Jednokierunkowe sieci wielowarstwowe

Multilayer Perceptron (MLP)

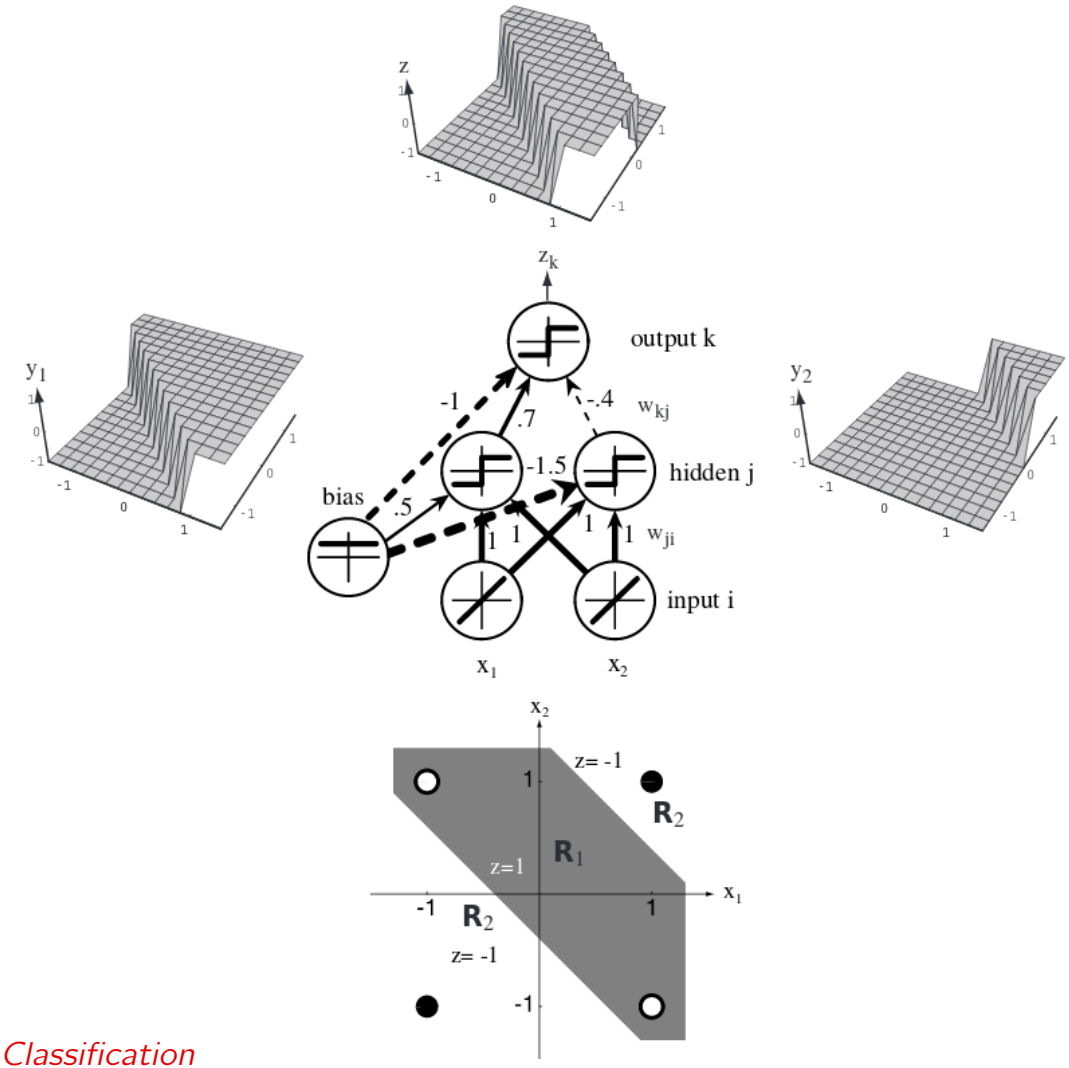
- Perceptrony jednowarstwowe potrafią rozwiązywać problemy separowalne liniowo
- Reguła delta dla sieci jednowarstwowej:

$$\Delta w_{ij} = \eta \left(y_j - f_j(\mathbf{x}) \right) f_j'(\mathbf{x}) x_i$$

  gdzie w_{ij} waga pomiędzy i-tym wejściem a j-tym wyjściem
- Problemy bardziej złożone (np. XOR) wymagają dodania kolejnej warstwy (nieliniowej) przetwarzania
- Każda kolejna warstwa tworzy nowy obraz danych, który może być dalej analizowany przez kolejne warstwy

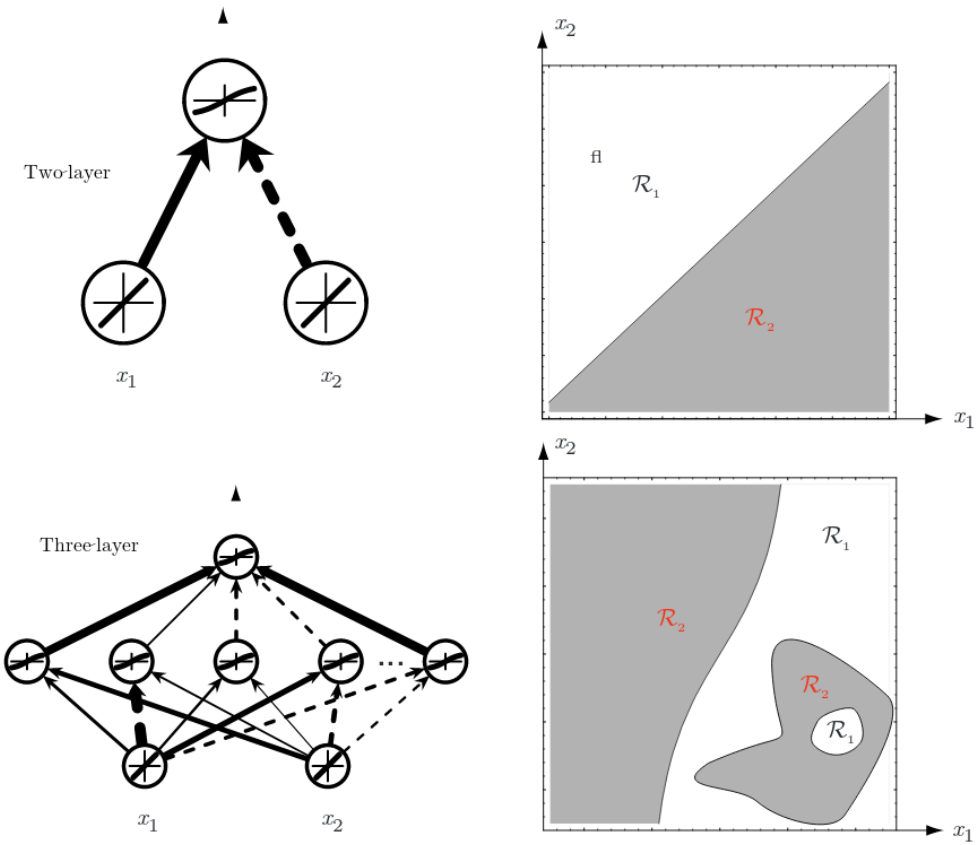
$$\mathbf{y} = f(\mathbf{x}) = f^{(M)} \left(f^{(M-1)} \left(\dots f^{(1)}(\mathbf{x}) \right) \right)$$

