Bartosz Żak 236713
Michał Leśniak 195642

Poniedziałek 17:45
10.02.2022r.

# Komputerowa Analiza Danych
## Zadanie 3

# 1. Cel

Zadanie polega na implementacji, w zależności od wybranego wariantu zadania:

a. Algorytmu K-średnich

b. Samorganizującej się sieci neuronowej w dwóch wariantach metody jej nauki:

- Algorytm Kohonena
- Algorytm gazu neuronowego

W celu wykonania zadania należy wygenerować 2 zbiory danych:

a. Punkty leżące na okręgu o średnicy 2 ze środkiem w punkcie (0, 0) (200 punktów)

b. Punkty leżące na na jednym z dwóch okręgów o średnicy 1 ze środkami w punktach: (-3, 0) oraz (3, 0) (po 100 punktów na każdą figurę).

Dostępne warianty zadania opisane zostały poniżej. Wykonanie poszczególnych części determinuje maksymalną możliwą do uzyskania ocenę (należy przy tym również wykonać wszystkie części wymagane na ocenę niższą):

1) **Kwantyzacja przestrzeni za pomocą samorganizującej się sieci neuronowej (maksymalna ocena 4)**

W wariancie tym należy stworzyć aplikację pozwalającą w oparciu o te dane przeprowadzić proces uczenia samoorganizującej się sieci neuronowej. W tym celu należy zaimplementować i przeanalizować działanie algorytmu Kohonena lub algorytmu gazu neuronowego. Stworzony program powinien umożliwiać wizualizację rozkładu punktów treningowych oraz wizualizację procesu nauki neuronów (animacja prezentująca przebieg nauki). W sprawozdaniu należy zwrócić uwagę na następujące rzeczy:

- Jak wpływają parametry nauki (współczynnik nauki, sąsiedztwo) na jakość kwantyzacji?
- Jak na jakość kwantyzacji wpływa sposób inicjalizacji wag neuronów?
- Jak na jakość kwantyzacji wpływa liczba neuronów i jak dobrać ich optymalną liczbę?
- Czy wszystkie neurony biorą udział w kwantyzacji zbioru danych?

2) **Kwantyzacja przestrzeni za pomocą algorytmu K-średnich (maksymalna ocena 5)**

Wariant ten jest identyczny z wariantem poprzednim przy czym zamiast samorganizującej się sieci neuronowej należy skorzystać z algorytmu K-średnich. W sprawozdaniu należy zwrócić uwagę na następujące rzeczy:

- Jak na jakość kwantyzacji wpływa sposób inicjalizacji centrów?
- Jak na jakość kwantyzacji wpływa liczba centrów i jak dobrać ich optymalną liczbę?
- Czy wszystkie centa biorą udział w kwantyzacji zbioru danych?

# 2. Wyniki

## 2.1 Wykresy i tabele

### 2.1.1 Samoorganizująca się sieć neuronów.

| Algorytm | Kohonen | | Gaz Neuronowy | |
|---|---|---|---|---|
| Metoda Inicjalizacji | Random | Zeros | Random | Zeros |
| Średni błąd kwadratowy | 0.144 | 0.145 | 0.144 | 0.145 |
| Odchylenie standardowe błędu | 0.003 | 0.003 | 0.003 | 0.002 |
| Minimalny błąd kwantyzacji | 0.138 | 0.138 | 0.139 | 0.139 |
| Średni liczba martwych neuronów | 0.00 | 0.00 | 0.00 | 0.00 |
| Stand. odch. liczby martwych neuronów | 0.00 | 0.00 | 0.00 | 0.00 |

