28.08.2020

UNIWERSYTET
MIKOŁAJA KOPERNIKA
W TORUNIU

Kurs PYTHON

Zmniejszenie wymiarowości danych

(Reducing data dimensionality)

Adrian Dzwonkowski
index: 272963
II rok, fizyka techniczna (s2), sp. cyfrowe systemy automatyki

Wstęp

Świat naukowy, niezależnie od jego kategorii jest nierozerwalnie powiązany z wszelkiego typu badaniami. Badania te pozwalają na nowe odkrycia czy bądź potwierdzenie jakiejś tezy. Obecnie wyniki eksperymentów najczęściej zapisywane są do postaci cyfrowej, gdzie dane te często są w postaci wielowymiarowej. Redukcja owej wymiarowości pozwala na prostszą analizę otrzymanych wyników, a osiąga się to np. poprzez uczenie maszynowe.
Celem projektu było napisanie programu w środowisku Python, którego zadaniem było zmniejszanie wymiarowości danych. Następnie należało wykreślić dane zredukowane do dwóch wymiarów i porównać z wykresem podanym w opisie projektu.

Opis projektu

Program wykorzystuje pakiety zawierające odpowiednie funkcje i parametry, które wymagane są do prawidłowego działania skryptu. Wykorzystywane są:

- Mdshare - pakiet mdshare zawiera oparty na Pythonie program do pobierania danych z dynamiki molekularnej z publicznego serwera FTP w FU Berlin. Projekt zakładał użycie danych: „alanine-dipeptide3x250ns-heavy-atom-distances.npz”. Występowały problemy z ich załadowaniem (więcej we wnioskach),
- NumPy - to biblioteka Pythona używana do pracy z tablicami. Posiada również funkcje do pracy w dziedzinie algebry liniowej, transformaty Fouriera i macierzy. Jest to pakiet typu ‘open source’ i można z niego swobodnie korzystać. Skrót NumPy oznacza numeryczny Python,
- Matplotlib - to biblioteka kreśląca dla języka programowania Python i jego rozszerzenia NumPy do matematyki numerycznej. Zawiera interfejs API do osadzania wykresów w aplikacjach przy użyciu zestawów narzędzi GUI ogólnego przeznaczenia, takich jak Tkinter, wxPython, Qt lub GTK +,
- argparse – moduł, którego zadaniem jest przygotowanie interfejsu do uruchomienia skryptu przy użyciu linii poleceń,
- t-SNE - technika redukcji wymiarowości, która szczególnie dobrze nadaje się do wizualizacji wielowymiarowych zbiorów danych. Technikę można zaimplementować za pomocą przybliżeń Barnes-Hut, co pozwala na zastosowanie jej w dużych, rzeczywistych zbiorach danych.

Skrypt podzielony jest na etapy, każdy z kroków ma przypisane odpowiednie komentarze tak, by łatwo było zrozumieć, która komenda, za co odpowiada. Na samym początku zaimplementowano pobieranie danych z serwera FTP. Z niewiadomych przyczyn funkcja ta nie działała, więc pobrano plik bezpośrednio do folderu z programem. W części zakomentowanej znajduje się kod odpowiadający za wczytanie danych z komputera PC. Dane zwracane są jako ndarray.

Projekt zakładał, że skrypt ma działać w trybie command-line. W tym celu wykorzystano pakiet argparse do przygotowania interfejsu za pomocą, którego użytkownik może wywołać program zmieniając dowolnie parametry. Aby uruchomić program i wywołać pomoc opisującą wszystkie parametry pobierane przez skrypt należy użyć polecenia: /python skrypt_python.py -h, które wywoła pomoc opisującą wszystkie parametry pobierane przez skrypt.

