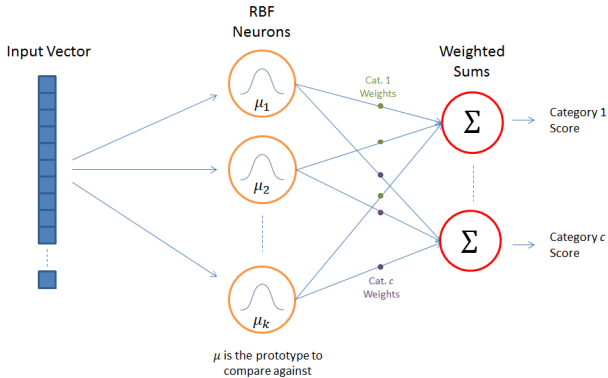


RBF Radial Basis Functions

Sieci neuronowe z radialnymi
funkcjami bazowymi

Radial Basis Function Network



Rys: Chris McCormick, Radial Basis Function Network (RBFN) Tutorial

Sieć RBF

Funkcja realizowana przez sieć RBF z jednym wyjściem:

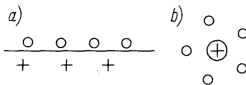
$$f(\mathbf{x}) = \sum_{i=1}^K w_i \varphi(\|\mathbf{x} - \mathbf{c}_i\|) + w_0$$

gdzie K - liczba funkcji bazowych z centrami w $\mathbf{c}_i$

- jedna warstwa ukryta, neurony z radialnymi funkcjami
- funkcja bazowa (radialna) φ określa podobieństwo przypadku (prototypy)
- wagi połączeń neuronów ukrytych określają centrum $\mathbf{c}_i$ funkcji radialnej
- wyjścia liniowe - ważona suma pobudzeń neuronów radialnych

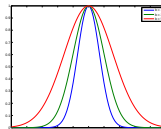
Funkcje radialne

- funkcje radialne mapują relacje lokalne, gdzie perceptrony (np. neurony sigmoidalne) mapują relacje globalne



- argumentem funkcji radialnej jest odległość od pewnego centrum (od prototypu) $\|\mathbf{x} - \mathbf{c}\|$. Zazwyczaj stosuje się metrykę euklidesową.
- funkcja radialna $\varphi(\|\mathbf{x} - \mathbf{c}\|)$ zanika w obszarach dalekich od centrum $\mathbf{c}$
- Przykład: funkcja Gaussa

$$\varphi(x) = e^{-\frac{x^2}{\sigma}}$$



- współczynnik dyspersji σ decyduje o zasięgu funkcji bazowej

Przykłady funkcji radialnych

Niech $r = \|\mathbf{x} - \mathbf{c}_i\|$

$$h(r) = r$$

liniowa

$$h(r) = e^{-\left(\frac{r}{\sigma}\right)^2}$$

gaussowska

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

multiquadratic (wielokwadratowa)

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

inverse multiquadratic

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

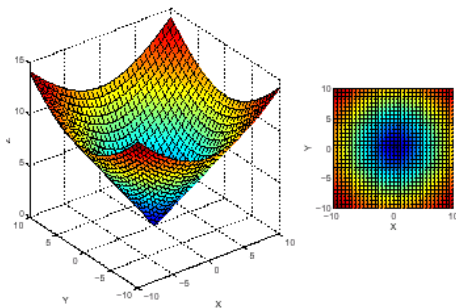
thin-plate spline (cienkiej płytki)

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

wielomianowa

Funkcja liniowa współrzędnej radialnej

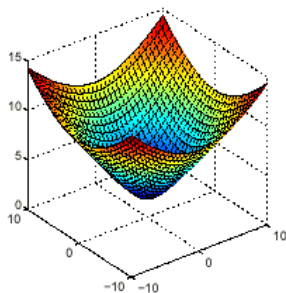
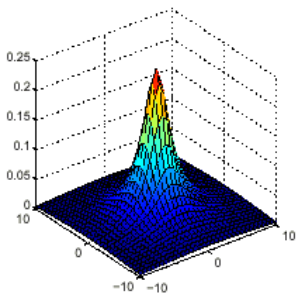
$$h_i(r) = r = \|\mathbf{x} - \mathbf{c}_i\|$$



Funkcje wielokwadratowe

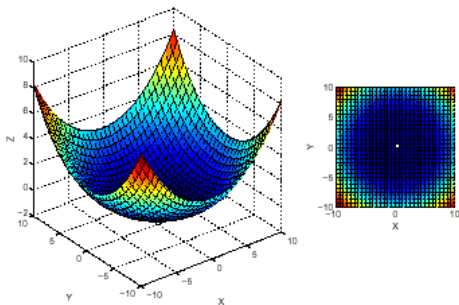
$$h(r) = (\sigma^2 + r^2)^{-\alpha}, \quad \alpha = 1$$

$$h(r) = (\sigma^2 + r^2)^{\beta}, \quad \beta = 0.5$$



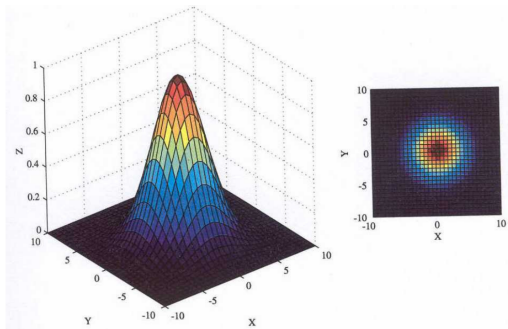
Funkcja cienkiej płytki

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



Funkcja Gaussa

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

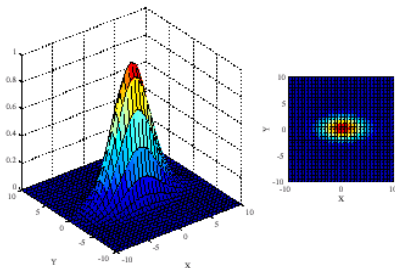


Najczęściej używana w roli funkcji bazowej RBF

Funkcja Gaussa i z pełną macierzą rozmycia

$$h(r) = e^{-\frac{r^2}{2\sigma^2}} = e^{-\frac{1}{2}(\mathbf{x}-\mathbf{x}_i)^T \Sigma^{-1}(\mathbf{x}-\mathbf{x}_i)}$$

Funkcja Gaussa wielu zmiennych



Macierz Σ określa rozmycia i rotacje funkcji

Uczenie RBF jako problem aproksymacji

Dla N punktów $\mathbf{x}_i, y_i$ znajdź funkcję spełniającą

$$f(\mathbf{x}_i) = y_i \quad i = 1, \dots, N$$

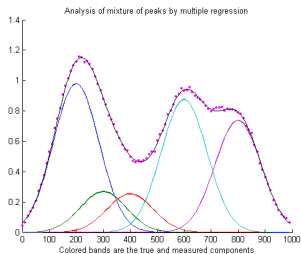
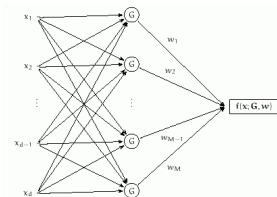
funkcja RBF

$$f(\mathbf{x}) = \sum_{i=1}^K w_i \varphi(\|\mathbf{x} - \mathbf{x}_i\|)$$

funkcja błędu MSE

$$E = \sum_{i=1}^N (y_i - f(\mathbf{x}_i))^2$$

RBF z jedną warstwą ukrytą i linowym wyjściem jest uniwersalnym aproksymatorem



Rys: <https://terpconnect.umd.edu/~toh/spectrum/CurveFittingB.html>

Rozwiązanie RBF

Dla $K = N$, funkcje bazowe o ustalonych centrach w punktach treningowych $\mathbf{x}_i = \mathbf{c}_i$

$$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,N} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ h_{N,1} & h_{N,2} & \cdots & h_{N,N} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

$$h_{ij} = \varphi(\|\mathbf{x}_i - \mathbf{x}_j\|)$$

$$\mathbf{H}\mathbf{w} = \mathbf{y}$$

dla $\mathbf{x}_1 \neq \mathbf{x}_2 \neq \dots \neq \mathbf{x}_N$ macierz $\mathbf{H}$ jest nieosobliwa i dodatnio określona, więc istnieje

$$\mathbf{w} = \mathbf{H}^{-1}\mathbf{y}$$

RBF jako problem aproksymacji

- dla radialnych f. bazowych macierz interpolacji **H** dodatnio określona (Light 1992)
- dla wąskich f. Gaussowskich istnieje idealne rozwiązanie, ale zła generalizacja

$$\mathbf{w} = \mathbf{y}$$

- problem źle określony, przewymiarowanie,
- hiperpłaszczyzna interpolacji nie jest gładka
- większe dyspersje σ i mniej funkcji ($K < N$)
→ lepsza generalizacja

Aproksymacja RBF

Dla ustalonych $\mathbf{c}_i$ oraz σ_i gdy $K < N$

$$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,K} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ h_{N,1} & h_{N,2} & \cdots & h_{N,K} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_K \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_K \end{pmatrix}$$

$$h_{ij} = \varphi(\|\mathbf{x}_i - \mathbf{c}_j\|)$$

$$\mathbf{H}\mathbf{w} = \mathbf{y}$$

rozwiązanie układu równań przez pseudoinwersję

$$\mathbf{w} = \mathbf{H}^+ \mathbf{y}, \quad \text{gdzie} \quad \mathbf{H}^+ = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T$$

W praktyce pseudoinwersję realizuje się za pomocą dekompozycji SVD

Rozwiązanie RBF

$$\mathbf{w} = \mathbf{H}^+ \mathbf{y}$$

dokonując rozkładu

$$\mathbf{H}^+ = \mathbf{USV}^T$$

gdzie $\mathbf{U}$ i $\mathbf{V}$ są ortogonalne o wymiarach $N \times N$ oraz $K \times K$
Macierz $\mathbf{S}$ o rozmiarach $N \times K$ jest pseudodiagonalna, gdzie

$$s_1 \geq s_2 \geq \dots \geq s_k \geq 0$$

Redukując liczbę kolumn do r naistotniejszych wartości

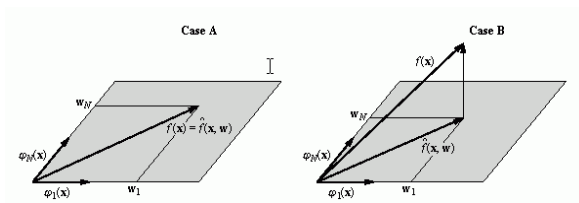
$$\mathbf{w} \approx \mathbf{V}_r \mathbf{S}_r^{-1} \mathbf{U}_r^T \mathbf{y}, \quad \text{gdzie} \quad s_r^{-1} = \left[\frac{1}{s_1}, \frac{1}{s_2}, \dots, \frac{1}{s_r} \right]$$

- dobór wag w jednym kroku dla ustalonych funkcji radialnych
- macierze $\mathbf{V}_r$ i $\mathbf{U}_r$ są ortogonalne, więc problem jest dobrze uwarunkowany
- głównym problemem jest dobór centrów i rozmyć funkcji φ

Interpretacja geometryczna

Jeśli prawdziwa aproksymowana funkcja $f(\mathbf{x})$ leży w przestrzeni rozpiętej przez wektory bazowe $\Phi(\mathbf{x})$ to możliwe jest rozwiązanie bez błędu, w przeciwnym razie aproksymowana jest projekcja ortogonalna (błąd jest ortogonalny do p-ni bazowej).