Jeszcze raz XOR



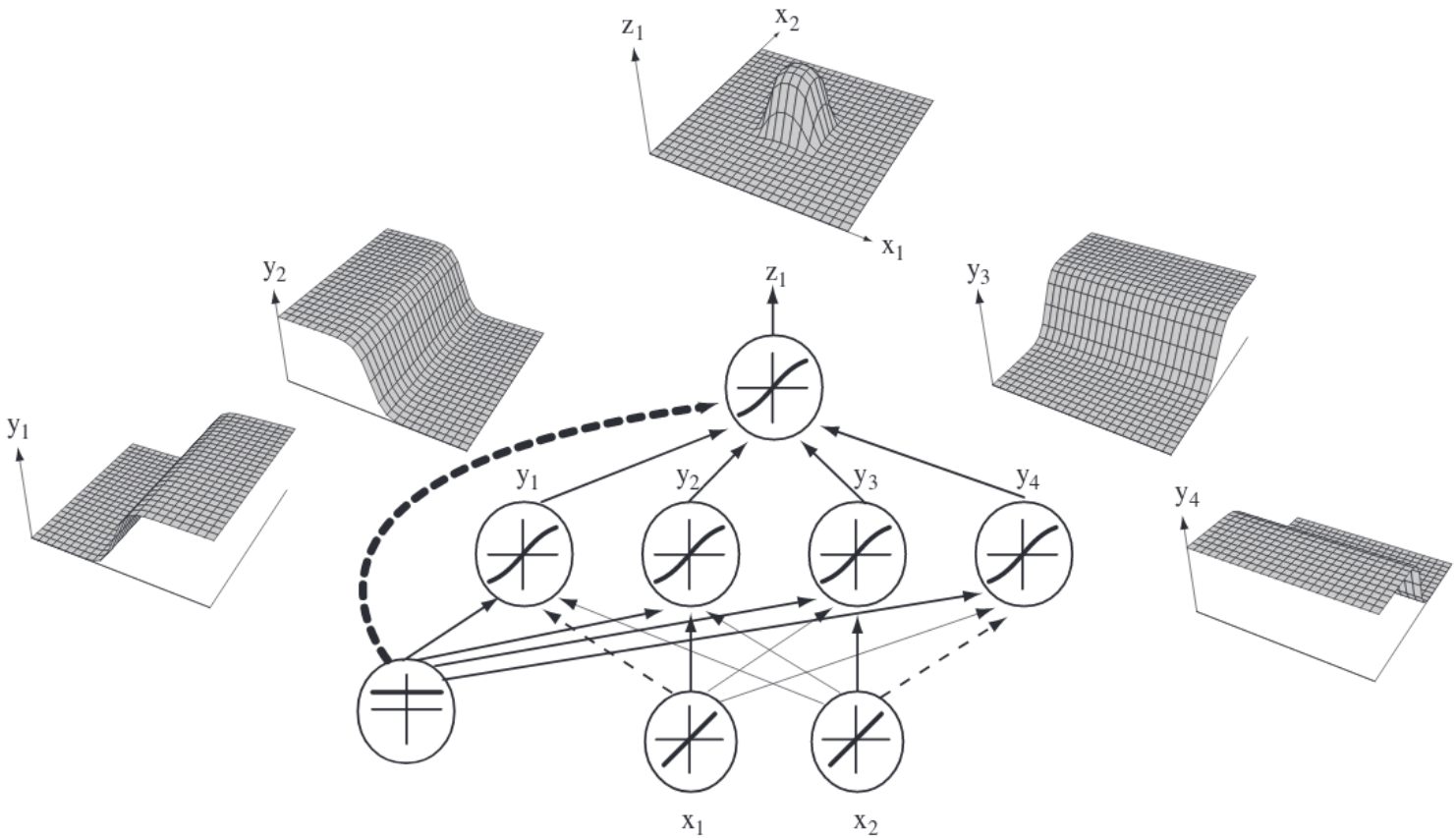
Rys: Duda, Hart, Pattern Classification

Warstwa ukryta a granice decyzji



Rys: Duda, Hart, Pattern Classification

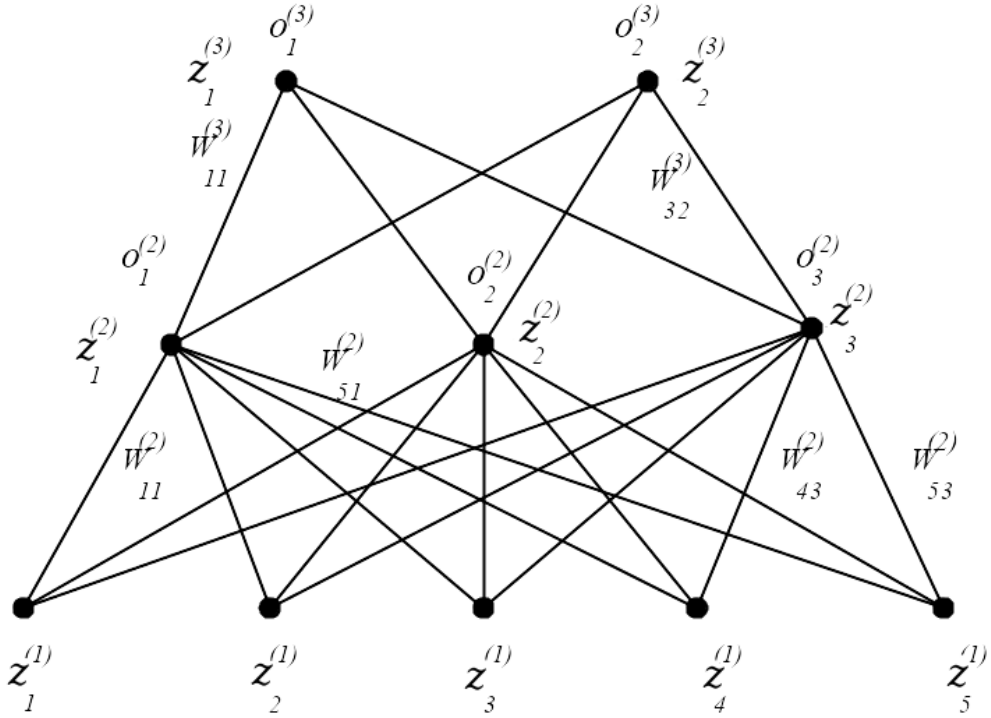
Przykład: sieć MLP 2-4-2



Rys: Duda, Hart, Pattern Classification

Perceprton 3 warstwowy

- Warstwy: wejściowa, ukryte, wyjściowa

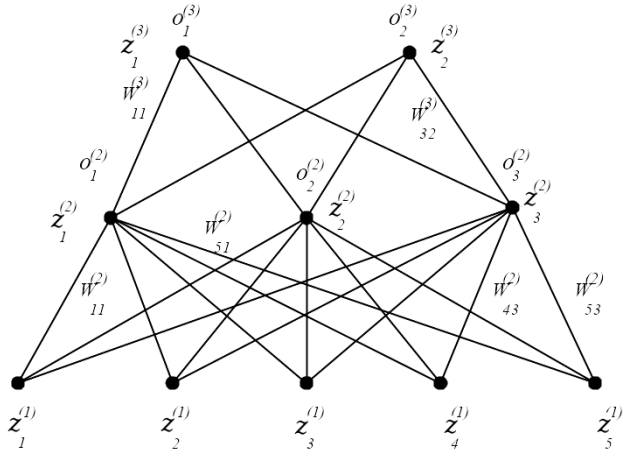


- Warstwy w pełni połączone (wagi tworzą macierz)

$$\mathbf{o}^{(K)} = \sigma\left(\mathbf{W}^{(K)}\mathbf{o}^{(K-1)}\right)$$

- Waga $w_{ij} = 0$ może być interpretowana jako brak połączenia

MLP - oznaczenia



M - liczba warstw ($M = 3$)

$w_{ij}^{(l)}$ - waga łącząca element i należący do warstwy $l-1$ oraz element j z warstwy l

$z_j^{(l)}$ - aktywacja neuronu j w warstwie l

$$z_j^{(l)} = \sum_i w_{ij}^{(l)} o_i^{(l-1)}$$

$o_j^{(l)}$ - sygnał wychodzący z elementu j należącego do warstwy l

$$o_j^{(l)} = \sigma(z_j^{(l)}) = \sigma\left(\sum_i w_{ij}^{(l)} o_i^{(l-1)}\right)$$