*Tabela 1. Statystyki SOM dla pierwszego zestawu danych, współczynnika nauki 0.1, promienia 1 oraz 20 neuronów.*

| Algorytm | Kohonen | | Gaz Neuronowy | |
|---|---|---|---|---|
| Metoda Inicjalizacji | Random | Zeros | Random | Zeros |
| Średni błąd kwadratowy | 0.133 | 0.132 | 0.137 | 0.136 |
| Odchylenie standardowe błędu | 0.002 | 0.002 | 0.002 | 0.002 |
| Minimalny błąd kwantyzacji | 0.130 | 0.130 | 0.131 | 0.132 |
| Średni liczba martwych neuronów | 0.00 | 0.00 | 0.00 | 0.00 |
| Stand. odch. liczby martwych neuronów | 0.00 | 0.00 | 0.00 | 0.00 |

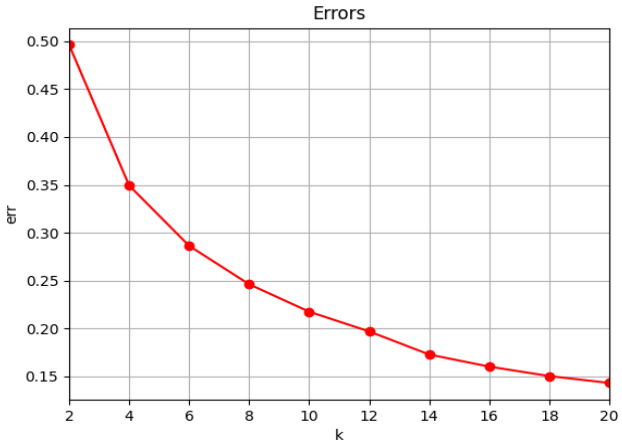
*Tabela 2. Statystyki SOM dla pierwszego zestawu danych, współczynnika nauki 0.33, promienia 0.5 oraz 20 neuronów.*

| Algorytm | Kohonen | | Gaz Neuronowy | |
|---|---|---|---|---|
| Metoda Inicjalizacji | Random | Zeros | Random | Zeros |
| Średni błąd kwadratowy | 0.112 | 0.113 | 0.114 | 0.112 |
| Odchylenie standardowe błędu | 0.005 | 0.005 | 0.003 | 0.002 |
| Minimalny błąd kwantyzacji | 0.103 | 0.104 | 0.106 | 0.106 |
| Średni liczba martwych neuronów | 2.32 | 2.38 | 2.26 | 2.00 |
| Stand. odch. liczby martwych neuronów | 1.24 | 1.38 | 0.66 | 0.00 |

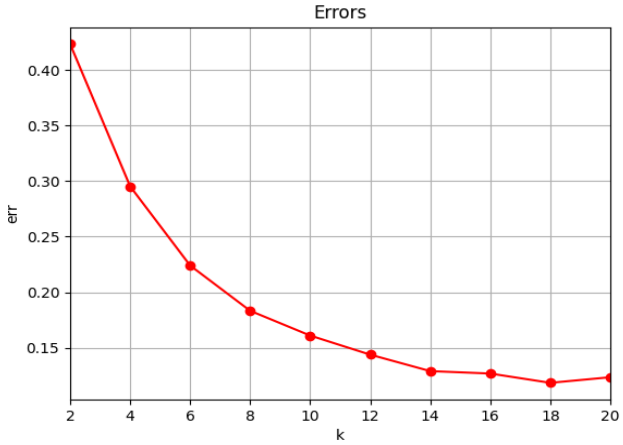
*Tabela 3. Statystyki SOM dla drugiego zestawu danych, współczynnika nauki 0.1, promienia 1 oraz 20 neuronów.*

| Algorytm | Kohonen | | Gaz Neuronowy | |
|---|---|---|---|---|
| Metoda Inicjalizacji | Random | Zeros | Random | Zeros |
| Średni błąd kwadratowy | 0.103 | 0.116 | 0.109 | 0.106 |
| Odchylenie standardowe błędu | 0.005 | 0.004 | 0.003 | 0.002 |
| Minimalny błąd kwantyzacji | 0.096 | 0.105 | 0.100 | 0.101 |
| Średni liczba martwych neuronów | 1.89 | 4.41 | 2.77 | 2.00 |
| Stand. odch. liczby martwych neuronów | 1.01 | 0.82 | 0.69 | 0.00 |

