AKADEMIA INNOWACYJNYCH ZASTOSOWAŃ TECHNOLOGII CYFROWYCH (AI TECH)

„Uczenie maszynowe” – laboratorium

Laboratorium 4 - Algorytmy Grupowania Danych

Data aktualizacji: 27.04.2024

Cel ćwiczenia

Celem ćwiczenia jest zapoznanie się z wybranymi algorytmami uczenia nienadzorowanego w ramach zadania grupowania (ang. *clustering*, czasem zwane klasteryzacją). Treść zadania obejmuje pracę z dwoma algorytmami grupowania: k-means (k-średnich) oraz DBSCAN. W ramach realizacji zdania zbadane zostaną różne konfiguracje algorytmów, użyte zostaną 3 miary oceny jakości grupowania oraz metody wizualizacji danych.

Wprowadzenie

Uczenie nadzorowane (lub uczenie z nauczycielem) związane jest z tworzeniem modelu na podstawie danych, który dokładnie wie jakie jest wejście i wyjście modelu (np. w klasyfikacji mamy oznaczenie klasy dla każdego rekordu), a algorytm w procesie uczenia budując model wykorzystuje tę wiedzę. W uczeniu nienadzorowanym brakuje takiej informacji (brak nauczyciela) i właśnie z taką sytuacją mamy do czynienia w zadaniu grupowania. Mamy dane (zbiór rekordów), dla których brak jest oznaczeń (klas), a zadanie sprowadza się do ich pogrupowania.

Przy zadaniu grupowania pojawiają się pytania. Jak pogrupować? Co znaczy dobre pogrupowanie danych? Czy powinno być dużo klastrów? Czy klastry powinny skupiać tylko podobne dane? Jak bardzo klastry powinny być „daleko” w przestrzeni danych? Na te pytania

odpowiedzi są różne, w zależności od analizowanego zbioru danych oraz przyjętej miary jakości klasteryzacji.

Przebieg Ćwiczenia

1. Wczytanie zbioru danych Abalone [1]. Dokładna analiza zbioru danych, można wykorzystać narzędia typu AutoEDA, np. [2, 3]. Napisać wnioski.
2. Uruchomienie algorytmu K-Means [4] oraz DBSCAN [5]. Policzenie miar ewaluacji modelu, np. Silhouette, Variance Ratio Criterion, Davies-Bouldin Index. Więcej informacji w [6].
3. Dokładne zapoznanie się ze sposobem działania metod oraz metryk.
4. Wizualizacja i porównanie wyników obu metod.
    1. Wykorzystując tylko dwie wybrane zmienne.
    2. Wykorzystując do wizualizacji algorytm redukcji wymiarowości.
5. Analiza hiperparametrów metod wraz z odpowiednimi wizualizacjami danych.

Punktacja

Przy realizacji zadania student może otrzymać **max 10 punktów** wedle poniższej tabeli.

2	Realizacja Ćwiczenia 1.
2	Realizacja Ćwiczenia 2.
2	Realizacja Ćwiczenia 3.
2	Realizacja Ćwiczenia 4.
2	Realizacja Ćwiczenia 5..

Pytania Pomocnicze

1. Czy przy grupowaniu potrzebna jest normalizacja/standaryzacja danych?

2. Co różni oba algorytmy z punktu widzenia reprezentacji klastra?
3. Który z algorytmów jest mniej odporny na szum i wartości odstające (ang. *outliers*)? Dlaczego?
4. Czy w zadaniu grupowania powinniśmy użyć walidacji krzyżowej?
5. Czy wyniki badanych algorytmów klasteryzacji powinny być powtarzane i uśredniane?
6. Co mierzą miary klasteryzacji podane w treści zdania?
7. Jak algorytmy zachowują się dla skrajnych wartości ilości klastrów?

Literatura

1. Zbiór danych Abalone. https://archive.ics.uci.edu/dataset/1/abalone
2. AutoEDA (ydata-profiling) - https://github.com/ydataai/ydata-profiling
3. AutoEDA (sweetviz) - https://github.com/fbdesignpro/sweetviz
4. K-Means - https://scikit-learn.org/stable/modules/clustering.html#k-means
5. DBSCAN - https://scikit-learn.org/stable/modules/clustering.html#dbscan
6. Miary ewaluacji metod klastrowania - https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation