Bootcamp Data Science

Przemysław Spurek

$$\mathbb{R}^n$$

Korelacja

Definicja

Kowariancją całkowalnych zmiennych losowych X i Y, spełniających warunek $E|XY| < \infty$, nazywamy wielkość:

$$\operatorname{cov}(X, Y) = E[(X - EX)(Y - EY)] = E(XY) - EXEY.$$

Definicja

Współczynnik korelacji:

$$\operatorname{cor}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{D^2X \cdot D^2Y}}$$

Korelacja

Dla dwóch powiązanych zmiennych korelacja mierzy zależność między nimi.

Dla dwóch powiązanych zmiennych regresja liniowa służy do przewidywania wartości jednej zmiennej.

Korelacja

Współczynnik korelacji między dwiema zmiennymi odpowiada na pytanie:

"Czy dwie zmienne są powiązane?"

Oznacza to, że jeśli jedna się zmienia, to druga też? Jeżeli dwie próbki pochodzą z rozkładu normalnego, to miarą korelacji jest korelacja Pearsona:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Korelacja

Można ją zapisać również w postaci:

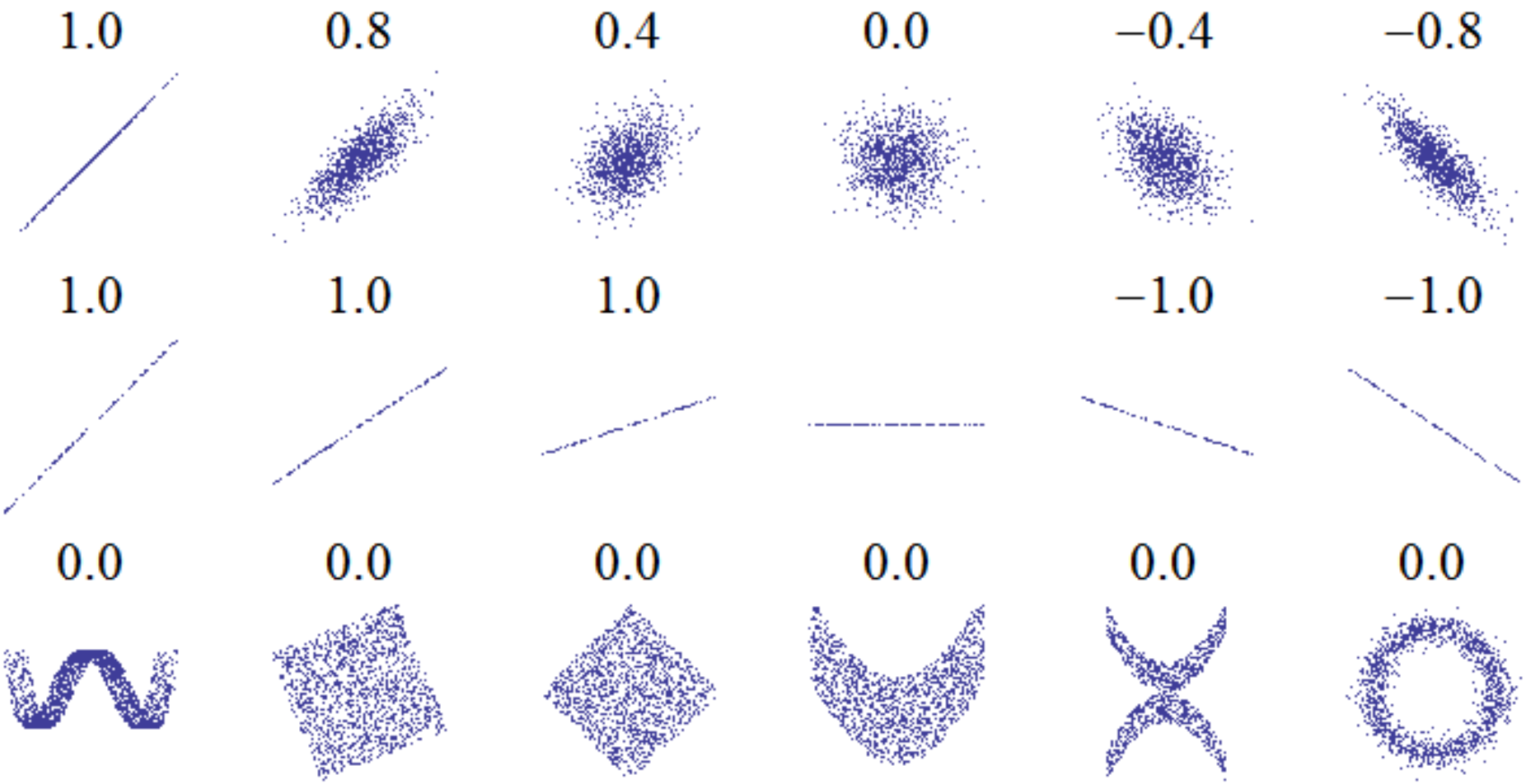
$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

gdzie s_{xy} oznacza kowariancję danych:

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

a s_x i s_y to odchylenia standardowe.

Korelacja



Współczynnik korelacji Pearsona

Współczynnik korelacji Pearsona - współczynnik określający poziom zależności liniowej między zmiennymi losowymi.

Korelacja rang Spearmana

Korelacja rang Spearmana (lub: korelacja rangowa Spearmana, rho Spearmana) jedna z nieparametrycznych miar monotonicznej zależności statystycznej między zmiennymi losowymi.

Korelacja rangowa przyjmuje zawsze wartości z przedziału $[-1, +1]$. Ich interpretacja jest podobna do klasycznego współczynnika korelacji Pearsona, z jednym zastrzeżeniem: w odróżnieniu od współczynnika Pearsona, który mierzy liniową zależność między zmiennymi, a wszelkie inne związki traktuje jak zaburzone zależności liniowe, korelacja rangowa pokazuje dowolną monotoniczną zależność (także nieliniową).

Korelacja Tau Kendalla

Tau Kendalla – statystyka będąca jedną z miar monotonicznej zależności dwóch zmiennych losowych. Służy w praktyce do opisu korelacji między zmiennymi porządkowymi.

Tau Kendalla przyjmuje wartości od -1 do 1 włącznie. +1 oznacza, że każda ze zmiennych rośnie przy wzroście drugiej. -1 oznacza, że każda maleje przy wzroście drugiej. Tym samym tau Kendalla, podobnie jak korelacja rangowa jest miarą monotonicznej zależności zmiennych losowych.

Korelacja

Zadanie w Jupyter

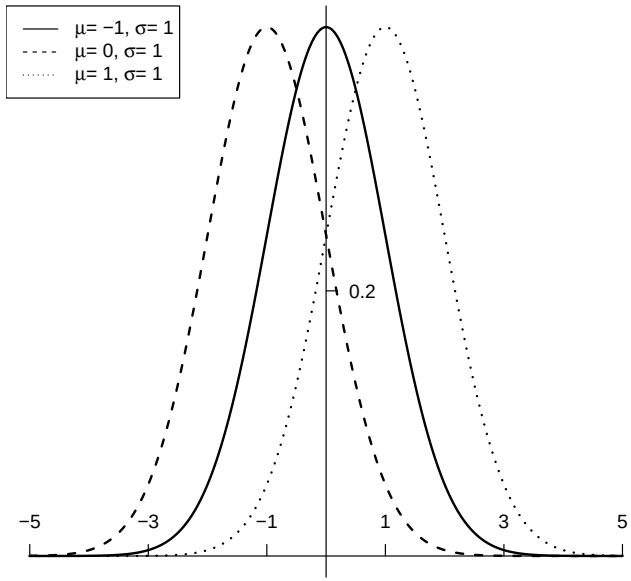
https://github.com/przem85/statistic_4/blob/master/D09_Z02_correlation.ipynb

https://github.com/przem85/statistic_4/blob/master/D09_Z05_correlation.ipynb

Wielowymiarowy rozkład normalny

Rozkład normalny

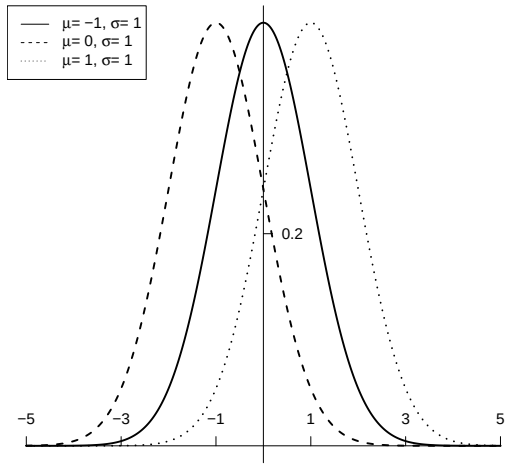
$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$$



Rozkład normalny

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$$

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\mu)(\sigma^2)^{-1}(x-\mu)\right), \quad x \in \mathbb{R}$$



N-wymiarowy rozkład normalny

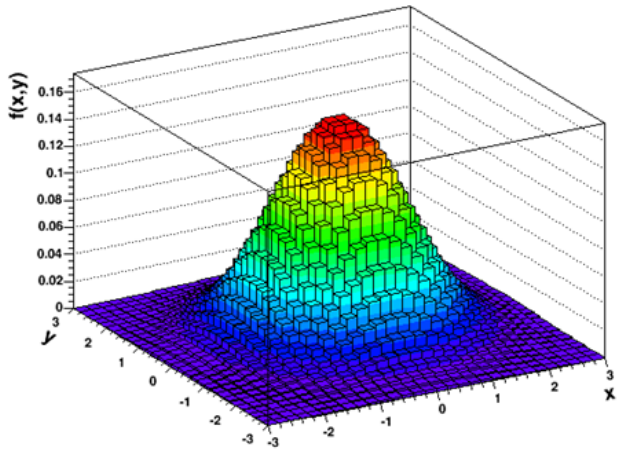
N-wymiarowy rozkład normalny dla macierzy kowariancji Σ – symetryczna dodatnia określona oraz średniej μ ma gęstość:

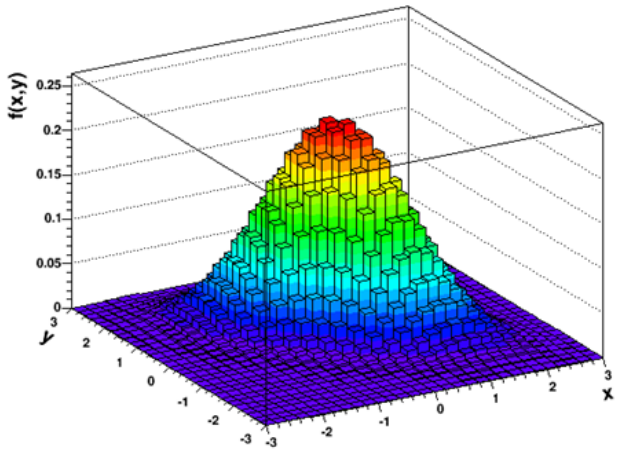
$$f(x_1, \ldots, x_n, \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right).$$

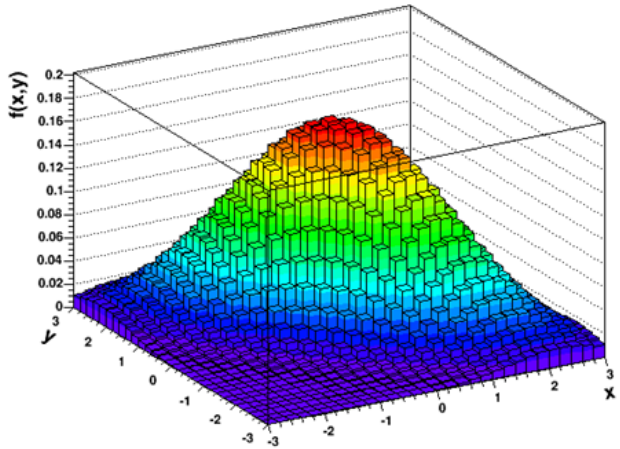
DODATKOWE:

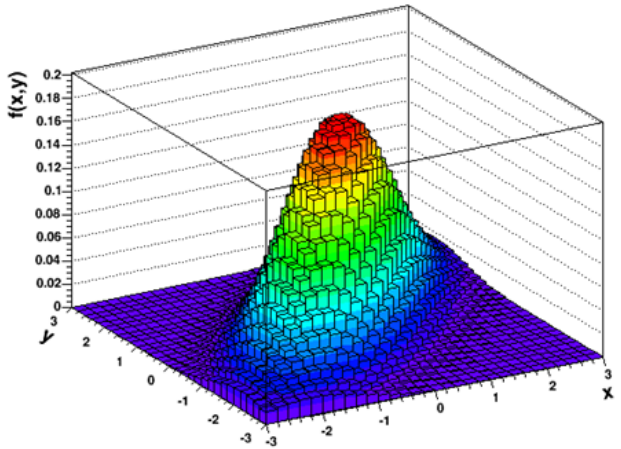
`https://github.com/przem85/statistic_4/blob/master/D09_Z03_normal_distribution_2D.ipynb`

2-wymiarowy rozkład normalny

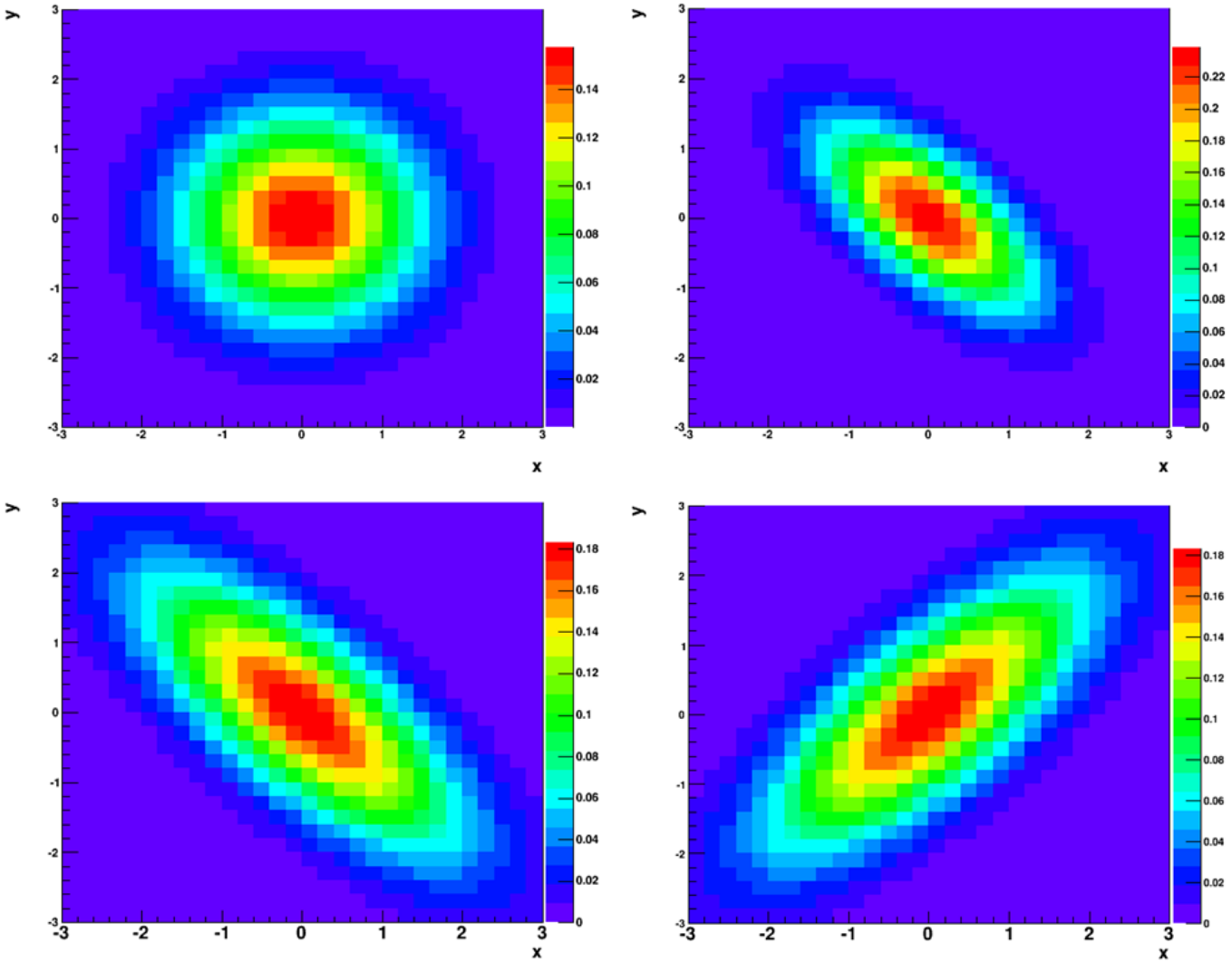








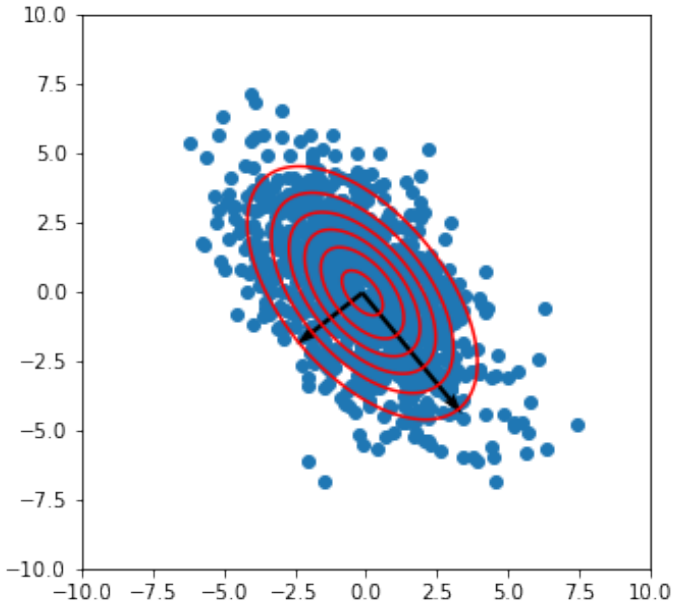
2-wymiarowy rozkład normalny



2-wymiarowy rozkład normalny

DODATKOWE:

https://github.com/przem85/statistic_4/blob/master/D09_Z04_normal_distribution_2D.ipynb



Modelowanie statystyczne

Modelowanie statystyczne

Choć testowanie hipotez może rozstrzygnąć problem, czy dwie lub więcej próbek danych pochodzi z tej samej populacji czy z różnych, nie potrafią określić zależności/relacji między nimi. Pokrewnym zagadnieniem jest przewidywanie ilościowe zmiennych.

Regresja liniowa

Istnieje znaczna różnica pomiędzy testowaniem hipotez, a modelowaniem statystycznym.

Testowanie hipotez zaczyna się od postawienia hipotezy zerowej. Opierając się na pytaniu i danych, wybieramy odpowiedni test statystyczny, a także pożądany poziom istotności i akceptujemy lub odrzucamy hipotezę zerową.

Natomiast modelowanie statystyczne zazwyczaj wymaga bardziej interaktywnej analizy danych. Rozpoczyna się wizualna analiza danych, szukając korelacji i/lub relacji między zmiennymi. Na podstawie tego pierwszego spojrzenia wybieramy model statystyczny.

Regresja

czyli znamy przykładowe wartości (x_i, y_i),
ktoś nam podaje nowy punkt x_0
i
chcemy przewidzieć wartość y_0.

Metoda Najmniejszych Kwadratów

W roku 1801 astronomowie zgubili z oczu asteroidę i chodziło o to, by odszukać ją z powrotem na niebie.
Gauss stworzył MNK (Metoda Najmniejszych Kwadratów) właśnie w celu jej odszukania, co mu się udało – znalazła się dokładnie tam, gdzie Gauss przewidział, że będzie.

Regresja liniowa

Dokonaliśmy pomiarów pewnej funkcji:

x_i	1	2	4
y_i	1	2	3

.

Podejrzewamy, że dane mogą być dobrze przybliżone za pomocą funkcji liniowej:

$$y = ax + b.$$

W związku z tym szukamy takich parametrów $a, b \in \mathbb{R}$ aby przybliżenie:

$$y \approx ax + b \text{ dla } i = 1, \ldots, n.$$

(gdzie w naszym przypadku $n = 3$) było optymalne.

Regresja liniowa

Inaczej mówiąc szukamy takich a, b aby:

$$\begin{cases} 1a + b \approx 1 \\ 2a + b \approx 2 \\ 4a + b \approx 3 \end{cases}$$

co w zapisie macierzowym możemy przedstawić:

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} \approx \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

Regresja liniowa

Pojawia się problem co to znaczy optymalne, oraz (gdy to już doprecyzujemy) jak to optymalnie znaleźć. Precyzyjniej mówiąc, potrzebujemy dookreślić jaką funkcję kosztu tu dopasujemy, która będzie mówiła ile "kosztuje" nas dany błąd (w zależności od doboru parametrów a, b).

Regresja liniowa

Pojawia się problem co to znaczy optymalne, oraz (gdy to już doprecyzujemy) jak to optymalnie znaleźć. Precyzyjniej mówiąc, potrzebujemy dookreślić jaką funkcję kosztu tu dopasujemy, która będzie mówiła ile "kosztuje" nas dany błąd (w zależności od doboru parametrów a, b).

Naturalnym wydawałoby się posumowanie modułów błędów:

$$error(a, b) = \sum |y_i - (ax_i + b)|.$$

Regresja liniowa

Pojawia się problem co to znaczy optymalne, oraz (gdy to już doprecyzujemy) jak to optymalnie znaleźć. Precyzyjniej mówiąc, potrzebujemy dookreślić jaką funkcję kosztu tu dopasujemy, która będzie mówiła ile "kosztuje" nas dany błąd (w zależności od doboru parametrów a, b).

Naturalnym wydawałoby się posumowanie modułów błędów:

$$error(a, b) = \sum |y_i - (ax_i + b)|.$$

Tak się czasami robi, ale takie podejście ma wadę, bo nie da się tych współczynników wyliczyć jawnym wzorem. W związku z tym, dokonamy naturalnej modyfikacji, zastępując moduł kwadratem (Square Error):

$$se(a, b) = \sum (y_i - (ax_i + b))^2.$$

Regresja liniowa

Można pokazać, że:

$$se(a, b) = \sum (y_i - (ax_i + b))^2.$$

przyjmuje minimum w:

$$a = \frac{\sum y_i x_i - b \sum x_i}{\sum x_i^2}$$

oraz

$$b = \frac{\sum y_i - a \sum x_i}{n}.$$

Regresja liniowa

Można pokazać, że:

$$se(a,b) = \sum (y_i - (ax_i + b))^2.$$

przyjmuje minimum w:

$$a = \frac{\sum y_i x_i - b \sum x_i}{\sum x_i^2}$$

oraz

$$b = \frac{\sum y_i - a \sum x_i}{n}.$$

W naszym przypadku otrzymujemy układ:

$$a = \frac{17 - 7b}{21}, \quad b = \frac{6 - 7b}{3}.$$

i wynik: $a = 9/10$, $b = 1/2$.

Regresja liniowa

Twierdzenie

Niech $A \in \mathbb{R}^{N \times K}$ będzie macierzą, która przekształca $\mathbb{R}^K$ w $\mathbb{R}^N$, i niech $y \in \mathbb{R}^K$. Wtedy punkt $x_0 \in \mathbb{R}^K$ spełnia:

$$x_0 = \operatorname{argmin}\{x \in \mathbb{R}^K : \|Ax - y\|^2\}$$

wtedy i tylko wtedy gdy:

$$A^T A x_0 = A^T y.$$

Regresja liniowa

Wróćmy do naszego przykładu:

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} \approx \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

Aby znaleźć optymalne przybliżenie, mnożymy obie strony przez A^T dostając równanie:

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

czyli rozwiązujemy:

$$\begin{bmatrix} 3 & 7 \\ 7 & 21 \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 6 \\ 17 \end{bmatrix}$$

i wynik $a = 9/10,\ b = 1/2$.

Regresja liniowa

Zadanie w Jupyter

https://github.com/przem85/statistic_4/blob/master/D09_Z01_cost_function_of_linear_regression.ipynb