$$\hat{f}(\mathbf{x}; \mathbf{w}) = \sum_i w_i \varphi_i(\mathbf{x})$$



Rozwiązanie z regularyzacją

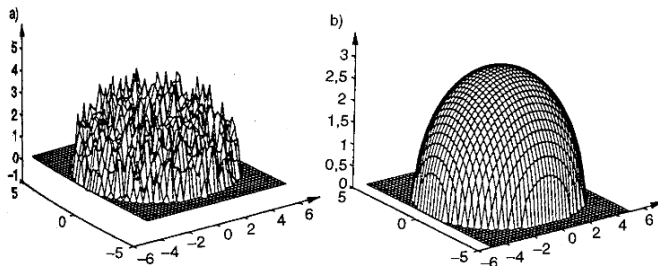
Minimalizacja funkcji błędu z członem regularyzacyjnym

$$E = \sum_{i=1}^N (y_i - f(\mathbf{x}_i))^2 + \lambda \|\hat{P}f\|^2$$

gdzie operator $\hat{P}$ wymusza np. gładkość funkcji aproksymującej

$$\|\hat{P}F_{\mathbf{w}}\|^2 = \int_{\mathbf{R}^N} \left\| \frac{\partial^2 F_{\mathbf{w}}(\mathbf{x})}{\partial x_i \partial x_j} \right\| d\mathbf{x}$$

Wpływ regularyzacji



Nadmiarowa liczba funkcji bazowych o małej dyspersji bez regularyzacji i po regularyzacji (Ossowski 1996)

Separacja w wielowymiarowej przestrzeni

Twierdzenie (Cover 1965):

Jeśli przekształcić wzorce $\mathbb{X} = \{\mathbf{x}_i\}, i = 1, \dots, N$, nieliniową funkcją na wektory $\Phi(\mathbf{x}_i) = [\varphi_1(\mathbf{x}_i), \varphi_2(\mathbf{x}_i), \dots, \varphi_K(\mathbf{x}_i)]^T$ gdzie $K > N$ to rośnie prawdopodobieństwo liniowej separacji, tj. istnieje płaszczyzna

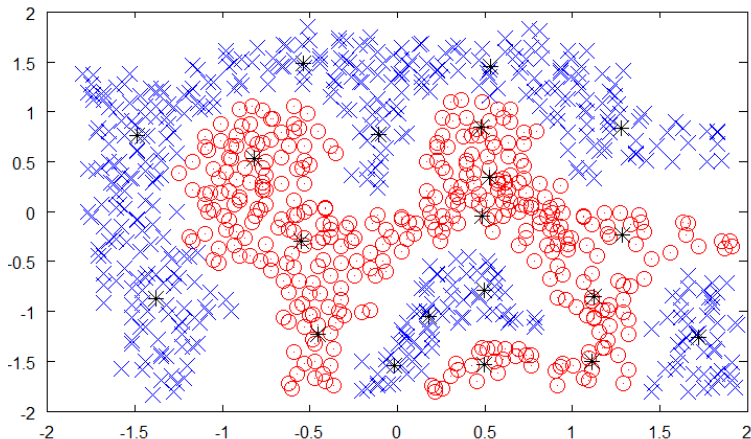
$$\begin{aligned} \mathbf{w}^T \Phi(\mathbf{x}_i) &\geq 0 \quad \text{dla} \quad \mathbf{x}_i \in \mathbb{C}_1 \\ \mathbf{w}^T \Phi(\mathbf{x}_i) &< 0 \quad \text{dla} \quad \mathbf{x}_i \in \mathbb{C}_2 \end{aligned}$$

granica decyzyjna zdefiniowana jest równaniem

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

Przy dostatecznie dużej wymiarowości przestrzeni do której rzutujemy wzorce prawdopodobieństwo separowalności liniowej rośnie do 1.

