Sztuczne sieci neuronowe

Krzysztof A. Cyran
POLITECHNIKA ŚLĄSKA
Instytut Informatyki, p. 311

Wykład 6

PLAN:

- **Repetitio (brevis)**
- **Sieci neuronowe z radialnymi funkcjami bazowymi**

Repetitio

- *W aspekcie architektury*: zajmowaliśmy się tylko sieciami typu *feed-forward*
- *W aspekcie działania pojedynczego neuronu*: rozważaliśmy tylko neurony obliczające sumy ważone wejść (ze względu na najczęściej stosowaną funkcję aktywacji neurony te są zwane *neuronami sigmoidalnymi*)

Repetitio (cd.)

W aspekcie uczenia: szczegółowo przedstawiono tylko gradientowe algorytmy uczenia nadzorowanego:

- *backpropagation* (wraz z modyfikacjami inercyjnymi) (algorytm I rzędu)
- *metodę zmiennej metryki* (algorytm II rzędu)
- *metodę gradientów sprzężonych* (algorytm I rzędu)

Repetitio (cd.)

Ponadto:

- Podano typowe reguły uczenia (nienadzorowanego i nadzorowanego) jako przypadki szczególne uogólnionej reguły uczenia
- Przedstawiono metody:
  - doboru współczynników uczenia
  - inicjalizacji wag
  - doboru architektury sieci
  - zwiększania zdolności generalizujących

Sieci neuronowe z radialnymi funkcjami bazowymi

- Sieci o radialnych funkcjach bazowych (RBF) składają się z jednej warstwy ukrytej oraz jednej warstwy wyjściowej.
- Neurony warstwy wyjściowej są liniowe
- Neurony ukryte realizują funkcję φ zmieniającą się radialnie (stąd nazwa sieci) wokół wybranego centrum $\mathbf{c}$

Sieci RBF (cd.)

Formalnie:

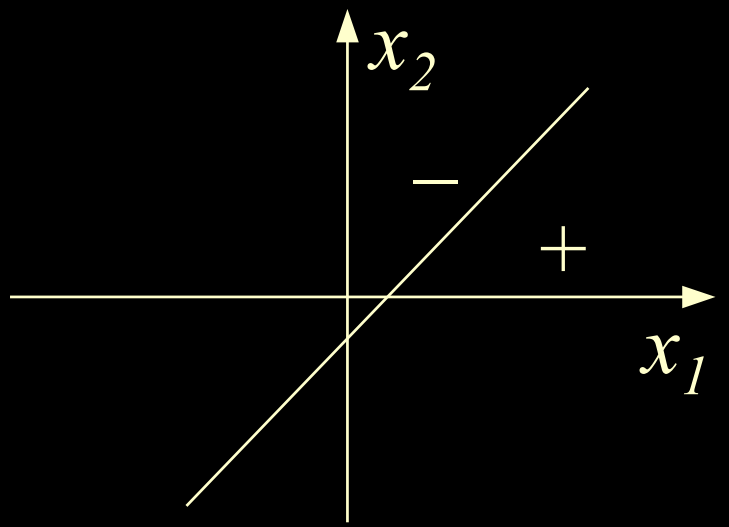
Wyjścia neuronów ukrytych (radialnych) sieci RBF generują wektor sygnałów **y** dany równaniem:

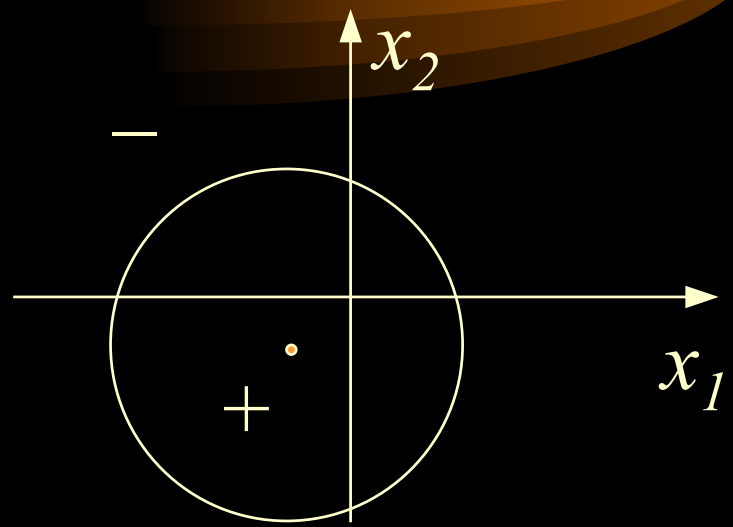
$$\mathbf{y} = \varphi\left(\|\mathbf{x} - \mathbf{c}\|\right)$$

Porównanie działania neuronów sigmoidalnych i radialnych

- Neuron sigmoidalny występujący w MLP reprezentuje w przestrzeni wejściowej hiperpłaszczyznę dzielącą tę przestrzeń na dwie otwarte klasy
- Neuron radialny reprezentuje hipersferę dokonującą podziału kołowego wokół punktu centralnego

Graficzne porównanie działania neuronu sigmoidalengo i radialnego





$$y_i = f\left(\sum_j w_{ij} x_j\right)$$

$$y_i = \varphi_i\left(\sum_j \left(x_j - t_{ij}\right)^2\right)$$

Sieci RBF (cd.)

- W zadaniach zawierających symetrie kołowe zastosowanie neuronów radialnych pozwala znacznie zmniejszyć liczbę neuronów ukrytych (a zatem zwiększyć generalizację)
- Ponadto wystarczy zawsze jedna warstwa ukryta

Nieliniowa φ-separowalność

- Niech $\varphi(x) = [\varphi_1(x), \varphi_2(x), ..., \varphi_k(x)]^T$ będzie wektorem funkcji radialnych w N wymiarowej przestrzeni wejściowej.
- Podział tej przestrzeni na klasy X^+ oraz X^- jest *nieliniowo φ - separowalny* jeśli istnieje taki wektor **w**, że:

$$\mathbf{w}^T\varphi(\mathbf{x}) > 0 \quad \text{dla} \quad \mathbf{x} \in X^+$$

$$\mathbf{w}^T\varphi(\mathbf{x}) < 0 \quad \text{dla} \quad \mathbf{x} \in X^-$$

Nieliniowa φ-separowalność (cd.)

Zatem: Problem jest nieliniowo φ-separowalny w przestrzeni wejściowej $\mathbf{x}$ wymiaru N jeśli jest liniowo separowalny w przestrzeni rzutowania $\varphi(\mathbf{x})$ wymiaru k. Granica między obu klasami w przestrzeni $\varphi(\mathbf{x})$ jest zdefiniowana za pomocą hiperpłaszczyzny o równaniu:

$$\mathbf{w}^{\mathrm{T}}\varphi(\mathbf{x}) = 0$$

Twierdzenie o nieliniowej φ-separowalności

Dowolny zbiór wzorców jest nieliniowo φ-separowalny pod warunkiem przyjęcia odpowiednio dużego wymiaru k przestrzeni rzutowania

Wniosek z twierdzenia o nieliniowej φ-separowalności

Przyjęcie dostatecznie dużej liczby neuronów radialnych realizujących funkcje $\varphi_i(x)$ zapewnia rozwiązanie dowolnego problemu klasyfikacyjnego przy użyciu dwu warstw:

- Warstwy ukrytej realizującej wektor $\varphi(x)$ oraz
- Warstwy wyjściowej realizowanej przez neuron liniowy z wektorem wagowym $\mathbf{w}$.

