

Bootcamp Data Science

Zajęcia 3

Przemysław Spurek

$$\mathbb{R}^n$$

Definicja

Rozkładem prawdopodobieństwa zmiennej losowej X o wartościach w \mathbb{R}^n nazywamy rozkład prawdopodobieństwa μ_X , określony na $\mathcal{B}(\mathbb{R}^n)$ zależnością:

$$\mu_X(B) = P(X^{-1}(B)), \quad B \in \mathcal{B}(\mathbb{R}^n).$$

Uwaga

W naszych rozważaniach będziemy używać dwuwymiarowej zmiennej losowej:

$$(X, Y): \Omega \times \Omega \rightarrow \mathbb{R}^2.$$

Definicja

Dwuwymiarową zmienną losową nazywamy parę zmiennych losowych (X, Y) określonych na tej samej przestrzeni zdarzeń elementarnych Ω i przyjmujących wartości rzeczywiste. Symbolicznie:

$$(X, Y): \Omega \times \Omega \rightarrow (-\infty, \infty) \times (-\infty, \infty).$$

Definicja

Łącznym rozkładem prawdopodobieństwa (lub rozkładem łącznym) pary zmiennych losowych (X, Y) określonych na tej samej przestrzeni zdarzeń elementarnych nazywamy przyporządkowanie:

$$A \rightarrow P((X, Y) \in A),$$

gdzie A - dowolny podzbiór zbioru par wartości zmiennych X, Y .

Definicja

Dystrybuantą zmiennej losowej (X,Y) nazywamy funkcję:

$$F(x,y) = P(X \leq x, Y \leq y),$$

gdzie $-\infty < x < \infty, -\infty < y < \infty$.

Definicja

Funkcją prawdopodobieństwa (łącznego) dwuwymiarowej zmiennej losowej dyskretnej nazywamy funkcję:

$$f(x, y) = P(X = x, Y = y).$$

Własność

Funkcja prawdopodobieństwa f oraz jej związek z dystrybuantą:

- ❶ $f(x, y) \geq 0$, dla dowolnej pary wartości (x, y) ,
- ❷ $\sum_x \sum_y f(x, y) = 1$, gdzie sumowanie odbywa się po wszelkich możliwych parach wartości zmiennych X i Y odpowiednio,
- ❸ $P((X, Y) \in A) = \sum_{(x, y) \in A} f(x, y)$,
- ❹ $F(x, y) = \sum_{s \leq x} \sum_{t \leq y} f(s, t)$.

Zadanie

W każdym z dwóch etapów teleturnieju można otrzymać 0, 1, lub 2 punkty. Niech zmienne losowe X , Y oznaczają odpowiednio liczby punktów uzyskane w etapie I i II przez losowo wybranego uczestnika. Funkcję prawdopodobieństwa łącznego określa tabela:

y	0	1	2
x			
0	0,5	0,05	0,01
1	0,2	0,1	0,06
2	0,02	0,03	?

Prawdopodobieństwo $P(X = x \text{ i } Y = y)$ podane jest na przecięciu wiersza $X = x$ i kolumny $Y = y$, na przykład $P(X = 1 \text{ i } Y = 0) = 0,2$.

Znajdziemy:

- a) $f(2, 2) = P(X = 2, Y = 2)$,
- b) $P(Y = 2)$,
- c) $F(1, 1)$.

a)

$$\begin{aligned} f(2, 2) &= P(X = 2, Y = 2) = \\ &= 1 - (0.5 + 0.05 + 0.01 + 0.2 + 0.1 + 0.06 + 0.02 + 0.03) = \\ &= 1 - 0.97 = 0.03. \end{aligned}$$

b)

$$P(Y = 2) = 0.01 + 0.06 + 0.03 = 0.1.$$

c)

$$F(1, 1) = P(X \leq 1, Y \leq 1) = 0.5 + 0.2 + 0.05 + 0.1 = 0.85$$

y	0	1	2
x			
0	0.5	0.05	0.01
1	0.2	0.1	0.06
2	0.02	0.03	0.03

Definicja

Dwuwymiarowa zmienna losowa (X, Y) nazywana jest ciągłą zmienną losową (krócej - zmienną ciągłą), jeśli jej łączny rozkład prawdopodobieństwa określony jest przez funkcję gęstości łącznej (łączną gęstość prawdopodobieństwa) taką, że:

- i) $f(x, y) \geq 0$,
- ii) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \, dx \, dy = 1$,
- iii) $P((X, Y) \in A) = \int \int_A f(x, y) \, dx \, dy$.

Zadanie

Dwuwymiarowa zmienna losowa ma gęstość łączną postaci:

$$f(x, y) = \begin{cases} Cx^2 & \text{gdzie } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{w przeciwnym przypadku} \end{cases}$$

dla pewnej stałej C. Znajdź wartość tej stałej C.

Ponieważ f jest gęstością to:

$$\begin{aligned} 1 &= \int_0^1 \int_0^1 Cx^2 dx dy = \int_0^1 \left(\int_0^1 Cx^2 dx \right) dy = \\ &= C \int_0^1 \left[\frac{1}{3} x^3 \right]_0^1 dy = C \int_0^1 \frac{1}{3} dy = \frac{C}{3}. \end{aligned}$$

Czyli:

$$1 = \frac{C}{3}$$

$$C = 3.$$

Definicja

Rozkładem brzegowym pary (X, Y) nazywamy rozkład prawdopodobieństwa zmiennej losowej X lub zmiennej losowej Y :

- a. dla dyskretnych zmiennych X, Y , brzegowe funkcje prawdopodobieństwa są postaci:

$$f_X(x) = P(X = x) = \sum_y f(x, y), \quad f_Y(y) = P(Y = y) = \sum_x f(x, y)$$

- b. dla ciągłych zmiennych X, Y , brzegowe gęstości są postaci:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy, \quad f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

Zadanie

Dwuwymiarowa zmienna losowa (X, Y) ma gęstość:

$$f(x, y) = \begin{cases} \frac{3}{8}(x - y)^2 & \text{gdy } -1 \leq x \leq 1, -1 \leq y \leq 1 \\ 0 & \text{w przeciwnym przypadku} \end{cases}$$

Znajdziemy gęstość zmiennej losowej X .

Zadanie

Dwuwymiarowa zmienna losowa (X, Y) ma gęstość:

$$f(x, y) = \begin{cases} \frac{3}{8}(x - y)^2 & \text{gdy } -1 \leq x \leq 1, -1 \leq y \leq 1 \\ 0 & \text{w przeciwnym przypadku} \end{cases}$$

Znajdziemy gęstość zmiennej losowej X .

Wyznamy gęstość zmiennej losowej X :

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_{-1}^1 \frac{3}{8}(x - y)^2 dy =$$

Zadanie

Dwuwymiarowa zmienna losowa (X, Y) ma gęstość:

$$f(x, y) = \begin{cases} \frac{3}{8}(x - y)^2 & \text{gdy } -1 \leq x \leq 1, -1 \leq y \leq 1 \\ 0 & \text{w przeciwnym przypadku} \end{cases}$$

Znajdziemy gęstość zmiennej losowej X .

Wyznamy gęstość zmiennej losowej X :

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x, y) dy = \int_{-1}^1 \frac{3}{8}(x - y)^2 dy = \\ &= \frac{3}{8} \int_{-1}^1 (x^2 - 2xy + y^2) dy = \frac{3}{8} [x^2 y - xy^2 + \frac{1}{3}y^3]_{-1}^1 = \frac{1}{4}(3x^2 + 1). \end{aligned}$$

Zadanie

Kontynuacja zadania 1. Uzupełnij tabelę:

y x	0	1	2	$f_X(x)$
0	0.5	0.05	0.01	?
1	0.2	0, 1	0.06	?
2	0.02	0.03	?	?
$f_Y(y)$?	?	?	

Zadanie

Kontynuacja zadania 1. Uzupełnij tabelę:

y x	0	1	2	$f_X(x)$
0	0.5	0.05	0.01	?
1	0.2	0, 1	0.06	?
2	0.02	0.03	?	?
$f_Y(y)$?	?	?	

$$f_Y(0) = 0.5 + 0.2 + 0.02 = 0.72$$

Zadanie

Kontynuacja zadania 1. Uzupełnij tabelę:

y x	0	1	2	$f_X(x)$
0	0.5	0.05	0.01	?
1	0.2	0,1	0.06	?
2	0.02	0.03	?	?
$f_Y(y)$?	?	?	

$$f_Y(0) = 0.5 + 0.2 + 0.02 = 0.72$$

$$f_Y(1) = 0.05 + 0.1 + 0.03 = 0.18$$

Zadanie

Kontynuacja zadania 1. Uzupełnij tabelę:

y x	0	1	2	$f_X(x)$
0	0.5	0.05	0.01	?
1	0.2	0,1	0.06	?
2	0.02	0.03	?	?
$f_Y(y)$?	?	?	

$$f_Y(0) = 0.5 + 0.2 + 0.02 = 0.72$$

$$f_Y(1) = 0.05 + 0.1 + 0.03 = 0.18$$

$$f_Y(3) = 0.01 + 0.06 + 0.03 = 0.1$$

Zadanie

Kontynuacja zadania 1. Uzupełnij tabelę:

y x	0	1	2	$f_X(x)$
0	0.5	0.05	0.01	?
1	0.2	0, 1	0.06	?
2	0.02	0.03	?	?
$f_Y(y)$?	?	?	

$$f_Y(0) = 0.5 + 0.2 + 0.02 = 0.72$$

$$f_Y(1) = 0.05 + 0.1 + 0.03 = 0.18$$

$$f_Y(3) = 0.01 + 0.06 + 0.03 = 0.1$$

$$f_X(0) = 0.5 + 0.05 + 0.01 = 0.56$$

Zadanie

Kontynuacja zadania 1. Uzupełnij tabelę:

y x	0	1	2	$f_X(x)$
0	0.5	0.05	0.01	?
1	0.2	0,1	0.06	?
2	0.02	0.03	?	?
$f_Y(y)$?	?	?	

$$f_Y(0) = 0.5 + 0.2 + 0.02 = 0.72$$

$$f_Y(1) = 0.05 + 0.1 + 0.03 = 0.18$$

$$f_Y(3) = 0.01 + 0.06 + 0.03 = 0.1$$

$$f_X(0) = 0.5 + 0.05 + 0.01 = 0.56$$

$$f_X(1) = 0.2 + 0.1 + 0.06 = 0.36$$

Zadanie

Kontynuacja zadania 1. Uzupełnij tabelę:

y x	0	1	2	$f_X(x)$
0	0.5	0.05	0.01	?
1	0.2	0,1	0.06	?
2	0.02	0.03	?	?
$f_Y(y)$?	?	?	

$$f_Y(0) = 0.5 + 0.2 + 0.02 = 0.72$$

$$f_Y(1) = 0.05 + 0.1 + 0.03 = 0.18$$

$$f_Y(3) = 0.01 + 0.06 + 0.03 = 0.1$$

$$f_X(0) = 0.5 + 0.05 + 0.01 = 0.56$$

$$f_X(1) = 0.2 + 0.1 + 0.06 = 0.36$$

$$f_X(3) = 0.02 + 0.03 + 0.03 = 0.08.$$

Czyli mamy:

x/y	0	1	2	$f_X(x)$
0	0.5	0.05	0.01	0.56
1	0.2	0.1	0.06	0.36
2	0.02	0.03	0.03	0.08
$f_Y(y)$	0.72	0.18	0.1	

Definicja

Rozkład zmiennej losowej $(X|Y = y)$ nazywamy rozkładem warunkowym zmiennej losowej X przy ustalonej wartości zmiennej losowej Y .

Definicja

Niech (X, Y) będzie dyskretnym wektorem losowym. Mówimy, że funkcja $p(X|Y = y): \mathcal{R} \rightarrow [0, 1]$ jest warunkową funkcją prawdopodobieństwa zmiennej losowej X pod warunkiem, że $Y = y$ jeśli dla każdego $x \in \mathcal{R}$

$$p(X|Y = y)(x) = P(X = x|Y = y).$$

Definicja

Niech (X, Y) będzie dyskretnym wektorem losowym. Mówimy, że funkcja $p(X|Y = y): R \rightarrow [0, 1]$ jest warunkową funkcją prawdopodobieństwa zmiennej losowej X pod warunkiem, że $Y = y$ jeśli dla każdego $x \in \mathbb{R}$

$$p(X|Y = y)(x) = P(X = x|Y = y).$$

Własność

Niech (X, Y) będzie dyskretnym wektorem losowym. Warunkowa funkcja prawdopodobieństwa zmiennej losowej X pod warunkiem, że $Y = y$ dana jest wzorem

$$p(X|Y = y) = \frac{p_{XY}(x, y)}{p_Y(y)},$$

o ile $p_Y(y) > 0$.