$f_i(\mathbf{x}) = o_i^{(M)}$ funkcja realizowana przez MLP (i-te wyjście)

Algorytm wstecznej propagacji błędu
(1974, 1986)

Miara błędu sieci dla pojedynczego wzorca wejściowego $\mathbf{x}$ i pożądanej odpowiedzi $\mathbf{y} = [y_1, \ldots, y_n]$

$$E(\mathbf{x}; \mathbf{W}) = \frac{1}{2} \sum_{i=1}^{n} (y_i - f_i(\mathbf{x}; \mathbf{W}))^2$$

gdzie n to liczba wyjść

Minimalizacja błędu metodą spadku gradientu:

$$\Delta w = -\eta \frac{\partial E(\mathbf{x}; \mathbf{W})}{\partial w} = \eta \sum_{i=1}^{n} (y_i - f_i(\mathbf{x}; \mathbf{W})) \frac{\partial f_i(\mathbf{x}; \mathbf{W})}{\partial w}$$

Zmiana wag dla warstwy wyjściowej

Dla wagi $w_{jk}^{(M)}$ łączącej wyjście $o_j^{(M-1)}$ z k-tym wyjściem sieci

$$\begin{aligned}
\Delta w_{jk}^{(M)} &= \eta \sum_{i=1}^{n} \left(y_i - f_i(\mathbf{x};\mathbf{W})\right) \frac{\partial f_i(\mathbf{x};\mathbf{W})}{\partial w_{jk}^{(M)}} \\
&= \eta \left(y_k - f_k(\mathbf{x};\mathbf{W})\right) \frac{\partial f_k(\mathbf{x};\mathbf{W})}{\partial w_{jk}^{(M)}} \\
&= \eta \left(y_k - f_k(\mathbf{x};\mathbf{W})\right) \frac{\partial \sigma(z_k^{(M)})}{\partial z_k^{(M)}} \frac{\partial z_k^{(M)}}{\partial w_{jk}^{(M)}} \\
&= \eta \left(y_k - f_k(\mathbf{x};\mathbf{W})\right) \sigma'(z_k^{(M)}) o_j^{(M-1)}
\end{aligned}$$

$$\Delta w_{jk}^{(M)} = \eta \left(y_k - f_k(\mathbf{x}; \mathbf{W}) \right) \sigma'(z_k^{(M)}) o_j^{(M-1)}$$

Niech $\delta_i^{(M)}$ oznacza błąd „lokalny” dla warstwy M

$$\delta_k^{(M)} = \sigma'(z_k^{(M)}) \left(y_k - f_k(\mathbf{x}; \mathbf{W}) \right)$$

Zmiana wag dla warstwy wyjściowej:

$$\Delta w_{jk}^{(M)} = \eta \delta_k^{(M)} o_j^{(M-1)}$$

$$\Delta w_{0k}^{(M)} = \eta \delta_k^{(M)}$$

Zmiana wag dla warstwy ukrytej $(M-1)$

Dla wagi $w_{jk}^{(M-1)}$ z warstwy ukrytej $M-1$:

$$\Delta w_{jk}^{(M-1)} = -\eta \frac{\partial E(\mathbf{W})}{\partial w_{jk}^{(M-1)}} = \eta \sum_{i=1}^{n} (y_i - f_i(\mathbf{x};\mathbf{W})) \frac{\partial f_i(\mathbf{x};\mathbf{W})}{\partial w_{jk}^{(M-1)}}$$

gradient

$$\begin{aligned}
\frac{\partial f_i(\mathbf{x};\mathbf{W})}{\partial w_{jk}^{(M-1)}} &= \frac{\partial \sigma(z_i^{(M)})}{\partial z_i^{(M)}} \frac{\partial z_i^{(M)}}{\partial w_{jk}^{(M-1)}} = \sigma'\left(z_i^{(M)}\right) \frac{\partial \left(\sum_l w_{li}^{(M)} o_l^{(M-1)} \right)}{\partial w_{jk}^{(M-1)}} \\
&= \sigma'\left(z_i^{(M)}\right) w_{ki}^{(M)} \frac{\partial o_k^{(M-1)}}{\partial w_{jk}^{(M-1)}} \\
&= \sigma'\left(z_i^{(M)}\right) w_{ki}^{(M)} \sigma'\left(z_k^{(M-1)}\right) o_j^{(M-2)}
\end{aligned}$$

$$\frac{\partial f_i(\mathbf{x};\mathbf{W})}{\partial w_{jk}^{(M-1)}} = \sigma'\left(z_i^{(M)}\right) w_{ki}^{(M)} \sigma'\left(z_k^{(M-1)}\right) o_j^{(M-2)}$$

zmiana wag dla warstwy $M-1$

$$\begin{aligned}
\Delta w_{jk}^{(M-1)} &= \eta \sum_{i=1}^{n} \left(y_i - f_i(\mathbf{x};\mathbf{W})\right) \frac{\partial f_i(\mathbf{x};\mathbf{W})}{\partial w_{jk}^{(M-1)}} \\
&= \eta \sum_{i=1}^{n} \left(y_i - f_i(\mathbf{x};\mathbf{W})\right) \sigma'\left(z_i^{(M)}\right) w_{ki}^{(M)} \sigma'\left(z_k^{(M-1)}\right) o_j^{(M-2)} \\
&= \eta \sigma'\left(z_k^{(M-1)}\right) \sum_{i=1}^{n} w_{ki}^{(M)} \delta_i^{(M)} o_j^{(M-2)} \\
&= \eta \delta_k^{(M-1)} o_j^{(M-2)}
\end{aligned}$$

$$\Delta w_{0k}^{(M-1)} = \eta \delta_k^{(M-1)} \qquad \text{gdzie} \qquad \delta_k^{(M-1)} = \sigma'\left(z_k^{(M-1)}\right) \sum_{i=1}^{n} w_{ki}^{(M)} \delta_i^{(M)}$$

