Metody inteligencji obliczeniowej w analizie danych
Sieci neuronowe II

Jan Karwowski

Wydział matematyki i Nauk Informacyjnych PW

9 marca 2021

Fundusze Europejskie
Wiedza Edukacja Rozwój

Rzeczpospolita Polska

Politechnika Warszawska

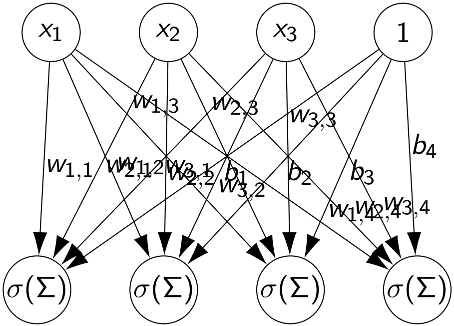
Unia Europejska
Europejski Fundusz Społeczny

Zadanie 10 pn. „Modyfikacja programów studiów na kierunkach prowadzonych przez Wydział Matematyki i Nauk Informacyjnych" realizowane jest w ramach projektu „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca" współfinansowanego ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego

Ustalić jak z oddawaniem sprawozdań.

Sieć neuronowa w zadaniu regresji

- Wektor cech: $X \in [0,1]^N$, wektor odpowiedzi $Y \in [0,1]^M$
- k-elementowy zbiór testowy $\subset X \times Y$
- Model (sieć) N i parametry (wagi) Θ
- $MSE = \sum_i \frac{(N(\mathbf{x}_i, \Theta) - \mathbf{y}_i)^2}{k}$
- Skąd wziąć wektor Θ?



$\Theta = (b_1, b_2, b_3, b_4, w_{1,1}, w_{1,2}, \ldots)$

Twierdzenie

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma \left(\sum_{i}^{n} x_i w_{i,j} + b_j \right)$$

Tw. Dla funkcji σ – sigmoidalnej – zbiór funkcji w postaci G (dla dowolnych N, α_j, w_i, b_j) jest gęsty w przestrzeni wszystkich funkcji ciągłych $[0,1]^n \to [0,1]$.

G. Cybenko. "Approximation by superpositions of a sigmoidal function". In: *Mathematics of Control, Signals, and Systems* 2.4 (1989), 303–314. ISSN: 1435-568X. DOI: 10.1007/bf02551274. URL: http://dx.doi.org/10.1007/bf02551274

MLP – Nomenklatura

- Warstwa
  - (Wejściowa)
  - Ukryta
  - Wyjściowa
- Aktywacja neuronu
- Uczenie sieci
- Wzorzec uczący – para $(\mathbf{x}, \mathbf{y})$ ze zbioru uczącego

Reguła Hebba I

Postulat: neruony, które często są pobudzone w tym samym momencie mają silne połączenia synaptyczne.

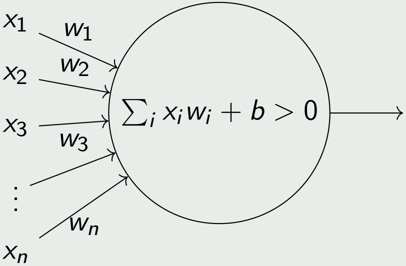
Donald O Hebb. *The organization of behavior*. 1949

Reguła Hebba II

Interpretacja dla perceptronu (pojedynczego)

Zakładamy, że wszystkie wartości w zbiorze uczącym są ze zbioru $\{-1, 1\}$.

$$w_i = \frac{\sum_{k=1}^{N} x_i^k y^k}{N}$$

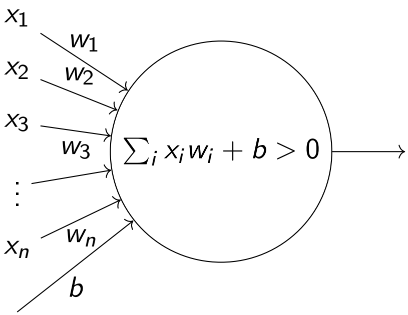


Reguła Hebba III

Efekt

Dla bardzo małej liczby wzorców uczących efekt jest zadowalający, w większych zbiorach sieć traci zdolność do uczenia się.

Reguła perceptronu I



Skokowa funkcja aktywacji

$$f(x) = \begin{cases} 0 & x \leqslant 0 \\ 1 & x > 0 \end{cases}$$

- Problem: wejście – zmienne ciągłe, wyjście – zmienne binarne.

Reguła perceptronu II

- Intuicja: zmienić wagi tak, że jeśli neuron nie jest aktywowany, a powinien być, to zmieniamy wagę tak, żeby zwiększyć jego szansę aktywacji.

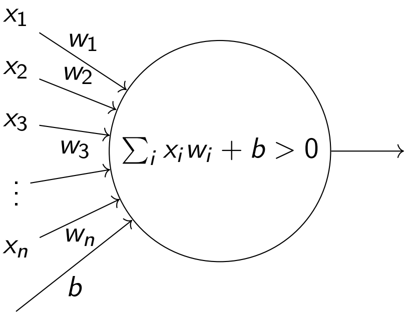
Metoda uczenia

y – klasa, $\hat{y}$ – wartość wyliczona przez perceptron, x_i i-te wejście. Rozważamy pojedynczy wzorzec uczący, prezentujemy kolejne wzorce, często wielokrotnie.

$$\Delta w_i = \begin{cases} x_i & \hat{y} < y \\ -x_i & \hat{y} > y \\ 0 & \hat{y} = y \end{cases}$$

$$\Delta_b = \begin{cases} 1 & \hat{y} < y \\ -1 & \hat{y} > y \\ 0 & \hat{y} = y \end{cases}$$

Jak można uprościć ten zapis?



Reguła delta I

- Wejście – ciągłe, wyjście ciągłe, funkcja aktywacji *różniczkowalna*.
- η – parametr: krok uczenia (także współczynnik uczenia)
- Ulepszenie reguły perceptronu
- W aktualizacji uwzględniamy wpływ nachylenia funkcji aktywacji na zmianę wyniku

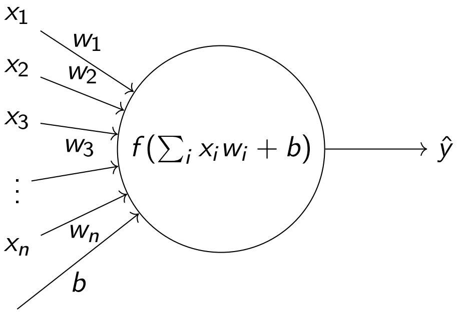
Reguła uczenia

y – wyjście wg. wzorca uczącego, $\hat{y}$ – wartość wyliczona przez perceptron, x_i i-te wejście.Rozważamy pojedynczy wzorzec uczący, prezentujemy kolejne wzorce, często wielokrotnie.