Program pobiera flagi:

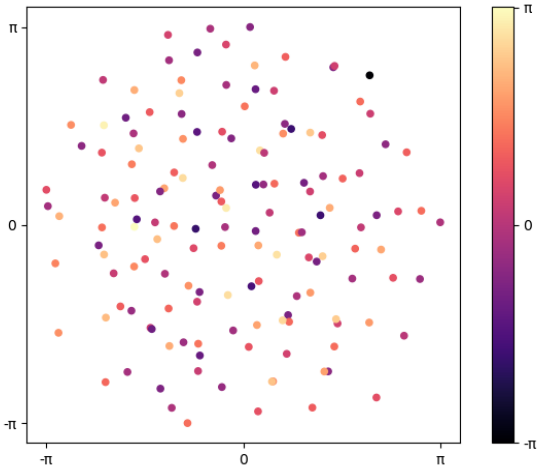
- step (-s) – Ustawienie kroku wczytywania danych. Im wyższy parametr tym więcej próbek zostanie zignorowanych. Bazowa wartość nastawiona na 500,
- dot_size (-ds) - Rozmiar punktów na wykresie. Bazowa wartość nastawiona na 20,
- x-scale (-x) - Wyświetla wykres w przedziale od -a do a (w osi x),
- y_scale (-y) - Wyświetla wykres w przedziale od -a do a (w osi y),
- alpha (-a) – Odpowiada za przezroczystość punktów. Parametr od 0 do 1; 0 - niewidoczne punkty; 1- nieprzezroczyste punkty. Bazowo ustawiono na 1.

Następnie projekt zakładał użycie funkcji $Y = fit(X)$, której zadaniem było przejęcie wielowymiarowego tensora i zwrócić jego rzutowanie na przestrzeń o niskim wymiarze. Wykorzystując pakiet t-SNE funkcja ta zmniejsza wymiarowość podanej tablicy ndarray. Odbywa się to za pomocą funkcji TSNE (). Funkcja ta przyjmuje różne argumenty, jednakże interesuje nas jedynie parametr odpowiadający za liczbę wymiarów, do których mają zostać zredukowane dane. W tym przypadku funkcja zwraca dwuwymiarową tablicę. Dane te są następnie skalowane, aby mieściły się w zakresie od −π do π i zwracane przez funkcję fit (X).

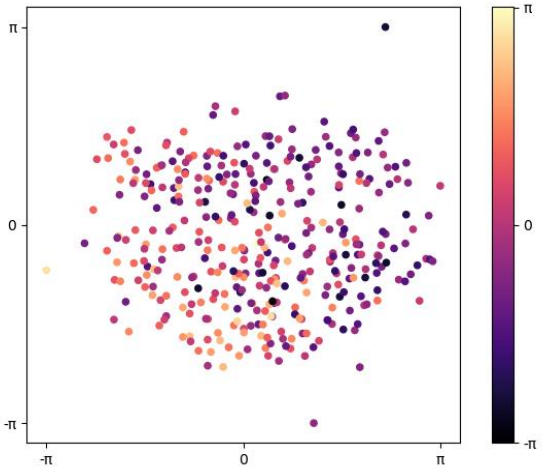
Na sam koniec otrzymane parametry przedstawiane są na wykresie punktowym wykonanym tak by był upodobniony do wykresu podanego w założeniach do projektu. Na koniec należało porównać oba te wykresy.