Struktura wzoru w kolejnych warstwach jest taka sama

Wsteczna propagacja błędu - podsumowanie

Funkcja realizowana przez sieć wielowarstwową

$$f_i(\mathbf{x};\mathbf{W}) = \sigma\left(\sum_j w_{ji}^{(M)}\sigma\left(\sum_k w_{kj}^{(M-1)}\sigma\left(\sum_l w_{lk}^{(M-2)}\ldots\sigma\left(\sum_n w_{nm}^{(2)}x_n\right)\right)\right)\ldots\right)$$

Aktualizacja wag w dowolnej warstwie K - **uogólniona reguła delta**

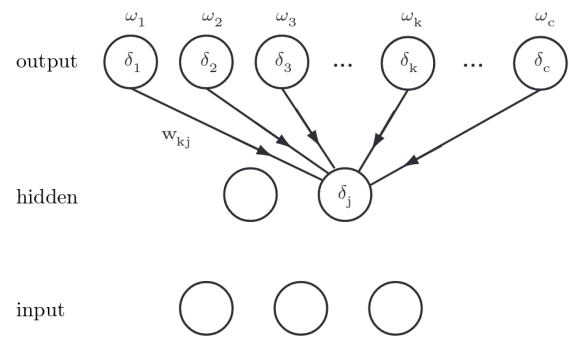
$$\Delta w_{ij}^{(K)} = \eta\delta_j^{(K)}o_i^{(K-1)}$$

gdzie sygnał wejściowy $o_i^{(1)} = x_i$
Sygnał błędu w warstwie K

$$\delta_i^{(K)} = \sigma'\left(z_i^{(K)}\right)\sum_j \delta_j^{(K+1)}w_{ij}^{(K+1)}$$

sygnał błędu w warstwie wyjściowej:

$$\delta_i^{(M)} = \sigma'(z_i^{(M)})\,(y_i - f_i(\mathbf{x};\mathbf{W}))$$



Algorytm wstecznej propagacji

1. Zainicjuj wagi $\mathbf{W}$ małymi losowymi wartościami
2. Dopóki nie spłoniony warunek stopu wykonuj
   2.1 dla pary uczącej $(\mathbf{x}, \mathbf{y})$ oblicz sygnał od wejścia do wyjścia

$$o_i^{(K)} = \sigma\left(\sum_j w_{ji}^{(K)} o_j^{(K-1)}\right)$$

   2.2 oblicz sygnał błędu od warstwy wyjściowej

$$\delta_i^{(M)} = \sigma'(z_i^{(M)})\,(y_i - f_i(\mathbf{x}))$$

   2.3 oblicz sygnał błędu od warstwy wyjściowej do wejściowej

$$\delta_i^{(K)} = \sigma'\left(z_i^{(K)}\right) \sum_j \delta_j^{(K+1)} w_{ij}^{(K+1)}$$

   2.4 aktualizuj wagi

$$\Delta w_{ij}^{(K)} = \eta \delta_j^{(K)} o_i^{(K-1)}$$

Algorytm wstecznej propagacji można w łatwy sposób rozszerzyć na:

- sieci o innej strukturze połączeń: np. dodatkowe połączenia pomiędzy wejściem lub wyjściem, lub innymi warstwami ukrytymi. Dla sieci rekurencyjnych odpowiednikiem jest BPTT (wsteczna propagacja w czasie)
- dowolną liczbę warstw
- różne nieliniowości w warstwach
- różne funkcje aktywacji dla poszczególnych neuronów
- różne stałe uczenia dla każdej warstwy lub węzła
- inne funkcje kosztu, nie tylko MSE, np. Cross Entropy

Warianty treningu

Dla zbioru zawierającego N przypadków i sieci zawierającej M wyjść

$$\hat{E} = \frac{1}{N}\sum_{i=1}^{N} E(\mathbf{x}_i) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - f_j(\mathbf{x}_i))^2$$

$$\Delta\hat{w} = \sum_{i=1}^{N} \Delta w_i$$

Batch gradient descent

- uśredniony błąd całego zbioru treningowego
- nie praktyczne dla dużych danych
- stabilny trening
- nie zależy od kolejności prezentacji wzorców

Warianty treningu

SGD - stochastic gradient descent (on-line)

- aktualizacja dla pojedynczego przypadku
- treningowego ($N = 1$), duża szybkość działania ale i duża wariancja,
- istotna jest kolejność prezentacji przypadków

MiniBatch gradient descent

- średnia z n losowych przypadków (*mini-batch*), trening on-line,
- często utożsamiany z SGD
- mniejsza wariancja niż SGD
- możliwe operacje na macierzach, łatwiejsze zrównoleglenie GPU/CPU

