

Rozdział 1

Miary zależności i miary bliskości między zmiennymi

W rozdziale tym przedstawiam parametry liczbowe do analizy danych 1 – 2 wymiarowych. Przypominam miary rozrzutu oraz omawiam miary zależności i miary bliskości między zmiennymi (cechami).

Główna różnica:

- miary zależności: min dla zmiennych niezależnych, max dla identycznych
- miary bliskości (odległości, zróżnicowania): min dla zmiennych identycznych

Wygodnie jest podzielić zmienne na ilościowe (liczbowe), porządkowe i jakościowe (nominalne). Miary wprowadzone dla danych liczbowych, wykorzystujące wartości liczbowe mają zastosowanie tylko do nich. Miary dla zmiennych porządkowych nadają się również dla zmiennych liczbowych, bo otrzymujemy je przez zamianę wartości cechy na kolejne liczby naturalne $1, 2, \dots, n$ lub ułamki jednostajnie rozłożone na $[0, 1]$, czyli $i/n - 1/2n$, gdzie $i = 1, 2, \dots, n$. Miary definiowane dla zmiennych jakościowych są oparte na gęstościach i mają zastosowanie do wszystkich zmiennych.

Przydaje się również podział miar na symetryczne i niesymetryczne (zależność czy błąd nie muszą być relacjami symetrycznymi). W skrócie:

	Zmienne ilościowe	Zmienne porządkowe	Zmienne jakościowe
Miary zależności symetryczne	Korelacja	Korelacja rang, Współczynnik Kendalla	Wspólna informacja
Miary zależności niesymetryczne			Współczynnik Goodmana-Kruskala
Miary odległości symetryczne	Błąd średniokwadratowy		
Miary odległości niesymetryczne	$\frac{\mathbb{E}(X-Y)^2}{\mathbb{E}Y^2}$		$\chi^2(p, q),$ $H(p q)$

1.1 Zmienne ilościowe

X, Y - zmienne losowe o wartościach rzeczywistych.

Błąd średniokwadratowy między X i Y :

$$\mathbb{E}(X - Y)^2.$$

Korelacja między X i Y :

$$\text{cor}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma(X)\sigma(Y)},$$

gdzie $\text{cov}(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y)$, $\sigma(X) = \sqrt{\text{cov}(X, X)}$.

1.1.1 Stwierdzenie. *Jeśli X, Y standaryzowane, tj. $\mathbb{E}(X) = \mathbb{E}(Y) = 0$, $\sigma(X) = \sigma(Y) = 1$, to $\mathbb{E}(X - Y)^2 = 2 - 2\text{cor}(X, Y)$.*

Dowód.

Dla standaryzowanych zmiennych mamy $\text{cor}(X, Y) = \mathbb{E}(XY)$, $\mathbb{E}X^2 = \mathbb{E}Y^2 = 1$.

Zatem $\mathbb{E}(X - Y)^2 = \mathbb{E}X^2 - 2\mathbb{E}XY + \mathbb{E}Y^2 = 2 - 2\text{cor}(X, Y)$. ■

Warunkowa wartość oczekiwana: $\hat{Y} := \mathbb{E}(Y|X)$

1.1.2 Stwierdzenie. *Niech $f(X)$ będzie funkcją zmienną losową X . Załóżmy, że $\mathbb{E}Y^2$ oraz $\mathbb{E}f(X)^2$ są skończone. Wtedy $\mathbb{E}(Y - f(X))^2$ osiąga minimum dla $f(X) = \hat{Y}(X)$.*

Dowód.

Zauważmy, że

$$\mathbb{E}(Y - \hat{Y})(\hat{Y} - f) = \mathbb{E}[(\hat{Y} - f)\mathbb{E}(Y - \hat{Y}|X)] = 0.$$

Zatem

$$\begin{aligned} \mathbb{E}(Y - f)^2 &= \mathbb{E}(Y - \hat{Y} + \hat{Y} - f)^2 \\ &= \mathbb{E}(Y - \hat{Y})^2 + \mathbb{E}(\hat{Y} - f)^2 + 2\mathbb{E}(Y - \hat{Y})(\hat{Y} - f) \\ &= \mathbb{E}(Y - \hat{Y})^2 + \mathbb{E}(\hat{Y} - f)^2 \\ &\geq \mathbb{E}(Y - \hat{Y})^2. \end{aligned}$$

