Sztuczne Sieci Neuronowe

Wykład 7

Sieci neuronowe
o radialnych funkcjach bazowych

wykład przygotowany na podstawie.
S. Osowski, “Sieci Neuronowe w ujęciu algorytmicznym ”, Rozdz. 5, PWNT, Warszawa 1996.

Bazowe funkcje radialne

Sieci neuronowe wielowarstwowe odgrywają, z punktu widzenia matematycznego, rolę aproksymacji stochastycznej funkcji wielu zmiennych, odwzorowując zbiór zmiennych wejściowych $\mathbf{x} \in \mathbf{R}^N$ w zbiór zmiennych $\mathbf{y} \in \mathbf{R}^M$.

Jest to aproksymacja typu globalnego

Ze względu na na charakter sigmoidalnej funkcji aktywacji neuron raz załączony (po przekroczeniu pewnej progowej wartości sygnału sumacyjnegu $\mathbf{u}_i$) pozostaje w tym stanie przy dowolnej wartości $\mathbf{u}_i$ większej od tego progu.

Odwzorowanie wartości funkcji w dowolnym punkcie przestrzeni jest więc dokonywane zbiorowym wysiłkiem wielu neuronów na raz (stąd aproksymacja globalna)

Bazowe funkcje radialne

Komplementarny sposób rozumienia odwzorowania zbioru wejściowego w wyjściowy, to odwzorowanie przez dopasowanie wielu pojedynczych funkcji aproksymujących do wartości zadanych, ważne jedynie w wąskim obszarze przestrzeni wielowymiarowej.

Jest to aproksymacja typu lokalnego

W takim rozwiązaniu odwzorowanie pełnego zbioru danych jest sumą odwzorowań lokalnych. Neurony ukryte stanowią zbiór funkcji bazowych typu lokalnego.

Bazowe funkcje radialne

Specjalną odmianę stanowią funkcje o radialnej funkcji bazowej, w której neuron ukryty realizuje funkcje zmieniającą się radialnie wokół wybranego ***centrum c***.

Funkcje takie oznaczane ogólnie **$\phi(||x-c||)$** nazywamy

radialnymi funkcjami bazowymi

Rola neuronu ukrytego będzie się sprowadzać w sieciach radialnych do odwzorowania radialnego przestrzeni wokół jednego punktu zadanego lub grupy takich punktów stanowiących klaster.

Superpozycja sygnałów pochodzących od wszystkich neuronów ukrytych, dokonywana przez neuron wyjściowy, umożliwia uzyskanie odwzorowania całej przestrzeni punktów.

Bazowe funkcje radialne

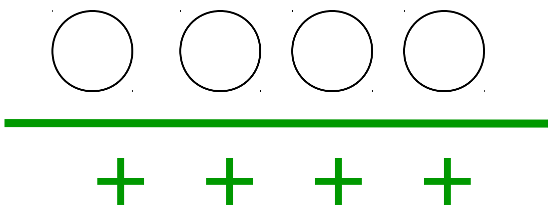
Sieci typu radialnego stanowią naturalne uzupełnienie sieci sigmoidalnych.

Neuron sigmoidalny reprezentował w przestrzeni wielowymiarowej ***hiperpłaszczyznę*** separującą tą przestrzeń na dwie kategorie (klasy), w których był spełniony odpowiedni warunek, albo
$\Sigma W_{ij} x_j > 0$ albo $\Sigma W_{ij} x_j < 0$.

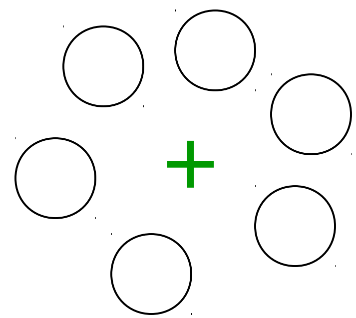
Neuron radialny z kolei reprezentuje ***hipersferę***, dokonującą podziału kołowego wokół punktu centralnego.

Bazowe funkcje radialne

Ilustracja podzialu przestrzeni danych



siec sigmoidalna



siec radialna

Stanowi to naturalne uzupełnienie neuronu sigmoidalnego, umożliwiające w wypadku wystąpienia naturalnej kołowej symetrii danych wydatne zmniejszenie liczby neuronów potrzebnych do realizacji zadania klasyfikacyjnego.

Bazowe funkcje radialne

W sieciach radialnych nie występuje potrzeba stosowania wielu warstw ukrytych.

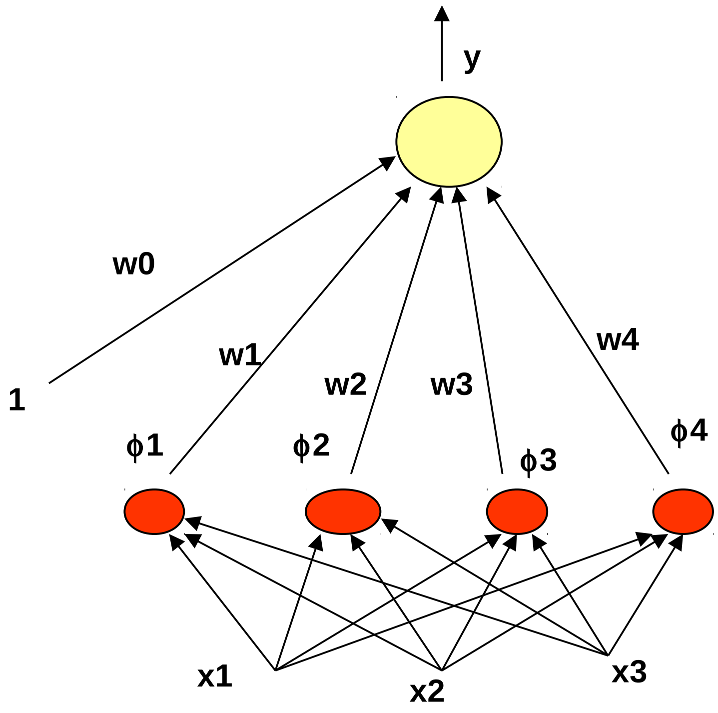
Typowa **sieć radialna** jest strukturą zawierającą

→ ***warstwę wejściową,*** na którą są podawane sygnały opisane wektorem wejściowym x,

→ ***warstwę ukrytą*** o neuronach radialnych

→ **warstwę wyjściową** zwykle złożoną z jednego neuronu (zwykle liniowy) którego jedyną rolą jest sumowanie wagowe sygnałów od neuronów ukrytych.

Sieć radialna



Bazowe funkcje radialne

Podstawą matematyczną ***funkcjonowania sieci radialnych*** jest twierdzenie Covera o separowalnosci wzorców, stwierdzające, że

złożony problem klasyfikacyjny zrzutowany nieliniowo w przestrzeń wielowymiarową ma większe prawdopodobieństwo być liniowo separowalny niż przy rzutowaniu w przestrzeń o mniejszej liczbie wymiarów.

Bazowe funkcje radialne

Przy oznaczeniu przez $\phi(x)$ wektora funkcji radialnych w **N**-wymiarowej przestrzeni wejściowej, podział tej przestrzeni na X^+, X^- przy czym X^+ i X^- reprezentują 2 klasy przestrzeni, jest liniowo ***ϕ-separowalny*** jeżeli istnieje taki zestaw wag **w**, że

$$w^T \phi(x) > 0 \quad x \in X^+$$
$$w^T \phi(x) < 0 \quad x \in X^-$$

Równanie $w^T\phi(x) = 0$ reprezentuje granicę między obu klasami.

Bazowe funkcje radialne

Udowodniono, że każdy zbiór wzorców losowo rozmieszczony w przestrzeni wielowymiarowej jest **ϕ*–separowalny*** z prawdopodobieństwem równym **1** pod warunkiem ***dużego wymiaru K*** przestrzeni rzutowania.

Oznacza to że przyjęcie ***dostatecznie dużej liczby neuronów ukrytych*** realizujących funkcje radialne **$\phi_i(x)$** zapewnia rozwiązanie problemu klasyfikacyjnego przy użyciu dwu warstw sieci:

$\rightarrow$ warstwy ukrytej realizującej wektor **$\phi(x)$**

$\rightarrow$ warstwy wyjściowej o jednym neuronie liniowym opisanym wektorem wagowym W (pozwala na superpozycję sygnału z warstwy ukrytej).

Bazowe funkcje radialne

Najprostsza sieć neuronowa typu radialnego działa na zasadzie wielowymiarowej interpolacji, której zadaniem jest odwzorowanie **p** różnych wektorów wejściowych **x_i (i=1,2,...,p)** z przestrzeni wejściowej **N-**wymiarowej w zbiór **p** liczb rzeczywistych **d_i (i=1,2,....,p)**, czyli określenie takiej funkcji radialnej **F(x)**, dla której spełnione są warunki interpolacji

$$F(x_i) = d_i$$

Przyjęcie **p** neuronów ukrytych połączonych wagami **w_i** z neuronami wyjściowymi (liniowymi) odpowiada tworzeniu sygnałów wyjściowych sieci jako sumy wagowej wartości poszczególnych funkcji bazowych.