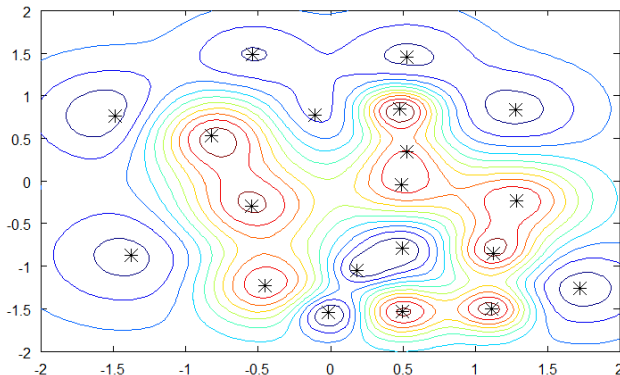
Przykład: 2 klasy



Rys: Chris McCormick, Radial Basis Function Network (RBFN) Tutorial

Przykład: 2 klasy

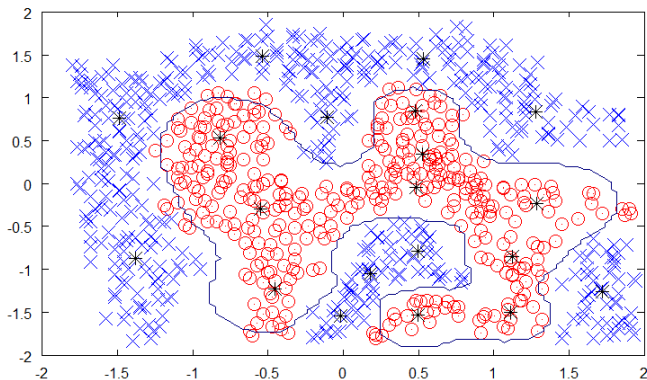
Wartości dla klasy 1, najczęściej wagi powiązane z funkcjami radialnymi dla klasy 1 mają wartości dodatnie a dla klasy 2 - ujemne



Rys: Chris McCormick, Radial Basis Function Network (RBFN) Tutorial

Przykład: 2 klasy

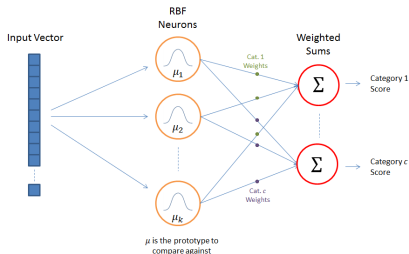
Granica decyzyjna



Rys: Chris McCormick, Radial Basis Function Network (RBFN) Tutorial

Sieć RBF

- jedna warstwa ukryta z radialnymi funkcjami + wyjścia liniowe
- parametry: wagi $\mathbf{w}$, centra $\mathbf{c}_i$, dyspersja σ_i (lub pełna macierz Σ), liczba funkcji radialnych K



Rys: Chris McCormick, Radial Basis Function Network (RBFN) Tutorial

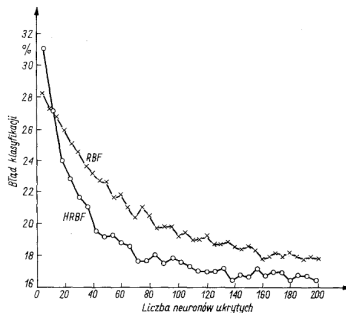
Sieć RBF

- Sieć **GRBF** (Generalized RBF) gdy $K < N$ liczba funkcji radialnych jest mniejsza niż liczba wzorców 1
 - liczba funkcji bazowych kluczowa, powinna być mniejsza od liczby wzorców
- Sieć **HRBF** (Hyper RBF) używa normy ważonej $\|\mathbf{x}\|_Q^2$ z pełną macierzą obrotów i skalowania $\mathbf{Q}$

$$\left\| \mathbf{x} - \mathbf{c}_i \right\|_Q^2 = (\mathbf{x} - \mathbf{c}_i)^T \mathbf{Q}^T \mathbf{Q} (\mathbf{x} - \mathbf{c}_i)$$

posiada o wiele więcej parametrów do uczenia ale zwykle wymaga mniejszej liczby funkcji bazowych.

Dobór liczby parametrów sieci RBF



HRBF uzyskuje podobną poprawność do RBF przy mniejszej liczbie neuronów

Trening RBF

- Parametry nieliniowe funkcji transferu: centra, dyspersje; + wagi.
- Trening 2 etapowy:
 - najpierw ustawienie funkcji bazowych
 - potem rozwiązywanie układu równań (pseudoinwersja)
- Możliwe jednoczesne uczenie centrów, dyspersji i wag, np. za pomocą metod gradientowych (wsteczna propagacja)
- Inicjalizacja początkowych centrów: losowa, klasteryzacja, samoorganizacja, probabilistyczna
- Inicjalizacja dyspersji: wartość stała, zależna od gęstości danych w różnych rejonach, średnie odległości od wektorów z innych klas

Ustalenie centrów

- losowe równomierne pokrycie przestrzeni wejściowej - może nie oddawać specyfiki problemu
- w problemie aproksymacji ustawiamy centra w miejscach minimum i maksimum a następnie usuwamy wzorce z otoczenia tych centrów a resztę rozmieszczemy równomiernie wśród pozostałych wzorców
- w pobliżu granic decyzyjnych (bliskie przypadki z różnych klas) lub w centrach skupisk wektorów z danej klasy
- losowy wybór spośród wzorców treningowych - prosta ale skuteczna metoda
- douczanie pozycji centrów metodami nadzorowanymi lub nienadzorowanymi: klasteryzacja, samoorganizacja, LVQ, ...

Ustalenie dyspersji

- wymagane gładkie odwzorowanie, rozmycie pełni rolę regularyzacji
- pola recepcyjne wszystkich funkcji bazowych powinny pokrywać cały obszar danych wejściowych
- pola recepcyjne powinny nakrywać się w niewielkim stopniu
- rozmycie σ jednakowe dla wszystkich funkcji np. (Haykin, 1994):

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

gdzie K ilość funkcji bazowych, d maksymalna odległość pomiędzy centrami.

Przykładowa funkcja gaussowska

$$\varphi(\|\mathbf{x} - \mathbf{c}_i\|) = e^{-\frac{K\|\mathbf{x} - \mathbf{c}_i\|^2}{d^2}}$$

- rozmycie równe średniej odległości od sąsiadujących centrów

Ustawienie dyspersji

- σ_i dla każdej funkcji radialnej równe odległości euklidesowej od najbliższego sąsiedniego centrum klastra (Tarasenko, 1994)
- uwzględniając odległość od P najbliższych sąsiadów (gdzie zwykle $P \leq 3$) (Moody, 1989)

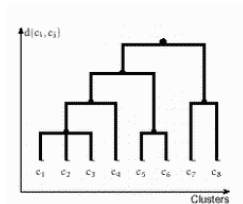
$$\sigma_i = \sqrt{\frac{1}{P} \sum_{k=1}^P \|\mathbf{c}_i - \mathbf{c}_k\|^2}$$

Incjalizacja RBF: klasteryzacja

Zastosuj algorytm klasteryzacji (analizy skupień) i przypisz centra funkcji radialnych centrom klastrów. Np. algorytm k -średnich, dendrogramy, łączenie histogramów

Metoda dendrogramów:

1. przypisz każdy $\mathbf{x}_i$ do odrębnego klastra
2. połącz najbliższą parę $\mathbf{x}_i$ i $\mathbf{x}_j$ w jeden klaster
3. powtarzaj 2 dopóki nie uzyska się zadowalającej liczby klastrów lub najmniejsza odległość między klastrami przekroczy ustalony poziom



Incjalizacja RBF: klasteryzacja

Algorytm inicjalizacji centrów przez klasteryzację (k -średnich), rozmieszcza centra w centrach skupisk:

1. Rozmieść k centrów równomiernie lub losowo ze zbioru treningowego
2. Dla centrum $\mathbf{c}_i$ znajdź zbiór wszystkich punktów z leżących najbliżej tego centrum
3. Aktualizuj położenie centrum $\mathbf{c}_i$ jako średnią dla punktów znalezionego zbioru

$$\mathbf{c}_i(k+1) = \frac{1}{N_i} \sum_{j=1}^{N_i} \bar{\mathbf{x}}_j(k)$$

4. Powtarzaj dwa ostatnie kroki aż do zbieżności

Inicjalizacja RBF: samoorganizacja

Algorytm inicjalizacji centrów przez samoorganizację (metoda on-line k -średnich):

1. Wybierz losowo k centrów skupisk
2. Dla każdego $\mathbf{x}$ ze zbioru treningowego, znajdź najbliższe centrum $\mathbf{c}_j$
3. Zmień położenie centrum $\mathbf{c}_j$ zwycięzcy

$$\mathbf{c}_i(k+1) = \mathbf{c}_i(k) + \eta_k(\mathbf{x} - \mathbf{c}_i(k))$$

4. Powtarzaj dla kolejnych $\mathbf{x}$ aż do uzyskania zbieżności

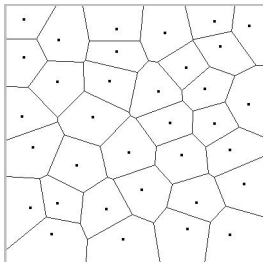
Stała uczenia η_k zanika w miarę wzrostu k , np:

$$\eta_k = \frac{\eta_0}{1 + \frac{k}{T}}$$

gdzie T - stała liczba epok

Diagram Voronoi

W drodze samoorganizacji przestrzeń dzielona jest na obszary Voronoi z centrum funkcji radialnej w każdej komórce definiującymi teselację

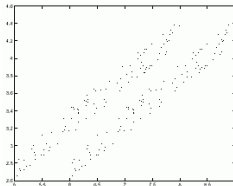


Uczenie - obroty i rozmycia

- Pełna macierz transformacji **Q** dla dużej liczby cech posiada za dużo parametrów
- Obroty funkcji zlokalizowanych są przydatne, więc wystarczy $Q_{ii} \neq 0$ i $Q_{ii+1} \neq 0$, która realizuje dowolne obroty
- Inna metoda: iloczyn funkcji radialnej z bicentralną

$$\varphi(\mathbf{x}) \cdot (\sigma(\mathbf{w}\mathbf{x} + w_0) - \sigma(\mathbf{w}\mathbf{x} + w_0))$$

- Uproszczenie: kąt obrotu ustalony po inicjacji położen



Trening RBF: metoda probabilistyczna

Zał: rozkład równomierny danych w zbiorze treningowym,
diagonalne dyspersje Σ

$$\varphi_i(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x}-\mathbf{c}_i)^T \Sigma^{-1}(\mathbf{x}-\mathbf{c}_i)}$$

Centra oraz rozmycia optymalizowane są równocześnie

$$\mathbf{c}_i(k+1) = \frac{\mathbf{c}_i(k) + \eta_k(\mathbf{x}\varphi_i(\mathbf{x}) - \mathbf{c}_i(k))}{(1 - \eta_k) + \eta_k\varphi_i(\mathbf{x})}$$

$$\Sigma_i(k+1) = \frac{\Sigma_i(k) + \eta_k(\varphi_i(\mathbf{x})\|\mathbf{x} - \mathbf{c}_i(k)\|^2 - \Sigma_i(k))}{(1 - \eta_k) + \eta_k\varphi_i(\mathbf{x})}$$

Współczynnik uczenia maleje w czasie $\eta_k = \frac{\eta_0}{k}$

Uczenie metodami gradientowymi

Minimalizacja funkcji błędu

$$E = \sum_{i=1}^N (y_i - f(\mathbf{x}_i))^2 = \sum_{i=1}^N \left(y_i - \sum_{k=0}^K w_k \varphi_k(\|\mathbf{x}_i - \mathbf{c}_k\|) \right)^2$$

$$\Delta w_j = -\eta \frac{\partial E}{\partial w_j} = -\eta (y - f(\mathbf{x})) \varphi(\|\mathbf{x} - \mathbf{c}_j\|)$$

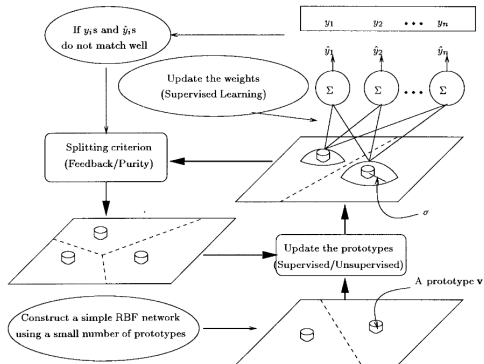
$$\Delta c_{ji} = -\eta \frac{\partial E}{\partial c_{ji}} = -\eta (y - f(\mathbf{x})) w_j e^{-\frac{\|\mathbf{x} - \mathbf{c}_j\|^2}{2\sigma_j^2}} \frac{(x_i - c_{ji})}{\sigma_j^2}$$

$$\Delta \sigma_j = -\eta \frac{\partial E}{\partial \sigma_j} = \eta (y - f(\mathbf{x})) w_j e^{-\frac{\|\mathbf{x} - \mathbf{c}_j\|^2}{2\sigma_j^2}} \frac{\|\mathbf{x} - \mathbf{c}_j\|^2}{\sigma_j^3}$$

Konstruktywny RBF

- GAL (Growing and Learning)
- GrRBF (Growing Radial Basis Function)
- FEN (Function Estimation Networks)
- RAN (Resource Allocation Networks)
- Klasyfikatory Gaussowskie, sieci probabilistyczne ...

Growing RBF



Nicolaos, *Growing Radial Basis Neural Networks: Merging Supervised and Unsupervised Learning with Network Growth Techniques*, 1997

Supervised Growing Cell Structures (Fritzke)

1. Zainicjuj małą sieć RBF
2. Wytrenuj sieć dowolną metodą
3. Wstaw nową funkcję radialną w punkcie o największym błędzie

$$\mathbf{c}_{new} = \frac{1}{2}(\mathbf{c}_m - \mathbf{c}_n)$$

gdzie $\mathbf{c}_m$ centrum odpowiadające największemu błędowi, $\mathbf{c}_n$ - najbliższe centrum względem $\mathbf{c}_m$

4. Powtarzaj dopóki błąd nie zmaleje do pożądanej wartości

Resource-Allocating Network (RAN)

- RAN (Pratt, 1991) algorytm rozrostu sieci RBF
- Startuje od pustej warstwy ukrytej i dla danego wektora $\mathbf{x}_t$ dodawany jest nowy neuron jeżeli spełnione są kryteria „odkrywczości”

$$\|\mathbf{x}_t - \mathbf{c}_i\| \geq \epsilon(t)$$

$$\|\mathbf{e}(t)\| = \|\mathbf{y}_t - f(\mathbf{x}_t)\| > e_{min}$$

gdzie $\mathbf{c}_i$ to centrum najbliższe $\mathbf{x}_t$

- Wartość progu $\epsilon(t)$ maleje z każdym krokiem

$$\epsilon(t) = \max\{\epsilon_{max} e^{-\frac{t}{T}}, \epsilon_{min}\}$$

- zakładając, że w kroku $t - 1$ istnieje k neuronów, wówczas nowy neuron dodawany jest z parametrami

$$\mathbf{c}_{k+1} = \mathbf{x}_t, \quad w_{k+1,j} = e_j(t), \quad \sigma_{k+1} = \alpha \|\mathbf{x}_t - \mathbf{c}_i\|$$

Porównanie MLP vs RBF

MLP	RBF
nielokalne relacje, wymagają do- uczania	lokalne efekty, stabilność, nie- wrażliwość na odstające przy- padki
jeden rodzaj parametrów (wagi)	kilka rodzajów parametrów: wagi, centra, rozmycia
trudna inicjalizacja	łatwa inicjalizacja
trudna interpretacja	łatwa interpretacja (prototypy)
uczenie tylko pod nadzorem	możliwe uczenie bez nadzoru
zawsze wie	czasami nie wie (przypadki w obszarach nie objętych polami receptyjnymi)
BP dość skomplikowane dla wielu warstw	uczenie łatwe bo 1 warstwa
granica decyzji perceptronu: hi- perpłaszczyzna	granica decyzyjna funkcji radial- nej: hipersfera
aktywacja: ważona suma sygna- łów $\mathbf{w}^T \mathbf{x}$	odległość od centrum funkcji bazowej $\ \mathbf{x} - \mathbf{x}\ $