Interpolacja wielowymiarowa w sieciach RBF

- Poszukujemy interpolacji wielowymiarowej odwzorowującej p różnych wektorów wejściowych $\mathbf{x}_i$ ($i = 1, 2, ..., p$) z przestrzeni wejściowej N wymiarowej w zbiór p liczb rzeczywistych d_i ($i = 1, 2, ..., p$) za pomocą sieci RBF

Interpolacja wielowymiarowa w sieciach RBF (cd.)

- Przedstawiona interpolacja jest równoważna poszukiwaniu takiej funkcji radialnej $F(\mathbf{x})$, dla której spełnione są warunki interpolacji:

$$F(\mathbf{x}_i) = d_i$$

$$\text{gdzie:} \quad F(\mathbf{x}) = \sum_{i=1}^{p} w_i \varphi(\|\mathbf{x} - \mathbf{x}_i\|)$$

- Wybór normy jest w zasadzie dowolny, choć w praktyce najczęściej stosuje się normę euklidesową. Wartości wektorów $\mathbf{x}_i$ stanowią centra funkcji radialnej.

Interpolacja wielowymiarowa w sieciach RBF – funkcje Greena

Jako funkcje φ przyjmuje się zazwyczaj funkcje radialne Greena $G(\mathbf{x}; \mathbf{x}_i)$ z których najpopularniejszą jest (nieznormalizowana) funkcja Gaussa:

$$G[\mathbf{x}; \mathbf{x}_i] = e^{-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2\sigma_i^2}} = e^{-\frac{1}{2\sigma_i^2}\sum_{k=1}^{N}(x_k - x_{i,k})^2}$$

w której $\mathbf{x}_i$ oznaczają wektory wartości średnich (centrów) a σ_i^2 wariancje

Interpolacja jako superpozycja wielowymiarowych funkcji Gaussa

- Po podstawieniu w miejsce funkcji radialnej wielowymiarowej nieznormalizowanej funkcji Gaussa otrzymuje się następujące równanie interpolujące:

$$F(\mathbf{x}) = \sum_{i=1}^{p} w_i e^{-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{2\sigma_i^2}}$$

- Powyższe równanie przedstawia superpozycję wielowymiarowych funkcji Gaussa z centrami (wartościami oczekiwanymi) ulokowanymi w $\mathbf{x}_i$ i szerokościami (odchyleniami standardowymi) σ_i

Ograniczenia

- Choć przedstawienie poprzednie jest zawsze możliwe, jest ono niepraktyczne ze względu na ilość neuronów ukrytych równą ilości próbek uczących p.
- Dlatego w praktyce stosuje się przybliżenie z ograniczeniem do K neuronów

Przybliżenie funkcji interpolującej

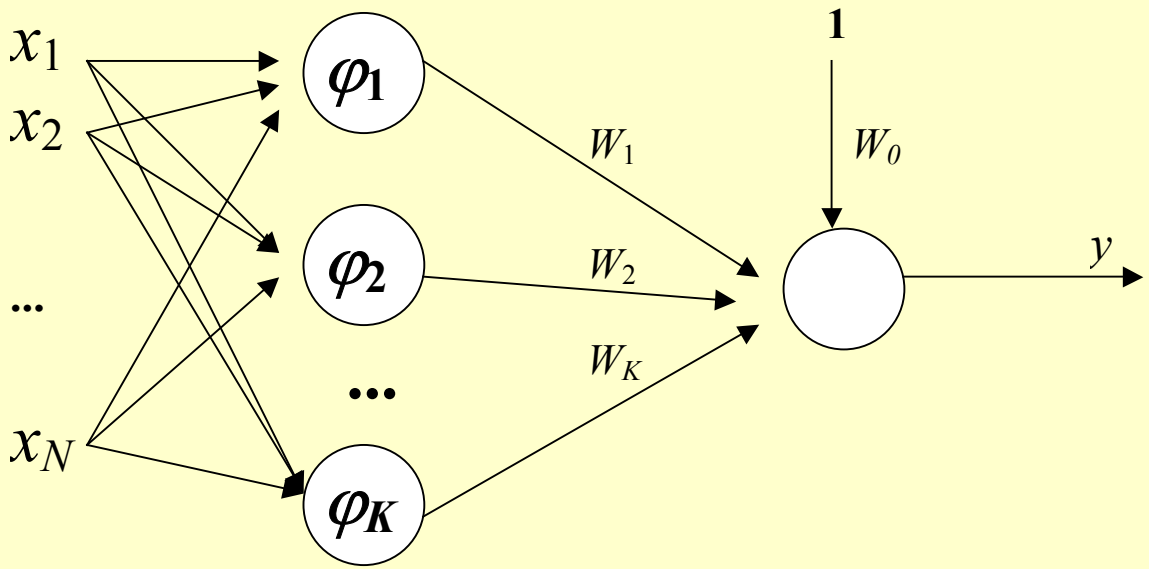
- Funkcję F przybliża się funkcją F^* daną:

$$F^*(\mathbf{x}) = \sum_{i=1}^{K} w_i G(\mathbf{x}; \mathbf{t}_i),$$

$$\text{gdzie}: G(\mathbf{x}; \mathbf{t}_i) = G(\|\mathbf{x} - \mathbf{t}_i\|), \quad K < p$$

- Wektory $\mathbf{t}_i$ ($i = 1, \ldots, K$) są centrami które należy wyznaczyć (w przypadku szczególnym , gdy $K = p$ otrzymuje się rozwiązanie dokładne dla którego oczywiście: $\mathbf{t}_i = \mathbf{x}_i$

Schemat sieci RBF



$$\varphi_i = G(\mathbf{x}; \mathbf{t}_i)$$

Uczenie sieci RBF

Uczenie sieci RBF polega na takim doborze

- wag w_i
- funkcji Greena $G(\mathbf{x}; \mathbf{t}_i)$

gdzie ($i = 1, ..., K$) aby funkcja F^* realizowana przez sieć RBF możliwie najlepiej przybliżała teoretyczną funkcję dokładną F.

Sieci z hiper-radialnymi funkcjami bazowymi (HRBF)

- W sieciach RBF funkcje bazowe Greena zależne są od normy euklidesowej
- W sieciach HRBF korzysta się z uogólnionej normy euklidesowej, dla której każdy wymiar ma swój odrębny współczynnik wagi, a nawet wektor wag

Sieci HRBF - uogólniona norma euklidesowa

- Uogólniona norma euklidesowa:

$$\|\mathbf{x}\|_{\mathbf{Q}}^2 = (\mathbf{Q}\mathbf{x})^{\mathbf{T}}(\mathbf{Q}\mathbf{x}) = \mathbf{x}^{\mathbf{T}}\mathbf{Q}^{\mathbf{T}}\mathbf{Q}\mathbf{x}$$

Przyjmując: $\mathbf{Q}^{\mathbf{T}}\mathbf{Q} = \mathbf{C} = [c_{kl}]$ dostajemy:

$$\|\mathbf{x}\|_{\mathbf{Q}}^2 = \sum_{k=1}^{N}\sum_{l=1}^{N} c_{kl}x_k x_l$$

Norma euklidesowa jako przypadek szczególny normy uogólnionej

- W szczególności jeśli **Q** jest macierzą diagonalną, wówczas:

$$\|\mathbf{x}\|_{\mathbf{Q}}^2 = \sum_{k=i}^{N} c_{ii} x_i^2$$

- Przy dodatkowym załażeniu, że **Q = 1** norma wagowa sprowadza się do normy klasycznej: $\|\mathbf{x}\|_{\mathbf{Q}}^2 = \|\mathbf{x}\|^2$

Działanie sieci HRBF

- Zastosowanie uogólnionej normy wagowej daje uogólnione wyrażenie rozwinięcia funkcji radialnych:

$$F^*(\mathbf{x}) = \sum_{i=1}^{K} w_i G\left(\|\mathbf{x} - \mathbf{t}_i\|_{\mathbf{Q}}\right)$$

