UCZENIE MASZYNOWE

NIEPARAMETRYCZNE METODY APROKSYMACJI FUNKCJI

Dr hab. inż. Grzegorz Dudek
Wydział Elektryczny
Politechnika Częstochowska

Projekt finansowany w ramach programu Ministra Nauki i Szkolnictwa Wyższego pod nazwą „Regionalna Inicjatywa Doskonałości" w latach 2019 - 2022 nr projektu 020/RID/2018/19 kwota finansowania 12 000 000 PLN

PAMIĘCIOWE METODY APROKSYMACJI FUNKCJI

- Jeśli postać modelu regresyjnego jest nieznana do aproksymacji funkcji możemy użyć metod pamięciowych (nieparametrycznych)
- Metody pamięciowe przechowują zbiór przykładów trenujących i na jego podstawie tworzą hipotezę dla nowego przykładu (punktu zapytania)
- Estymując funkcję regresji uwzględniamy w modelu jej własności lokalne, czyli dotyczące otoczenia punktu zapytania
- Model jest kombinacją liniową pewnych funkcji bazowych

MIARY ODLEGŁOŚCI

Do konstrukcji funkcji regresji w otoczeniu punktu zapytania niezbędne jest zdefiniowanie metryki mierzącej odległości pomiędzy dwoma punktami (przykładami).

Miara odległości pomiędzy wektorami $\mathbf{x}_a$ i $\mathbf{x}_b$ jest funkcją:

$$d : X \times X \to \mathbb{R},$$

taką, że:

$$\exists d_0 \in \mathbb{R} : -\infty < d_0 \le d(\mathbf{x}_a, \mathbf{x}_b) < +\infty, \quad \forall \mathbf{x}_a, \mathbf{x}_b \in X,$$
$$d(\mathbf{x}_a, \mathbf{x}_a) = d_0, \quad \forall \mathbf{x}_a \in X,$$
$$d(\mathbf{x}_a, \mathbf{x}_b) = d(\mathbf{x}_b, \mathbf{x}_a), \quad \forall \mathbf{x}_a, \mathbf{x}_b \in X,$$

Jeśli ponadto:

$$d(\mathbf{x}_a, \mathbf{x}_b) = d_0 \text{ wtedy i tylko wtedy, gdy } \mathbf{x}_a = \mathbf{x}_b,$$
$$d(\mathbf{x}_a, \mathbf{x}_c) \le d(\mathbf{x}_a, \mathbf{x}_b) + d(\mathbf{x}_b, \mathbf{x}_c), \quad \forall \mathbf{x}_a, \mathbf{x}_b, \mathbf{x}_c \in X,$$

d zwana jest metryczną miarą odległości.

Do określania odległości pomiędzy punktami najczęściej stosuje się:

- odległość euklidesową:

$$d(\mathbf{x}_a, \mathbf{x}_b) = [(\mathbf{x}_a - \mathbf{x}_b)^{\mathrm{T}}(\mathbf{x}_a - \mathbf{x}_b)]^{1/2} = \sqrt{\sum_{j=1}^{n}(x_{a,j} - x_{b,j})^2}$$

- odległość miejską (*city-block*, *Manhattan*):

$$d(\mathbf{x}_a, \mathbf{x}_b) = \sum_{j=1}^{n} | x_{a,j} - x_{b,j} |$$

Każda z tych funkcji stanowi szczególny przypadek odległości Minkowskiego:

$$d(\mathbf{x}_a, \mathbf{x}_b) = \left(\sum_{j=1}^{n} | x_{a,j} - x_{b,j} |^m \right)^{1/m}$$

Miary oparte na odległości Minkowskiego nie są niezmiennicze względem skali wartości atrybutów (atrybuty wyrażone są w różnych jednostkach lub zmieniają się w różnych zakresach). Zmiana skali powoduje zmianę odległości pomiędzy punktami. Aby temu zapobiec, zaleca się wcześniejszą normalizację obserwacji lub ważenie atrybutów przy wyznaczaniu odległości.

Funkcja ważonej odległości euklidesowej ma postać:

$$d(\mathbf{x}_a, \mathbf{x}_b) = [(\mathbf{x}_a - \mathbf{x}_b)^{\mathrm{T}} \mathbf{W}^{-1} (\mathbf{x}_a - \mathbf{x}_b)]^{1/2}$$

gdzie **W** = diag{$\sigma_1^2, \sigma_2^2, \ldots, \sigma_T^2$}, σ_t są odchyleniami standardowymi poszczególnych atrybutów.

Aby miara uwzględniała również korelacje między atrybutami, stosuje się odległość Mahalanobisa:

$$d(\mathbf{x}_a, \mathbf{x}_b) = [(\mathbf{x}_a - \mathbf{x}_b)^{\mathrm{T}} \mathbf{S}^{-1} (\mathbf{x}_a - \mathbf{x}_b)]^{1/2}$$

gdzie **S** jest estymatorem macierzy kowariancji.

Ważenie stosuje się także w przypadku, gdy chcemy zróżnicować udział składowych. Funkcja ważonej odległości Minkowskiego ma postać:

$$d(\mathbf{x}_a, \mathbf{x}_b) = \left(\sum_{j=1}^{n} w_j \mid x_{a,j} - x_{b,j} \mid^m \right)^{1/m}$$

gdzie $w_j \geq 0$ jest wagą j-tej składowej (czynnikiem skalującym); często $\sum_{j=1}^{n} w_j = 1$.

Czynnik skalujący w_j rozciąga lub skraca przestrzeń obrazów wzdłuż jej osi. W przypadku gdy w_j ={0, 1}, niektóre wymiary tej przestrzeni mogą być wyeliminowane (selekcja atrybutów).

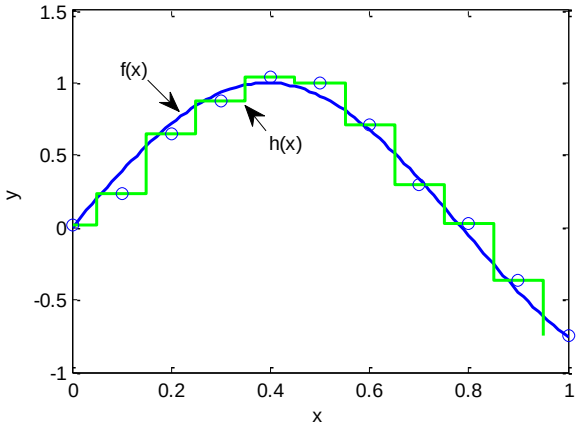
METODA NAJBLIŻSZEGO SĄSIADA

Metoda najbliższego sąsiada (NS, *nearest neighbor*) jest najprostszym pamięciowym podejściem do aproksymacji funkcji. W celu odpowiedzi na zapytanie dotyczące przykładu **x*** znajduje się najbliższy mu przykład trenujący $\langle \mathbf{x}_i, y_i \rangle$ i przyjmuje się jego etykietę za hipotezę:

$$h(\mathbf{x}^*) = \{y_i \mid \arg\min_i d(\mathbf{x}_i, \mathbf{x}^*)\}$$

Zaletą metody NS jest jej prostota i uniwersalność. Algorytm NS nie wymaga założeń dot. dziedziny i reprezentacji przykładów, poza tym, że jest na nich określona pewna miara odległości.

Mankamentem metody NS jest tendencja do nadmiernego dopasowania i schodkowa aproksymanta.

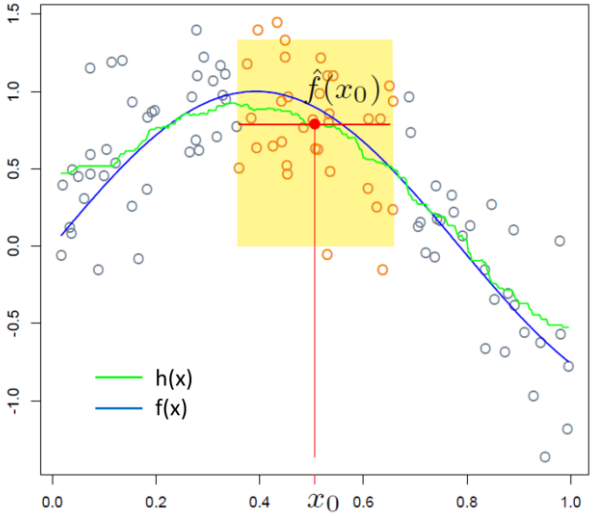


METODA *K* NAJBLIŻSZYCH SĄSIADÓW

Uogólnieniem metody NS jest metoda *k* najbliższych sąsiadów (*k*-NS), gdzie hipotezę tworzy się na podstawie etykiet *k* najbliższych przykładów do przykładu **x***, którego dotyczy zapytanie:

$$h(\mathbf{x}^*) = \frac{1}{k}\sum_{i\in\Omega} y_i$$

gdzie Ω jest zbiorem indeksów *k* najbliższych sąsiadów **x*** w zbiorze trenującym.

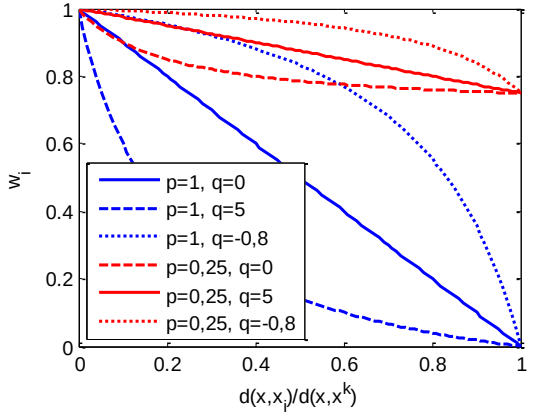


Dalszym uogólnieniem metody NS jest wprowadzenie wag zależnych od odległości przykładu $\mathbf{x}_i$ ze zbioru *k*-NS od $\mathbf{x}^*$, np.:

$$h(\mathbf{x}^*) = \frac{\sum_{i\in\Omega} w_i(\mathbf{x}) y_i}{\sum_{i\in\Omega} w_i(\mathbf{x})}$$

gdzie $w_i(\mathbf{x})$ jest funkcja ważącą, zależną od odległości, zwykle monotonicznie malejącą, osiągającą maksymalną wartość w zerze, a minimalną (nieujemną) – dla odległości do *k*-tego najbliższego sąsiada ($\mathbf{x}^k$), np.:

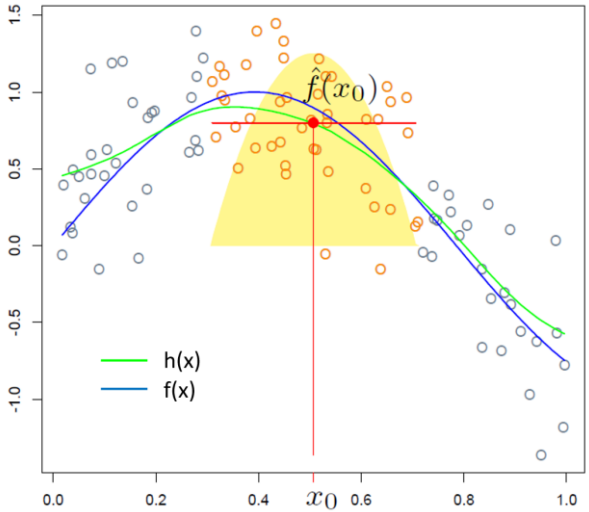
$$w_i(\mathbf{x}) = p\left(\frac{1-\frac{d(\mathbf{x},\mathbf{x}_i)}{d(\mathbf{x},\mathbf{x}^k)}}{1+q\frac{d(\mathbf{x},\mathbf{x}_i)}{d(\mathbf{x},\mathbf{x}^k)}} - 1\right) + 1$$



Generalnie im gładsza jest funkcja wagowa, tym gładsza jest funkcja estymowana. Przykład $\mathbf{x}_i$ ze zbioru *k*-NS uwzględniany jest w formowaniu hipotezy dla przykładu $\mathbf{x}^*$ w stopniu zależnym od odległości od $\mathbf{x}^*$. Stopień ten jest wyrażony wagą $w_i(\mathbf{x})$.

Inne typy funkcji wagowych:

- gaussowska – $w(\mathbf{x}) = \exp\left(-\frac{d(\mathbf{x},\mathbf{x}_i)^2}{k^2}\right)$
- hiperbola kwadratowa – $w(\mathbf{x}) = \frac{1}{1+m(d(\mathbf{x},\mathbf{x}_i)/k)^2}$



Estymator Nadarayi–Watsona

Dalszym uogólnieniem jest wprowadzenie funkcji wagowych, zwanych jądrami K^* (*kernel*), dla każdego przykładu trenującego. Dla przypadku jednowymiarowego hipoteza ma postać (tzw. estymator Nadarayi–Watsona):

$$h(x) = \frac{\sum_{i=1}^{N} K\left(\frac{x - x_i}{s}\right) y_i}{\sum_{i=1}^{N} K\left(\frac{x - x_i}{s}\right)}$$

gdzie $s \in \mathbb{R}^+$ jest parametrem wygładzania.

* Patrz wykład na temat klasyfikatorów Bayesa, slajdy 13–16.

ESTYMATOR NADARAYI–WATSONA – PRZYKŁAD

Na podstawie przykładów trenujących:

x	3.0	3.3	3.9	4.2	4.9	6.9	7.2
y	0.30	0.34	0.42	0.48	0.56	0.64	0.62

wyznacz wartość funkcji w punkcie x* = 6.5.

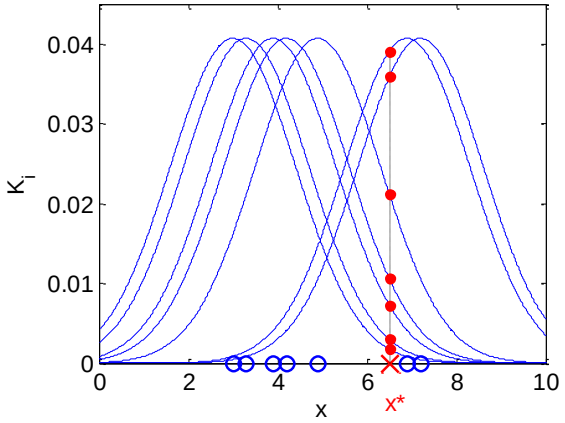
Jako jądra przyjmijmy funkcje gaussowskie:

$$K\left(\frac{x-x_i}{s}\right)=\frac{1}{s\sqrt{2\pi}}\exp\left(-\frac{(x-x_i)^2}{2s^2}\right)$$

gdzie za *s* przyjęto 1.4.

Wartości jąder kolejnych punktów trenujących w punkcie x* ($K((6.5-x_i)/1.4)$ wynoszą:

K_i = [0.0018, 0.0030, 0,0073, 0.0106, 0.0212, 0.0391, 0.0359]



Rysunek. Funkcje jądrowe rozpięte nad przykładami trenującymi i ich wartości w punkcie x.*

ESTYMATOR NADARAYI–WATSONA – PRZYKŁAD

Estymator Nadarai–Watsona można zapisać jako:

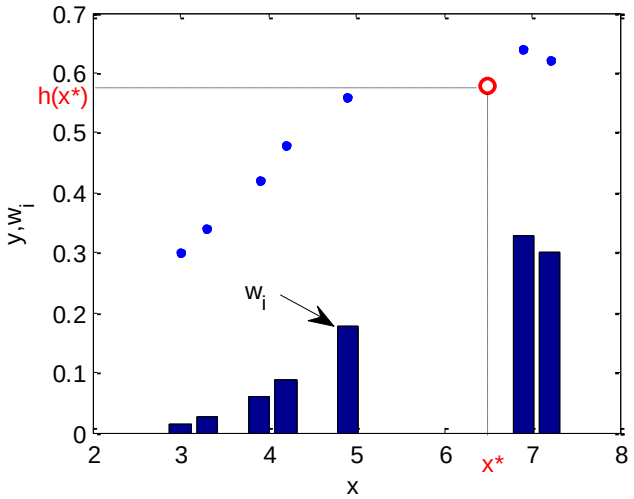
$$h(x) = \sum_{i=1}^{N} w_i(x) y_i \text{ , gdzie } w_i(x) = \frac{K\left(\frac{x - x_i}{s}\right)}{\sum_{k=1}^{N} K\left(\frac{x - x_k}{s}\right)}$$

Wartości wag w_i dla kolejnych punktów uczących wynoszą:

w_i = [0.0151, 0.0251, 0.0610, 0.0892, 0.1784, 0.3291, 0.3021]

Estymowana wartość funkcji w punkcie x^*:

$$h(x) = 0.0151 \cdot 0{,}30 + 0{,}0251 \cdot 0{,}34 + 0{,}0610 \cdot 0{,}42 + \\ + 0{,}0892 \cdot 0{,}48 + 0{,}1784 \cdot 0{,}56 + 0{,}3291 \cdot 0{,}64 + \\ 0{,}3021 \cdot 0{,}62 = 0{,}5793$$



Rysunek. Aproksymacja metodą Nadarayi-Watsona.

ESTYMATOR NADARAYI–WATSONA

W przypadku wielowymiarowym stosuje się jądra produktowe (iloczyny jednowymiarowych jąder dla poszczególnych współrzędnych/atrybutów). Wtedy otrzymujemy:

$$h(\mathbf{x}) = \frac{\sum_{i=1}^{N} \prod_{j=1}^{n} K\left(\frac{x_j - x_{i,j}}{s_j}\right) y_i}{\sum_{i=1}^{N} \prod_{j=1}^{n} K\left(\frac{x_j - x_{i,j}}{s_j}\right)}$$

Dla jąder gaussowskich produktowy estymator Nadarayi–Watsona przybiera postać:

$$h(\mathbf{x}) = \frac{\sum_{i=1}^{N} \exp\left(-\sum_{j=1}^{n} \frac{(x_j - x_{i,j})^2}{2s_j^2}\right) y_i}{\sum_{i=1}^{N} \exp\left(-\sum_{j=1}^{n} \frac{(x_j - x_{i,j})^2}{2s_j^2}\right)}$$

W tym przypadku parametry wygładzania dobiera się indywidualnie dla każdego wymiaru/atrybutu.

ESTYMATOR NADARAYI–WATSONA

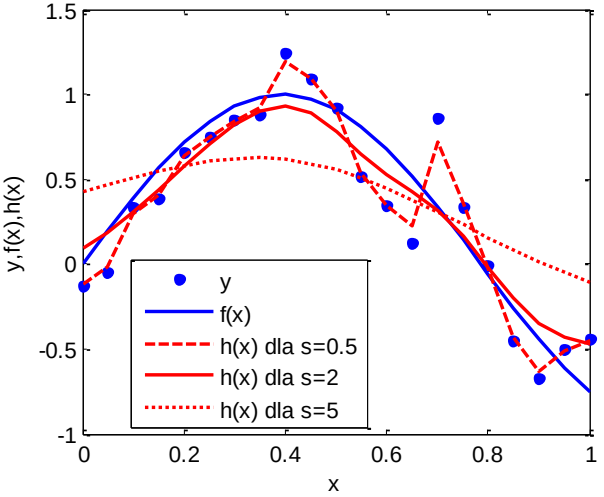
W pierwszym przybliżeniu parametry wygładzania można wyznaczyć ze wzoru:

$$s_j = \sigma_j \left(\frac{4}{(n+2)N} \right)^{\frac{1}{n+4}} \approx \sigma_j N^{-\frac{1}{n+4}}$$

gdzie σ_j jest odchyleniem standardowym x_j estymowanym z próby,

i dostroić w procedurze kroswalidacji.

Decydując o kompromisie pomiędzy obciążeniem i wariancją estymatora, parametry wygładzania s_j istotnie wpływają na jego jakość, stąd bardzo ważny jest precyzyjny dobór ich wartości. Zbyt małe wartości s_j skutkują nadmiernym dopasowaniem modelu do danych uczących, natomiast zbyt duże – nadmiernym wygładzeniem estymatora, maskującym specyficzne cechy aproksymowanej funkcji[†].



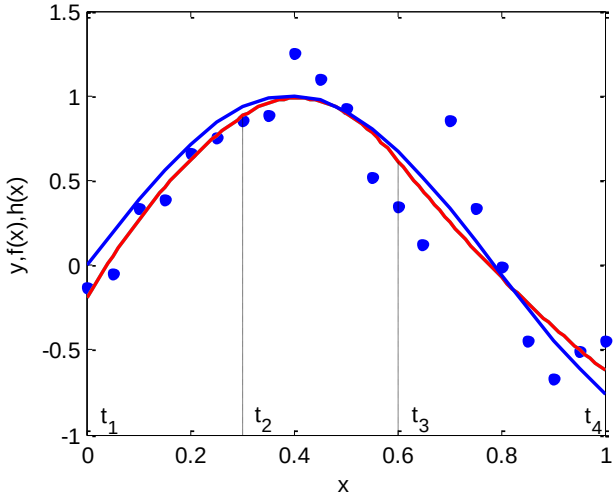
[†]Skrypt implementujący estymator Nadarayi-Watsona: http://www.mathworks.com/matlabcentral/fileexchange/39361-nadaraya-watson-smoothing/content/smoothing.m

APROKSYMACJA ZA POMOCĄ FUNKCJI SKLEJANYCH

Funkcja sklejana[‡] zwana splajnem określona na przedziale $[t_1, t_{m+1}]$ jest

- wielomianem stopnia co najwyżej q w każdym podprzedziale $[t_k, t_{k+1}]_{k = 1, 2, \ldots, m}$ (brzegi przedziałów t_k nazywamy węzłami) i
- posiada ciągłe pochodne rzędu 1, 2, …, q–1 dla wszystkich argumentów z przedziału $[t_1, t_{m+1}]$.

Czyli jest to funkcja gładka, odcinkowo wielomianowa. Gładkość zapewniona jest przez jednakowe pochodne na granicach podprzedziałów.



Funkcja sklejana ma postać:

$$h(x) = \sum_{l=1}^{q} \alpha_l x^{l-1} + \sum_{k=1}^{m} \beta_k (x - t_k)_+^q \qquad (*)$$

[‡] Koronacki J., Ćwik J.: Statystyczne systemy uczące się. WNT 2005.

APROKSYMACJA ZA POMOCĄ FUNKCJI SKLEJANYCH

gdzie: α_l, β_k – współczynniki, $(x-t_k)_+^q = \begin{cases} (x-t_k)^q, & \text{dla } x \geq t_k \\ 0, & \text{dla } x < t_k \end{cases}$

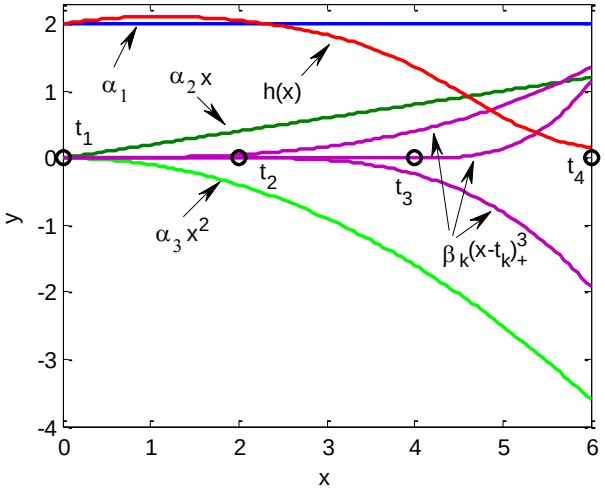
Przykładowo dla $q = 3$ i trzech podprzedziałów (patrz rysunek powyżej) funkcja sklejana ma postać:

$$h(x) = \alpha_1 + \alpha_2 x + \alpha_3 x^2 + \beta_1 (x-t_1)_+^3 + \beta_2 (x-t_2)_+^3 + \beta_3 (x-t_3)_+^3$$

Jest to suma:

- wyrazu wolnego α_1
- prostej $\alpha_2 x$
- paraboli $\alpha_3 x^2$ oraz
- wielomianów $\beta_k (x-t_k)_+^3$

Wielomian $\beta_k (x-t_k)_+^3$ „zaczyna” się w punkcie t_k i jest niezerowy na prawo od tego punktu.



Dobór liczby przedziałów m oraz węzłów t_k odbywa się adaptacyjnie, np. w procedurze kroswalidacji.

Dla zadanych wartości m oraz t_k współczynniki α_l, β_k estymuje się metodą najmniejszych kwadratów[§].

Inny sposób estymacji parametrów polega na minimalizacji błędu powiększonego o składnik kary (regularyzacja)[**]:

$$\sum_{i=1}^{N}(y_i - h(x_i))^2 + \lambda\int_R (h''(x))^2\,dx$$

gdzie λ jest tzw. współczynnikiem wygładzania, a $h''(x)$ jest drugą pochodną hipotezy.

Całka z kwadratu drugiej pochodnej hipotezy jest tym większa im bardziej oscylacyjny jest przebieg tej funkcji. Im większa wartość λ, tym większa jest kara za niegładkość (oscylacyjność). Kara pozwala na wygładzenie estymatora (łagodniejszą zmienność).

[§] zauważ że funkcja (*) jest modelem liniowym (patrz ostatni slajd wykładu 6 – rozszerzona reprezentacja)
[**] Koronacki J., Ćwik J.: Statystyczne systemy uczące się. WNT 2005.

Często przyjmuje się, że węzłami są wszystkie punkty trenujące, a q = 3 (splajny kubiczne). Wtedy jedynym parametrem do estymacji jest współczynnik wygładzania λ. Jego wartość dobiera się w kroswalidacji.

W przypadku wielowymiarowym (przykłady są wektorami) hipoteza jest rozwiązaniem następującego zadania minimalizacji:

$$\hat{h}(\mathbf{x}) = \arg\min_{h(.)} \left(\sum_{i=1}^{N} (y_i - h(\mathbf{x}_i))^2 + \lambda R(h) \right)$$

gdzie kara za niegładkość ma postać:

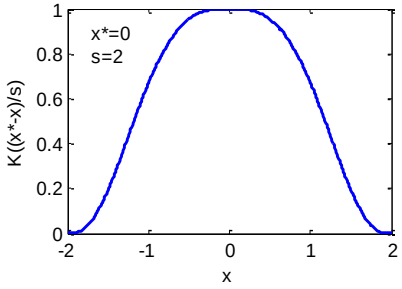
$$R(h) = \sum_{j=1}^{n} \sum_{k=1}^{n} \int \frac{\partial^2 h}{\partial x_j \partial x_k} d\mathbf{x}$$

LOKALNA REGRESJA LINIOWA

W lokalnej regresji liniowej (LOWESS, *locally weighted scatterplot smoothing*) dla przykładu **x*** (zapytania) tworzy się model liniowy na podstawie zbioru trenującego, przy czym w minimalizowanej funkcji błędu uwzględnia się udział każdego przykładu w stopniu zależnym od odległości od **x***:

$$\sum_{i=1}^{N} K\left(\frac{x^*-x_i}{s}\right)(y_i - h(x^*))^2 \qquad (**)$$

Odległość od $\mathbf{x}_i$ od **x*** wyrażona jest za pomocą jądra *K* (tradycyjnie używa się jądra kubicznego: $K(x_i, x^*, s) = \left(1-\left(\frac{|x^*-x_i|}{s}\right)^3\right)^3$).



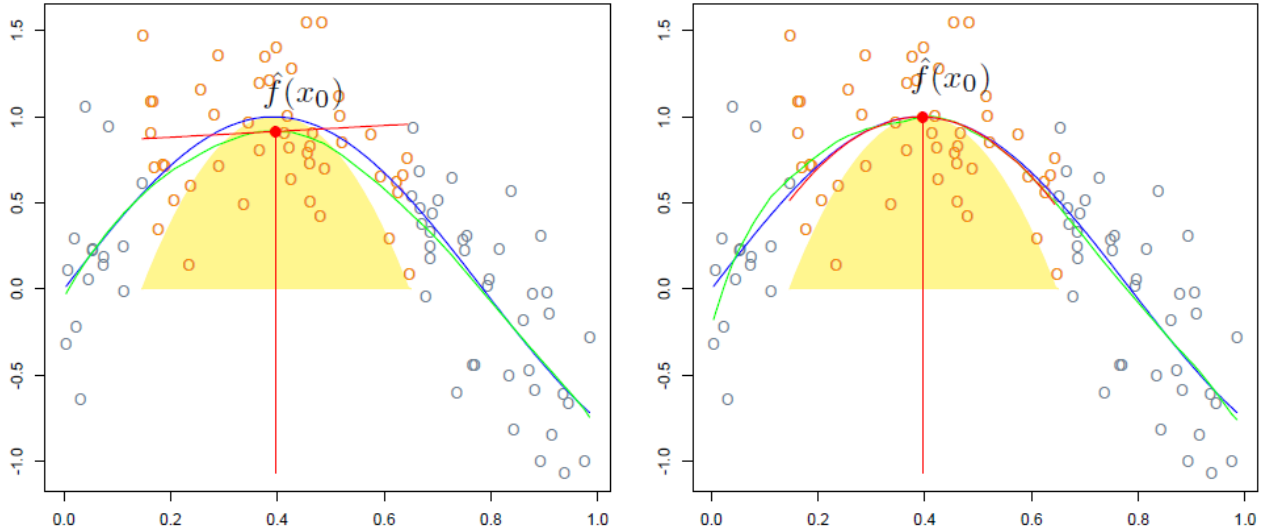
Hipoteza ma postać liniową:

$$h(x) = a_0(x^*) + a_1(x^*)x$$

Model liniowy dopasowywany jest lokalnie do przykładów w otoczeniu zapytania *x** metodą ważonych najmniejszych kwadratów minimalizującą funkcję (**).

LOKALNA REGRESJA LINIOWA I WIELOMIANOWA

Zamiast hipotezy liniowej można użyć wielomianu, np.: $h(x) = a_0(x^*) + a_1(x^*)x + a_2(x^*)x^2$. Pozwala to na dokładniejszą aproksymację, uwzględniającą lokalną nieliniowość funkcji docelowej.



Parametr *s* dobiera się w kroswalidacji.

Metodę regresji lokalnej można uogólnić na przypadek wielowymiarowy.