

UCZENIE MASZYNOWE

LINIOWE METODY APROKSYMACJI

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

Metody aproksymacji pod względem sposobu reprezentacji funkcji aproksymowanej dzielimy na:

- parametryczne, w których hipoteza reprezentowana jest przez wektor liczb rzeczywistych – parametrów modyfikowanych w trakcie uczenia, na podstawie których obliczana jest wartość funkcji dla dowolnego punktu z dziedziny,
- pamięciowe (*memory-based*, *instance-based*, *similarity-based*, *lazy learners*), przechowujące zbiór przykładów uczących i wyznaczające wartość hipotezy na podstawie przykładów najbardziej podobnych do przykładu wejściowego,
- symboliczne, oparte na symbolicznej reprezentacji funkcji, np. w postaci drzew decyzyjnych.

Założenie

- przykłady opisywane są wektorami liczb rzeczywistych (atributów) $\mathbf{x} = [1, x_1, x_2, \dots, x_n]^T$
- hipoteza reprezentowana jest wektorem liczb rzeczywistych (parametrów, wag). W modelu liniowym: $\mathbf{w} = [w_0, w_1, w_2, \dots, w_n]^T$

Uczenie się

Uczenie się polega na modyfikowaniu wag na podstawie przykładów trenujących.

Cel – zmniejszenie błędu aproksymacji, np.:

$$E(h_{\mathbf{w}}) = \frac{1}{2} \sum_{\mathbf{x} \in P} (f(\mathbf{x}) - h_{\mathbf{w}}(\mathbf{x}))^2$$

gdzie: $f(\mathbf{x})$ jest wartością funkcji docelowej dla przykładu $\mathbf{x}$, a $h_{\mathbf{w}}(\mathbf{x}) = F(\mathbf{x}, \mathbf{w})$ jest wartością hipotezy przy wagach $\mathbf{w}$ dla przykładu $\mathbf{x}$.

Ponieważ $f(\mathbf{x})$ jest zwykle nieznana w jej miejsce podstawiamy y ($y = f(\mathbf{x}) + \varepsilon$).

Proces uczenia się często wykorzystuje regułę spadku gradientu (patrz – perceptron).

Gradient $\nabla E(h_w) = \left[\frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}, \dots, \frac{\partial E}{\partial w_n} \right]$ ze znakiem ujemnym wskazuje kierunek "przesunięcia"

wag:

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla E(h_w)$$

$$\mathbf{w} \leftarrow \mathbf{w} + \eta \sum_{\mathbf{x} \in P} (f(\mathbf{x}) - h_w(\mathbf{x})) \nabla_w h_w(\mathbf{x})$$

gdzie $\eta > 0$ jest współczynnikiem uczenia.

Powyższa reguła nosi nazwę **uogólnionej reguły delta**. Reprezentuje ona ogólny algorytm uczenia się w parametrycznych aproksymatorach funkcji (tryb epokowy – wagi modyfikowane są po każdej epoce). Jego konkretyzacja zależy od funkcji F , czyli od charakteru zależności hipotezy od $\mathbf{w}$.

Regułę tę można zastosować w trybie inkrementacyjnym (adaptacja wag po prezentacji każdego przykładu): $\mathbf{w} \leftarrow \mathbf{w} + \eta (f(\mathbf{x}) - h_w(\mathbf{x})) \nabla_w h_w(\mathbf{x})$

APROKSYMATOR LINIOWY (PERCEPTRON)

Hipoteza ma postać:

$$h_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T \mathbf{x} = \sum_{i=0}^n w_i x_i$$

Gradient hipotezy:

$$\nabla h_{\mathbf{w}}(\mathbf{x}) = \left[\frac{\partial h_{\mathbf{w}}}{\partial w_0}, \frac{\partial h_{\mathbf{w}}}{\partial w_1}, \dots, \frac{\partial h_{\mathbf{w}}}{\partial w_n} \right] = [1, x_1, x_2, \dots, x_n] = \mathbf{x}$$

Reguła delta (zwana reguła Widrowa-Hoffa, adaline lub LMS) dla hipotezy liniowej:

$$\begin{array}{ll} \text{tryb epokowy} & \text{tryb inkrementacyjny} \\ \mathbf{w} \leftarrow \mathbf{w} + \eta \sum_{\mathbf{x} \in P} (y - \mathbf{w}^T \mathbf{x}) \mathbf{x} & \text{lub } \mathbf{w} \leftarrow \mathbf{w} + \eta (y - \mathbf{w}^T \mathbf{x}) \mathbf{x} \end{array}$$

UCZENIE PERCEPTRONU

Tryb epokowy	Tryb inkrementacyjny
1. Wybierz losowo $\mathbf{w}$, ustal η 2. Powtarzaj 2.1. $\Delta = 0, E = 0$ 2.2. Powtarzaj dla $i = 1, 2, \dots, N$$\Delta = \Delta + (y_i - \mathbf{w}^T \mathbf{x}_i) \mathbf{x}_i$$E = E + (y_i - \mathbf{w}^T \mathbf{x}_i)^2$ 2.3. $\mathbf{w} \leftarrow \mathbf{w} + \eta \Delta$ 2.4. Jeśli osiągnięto warunek stopu, to zakończ	1. Wybierz losowo $\mathbf{w}$, ustal η 2. Powtarzaj 2.1. $E = 0$ 2.2. Powtarzaj dla $i = 1, 2, \dots, N$$E = E + (y_i - \mathbf{w}^T \mathbf{x}_i)^2$$\mathbf{w} \leftarrow \mathbf{w} + \eta (y_i - \mathbf{w}^T \mathbf{x}_i) \mathbf{x}_i$ 2.3. Jeśli osiągnięto warunek stopu, to zakończ

Warunek stopu: osiągnięto $k_{\max}$ iteracji lub przez ostatnie K iteracji nie osiągnięto znaczącej poprawy rezultatu lub $E < E_{\max}$.

W regresji wagi wyznacza się w sposób **analityczny**.

Błąd aproksymacji możemy zapisać:

$$E(h_w) = \sum_{\mathbf{x} \in P} (y - h_w(\mathbf{x}))^2 = (\mathbf{Y} - \mathbf{X}\mathbf{w})^T (\mathbf{Y} - \mathbf{X}\mathbf{w}) = \mathbf{Y}^T \mathbf{Y} - 2(\mathbf{X}\mathbf{w})^T \mathbf{Y} + (\mathbf{X}\mathbf{w})^T \mathbf{X}\mathbf{w}$$

gdzie:

$$\mathbf{Y} = [y_1, y_2, \dots, y_N]^T \text{ – wektor pożądaných odpowiedzi, } \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,n} \\ 1 & x_{2,1} & \dots & x_{2,n} \\ \dots & \dots & \dots & \dots \\ 1 & x_{N,1} & \dots & x_{N,n} \end{bmatrix} \text{ – macierz przykładów}$$

Przyrównując pochodną błędu do zera $-\mathbf{X}^T \mathbf{Y} + \mathbf{X}^T \mathbf{X} \mathbf{w} = 0$ otrzymujemy wagi modelu regresji liniowej:

$$\mathbf{w} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

Uwaga: Jeśli kolumny $\mathbf{X}$ są liniowo zależne, to wyznacznik macierzy $\mathbf{X}^T \mathbf{X}$ jest zerowy i macierzy tej nie można odwrócić.

Dla przypadku jednowymiarowego ($n = 1$) otrzymujemy:

$$w_1 = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (y_i - \bar{y})}, \quad w_0 = \bar{y} - w_1 \bar{x}$$

gdzie i oznacza numer przykładu uczącego, $\bar{x}$ i $\bar{y}$ to wartości średnie w zbiorze uczącym.

Problem: duża liczba atrybutów (zależnych od siebie, nie wpływających na zmienną wyjściową y)

Cel: wybrać niewielki podzbiór atrybutów, który pozwoli uprościć model zachowując jego dokładność.

Regresja krokowa (*stepwise regression*) pozwala wyłonić w procedurze krokowej atrybuty istotne i zbudować na nich model liniowy, który zapewnia najmniejszy błąd regresji.

Konstrukcja modelu regresji krokowej może przebiegać w trzech trybach:

- Krokowe dodawanie atrybutów –
 1. Ustal $\Phi = \{x_1, x_2, \dots, x_n\}$ – zbiór atrybutów kandydujących i $\Omega = \emptyset$ – zbiór atrybutów istotnych
 2. Powtarzaj dla każdego $x_i \in \Phi$:
 - 2.1. Zbuduj model z wykorzystaniem wszystkich atrybutów z Ω i i -tego atrybutu z Φ ; odnotuj błąd tego modelu
 3. Jeśli nie nastąpiła poprawa modelu w p. 2 – zakończ
 4. Wybierz atrybut z Φ , dla którego nastąpiła największa poprawa modelu i przenieś go z Φ do Ω
 5. Powtórz kroki 2-5

- Krokowa eliminacja atrybutów –
 1. Ustal $\Omega = \{x_1, x_2, \dots, x_n\}$ – zbiór atrybutów istotnych
 2. Powtarzaj dla każdego $x_i \in \Omega$:
 - 2.1. Zbuduj model z wykorzystaniem atrybutów z Ω pomijając i -ty atrybut; odnotuj błąd tego modelu
 3. Jeśli nie nastąpiła poprawa modelu w p. 2 – zakończ
 4. Wybierz atrybut z Ω , po pominięciu którego nastąpiła największa poprawa modelu i usuń go z Ω
 5. Powtórz kroki 2-5
- Naprzemienne użycie dwu powyższych trybów

Miarą poprawy modelu jest tzw. p -wartość (liczbowe wyrażenie istotności statystycznej) testu statystycznego F -Snedecora* (stosuje się też inne kryteria).

* patrz: Józwiak J., Podgórski J.: Statystyka od podstaw. PWE 2001, str. 403

Problem: duże wagi w_i (co do modułu) sprawiają, że wyjście y jest wrażliwe na małe zmiany wejść x_i

Cel: zmniejszyć wagi (co do modułu)

W **regresji grzbietowej** (*ridge regression*) kryterium zawiera sumę kwadratów wag jako składnik kary[†]:

$$E(h_w) = \sum_{\mathbf{x} \in P} (y - h_w(\mathbf{x}))^2 + \lambda \sum_{i=1}^n w_i^2 = (\mathbf{Y} - \mathbf{X}\mathbf{w})^T (\mathbf{Y} - \mathbf{X}\mathbf{w}) + \lambda \mathbf{w}^T \mathbf{w}$$

gdzie $\lambda \geq 0$ jest parametrem określającym stopień uwzględnienia kary w kryterium.

Dla $\lambda = 0$ otrzymujemy zwykły model regresji liniowej; dla $\lambda = \infty$ otrzymujemy zerowe wagi.

Aby wyrównać wpływ poszczególnych wag na wartość kary przed wykonaniem obliczeń należy sprowadzić wartości wszystkich atrybutów do tej samej skali (wariancja próbkowa wszystkich atrybutów powinna wynosić 1).

[†] Zapis w postaci macierzowej wymaga wcześniejszego wyeliminowania wyrazu wolnego w_0 i scentrowania atrybutów. Wtedy macierz $\mathbf{X}$ ma rozmiary $N \times n$, a $\mathbf{w} - N \times 1$. Szczegóły w [Tib].

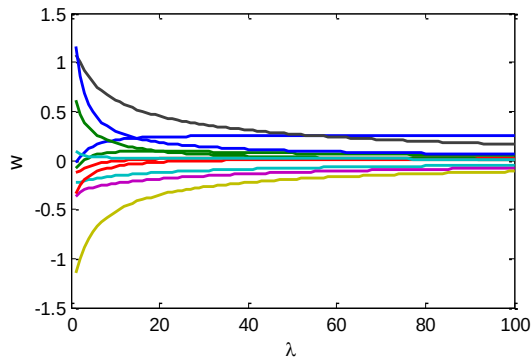
Wagi minimalizujące powyższe kryterium wyznacza się ze wzoru:

$$\mathbf{w} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{Y}$$

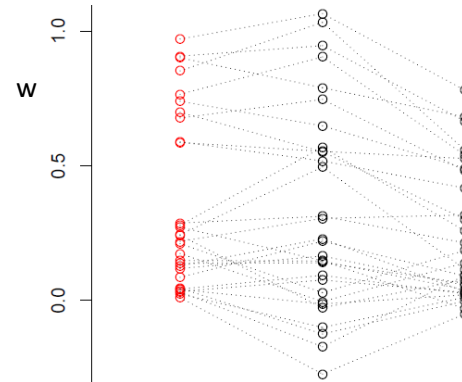
gdzie $\mathbf{I}$ jest macierzą jednostkową (z jedynkami na przekątnej) o wymiarach $n \times n$.

Optymalną wartość parametru λ dobiera się w procedurze kroswalidacji.

Rys. Przykładowe wartości wag dla różnych wartości λ .



Rys. Wartości wag (od lewej): rzeczywiste, estymowane w regresji liniowej, estymowane w regresji grzbietowej



REGULARYZACJA

Wzbogacenie kryterium o karę w postaci sumy kwadratów wag nazywa się **regularyzacją Tichonową**. Regularyzacja zapobiega przeuczeniu modelu (nadmiernemu dopasowaniu).

Regularyzacja pozwala zredukować błąd średniokwadratowy (MSE). Odbywa się to poprzez redukcję wariancji, chociaż obciążenie modelu wzrasta.

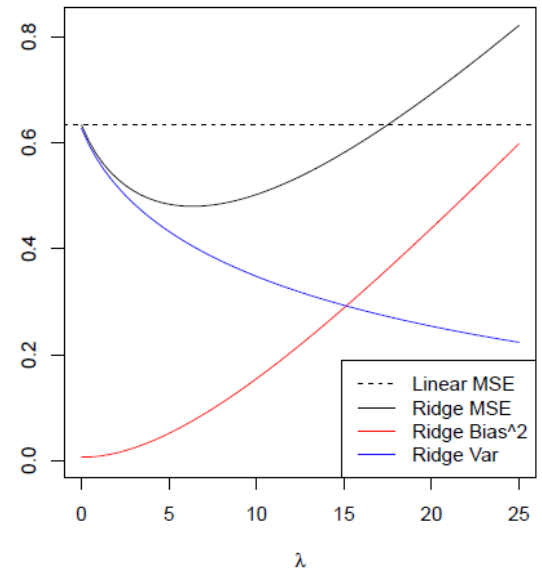
Błąd MSE można zapisać jako:

$$E[(f(\mathbf{x}) - h(\mathbf{x}))^2] = \underbrace{(E[h(\mathbf{x})] - f(\mathbf{x}))^2}_{\text{MSE (mean squared error)}} + \underbrace{E[(h(\mathbf{x}) - E[h(\mathbf{x})])^2]}_{\text{Kwadrat obciążenia (bias)^2}} + \underbrace{E[(h(\mathbf{x}) - E[h(\mathbf{x})])^2]}_{\text{Wariancja (var)}}$$

gdzie $E(.)$ oznacza wartość oczekiwaną.

Wariancja informuje jak wrażliwy jest model na drobne zmiany w zbiorze uczącym.

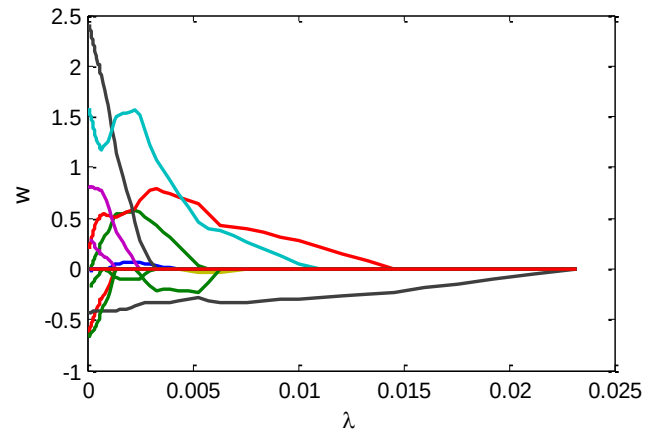
Obciążenie informuje jak dokładny jest model dla różnych zbiorów uczących.



LASSO (*Least Absolute Shrinkage and Selection Operator*) jest metodą regularyzacji modelu regresji liniowej, w której kryterium zawiera sumę modułów wag jako składnik kary:

$$E(h_w) = \sum_{\mathbf{x} \in P} (y - h_w(\mathbf{x}))^2 + \lambda \sum_{i=1}^n |w_i|$$

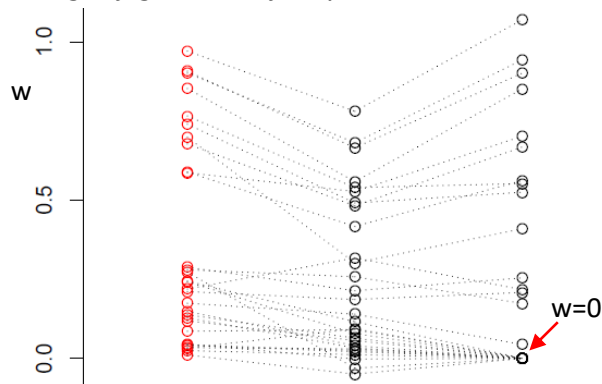
Wprowadzenie kary w postaci sumy modułów zamiast sumy kwadratów wag ma ciekawe konsekwencje. W regresji grzbietowej wagi zmniejszają się wraz ze wzrostem λ , ale nigdy nie osiągają zera. W LASSO wagi mogą się zerować przy odpowiednio dużych wartościach λ . LASSO jest jednocześnie algorytmem **regularyzacji** i **selekcji atrybutów** (atrybuty z zerowymi wagami nie są uwzględniane w modelu).



Podobnie jak w regresji grzbietowej redukcja MSE odbywa się poprzez redukcję wariancji, chociaż obciążenie modelu wzrasta.

Wyznaczenie wartości wag minimalizujących kryterium używane w LASSO nie jest możliwe na drodze analitycznej jak w regresji grzbietowej lecz wymaga algorytmu iteracyjnego.

Rys. Wartości wag (od lewej): rzeczywiste, estymowane w regresji grzbietowej, estymowane w LASSO



Elastyczna sieć (*elastic net*) łączy regresję grzbietową z LASSO. Składnik kary ma tutaj postać:

$$\lambda \sum_{i=1}^n (\alpha |w_i| + (1 - \alpha) w_i^2)$$

gdzie $\alpha \in [0, 1]$. Dla $\alpha = 0$ otrzymujemy regresję grzbietową, dla $\alpha = 1$ otrzymujemy LASSO.

ROZSZERZONA REPREZENTACJA

Rozszerzona reprezentacja polega na wzbogaceniu przykładów o dodatkowe atrybuty, które są funkcjami atrybutów oryginalnych, np: x_i^2 , x_i^3 , $x_i x_j$, $\log(x_i)$, $\sin(x_i)$, $1/x_i$ itd.

Model zbudowany na rozszerzonych przykładach, np. $\mathbf{x} = [x_1, x_2, x_1^2, x_2^2, x_1 x_2]$ postaci:

$$h_w(\mathbf{x}) = w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2 + w_5 x_1 x_2 + w_0$$

to nadal model **regresji liniowej**, choć zależność pomiędzy $\mathbf{x}$ a y , którą wyraża **nie jest liniowa**! Wagi takiego modelu estymujemy opisanymi powyżej metodami regresji i aproksymacji liniowej.