Wyniki

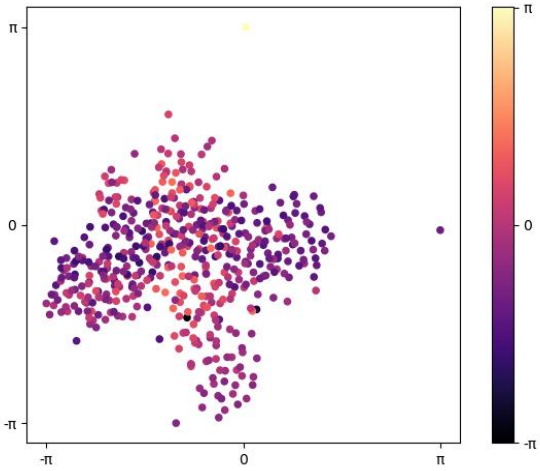
W tym punkcie zostaną przedstawione i omówione otrzymane wyniki. Otóż dane wykorzystane do projektu składają się z bardzo dużej ilości próbek. Jest to problematyczne, gdyż bazowanie na całości danych powoduje bardzo długi czas obliczeń, a czasami wręcz powoduje zacięcie całego komputera. Z tej też przyczyny należy odpowiednio ustawić parametr ‘step’, na którym głównie bazujemy podczas zmian jakichkolwiek parametrów. Należy tak dobrać parametr by otrzymane dane były wizualnie satysfakcjonujące oraz tak by czas obliczeń był jak najkrótszy. Przeprowadzono serię obliczeń, zdecydowano, że najlepszą nastawą ‘step’ będzie wartość w zakresie 1000-500. Seria została przeprowadzona dla wartości od 5000 do 50. Wartości powyżej 1000 ukazują zbyt małą gęstość punktów na wykresach, a dla wartości poniżej 500 czas obliczeń znacząco się wydłuża, szczególnie zaobserwowano to dla wartości 150-50, gdzie generowanie wykresu trwało kilka minut. Wyniki zilustrowano na obrazach (Rys. 1 – 9). Następnie wyniki przyrównano do obrazu bazowego (Rys. 10).



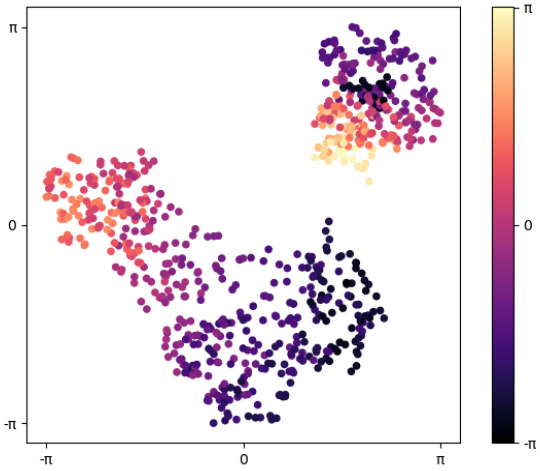
Rys.1 Wykres dla kroku 5000.



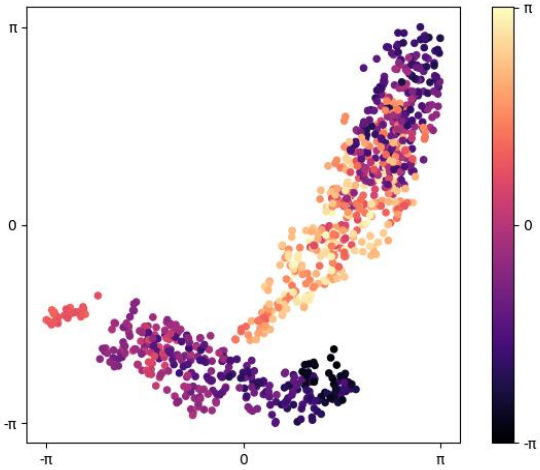
Rys.2 Wykres dla kroku 2000.



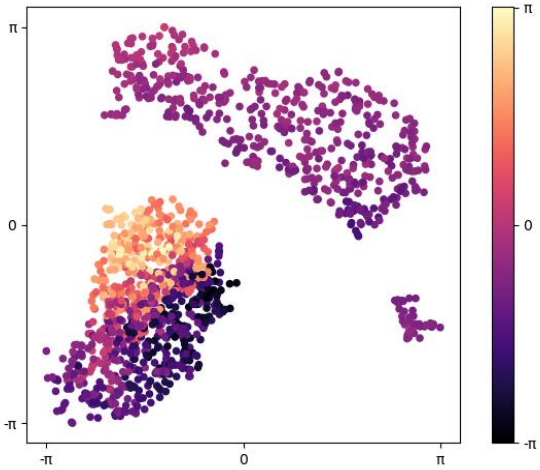
Rys.3 Wykres dla kroku 1500.



