AKADEMIA INNOWACYJNYCH ZASTOSOWAŃ TECHNOLOGII CYFROWYCH (AI TECH)

„Uczenie maszynowe" – laboratorium

Laboratorium 1

Przetwarzanie danych

data aktualizacji: 05.03.2023

Cel ćwiczenia

Przetwarzanie danych na potrzeby budowania modeli uczenia maszynowego.

Zagadnienia opracowywane w zadaniu: redukcja wymiarów poprzez PCA / t-SNE, transformacje danych, czyszczenie danych, problem wartości odstających (outlierów). Uruchomienie 2 wybranych modeli klasyfikacji/grupowania. Analiza uzyskanych wyników.

Wprowadzenie

Przy realizacji tego zadania nacisk jest położony na zbiór „Polish companies bankruptcy" [3], który jest „trudniejszy", tj. przed budową modelu klasyfikacji (lub grupowania / klastrowania) wymaga dodatkowych prac związanych z obróbką danych. Aby tego dokonać należy zapoznać się ze zbiorem i przeanalizować jego zawartość.

Zbiór składa się z 5 plików (każdy z plików reprezentuje inny rok analizy) danych opisanych 64 atrybutami dotyczących analizy fundamentalnej spółek polskich. Zadaniem jest zbudowanie modelu klasyfikacji, który wskaże czy dana spółka zbankrutuje.

Wszystkie atrybuty są numeryczne i obejmują dane analizy finansowej (np. dochód, zysk, kapitał własny, sprzedaż itp.). Są rekordy z brakującymi danymi, a także wartości ujemne oraz o różnych dziedzinach.

Do trudności w realizacji tego zadania należy poradzenie sobie z tymi „niedogodnościami" i zbudować model klasyfikacji (lub klasteryzacji). Wstępny opis zbioru (i zbiór) znajduje się w [3], dokładniejszy opis i przykładowe analizy (i klasyfikacje) podano w pracy [8].

Redukcja wymiarów

Przy realizacji zadania sugeruje się użycie PCA oraz t-SNE.

Algorytmy różnią się znacząco modelem, działaniem[1] i uzyskiwanymi wynikami. Warto wykonać prace z dwoma modelami, ich wizualizację oraz porównanie efektów działania.

Sugerowane narzędzia

Zadanie może być realizowane przy pomocy języka python i bibliotek użytecznych przy manipulacji danymi oraz wizualizacji wyników (tabele/wykresy).

Wskazane narzędzia (python): jupyter, numpy, pandas, matplotlib / seaborn / plotly, w szczególności sklearn:

https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

Przebieg ćwiczenia

1. Badanie cech zbioru wraz z wizualizacją.
2. Wizualizacja zbiorów przy redukcji wymiarów poprzez PCA / t-SNE. Analiza wyników.
3. Preprocessing danych – modyfikacja składowych (atrybutów) zbioru - transformacja, czyszczenie danych, problem wartości odstających (outlierów). Analiza oraz porównanie różnych metod.
4. Przygotowanie zbioru dla budowy modelu.

[1] Porównanie działania PCA i t-SNE znajduje się w [4] i [5].

5. Uruchomienie 2 wybranych modeli klasyfikacji lub grupowania. Analiza wyników oraz wpływu decyzji podjętych w obszarze czyszczenia zbioru.

Punktacja

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

2	Analiza zbioru danych - wizualizacja
1	Użycie PCA + wizualizacja wyników
1	Użycie t-SNE + wizualizacja wyników
4	Czyszczenie zbioru: brakujące dane, standaryzacja / normalizacja, analiza danych – outliery.
2	Użycie dwóch wybranych algorytmów klasyfikacji lub grupowania. Analiza wyników oraz wpływu decyzji podjętych w obszarze czyszczenia zbioru.

Przy realizacji tego zadania wystarczy prosty raport PDF utworzony przy użyciu Jupyter. Przy wynikach badań należy dać komentarz, podać wnioski i podsumowanie.

Pytania dodatkowe

1. Na czym polega metoda PCA?
2. Na czym polega metoda t-SNE? Jaka jest fundamentalna różnica względem PCA?
3. Na czym polega standaryzacja danych oraz normalizacja danych? Jakie są różnice pomiędzy tymi metodami? Jaki wpływ mają poszczególne transformacje danych na ostateczne wyniki modeli?
4. Na czym polega wybrana metoda detekcji obserwacji odstających? Jaki wpływ na wyniki ma wybrana metoda? Jak wybór metody obserwacji odstających wpływa na podział zbioru danych na zbiór treningowy oraz zbiór testowy?
5. Na czym polegają wybrane metody klasyfikacji lub grupowania danych?

Literatura

1. Materiały do wykładu
2. Cichosz P. ”Systemy uczące się”, WNT Warszawa
3. Zbiór danych https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data
4. Różnica PCA a t-SNE
   https://www.geeksforgeeks.org/difference-between-pca-vs-t-sne/
5. Kiedy używać PCA czy t-SNE?
   https://stats.stackexchange.com/questions/238538/are-there-cases-where-pca-is-more-suitable-than-t-sne
6. Tutorial do t-SNE
   https://www.datacamp.com/community/tutorials/introduction-t-sne
7. Tutorial do PCA/t-SNE w python

   https://towardsdatascience.com/an-introduction-to-t-sne-with-python-example-5a3a293108d1
8. Zieba, M., Tomczak, S. K., & Tomczak, J. M. (2016). Ensemble Boosted Trees with Synthetic Features Generation in Application to Bankruptcy Prediction. Expert Systems with Applications
   https://www.ii.pwr.edu.pl/~tomczak/PDF/%5BMZSTJT%5D.pdf
9. Zasoby Internetu, słowa kluczowe: uczenie maszynowe (machine learning), data mining, missing data, data cleaning, preprocessing, transformation, data standarization, data normalization, PCA, t-SNE