Zadanie 3 - Inteligentna Analiza Danych

Adam Zambrzycki
`Nr indeksu: 216933`

Konrad Stępniak
`Nr indeksu: 216892`

6 czerwca 2019

Kierunek	Informatyka
Rok akademicki	2018/19
Semestr	4
Grupa dziekańska	2

Symbol α będzie oznaczał współczynnik nauki, a K liczbę centrów. Współczynnik skalujący w sieci będzie nazywany sigmą.

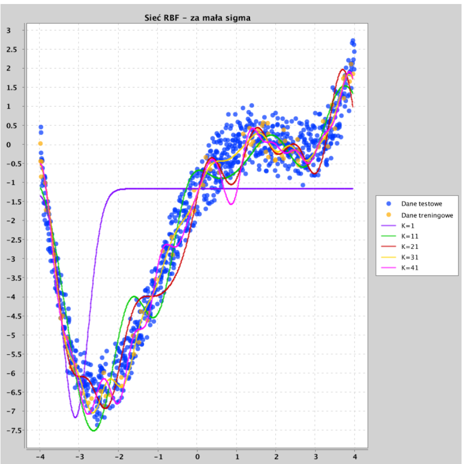
1 Osobna nauka warstw - Aproksymacja

Do nauki wykorzystano następujące parametry: $\alpha = 0.05$, liczba iteracji = 20000. Optymalną sigmę uzyskano ze wzoru:

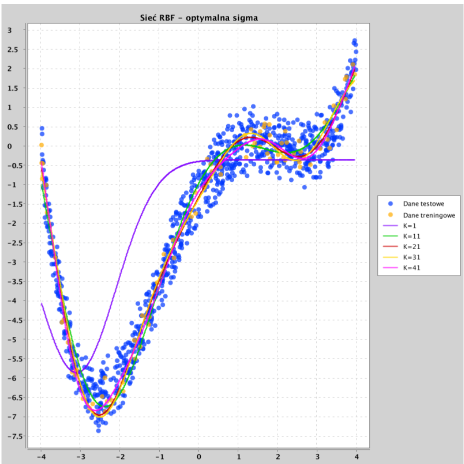
$$\sigma = \frac{d}{\sqrt{2M}} \tag{1}$$

Gdzie d - maksymalna odległość między centrami, a M to liczba centrów.

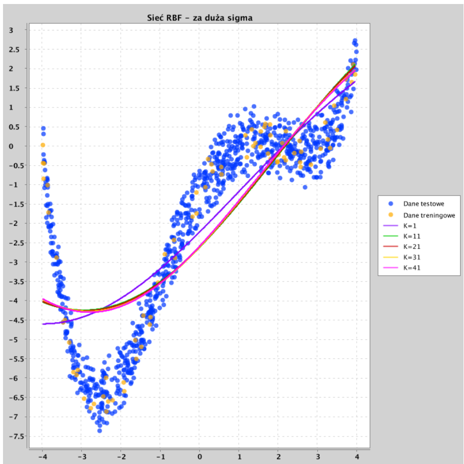
1.1 Podzadanie 1



Rysunek 1: Za mała sigma



Rysunek 2: Optymalna sigma

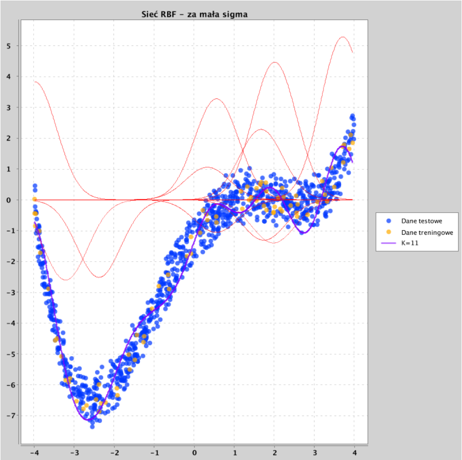


Rysunek 3: Za duża sigma

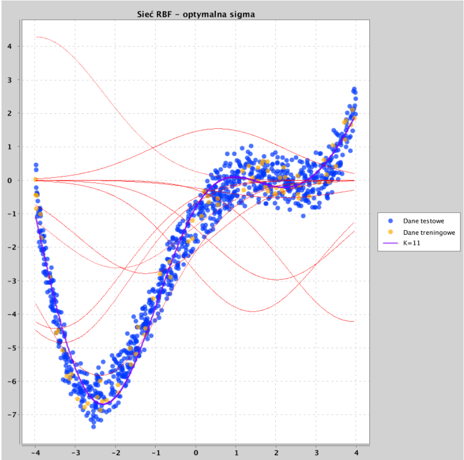
Dla liczby $K >= 11$, gdy współczynnik skalujący jest zbyt mały sieć przybliża dane treningowe do pewnego stopnia, jednak tworzy pewne zakłócenia, co daje niedokładne przybliżenie. Z drugiej strony, gdy sigma jest zbyt duża sieć praktycznie wcale nie aproksymuje danych jedynie je przecina w pewien sposób. Sieć o $K = 1$ daje takie same wyniki, jak $K = 41$. Gdy sigma jest optymalna, sieć o każdej liczbie neuronów oprócz $K = 1$, aproksymuje tak samo dokładnie.

1.2 Podzadanie 2

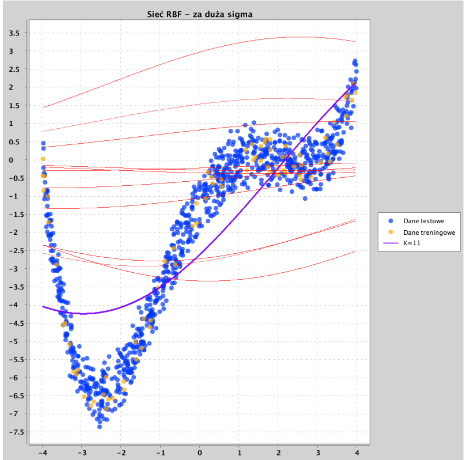
Czerwone linie reprezentują funkcje pojedynczych neuronów. Na podstawie poprzednich wyników wybrano $K = 11$. Na Rysunku 4 widać powód zakłóceń wykrytych na Rysunku 1. Gdy sigma jest za mała funkcje aktywacji neuronów mają zbyt wąski obszar aktywacji. Z drugiej strony na Rysunku 6 widać, że przy dużym



Rysunek 4: Za mała sigma



Rysunek 5: Optymalna sigma



Rysunek 6: Za duża sigma

współczynniku skalującym funkcje neuronów są prawie stałe. Nie da się stworzyć z kombinacji funkcji stałych dowolnej funkcji. Jeśli zaś dobierzemy sigmę optymalnie, neurony będą realizowały funkcje nie za wąską i nie ze szeroką, co umożliwi bardzo dokładną aproksymację jak na Rysunku 5.

1.3 Podzadanie 3

Symbole ϵ_a, ϵ_b oznaczają błędy średniokwadratowe odpowiednio dla zbioru treningowego i testowego. Symbol σ w Tabeli 1 oznacza odchylenie standardowe. Losowy wybór centrów neuronów wprowadza dodatkowy błąd

Tabela 1: Błąd średniokwadratowy oraz odchylenie dla zbioru treningowego i testowego dla 100 prób nauki

K	$avg(\epsilon_a)$	$\sigma(\epsilon_a)$	$avg(\epsilon_b)$	$\sigma(\epsilon_b)$
1	2.249	0.626	1.955	0.562
6	0.286	0.286	0.252	0.196
11	0.151	0.117	0.168	0.075
16	0.078	0.025	0.115	0.018
21	0.061	0.016	0.107	0.010
26	0.058	0.009	0.108	0.008
31	0.052	0.006	0.102	0.005
36	0.047	0.003	0.098	0.003
41	0.046	0.003	0.097	0.003

