Cvičení 9

Příklad 1: Připomeňte si pojmy vektorový prostor, metrický prostor, metrika, vzdálenost a podobnost.

a) Jaké znáte metriky pro určení vzdálenosti mezi dvěma reálnými vektory stejné dimenze?
b) Jaké znáte metriky pro určení vzdálenosti mezi dvěma binárními vektory stejné dimenze?
c) Jaké znáte metriky pro určení vzdálenosti mezi dvěma slovy nad vybranou abecedou?
d) Jak můžete převést vzdálenost na podobnost?

Příklad 2: Rozhodněte, které z níže uvedených zobrazení jsou metrikou, ultrametrikou, pseudometrikou, podobností a nepodobností.

a) $d_1 : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $d_1(x, y) = |\log(y/x)|$,
b) $d_2 : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, $d_2(x, y) = ||x||_2 + ||y||_2$ pro $x \neq y$ a $d_2(x, y) = 0$ pro $x = y$,
c) $d_3 : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, $d_3(x, y) = ||x||_3 \cdot ||y||_3$ pro $x \neq y$ a $d_3(x, y) = 0$ pro $x = y$,
d) $d_4 : 2^A \times 2^A \to \mathbb{R}$, kde A je libovolná neprázdná množina a $d_4(x, y) = 1$ pro $x \neq y$ a $d_x(x, y) = 0$ pro $x = y$,
e) $d_5 : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $d_5(x, y) = e^{|y-x|} - 1$,
f) $d_6 : \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n} \to \mathbb{R}$, $d_6(X, Y) = \mathrm{rank}(Y - X)$, kde $\mathrm{rank}(X)$ = hodnost matice X,
g) $d_7 : \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n} \to \mathbb{R}$, $d_7(X, Y) = \det(Y + X)$,
h) $d_8 : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $d_8(x, y) = |x^2 - y^2|$,
i) $d_9 : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, $d_9(x, y) = e^{-||y-x||_2}$,
j) $d_{10} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, $d_{10}(x, y) = \frac{\sum_{i=1}^n (x_i \cdot y_i)}{(\sum_{i=1}^n x_i^2)^{1/2} \cdot (\sum_{i=1}^n y_i^2)^{1/2}} = \frac{x \cdot y}{||x||_2 \cdot ||y||_2}$,
k) $d_{11} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, $d_{11}(x, y) = 1 - \frac{x \cdot y}{||x||_2 \cdot ||y||_2}$.

Příklad 3: Navrhněte způsoby kategorizace a navrhněte způsoby normalizace níže uvedených dat.

a) Vstupní vektorový model má na řádku osobu a ve sloupcích jsou informace: pohlaví - m/ž, rok narození, dosažené vzdělání, místo trvalého pobytu, rok získání řidičského průkazu, počet zaplacených pokut, počet nezaplacených pokut.

jméno	m/ž	narození	vzděláni	pobyt	získání ř.p.	zaplacené	nezaplacené
P1	ž	1942	sš	Ostrava	1993	0	1
P2	m	1968	vš	Ostrov nad Ohří	1994	4	0
P3	ž	1957	vš	Ostrovačice	1987	0	0
P4	m	1986	sš	Ostravice	2010	1	0

b) Je dán následující vektorový model. Řádky tvoří dokumenty, sloupce tvoří frekvence nalezených termů.

	t1	t2	t3	t4	t5	t6	t7
d1	1	0	0	1	3	0	1
d2	1	0	0	2	2	0	0
d3	0	2	1	1	1	1	0
d4	1	0	3	1	0	1	0

Příklad 4: Výjmenujte a použijte metody pro výpočet vzdálenosti či podobnosti nad nenormovanými a normovanými daty v příkladu 3.

a) $d(P1, P4) =$, $d(P1, P3) =$, $s(P1, P4) =$, $s(P1, P3) =$
b) $d(d1, d2) =$, $d(d1, d3) =$, $s(d1, d2) =$, $s(d1, d3) =$

Příklad 5: Pokud je možné, tak použijte dále uvedené metody pro výpočet vzdálenosti a podobnosti mezi dvojici slov, mezi dvojici množin nebo mezi dvojici binárních vektorů. Hammingova vzdálenost, editační vzdálenost (Levenshtein), longest common subsequence, kosinova podobnost, Jaccardova podobnost a Diceova podobnost.

a) lesnictví - lesněnky
b) lesněnky - žvýkačky
c) logika - geolog
d) 011110 - 111111
e) kosmický let - komický ret
f) ACCGATTA - ACAGATAA
g) 10101 - 01010

Příklad 6: Určete, které prvky patří do uzavřené koule $K_L(les, 1)$ a které do $K_{LCS}(les, 1)$ nad množinou A={lem, len, les lesy, lep, lis, los, lup, ples, nes, nos, ves, vos, vez}.

Příklad 7: Na množině $A = \{a, b, c, d, e, f\}$ jsou dány následující relace:

a) ekvivalence ρ_1, která indukuje rozklad $A/\rho_1 = \{\{a, b, c, d\}, \{e, f\}\}$,
b) ekvivalence ρ_2, která indukuje rozklad $A/\rho_2 = \{A\}$,
c) tolerance ρ_3, která indukuje pokrytí $A/\rho_3 = \{\{b, d, e, f\}, \{a, b, c, d\}\}$,
d) $\rho_4 = \{(a, b), (b, c), (c, d), (d, e), (e, f), (f, a)\}$.

Nakreslete grafy, které reprezentují relace ρ_i a určete, zda délka nejkratší cesty v daném grafu je metrikou. Pokud ano, tak určete typ daného metrického prostoru a vypočtěte průměr tohoto prostoru ($diam_i(A) =$).