Wstęp do metod numerycznych

Aproksymacja i zagadnienie najmniejszych kwadratów

P. F. Góra

http://th-www.if.uj.edu.pl/zfs/gora/

2018

Aproksymacja

Termin *aproksymacja* występuje w dwu znaczeniach:

Aproksymacja punktowa: Mając N punktów, staramy się znaleźć funkcję *należącą do znanej kategorii*, która będzie przebiegać możliwie "najbliżej" tych punktów. Podkreślam, że funkcja jest znana co do swego kształtu (np. wielomian ustalonego stopnia, kombinacja funkcji trygonometrycznych, funkcja opisująca jakiś rozkład prawdopodobieństwa itp), a tylko nieznane są jej parametry.

Aproksymacja ciągła: Mając ustaloną funkcję $g(x)$, której sposób obliczania jest trudny, skonstruować inną funkcję, która będzie w pewnym sensie bliska funkcji wyjściowej, a jednocześnie obliczeniowo prostsza.

Aproksymacja punktowa

Interpolacja punktowa najczęściej kojarzy się z dopasowaniem funkcji do danych doświadczalnych. Mamy N par punktów $\{(x_i, y_i)\}_{i=1}^{N}$, gdzie x_i jest dokładną wartością argumentu, y_i zmierzoną (lub obliczoną na jakimś wcześniejszym etapie) wartością funkcji. Skrajnym przypadkiem aproksymacji punktowej jest interpolacja — funkcja przechodzi przez *wszystkie* punkty doświadczalne, ale jest "trudną" funkcją: wielomianem wysokiego stopnia, funkcją sklejaną, skomplikowaną funkcją wymierną, my tymczasem chcemy mieć jakąś "prostą" funkcję, przechodzącą dostatecznie blisko wszystkich punktów.

Z teorii możemy wiedzieć, że zależność pomiędzy x a y *powinna mieć* charakter $y = y(x)$, jednak zmierzone (lub obliczone) wartości nie odpowiadają dokładnie wartościom teoretycznym, gdyż są obarczone błędami pomiarowymi (obliczeniowymi).

Liniowe zagadnienie najmniejszych kwadratów

Każdej zmierzonej (i obarczonej błędem) wartości y_i odpowiada wartość teoretyczna $\tilde{y}_i$, jaką zmienna y "powinna" przybrać dla danej wartości zmiennej x. Przyjmujemy, że wartość teoretyczna jest **kombinacją liniową** pewnych znanych funkcji:

$$\tilde{y}_i = a_1 \cdot f_1(x_i) + a_2 \cdot f_2(x_i) + \cdots + a_s \cdot f_s(x_i) \tag{1}$$

Zespół *wszystkich* wartości teoretycznych możemy zatem przedstawić jako

$$\tilde{\mathbf{y}} = \mathbf{A}\mathbf{p}\,, \tag{2a}$$

gdzie

$$\mathbf{A} = \begin{bmatrix} f_1(x_1) & f_2(x_1) & f_3(x_1) & \cdots & f_s(x_1) \\ f_1(x_2) & f_2(x_2) & f_3(x_2) & \cdots & f_s(x_2) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ f_1(x_n) & f_2(x_n) & f_3(x_n) & \cdots & f_s(x_n) \end{bmatrix} \in \mathbb{R}^{n \times s}\,, \tag{2b}$$

$$\mathbf{p} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_s \end{bmatrix} \in \mathbb{R}^{s}\,. \tag{2c}$$

Problemem numerycznym, który chcemy rozwiązać, jest znalezienie "najlepszego" wektora parametrów $\mathbf{p}$.

Przykład

Do n punktów pomiarowych (x_1, y_1), (x_2, y_2),...,(x_n, y_n) dopasowujemy wielomian drugiego stopnia $\tilde{y} = ax^2 + bx + c$. Wartości teoretyczne możemy zapisać jako

$$\tilde{\mathbf{y}} = \begin{bmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ \dots & \dots & \dots \\ x_n^2 & x_n & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix}. \tag{3}$$

Zauważmy, że dopasowywanie do danych wielomianu ustalonego stopnia (nie tylko linii prostej!) jest zagadnieniem liniowym!

Błędy pomiarowe

Różnica pomiędzy wartością zmierzoną y_i a wartością teoretyczną $\tilde{y}_i$ jest spowodowana błędem pomiarowym: $y_i - \tilde{y}_i = \xi_i$. Przyjmujemy, że liczby ξ_i, $\langle \xi_i \rangle = 0$, są liczbami losowymi, pochodzącymi z rozkładu normalnego (Gaussa). Oznaczmy wektor wszystkich błędów pomiarowych przez $\boldsymbol{\xi} = [\xi_1, \xi_2, \ldots, \xi_n]^T \in \mathbb{R}^n$. Dalej, przyjmijmy, że łącznie wszystkie błędy tworzą n-wymiarowy rozkład Gaussa o macierzy kowariancji $\mathbf{G}$:

$$\left\langle \boldsymbol{\xi}\boldsymbol{\xi}^T \right\rangle = \mathbf{G}\,, \tag{4}$$

gdzie $\langle \cdots \rangle$ oznacza średniowanie po realizacjach zmiennych losowych. Macierz $\mathbf{G}$ jest symetryczna i dodatnio określona.