■

W szczególności dla $f(Y) = \mathbb{E}Y$ otrzymujemy z powyższego dowodu

$$(1.1.3) \quad \text{var}Y = \text{var}\hat{Y} + \mathbb{E}(Y - \hat{Y})^2.$$

Wariancja warunkowa: $\text{var}(Y|X) := \mathbb{E}[(Y - \mathbb{E}(Y|X))^2|X]$

Dekompozycja wariancji Y dana przez 1.1.3 jest równoważna

$$(1.1.4) \quad \text{var}Y = \mathbb{E}\text{var}(Y|X) + \text{var}\mathbb{E}(Y|X).$$

Jeszcze w innym zapisie mamy

$$1 = \frac{\text{var}\hat{Y}}{\text{var}Y} + \frac{\mathbb{E}(Y - \hat{Y})^2}{\text{var}Y},$$

gdzie $\frac{\text{var}\hat{Y}}{\text{var}Y}$ - współczynnik dopasowania X do Y (miara dokładności predykcji),
 $\frac{\mathbb{E}(Y - \hat{Y})^2}{\text{var}Y}$ - względny błąd średniokwadratowy predykcji.

1.1.5 Stwierdzenie. Niech f będzie funkcją mierzalną X . Wtedy $\rho^2(Y, f) := \text{cor}^2(Y, f)$ osiąga maksimum dla $f = \hat{Y}$.

Dowód.

$$\begin{aligned} \text{cov}(Y, f) &= \mathbb{E}(Y - \mathbb{E}Y)(f - \mathbb{E}f) = \mathbb{E}Y(f - \mathbb{E}f) \\ &= \mathbb{E}[(f - \mathbb{E}f)\mathbb{E}(Y|X)] = \mathbb{E}(f - \mathbb{E}f)\hat{Y} \\ &= \mathbb{E}(f - \mathbb{E}f)(\hat{Y} - \mathbb{E}\hat{Y}) = \text{cov}(\hat{Y}, f) \end{aligned}$$

W szczególności dla $f = \hat{Y}$ mamy $\text{cov}(Y, \hat{Y}) = \text{var}\hat{Y} = \sigma_{\hat{Y}}^2$.

Zatem

$$\begin{aligned} \rho(Y, f) &= \frac{\text{cov}(Y, f)}{\sigma_Y \sigma_f} \\ &= \frac{\sigma_{\hat{Y}}^2}{\sigma_Y \sigma_{\hat{Y}}} \cdot \frac{\text{cov}(\hat{Y}, f)}{\sigma_{\hat{Y}} \sigma_f} \\ &= \frac{\text{cov}(Y, \hat{Y})}{\sigma_Y \sigma_{\hat{Y}}} \cdot \frac{\text{cov}(\hat{Y}, f)}{\sigma_{\hat{Y}} \sigma_f} \\ &= \rho(Y, \hat{Y}) \rho(\hat{Y}, f) \end{aligned}$$

Po podniesieniu stronami do kwadratu dostajemy:

$$\rho^2(Y, f) = \rho^2(Y, \hat{Y}) \rho^2(\hat{Y}, f).$$

Maksimum tego wyrażenia jest osiągane gdy $\rho^2(\hat{Y}, f) = 1$, a więc dla $f = \hat{Y}$. ■

Zauważmy, że maksimum w powyższym stwierdzeniu jest równe współczynnikowi dopasowania

$$\rho^2(Y, \hat{Y}) = \frac{\text{cov}^2(Y, \hat{Y})}{\text{var}Y \text{var}\hat{Y}} = \frac{\text{var}\hat{Y}}{\text{var}Y}.$$

1.2 Zmienne porządkowe

Niech x_1, x_2, \dots, x_n będzie ciągiem danych liczbowych, zwanych próbą (niekoniecznie losową).

Niech $x_{1:n} \leq x_{2:n} \leq \dots \leq x_{n:n}$ będzie uporządkowaną próbą losową.

Przyjmijmy $x_0 := -\infty$; $x_{n+1} := +\infty$. Wtedy

$$\forall x_i \exists! \text{ para indeksów } (k, \leq k) \quad x_{k-1:n} < x_{k:n} = x_i = x_{k:n} < x_{k+1:n}.$$

Powyzsza obserwacja uzasadnia następującą definicję. **Rangą** obserwacji x_i w próbie nazywamy:

$$R_i \equiv R(x_i) = \frac{k + k'}{2}$$

Np.

$$\begin{array}{l} x: \quad 2 \quad 3 \quad 2.5 \quad 2.5 \quad 1.5 \quad 5 \quad 4 \\ R: \quad 2 \quad 5 \quad 3.5 \quad 3.5 \quad 1 \quad 7 \quad 6 \end{array}$$

Korelacja rang (Spearmana):

$$\rho(X, Y) = \text{cor}(R(X), R(Y))$$

gdzie $R(X) = (R(x_1), R(x_2), \dots, R(x_n))$, $R(Y) = (R(y_1), R(y_2), \dots, R(y_n))$.

1.2.1 Stwierdzenie. Załóżmy, że elementy próby są parami różne. Wtedy

$$(1) \hat{R} := \frac{1}{n} \sum_{i=1}^n R(x_i) = \frac{n+1}{2},$$

$$(2) \text{var}(R(x)) := \frac{1}{n-1} \sum_{i=1}^n (R(x_i) - \hat{R})^2 = \frac{n(n+1)}{12},$$

$$(3) \rho(X, Y) = \frac{12}{n(n^2-1)} \sum_{i=1}^n R(x_i)R(y_i) - \frac{3(n+1)}{n-1}$$

Współczynnik Kendala zależności między X a Y:

Założmy, że X_1, X_2 niezależne o rozkładzie X, Y_1, Y_2 niezależne o rozkładzie Y.

$$X_{12} = \begin{cases} 1 & X_1 > X_2 \\ 0 & X_1 = X_2 \\ -1 & X_1 < X_2 \end{cases} \quad Y_{12} = \begin{cases} 1 & Y_1 > Y_2 \\ 0 & Y_1 = Y_2 \\ -1 & Y_1 < Y_2 \end{cases}$$

$$\tau_k = \text{cor}(X_{12}, Y_{12}) = \frac{\mathbb{P}((X_1 - X_2)(Y_1 - Y_2) > 0) - \mathbb{P}((X_1 - X_2)(Y_1 - Y_2) < 0)}{\sqrt{\mathbb{P}(X_1 \neq X_2)\mathbb{P}(Y_1 \neq Y_2)}}$$

Jeśli X, Y mają ciągle dystrybuanty, to

$$\tau_k(X, Y) = \mathbb{P}((X_2 - X_1)(Y_2 - Y_1) = 1) - \mathbb{P}((X_2 - X_1)(Y_2 - Y_1) = -1)$$

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ - próba z (X, Y)

τ próbkowy:

$$\tau_k = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \text{sgn}((x_i - x_j)(y_i - y_j))$$

Uwaga. W R: `cor(X, Y, method= "pearson"`
`"sperman"`)
`"kendall"`

Domyślnie ustawiony jest "pearson".

1.3 Zmienne jakościowe

W tej części opisuję miary rozrzutu i bliskości oparte na gęstościach prawdopodobieństwa wykorzystywane przede wszystkim do analizy cech jakościowych.

1.3.1 Miary rozrzutu oparte na gęstościach

Entropia rozkładu p o nośniku Ω

$$H(p) = - \int_{\Omega} [\log p(v)]p(v)dv$$

Jeśli X - zmienna losowa o gęstości p_X , to $H(X) := H(p_X)$.

Uwaga. Różnice i podobieństwa między H i var :

(1) Załóżmy, że $0 \leq X \leq 1$. Wtedy

$$var X = \mathbb{E}X^2 - (\mathbb{E}X)^2 \leq \mathbb{E}X - (\mathbb{E}X)^2 \leq \frac{1}{4}.$$

Zatem var jest największa dla rozkładu dwupunktowego: $p_0 = \frac{1}{2} = p_1$. Entropia natomiast jest największa dla rozkładu jednostajnego.

(2) Załóżmy teraz, że $X \sim N(\mu, \sigma)$. Mamy

$$\ln f_X(x) = \ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{(x-\mu)^2}{2\sigma^2}$$

$$var(X) = \int (x - m)^2 f_X(x) dx = \sigma^2$$

$$H(X) = -\ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) + \frac{1}{2} = \frac{1}{2} + \ln\sqrt{2\pi} + \ln\sigma$$

Współczynnik Giniego, $V(p)$ dla gęstości p

$$V(p) = \int [1 - p(v)]p(v)dv = 1 - \int p^2(v)dv$$

Jeśli X - zmienna losowa o gęstości p_X , to $V(X) := V(p_X)$.

V jest liniowym (rozwiniecie Taylora dla \log) przybliżeniem H .

1.3.2 Miary bliskości

Dla prostoty ograniczymy się w dalszej części rozdziału do rozkładów dyskretnych p, q o wspólnym nośniku Ω

Odległość Kullbacka-Leiblera (względna entropia):

$$H(p||q) = \sum_{i \in \mathcal{X}} (\log \frac{p_i}{q_i})p_i$$

1.3.1 Stwierdzenie. (1) $H(p||q) \geq 0$

(2) $H(p||q) = 0 \Leftrightarrow p = q$

Dowód.

$$H(p||q) = \sum_i (\log \frac{p_i}{q_i})p_i = - \sum_i (\log \frac{q_i}{p_i})p_i \geq \sum_i (1 - \frac{q_i}{p_i})p_i = 0. \text{ Stąd wynika (1).}$$

Równość w ostatniej nierówności jest równoważna warunkowi $\frac{q_i}{p_i} = 1$ dla wszystkich i . Stąd otrzymujemy (2). ■

Odległość χ^2 między rozkładami prawdopodobieństwa:

$$\chi^2(p, q) = \sum_{i \in \Omega} \left(\frac{p_i - q_i}{p_i} \right)^2 p_i$$

Odległość χ^2 jest kwadratowym (rozwińcie Taylora dla \log) przybliżeniem $H(p||q)$.

1.3.3 Miary zależności

Niech (X, Y) będą zmiennymi o rozkładzie dyskretnym. Ponadto

$$p_{ij} := \mathbb{P}(X = i, Y = j),$$

$$p_{j|i} := \mathbb{P}(Y = j | X = i),$$

$$p_{i.} := \mathbb{P}(X = i),$$

$$p_{.j} := \mathbb{P}(Y = j),$$

$$V(Y|X = i) := 1 - \sum_j p_{j|i}^2,$$

$$\mathbb{E}(V(Y|X)) := \sum_i V(Y|X = i)p_{i.} = 1 - \sum_i p_{i.} \sum_j p_{j|i}^2.$$

Współczynnik Goodmana-Kruskala:

$$\tau(Y|X) = \frac{V(Y) - \mathbb{E}(V(Y|X))}{V(Y)}$$

Zakładamy, że rozkład Y nie jest zdegenerowany, czyli, że $V(Y) > 0$.

1.3.2 Stwierdzenie. (1) $0 \leq \tau \leq 1$

(2) $\tau = 0 \Leftrightarrow X, Y$ niezależne

Dowód.

Oczywiście $\tau \leq 1$.

Dla dowodu, że $\tau \geq 0$, zauważmy, że

$$\mathbb{E}V(Y|X) = \sum_i V(Y|X = x_i)p_i = 1 - \sum_i p_{i.} \sum_j p_{j|i}^2.$$

Wystarczy pokazać, że $\sum_j p_{.j}^2 \leq \sum_j \sum_i p_{i.} p_{j|i}^2$

Z kolei wystarczy pokazać, że $p_{.j}^2 \leq \sum_i p_{j|i}^2 p_{i.}$

Lewa = $p_{.j}^2 = (\sum_i p_{j|i} p_{i.})^2$, więc (1) wynika z nierówności Jensena.

Dla dowodu (2) zauważmy, że "=" w nierówności Jensena wyrazów $p_{j|i} = p_{.j}, \forall i, j$ jest równoważna niezależności X, Y . ■

Wspólna informacja zawarta w X i Y :

$$M(X, Y) = \sum_{j=1}^l \sum_{i=1}^k p_{ij} \log \frac{p_{ij}}{p_{i.} p_{.j}}$$

1.3.3 Stwierdzenie. (1) $M(X, Y) \geq 0$

(2) $M(X, Y) = 0 \Leftrightarrow X, Y$ niezależne

Dowód.

Wynika z poprzedniego stwierdzenia, bo M jest równa $H((p_{ij}) || (p_{i \cdot}, p_{\cdot j}))$

■

Uwaga. Przy okazji

$$\begin{aligned} M(X, Y) &= -\sum_j \sum_i p_{ij} \log\left(\frac{p_{i \cdot} p_{\cdot j}}{p_{ij}} - 1 + 1\right) \\ &\approx -\left[\sum_{ji} \left(\frac{p_{i \cdot} p_{\cdot j}}{p_{ij}} - 1\right) p_{ij} - \frac{1}{2} \sum_{ji} \left(\frac{p_{i \cdot} p_{\cdot j}}{p_{ij}} - 1\right)^2 p_{ij}\right] \\ &= \frac{1}{2} \sum_{ij} \frac{(p_{i \cdot} p_{\cdot j} - p_{ij})^2}{p_{ij}} \end{aligned}$$

Ostatnie wyrażenie oraz statystyką $\chi^2 = \sum_{ij} \frac{(p_{ij} - p_{i \cdot} p_{\cdot j})^2}{p_{i \cdot} p_{\cdot j}}$ dla testowania niezależności mają podobną interpretację, ale różnica w treści matematycznej jest (przynajmniej na pierwszy rzut oka) zasadnicza. Być może o podobieństwie wyrażen decydują własności błędu względnego: jeśli błąd względny oszacowania a za pomocą b jest nie większy od ε , to błąd względny oszacowania b za pomocą a jest nie większy niż $\varepsilon/(1 - \varepsilon)$. Przy małym ε wyrażenia te są porównywalne.

$$\begin{aligned} \text{Oznaczmy } H(Y) &:= -\sum_j p_{\cdot j} \log p_{\cdot j}, \\ H(Y|X = x_i) &:= -\sum_j p_{j|i} \log p_{j|i}, \\ \mathbb{E}H(Y|X) &:= \sum_i H(Y|X = x_i) p_{i \cdot}. \end{aligned}$$

1.3.4 Stwierdzenie.

$$M(Y|X) := \frac{H(Y) - \mathbb{E}H(Y|X)}{H(Y)} = \frac{M(X, Y)}{H(Y)}$$

Dowód.

$$\begin{aligned} 1. \mathbb{E}H(Y|X) &= -\sum_i p_{i \cdot} \sum_j p_{j|i} \log p_{j|i} \\ &= -\left[\sum_{ij} p_{ij} \log p_{j|i}\right] \\ &= -\left[\sum_{ij} p_{ij} \log p_{ij} - \sum_{ij} p_{ij} \log p_{i \cdot}\right] \\ &= H(X, Y) + \sum_i \log p_{i \cdot} \sum_j p_{ij} = H(X, Y) + \sum_i \log p_{i \cdot} \cdot p_{i \cdot}. \\ &= H(X, Y) - H(X) \end{aligned}$$

$$\text{Zatem } M(Y|X) = \frac{H(Y) + H(X) - H(X, Y)}{H(Y)}.$$

$$\begin{aligned} 2. M(X, Y) &= \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{p_{i \cdot} p_{\cdot j}} \\ &= \sum_i \sum_j p_{ij} \log p_{ij} - \sum_i \sum_j p_{ij} \log p_{i \cdot} p_{\cdot j} \\ &= -H(X, Y) + \sum_i \sum_j p_{ij} \log p_{i \cdot} + \sum_j \sum_i p_{ij} \log p_{\cdot j} \\ &= -H(X, Y) + \sum_i (\log p_{i \cdot}) p_{i \cdot} + \sum_j (\log p_{\cdot j}) p_{\cdot j} \\ &= -H(X, Y) + H(X) + H(Y) \end{aligned}$$

$$\text{Zatem } M(X, Y) = H(X) + H(Y) - H(X, Y).$$

Z 1. i 2. otrzymujemy tezę. ■