Bazowe funkcje radialne

Niech dana będzie sieć radialna o jednym wyjściu i p parach uczących (x_i, d_i). Przy założeniu p centrów umieszczonych w kolejnych wektorach x_i, to jest $c_i = x_i$, otrzymuje się układ równań liniowych względem wag w_i.

Układ ten można zapisać w postaci macierzowej

$$\Phi\, w = d$$

Wykazano, że dla szeregu funkcji radialnych przy założeniu $x_1 \neq x_2 \neq ... \neq x_p$ kwadratowa macierz Φ jest nieosobliwa, a przy tym ujemnie określona.
Istnieje zatem rozwiązanie równania

$$w = \Phi^{-1}\, d$$

Pozwalające określić wektor wag w neuronu wyjściowego sieci.

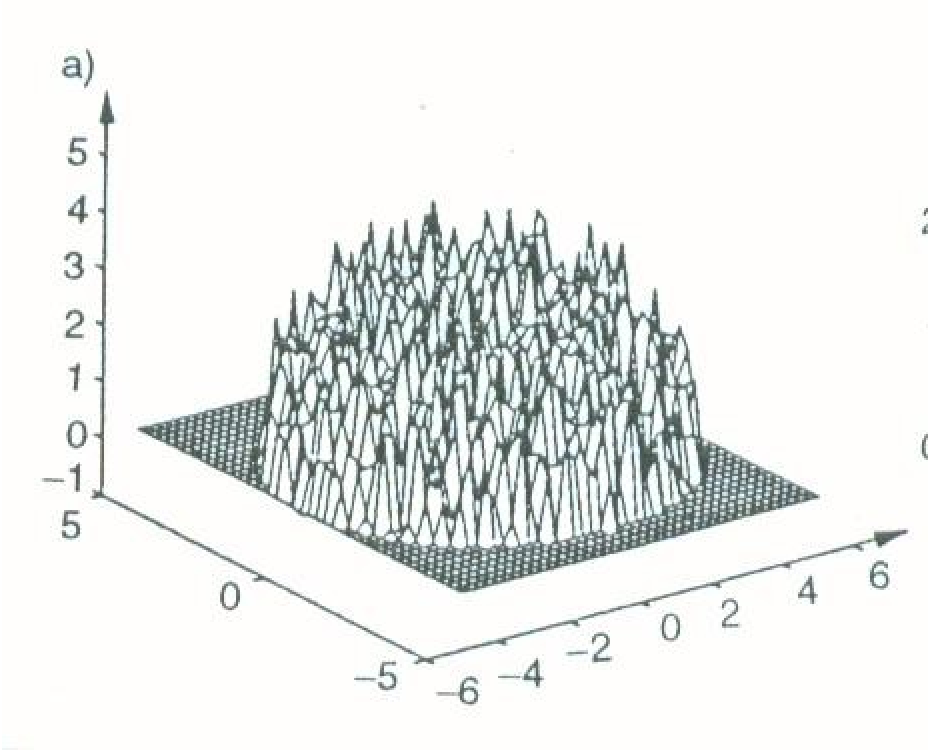
Bazowe funkcje radialne

Rozwiązanie „teoretyczne” problemu przedstawione na poprzedniej stronie nie byłoby właściwe ze względu na nieodpowiednie własności generalizacyjne sieci.

Przy dużej liczbie danych uczących i równej im liczbie funkcji radialnych problem staje się przewymiarowany (źle postawiony), gdyż liczba równań przewyższa ilość stopni swobody procesu fizycznego modelowanego równaniem $\phi \mathbf{W} = \mathbf{d}$.

Rezultatem nadmiarowości wag będzie dopasowanie się modelu do różnego rodzaju szumów lub nieregularności występujących w danych uczących. W efekcie hiperplaszczyzna interpolująca dane uczące będzie niegładka, a zdolności uogólniające sieci niewielkie.

Bazowe funkcje radialne



Odwzorowanie danych przy nadwymiarowej liczbie funkcji bazowych.

Bazowe funkcje radialne

Metody stabilizacji wywodzą się od Tichonowa, polegają na stabilizacji procesu poprzez ***dołączenie do równania podstawowego*** dodatkowych warunków w postaci ***równań więzów***, ograniczających stopnie swobody dobieranych parametrów.

$$L(F) = \frac{1}{2} \sum [F(x_i) - d_i]^2 + \frac{1}{2} \lambda \, || PF||^2$$

gdzie **|| PF||** jest czynnikiem regularyzacyjnym , odpowiednikiem funkcji kary.