Metoda najmniejszych kwadratów

Twierdzenie 1. *Jeżeli błędy pomiarowe pochodzą z rokładu Gaussa o macierzy kowariancji* $\mathbf{G}$*, estymator największej wiarygodności odpowiada minimum formy kwadratowej*

$$Q = \frac{1}{2}\boldsymbol{\xi}^T \mathbf{G}^{-1} \boldsymbol{\xi}\,. \tag{5}$$

Zauważmy, że ponieważ $\mathbf{G}$ jest symetryczna i dodatnio określona, także $\mathbf{G}^{-1}$ jest symetryczna i dodatnio określona, a zatem forma kwadratowa (5) z całą pewnością posiada minimum.

Obecność *odwrotności* macierzy kowarianci w wyrażeniu (5) oznacza, że pomiary obarczone większym błędem dają mniejszy wkład do Q.

Forma kwadratowa estymatorów

$$
\begin{aligned}
Q &= \frac{1}{2}\boldsymbol{\xi}^T \mathbf{G}^{-1} \boldsymbol{\xi} = \frac{1}{2}(\mathbf{y}-\tilde{\mathbf{y}})^T \mathbf{G}^{-1}(\mathbf{y}-\tilde{\mathbf{y}}) \\
&= \frac{1}{2}\left(\mathbf{y}-\mathbf{A}\mathbf{p}\right)^T \mathbf{G}^{-1}\left(\mathbf{y}-\mathbf{A}\mathbf{p}\right) \\
&= \frac{1}{2}\left[\mathbf{y}^T\mathbf{G}^{-1}\mathbf{y} - (\mathbf{A}\mathbf{p})^T\mathbf{G}^{-1}\mathbf{y} - \mathbf{y}^T\mathbf{G}^{-1}\mathbf{A}\mathbf{p} + (\mathbf{A}\mathbf{p})^T\mathbf{G}^{-1}\mathbf{A}\mathbf{p}\right] \\
&= \frac{1}{2}\mathbf{p}^T\mathbf{A}^T\mathbf{G}^{-1}\mathbf{A}\mathbf{p} - \mathbf{p}^T\mathbf{A}^T\mathbf{G}^{-1}\mathbf{y} + \underbrace{\frac{1}{2}\mathbf{y}^T\mathbf{G}^{-1}\mathbf{y}}_{=\text{const}} \qquad (6)
\end{aligned}
$$

W **liniowym** zagadnieniu najmniejszych kwadratów minimalizowana funkcja jest *formą kwadratową w parametrach*. Dzięki temu wiemy, że minimum istnieje i jest jednoznaczne. *Liniowość* oznacza tutaj, że funkcja "teoretyczna" zależy *liniowo* od parametrów, nie od argumentu!

Minimum formy kwadratowej

Aby znaleźć estymator, należy znaleźć taki wektor $\mathbf{p}$, że forma kwadratowa (6) przybiera najmniejszą możliwą wartość. Można to zrobić albo bezpośrednio, metodą zmiennej metryki lub gradientów sprzężonych, albo formalnie rozwiązując równanie $\nabla Q = 0$, gdzie różniczkujemy po składowych wektora $\mathbf{p}$. Otrzymujemy

$$\mathbf{A}^T \mathbf{G}^{-1} \mathbf{A} \mathbf{p} = \mathbf{A}^T \mathbf{G}^{-1} \mathbf{y} \tag{7}$$

Tego równanie nie można "uprościć" pozbywając się członu $\mathbf{A}^T \mathbf{G}^{-1}$, bo jest to macierz niekwadratowa, dla której nie da się zdefiniować odwrotności. Natomiast samo równanie (7) jest dobrze określone, gdyż macierz tego równania $\mathbf{A}^T \mathbf{G}^{-1} \mathbf{A}$ jest symetryczna i dodatnio określona.

Własności wektora estymatorów

Wektor $\mathbf{p}$ obliczamy z (7) dla takich wartości pomiarów, jakie faktycznie mamy. Tak obliczony wektor $\mathbf{p}$ jest wektorem estymatorów. Pamiętajmy, że pomiary są obarczone błędami losowymi, a więc także obliczone estymatory są, formalnie, *gaussowskimi liczbami losowymi*. Co można powiedzieć o tych liczbach? $\mathbf{y} = \mathbf{A}\mathbf{p}^* + \boldsymbol{\xi}$, gdzie $\mathbf{p}^*$ jest zbudowany z "prawdziwych", nieznanych wartości $[a_1, \ldots, a_s]^T$ z równania (1). Widać, że

$$\mathbf{A}^T\mathbf{G}^{-1}\mathbf{A}(\mathbf{p} - \mathbf{p}^*) = \mathbf{A}^T\mathbf{G}^{-1}\boldsymbol{\xi} \tag{8}$$

wobec czego

$$\langle \mathbf{p} \rangle = \mathbf{p}^* , \tag{9}$$

gdyż $\langle \boldsymbol{\xi} \rangle = 0$. Obliczone estymatory są przybliżeniem "prawdziwych" wartości parametrów w sensie równania (9).

Macierz kowariancji estymatorów wynosi

$$\mathbf{C_p} = \left\langle (\mathbf{p}-\mathbf{p}^*)(\mathbf{p}-\mathbf{p}^*)^T \right\rangle$$
$$= \left(\mathbf{A}^T\mathbf{G}^{-1}\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{G}^{-1}\left\langle \boldsymbol{\xi}\boldsymbol{\xi}^T\right\rangle \mathbf{G}^{-1}\mathbf{A}\left(\mathbf{A}^T\mathbf{G}^{-1}\mathbf{A}\right)^{-1}, \qquad (10)$$

gdzie skorzystaliśmy z symetrii macierzy $\mathbf{G}^{-1}$ i macierzy $\mathbf{A}^T\mathbf{G}^{-1}\mathbf{A}$. Korzystając z równania (4) otrzymujemy

$$\mathbf{C_p} = \left(\mathbf{A}^T\mathbf{G}^{-1}\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{G}^{-1}\overbrace{\underbrace{\mathbf{G}}_{\left\langle \boldsymbol{\xi}\boldsymbol{\xi}^T\right\rangle}\mathbf{G}^{-1}}^{\mathbb{I}}\mathbf{A}\left(\mathbf{A}^T\mathbf{G}^{-1}\mathbf{A}\right)^{-1},$$
$$= \left(\mathbf{A}^T\mathbf{G}^{-1}\mathbf{A}\right)^{-1}\underbrace{\mathbf{A}^T\mathbf{G}^{-1}\mathbf{A}\left(\mathbf{A}^T\mathbf{G}^{-1}\mathbf{A}\right)^{-1}}_{\mathbb{I}}$$
$$= \left(\mathbf{A}^T\mathbf{G}^{-1}\mathbf{A}\right)^{-1}. \qquad (11)$$

Nadokreślony układ równań

Zamiast minimalizować formę kwadratową (5), moglibyśmy *zażądać*, aby równanie $y_i = \tilde{y}_i$ było *ściśle* spełnione dla wszystkich punktów pomiarowych (x_i, y_i). Wobec równania (2a) oznacza to, że chcemy rozwiązać układ równan liniowych

$$\mathbf{A}\mathbf{p} = \mathbf{y}\,. \tag{12}$$

Jest to nadokreślony układ równań (s niewiadomych i n, $n > s$, równań) i, poza wyjątkowymi przypadkami, nie ma on ścisłego rozwiązania. Jak jednak wiemy z poprzednich wykładów, metoda *SVD* (*Singular Value Decomposition*) dostarcza przybliżonego rozwiązania takich układów, optymalnego w sensie najmniejszych kwadratów. Jeżeli macierz kowariancji $\mathbf{G}$ jest proporcjonalna do macierzy jednostkowej, $\mathbf{G} = \sigma^2\mathbb{I}$, co odpowiada pomiarom nieskorelowanym i obarczonym takimi samymi błędami i w praktyce zdarza się bardzo często, przybliżone rozwiązanie (12) uzyskane za pomocą *SVD* jest (w arytmetyce dokładnej) tym samym rozwiązaniem, które otrzymalibyśmy minimalizując formę kwadratową (6) lub rozwiązując układ równań (7).

Gdybyśmy, zamiast (12), zażądali spełnienia układu równań

$$\mathbf{G}^{-1}\mathbf{A}\mathbf{p} = \mathbf{G}^{-1}\mathbf{y}\,, \tag{13}$$

również nadokreślonego, ale uwzględniającego różne wagi poszczególnych pomiarów, rozwiązanie optymalne w sensie SVD byłoby równoważne rozwiązaniu równania (7).

Pomiary nieskorelowane

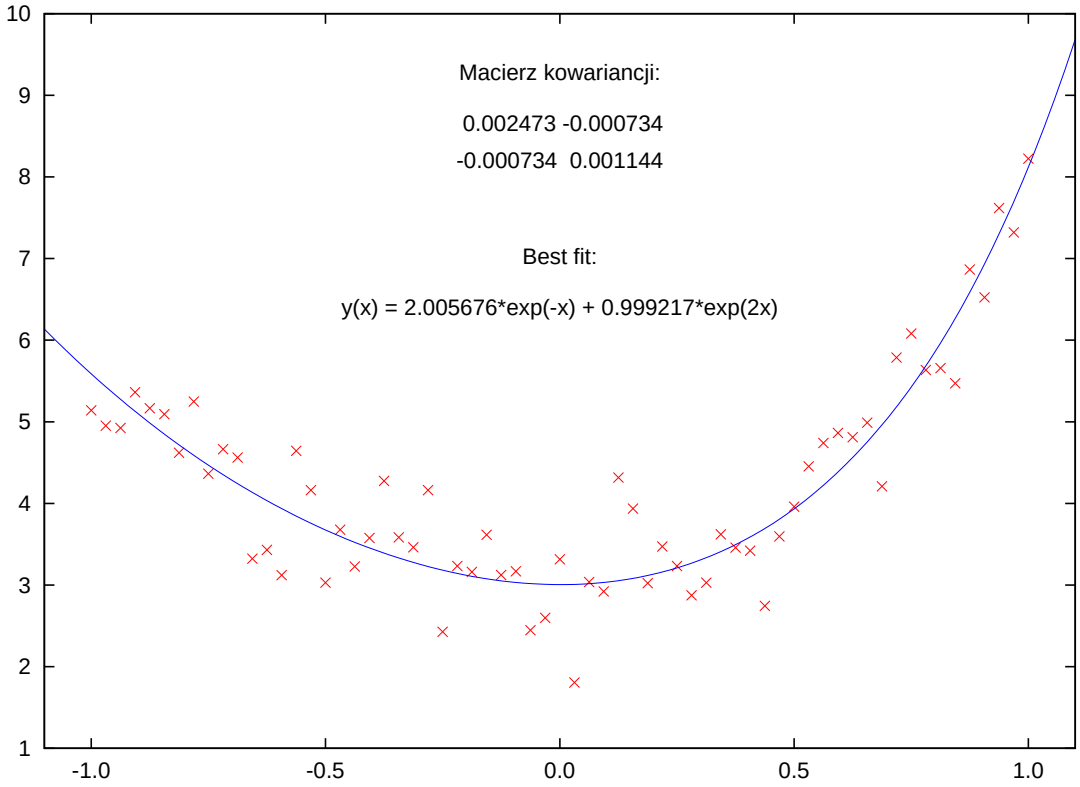
Na ogół (i na ogół z dobrym uzasadnieniem) zakłada się, że pomiary są niezależne, a ich wyniki nieskorelowane. Wówczas elementy pozadiagonalne macierzy $\mathbf{G}$ znikają, $\mathbf{G} = \mathrm{diag}\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2\}$. Minimalizowana forma kwadratowa (5) upraszcza się do

$$Q = \sum_{i=1}^{n} \frac{(a_1 f_1(x_i) + a_2 f_2(x_i) + \cdots + a_s f_s(x_i) - y_i)^2}{\sigma_i^2} \,. \tag{14}$$

W dość częstym przypadku pomiarów nieskorelowanych i identycznych $\forall i = 1, \ldots, n \colon \sigma_i^2 = \sigma^2$, a zatem $\mathbf{G} = \sigma^2 \mathbb{I}$. W tym wypadku estymatory nie zależą od macierzy kowariancji pomiarów, gdyż macierz $\mathbf{G}$ wypada z równania (7), natomiast macierz kowariancji estymatorów upraszcza się do

$$\mathbf{C_p} = \sigma^2 \left(\mathbf{A}^T \mathbf{A} \right)^{-1} \,. \tag{15}$$

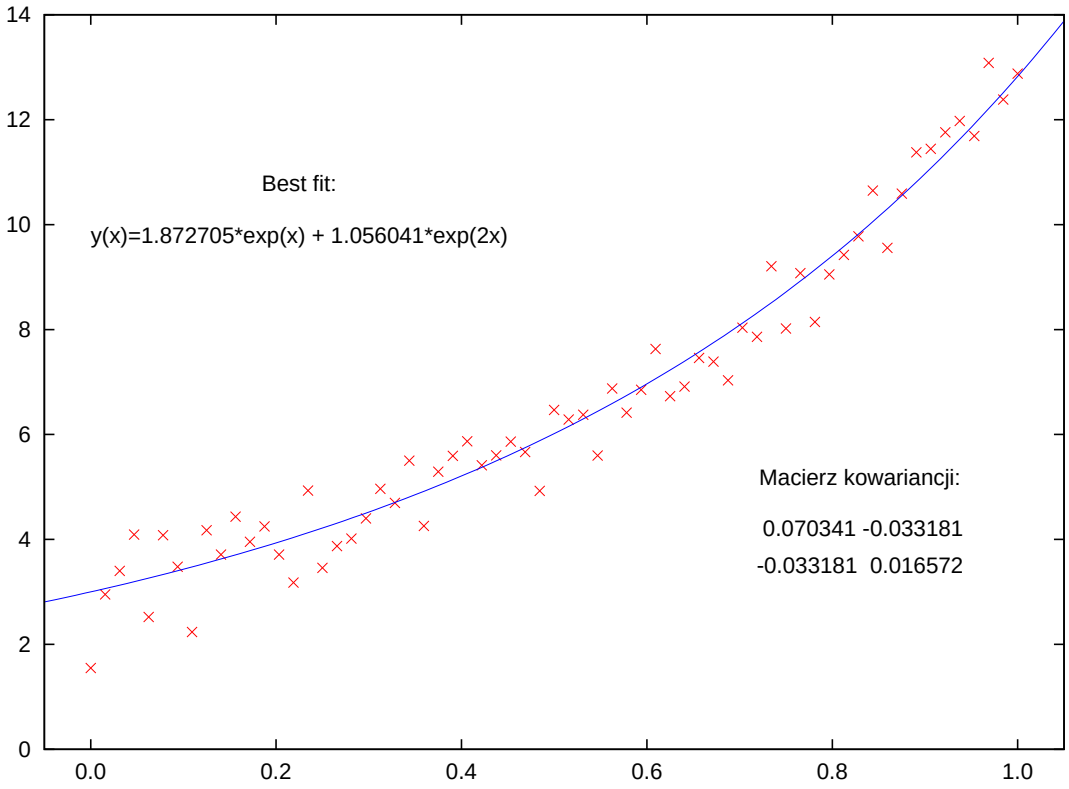
Przykład



Do zaznaczonych punktów dopasowano krzywą $y = ae^{-x} + be^{2x}$ za pomocą *liniowej* metody najmniejszych kwadratów. Przyjęto, że pomiary są identyczne i nieskorelowane, o stałym błędzie $\sigma^2 = 0.305226$.

W powyższym przykładzie współczynniki korelacji estymatorów są bardzo małe, gdyż, efektywnie, obie funkcje dopasowują się do innych zakresów danych (funkcja e^{2x} jest mała dla $x < 0$, funkcja e^{-x} jest mała dla $x > 0$). Jeżeli rózne funkcje bazowe "konkurują" o te same dane, współczynnik korelacji jest, co do wartości bezwzględnej, większy. Ujemny współczynnik korelacji pomiędzy estymatorami oznacza, że *prawie* tak samo dobre dopasowanie można uzyskać zmniejszając jeden, zwiększając zaś drugi.

Przykład



Do zaznaczonych punktów dopasowano krzywą $y = ae^x + be^{2x}$ za pomocą *liniowej* metody najmniejszych kwadratów. Przyjęto, że pomiary są identyczne i nieskorelowane, o stałym błędzie $\sigma^2 = 0.8152$.

Kryterium Akaike

Czasami nie wiadomo ile funkcji bazowych $f_i(x)$ nalezy uwzględnić w dopasowaniu, czyli we wzorze (1). W szczególności, jeśli do danych doświadczalnych dopasowujemy wielomian, niekiedy — jeśli nie mamy dobrego modelu teoretycznego — nie wiemy, jaki stopień wielomianu wybrać. Jest jasne, że im wyższy stopień wielomianu, tym dopasowanie będzie "lepsze" (wielomian interpolacyjny będzie przechodził dokładnie przez wszystkie punkty!), ale zawsze staramy się dobrać model o jak najmniejszej liczbie parametrów.

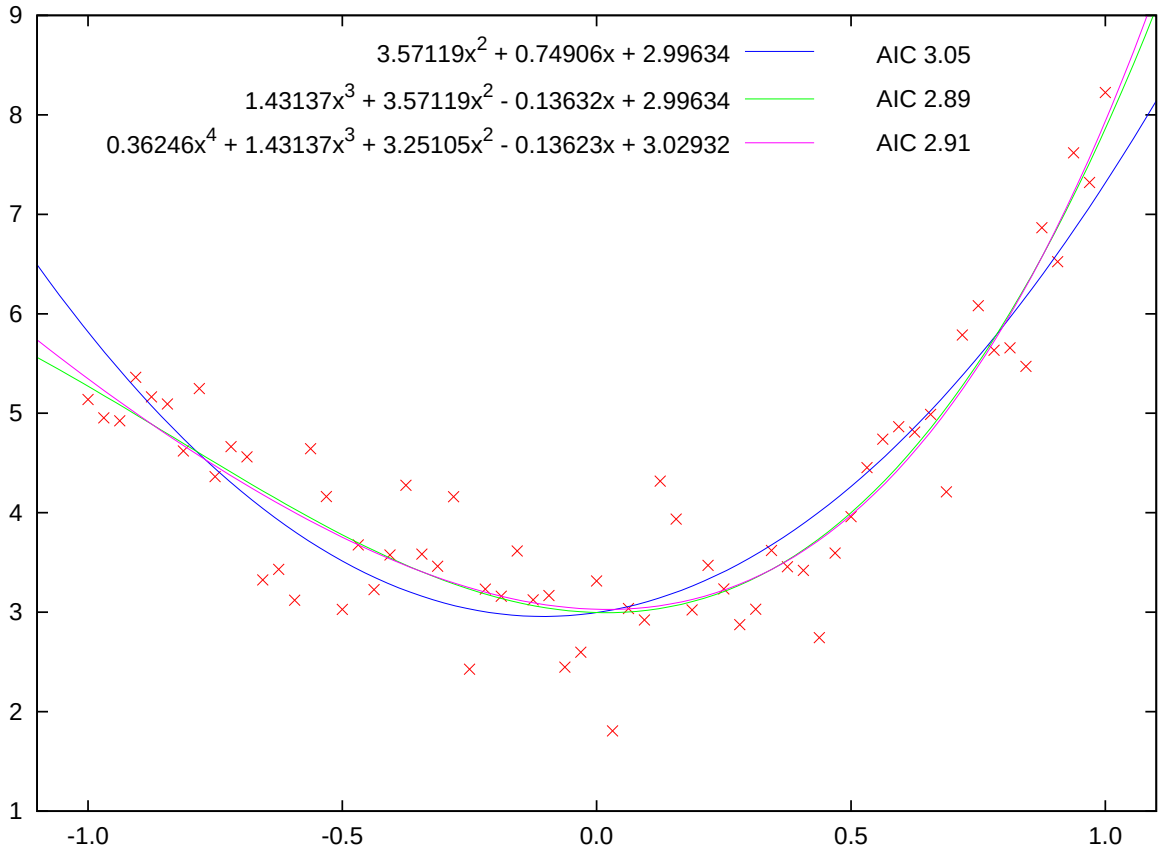
Jak zbalansować jak najlepsze dopasowanie z postulatem jak najmniejszej liczby parametrów?

Hirotugu Akaike zaproponował kryterium, które nagradza za jak najlepsze dopasowanie, ale karze za zbyt wiele parametrów: Należy zminimalizować wielkość

$$AIC = \ln Q + \frac{2s}{N}\,, \tag{16}$$

gdzie Q jest wartością minimaliowanej formy kwadratowej (6) w minimum, zwaną błędem rezydualnym, s liczbą parametrów, N liczbą punktów, do których dopasowujemy. *AIC* jest akronimem od *Akaike Information Criterion*.

Przykład



Wielomiany drugiego, trzeciego i czwartego stopnia dopasowane do tych samych danych

Nieliniowe zagadnienie najmniejszych kwadratów

Przypuśćmy, że dopasowywana do danych pomiarowych zależność teoretyczna zależy od parametrów w sposób nieliniowy,

$$\tilde{y}_i = f(x_i; \mathbf{p}) \tag{17}$$

gdzie $\mathbf{p} \in \mathbb{R}^s$ jest wektorem parametrów. Zakładamy, że $f(\cdot; \mathbf{p})$ jest *znaną* funkcją, a tylko jej parametry są nieznane. Na przykład do danych doświadczalnych dopasowujemy funkcję Gaussa

$$y(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\bar{x})^2}{2\sigma^2}\right). \tag{18}$$

Parametrami będą w tym wypadku $\bar{x}$ oraz σ^2. Widać, że funkcja (18) zależy od nich *nieliniowo*.

Zakładamy, że błędy pomiarowe są gaussowskie, o macierzy kowariancji $\mathbf{G}$. Wówczas tworzymy wektor $\mathbf{u} = [u_1, u_2, \ldots, u_N]^T \in \mathbb{R}^N$, gdzie $u_i = y_i - \tilde{y}_i = y_i - f(x_i; \mathbf{p})$. Żądamy, aby funkcja

$$Q = \frac{1}{2}\mathbf{u}^T \mathbf{G}^{-1} \mathbf{u} \tag{19a}$$

osiągała minimum jako funkcja parametrów $\mathbf{p}$.

W najczęstszym przypadku pomiarów nieskorelowanych, obarczonych identycznymi błędami, funkcja (19a) redukuje się do postaci

$$Q = \text{const} \cdot \frac{1}{2} \sum_{i=1}^{N} \left(y_i - f(x_i; \mathbf{p}) \right)^2 . \tag{19b}$$

Ani funkcja (19a), ani jej szczególna postać (19b), nie są formami kwadratowymi w parametrach!

Dodatnia określoność funckji Q, $Q \geqslant 0$, *w praktyce* gwarantuje istnienie minimum. Nie da się jednak zagwarantować, że minimum jest tylko jedno.

Poza **bardzo nielicznymi** przypadkami, w których *łatwo* można rozwiązać układ równań $\nabla_{\mathbf{p}} Q = 0$, minimum funkcji $Q(\mathbf{p})$ należy znaleźć **numerycznie**, przy pomocy metody Levenberga-Marquardta.

Pseudolinearyzacja

Czasami do znalezienia minimum Q stosuje się metodę *pseudolinearyzacji*. Przypuśćmy, że $\mathbf{p}_n$ jest aktualnym przybliżeniem poszukiwanej wartości parametrów $\mathbf{p}$. Stawiamy hipotezę, iż "prawdziwe" wartości parametrów są małą poprawką w stosunku do $\mathbf{p}_n$: $\mathbf{p} \simeq \mathbf{p}_n + \boldsymbol{\delta}\mathbf{p}$ i rozwijamy (17) w szereg Taylora do pierwszego rzędu:

$$\tilde{y}_i = f(x_i; \mathbf{p}_n + \boldsymbol{\delta}\mathbf{p}) \simeq f(x_i; \mathbf{p}_n) + \left[\nabla_{\mathbf{p}} f\big|_{\mathbf{p}_n}\right]^T \boldsymbol{\delta}\mathbf{p}\,. \tag{20}$$

Podstawiamy to rozwinięcie (dla uproszczenia zakładamy nieskorelowane, identyczne pomiary) do (19b).

Funkcja

$$Q = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - f(x_i;\mathbf{p}_n) - \left[\nabla_{\mathbf{p}} f|_{\mathbf{p}_n}\right]^T \boldsymbol{\delta}\mathbf{p}\right)^2 \tag{21}$$

jest formą kwadratową w poprawkach $\boldsymbol{\delta}\mathbf{p}$. Po znalezieniu znanymi metodami wartości $\boldsymbol{\delta}\mathbf{p}_{\text{min}}$, odpowiadających (jedynemu) minimum (21), podstawiamy $\mathbf{p}_{n+1} = \mathbf{p}_n + \boldsymbol{\delta}\mathbf{p}_{\text{min}}$ i powtarzamy całą procedurę.

Taka procedura dość dobrze działa w wypadku nieliniowej metody najmniejszych kwadratów, choć nie należy jej polecać jako ogólnej metody minimalizacji. Pseudolinearyzacja ma tylko jedno niewątpliwe zastosowanie: Po znalezieniu ostatecznych wartości minimalizujących funkcję (19a), za pomocą pseudolinearyzacji wokół tego punktu znajdujemy macierz kowariancji estymatorów, będącą charaktestystyką *liniową*.

Przykład

Przypuśmy, że do danych dopasowujemy funkcję Gaussa (18), zaś aktualnymi przybliżeniami parametrów są $\bar{x}_n$, σ_n^2. Obliczamy

$$f(x;\bar{x},\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\bar{x})^2}{2\sigma^2}\right), \tag{22a}$$

$$\left.\frac{\partial f}{\partial \bar{x}}\right|_{\bar{x}_n,\sigma_n^2} = \frac{1}{\sqrt{2\pi\sigma_n^2}}\exp\left(-\frac{(x-\bar{x}_n)^2}{2\sigma_n^2}\right)\cdot\frac{x-\bar{x}_n}{\sigma_n^2}, \tag{22b}$$

$$\left.\frac{\partial f}{\partial (\sigma^2)}\right|_{\bar{x}_n,\sigma_n^2} = \frac{1}{\sqrt{2\pi\sigma_n^2}}\exp\left(-\frac{(x-\bar{x}_n)^2}{2\sigma_n^2}\right)\cdot\left[\frac{(x-\bar{x}_n)^2}{4\left(\sigma_n^2\right)^2}-\frac{1}{2\sigma_n^2}\right] \tag{22c}$$

Wyrażenie

$$Q = \frac{1}{2}\sum_{i=1}^{N}\left(y_i - \frac{1}{\sqrt{2\pi\sigma_n^2}}\exp\left(-\frac{(x_i-\bar{x}_n)^2}{2\sigma_n^2}\right)\right.$$
$$\left.\cdot\left\{1 + \frac{x_i-\bar{x}_n}{\sigma_n^2}\,\delta\bar{x} + \left[\frac{(x_i-\bar{x}_n)^2}{4\left(\sigma_n^2\right)^2} - \frac{1}{2\sigma_n^2}\right]\delta\sigma^2\right\}\right)^2 \qquad (23)$$

jest formą kwadratową w zmiennych $\delta\bar{x}$, $\delta\sigma^2$.

Z punktu widzenia Big Data...

W ciągu ostatnich kilkunastu lat rozmiary dostępnych zbiorów danych rosną *szybciej* niż rośnie prędkość procesorów (nawet po uwzględnieniu paralelizacji i obliczeń na GPU). Z tego punktu widzenia możliwości uczenia maszynowego są ograniczone raczej przez możliwości obliczeniowe niż przez dostępne zbiory danych. Tę klasę problemów zwyczajowo określa się jako *Big Data*.

Uczenie maszynowe bardzo często oznacza konieczność minimalizowania funkcji

$$Q(\mathbf{p}) = \frac{1}{N}\sum_{i=1}^{N} Q_i(\mathbf{p}; x_i, y_i) \tag{24}$$

gdzie

- $\forall i: Q_i \geqslant 0$,
- Q_i mają taką samą postać funkcyjną, różnią się tylko elementami x_i, y_i (elementami zbioru uczącego),
- $N \gg 1$ (N jest zazwyczaj **BARDZO** duże, rzędu kilkudziesięciu, niekiedy kilkuset tysięcy).

$\mathbf{p}$ jest wektorem estymatorów, których wartości chcemy znaleźć. Tego typu problemy pojawiają się w zagadnieniu najmniejszych kwadratów.

Problem (24) można rozwiązać za pomocą którejś z już omówionych metod (minimalizacja funkcji wielu zmiennych lub rozwiązywanie równań typu (7), jeżeli mowa o liniowym zagadnieniu najmniejszych kwadratów). Jednak dla bardzo dużych N, czyli dla bardzo dużych zbiorów uczących, inne podejście może być bardziej efektywne.

Punktem wyjścia jest metoda najszybszego spadku (ang. *steepest gradient descent*): W każdym kroku podążamy w kierunku minus gradientu funkcji $Q(\mathbf{p})$, aż do osiągnięcia minimum kierunkowego.

Jak jednak pamiętamy, daleko od minimum nie minimalizujemy, tylko podążamy z arbitralnie dobranym krokiem w kierunku ujemnego gradientu. Dlatego zamiast minimalizować, w n-tym kroku iteracji przyjmujemy

$$\begin{aligned} \mathbf{p}_{n+1} &= \mathbf{p}_n - \gamma \cdot \nabla_{\mathbf{p}} Q(\mathbf{p})|_{\mathbf{p}_n} \\ &= \mathbf{p}_n - \gamma \cdot \frac{1}{N} \sum_{i=1}^{N} \nabla_{\mathbf{p}} Q_i(\mathbf{p}; x_i, y_i)|_{\mathbf{p}_n} \ . \end{aligned} \tag{25}$$

$\gamma =$ const i jest nazywane *prędkością uczenia* (ang. *learning rate*).

Uwaga: Przyjęcie stałego kroku nie jest aż tak naiwne, jak by się to mogło wydawać. Jeśli rozpatrzymy układ równań różniczkowych typu

$$\frac{d\mathbf{y}}{dx} = \mathbf{f}(\mathbf{y}) \tag{26}$$

to najprostszą (zdecydowanie nie najlepszą, ale najprostszą i najszybszą) metodą jego numerycznego rozwiązywania jest *metoda Eulera*

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h \cdot \mathbf{f}(\mathbf{y}_n)\,, \tag{27}$$

gdzie h jest stałym krokiem narastania zmiennej niezależnej x. Metoda najszybszego spadku (25) ze stałym krokiem γ ma taką samą postać, co metoda Eulera, po utożsamieniu $f \equiv -\nabla Q$.

Stochastic Gradient Descent

Jeśli w (25) $N \gg 1$, czas obliczania sumy gradientów cząstkowych $\sum_i^N \nabla Q_i$ może być bardzo znaczny. Zarazem wyrażenie $\frac{1}{N}\sum_i^N \nabla Q_i$ ma postać średniego gradientu cząstkowego. Jeśli założymy, że poszczególne elementy tej sumy nie odbiegają zbytnio od średniej, możemy średni, kosztowny w wyliczaniu gradient, zastąpić *losowo wybranym* gradientem cząstkowym. Otrzymujemy zatem algorytm *Stochastic Gradient Descent*:

- wybierz losowo $k \in \{1, 2, \ldots, N\}$
- przyjmij

$$\mathbf{p}_{n+1} = \mathbf{p}_n - \gamma \, \nabla_{\mathbf{p}} Q_k(\mathbf{p}; x_k, y_k)|_{\mathbf{p}_n} \tag{28}$$

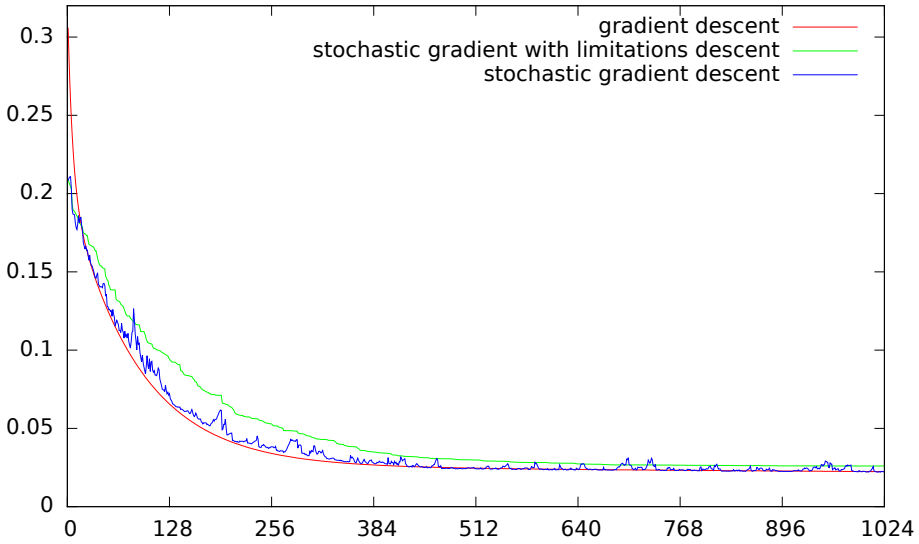
- Dodatkowo akceptujemy powyższy krok tylko jeżeli $Q(\mathbf{p}_{n+1}) < Q(\mathbf{p}_n)$. Jeśli nie, nie wykonujemy kroku, ale losujemy nowe k. Ten wariant nazwiemy *Stochastic Gradient Descent z ograniczeniem*.

Iterację kończymy albo po wykonaniu z góry określonej liczby kroków, albo — częściej — gdy różnica $\left|Q(\mathbf{p}_n) - Q(\mathbf{p}_{n+1})\right| < \varepsilon$, gdzie ε jest ustaloną tolerancją.

Wartość *prędkości uczenia* γ oraz tolerancję ε dobieramy "eksperymentalnie", to znaczy kierując się znajomością natury problemu i doświadczeniem uzyskanym przy rozwiązywaniu podobnych problemów.

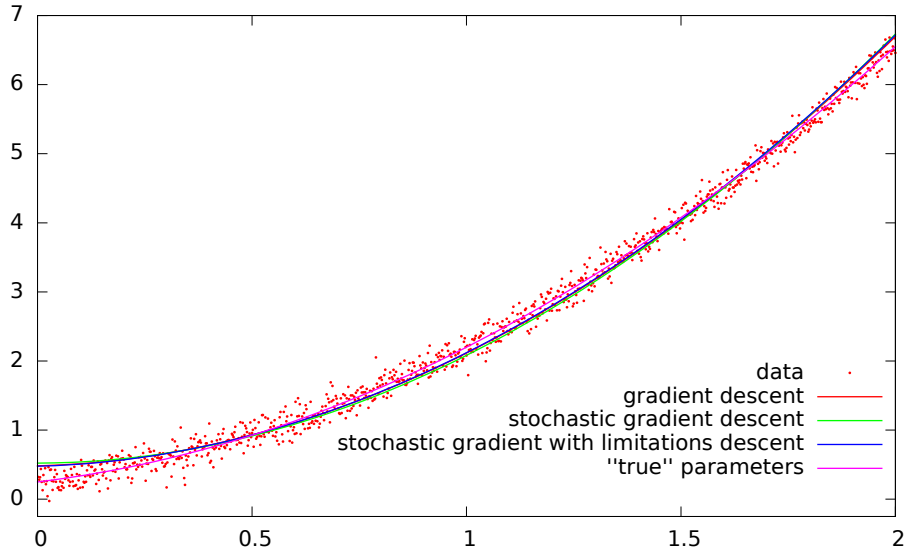
Okazuje się, że metoda *Stochastic Gradient Descent* działa zdumiewająco dobrze dla bardzo dużych zbiorów uczących. Co ciekawe, przyjęcie warunku, iż wartość funkcji Q musi maleć po wykonaniu kroku na ogół *pogarsza* wyniki: Od czasu do czasu można/trzeba wykonać krok "w złą stronę".

Przykład



Wygenerowano $N = 1024$ punktów $y_i = ax_i^2 + bx_i + c + \xi_i$, gdzie $\{\xi\}_{i=1}^N$ są liczbami o rozkładzie normalnym, natomiast $x_i \in [0, 2]$. Rysunek przedstawia kolejne wartości funkcji $Q(a, b, c)$ uzyskane w metodzie najszybszego spadku (czerwony), w metodzie *Stochastic Gradient Descent* (niebieski) oraz *Stochastic Gradient Descent* z ograniczeniami (zielony). $\gamma = 1/128$. W przypadku *Stochastic Gradient Descent* wartość minimalizowanej funkcji niekiedy rośnie, ale po dostatecznie dużej liczbie iteracji wyniki

tej metody są nieco *lepsze*, niż dla metody z ograniczeniem. Co ważniejsze, wyniki są porównywalne z wynikami uzyskanymi w obliczeniowo droższej metodzie najszybszego spadku.



Dopasowane krzywe są niemalże nieodróżnialne w obszarze odpowiadającym danym wejściowym (zawartości zbioru uczącego).