Definicja

Niech (X, Y) będzie absolutnie ciągłym wektorem losowym. Mówimy, że funkcja $f(X|Y = y): \mathbb{R} \rightarrow [0, 1]$ jest warunkową gęstością zmiennej losowej X pod warunkiem, że $Y = y$ jeśli dla każdego $[a, b] \subset \mathbb{R}$

$$P(X \in [a, b] | Y = y) = \int_a^b f(X|Y = y)(x) dx.$$

Definicja

Niech (X, Y) będzie absolutnie ciągłym wektorem losowym. Mówimy, że funkcja $f(X|Y = y): \mathbb{R} \rightarrow [0, 1]$ jest warunkową gęstością zmiennej losowej X pod warunkiem, że $Y = y$ jeśli dla każdego $[a, b] \subset \mathbb{R}$

$$P(X \in [a, b] | Y = y) = \int_a^b f(X|Y = y)(x) dx.$$

Własność

Niech (X, Y) będzie absolutnie ciągłym wektorem losowym. Warunkowa gęstość zmiennej losowej X pod warunkiem, że $Y = y$ dana jest wzorem

$$f(X|Y = y)(x) = \frac{f_{XY}(x, y)}{f_Y(y)},$$

o ile $f_Y(y) > 0$.

Definicja

Niech (X, Y) będzie parą zmiennych losowych o rozkładzie określonym przez funkcję $f(x, y)$ będącą funkcją prawdopodobieństwa łącznego lub gęstością. Zmienne losowe X, Y są niezależne, jeśli:

$$f(x, y) = f_X(x)f_Y(y),$$

dla wszystkich wartości x, y .

Zmienne X i Y , które nie są niezależne, nazywamy zależnymi zmiennymi losowymi.

Zadanie

Kontynuacja zadania 1. Czy liczby punktów uzyskane w I i II etapie teleturnieju przez losowo wybranego uczestnika są niezależnymi zmiennymi losowymi?

Zadanie

Kontynuacja zadania 1. Czy liczby punktów uzyskane w I i II etapie teleturnieju przez losowo wybranego uczestnika są niezależnymi zmiennymi losowymi?

Musimy sprawdzić:

$$f(x, y) \stackrel{?}{=} f_X(x)f_Y(y).$$

Wystarczy znaleźć jeden przykład dla którego powyższa nierówność nie zachodzi. Mamy:

$$0.5 = f(0, 0) \neq f_X(0)f_Y(0) = 0.56 \cdot 0.72 = 0.4032.$$

Zadanie

Czy X , Y są niezależnymi zmiennymi losowymi, jeśli ich łączna gęstość ma postać:

$$f(x, y) = \begin{cases} \frac{3}{8}(x - y)^2 & \text{gdy } -1 \leq x \leq 1, -1 \leq y \leq 1 \\ 0 & \text{w przeciwnym przypadku} \end{cases}$$

Musimy sprawdzić:

$$f(x, y) \stackrel{?}{=} f_X(x)f_Y(y).$$

Musimy sprawdzić:

$$f(x, y) \stackrel{?}{=} f_X(x)f_Y(y).$$

Mamy:

$$\begin{aligned} f_Y(y) &= \frac{3}{8} \int_{-1}^1 (x^2 - 2xy + y^2) dx = \frac{3}{8} \left(\frac{x^3}{3} - x^2 y + y^2 x \right) \Big|_{-1}^1 = \\ &= \frac{3}{8} \left[\left(\frac{1}{3} - y + y^2 \right) - \left(-\frac{1}{3} - y - y^2 \right) \right] = \frac{3}{8} \left(\frac{2}{3} + 2y^2 \right) = \\ &= \frac{1}{4} + \frac{3}{4}y^2 = \frac{1}{4}(3y^2 + 1) \end{aligned}$$