*Tabela 4 . Statystyki SOM dla drugiego zestawu danych, współczynnika nauki 0.33 , promienia 0.5 oraz 20 neuronów.*

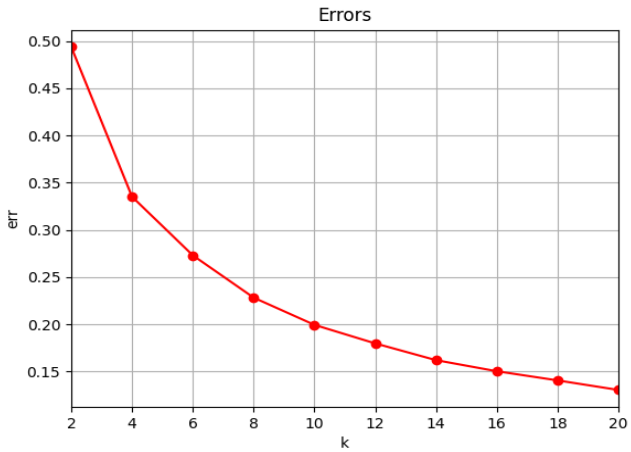



*Wykres 1. Zmiana średniego błędu kwantyzacji dla współczynnika nauki 0.1 , promienia sąsiedztwa 1, algorytmu Kohonena, losowej inicjalizacji i pierwszego zestawu danych.*




*Wykres 2. Zmiana średniego błędu kwantyzacji dla współczynnika nauki 0.1 , promienia sąsiedztwa 1, algorytmu Kohonena, losowej inicjalizacji i drugiego zestawu danych.*

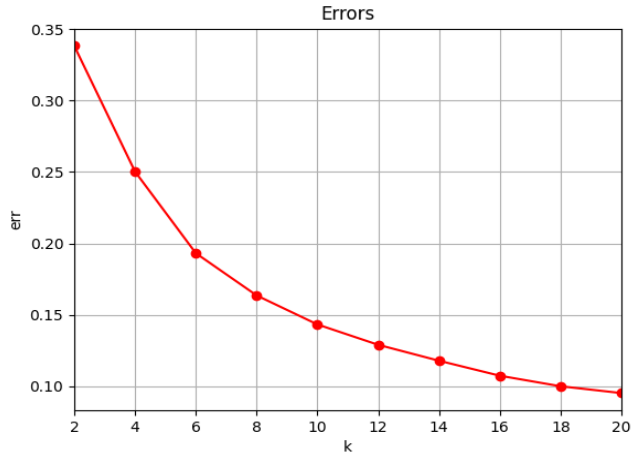
### 2.1.2 Algorytm K-średnich




*Wykres 3. Zmiana średniego błędu kwantyzacji dla różnych ilości centrów dla pierwszego zestawu danych i metody Forgy.*

| Metoda | Forgy | Random Partition |
|---|---|---|
| Średni błąd kwadratowy | 0.139 | 0.143 |
| Odchylenie standardowe błędu | 0.004 | 0.006 |
| Minimalny błąd kwantyzacji | 0.130 | 0.131 |
| Średni liczba pustych klastrów | 0.00 | 0.60 |
| Stand. odch. liczby pustych klastrów | 0.00 | 0.75 |

