|  |  |  |
| --- | --- | --- |
| Příjmení: **Holcman** | Jméno: **Jiří** | Obor: **NKYB** |
| Předmět: **AK0DA** | | Datum: **27.3.2022** |
| Název protokolu: **Semestrální úkol** | | |

# Zadání

Zvolte si libovolný programovací jazyk, libovolnou machine learning/datamining/neural network/artificial intelligence knihovnu a libovolný dataset, na kterém si vyzkoušíte některou z úloh dataminingu:

klasifikace

clustering

regrese

extrakce asociačních vzorů

detekce anomálií

Cílem je, abyste se vlastnoručně seznámili s dostupnými nástroji a prací s nimi. Výstup bude v prezentovatelné formě - např. textový dokument s vizualizacemi, ale klidně i jiný formát. Jen prosím takový, abych nemusel pro kontrolu úlohy instalovat 14 knihoven a specifické IDE.

# Vypracování

Pro tuto úlohu jsem si zvolil úlohu založenou na metodě hierarchického shlukování (clasteringu) dat za pomocí vzdálenosti mezi jednotlivými členy datasetu DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Nastavením dvou vstupních parametrů (minimálního počtu a vzdálenosti) se ze vstupního datasetu vytvoří graf obsahující všechny vstupní prvky a jejich seskupení do shluků (clastru).

Jako vstupní dataset byl použit soubor ve tvaru CSV (hodnoty oddělené čárkami), která osahuje seznam zákazníků obchodu s uvedením jejich příjmů a přiřazených hodnot nákupování. Soubor obsahuje i pohlaví a věk, ale s tím není v dané úloze počítáno. Dalo by se tím díky tomu úlohu dále dělit na podúlohy (dle věku, pohlaví…).

Vzor části souboru Zakaznici.csv

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Pohlaví | Věk | Příjem (tis EUR) | Skóre útraty (1-100) |
| 1 | Muž | 19 | 15 | 39 |
| 2 | Muž | 21 | 15 | 81 |
| 3 | Žena | 20 | 16 | 6 |
| 4 | Žena | 23 | 16 | 77 |
| 5 | Žena | 31 | 17 | 40 |
| 6 | Žena | 22 | 17 | 76 |
| 7 | Žena | 35 | 18 | 6 |
| 8 | Žena | 23 | 18 | 94 |
| 9 | Muž | 64 | 19 | 3 |
| 10 | Žena | 30 | 19 | 72 |

….

Celý program je realizován v **MATLABu R2022a** a spouští se skriptem **start.m**. Jako datamining knihovna byl použit Statistics and Machine Learning Toolbox. Tento toolbox poskytuje funkce a aplikace popis, analýzu a modelování dat, součástí toho je i funkce DBSCAN. Jako výstup z běhu skriptu jsou tři grafy.

Popis skriptu:

Skript je celý v jednom souboru s názvem **start.m**. Na začátku programu je definován počet dimenzí (každý prvek – zákazník obsahuje roční příjem a index útraty zákazníka v rozsahu 1-100 – to znamená počet dimenzí = 2):

Zdrojový kód

dimenze = 2;

Dále jsou načtena potřebná data z datasetu uloženém v souboru **Zakaznik.csv**:

Zdrojový kód

**% načíst dataset**

**jmenoSouboru = 'Zakaznici.csv';**

**dataset = readmatrix(jmenoSouboru,'Range','D2:E10000');**

Zde jsou načtena pouze potřebná data a to sloupec D a E a to od druhého řádku souboru (první řádek obsahuje hlavičku s informacemi co který sloupec znamená).

V dalším kroku z celého pole hodnot vypočítáme minimální vzdálenosti nejbližšího prvku pro každého zákazníka, tím dostaneme 200 hodnot, které seřadíme dle velikosti a z něj vytvoříme graf.

Zdrojový kód

**% minimální vzdálenost**

**vzdalenosti = zeros(200,1);**

**for a = 1:200**

**vysledek = 1000000000;**

**for b = 1:200**

**if (a ~= b)**

**vypocet = sqrt( ((dataset(b,1)-dataset(a,1))^2) + ((dataset(b,2)-dataset(a,2))^2) );**

**if (vysledek > vypocet)**

**vysledek = vypocet;**

**end**

**end**

**end**

**vzdalenosti(a,1) = vysledek;**

**end**

**f=figure('PaperOrientation','landscape');**

**plot(1:200,sort(vzdalenosti))**

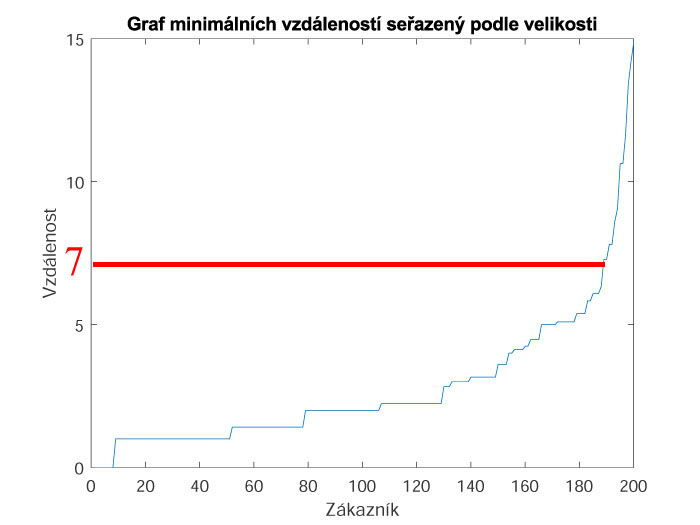
**title('Graf minimálních vzdáleností seřazený podle velikosti');**