$$\Delta w_i = \eta(y - \hat{y})f'(\sum_j x_j w_j + b)x_i$$

$$\Delta b = \eta(y - \hat{y})f'(\sum_j x_j w_j + b)$$

Reguła delta II

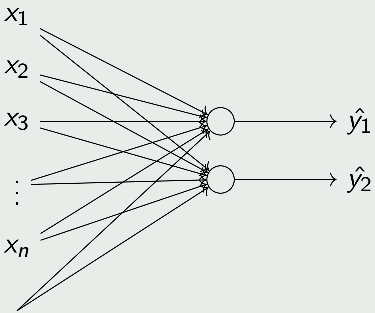


Reguła delta III

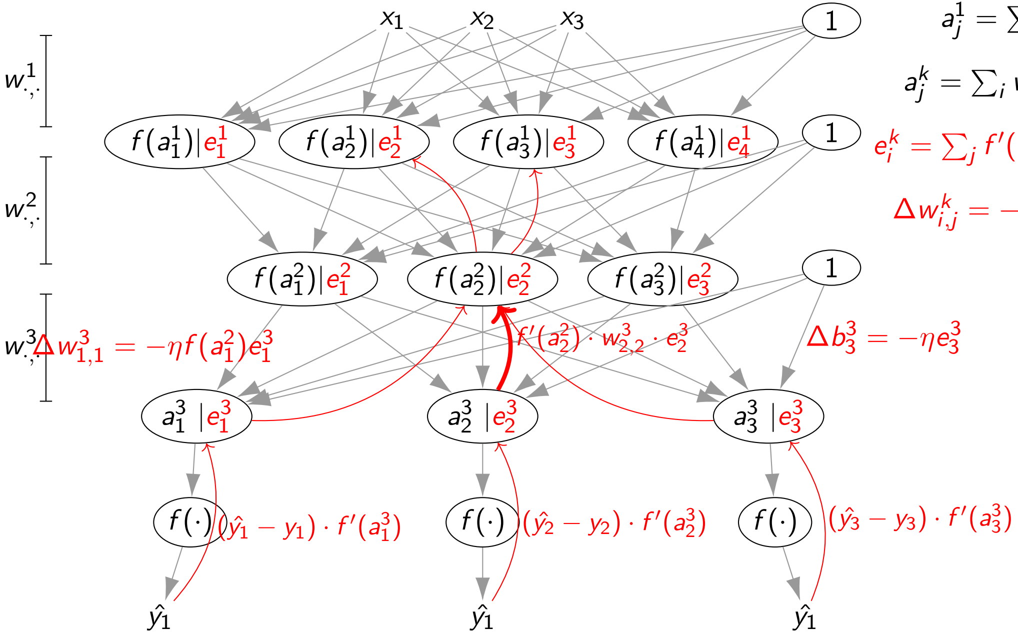
Perceptron – wiele wyjść

$$\Delta w_{i,j} = \alpha(y_j - \hat{y}_j)f'(\sum_k x_k w_k, j + b)x_i$$

$$\Delta b_j = \alpha(y_j - \hat{y}_j)f'(\sum_K x_K w_k, j + b)$$



Propagacja wsteczna błędu – MLP



$$a^1_j = \sum_i w^1_{i,j} x_i$$

$$a^k_j = \sum_i w^k_{i,j} f(a^{k-1}_i)$$

$$e^k_i = \sum_j f'(a^k_i) w^{k+1}_{i,j} e^{k+1}_j$$

$$\Delta w^k_{i,j} = -\eta f(a^{k-1}_i) e^k_j$$

Propagacja wsteczna błędu

Algorytm

$\Theta \leftarrow \mathit{InicjujLosowo}$;

while $\neg \mathit{StopCondition}$ **do**

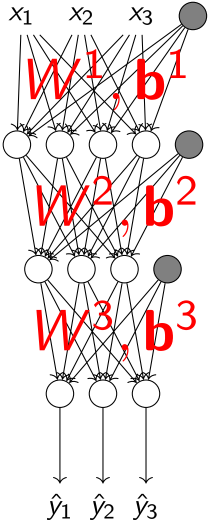
  for $(\mathbf{X}, \mathbf{Y}) \in$ *Zbiór uczący* **do**

    $\hat{Y} \leftarrow \mathit{FeedForward}(\Theta, X)$;

    $\Delta\Theta \leftarrow \mathit{Backpropagate}(\hat{Y}, Y)$;

    $\Theta \leftarrow \Theta + \Delta\Theta$;

Uczenie gradientowe I



η – krok uczenia (niektórzy używają α)
$\Theta = [W^1, \mathbf{b}^1, W^2, \mathbf{b}^2, \ldots \mathbf{b}^3]$

$$MSE(\Theta) = \frac{1}{2N} \sum_i (\hat{y}_1(\Theta, \mathbf{x}^i) - y_1^i)^2 + (\hat{y}_2(\Theta, \mathbf{x}^i) - y_2^i)^2 + (\hat{y}_3(\Theta, \mathbf{x}^i) - y_3^i)^2$$

$$\Delta\Theta = -\eta \nabla MSE(\Theta)$$

Wyprowadzenie pochodnej

$$MSE(\Theta) = \frac{1}{N}\sum_i \frac{1}{2}\left((\hat{y}_1(\Theta, \mathbf{x}^i) - y_1^i)^2 + (\hat{y}_2(\Theta, \mathbf{x}^i) - y_2^i)^2 + (\hat{y}_3(\Theta, \mathbf{x}^i) - y_3^i)^2\right)$$

Pochodna po wadze w ostatniej warstwie $w_{1,1}^3$

- $\frac{1}{2}\left((\hat{y}_1(\Theta, \mathbf{x}) - y_1)^2 + (\hat{y}_2(\Theta, \mathbf{x}) - y_2)^2 + (\hat{y}_3(\Theta, \mathbf{x}) - y_3)^2\right)$
- $\frac{1}{2}\frac{\partial(\hat{y}_1(\Theta,\mathbf{x})-y_1)^2}{\partial w_{1,1}^3} = \frac{1}{2}2(\hat{y}_1(\Theta, \mathbf{x}) - y_1) \cdot \frac{\partial \hat{y}_1(\Theta,\mathbf{x}^i)}{\partial w_{1,1}^3} =$
  $(\hat{y}_1(\Theta, \mathbf{x}) - y_1) \cdot \frac{\partial f(a_1^3)}{\partial w_{1,1}^3} =$

$\hat{y}_i = f(a_i^3)$
$a_j^k = \sum_i w_{i,j}^k f(a_i^{k-1})$
$a_j^1 = \sum_i w_{i,j}^1 x_i$

Wyprowadzenie pochodnej

$$MSE(\Theta) = \frac{1}{N}\sum_i \frac{1}{2}\left((\hat{y}_1(\Theta,\mathbf{x}^i) - y_1^i)^2 + (\hat{y}_2(\Theta,\mathbf{x}^i) - y_2^i)^2 + (\hat{y}_3(\Theta,\mathbf{x}^i) - y_3^i)^2\right)$$

Pochodna po wadze w ostatniej warstwie $w_{1,1}^3$

- $\frac{1}{2}\left((\hat{y}_1(\Theta,\mathbf{x}) - y_1)^2 + (\hat{y}_2(\Theta,\mathbf{x}) - y_2)^2 + (\hat{y}_3(\Theta,\mathbf{x}) - y_3)^2\right)$
- $\frac{1}{2}\frac{\partial(\hat{y}_1(\Theta,\mathbf{x}) - y_1)^2}{\partial w_{1,1}^3} = \frac{1}{2}2(\hat{y}_1(\Theta,\mathbf{x}) - y_1)\cdot\frac{\partial\hat{y}_1(\Theta,\mathbf{x}^i)}{\partial w_{1,1}^3} =$
  $(\hat{y}_1(\Theta,\mathbf{x}) - y_1)\cdot\frac{\partial f(a_1^3)}{\partial w_{1,1}^3} = (\hat{y}_1(\Theta,\mathbf{x}) - y_1)\cdot f'(a_1^3)\cdot\frac{\partial a_1^3}{\partial w_{1,1}^3} =$

$\hat{y}_i = f(a_i^3)$
$a_j^k = \sum_i w_{i,j}^k f(a_i^{k-1})$
$a_j^1 = \sum_i w_{i,j}^1 x_i$

Wyprowadzenie pochodnej

$$MSE(\Theta) = \frac{1}{N}\sum_i \frac{1}{2}\left((\hat{y}_1(\Theta,\mathbf{x}^i) - y_1^i)^2 + (\hat{y}_2(\Theta,\mathbf{x}^i) - y_2^i)^2 + (\hat{y}_3(\Theta,\mathbf{x}^i) - y_3^i)^2\right)$$

Pochodna po wadze w ostatniej warstwie $w_{1,1}^3$

- $\frac{1}{2}\left((\hat{y}_1(\Theta,\mathbf{x}) - y_1)^2 + (\hat{y}_2(\Theta,\mathbf{x}) - y_2)^2 + (\hat{y}_3(\Theta,\mathbf{x}) - y_3)^2\right)$
- $\frac{1}{2}\frac{\partial(\hat{y}_1(\Theta,\mathbf{x}) - y_1)^2}{\partial w_{1,1}^3} = \frac{1}{2}2(\hat{y}_1(\Theta,\mathbf{x}) - y_1)\cdot\frac{\partial \hat{y}_1(\Theta,\mathbf{x}^i)}{\partial w_{1,1}^3} =$
  $(\hat{y}_1(\Theta,\mathbf{x}) - y_1)\cdot\frac{\partial f(a_1^3)}{\partial w_{1,1}^3} = (\hat{y}_1(\Theta,\mathbf{x}) - y_1)\cdot f'(a_1^3)\cdot\frac{\partial a_1^3}{\partial w_{1,1}^3} =$
  $(\hat{y}_1(\Theta,\mathbf{x}) - y_1)\cdot f'(a_1^3)\cdot\frac{\partial \sum_j w_{j,1}^3 a_j^2}{\partial w_{1,1}^3} =$

$\hat{y}_i = f(a_i^3)$
$a_j^k = \sum_i w_{i,j}^k f(a_i^{k-1})$
$a_j^1 = \sum_i w_{i,j}^1 x_i$

Wyprowadzenie pochodnej

$$MSE(\Theta) = \frac{1}{N}\sum_i \frac{1}{2}\left((\hat{y}_1(\Theta,\mathbf{x}^i) - y_1^i)^2 + (\hat{y}_2(\Theta,\mathbf{x}^i) - y_2^i)^2 + (\hat{y}_3(\Theta,\mathbf{x}^i) - y_3^i)^2\right)$$

Pochodna po wadze w ostatniej warstwie $w^3_{1,1}$

- $\frac{1}{2}\left((\hat{y}_1(\Theta,\mathbf{x}) - y_1)^2 + (\hat{y}_2(\Theta,\mathbf{x}) - y_2)^2 + (\hat{y}_3(\Theta,\mathbf{x}) - y_3)^2\right)$
- $\frac{1}{2}\frac{\partial(\hat{y}_1(\Theta,\mathbf{x}) - y_1)^2}{\partial w^3_{1,1}} = \frac{1}{2}2(\hat{y}_1(\Theta,\mathbf{x}) - y_1)\cdot\frac{\partial \hat{y}_1(\Theta,\mathbf{x}^i)}{\partial w^3_{1,1}} =$
  $(\hat{y}_1(\Theta,\mathbf{x}) - y_1)\cdot\frac{\partial f(a^3_1)}{\partial w^3_{1,1}} = (\hat{y}_1(\Theta,\mathbf{x}) - y_1)\cdot f'(a^3_1)\cdot\frac{\partial a^3_1}{\partial w^3_{1,1}} =$
  $(\hat{y}_1(\Theta,\mathbf{x}) - y_1)\cdot f'(a^3_1)\cdot\frac{\partial \sum_j w^3_{j,1}a^2_j}{\partial w^3_{1,1}} =$
  $(\hat{y}_1(\Theta,\mathbf{x}) - y_1)\cdot f'(a^3_1)\cdot\frac{\partial(w^3_{1,1}f(a^2_1) + w^3_{2,1}f(a^2_2) + w^3_{3,1}f(a^2_3))}{\partial w^3_{1,1}} =$

$\hat{y}_i = f(a^3_i)$
$a^k_j = \sum_i w^k_{i,j} f(a^{k-1}_i)$
$a^1_j = \sum_i w^1_{i,j} x_i$

Wyprowadzenie pochodnej

$$MSE(\Theta) = \frac{1}{N}\sum_i \frac{1}{2}\left((\hat{y}_1(\Theta,\mathbf{x}^i) - y_1^i)^2 + (\hat{y}_2(\Theta,\mathbf{x}^i) - y_2^i)^2 + (\hat{y}_3(\Theta,\mathbf{x}^i) - y_3^i)^2\right)$$

Pochodna po wadze w ostatniej warstwie $w^3_{1,1}$

- $\frac{1}{2}\left((\hat{y}_1(\Theta,\mathbf{x}) - y_1)^2 + (\hat{y}_2(\Theta,\mathbf{x}) - y_2)^2 + (\hat{y}_3(\Theta,\mathbf{x}) - y_3)^2\right)$
- $\frac{1}{2}\frac{\partial(\hat{y}_1(\Theta,\mathbf{x}) - y_1)^2}{\partial w^3_{1,1}} = \frac{1}{2}2(\hat{y}_1(\Theta,\mathbf{x}) - y_1)\cdot\frac{\partial \hat{y}_1(\Theta,\mathbf{x}^i)}{\partial w^3_{1,1}} =$

  $(\hat{y}_1(\Theta,\mathbf{x}) - y_1)\cdot\frac{\partial f(a^3_1)}{\partial w^3_{1,1}} = (\hat{y}_1(\Theta,\mathbf{x}) - y_1)\cdot f'(a^3_1)\cdot\frac{\partial a^3_1}{\partial w^3_{1,1}} =$

  $(\hat{y}_1(\Theta,\mathbf{x}) - y_1)\cdot f'(a^3_1)\cdot\frac{\partial\sum_j w^3_{j,1}a^2_j}{\partial w^3_{1,1}} =$

  $(\hat{y}_1(\Theta,\mathbf{x}) - y_1)\cdot f'(a^3_1)\cdot\frac{\partial(w^3_{1,1}f(a^2_1) + w^3_{2,1}f(a^2_2) + w^3_{3,1}f(a^2_3))}{\partial w^3_{1,1}} =$

  $(\hat{y}_1(\Theta,\mathbf{x}) - y_1)\cdot f'(a^3_1)\cdot f(a^2_1) =$

$\hat{y}_i = f(a^3_i)$
$a^k_j = \sum_i w^k_{i,j} f(a^{k-1}_i)$
$a^1_j = \sum_i w^1_{i,j} x_i$

Wyprowadzenie pochodnej

$$MSE(\Theta) = \frac{1}{N}\sum_i \frac{1}{2}\left((\hat{y}_1(\Theta,\mathbf{x}^i) - y_1^i)^2 + (\hat{y}_2(\Theta,\mathbf{x}^i) - y_2^i)^2 + (\hat{y}_3(\Theta,\mathbf{x}^i) - y_3^i)^2\right)$$