Ponieważ funkcja f jest symetryczna ze względu na parametry x i y , to:

$$f_X(x) = \frac{3}{8} \int_{-1}^1 (x^2 - 2xy + y^2) dy = \frac{1}{4}(3x^2 + 1).$$

Policzmy:

$$f_X(x)f_Y(y) = \frac{1}{4}(3x^2 + 1)\frac{1}{4}(3y^2 + 1) = \frac{1}{16}(3x^2 + 1)(3y^2 + 1).$$

Policzmy:

$$f_X(x)f_Y(y) = \frac{1}{4}(3x^2 + 1)\frac{1}{4}(3y^2 + 1) = \frac{1}{16}(3x^2 + 1)(3y^2 + 1).$$

Wystarczy znaleźć jeden przykład, dla którego powyższa nierówność nie zachodzi. Sprawdźmy więc $x = 0$ i $y = 0$. Mamy:

$$f(x, y) = 0$$

oraz

$$f_X(x)f_Y(y) = \frac{1}{16}.$$

Policzmy:

$$f_X(x)f_Y(y) = \frac{1}{4}(3x^2 + 1)\frac{1}{4}(3y^2 + 1) = \frac{1}{16}(3x^2 + 1)(3y^2 + 1).$$

Wystarczy znaleźć jeden przykład, dla którego powyższa nierówność nie zachodzi. Sprawdźmy więc $x = 0$ i $y = 0$. Mamy:

$$f(x, y) = 0$$

oraz

$$f_X(x)f_Y(y) = \frac{1}{16}.$$

Zmienne losowe X oraz Y są zależnymi ponieważ:

$$f(x, y) \neq f_X(x)f_Y(y).$$

Definicja

Wartością oczekiwaną (średnią) zmiennej losowej $g(X, Y)$ nazywamy:

$$E[g(X, Y)] = \sum_x \sum_y g(x, y) f(x, y)$$

gdy X, Y są dyskretne,

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy$$

gdy X, Y są ciągłe.

Własność

Jeśli zmienne losowe X, Y są niezależne, to:

$$E(XY) = EX \cdot EY.$$

Definicja

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

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

Definicja

Współczynnik korelacji:

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

Dla dwóch powiązanych zmiennych korelacja mierzy zależność między nimi.

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

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:

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

Można ją zapisać również w postaci:

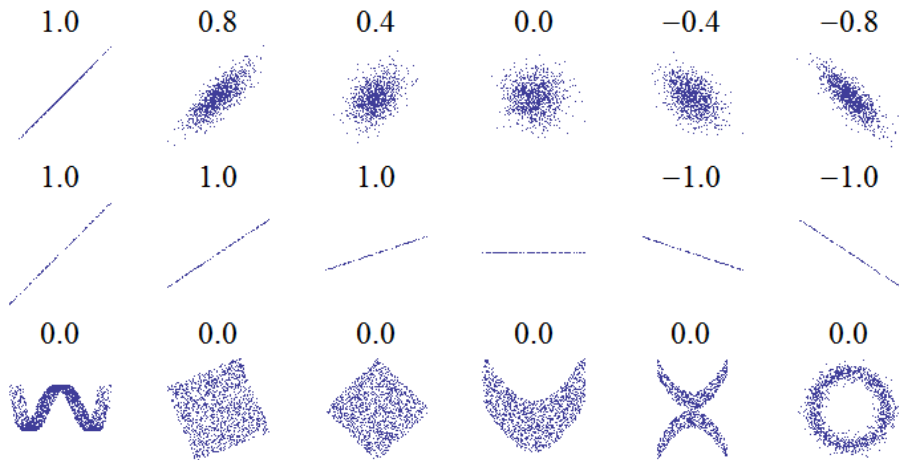
$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

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.

Zadanie w Jupyter

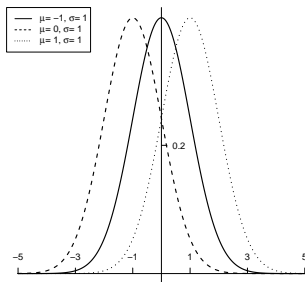
https://github.com/przem85/bootcamp/blob/master/statistics/D09_Z02.ipynb

https://github.com/przem85/bootcamp/blob/master/statistics/D09_Z05.ipynb

Wielowymiarowy rozkład normalny

Rozkład normalny

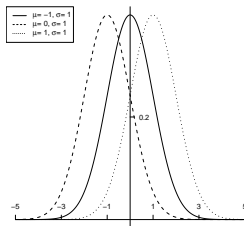
$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$$



Rozkład normalny

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

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



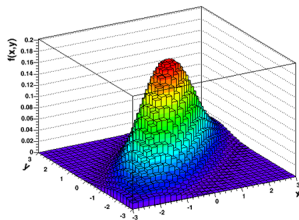
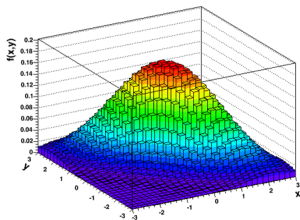
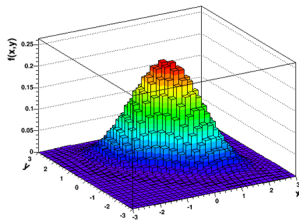
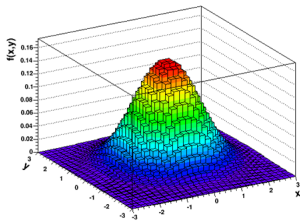
N-wymiarowy rozkład normalny

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

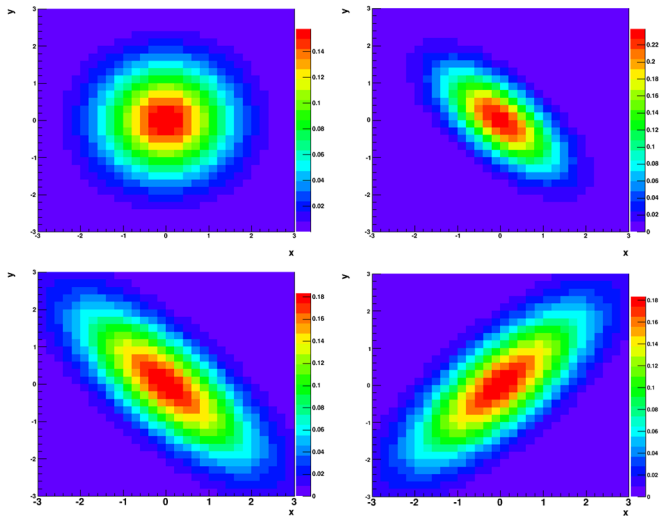
$$f(x_1, \dots, x_n, \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp \left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right).$$

https://github.com/przem85/bootcamp/blob/master/statistics/D09_Z03.ipynb

2-wymiarowy rozkład normalny

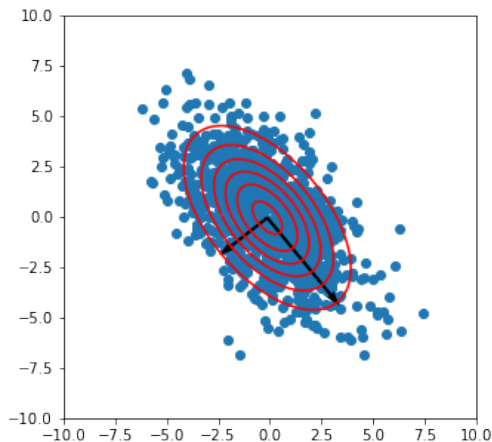


2-wymiarowy rozkład normalny



2-wymiarowy rozkład normalny

https://github.com/przem85/bootcamp/blob/master/statistics/D09_Z04.ipynb



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.

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 odzyskania, co mu się udało – znalazła się dokładnie tam, gdzie Gauss przewidział, że będzie.

Dokonaliśmy pomiarów pewnej funkcji:

$$\begin{array}{c|ccc} x_i & 1 & 2 & 4 \\ \hline y_i & 1 & 2 & 3 \end{array}.$$

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:

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

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

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:

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

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).

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:

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

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

Twierdzenie

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

$$A^T Ax_0 = A^T y.$$

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 $a = 9/10$, $b = 1/2$.

Zadanie w Jupyter

https://github.com/przem85/bootcamp/blob/master/statistics/D09_Z01.ipynb