I-SUNS: Zadanie č.1

Metriky a zhlukovanie

Vo vybranom programovacom jazyku implementujte program, bude hľadať podobnosti medzi krajinami. K dispozícii v IS budete mať rôzne štatistické ukazovatele o vybraných krajinách a vašou úlohou je najprv pomocou rôznych metrík nájsť navzájom najviac/najmenej podobné a následne využiť metódy učenia bez učiteľa na zoskupovanie krajín podľa istej podobnosti. **Prosím začnite prečítaním celého zadania vrátane spresnení, ktoré sú uvedené na konci.**

Čas odovzdania je určený časom vloženia do AIS. Deadline pre získanie 8 bodov je **18.10.2019 o 07:59** pre cvičenie o 08:00 a **18.10.2019 o 09:59** pre cvičenie o 10:00. Každý týždeň omeškania je penalizovaný stratou dvoch bodov.

Dáta

Dáta dostanete v *.csv* súbore, kde prvý riadok je popis stĺpcov a nasledujú číselné hodnoty pre jednotlivé krajiny. Pár poznámok:

- Dáta boli pozliepané z dvoch zdrojov:
  - **World Happiness Index 2017** (po stĺpec Trust in Government) o projekte; východiskové spracovanie.
  - **UNdata 2018** o projekte; východiskové spracovanie.
- Celkovo obsahuje súbor 51 ukazovateľov pre 145 krajín.
- Pre niektoré ukazovatele (v posledných 12 stĺpoch) nie sú dáta kompletné - nedali sa získať - *je na vás, ako si s tým poradíte.*
- Ak sa stretnete so stĺpcom s textovou hodnotou, je potreba ju nahradiť za číselnú.
- Bližší popis jednotlivých stĺpcov je na poslednej strane zadania.

Úlohy

1. **Načítajte dáta a pripravte ich na spracovanie**. Načítajte dáta z .csv do vášho programu - vyberte vhodnú štruktúru, aby s nimi vedeli vami zvolené algoritmy/modely/knižnice ďalej pracovať. Máte možnosť vybrať si podmožinu dát (do

rozumnej miery). Nezabudnite dáta normalizovať (inak výsledky nebudú výpovedné). **1b**

2. **Oboznámte sa so základnými používanými metrikami.** Vyberte si metriku (napr. L1, L2) a pomocou nej zostrojte maticu podobností krajín (stačí podmnožina krajín aj ukazovateľov). Potešíme sa, ak si vyberiete takú podmnožinu ukazovateľov, aby ste dostali čo najzaujímavejšie výsledky. **2b**

3. **Využite algoritmy zhlukovania na nájdenie skupín podobných krajín**. Vašim cieľom by malo byť, aby váš model rozdelil krajiny podľa ich geografickej polohy (stĺpec *Region* v originálnych dátach) →podľa toho vyberajte modely, parametre a vstupné dáta. Vyskúšajte aspoň jeden zhlukovací algoritmus s dopredu uvedeným počtom zhlukov (napr. *k-means*) a s dopredu neznámym počtom zhlukov (napr. *DBSCAN*). **1.5b+1.5b**

4. **Analyzujte získané výsledky**. Vedeli ste sa priblížiť ku geografickému rozloženiu? Ak nie, prečo? Čo ste v rámci zadania vyskúšali, aby tomu tak bolo? Analyzujte zhluky pre aspoň dve rôzne spustenia/nastavenia - aké sú veľké, čo je v nich, ktorá krajina je vhodným reprezentantom (je najbližšie k centroidu). **2b**

Nepovinné úlohy

- Vyskúšajte si zhlukovanie pomocou SOM (samoorganizujúcej sa mapy). **2b**
- Implementujte zhlukovací algoritmus „sami“ - teda vlastným algoritmom. **1b**

Poznámky, spresnenia, odkazy

- Zadanie má tri časti:

  1. vytvorenie kódu a spracovanie vzoriek
  2. napísanie dokumentácie
  3. osobné odovzdanie na cvičení

  Aby bolo zadanie považované za odovzdané je potreba spraviť každú časť. Nepodceňte dokumentáciu - je potreba sa v nej vyjadriť ku vstupom a výstupom a každej podúlohe v zadaní (aspoň niekoľkými slovami - metódy, výsledky, ak sa to hodí aj obrázky).

- Je potrebné, aby bolo zadanie pred cvičením, na ktorom sa chystáte odovzdávať, nahraté v AIS v prislúchajúcom mieste odovzdania.
- Zadanie bude obodované na cvičení po prezentácii pred cvičiacim. **Pochopenie použitých metód a funkcií sa chápe ako prirodzená súčasť zadania a neschopnosť zodpovedať na otázky o týchto metódach je penalizované stratou bodov z danej časti,** a to aj v prípade, že kód je funkčný a správny.
- Dobre čítajte dokumentáciu metód, ktoré používate - napr. koľkorozmerný vstup očakávajú - upravte podľa toho vstupy.
- Nie ste hodnotení na základe úspešnosti vašich modelov, ale pri zlých výsledkoch je očakávaná aspoň snaha ich zlepšiť a pochopenie, prečo tomu tak bolo.
- Môže vám pomôcť:
  - towardsdatascience.com - Importance of distance metrics in machine learning modelling.
  - medium.com - How, When and Why Should You Normalize / Standardize / Rescale Your Data?
  - Stanford CS221 - K means
  - naftaliharris.com - Visualizing K-Means Clustering
  - OpenCV dokumentácia - kmeans - python
  - OpenCV dokumentácia - kmeans - c++
  - DLib dokumentácia - kmeans - C++
  - naftaliharris.com - Visualizing DBSCAN Clustering
  - Scikit-learn dokumentácia - dbscan - python
  - DLib dokumentácia - Chinese whispers - C++
  - Chris Biemann - Chinese Whispers - an Efficient Graph Clustering Algorithm and its Application to Natural Language Processing Problems

Popis datasetu

0. *Country* - názov krajiny, zrejme treba priradiť ID a pri počítaní vzdialenosti neuvažovať.

1. *Region* - oblasť podľa World Happiness Index; pri clusteringu neuvažujem, pri vzdialenostiach môžem (ale nie triviálne) - spolu 10 svetových oblastí.

2. *Happiness rank* - poradie krajiny v rebríčku najštastnejších krajín

3. *Happiness score* - dosiahnuté skóre krajiny - súčet stĺpcov 4-9 + dystopian residual.

4–9. *Ukazovatele šťastia* - samohodnotené ľudmi z danej krajiny.

10 *Region* oblasť podľa UNData - spolu 19 oblastí.

11 *Surface area* - rozloha v km^2.

12–18. *Obyvateľstvo krajiny* - počet, rozdelenie muži-ženy, vekové rozdelenie, rýchlosť rastu počtu, podiel imigrantov.

19–27. *Ekonomické ukazovatele* - HDP, podiel sektorov, import/export.

28–31. *Zamestnanosť* - podiel pracujúcich v sektorov, výška nezamestnanosti.

32–35. *Kvalita života* - dĺžka života (m/ž), pôrodnosť, novorodenecká úmrtnosť.

36–39. *Vyspelosť krajiny* - populácia v mestách, telekomuničné prostriedky.

40–42. *Životné prostredie* - emisie, ohrozené druhy, spotreba energie

43–54. *Neúplné (ale zaujímavé informácie)* - podiel žiakov v školách, počet žien v parlamente, počet lekárov, počet utečencov, zalesnené oblasti