Własności MLP

- **Uniwersalny aproksymator:** sieć MLP z jedną warstwą ukrytą jest w stanie aproksymować dowolną funkcję ciągła z dowolną dokładnością.
  Dwie warstwy ukryte rozszerzają możliwości na funkcje nieciągłe.
  Klasyfikator potrafi zrealizować dowolne granice decyzji (obszary nie muszą być wypukłe ani połączone)
- Neurony ukryte: transformacja nieliniowa do przestrzeni odwzorowań, tworząca nowe cechy za pomocą nieliniowych kombinacji
- Wiele zastosowań: klasyfikacja, wielowymiarowa regresja

Problemy I

- Optymalizacja nieliniowych funkcji zawsze sprawia problemy, tu mamy złożenie (czasem wielu) funkcji nieliniowych
- Dobór architektury sieci: Ile węzłów w warstwie ? Jakie funkcje aktywacji? Sieci ontogeniczne (rozrastające się) dostosowujące rozmiar do złożoności problemu
- Przeuczenie i generalizacja - zbyt duża liczba optymalizowanych parametrów powoduje przeuczenie, zbyt mała - może generować zbyt proste rozwiązania
- Regularyzacja - metody ograniczenia zjawiska przeuczenia, np. modyfikacje funkcji kosztu, ograniczenie liczby parametrów
- Inicjalizacja parametrów - szybkość treningu i wynik zależy od punktu startowego

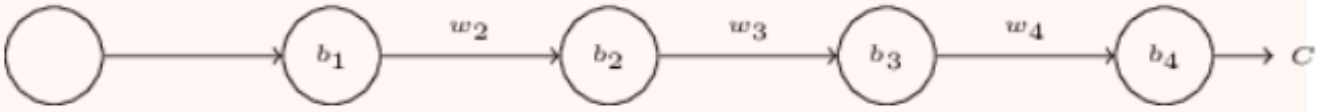
Problemy II

- Minima lokalne i plateau, wąskie „rynny” - np. wielokrotny start
- Wpływ nowych wzorców na już nauczone – zapominanie
- Dobór stałej uczenia
- Znikający gradient, eksplodujący gradient
- Przygotowanie danych uczących: normalizacja, standaryzacja, kodowanie wyjść, ...
- Ocena modelu: zbiory walidacyjne, testowe, kroswalidacja

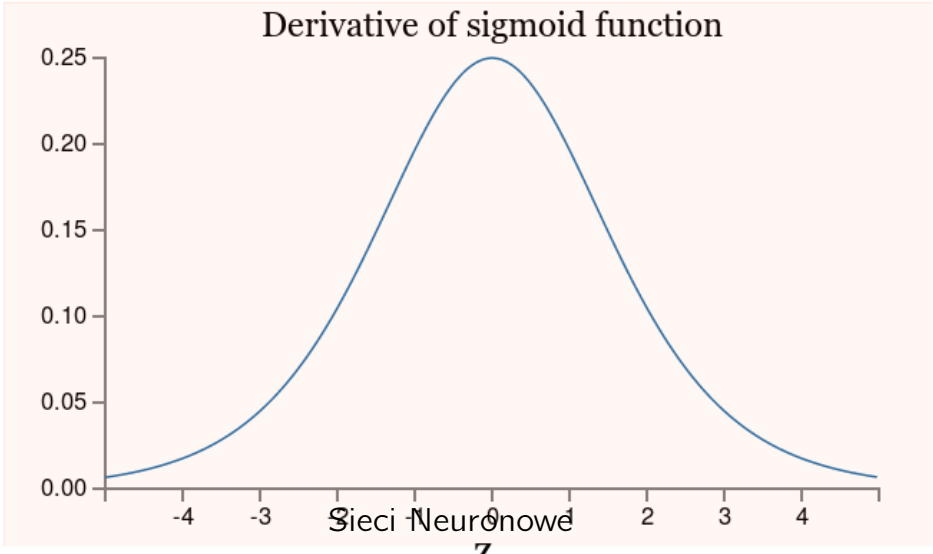
Znikający gradient

Szybkość uczenia (zmiany wag) drastycznie maleje w głębszych warstwach - sygnał błędu zanika
Przykład: sieć 1 x 1 x 1 x 1

$$\frac{\partial C}{\partial b_1} = \sigma'(z_1) \times w_2 \times \sigma'(z_2) \times w_3 \times \sigma'(z_3) \times w_4 \times \sigma'(z_4) \times \frac{\partial C}{\partial a_4}$$



gdy $|w_j| < 1$ wówczas $|w_j \sigma'(z_j)| < \frac{1}{4}$



Nielsen, 2016

Problem niestabilnego gradientu

- gradient w niższych warstwach zależy od wartości wag, aktywacji i gradientów w warstwach wyższych
- **eksplodujący gradient**
  gdy $w_i >> 1$ oraz $z_i \approx 0$ wówczas $|w_i \sigma'(z_i)| > 1$, może to owocować bardzo dużymi gradientami w początkowych warstwach
- znikający i eksplodujący gradient to przypadki szczególne **problemu niestabilnego gradientu** -> wartości gradientów w poszczególnych warstwach mogą znacznie się różnić, warstwy uczą się z różnym tempem

Specyfikacja sieci i metody treningu

- Specyfikacja modelu:
  - specyfikacja architektury
  - typ neuronów ukrytych: liniowe, sigmoidalne, ReLU, ...
  - typ neuronów wyjściowych: liniowe, sigmoidalne, softmax, ...
- Funkcja kosztu:
  - mean squared error (MSE),
  - cross-entropy (CE),
- Optymalizacja metodami spadku gradientu lub innymi:
  - SGD, RPROP, Quckprop,
  - Adam, AdaGrad, ...
- Metody regularyzacji:
  - momentum, L_1, L_2 (weight decay), ...

