní** shlukování (pro N bodů začínám s N shluky)
   (b) `top-down` – **divizivní** shlukování (začínám s jedním shlukem, např. k-means, Kernighan-Lin)
2. **nehierarchické** shlukování (mezi shluky není uspořádání)

Podle překrývání shluků:

1. **nepřekrývající** se *(crisp, sharp)* $\Rightarrow$ množina shluků tvoří `rozklad` na množině dat (relace ekvivalence) $[(\forall i \neq j) : C_i \cap C_j = \emptyset] \wedge [\bigcup C_i = \mathcal{X}]$
2. **překrývající** se *(soft, overlap)* $\Rightarrow$ množina shluků tvoří `pokrytí` (relace tolerance)

Aglomerativní shlukování

1. **Single Linkage** $d_{SL}(C_i, C_j) = \min\{d(x,y) \mid x \in C_1, y \in C_2\}$
2. **Complete Linkage** $d_{CL}(C_i, C_j) = \max\{d(x,y) \mid x \in C_1, y \in C_2\}$
3. **Average Linkage** $d_{AL}(C_i, C_j) = \dfrac{\sum_{x \in C_1, y \in C_2} d(x,y)}{|C_1| \cdot |C_1|}$
4. **Centroid method** $d_C(C_i, C_j) = d(c_i, c_j)$, kde $c_i = \dfrac{\sum_{x \in C_i} x}{|C_i|}$
5. **Ward's method** $d_W(C_i, C_j) = \dfrac{|C_i| \cdot |C_j|}{|C_i| + |C_j|} d^2(C_i, C_j)$

DBSCAN

DBSCAN je zástupce shlukovacích metod založených na hustotě. Má dva hyperparametry:

- ε – velikost okolí
- `min_sample` – minimální počet objektů v okolí

Overlap clustering

- Fuzzy-k-means
- Rough-set-k-means
- CPM - Clique Percolation Method

Interní validace shlukování

- Koeficient siluety (Silhouette index) – větší lepší
- Dunn index – větší lepší (identifikuje husté a dobře oddělené shluky)
- Davies-Bouldin index – menší lepší
- Calinski-Harabasz index – větší lepší, ale musí tam být 'peak' (identifikuje husté a dobře oddělené shluky)

Externí validace shlukování

- Entropie
- Čistota (purity)
- Rand index ("accuracy")
- F-míra (F-measure)
- Jaccardův index (IoU)

Jádro grafu

Buď $G = (V, E)$ orientovaný graf. Množinu $C \subseteq V$ nazveme **jádrem grafu** $G \iff$

1. Neexistují hrany mezi vrcholy jádra grafu, tzn. $(\forall e \in E, e = (v_1, v_2)) : v_1 \in C \Rightarrow v_2 \notin C$.
2. Z každého vrcholu mimo jádro grafu vede hrana do jádra grafu, tzn. $v_1 \notin C \Rightarrow (\exists v_2 \in C) : e = (v_1, v_2)$.

Algoritmus nalezení jádra G:

1. Vrchol(y) s nulovým **výstupním** stupněm přidám do C.
2. Odstraním všechny hrany incidentní s vrcholy C.
3. Opakuj 1,2.

Topologické uspořádání

Topologické uspořádání vrcholů orientovaného grafu je očíslování vrcholů přirozenými čísly tak, že každá hrana vede z vrcholu s nižším číslem do vrcholu s vyšším číslem. Pouze pro acyklické grafy.

1. Najdi vrchol(y) s nulovým **vstupním** stupněm.
2. Odstraň incidentní hrany.
3. Opakuj 1,2.

Doplnění ke svazům

Hasseův diagram, který obsahuje strukturu „motýlek“ není svaz.