Bazowe funkcje radialne

Teoretyczne rozwiązanie problemu regularyzacji uzyskano przy zastosowaniu funkcji Greena, w której c_i jest centrum rozwinięcia, a wagi

$$\mathbf{W_i = [d_i - F(c_i)] / \lambda}$$

reprezentują nie znane współczynniki rozwinięcia.
Funkcje Greena odgrywają rolę funkcji bazowych, a rozwiązanie problemu interpolacji jest wyrażone w standartowej postaci

$$\mathbf{F(x) = \Sigma\, W_i\, G(x,c_i)}$$

Najbardziej popularnym typem radialnej funkcji Greena jest funkcja Gaussa.

$$\mathbf{G(x,x_i) = \exp(\, - \|x-c_i\|/2\sigma^2\,) = \exp\,(\, -1/2\sigma_i^2\, \Sigma\, (x_k - c_{i,k})^2\,)}$$

przy czym $\mathbf{c_i}$ oznacza wektor wartości średnich (centrów),
a $\mathbf{\sigma_i^2}$ **–** wariancje.

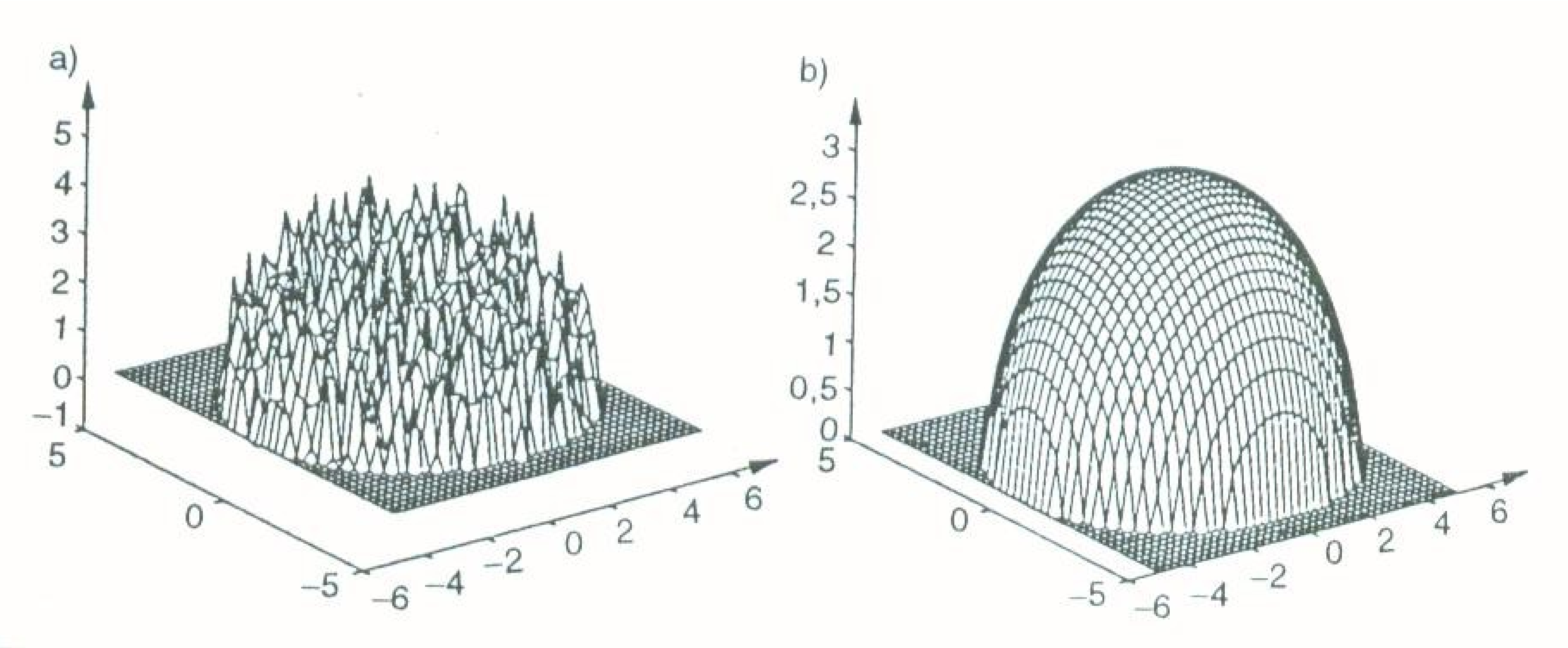
Bazowe funkcje radialne

Przy gaussowskiej postaci funkcji Greena regularyzowalne rozwiązanie problemu aproksymacji za pomocą funkcji radialnych przyjmie postać

$$F(x) = \Sigma\, W_i \exp(- ||x-c_i||^2 / 2\, \sigma_i^2)$$

będący superpozycją wielowymiarowych funkcji Gaussa z centrami ulokowanymi w c_i (przy wartości **p** równej liczbie danych uczących, wektor centrum odpowiada współrzędnym c_i wektora uczącego) i szerokością σ_i.