Funkcja kosztu MSE

Błąd średniokwadratowy (*Mean Squared Error*)

$$J(\Theta) = \frac{1}{N} \sum (\hat{y} - y)^2$$

Gradient funkcji kosztu:

$$\frac{\partial J}{\partial \Theta} = \frac{2}{N} \sum (\hat{y} - y) \frac{\partial \hat{y}}{\partial \Theta}$$

gdzie
$\hat{y}$ - wyjścia sieci,
Θ - parametry sieci,
y - oczekiwane wyjścia

Funkcja kosztu Cross Entropy

Cross entropy

$$J(\Theta) = -\sum y \log \hat{y}$$

- miara odległości miedzy rozkładami y i $\hat{y} \in (0, 1)$
- gdy $y \approx \hat{y}$ to funkcja kosztu bliska zeru
- $J > 0$

Gradient funkcji kosztu:

$$\frac{\partial J}{\partial \Theta} = -\sum y \frac{1}{\hat{y}} \frac{\partial \hat{y}}{\partial \Theta}$$

☞ Wizualizacja funkcji CE

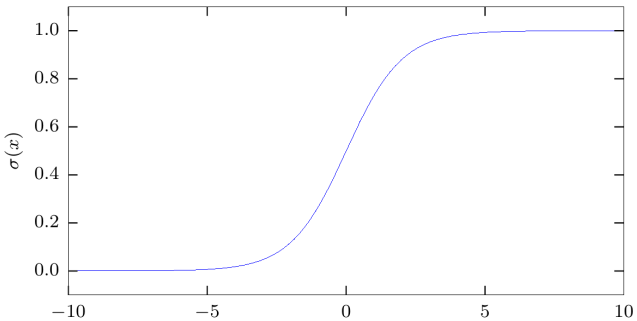
Funkcja liniowa w warstwie wyjściowej

$$z = \mathbf{w}^T \mathbf{h} + b$$

- nieograniczone wartości $(-\infty, +\infty)$
- jednostki nie nasycają się ale bardzo duże wartości aktywacji również mogą przysporzyć problemów w trakcie uczenia
- stabilniejsze w optymalizacji metodami gradientowymi od innych typowych funkcji nieliniowych stosowanych w sieciach
- zastosowanie wraz funkcją kosztu MSE do problemów regresji (dane wyjściowe ciągłe), modelowanie rozkładów gusowskich

Funkcja logistyczna w warstwie wyjściowej

$$\sigma(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$$



- wartości ograniczone $(0, 1)$
- nasycają się, aktywne uczenie tylko w okolicach $x \approx 0$
- zastosowanie: klasyfikacja 2 klas, modelowanie rozkładu Bernoullego, binarne wartości wyjściowe (*multi-label classification*)
- binarna funkcja kosztu Cross-Entropy, na każdym wyjściu oczekujemy stanu 0 lub 1

$$J_{CE} = -\sum y \ln \hat{y} = -\left[y \ln \hat{y} + (1-y) \ln(1-\hat{y})\right]$$

Funkcja sigmoidalna w warstwie wyjściowej

- podczas minimalizacji MSE błąd zanika, gdy funkcja się nasyca (także dla niepoprawnej odpowiedzi), wówczas $\frac{\partial \hat{y}}{\partial \Theta} \approx 0$

$$\frac{\partial J_{MSE}}{\partial \Theta} = \frac{2}{N} \sum (\hat{y} - y) \frac{\partial \hat{y}}{\partial \Theta}$$

- problem znika przy zastosowaniu kosztu Cross Entropy (CE)

$$\frac{\partial J_{CE}}{\partial \Theta} = \sum (\hat{y} - y) \frac{\partial z}{\partial \Theta}$$

gdzie z jest liniową aktywacją (*logit*) neuronu wyjściowego

Softmax w warstwie wyjściowej

$$softmax(\mathbf{z})_i = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

- wartości $(0, 1)$ unormowane

$$\sum_i softmax(\mathbf{z})_i = 1$$

- Zastosowanie: klasyfikacja n rozłącznych klas (*multi-class classification*),
  każde wyjście sieci związane jest z jedną klasą, wartości kodowane w postaci wektora binarnego (kodowanie *one hot*)