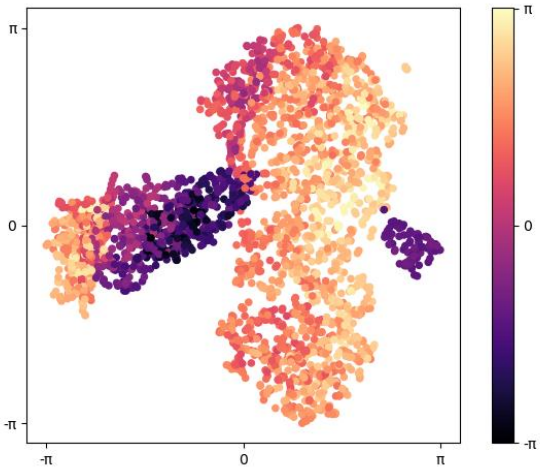
Rys.4 Wykres dla kroku 1000.



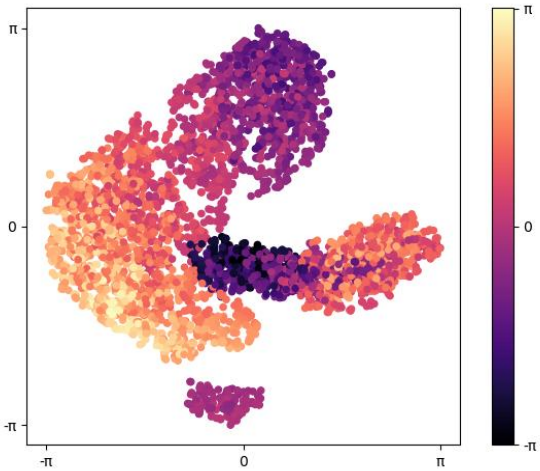
Rys.5 Wykres dla kroku 750.



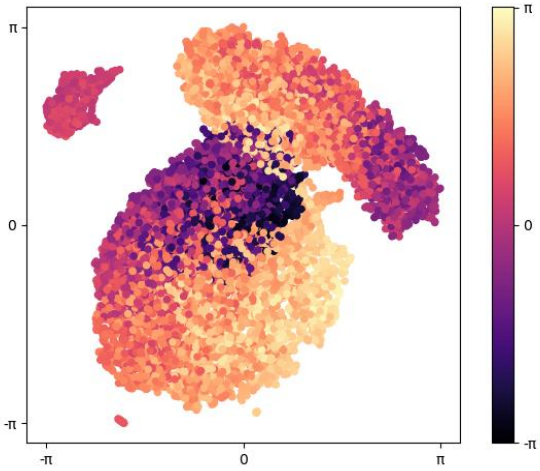
Rys.6 Wykres dla kroku 500.



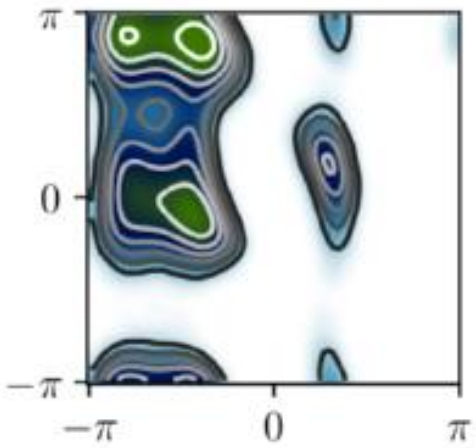
Rys.7 Wykres dla kroku 250.



Rys.8 Wykres dla kroku 150.



Rys.9 Wykres dla kroku 50.



Rys.10 Wykres bazowy.

Wnioski

Celem projektu było sprawdzenie nabytych umiejętności studenta, jego korzystania ze środowiska Python oraz porównanie projekcji wyników eksperymentu, na którego danych bazowano. Należało sprawdzić czy projekcja wykonana poprzez nauczanie maszynowe odda wyniki porównywalne do wyników bazowych. Przyglądając się otrzymanym wykresom można śmiało stwierdzić, że wyniki te znacznie się od siebie różnią, większość obszarów się ze sobą nie pokrywa, jednakże zaobserwowano wyraźne skupiska danych w pewnych obszarach, ich zagęszczenie zależne od nastawy parametru ‘step’.