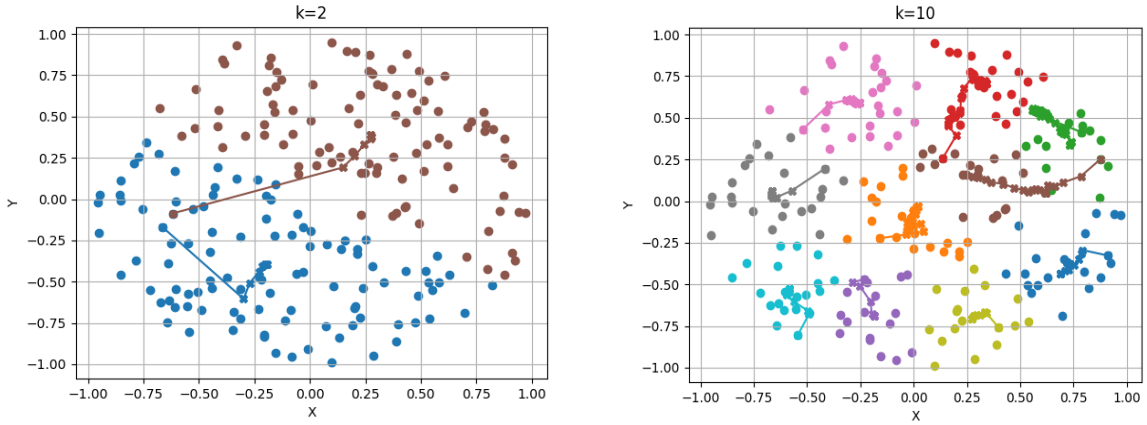
*Tabela 5. Statystyki dla pierwszego zestawu danych dla dwóch metod i k=20*



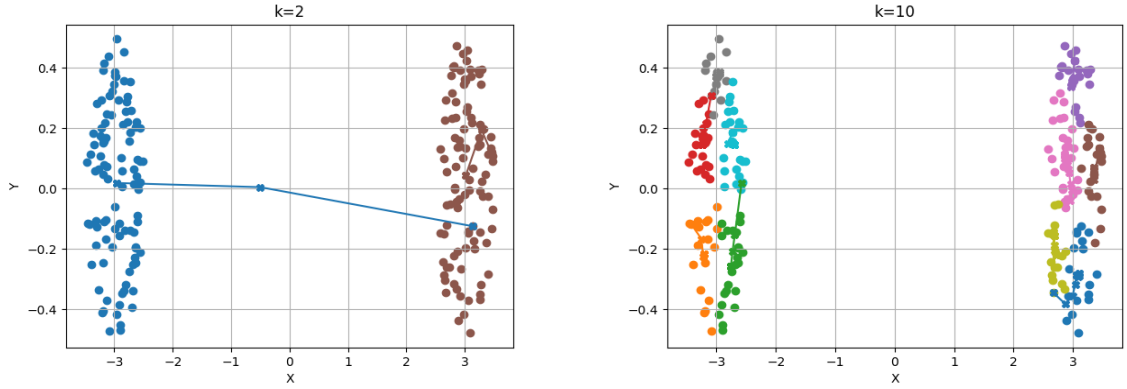

*Wykres 4. Zmiana średniego błędu kwantyzacji dla różnych ilości centrów dla drugiego zestawu danych i metody Forgy.*

| Metoda | Forgy | Random Partition |
|---|---|---|
| Średni błąd kwadratowy | 0.103 | 0.328 |
| Odchylenie standardowe błędu | 0.007 | 0.019 |
| Minimalny błąd kwantyzacji | 0.095 | 0.252 |
| Średni liczba pustych klastrów | 0.0 | 17.76 |
| Stand. odch. liczby pustych klastrów | 0.0 | 0.45 |

*Tabela 6. Statystyki dla drugiego zestawu danych dla dwóch metod i k=20*




*Rysunek 1. Zmiany położeń centrów dla metody Forgy i k=2 oraz k=10*




*Rysunek 2. Zmiany położeń centrów dla metody Random Partition  i k=2 oraz k=10*

## 2.2 Analiza wyników

### 2.2.1 Samoorganizująca się sieć neuronów

Do wykonania zadania użyliśmy dla porównania dwóch różnych algorytmów sieci neuronowych – Kohonena oraz algorytm gazu neuronowego. Dodatkowo inicjowaliśmy wagi neuronów również na dwa różne sposoby. Pierwszy o nazwie „random” polega na losowym przydziale wag neuronom z zakresu od najmniejszej możliwej wartości z danych treningowych do największej. Druga z nich o nazwie „zeros” po prostu przydziela każdemu neuronowi wektor zerowy.