Bazowe funkcje radialne



Wplyw regularyzacji na odwzorowanie danych przy nadwymiarowej liczbie funkcji bazowych : a) brak regularyzacji ; b) wynik z regularyzacją.

Sieć neuronowa radialna

Zastosowanie w rozwinięciu p funkcji bazowych, przy czym p oznacza liczbę wzorców uczących, jest niedopuszczalne z praktycznego punktu widzenia gdyż liczba tych wzorców jest bardzo duża i złożoność obliczeniowa algorytmu niepomiernie wzrasta.

Rozwiązanie układu równań $p \times p$ przy dużych wartościach p staje się niezwykle trudne, bardzo duże macierze zwykle są bardzo źle uwarunkowane (współczynnik uwarunkowania nawet 10^{20}).

Sieć neuronowa radialna

Podobnie jak w przypadku sieci wielowarstwowych potrzebna jest redukcja wag, tutaj sprowadzająca się do redukcji ilości funkcji bazowych do liczby **K**.

Rozwiązania równania aproksymującego można przedstawić w postaci

$$F^*(x) = \sum_{i=1}^{K} W_i \, G(x, c_i)$$

gdzie $G(x,c_i) = G(||x-c_i||)$, $K<p$, a c_i $(i=1,...,K)$ jest zbiorem centrów które należy wyznaczyć.

W szczególnym przypadku, dla $K=p$ otrzymuje się rozwiązanie dokładne.

Sieć neuronowa radialna

Zadanie aproksymacji polega na dobraniu odpowiednich funkcji Greena **$G(x,c_i)$** i takim doborze wag **W_i (i=1,2,....,K)** , aby rozwiązanie najlepiej przybliżało rozwiązanie dokładne. Rozwiązuje się to przy pomocy ***metody wariacyjnej (Galerkina)*** minimalizując funkcjonał

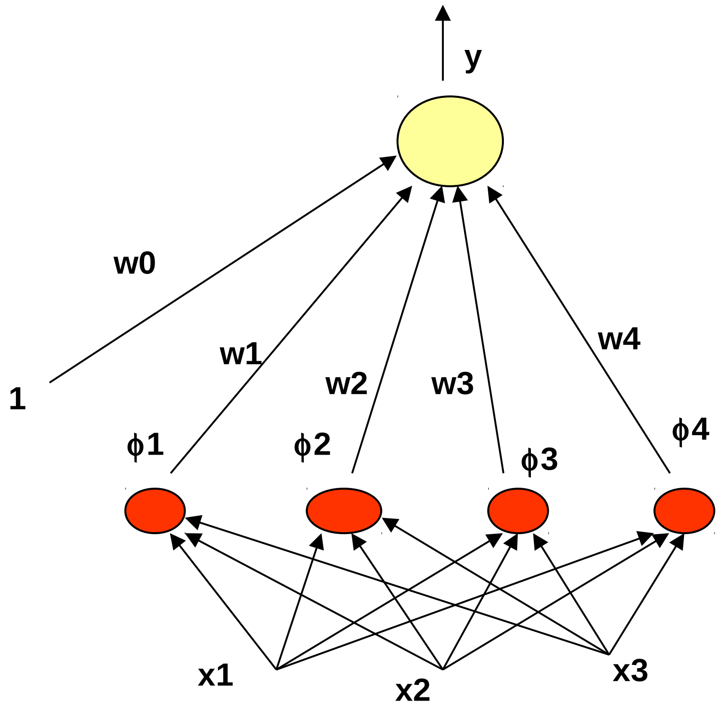
$$L(F^*) = \sum_{i=1}^{p} \left[\sum_{j=1}^{K} W_j \, G(|| x_i - c_j ||) - d_i \right]^2 + \lambda \, || PF^* ||^2$$

Uzyskane rozwiązanie wyrażające funkcję aproksymującą w przestrzeni wielowymiarowej może być zinterpretowane w postaci sieci neuronowej radialnej ***zwanej siecią RBF (Radial Basic Function)***.