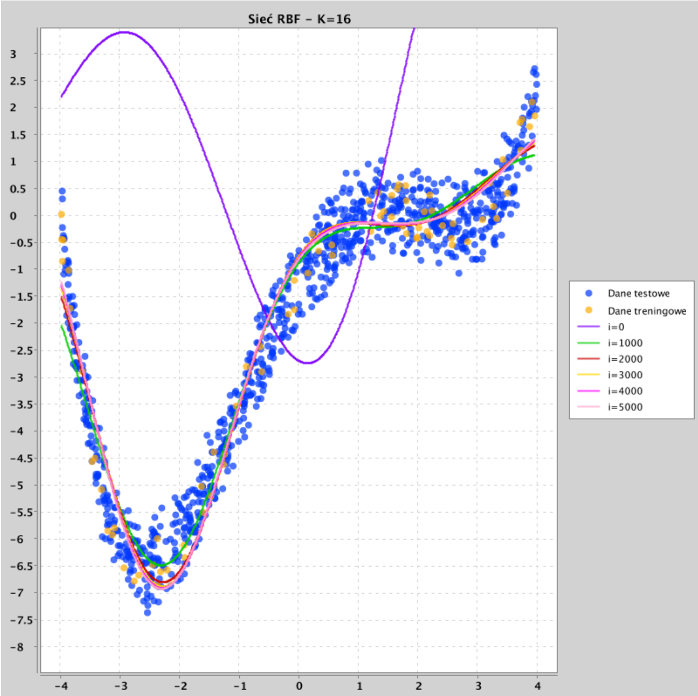
do sieci, gdyż przy małej ich liczbie mogą one zostać rozłożone nierównomiernie po całym zbiorze. Wynika stąd bardzo duże odchylenie standardowe dla $K = 1, 6, 11$. Gdy $K = 11$, błąd jest już relatywnie mały jednak nauka jest niestabilna, przez wspomniane losowanie centrów. Zauważmy, że dla $K = 16$, błąd zmniejsza się już dwukrotnie, a odchylenie standardowe jest bardzo małe. Taka liczba neuronów zapewnia już reprezentatywność danych treningowych. Przy $K = 26, 31, 36, 41$ średni błąd i odchylenie zmienia się nieznacząco w stosunku do $K = 16$. Za duża liczba neuronów nie zwiększa znacząco możliwości aproksymacyjnych sieci. Błąd i odchylenie na zbiorze testowym zawsze jest większe niż na zbiorze treningowym.

1.4 Podzadanie 4

Na podstawie obserwacji z Tabeli 1, Rysunek 7 został wygenerowany z parametrami $K = 16$, liczba iteracji $= 5000$. Można na nim wyraźnie zaobserwować, że już po 2000 iteracji funkcja realizowana przez sieć wyglądała tak samo jak dla $i = 5000$. Gdy błąd przestanie się znacząco zmniejszać należy przestać uczyć sieć, gdyż nie da to lepszych efektów.

1.5 Podsumowanie

- Błąd średniokwadratowy z każdą iteracją spada dosyć szybko, jednak w pewnym momencie osiąga próg, po którym dalsze uczenie sieci nie daje znaczących efektów.