Pochodna po wadze w ostatniej warstwie $w_{1,1}^3$

- $\frac{1}{2}\left((\hat{y}_1(\Theta,\mathbf{x}) - y_1)^2 + (\hat{y}_2(\Theta,\mathbf{x}) - y_2)^2 + (\hat{y}_3(\Theta,\mathbf{x}) - y_3)^2\right)$
- $\frac{1}{2}\frac{\partial(\hat{y}_1(\Theta,\mathbf{x})-y_1)^2}{\partial w_{1,1}^3} = \frac{1}{2}2(\hat{y}_1(\Theta,\mathbf{x}) - y_1)\cdot\frac{\partial \hat{y}_1(\Theta,\mathbf{x}^i)}{\partial w_{1,1}^3} =$
  $(\hat{y}_1(\Theta,\mathbf{x}) - y_1)\cdot\frac{\partial f(a_1^3)}{\partial w_{1,1}^3} = (\hat{y}_1(\Theta,\mathbf{x}) - y_1)\cdot f'(a_1^3)\cdot\frac{\partial a_1^3}{\partial w_{1,1}^3} =$
  $(\hat{y}_1(\Theta,\mathbf{x}) - y_1)\cdot f'(a_1^3)\cdot\frac{\partial\sum_j w_{j,1}^3 a_j^2}{\partial w_{1,1}^3} =$
  $(\hat{y}_1(\Theta,\mathbf{x}) - y_1)\cdot f'(a_1^3)\cdot\frac{\partial(w_{1,1}^3 f(a_1^2) + w_{2,1}^3 f(a_2^2) + w_{3,1}^3 f(a_3^2))}{\partial w_{1,1}^3} =$
  $(\hat{y}_1(\Theta,\mathbf{x}) - y_1)\cdot f'(a_1^3)\cdot f(a_1^2) = e_1^3\cdot f(a_2^1)$

$\hat{y}_i = f(a_i^3)$
$a_j^k = \sum_i w_{i,j}^k f(a_i^{k-1})$
$a_j^1 = \sum_i w_{i,j}^1 x_i$

Wyprowadzenie pochodnej II

$$MSE(\Theta) = \frac{1}{N}\sum_i \frac{1}{2}(\hat{y}_1(\Theta,\mathbf{x}^i) - y_1^i)^2 + (\hat{y}_2(\Theta,\mathbf{x}^i) - y_2^i)^2 + (\hat{y}_3(\Theta,\mathbf{x}^i) - y_3^i)^2$$

Pochodna po wadze w przedostatniej warstwie $w^2_{1,1}$

- $\frac{1}{2}\frac{\partial\left((\hat{y}_1(\Theta,\mathbf{x})-y_1)^2+(\hat{y}_2(\Theta,\mathbf{x})-y_2)^2+(\hat{y}_3(\Theta,\mathbf{x})-y_3)^2\right)}{\partial w^2_{1,1}} = \dots$
- $\frac{\partial \hat{y}_i(\Theta,\mathbf{x})}{\partial w^2_{1,1}} = \frac{\partial f\left(\sum_j w^3_{j,i}a^2_j\right)}{\partial w^2_{1,1}} = f'(\sum_j w^3_{j,i}a^2_j)\frac{\partial \sum_j w^3_{j,i}a^2_j}{\partial w^2_{1,1}} =$
  $f'(a^3_i)\frac{\partial\left(w^3_{1,i}f(a^2_1)+w^3_{2,i}f(a^2_2)+\dots\right)}{\partial w^2_{1,1}} = f'(a^3_i)w^3_{1,i}\frac{\partial f(a^2_1)}{\partial w^2_{1,1}} =$
  $f'(a^3_i)w^3_1 f'(a^2_1)\frac{\partial \sum_j w^2_{j,1}f(a^1_j)}{\partial w^2_{1,1}} = f'(a^3_i)w^3_{1,i}f'(a^2_1)\cdot f(a^1_1)$

$$\hat{y}_i = f(a^3_i)$$
$$a^k_j = \sum_i w^k_{i,j}f(a^{k-1}_i)$$
$$a^1_j = \sum_i w^1_{i,j}x_i$$
$$e^3_i = (\hat{y}_i(\Theta,\mathbf{x}) - y^3_i)f'(a^3_i)$$

Wyprowadzenie pochodnej II

$$MSE(\Theta) = \frac{1}{N}\sum_i \frac{1}{2}(\hat{y}_1(\Theta, \mathbf{x}^i) - y_1^i)^2 + (\hat{y}_2(\Theta, \mathbf{x}^i) - y_2^i)^2 + (\hat{y}_3(\Theta, \mathbf{x}^i) - y_3^i)^2$$

Pochodna po wadze w przedostatniej warstwie $w_{1,1}^2$

- $\frac{1}{2}\frac{\partial\left((\hat{y}_1(\Theta,\mathbf{x})-y_1)^2+(\hat{y}_2(\Theta,\mathbf{x})-y_2)^2+(\hat{y}_3(\Theta,\mathbf{x})-y_3)^2\right)}{\partial w_{1,1}^2} = \dots$

$$\hat{y}_i = f(a_i^3)$$
$$a_j^k = \sum_i w_{i,j}^k f(a_i^{k-1})$$
$$a_j^1 = \sum_i w_{i,j}^1 x_i$$
$$e_i^3 = (\hat{y}_i(\Theta, \mathbf{x}) - y_i^3) f'(a_i^3)$$
$$\frac{\partial \hat{y}_i(\Theta,\mathbf{x})}{\partial w_{1,1}^2} =$$
$$f'(a_i^3) w_{1,i}^3 f'(a_1^2) \cdot f(a_1^1)$$

Wyprowadzenie pochodnej II

$$MSE(\Theta) = \frac{1}{N}\sum_i \frac{1}{2}(\hat{y}_1(\Theta, \mathbf{x}^i) - y_1^i)^2 + (\hat{y}_2(\Theta, \mathbf{x}^i) - y_2^i)^2 + (\hat{y}_3(\Theta, \mathbf{x}^i) - y_3^i)^2$$

Pochodna po wadze w przedostatniej warstwie $w_{1,1}^2$

- $\frac{1}{2}\frac{\partial \sum_j (\hat{y}_j(\Theta,\mathbf{x}) - y_j)^2}{\partial w_{1,1}^2} = \frac{1}{2} \cdot 2\sum_j (\hat{y}_j(\Theta, \mathbf{x}) - y_j)\frac{\partial \hat{y}_j(\Theta,\mathbf{x})}{\partial w_{1,1}^2} =$
  $\sum_j (\hat{y}_j(\Theta, \mathbf{x}) - y_j) f'(a_j^3) w_{1,j}^3 f'(a_1^2) f(a_1^1) =$
  $\left(\sum_j e_j^3 w_{1,j}^3\right) f'(a_1^2) f(a_1^1)$

$\hat{y}_i = f(a_i^3)$
$a_j^k = \sum_i w_{i,j}^k f(a_i^{k-1})$
$a_j^1 = \sum_i w_{i,j}^1 x_i$
$e_i^3 = (\hat{y}_i(\Theta, \mathbf{x}) - y_i^3) f'(a_i^3)$
$\frac{\partial \hat{y}_i(\Theta,\mathbf{x})}{\partial w_{1,1}^2} =$
$f'(a_i^3) w_{1,i}^3 f'(a_1^2) \cdot f(a_1^1)$