Sieć RBF

Sieć RBF

Ma ona strukturę dwuwarstwową, warstwa ukryta realizuje odwzorowanie nieliniowe realizowane przez neurony radialnej funkcji bazowej. Neuron wyjściowy jest liniowy, a jego rolą jest sumowanie wagowe sygnałów pochodzących od neuronów warstwy ukrytej.



Ogólna postać
sieci radialnej RBF

Siec neuronowa radialna

Uzyskana architektura sieci radialnych ma strukturę analogiczną do struktury wielowarstwowej sieci sigmoidalnych o jednej warstwie ukrytej. Rolę neuronów ukrytych odgrywają radialne funkcje bazowe różniące się kształtem od funkcji sigmoidalnych.

Istotne różnice miedzy obydwoma typami sieci:

→ ***sieć radialna*** ma strukturę ustaloną o jednej warstwie ukrytej i liniowym neuronie wyjściowym;

→ ***sieć sigmoidalna*** może mieć różną liczbę warstw, a neurony wyjściowe mogą być liniowe albo nieliniowe;

Sieć neuronowa radialna

W przypadku stosowania funkcji radialnych występuje większe zróżnicowanie w doborze ich kształtu.

Najbardziej popularnym odpowiednikiem funkcji sigmoidalnej jest funkcja bazowa Gaussa określona zależnością

$$\phi(r) = \exp(-r^2/2\sigma^2),$$

przy czym $r = || x-c ||$, a $\sigma>0$ jest parametrem.

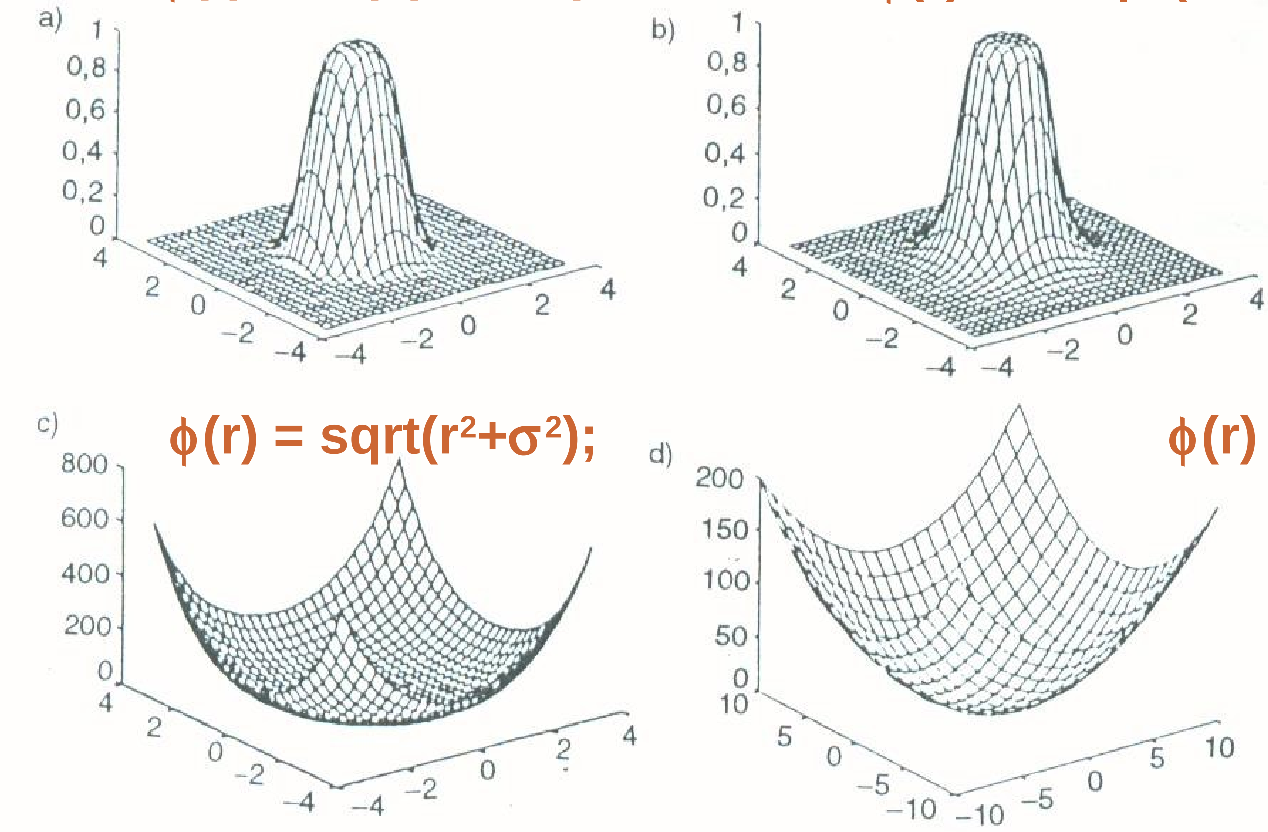
Stosuje się również funkcje potęgowe:

$$\phi(r) = 1/\mathrm{sqrt}(r^2+\sigma^2);$$
$$\phi(r) = \mathrm{sqrt}(r^2+\sigma^2);$$
$$\phi(r) = r^{2n+1};$$