- Jednocześnie powyższe równanie opisuje działanie sieci HRBF

Uogólniona (nieznormalizowana) funkcja Gaussa

Nieznormalizowana uogólniona funkcja Gaussa otrzymuje się poprzez przyjęcie dla i-tej funkcji bazowej:

- gaussowskiej funkcji radialnej Greena o centrum $\mathbf{t}_i$ oraz
- Macierzy wagowej $\mathbf{Q}_i$

Dana jest ona wzorem:

$$G\left(\left\|\mathbf{x}-\mathbf{t}_i\right\|_{\mathbf{Q}_i}\right)=e^{\left(-[\mathbf{x}-\mathbf{t}_i]^{\mathrm{T}}\mathbf{Q}_i^{\mathrm{T}}\mathbf{Q}_i[\mathbf{x}-\mathbf{t}_i]\right)}=e^{\left(-\frac{1}{2}[\mathbf{x}-\mathbf{t}_i]^{\mathrm{T}}\mathbf{S}_i^{-1}[\mathbf{x}-\mathbf{t}_i]\right)}$$

Uogólniona (nieznormalizowana) funkcja Gaussa (cd.)

- Nieznormalizowana uogólniona funkcja Gaussa dana jest zatem wzorem:

$$G\left(\left\|\mathbf{x}-\mathbf{t}_i\right\|_{\mathbf{Q}_i}\right)=e^{\left(-[\mathbf{x}-\mathbf{t}_i]^{\mathrm{T}}\mathbf{Q}_i^{\mathrm{T}}\mathbf{Q}_i[\mathbf{x}-\mathbf{t}_i]\right)}=$$

$$=e^{\left(-\frac{1}{2}[\mathbf{x}-\mathbf{t}_i]^{\mathrm{T}}\mathbf{S}_i^{-1}[\mathbf{x}-\mathbf{t}_i]\right)}$$

- Widać, że wyrażenie $\frac{1}{2}\,\mathbf{S}_i^{-1}=\mathbf{Q}_i^{\mathrm{T}}\mathbf{Q}_i$ pełni funkcję czynnika $1/(2\sigma_i^2)$ standardowej wielowymiarowej funkcji Gaussa

Porównanie sieci RBF z sieciami sigmoidalnymi

- W sieciach RBF parametr funkcji aktywacji σ_i jest zależny od neuronu i podlega uczeniu, natomiast w sieciach sigmoidalnych analogiczny parametr β jest stały i jednakowy dla wszystkich neuronów.
- Argumentem funkcji radialnej jest odległość danej próbki $\mathbf{x}$ od centrum $\mathbf{t}_i$, a w sieci sigmoidalnej jest to iloczyn skalarny wektorów $\mathbf{w}^T\mathbf{x}$.
- Neurony radialne dzielą przestrzeń na obszary lokalne poprzez hipersfery, natomiast neurony sigmodalne dzielą przestrzeń na obszary globalne poprzez hiperpłaszczyzny.

Uczenie sieci (H)RBF

- Uczenie sieci (hiper)radialnych składa się z dwóch etapów:
  1. dobór centrów i parametrów kształtu funkcji bazowych
  2. dobór wag liniowych neuronów wyjściowych

Uczenie sieci (H)RBF (cd.)

- Ponieważ zadanie drugie (doboru wag wyjściowych neuronów liniowych) może zostać rozwiązane algebraicznie, o ile znane jest rozwiązanie zadania pierwszego, zatem to zadanie pierwsze jest podstawowym zadaniem w uczeniu sieci (H)RBF

Metody wyznaczania centrów i parametrów kształtu funkcji bazowych

- Losowy wybór centrów funkcji bazowych
- Samoorganizujący się proces podziału na klastery
- Uczenie pod nadzorem (oparte o algorytmy propagacji wstecznej)

Losowy wybór centrów funkcji bazowych

- Rozwiązanie najprostsze, lecz dopuszczalne dla sieci radialnych przy założeniu, że rozkład danych uczących dobrze odzwierciedla specyfikę problemu (a tak zawsze powinno być)
- Wówczas dobór stałych parametrów funkcji bazowych jest dokonywany losowo przy rozkładzie równomiernym

Dobór losowy centrów (cd.)