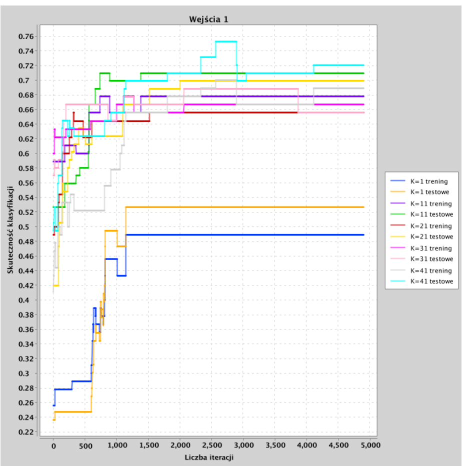


Rysunek 7: Zmiana funkcji sieci w różnych momentach

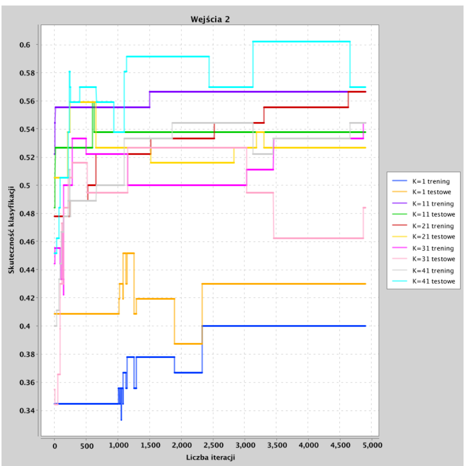
- Potrzebna liczba neuronów w warstwie radialnej jest zależna od charakterystyki danych jakie aproksymujemy. Za mała nie przybliży jej wcale, a za duża daje prawie takie same wyniki jak optymalna, którą należy dobrać eksperymentalnie.
- Sieć można uznać za nauczoną, gdy jej błąd przestanie znacząco maleć. Zależy to od charakterystyki danych. Jest to dosyć ważny aspekt, gdyż wtedy można trenować sieć np. przy 5000 iteracjach, a nie 20000 iteracji, co zabiera czas i moc obliczeniową.
- Jeśli centra zostaną źle wylosowane to błąd znacząco się zwiększa, dlatego należy wybierać taką ich liczbe aby minimalizować błąd stworzony przez losowanie.
- Współczynnik skalujący ma ogromny wpływ na funkcje realizowaną przez sieć, gdy dobierzemy go źle to przy za małej wartości sieć będzie niedokładna. Przy za dużym współczynniku nie będzie aproksymowała wcale. Należy go dobierać tak, aby był optymalny.