Analizując wykres 1 i 2 można zauważyć, że wraz ze wzrostem liczby neuronów zmniejsza się średni błąd kwantyzacji. Na podstawie tych wykresów można dobrać optymalną liczbę neuronów wybierając miejsce, w którym wykres się stabilizuje i wraz z kolejnymi neuronami błąd już nie zmniejsza się.

Nie wszystkie neurony muszą brać udział w kwantyzacji. Może się zdarzyć taka sytuacja, że niektóre neurony nigdy nie wygrają konkursu o reprezentowanie punktu treningowego. Wtedy taki neuron nazywamy martwym. W celu minimalizacji liczby martwych neuronów wykorzystaliśmy mechanizm zmęczenia, który nakłada „karę” na zwycięski neuron, która powoduje zwiększenie jego faktycznej odległości od punktu treningowego, ale tylko przy porównywaniu odległości w celu znalezienia zwycięskiego neuronu. W innych przypadkach używana jest standardowa odległość bez „kary”. Na podstawie danych z tabeli 1 i 2 widać, że dla pierwszego zestawu danych mechanizm całkowicie eliminuje martwe neurony. Dla 2 zestawu danych jedynie minimalizuje ich liczbę. Istotnie jeśli spojrzymy na tabelę 3 i 4 liczba martwych neuronów raczej nie przekracza 2 na 20, ale zazwyczaj jest nawet mniejsza.

Widać również, że sposób inicjalizacji wag neuronów wpływa w naszym przypadku w minimalny sposób na jakość kwantyzacji, gdyż różnice są niewielkie. Duży wpływ na jakość mają za to parametry nauki takie jak współczynnik nauki czy promień sąsiedztwa. Nie należy przesadzać z wartością współczynnika nauki gdyż za duży współczynnik może za bardzo zmieniać wagi i zaprzepaścić wcześniejsze postępy w nauce. Korzystamy z mechanizmu dynamicznego zmniejszania współczynnika nauki oraz promienia sąsiedztwa z wyżej wymienionych powodów.

### 2.2.2 Algorytm K-średnich

Do wykonania zadania użyliśmy dwóch metod inicjalizacji centrów (Forgy i Random Partition). W metodzie Forgy inicjalizujemy centra losowo wybranymi wartościami z danych treningowych, podczas gdy w metodzie Random Partition najpierw losowo przydzielamy do każdego z centrów taką samą liczbę danych treningowych, a następnie uśredniamy ich wartości by otrzymać nowe położenie centrów.

Porównując obydwie metody dla dwóch zestawów danych (tabele 5. i 6.) możemy zauważyć, że metoda Forgy jest lepsza w przypadku drugiego zestawu danych czyli ma mniejsze błędy, a metoda Random Partition lepiej się sprawdza w przypadku pierwszego zestawu danych.

Należy zauważyć, że w przypadku metody Random Partition występują nieaktywne centra z czego jest ich znacznie więcej dla drugiego zestawu danych.

Jak wspomnieliśmy w poprzednim akapicie nie zawsze wszystkie centra biorą udział w kwantyzacji zbioru danych, przez co ważne jest dobranie ich optymalnej liczby i oczywiście dobrej metody ich inicjalizacji.

Wyraźnie widać na wykresach 3 i 4, że wraz ze wzrostem liczby centrów polepsza się jakość kwantyzacji poprzez zmniejszanie się średniego błędu kwantyzacji. Jednak czasami może nam zależeć na mniejszej liczbie klastrów. Zależy to od danej sytuacji. Popularną metodą wyboru optymalnej liczby centrów jest ‘elbow method’ polegająca na obserwacji zmiany błędów kwantyzacji wraz ze wzrostem liczby centrów i wyboru najmniejszej liczby centrów, po której wykres się stabilizuje. Na przykład dla pierwszego zestawu danych i metody Forgy byłaby to liczba 8.