Sieć neuronowa radialna

$\phi(r) = \exp(-r^2/2\sigma^2)$, $\phi(r) = 1/\mathrm{sqrt}(r^2+\sigma^2)$,

$\phi(r) = \mathrm{sqrt}(r^2+\sigma^2)$; $\phi(r) = r^2$



Wykresy funkcji bazowych: a) funkcja Gaussa; b) funkcja Hardy’ego; c) funkcja wielomianowa; d) funkcja potęgowa.

Sieć neuronowa radialna

→ Funkcja nieliniowa radialna dla każdego neuronu ukrytego ma inne parametry c_i oraz σ_i, w sieci sigmoidalnej natomiast stosuje się zwykle standartowe funkcje aktywacji o parametrze β identycznym dla każdego neuronu.

→ Argumentem funkcji radialnej jest odległość danej próbki x od centrum c_i, a w sieci sigmoidalnej jest to iloczyn skalarny wektorów $W^T x$.

Metody uczenia

Problem uczenia sieci przy wybranym typie radialnej funkcji bazowej składa się z dwu etapów:

$\rightarrow$ doboru centrów i parametrów kształtu funkcji bazowych
$\rightarrow$ doboru wag neuronów warstwy wyjściowej

Podstawowa trudność to etap pierwszy.

Stosuje się najczęściej:
- $\rightarrow$ **wybór losowy**,
- $\rightarrow$ **samoorganizujący się proces podziału na klastry**
- $\rightarrow$ **uczenie pod nadzorem**.

Losowy wybór centrów funkcji bazowych

Wybór stałych parametrów funkcji bazowych jest dokonywany

losowo przy rozkładzie równomiernym

Jest to najprostsze rozwiązanie.

Jest to podejście dopuszczalne dla klasycznych sieci radialnych pod warunkiem, że rozkład danych uczących dobrze odzwierciedla specyfikę problemu.

Przy wyborze gaussowskiej funkcji radialnej zakłada się wartość odchylenia standartowego funkcji zależną od rozrzutu dobranych losowo centrów c_i, wobec czego gaussowska funkcja bazowa przyjmuje postać

$$G(|| x - c_i||^2) = \exp (- ||x-c_i||^2 / d^2 / K)$$

dla **$i=1,2, \ldots, K$,** przy czym d oznacza maksymalną odległość miedzy centrami **c_i.** Jak wynika z powyższego wzoru **$\sigma = d/\mathrm{sqrt}(2K)$** i jest jednakowe dla każdej funkcji bazowej, **d –** maksymalna odległość miedzy centrami.

Losowy wybór centrów funkcji bazowych

Po ustaleniu parametrów funkcji bazowych pozostaje problem doboru wag neuronu wyjściowego, sprowadzający się do rozwiązania równania

$$\mathbf{G\ W = d}$$

gdzie

$$\mathbf{G_{ij} = exp(- ||x_i - c_j||^2 / d^2 / K)}$$

gdzie $\mathbf{i=1,...,p}$ oraz $\mathbf{j=1,....,K}$.

Dobor parametrow funkcji radialnych

Znacznie lepsze rezultaty można uzyskać przez zastosowanie

samoorganizujacego się procesu podziału danych uczących

na klastery w jednej z jego licznych odmian.

$\rightarrow$ Centrum klastera jest utożsamiane z centrum odpowiedniej funkcji radialnej.

$\rightarrow$ Liczba tych funkcji jest równa liczbie klasterów i może być korygowana przez algorytm samo organizacji.

$\rightarrow$ Proces podziału danych na klastery może być przeprowadzany metodą K-uśrednień. Aparat matematyczny zaangażowany w tą procedurę jest dość skomplikowany....

Algorytmy uczące oparte na propagacji wstecznej

Odmienną klasą algorytmów uczących dla sieci radialnych są
metody gradientowe realizujące uczenie z nadzorem,
wykorzystujące algorytm propagacji wstecznej.

→ Dla metod gradientowych podstawowa trudność to wyznaczenie składowych gradientu względem wszystkich parametrów.

