RBF – sieci neuronowe o radialnych funkcjach bazowych

Jerzy Stefanowski

Zakład Inteligentnych Systemów Wspomagania Decyzji
Instytut Informatyki
Politechnika Poznańska

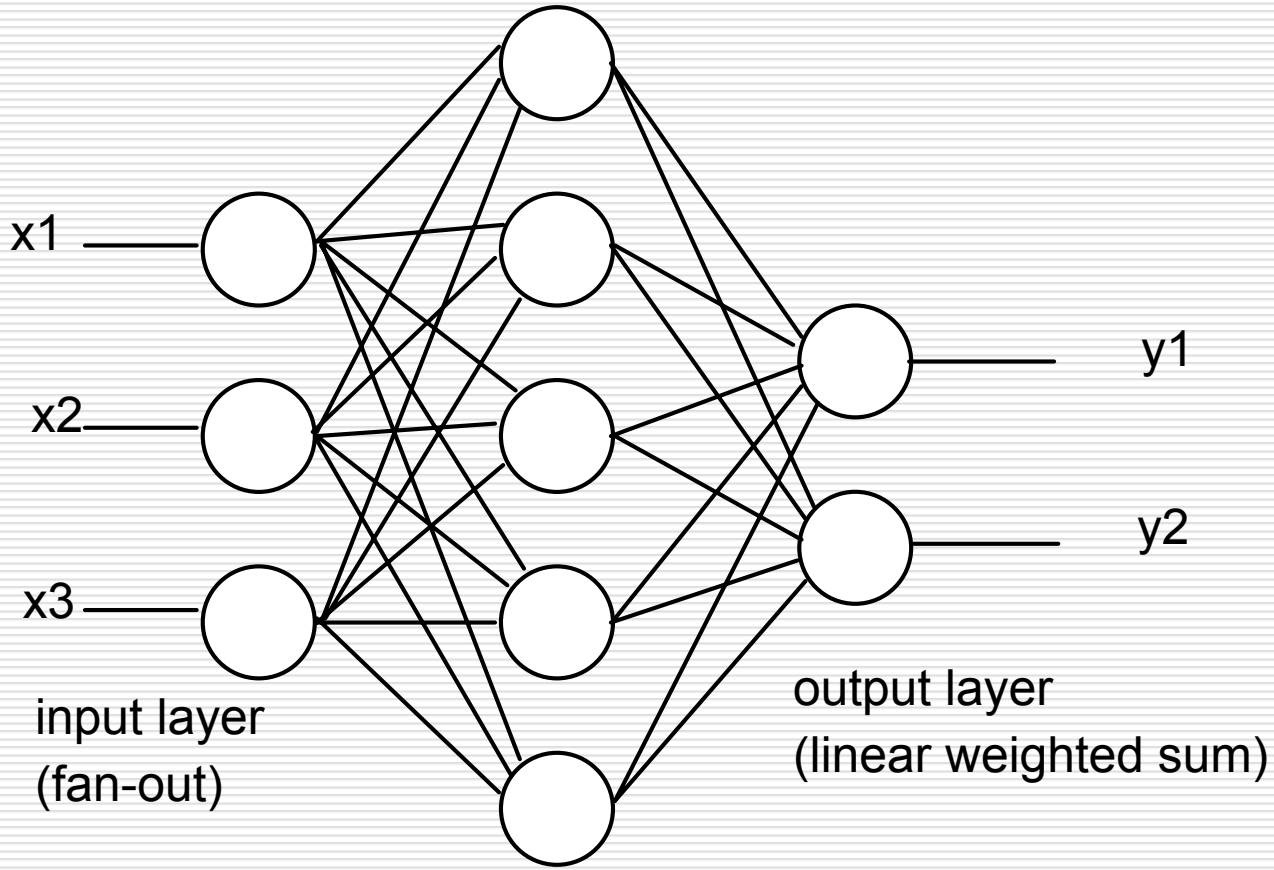
Wprowadzenie

- Idea sieci RBF opiera się na rozwiązaniach statystycznych metod aproksymacji funkcji zmiennej liczbowej
  - RBF – ang. Radial Basis Function
  - Statystyka – aproksymacja nieparametryczna specjalnymi metodami lokalnymi
    - Estymatory jądrowe oparte na funkcjach o symetrii kołowej
- Są także stosowane w zadaniach klasyfikacji, analizie szeregów czasowych, predykcji zmiennej liczbowej i niektórych numerycznych problemach modelowania złożonych systemów
- Alternatywne podejście niż MLP

Różnice MLP i RBF

- Aproksymacja lokalna vs. globalna
  - Twierdzenia matem. → uniwersalny aproksymacja „dowolnej” funkcji ciągłej z wystarczającą dokładnością.
  - Twierdzenie Covera o liniowej separowalności wzorców w przekształconej przestrzeni wysoce wielowymiarowej
- Zdecydowanie prostsza topologia
  - Warstwa ukryta z nieliniowym przekształceniem z funkcjami RBF
  - Warstwa wyjściowa – proste (liniowe) ważone sumowanie
- Prostsze algorytmy uczeni sieci RBF

Typowa topologia RBF



RBF ...

- Sieć RNF oprócz warstwy wejściowej dostarczającej dane x składa się z 2 warstw
- Z warstwy neuronów ukrytych (implementujących funkcje RBF)
- Warstwy wyjściowej (wiele neuronów) implementującej liniowe sumowanie (podobnie jak MLP)
- Uczenie sieci dwu etapowe

Zadanie aproksymacji numerycznej

- Niech będzie dany zbiór N punktów

$$\{\mathbf{x}_i \in R^m \,|\, i = 1,\ldots,N\}$$

- i odpowiadający mu zbiór N liczb rzeczywistych

$$\{d_i \in R \,|\, i = 1,\ldots,N\}$$

- Zadanie aproksymacji polega na znalezieniu funkcji

$$f(\mathbf{x}_i) = d_i \quad \forall i = 1,\ldots,N$$

Statystyka – regresja nieparametryczna

- Statystka – modele liniowe i zaawansowane (rozwijane od prawie 200 lat)
- Porównanie pojęć (Mark Orr: Introd. To RBF)

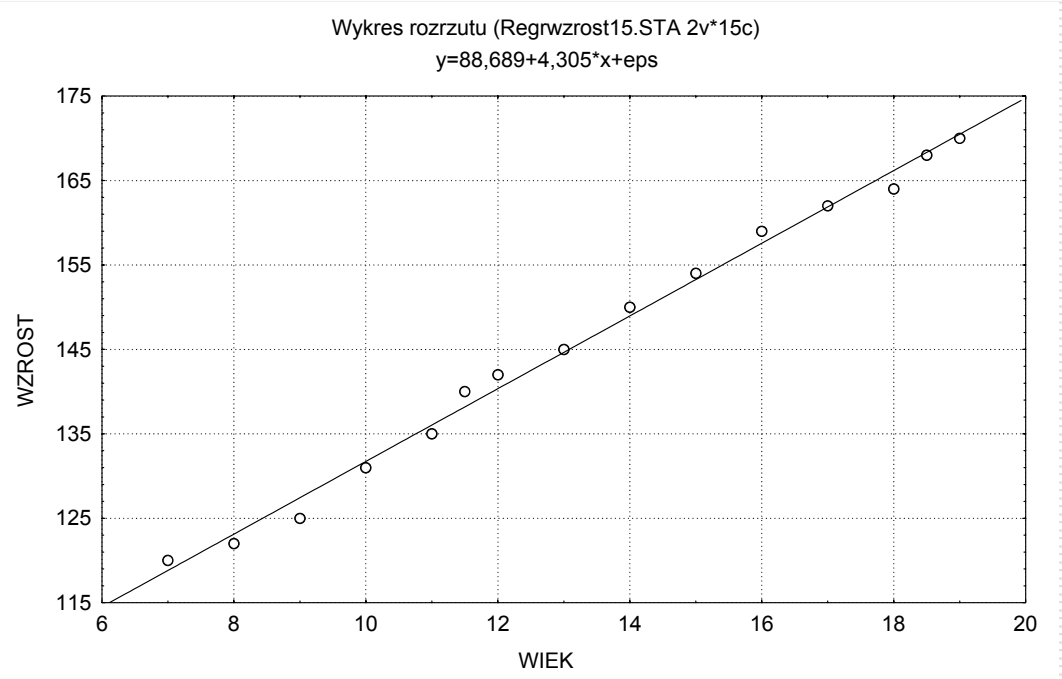
statistics	*neural networks*
model	network
estimation	learning
regression	supervised learning
interpolation	generalisation
observations	training set
parameters	(synaptic) weights
independent variables	inputs
dependent variables	outputs
ridge regression	weight decay

Table 1: Equivalent terms in statistics and neural networks.

Klasyczne modele liniowe

- Aproksymacja funkcji
- Regresja parametryczna MNK

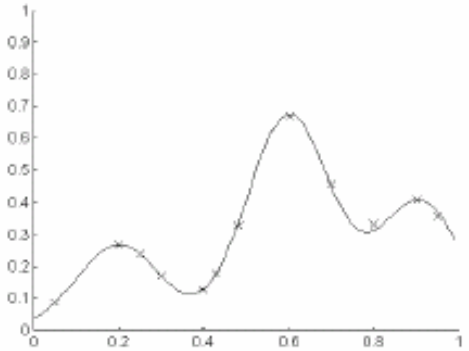
Dane pomiarowe:
(7,120),(9,125),(18,16),
(11.5,140),
(8,122),(11,135),
(13,145), (17,162),
(10,131), (19,170),
(14,150), (12,142),
(18.5,168), (15,154),
(16,159)

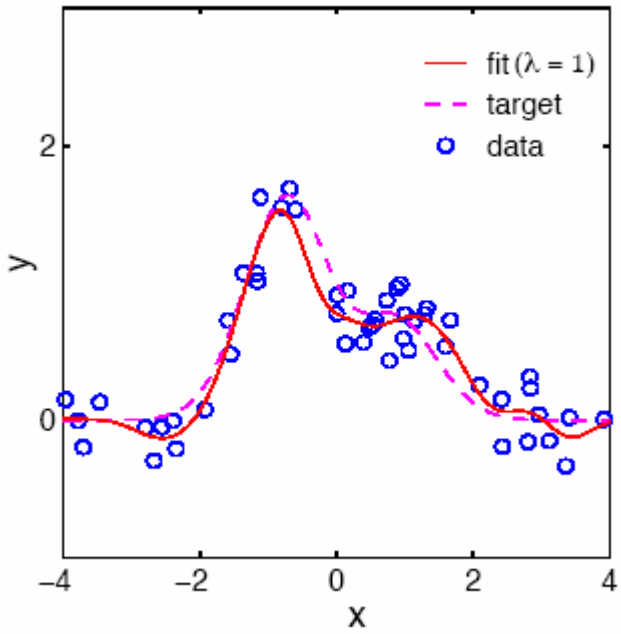


Różne funkcje

- Trudniejsze funkcje – regresja nieparametryczna

data no	1	2	3	4	5	6	7	9	10	11	12	13
x	0.0500	0.2000	0.2500	0.3000	0.4000	0.4300	0.4800	0.6000	0.7000	0.8000	0.9000	0.9500
$f(x)$	0.0863	0.2662	0.2362	0.1687	0.1260	0.1756	0.3290	0.6694	0.4573	0.3320	0.4063	0.3535





Aproksymacje lokalne – regresja nieparametryczna

- Więcej J.Koronacki, J.Ćwik: Statystyczne systemy uczące się.
- Estymując funkcję regresji staramy się uwzględnić w modelu własności lokalne
- Składanie funkcji bazowych zdolnych lokalnie przybliżyć własności pewnych podobszarów dziedziny
- Regresyjne funkcje sklejane z węzłami

Locally weighted regression

$$y = \alpha + \sum_{j=1}^{p} f_j(\mathbf{x}, \beta)$$

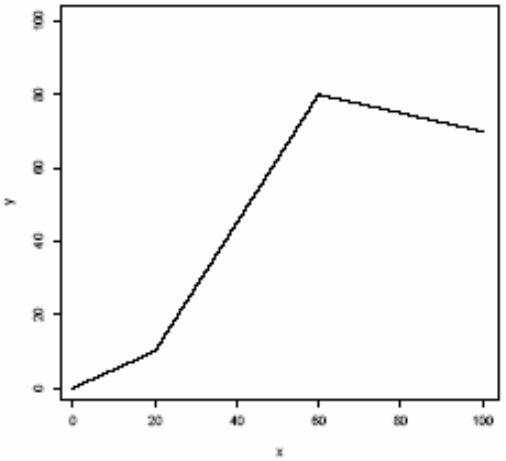


Figure 1.1. An Illustration of Linear Regression Splines with Two Knots

Aproksymacja radialnymi funkcjami kołowymi

- Zadanie aproksymacji $f(\mathbf{x}_i) = d_i \quad \forall i = 1,\ldots,N$
- Przyjmijmy funkcje liniową względem parametrów w wykorzystującą funkcje o symetrii kołowej RBF

$$f(\mathbf{x}) = \sum_{i=1}^{m} w_i \cdot \varphi(\|\mathbf{x} - c_i\|)$$

- Radialna funkcja bazowa $\rightarrow$ funkcja φ o postaci $\varphi(\mathbf{x},\mathbf{c}) = \varphi(r(\mathbf{x},\mathbf{c}))$, gdzie r jest odległością między punktami **x** i **c.** Punkt c nazywamy centrum
- Związek funkcji radialnym z funkcjami jądrowymi (kernals), z parametrem σ szerokością jądra

Własności radialnych funkcji bazowych

- Dyskusja w Koronacki,Ćwik rozdz. 5
- Funkcja φ(r(**x**,**c**)) jak typowa funkcja jądrowa powinna być symetryczna i może mieć maksimum ok. **c**. Przykład typowej funkcji

$$r(u) = \exp(-\frac{u^2}{2}) \quad -\infty < u < \infty$$

- Jeżeli dodatkowo jądro φ jest całkowalne na prostej i całka nie równa się zeru, to rodzina funkcji f(x) jest „gęsta” w przestrzeniu Lq funkcji rzeczywistych Rm całkowalnych w q-tej potędzie.
- Funkcje ciągłe w tej przestrzeni mogą być przybliżone funkcjami radialnymi – choć liczba składników w funkcji może być duża.

Dokładne podejście analityczne do ustalenia wag wi

- Równanie funkcji $f(\mathbf{x})$ z RBF φ dla kolejnych i można przekształcić do postaci macierzowej

$$\begin{bmatrix} \varphi_{11} & \varphi_{12} & \cdots & \varphi_{1N} \\ \varphi_{21} & \varphi_{22} & \cdots & \varphi_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ \varphi_{N1} & \varphi_{N2} & \cdots & \varphi_{NN} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \cdots \\ w_N \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \cdots \\ d_N \end{bmatrix}$$

Jeden „węzeł” sieci na jeden wektor treningowy x, bez ”regularyzacji” (centrum na wektorze)

Dla odpowiednio wąskich funkcji φ zastosujmy standardowe przekształcenie

Wymaganie wobec macierzy Φ

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

Lecz to rozwiązanie nie jest najlepsze

Mocno podane na przeuczenie i o słabych własnościach generalizujących

Niedogodne / nie efektywne obliczeniowo

Filozofia RBF

MLP - dyskryminacja, LDA, aproksymacja stochastyczna.

RBF = Radial Basis Functions (1988) - inne podejście.

Uczenie jako problem aproksymacji, najlepszego dopasowania (rekonstrukcji) hiperpowierzchni do danych treningowych.

- Twierdzenie (Cover 1965):

Jeśli przekształcić wzorce $\mathbf{X}=\{X^{(i)}\}$, i=1.. p, nieliniową funkcją na wektory $\Phi(X^{(i)})=\{h(X^{(i)})_k\}$, $k = 1..M$, $M > p$
wzorce prawdopodobnie staną się liniowo separowalne: tj. istnieje płaszczyzna

$\mathbf{W}^T \Phi(X^{(i)}) \geq 0$ dla $X^{(i)} \in C_1$,
$\mathbf{W}^T \Phi(X^{(i)}) < 0$ dla $X^{(i)} \in C_2$

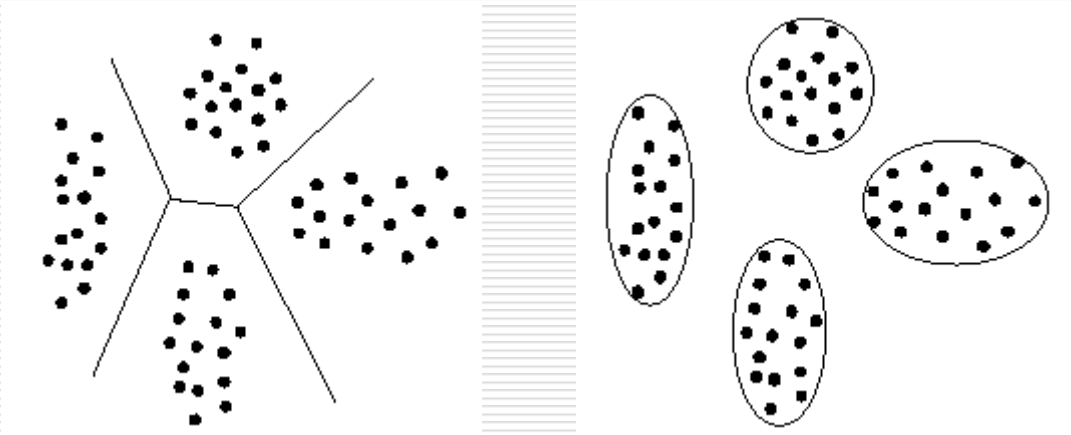
Separowalność wielomianowa

Jeśli wziąć funkcje wielomianowe:

$$\Phi_A(X) = \sum_{0 \le i_1 i_2 \ldots i_r \le p} A_{i_1 i_2 \ldots i_r} X_{i_1} X_{i_1} \cdots X_{ir}$$

to zamiast sep. liniowej mamy sep. wielomianową.

Functional Link Networks (Pao), SVM i Kernel Methods - optymalizacja nieliniowego przekształcenia



Funkcje radialne

Przykłady:

$$h(r) = r = \|X - X_i\|$$

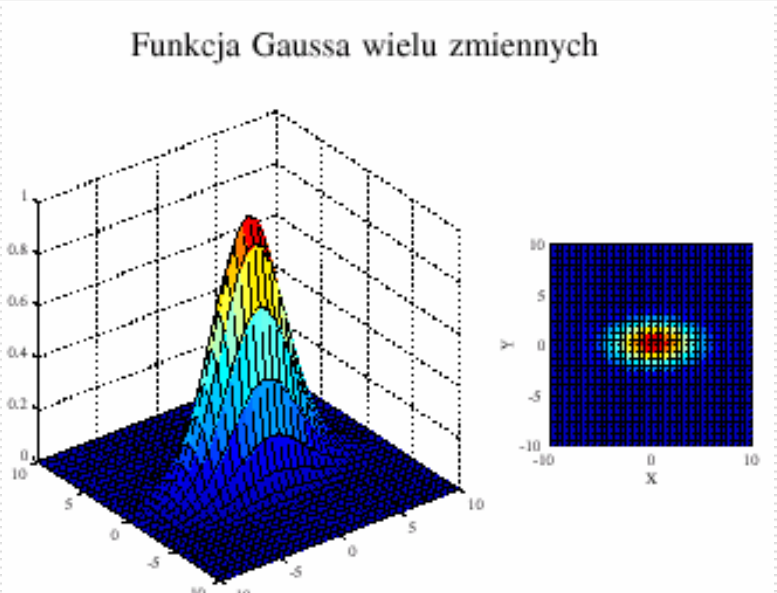
$$h(r) = \left(\sigma^2 + r^2\right)^{-\alpha}, \quad \alpha > 0$$

$$h(r) = \left(\sigma^2 + r^2\right)^{\beta}, \quad 1 > \beta > 0$$

$$h(r) = e^{-(r/\sigma)^2}$$

$$h(r) = (\sigma r)^2 \ln(\sigma r)$$

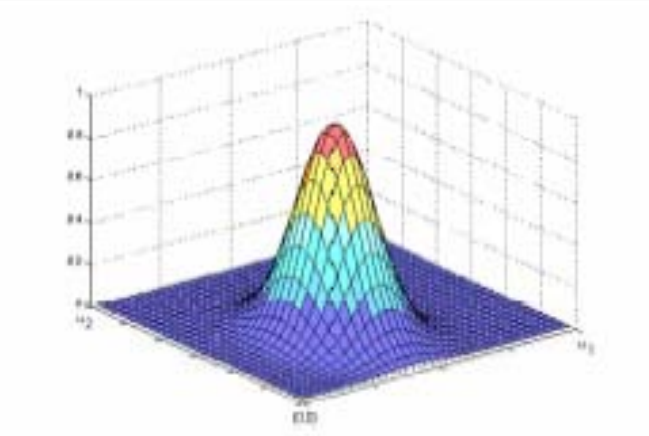
Radialna
Inverse multiquadratic
Multiquadratic
Gauss
Thin splines (cienkiej płytki)



Funkcja Gaussa

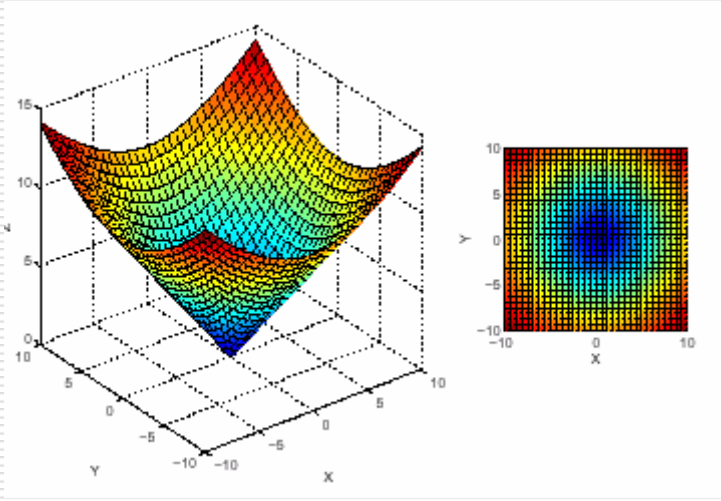
Jedyna lokalna i separowalna f. radialna

$$h(r) = e^{-r^2/2\sigma^2} = e^{-\frac{1}{2}\left(X - X^{(i)}\right)^T \Sigma^{-1} \left(X - X^{(i)}\right)}$$



Funkcja współrzędnej radialnej

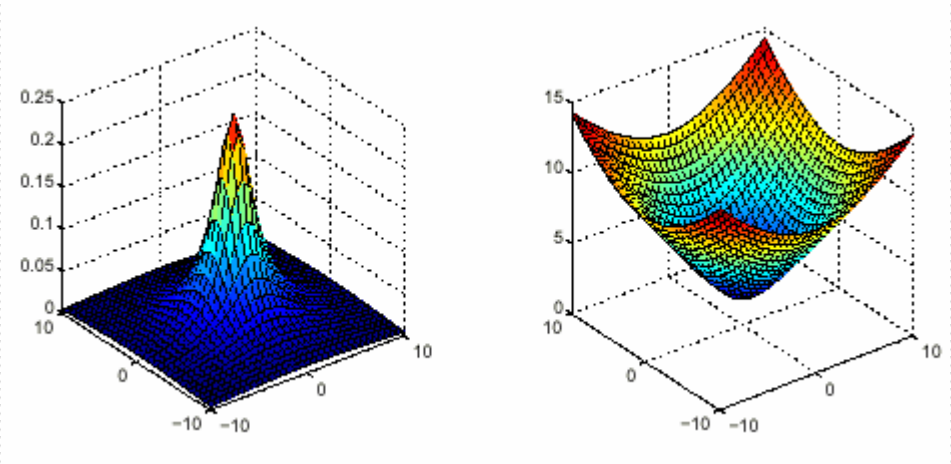
$$h_i(r) = r = \left\| X - X^{(i)} \right\|$$



Funkcje wielokwadratowe

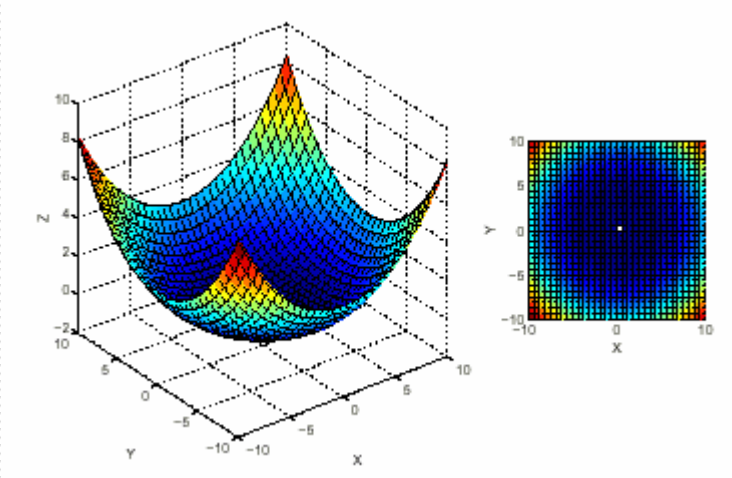
$$h(r) = \left(\sigma^2 + r^2\right)^{-\alpha}, \quad \alpha = 1;$$

$$h(r) = \left(\sigma^2 + r^2\right)^{\beta}, \quad \beta = 1/2$$



Funkcje cienkiej płytki

$$h(r) = (\sigma r)^2 \ln(\sigma r)$$



Budowanie sieci RBF

- Powrót do zadań konstruowania sieci RBF
- Zadanie dokładnego rozwiązanie – odrzucamy!
- Możliwości
  - Liczba *K* funkcji bazowych (neuronów ukrytych) powinna być dużo mniejsza niż n przykładów
  - Centra funkcji bazowych **c** nie opieramy na danych treningowych – ustalane w trakcie uczenia
  - Funkcje bazowe nie muszą mieć takiej samej szerokości $\sigma \rightarrow$ podobnie jak poprzednio dobieramy przez algorytm uczący
  - Można wprowadzić dodatkowy element / próg

Ulepszona konstrukcja sieci RBF

- Funkcja

$$f(\mathbf{x}) = w_0 + \sum_{i=1}^{K} w_i \cdot \varphi(\|\mathbf{x} - c_i\|)$$

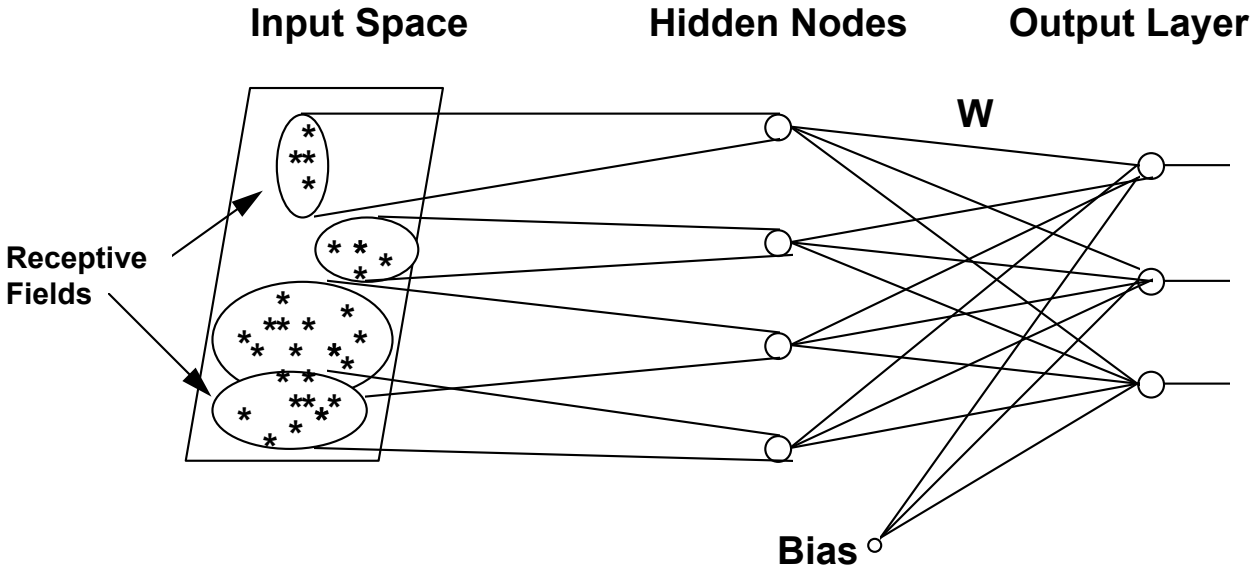
- Przykładowo

$$\varphi(\mathbf{x};\mathbf{c},\sigma) = \exp(\frac{\|\mathbf{x}-\mathbf{c}\|^2}{2\sigma^2})$$

- Jak konstruować sieć RBF (ile neuronów ukrytych)
- Jak dobierać parametry **w**, **c**, σ → algorytmy uczenia sieci

Działanie sieci RBF

- **x** w różnym stopniu pobudza lokalne funkcje



Uczenie sieci RBF

- Wyznaczenie centrów i parametrów funkcji φ
- Uczenie neuronów wyjściowych
  - Jedno wyjście $\rightarrow$ wyznaczenie wektora wag **w**

Wyznaczenia centrów

- *Randomly Fixed (Lowe, 1989):* Centra wybierane z danych treningowych w reprezentatywny sposób
- *Using K-Means Cluster Centers (MacQueen, 1967, Moody and Darken, 1989): k* RBF
- Samoorganizacja – sieci Kohonena

Wybór szerokości σ funkcji bazowej

- *Fixed:* (Haykin, 1994), where *d* is the maximum distance between the chosen centers and *M* is the number of centers (RBF's).

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

- *Distance Averaging (Moody and Darken, 1989):* a "reasonable" estimate for the global width parameter is the average , which represents a global average over all Euclidean distances between the center of each unit *i* and that of its nearest neighbor *j*.

$$\sigma_j = \left\langle \left\| \mu_i - \mu_j \right\| \right\rangle$$

$$\sigma_j = \alpha \left\| \mu_i - \mu_j \right\|$$

- Inne

Uczenie sieci z jednym wyjściem

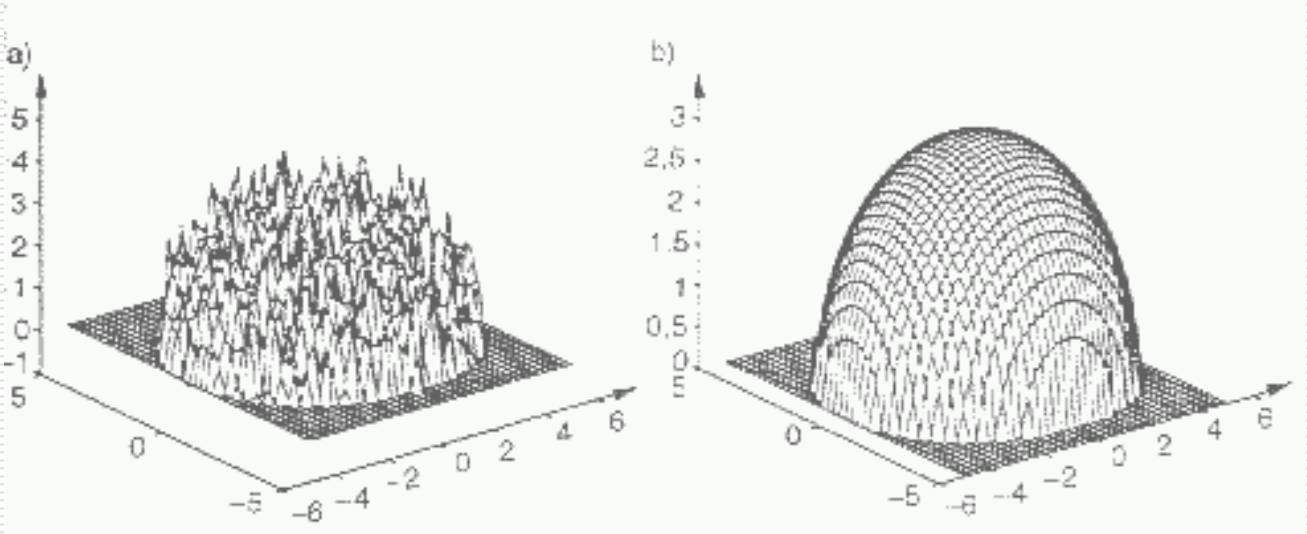
- Dany jest zbiór uczący $\{\mathbf{x}_i, t_i\} | i = 1, \ldots, N$
- Dla każdego wzorca warstwa ukryta dostarcza $\varphi_1(\mathbf{x}), \ldots, \varphi_K(\mathbf{x})$
- Z których obliczamy odpowiedź $y(\mathbf{x}) = w_1 \cdot \varphi_1(\mathbf{x}) + \ldots + w_2 \cdot \varphi_K(\mathbf{x}) + w_0$
- Porównując odpowiedź z wartością docelową obliczamy ogólny błąd odpowiedzi $E = \frac{1}{2} \sum_{i=1}^{N} (t_i - f(\mathbf{x}_i))^2$
- Wektor wag minimalizujący E poszukuje się specjalnymi metodami rozwiązującymi układ równań liniowych (pseudoinwersja, specjalne wersje SVD)

Regularyzacja w sieciach RBF

- Nadmiar neuronów ukrytych $\rightarrow$ przeuczenia i słaba generalizacja + niestabilność numeryczna funkcji aproksymacyjnej f
- Trick ze statystyki $\rightarrow$ regularyzacja
- Nałożenie na funkcje kosztu/błędu tzw. kary powodującej wygładzenie funkcji f

$$E = \frac{1}{2}\sum_{i=1}^{N}(t_i - f(\mathbf{x}_i))^2 + \frac{1}{2}\lambda\sum_K \int \left|Py_k(x)\right|^2 dx$$

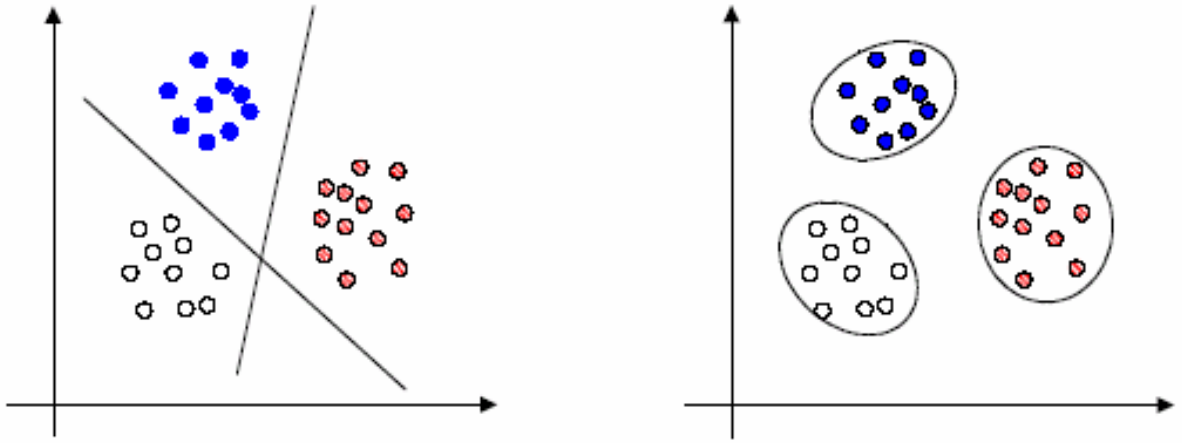
Wpływ regularyzacji



Duża liczba f. bazowych o małej dyspersji bez regularyzacji i po regularyzacji (Ossowski 1996)

RBF w zadaniach klasyfikacje

- Globalne działanie MLP (hiperpłaszczyzny separujące) vs. lokalne sąsiedztwa RBF



An MLP would naturally separate the classes with hyper-planes in the input space (as on the left). An alternative approach would be to model the separate class distributions by localised radial basis functions (as on the right).

Inaczej o różnicach z MLP

- MLP vs. RBF