2 Osobna nauka warstw - klasyfikacja

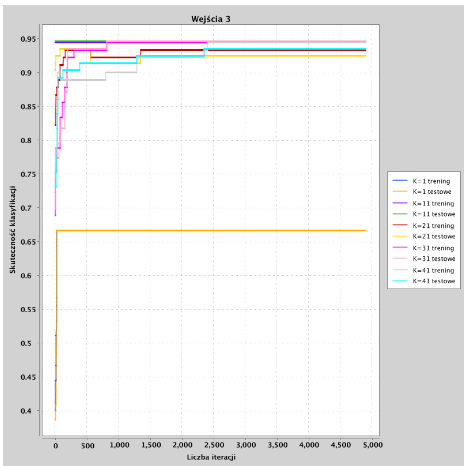
2.1 Podzadanie 1



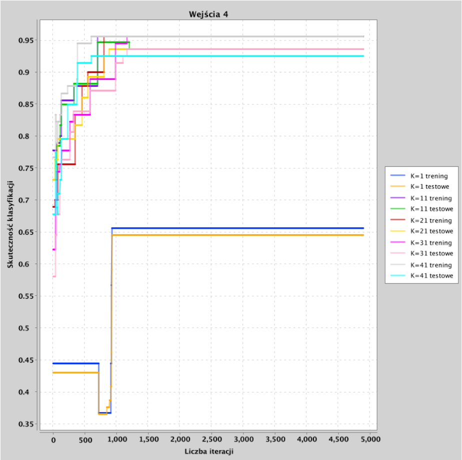
Rysunek 8



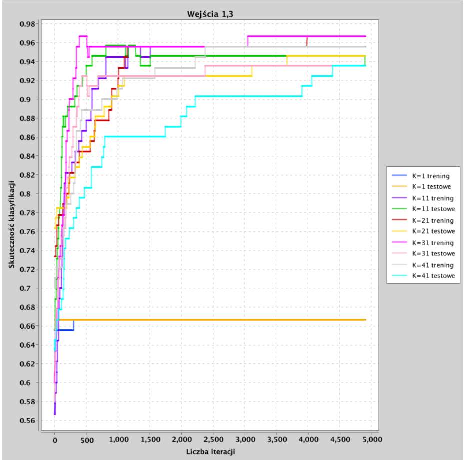
Rysunek 9



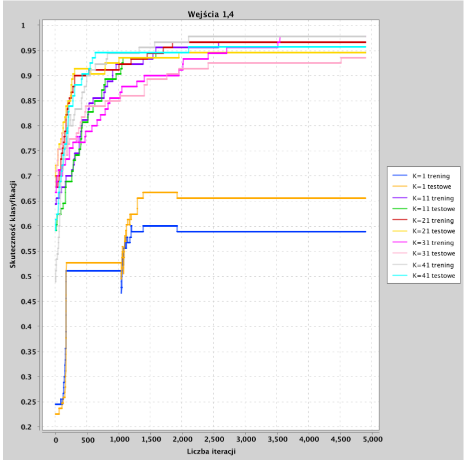
Rysunek 10



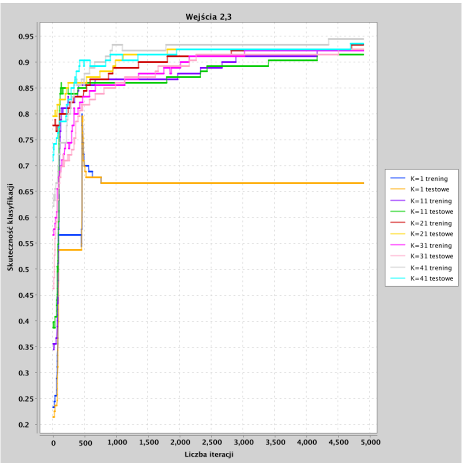
Rysunek 11



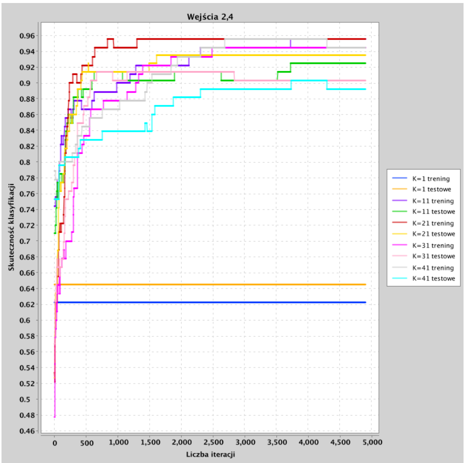
Rysunek 12



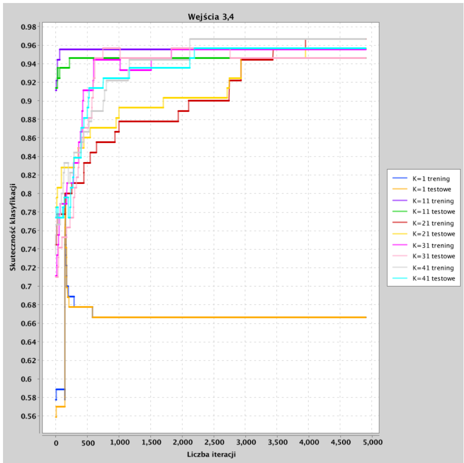
Rysunek 13



Rysunek 14



Rysunek 15



Rysunek 16

3 Nauka obu warstw

$$\frac{\partial}{\partial \sigma_k} = \frac{1}{N}\sum_{j=1}^{N}(f(x^j) - y^j) z_k(x^j) d^2(x^j, c_k)\frac{1}{\sigma_k^3} \tag{2}$$

3.1 Pochodna po współczynniku skalującym

$$\frac{\partial E}{\partial \sigma_k} = \frac{1}{2N}\sum_{j=1}^{N}(f(x^j) - y^j)^2 \tag{3}$$

$$= \frac{1}{2N}\sum_{j=1}^{N}2(f(x^j) - y^j)f'(x^j) \tag{4}$$

$$= \frac{1}{N}\sum_{j=1}^{N}f(x^j - y^j)\frac{\partial}{\partial \sigma_k}\sum_{k=0}^{K}w_k z_k(x^j) \tag{5}$$

$$\frac{\partial}{\partial \sigma_k}\sum_{k=0}^{K}w_k z_k(x^j) = \sum_{k=0}^{K}w_k z_k(x^j) = w_k\frac{\partial}{\partial \sigma_k}z_k(x^j) \tag{6}$$

Ponieważ tylko $w_k \frac{\partial}{\partial \sigma_k} z_k(x^j)$ zależy od σ_k, dla $k = 0, ..., K$

$$\frac{\partial}{\partial \sigma_k} z_k(x^j) = \frac{\partial}{\partial \sigma_k} exp(-d^2(x^j, c_k)\frac{1}{2\sigma_k^2}) \tag{7}$$

$$= z_k(x^j)\frac{\partial}{\partial \sigma_k}(\frac{1}{2\sigma_k^2}) \tag{8}$$

$$= z_k(x^j)\frac{-1}{2}d^2(x^j, c_k)\frac{\partial}{\partial \sigma_k}\sigma_k^{-2} \tag{9}$$

$$= z_k(x^j)\frac{-1}{2}d^2(x^j, c_k) - 2\sigma_k^{-3} \tag{10}$$

$$= z_k(x^j)d^2(x^j, c_k)\frac{1}{\sigma_k^3} \tag{11}$$

Ostatecznie

$$\frac{\partial E}{\partial \sigma_k} = \frac{1}{N}\sum_{j=1}^{N} f(x^j - y^j)z_k(x^j)w_k d^2(x^j, c_k)\frac{1}{\sigma_k^3} \tag{12}$$

3.2 Pochodna po koordynacie centrum neuronu

$$\frac{\partial E}{\partial c_{k,i}} = \frac{1}{2N}\sum_{j=1}^{N}(f(x^j) - y^j)^2 \tag{13}$$

$$= \frac{1}{2N}\sum_{j=1}^{N} 2f(x^j - y^j)f'(x^j) \tag{14}$$

$$= \frac{1}{N}\sum_{j=1}^{N} f(x^j - y^j)\frac{\partial}{\partial c_{k,i}}\sum_{k=0}^{K} w_k z_k(x^j) \tag{15}$$

$$\frac{\partial}{\partial c_{k,i}}\sum_{k=0}^{K} w_k z_k(x^j) = \sum_{k=0}^{K} w_k z_k(x^j) = w_k \frac{\partial}{\partial c_{k,i}} z_k(x^j) \tag{16}$$

Ponieważ tylko $w_k \frac{\partial}{\partial c_{k,i}} z_k(x^j)$ zależy od $c_{k,i}$, dla $k = 1, ..., K$

$$w_k \frac{\partial}{\partial c_{k,i}} z_k(x^j) = w_k \frac{\partial}{\partial c_{k,i}} exp(-d^2(x^j, c_k)\frac{1}{2\sigma_k^2}) \tag{17}$$

$$= w_k z_k(x^j)\frac{\partial}{\partial c_{k,i}} - d^2(x^j, c_k)\frac{1}{2\sigma_k^2} \tag{18}$$

$$= w_k z_k(x^j)\frac{-1}{2\sigma_k^2}\frac{\partial}{\partial c_{k,i}} d^2(x^j, c_k) \tag{19}$$

$$\frac{\partial}{\partial c_{k,i}} d^2(x^j, c_k) = \frac{\partial}{\partial c_{k,i}} \sqrt{\sum_{i=1}^{n}(x_i - c_{k,i})^2}^{\,2} \tag{20}$$

$$= \frac{\partial}{\partial c_{k,i}} \sum_{i=1}^{n}(x_i - c_{k,i})^2 \tag{21}$$

$$= \sum_{i=1}^{n} 2(x_i - c_{k,i})(-1) \tag{22}$$

$$= -2\sum_{i=1}^{n}(x_i - c_{k,i}) \tag{23}$$

$$= -2(x_i - c_{k,i}) \tag{24}$$

Ponieważ tylko $x_i - c_{k,i}$ zależy od $c_{k,i}$.

$$w_k z_k'(x^j) = w_k z_k(x^j)\frac{-1}{2\sigma_k^2} - 2(x_i - c_{k,i}) \tag{25}$$

$$w_k z_k'(x^j) = w_k z_k(x^j)\frac{1}{\sigma_k^2}(x_i - c_{k,i}) \tag{26}$$

$$\tag{27}$$

Ostateczny wzór:

$$\frac{\partial E}{\partial c_{k,i}} = \frac{1}{N}\sum_{j=1}^{N} f(x^j - y^j) w_k z_k(x^j) \frac{1}{\sigma_k^2}(x_i - c_{k,i}) \tag{28}$$