- Po dokonaniu losowego wyboru centrów $\mathbf{t}_i$ oblicza się wartość parametru związanego z odchyleniem standardowym:

$$\frac{1}{2\sigma^2} = \frac{d^2}{K} \quad \Rightarrow \quad \sigma = \frac{\mathrm{d}}{\sqrt{2\mathrm{K}}}$$

- W powyższym wzorze *d* oznacza maksymalną odległość między centrami $\mathbf{t}_i$
- Funkcje bazowe są zatem postaci:

$$G\left(\left\|\mathbf{x}-\mathbf{t}_i\right\|^2\right) = e^{\left(-\frac{\left\|\mathbf{x}-\mathbf{t}_i\right\|^2}{\frac{d^2}{K}}\right)}$$

Samoorganizujący się proces podziału na klastery

- Dane wejściowe dzieli się na klastery (np. za pomocą algorytmu K-uśrednień)
- Liczba funkcji bazowych równa jest ilości klasterów
- Do centrum każdego klastera przyporządkowuje się centrum funkcji bazowej

Gradientowe uczenie pod nadzorem sieci HRBF

W metodzie tej modyfikuje się równocześnie:

- centra funkcji bazowych,
- ich parametry oraz
- wartości wektora wag neuronów liniowych warstwy wyjściowej

Gradientowe uczenie pod nadzorem sieci HRBF (cd.)

- Zdefiniujmy błąd E jako:

$$E = \frac{1}{2}\left\{\sum_{i=0}^{K}\left[w_i\varphi_i(x)\right] - d\right\}^2$$

- Przy czym wyjście sieci y jest określone jako:

$$y = \sum_{i=0}^{K} w_i\varphi_i(\mathbf{x}),$$

$$\varphi_0(\mathbf{x}) = 1, \quad \varphi_i(\mathbf{x}) = e^{\left(-\frac{1}{2}[\mathbf{Q}_i(\mathbf{x}-\mathbf{t}_i)]^{\mathrm{T}}[\mathbf{Q}_i(\mathbf{x}-\mathbf{t}_i)]\right)}$$

Składowe gradientu względem parametrów podlegających uczeniu

$$j = 1..N,\ i = 1..K,\ k = 1..N$$

$$\frac{\partial E}{\partial w_0} = y - d, \quad \frac{\partial E}{\partial w_i} = e^{-\frac{1}{2}u_i}(y - d)$$

$$\frac{\partial E}{\partial t_j^{(i)}} = e^{-\frac{1}{2}u_i} w_i (y - d) \sum_{k=1}^{N} Q_{kj}^{(i)} z_k^{(i)}$$

$$\frac{\partial E}{\partial Q_{jk}^{(i)}} = -e^{-\frac{1}{2}u_i} w_i (y - d)\left(x_k - t_k^{(i)}\right) z_j^{(i)}$$

Zmienne wykorzystywane do określania składowych gradientu

$$j = 1..N,\ i = 1..K,\ k = 1..N$$

$$z_j^{(i)} = \sum_{k=1}^{N} Q_{jk}^{(i)} \left(x_k - t_k^{(i)} \right)$$

$$u_i = \sum_{k=1}^{N} \left(z_k^{(i)} \right)^2$$

Gradientowe uczenie pod nadzorem sieci HRBF (cd.)

Znając składowe gradientu funkcji błędu względem wszystkich parametrów podlegających uczeniu można stosować dowolną metodę optymalizacji gradientowej, np. największego spadku, w której kierunek poszukiwań $\mathbf{p} = -\eta \nabla E$