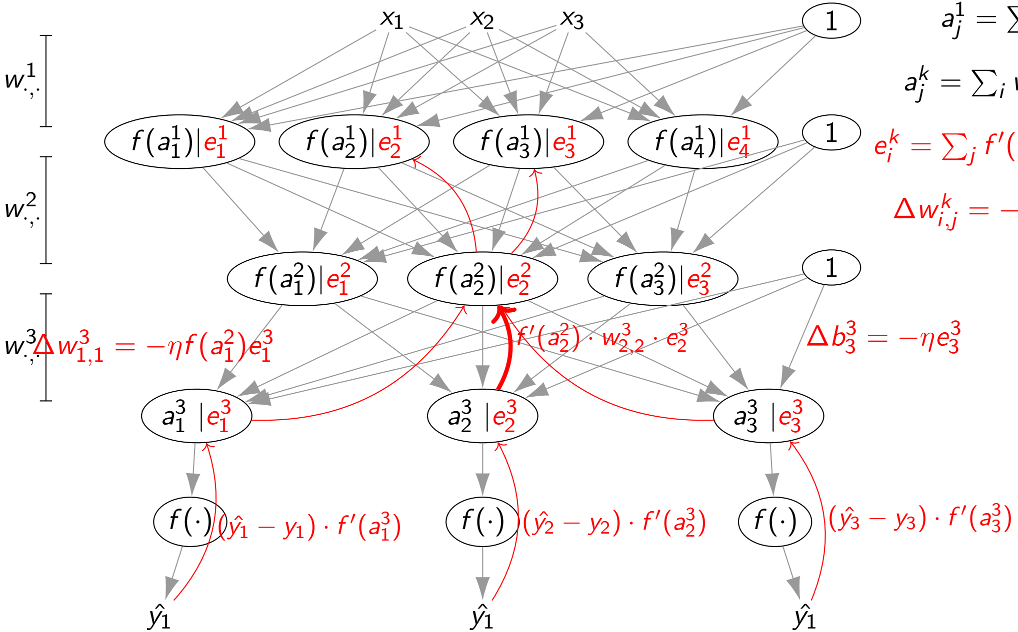
Wyprowadzenie pochodnej II

$$MSE(\Theta) = \frac{1}{N}\sum_i \frac{1}{2}(\hat{y}_1(\Theta, \mathbf{x}^i) - y_1^i)^2 + (\hat{y}_2(\Theta, \mathbf{x}^i) - y_2^i)^2 + (\hat{y}_3(\Theta, \mathbf{x}^i) - y_3^i)^2$$

Pochodna po wadze w przedostatniej warstwie $w^2_{1,1}$

- $\frac{1}{2}\frac{\partial \sum_j (\hat{y}_j(\Theta,\mathbf{x})-y_j)^2}{\partial w^2_{1,1}} = \frac{1}{2}\cdot 2\sum_j(\hat{y}_j(\Theta,\mathbf{x}) - y_j)\frac{\partial \hat{y}_j(\Theta,\mathbf{x})}{\partial w^2_{1,1}} =$
  $\sum_j(\hat{y}_j(\Theta,\mathbf{x}) - y_j)f'(a^3_j)w^3_{1,j}f'(a^2_1)f(a^1_1) =$
  $\left(\sum_j e^3_j w^3_{1,j}\right) f'(a^2_1)f(a^1_1)$
- $e^2_1 = \sum_j e^3_j w^3_{1,j} f'(a^2_1)$

$\hat{y}_i = f(a^3_i)$
$a^k_j = \sum_i w^k_{i,j} f(a^{k-1}_i)$
$a^1_j = \sum_i w^1_{i,j} x_i$
$e^3_i = (\hat{y}_i(\Theta,\mathbf{x}) - y^3_i)f'(a^3_i)$
$\frac{\partial \hat{y}_i(\Theta,\mathbf{x})}{\partial w^2_{1,1}} =$
$f'(a^3_i)w^3_{1,i}f'(a^2_1)\cdot f(a^1_1)$



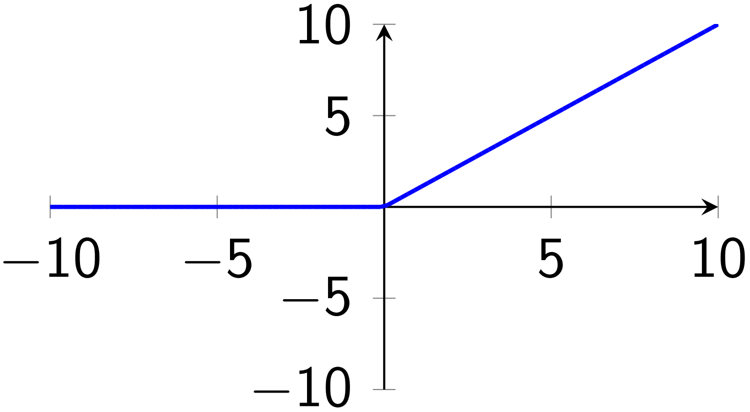
$$a^1_j = \sum_i w^1_{i,j} x_i$$

$$a^k_j = \sum_i w^k_{i,j} f(a^{k-1}_i)$$

$$e^k_i = \sum_j f'(a^k_i) w^{k+1}_{i,j} e^{k+1}_j$$

$$\Delta w^k_{i,j} = -\eta f(a^{k-1}_i) e^k_j$$

$$\mathrm{ReLU}(x) = \max\{x, 0\}$$



Krok uczenia η

$\Delta\Theta = -\eta\nabla MSE(\Theta)$

- Zwykle 10^{-3} lub mniej

Miary błędu

- Mean Absolute Error $\frac{1}{N}\sum_i \left|\hat{y}^i - y^i\right|$
- Mean Squared Error $\frac{1}{N}\sum_i \left(\hat{y}^i - y^i\right)^2$
- Categorical crossentropy $\frac{1}{N}\sum_i -\sum_j y^i_j \log \hat{y}^i_j$

Stochastic Gradient Descent

$$MSE(\Theta) = \frac{1}{N}\sum_{i}\frac{1}{2}(\hat{y}_1(\Theta,\mathbf{x}^i) - y_1^i)^2 + (\hat{y}_2(\Theta,\mathbf{x}^i) - y_2^i)^2 + (\hat{y}_3(\Theta,\mathbf{x}^i) - y_3^i)^2$$

SGD

$\Theta \leftarrow$ *InicjujLosowo*;
while $\neg$*StopCondition* **do**
  for $(\mathbf{X},\mathbf{Y}) \in$ *Zbiór uczący* **do**
    $\hat{Y} \leftarrow Network(\Theta, X)$;
    $\Delta\Theta \leftarrow -\eta\nabla Network(\Theta, X)$;
    $\Theta \leftarrow \Theta + \Delta\Theta$;

Podstawowy Gradient Descent

$\Theta \leftarrow$ *InicjujLosowo*;
while $\neg$*StopCondition* **do**
  $\Delta\Theta \leftarrow 0$;
  for $(\mathbf{X},\mathbf{Y}) \in$ *Zbiór uczący* **do**
    $\hat{Y} \leftarrow Network(\Theta, X)$;
    $\Delta\Theta \leftarrow \Delta\Theta - \eta\nabla Network(\Theta, X)$;
  $\Theta \leftarrow \Theta + \Delta\Theta$;

Trajektoria i szybkość zbieżności

- Podstawowy wariant bywa nazywany Batch Gradient Descent

$\Theta \leftarrow \mathit{InicjujLosowo}$;
while $\neg \mathit{StopCondition}$ **do**
  $\mathit{MiniBatches} \leftarrow \mathit{PodzielLosowoNaRozczeZbiory}(\mathit{ZbiorUczcy}, k)$;
  for $\mathit{batch} \in \mathit{MiniBatches}$ **do**
    $\Delta\Theta \leftarrow 0$;
    for $(\mathbf{X}, \mathbf{Y}) \in \mathit{Batch}$ **do**
      $\hat{Y} \leftarrow \mathit{Network}(\Theta, X)$;
      $\Delta\Theta \leftarrow \Delta\Theta - \eta \nabla \mathit{Network}(\Theta, X)$;
    $\Theta \leftarrow \Theta + \Delta\Theta$;

Uwaga o implementacji GD

W ostatniej warstwie (MSE)

$$\mathbf{e}^m = f'(\mathbf{a}^m) \odot (\hat{\mathbf{y}} - \mathbf{y})$$

Pozostałe warstwy

$$\mathbf{e}^{k-1} = f'(\mathbf{a}^{k-1}) \odot \left(\mathbf{e}^k \left(W^k\right)^T\right)$$