**xlabel( 'Zákazník' );**

**ylabel( 'Vzdálenost' );**

**saveas(f,'graf\_min\_vzdalenosti.pdf');**

Ve vytvořeném grafu je nalezeno místo, kde se hodnoty začínají prudce zvyšovat (v našem případě jde o hodnotu eps = 7) což nám signalizuje že další hodnoty (vzdálenosti) již oddělují hraniční a šumové body.



Nyní je vytvořen graf všech zákazníků, kteří ještě nejsou roztříděni do jednotlivých skupin podle hustoty (do tzv. clastrů).

Zdrojový kód

**% graf hodnot v datasetu**

**f=figure('PaperOrientation','landscape');**

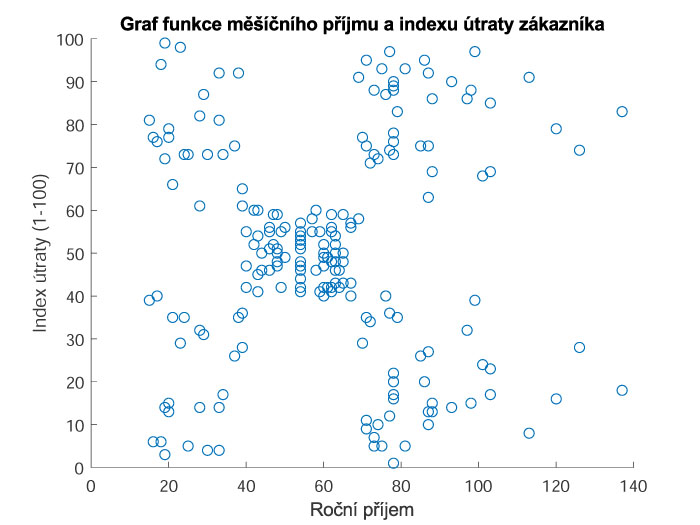
**scatter(dataset(:,1),dataset(:,2));**

**title('Graf funkce měšíčního příjmu a indexu útraty zákazníka');**

**xlabel( 'Roční příjem' );**

**ylabel( 'Index útraty (1-100)' );**

**saveas(f,'graf\_bez\_skupin.pdf');**



Nyní je na toto pole hodnot aplikována funkce DBSCAN, která na základě vstupních parametrů provede clastering, čily rozdělení do skupin dle hustoty.

Jako parametry jsou použity dvě hodnoty:

**Eps = 7** – tato hodnota je odhadnuta z grafu minimálních vzdáleností

**minBodu** – tato hodnota se běžně určuje jako dvojnásobek dimenze problému, v našem případě 2 x 2 = 4

Zdrojový kód

**% clastering pomocí funkce DBSCAN**

**eps = 7;**

**minBodu = dimenze \* 2;**

**skupina = dbscan(dataset,eps,minBodu);**

Na závěr je opět vytvořen graf, tentokrát již s barevně odlišenými shluky dat.

Zdrojový kód

**f=figure('PaperOrientation','landscape');**

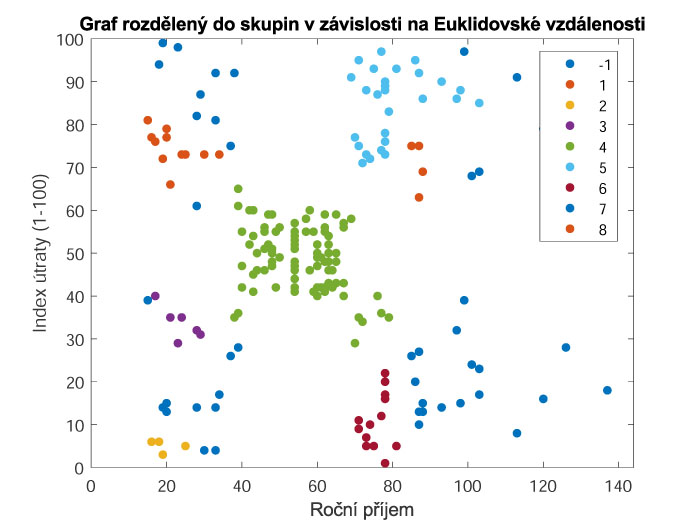
**gscatter(dataset(:,1),dataset(:,2),skupina);**

**title('Graf rozdělený do skupin v závislosti na Euklidovské vzdálenosti');**

**xlabel( 'Roční příjem' );**

**ylabel( 'Index útraty (1-100)' );**

**saveas(f,'graf\_skupiny.pdf');**

****

# Závěr

Cílem projektu bylo seznámit se z libovolnou metodou dataminingu a jeho praktické aplikace. Z výsledku této úlohy by se dalo přesněji a cíleněji zaměřit reklamu zasílanou zákazníkům obchodu a mít tak lepší přehled o tom v čem by se mohl obchod do budoucna zlepšit.