

# Laboratorium 7: Sieci neuronowe Kohonena (SOM)

W sieciach neuronowych Kohonena proces uczenia sieci przebiega bez nadzoru i opiera się wyłącznie na obserwacji danych wejściowych. Wykryte przez sieć skupienia danych reprezentowane są przez położone blisko siebie neurony warstwy wyjściowej. Neurony te tworzą mapę topologiczną danych wejściowych. Otrzymana mapa topologiczna stanowi podstawę analizy i oceny danych wejściowych i jest głównym wynikiem działania sieci. W ramach laboratorium własność samoorganizacji mapy topologicznej danych wejściowych zostaną wykorzystane do procesu klasteryzacji danych.

Sieci Kohonena zawsze posiadają warstwę wejściową i warstwę wyjściową (mapę topologiczną). Warstwa wyjściowa jest najczęściej dwuwymiarowa, o regularnej topologii.

W ramach laboratorium należy dokonać klasteryzacji następujących danych wykorzystując sieci Kohonena (polecenie `selforgmap`):

- 1) samodzielnie wygenerowanych sztucznych danych składających się z 3 klas abstrakcji
- 2) otrzymanego rzeczywistego zbioru danych mikromacierzowych, dla których dokonano redukcji wielowymiarowych danych metodą PLS. Dane składają się z macierzy o 3 kolumnach [supergen1, supergen2, D], gdzie supergen oznacza czynniki otrzymane metodą PLS, a D oznacza wektor przynależności obserwacji do klas określających typ raka.

Zbiór mikromacierzowy pochodzi z tkanek raka jajnika i zawiera trzy typy nowotworów ocenionych histopatologicznie: 1-clearcell, 2-endogenous, 3-serous. Istotą trudności w analizie danych mikromacierzowych jest ich wymiarowość. Najczęściej mamy do dyspozycji kilkadziesiąt do kilkuset próbek, przy kilku do kilkudziesięciu tysiącach analizowanych genów. Stąd konieczność zmniejszenia liczby analizowanych cech. Jedną ze stosowanych metod jest redukcja wymiarowości PLS. Metoda PLS pozwoliła dokonać liniowej transformacji wielowymiarowej przestrzeni cech (genów) do nowej dwuwymiarowej przestrzeni „super genów” stosując kryterium maksymalizacji kowariancji pomiędzy liniową kombinacją genów, a wektorem klas.

Dzięki temu w naszym przypadku możliwe jest zmniejszenie do 2 liczby analizowanych „super genów”.

Na podstawie wektora przynależności konkretnych próbek do klas należy ocenić wpływ

- liczby neuronów
- topologii sieci (proszę zmodyfikować strukturę sieci, lub zwiększyć liczbę klas abstrakcji)
- wyboru funkcji odległości
- liczby epok uczenia
- wymiarowości sieci

na prawidłowość podziału zbioru danych na zdefiniowane klasy.

**Sprawozdanie z ćwiczenia powinno zawierać:**

Wyniki symulacji

Kod napisanego programu!

Interpretację otrzymanych wyników wraz z ich graficznym przedstawieniem (plotsom)

Dwuetapową analizę danych (część nie jest konieczna do zaliczenia)

Wnioski

Czas oddania sprawozdania to 14 dni od daty laboratorium.