Pochodna wagi

$$\frac{\partial MSE}{\partial w_{i,j}^k} = e_j^k f\left(a_i^{k-1}\right)$$

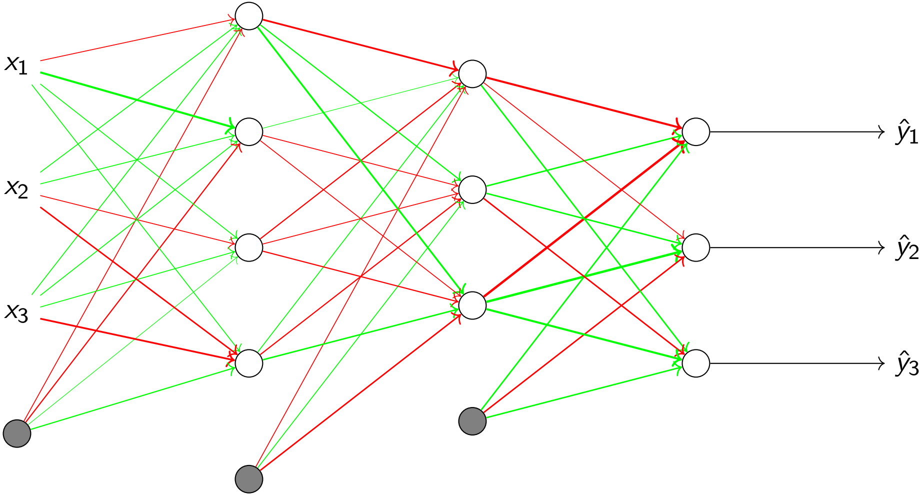
Początkowe wagi

- Duży wpływ na zbieżność
- Małe wartości dookoła zera
- Najprostsze warianty
  - Rozkład gaussowski
  - Rozkład jednostajny
- Związek z funkcją aktywacji
- **Xavier**

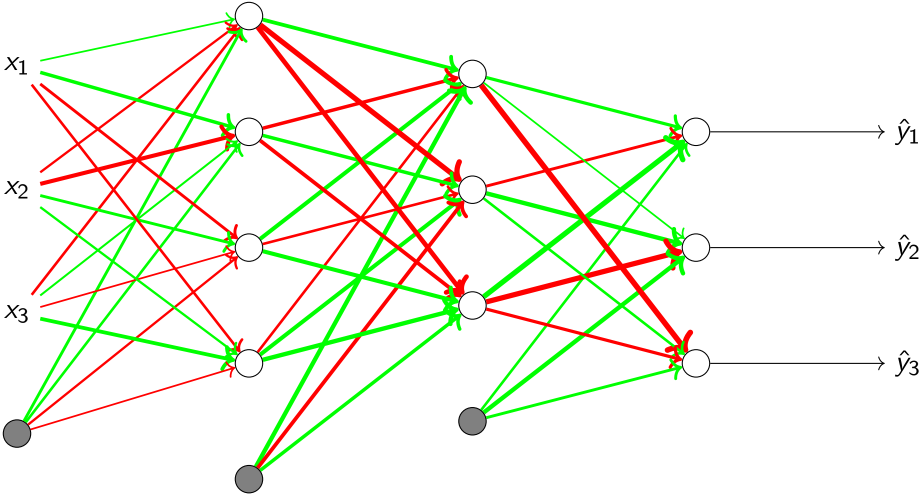
$$w_{i,j}^{k} \in \left[-\frac{\sqrt{6}}{\sqrt{IN + OUT}}, \frac{\sqrt{6}}{\sqrt{IN + OUT}} \right]$$

  Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*. Ed. by Yee Whye Teh and D. Mike Titterington. Vol. 9. JMLR Proceedings. JMLR.org, 2010, pp. 249–256
- **He** – inny wariant normalizacji wewnątrz warstwy Kaiming He et al. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". In: *2015 IEEE International Conference on Computer Vision (ICCV)* (2015). DOI: 10.1109/iccv.2015.123

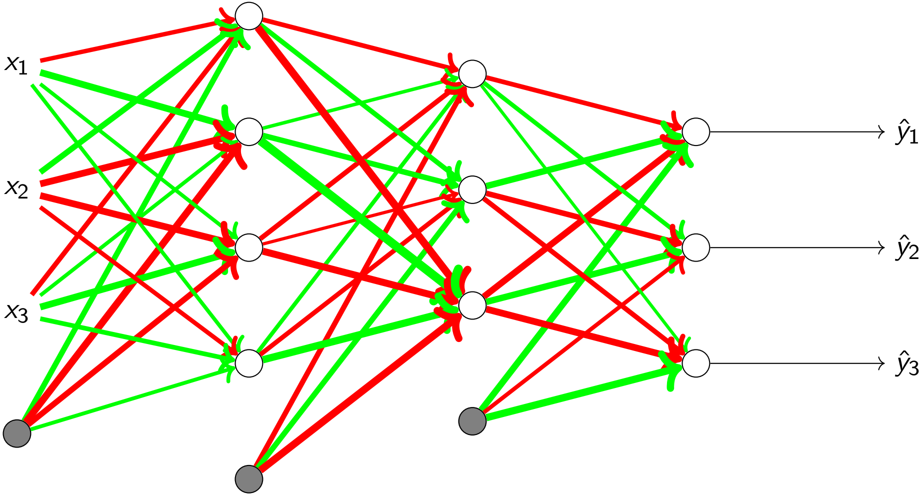
Eksplozja wag



Eksplozja wag



Eksplozja wag



Uczenie z momentem

- η – krok uczenia
- $\lambda \in (0, 1)$ – współczynnik wygaszania momentu

$\Theta \leftarrow \mathit{InicjujLosowo}$;
$\mathit{Momentum} \leftarrow [0, 0, \ldots 0]$;
while $\neg\mathit{StopCondition}$ **do**
  $\Delta\Theta \leftarrow [0, 0, \ldots 0]$;
  for $(\mathbf{X}, \mathbf{Y}) \in$ *Zbiór uczący* **do**
    $\hat{Y} \leftarrow \mathit{Network}(\Theta, X)$;
    $\Delta\Theta \leftarrow \Delta\Theta - \nabla \mathit{Network}(\Theta, X)$;
  $\mathit{Momentum} \leftarrow \Delta\Theta + \mathit{Momentum} \cdot \lambda$;
  $\Theta \leftarrow \Theta + \eta \mathit{Momentum}$;

Normalizacja gradientu

RMSProp

- η – krok uczenia
- β – współczynnik wygaszania

$\Theta \leftarrow \mathit{InicjujLosowo}$;
$\mathbb{E}\left[g^2\right] \leftarrow [0, 0, \ldots, 0]$;
while $\neg \mathit{StopCondition}$ **do**
  $g \leftarrow 0$;
  for $(\mathbf{X}, \mathbf{Y}) \in$ *Zbiór uczący* **do**
    $\hat{Y} \leftarrow \mathit{Network}(\Theta, X)$;
    $g \leftarrow g + \nabla \mathit{Network}(\Theta, X)$;
  $\mathbb{E}\left[g^2\right] \leftarrow \beta\mathbb{E}\left[g^2\right] + (1-\beta)g^2$`// Kwadrat po współrzędnych`
  $\Theta \leftarrow \Theta - \eta \left[\frac{g_i}{\sqrt{\mathbb{E}[g^2]_i}}\right]_{\forall i}$;

Także inne warianty, np. Adagrad, Adadelta.

