Kowalska Katarzyna
Nr albumu: 288720

Projekt końcowy
Programowanie w Python

Cel

Projekt polegał na zredukowaniu wymiarowości wprowadzonych danych, z wielowymiarowej reprezentacji danych X do dwu lub trójwymiarowej i przedstawieniu realizacji w postaci wykresu.

Narzędzia i wykonanie

- Celem realizacji projektu wykorzystano bazę danych z publicznego serwera Freie Universität Berlin – „*alanine-dipeptide-3x250ns-heavy-atom-distances.npz*" przy użyciu biblioteki *mdshare*. Jest to baza reprezentująca dane pochodzące z dynamiki molekularnej – trzech trajektorii dipeptydu alaniny MD w postaci 3 obiektów numpy.ndarray, każdy o kształcie = [250000,3], dtype = numpy.float32, za pomocą odpowiedniego kodu zapisana została w jednej macierzy danych.
- Ponadto w projekcie wykorzystano przede wszystkim funkcję *PCA()* z biblioteki *sklearn*, która jako metoda (*ang. Principal Component Analysis*) znajduje najlepsze dopasowanie linii dla wybranych punktów – danych poprzez maksymalizowanie sumy kwadratów odległości między punktem danej a początkiem nowego układu współrzędnych. W ten sposób uzyskuje się dla każdej zmiennej dopasowanie – *PC1, PC2 (Principal Component*) itd. Dopasowana linia mówi o liniowej kombinacji dwóch zmiennych. Metoda opiera się na *SVD (ang. Singular Value Decomposition*). Obliczona zostaje macierz kowariancji, na podstawie której oblicza się wektory własne macierzy danych, dające informacje o kierunku rozproszenia tych danych oraz wartości własne reprezentujące wielkość tego rozproszenia. Aby określić istotność wpływu danej zmiennej metoda oblicza wariancję zmiennych. Komponenty (*Principal Components*) o mniejszym znaczeniu zostają zignorowane, wówczas otrzymuje się zestaw danych o mniejszym wymiarze niż oryginalny.
- Chcąc zwizualizować uzyskany rezultat zastosowano elementy z biblioteki *matplotlib.*
- Dodatkowo, przy użyciu zasobów biblioteki *argparse* umożliwiono manipulację wybranymi parametrami podczas programu.

Po wykonaniu redukcji wymiarowości, graficzną reprezentację otrzymanych danych przedstawiono na wykresie w folderze *etc* w repozytorium *https://github.com/hikkatie/288720_PPSeminar*. Obraz (*image_3D.png*) przedstawiono w przestrzeni trójwymiarowej, gdzie trzeci wymiar stanowi skala koloru. Na wykresie zastosowano skalę [-π, +π] dla każdej z reprezentowanych osi. Ponadto w folderze *etc* zachowano wykres słupkowy (*screeplot.png*) prezentujący zależność procentowego udziału wyjaśnionej wariancji (ang. *Executed variance*) dla każdego składnika głównego (PC). Na tej podstawie zauważono, że największy wpływ mają pierwsze dwa komponenty. W folderze *src* znajduje się właściwy kod programu.

Podsumowanie

Uzyskano zredukowaną wymiarowość wprowadzonych danych przy użyciu metody PCA, która obok tSNE (ang. *Stochastic Neighbor Embedding*) jest jedną z najczęściej wykonywanych do tego celu metod. PCA to szybka metoda, oparta nie na probabilistycznym podejściu jak w przypadku tSNE, a raczej na geometrycznym (związane z algebrą liniową). Porównując z reprezentacją zaprezentowaną w przykładzie do projektu, nie uzyskano na wykresie izolinii oraz rozmieszczenie danych jest odmienne, punktowe. Mimo to założenie redukcji wielu wymiarów zostało zachowane.