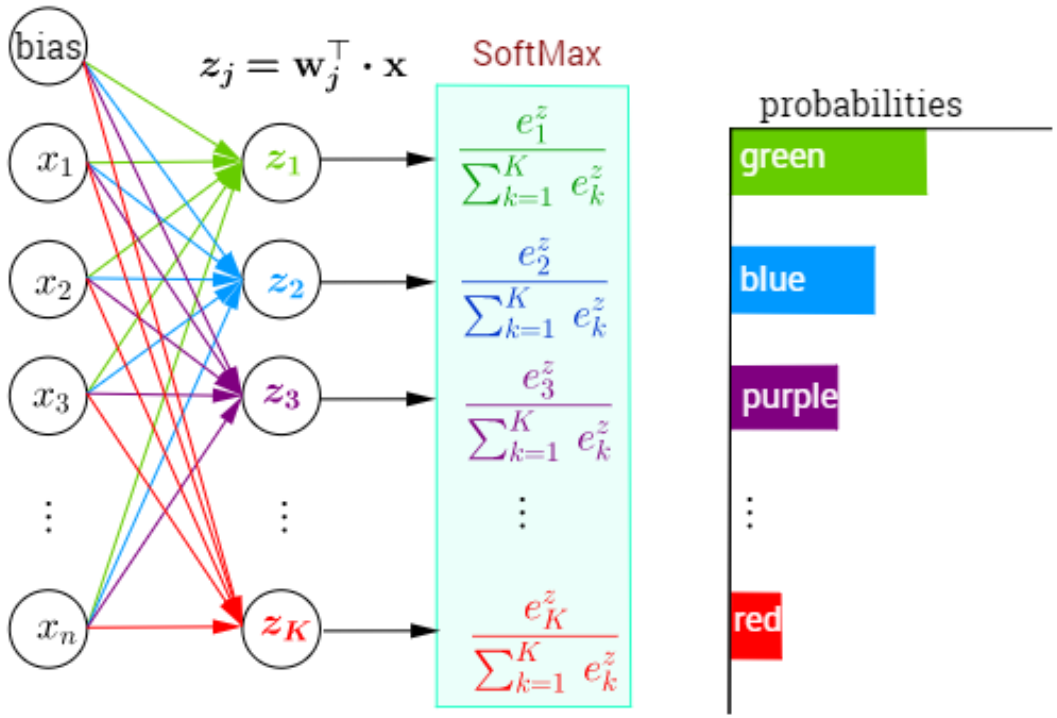
$$[0, \ldots, 0, 1, 0, \ldots, 0]$$

- wartości wyjściowe sieci reprezentują rozkład prawdopodobieństw zmiennej wyjściowej

$$\hat{y}_i = P(y = i|X), \qquad i = 1, \ldots, n$$

Multi-Class Classification with NN and SoftMax Function

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_K \end{bmatrix} = \begin{bmatrix} \mathbf{w}_1^\top \\ \mathbf{w}_2^\top \\ \mathbf{w}_3^\top \\ \vdots \\ \mathbf{w}_K^\top \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}$$



Źródło:
stats.stackexchange.com/questions/265905/derivative-of-softmax-with-respect-to-weights

- gradient

$$\frac{\partial softmax(\mathbf{z})_i}{\partial z_j} = softmax(\mathbf{z})_i(\delta_{ij} - softmax(\mathbf{z})_i)$$

- gradient funkcji kosztu CE gdy $\hat{y}_i = softmax(\mathbf{z})_i$

$$\frac{\partial J_{CE}}{\partial \Theta} = \sum(\hat{y} - y)\frac{\partial z}{\partial \Theta}$$

- obliczenia się upraszczają i znika problem zanikania błędu dla nasyconych wyjść, który występował dla funkcji kosztu MSE i wyjść sigmoidalnych

- softmax dobrze współgra z funkcją kosztu *cross entropy*, błąd nie znika nawet dla dużych wartości aktywacji z_i

$$\log softmax(\mathbf{z})_i = z_i - \log \sum_j e^{z_k}$$

- ważne są różnice wartości pomiędzy wyjściami a nie amplitudy poszczególnych wyjść

$$softmax(x) = softmax(x + c)$$

- minimalizacja kosztu CE powoduje zwiększenie wartości z_i na jednym z wyjść i zmniejszenie wszystkich pozostałych

$$\log \sum_k e^{z_k} \approx \max_i z_i$$

- z punktu widzenia neurobiologicznego softmax może być widziany jako realizacja mechanizmów hamowania występującego w grupach neuronów w korze, podobnie do mechanizmu zwycięzca bierze wszystko (*winner takes all*)

Dobór funkcji kosztu uzależniony jest od rozkładu wartości wyjściowych i typu neuronów wyjściowych

Output Type	Output Distribution	Output Layer	Cost Function
Binary	Bernoulli	Sigmoid	Binary cross-entropy
Discrete	Multinoulli	Softmax	Discrete cross-entropy
Continuous	Gaussian	Linear	Gaussian cross-entropy (MSE)
Continuous	Mixture of Gaussian	Mixture Density	Cross-entropy
Continuous	Arbitrary	See part III: GAN, VAE, FVBN	Various

Gooffellow, Deep Learning, 2016 [?]

Dobór neuronów ukrytych

- wybór funkcji realizowanych przez warstwy ukryte jest kluczowy dla przebiegu procesu uczenia
- funkcje liniowe - wiele warstw liniowych sprowadza się do pojedynczej transformacji liniowej

$$f(\mathbf{x}) = \mathbf{W}_n\mathbf{W}_{n-1}\ldots\mathbf{W}_1\mathbf{W}_1\mathbf{x} = \mathbf{A}\mathbf{x}$$

- funkcje ograniczone: sigmoidalne i tangens hiperboliczny powodują problemy z niestabilnym gradientem, uczenie może utknąć, gdy funkcja się nasyci
- obecnie najpopularniejszym typem aktywacji w sieciach jest ReLU

Demonstracje

- ☞ TensorFlow playground
- ☞ ConvNetJS - Deep Learning in your browser