Adam

Adaptive moment estimation

Moment+Normalizacja gradientu

- η – Krok uczenia
- $\beta_1, \beta_2 \in [0, 1]$ – Współczynniki wygaszania

Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015*. Ed. by Yoshua Bengio and Yann LeCun. 2015

$\Theta \leftarrow \mathit{InicjujLosowo}$, $m \leftarrow [0, 0, \ldots, 0]$, $v \leftarrow [0, 0, \ldots, 0]$;
$t \leftarrow 0$;
while $\neg \mathit{StopCondition}$ **do**
    $t \leftarrow t + 1$;
    $g \leftarrow [0, 0, \ldots, 0]$;
    for $(\mathbf{X}, \mathbf{Y}) \in \mathit{Zbiór\ uczący}$ **do**
        $\hat{Y} \leftarrow \mathit{Network}(\Theta, X)$;
        $g \leftarrow g + \nabla \mathit{Network}(\Theta, X)$;
    $m \leftarrow \beta_1 m + (1 - \beta_1) g$;
    $v \leftarrow \beta_2 v + (1 - \beta_2) g^2$ `// Kwadrat po współrzędnych`
    $\hat{m} \leftarrow m / (1 - (\beta_1)^t)$;
    $\hat{v} \leftarrow v / (1 - (\beta_2)^t)$;
    $\Theta \leftarrow \Theta - \eta \frac{\hat{m}}{\sqrt{\hat{v}}}$ `// Dzielenie, pierwiastek po współrzędnych`

Adam

Adaptive moment estimation

Moment+Normalizacja gradientu

- η – Krok uczenia
- $\beta_1, \beta_2 \in [0, 1]$ – Współczynniki wygaszania

Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015*. Ed. by Yoshua Bengio and Yann LeCun. 2015

$\Theta \leftarrow \mathit{InicjujLosowo}$, $m \leftarrow [0, 0, \ldots, 0]$, $v \leftarrow [0, 0, \ldots, 0]$;
$t \leftarrow 0$;
while $\neg \mathit{StopCondition}$ **do**
    $t \leftarrow t + 1$;
    $g \leftarrow [0, 0, \ldots, 0]$;
    for $(\mathbf{X}, \mathbf{Y}) \in$ *Zbiór uczący* **do**
        $\hat{Y} \leftarrow \mathit{Network}(\Theta, X)$;
        $g \leftarrow g + \nabla \mathit{Network}(\Theta, X)$;
    $m \leftarrow \beta_1 m + (1 - \beta_1) g$;
    $v \leftarrow \beta_2 v + (1 - \beta_2) g^2$ `// Kwadrat po współrzędnych`
    $\hat{m} \leftarrow m / (1 - (\beta_1)^t)$;
    $\hat{v} \leftarrow v / (1 - (\beta_2)^t)$;
    $\Theta \leftarrow \Theta - \eta \frac{\hat{m}}{\sqrt{\hat{v}} + \varepsilon}$ `// Dzielenie, pierwiastek po współrzędnych`

Moment+Normalizacja gradientu

- η – Krok uczenia
- $\beta_1, \beta_2 \in [0, 1]$ – Współczynniki wygaszania

Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015*. Ed. by Yoshua Bengio and Yann LeCun. 2015

$\hat{m}$, $\hat{v}$ – estymatory nieobciążone

Indeks dolny – wartość g w zadanej iteracji.

$$m = \sum_{n=0}^{t} \beta g_{t-n} (1-\beta)^n$$

Przybliżenie nieskończonej sumy (mamy tylko t iteracji, więc nie możemy jej policzyć):

$$m = \sum_{n=0}^{\infty} \beta g_{t-n} (1-\beta)^n$$

m wyliczone w algorytmie jest estymatorem obciążonym, bo „ogon" sumy jest wyzerowany. Estymator nieobciążony:

$$\hat{m} = \left(\sum_{n=0}^{t} \beta g_{t-n} (1-\beta)^n \right) / \left(1 - \beta^t\right)$$

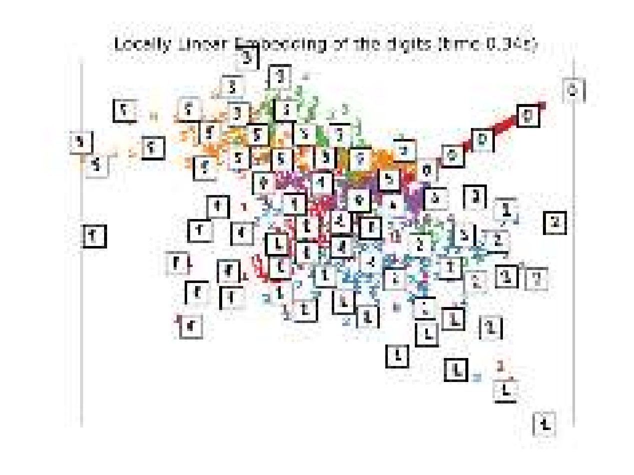
MLP, a wypukłość funkcji błędu

Problem analizy skupień I

- Zbiór punktów w przestrzeni $\mathbb{R}^n$
- Podział punktów na k rozłącznych podzbiorów, tak żeby punkty w każdym zbiorze były możliwie „podobne".

Problem analizy skupień II

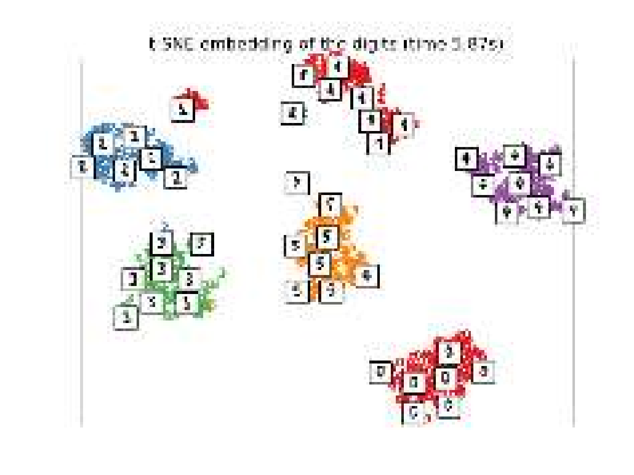
Locally linear embedding (LLE)



Źródło: `https://scikit-learn.org/stable/modules/manifold`

Problem analizy skupień III

t-distributed Stochastic Neighbor Embedding (t-SNE)



Źródło: `https://scikit-learn.org/stable/modules/manifold`

Problem analizy skupień IV

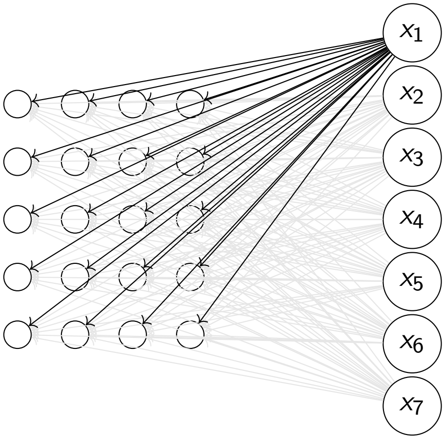
K-Means

Teuvo Kohonen. “Automatic formation of topological maps of patterns in a self-organizing system”. In: *Proceedings of the 2nd scandinavian Conference on Image Analysis*. 1981, pp. 214–220

Teuvo Kohonen. “Self-organized formation of topologically correct feature maps”. In: *Biological Cybernetics* 43.1 (1982), 59–69. ISSN: 1432-0770. DOI: 10.1007/bf00337288. URL: http://dx.doi.org/10.1007/bf00337288

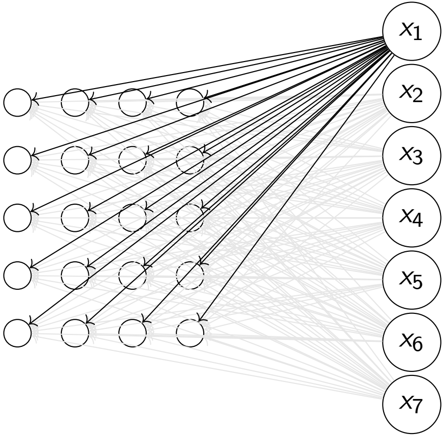
Teuvo Kohonen. “Self-Organization and Associative Memory”. In: *Springer Series in Information Sciences* (1988). ISSN: 0720-678X. DOI: 10.1007/978-3-662-00784-6. URL: http://dx.doi.org/10.1007/978-3-662-00784-6