→ Kolejna trudność to wybór wartości startowych parametrów. Przy starcie uczenia z wartości losowych prawdopodobieństwo utknięcia procesu w minimum lokalnym jest większe niż w przypadku sieci sigmoidalnych, ze względu na silną nieliniowość funkcji wykładniczych.

→ Losowe wartości punktów startowych stosuje się rzadko, zastępując je odpowiednią procedurą initializacji, wykorzystującą informację zawartą w zbiorze danych uczących. Do tego celu wykorzystuje się algorytmy samoorganizacji. Wartości parametrów funkcji radialnych otrzymane w wyniku ich działania przyjmuje się za wartości startowe.

Sieć radialna a sieć sigmoidalna

Sieci neuronowe o radialnych funkcjach bazowych znalazły zastosowanie zarówno w rozwiązywaniu problemów klasyfikacyjnych, zadaniach aproksymacji funkcji wielu zmiennych, jak i zagadnieniach predykcji..... tych obszarach zastosowań gdzie funkcje sigmoidalne mają ugruntowaną pozycję.

W stosunku do sieci wielowarstwowych o sigmoidalnych funkcjach aktywacji wyróżniają się pewnymi właściwościami szczególnymi, umożliwiającymi lepsze odwzorowanie cech charakterystycznych modelowanego procesu.

Przedyskutujmy te różnice....

Sieć radialna a sieć sigmoidalna

Sieć sigmoidalna:

Działanie funkcji rozciąga się od określonego punktu w przestrzeni aż do nieskończoności, reprezentuje aproksymację globalną funkcji zadanej. Nie ma niemożności fizycznego powiązania obszaru aktywności neuronu z odpowiednim obszarem danych uczących, trudności z określeniem optymalnego punktu startowego z procesie uczenia.

Sieć radialna:

Bazuje na funkcjach mających wartość niezerową jedynie w określonej przestrzeni tylko wokół centrów, realizuje aproksymację typu lokalnego, której zasięg działania jest bardzo ograniczony. Można się spodziewać że zdolności do uogólniania są gorsze niż dla sieci sigmoidalnych. Łatwość powiązania parametrów funkcji bazowych z fizycznym rozmieszczeniem danych w obszarze parametrów. Łatwość uzyskania dobrych wartości startowych w procesie uczenia pod nadzorem.

Sieć radialna a sieć sigmoidalna

Sieć radialna:

Przestrzenie decyzyjne tworzone w sieciach radialnych są stosunkowo proste i w sposób naturalny kształtowane. Sieć dostarcza nie tylko informacji do jakiej klasy należy wzorzec testujący, ale wskazuje również na ewentualną możliwość utworzenia oddzielnej klasy.

Na ogół uważa się, że sieci radialne lepiej niż sieci sigmoidalne nadają się do takich żądań klasyfikacyjnych jak wykrywanie uszkodzeń w różnego rodzaju systemach, rozpoznawanie wzorców, itp.

Znaczną zaletą sieci radialnych jest znacznie uproszczony algorytm uczenia. Przy istnieniu tylko jednej warstwy ukrytej i ***ścisłym powiązaniu aktywności neuronu z odpowiednim obszarem przestrzeni danych uczących***, punkt startowy uczenia jest znacznie bliżej rozwiązania optymalnego, niż jest to możliwe w sieciach wielowarstwowych.

Sieć radialna a sieć sigmoidalna

Sieć radialna cont.:

Dodatkowo, **możliwe jest odseparowanie etapu doboru parametrów funkcji bazowych od doboru wartości wag sieci (algorytm hybrydowy)**, co może przyśpieszyć i uprościć proces uczenia. Przy zastosowaniu ortogonalizacji proces optymalnego kształtowania struktury sieci jest stałym fragmentem uczenia, nie wymagającym żadnego dodatkowego wysiłku.

Liczba neuronów ukrytych decyduje w dużym stopniu o dokładności odwzorowania i zdolnościach uogólniania sieci. W przypadku sieci radialnej problem doboru liczby neuronów ukrytych jest o wiele prostszy niż w sieciach sigmoidalnych, ze względu na **lokalny charakter aproksymacji** reprezentowany przez poszczególne funkcje bazowe.

Zestaw pytań do testu

1. Co to znaczy że sieć radialna realizuje aproksymację typu lokalnego?
2. Podaj przykład bazowej funkcji radialnej.
3. Czy w sieci radialnej stosujemy wiele warstw ukrytych?
4. Co to jest siec RBF, narysuj schemat.
5. Wymień etapy procesu uczenia sieci radialnej.