Architektura sieci Kohonena



- Niech podobne wektory wejścia pobudzają sąsiednie neruony
- Niech neurony dla różniących się wejść będą odseparowane
- Wagi definiują pozycję neuronu w $\mathbb{R}^n$

Uczenie sieci Kohonena



Szkic algorytmu

1. Weź wektor ze zbioru danych
2. Wybierz neuron, który jest najbliżej tego wzorca
3. Przesuń najbliższy neuron w kierunku wzorca
4. Przesuń, z pewnym osłabieniem, sąsiadów neuronu w kierunku wzorca
   - Sąsiedzi w przestrzeni siatki 2D

Sieć Kohonena – uczenie

$\forall w_{i,j}^k = RAND$ `// Losowa inicjacja`
for $t \leftarrow 1 \ldots \lambda$ **do**
  for $\mathbf{x} \in \mathit{RandomShuffle}(\mathit{data})$ **do**
    $(i^*, j^*) \leftarrow \arg\min_{i,j} d(\mathbf{w}_{i,j}, \mathbf{x})$;
    for $(i, j) \in N \times M$ **do**
      $\mathbf{w_{i,j}} \leftarrow \mathbf{w_{i,j}} + \theta((i^*, j^*), (i, j), t)\alpha(t)(\mathbf{x} - \mathbf{w_{i,j}})$;

- $w_{i,j}^k$ – waga między wejściem k, a neuronem na pozycji (i, j)
- λ – liczba iteracji uczenia
- d – metryka w przestrzeni danych
- N, M – wymiary siatki
- $\alpha(t)$ – wygaszanie w czasie
- $\theta(n_1, n_2, t)$ – waga sąsiedztwa wygaszana w czasie

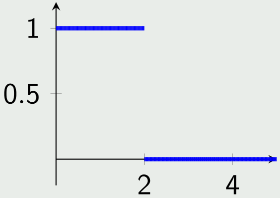
Sąsiedztwo

Względem odległości euklidesowej między punktami

$$\theta(n_1, n_2, t) = f(d(n_1, n_2), t)$$

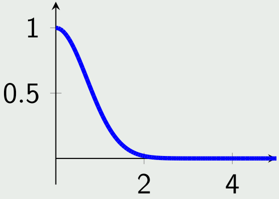
Koło

$$f(x, r) = \begin{cases} 1 & x \leqslant r \\ 0 & x > r \end{cases}$$



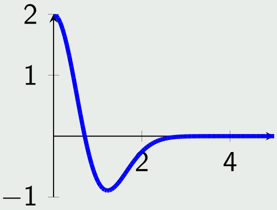
Funkcja Gaussa

$$f(x, t) = e^{-x^2}$$



„Meksykański kapelusz”

$$f(x, r) = (2 - 4x^2)e^{-x^2}$$

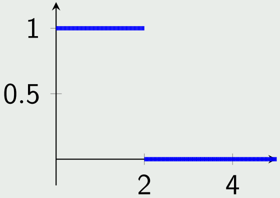


Uwaga: skalowanie szerokości.

$$\theta(n_1, n_2, t) = f(d(n_1, n_2), t)$$

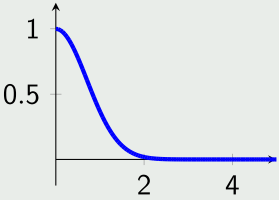
Koło

$$f(x, r) = \begin{cases} 1 & x \leqslant re^{-t/\lambda} \\ 0 & x > re^{-t/\lambda} \end{cases}$$



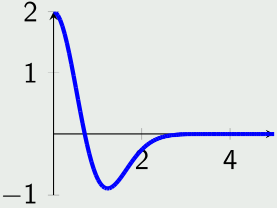
Funkcja Gaussa

$$f(x, t) = e^{-(xt)^2}$$



„Meksykański kapelusz"

$$f(x, r) = (2 - 4x^2)e^{-x^2}$$



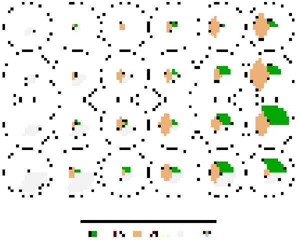
Uwaga: skalowanie szerokości.

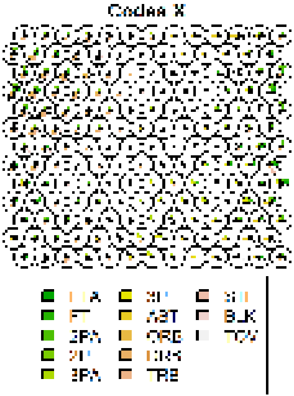
Wygaszanie w czasie – przykłady

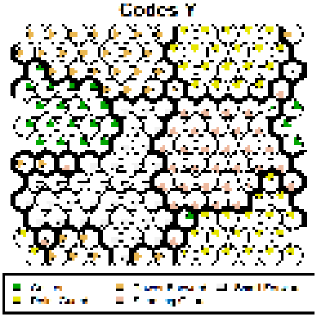
- $\alpha(t) = \eta e^{-t\lambda}$, gdzie η to parametr $\in \mathbb{R}$
  - $\alpha(t) = \eta^t$, $\eta \in (0, 1)$
- $\alpha(t) = \eta$ – brak redukcji w czasie

- Niekoniecznie siatka 2D
- **Siatka sześciokątna 2D**

https://clarkdatalabs.github.io/soms/SOM_NBA







Literatura I

- G. Cybenko. "Approximation by superpositions of a sigmoidal function". In: *Mathematics of Control, Signals, and Systems* 2.4 (1989), 303–314. ISSN: 1435-568X. DOI: 10.1007/bf02551274. URL: http://dx.doi.org/10.1007/bf02551274.
- Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*. Ed. by Yee Whye Teh and D. Mike Titterington. Vol. 9. JMLR Proceedings. JMLR.org, 2010, pp. 249–256.
- Kaiming He et al. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". In: *2015 IEEE International Conference on Computer Vision (ICCV)* (2015). DOI: 10.1109/iccv.2015.123.
- Donald O Hebb. *The organization of behavior*. 1949.
- Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015*. Ed. by Yoshua Bengio and Yann LeCun. 2015.

Literatura II

Teuvo Kohonen. "Automatic formation of topological maps of patterns in a self-organizing system". In: *Proceedings of the 2nd scandinavian Conference on Image Analysis*. 1981, pp. 214–220.

Teuvo Kohonen. "Self-Organization and Associative Memory". In: *Springer Series in Information Sciences* (1988). ISSN: 0720-678X. DOI: 10.1007/978-3-662-00784-6. URL: http://dx.doi.org/10.1007/978-3-662-00784-6.

Teuvo Kohonen. "Self-organized formation of topologically correct feature maps". In: *Biological Cybernetics* 43.1 (1982), 59–69. ISSN: 1432-0770. DOI: 10.1007/bf00337288. URL: http://dx.doi.org/10.1007/bf00337288.

Fundusze Europejskie
Wiedza Edukacja Rozwój

Rzeczpospolita Polska

Politechnika Warszawska

Unia Europejska
Europejski Fundusz Społeczny

Zadanie 10 pn. „Modyfikacja programów studiów na kierunkach prowadzonych przez Wydział Matematyki i Nauk Informacyjnych" realizowane jest w ramach projektu „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca" współfinansowanego ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego