Elementy Modelowania Matematycznego
Wykład 4
Regresja i dyskryminacja liniowa

Romuald Kotowski

Katedra Informatyki Stosowanej

PJWSTK 2009

Spis treści

Para zmiennych losowych

Wstęp

Bardzo często interesujący jest łączny probabilistyczny rozkład kilku zmiennych losowych. Tu ograniczymy się do przypadku tylko dwóch zmiennych losowych, ale łatwo zauważyć, że wszystkie ogólne rozważania na temat pary zmiennych losowych mają swoje naturalne i proste uogólnienia na przypadek ich większej liczby.

Para zmiennych losowych

Prawdopodobieństwo łączne

X, Y – dwie dyskretne zmienne losowe określone na tej samej przestrzeni zdarzeń elementarnych. Ich łączny rozkład jest dany **funkcją prawdopodobieństwa łącznego**

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

określającą prawdopodobieństwo jednoczesnego przyjęcia przez zmienną losową X wartości x i przez zmienną losową Y wartości y. Funkcja prawdopodobieństwa ma następujące własności:

1. $f(x,y) \geq 0$ dla wszystkich (x,y)
2. $\sum_x \sum_y f(x,y) = 1$
3. $P((X,Y) \in A) = \sum_{(x,y)\in A} f(x,y)$

Para zmiennych losowych

Prawdopodobieństwo łączne

Przykład
Funkcja prawdopodobieństwa łącznego dana jest wzorem

$$P(X=x, Y=y) = f(x,y) = \begin{cases} \frac{1}{30}(x+y) & \text{dla } x=0,1,2 \text{ oraz } y=0,1,2,3 \\ 0 & \text{w innym przypadku} \end{cases}$$

Tablica kondyngencji

X	Y			
	0	1	2	3
0	0	1/30	1/15	1/10
1	1/30	1/15	1/10	2/15
2	1/15	1/10	2/15	1/6

czyli np. $P(X=2, Y=0) = f(2,0) = \frac{1}{30}(2+0) = \frac{1}{15}$

Para zmiennych losowych

Prawdopodobieństwo łączne

Dystrybuantą łączną dyskretnych zmiennych losowych X i Y nazywamy funkcję

$$F(x,y) = P(X \leq x, Y \leq y) = \sum_{s \leq x} \sum_{t \leq y} f(s,t)$$

Dystrybuantą łączną ciągłych zmiennych losowych X i Y nazywamy funkcję

$$F(x,y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(s,t) ds\, dt$$

Para zmiennych losowych

Rozkład brzegowy

Rozkład brzegowy – interesuje nas tylko rozkład jednej zmiennej

Zmienna dyskretna

$$g(x) = \sum_{y} f(x,y) \qquad h(y) = \sum_{x} f(x,y)$$

Zmienna ciągła

$$g(x) = \int_{-\infty}^{\infty} f(x,y)dy \qquad h(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

Para zmiennych losowych

Rozkład brzegowy

Przykład c.d.

Rozkład brzegowy zmiennej losowej X jest dany funkcją prawdopodobieństwa

$$g(x) = P(X = x) = \sum_{y=0}^{3} f(x,y) = \frac{1}{30}\sum_{y=0}^{3}(x+y) = \frac{1}{15}(2x+3) \vee 0$$

Rozkład brzegowy zmiennej losowej Y jest dany funkcją prawdopodobieństwa

$$h(y) = P(Y = y) = \sum_{x=0}^{2} f(x,y) = \frac{1}{10}(y+1) \vee 0$$

Para zmiennych losowych

Rozkład brzegowy

Przykład c.d.

Tablica kondyngencji

X	Y				
	0	1	2	3	
0	0	1/30	1/15	1/10	1/5
1	1/30	1/15	1/10	2/15	1/3
2	1/15	1/10	2/15	1/6	7/15

Para zmiennych losowych

Rozkład warunkowy

Rozkład warunkowy zmiennej losowej X pod warunkiem, że zmienna losowa Y przyjęła wartość y, czyli że $Y = yg$, jest dany funkcją

$$f(x|y) = \frac{f(x,y)}{h(y)}$$

Para zmiennych losowych

Rozkład warunkowy

Przykład c.d.

$$f(x|y) = \begin{cases} \frac{\frac{1}{30}(x+y)}{\frac{1}{10}(y+1)} = \frac{x+y}{3(y+1)} & \text{dla } x = 0,1,2 \\ 0 & \text{w innym przypadku} \end{cases}$$

Dla $Y = 2$

$$f(x|2) = P(X = x|Y = 2) = \frac{x+2}{9}, \text{ dla } x = 0,1,2 \vee 0$$

$$P(X = 0|Y = 2) = \frac{2}{9}, P(X = 1|Y = 2) = \frac{1}{3}, P(X = 2|Y = 2) = \frac{4}{9}$$

Para zmiennych losowych

Zmienne niezależne

Dwie zmienne losowe X i Y o łącznym rozkładzie $f(\cdot,\cdot)$ nazywamy **niezależnymi** wtedy i tylko wtedy, gdy dla wszystkich par uporządkowanych (x, y) z zakresu wartości zmiennej losowej X oraz zmiennej losowej Y

$$f(x, y) = g(x)\, h(y)$$

Przykład zależnych zmiennych losowych

$$f(x, y) = \begin{cases} 8xy & \text{dla } 0 < x < y < 1 \\ 0 & \text{w innym przypadku} \end{cases}$$

Para zmiennych losowych

Wartość oczekiwana

$p(X, Y)$ – ustalona (rzeczywista) funkcję zmiennych losowych X i Y o łącznym rozkładzie $f(x, y)$. Wartością oczekiwaną zmiennej losowej $p(X, Y)$ nazywamy wielkość

$$\mu_{p(X,Y)} \equiv E[p(X, Y)] = \begin{cases} \sum_x \sum_y p(x, y) f(x, y), & \text{gdy } X, Y \text{ dyskretne} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y) f(x, y) dx\, dy, & \text{gdy } X, Y \text{ ciągłe} \end{cases}$$

Para zmiennych losowych

Wartość oczekiwana

Zastosowanie

Każdy z momentów pojedynczej zmiennej losowej, powiedzmy zmiennej X, może być przedstawiony jako wartość oczekiwana odpowiedniej funkcji $p(X, Y)$. Chcąc na przykład otrzymać wartość oczekiwaną zmiennej losowej X wystarczy za funkcję $p(X, Y)$ przyjąć funkcję $p(X, Y) = X$:

$$\mu_X \equiv E(X) = \begin{cases} \sum_x \sum_y xf(x,y) = \sum_x xg(x), & \text{gdy } X, Y \text{ dyskretne} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xf(x,y)dx\,dy = \int_{-\infty}^{\infty} xg(x)dx, & \text{gdy } X, Y \text{ ciągłe} \end{cases}$$

Wskaźniki rozproszenia

Wariancja w próbie przypomnienie

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{1}$$

$\bar{x}$ – średnia w próbie.

Odchylenie standardowe

$$s = \sqrt{s^2} \tag{2}$$

Para zmiennych losowych

Kowariancja

X i Y – zmienne losowe o łącznym rozkładzie $f(\cdot,\cdot)$, Kowariancją zmiennych X i Y nazywamy wielkość:

$$\sigma_{XY} = E[(X-\mu_X)(Y-\mu_Y)] = \begin{cases} \sum_x \sum_y (x-\mu_X)(y-\mu_Y) f(x,y), & X, Y \text{ dyskretne} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f(x,y)\,dx\,dy, & X, Y \text{ ciągłe} \end{cases}$$

μ_X, μ_Y – odpowiednio średnia wartość (oczekiwana) zmiennej X i zmiennej Y. Inne oznaczenie σ_{XY} to $Conv(X,Y)$. Zauważmy, że $\sigma_{XX} = \sigma_X^2$.

Para zmiennych losowych

Kowariancja

Kowariancja zmiennych X i Y jest dodatnia

- jeżeli 'dużym' wartościom zmiennej X (czyli większym od wartości średniej μ_X) towarzyszą zwykle 'duże' wartości zmiennej Y (większe od μ_Y) i...
- jeżeli 'małym' wartościom zmiennej X towarzyszą zwykle 'małe' wartości zmiennej Y (mniejsze od wartości średniej μ_X i μ_Y odpowiednio)

Kowariancja zmiennych X i Y jest ujemna

- jeżeli wartościom zmiennej X większym od μ_X towarzyszą zwykle wartości zmiennej Y mniejsze od μ_Y
- jeżeli wartościom zmiennej X mniejszym od μ_X towarzyszą zwykle wartości zmiennej Y większe od μ_Y

Para zmiennych losowych

Kowariancja

Kowariancja umożliwia zatem skonstruowanie wskaźnika mówiącego o istnieniu (lub nieistnieniu) zależności 'dodatniej' lub 'ujemnej' między zmiennymi losowymi.

Przydatne zależności

$$Cov(X, Y) = E(XY) - \mu_X \mu_Y$$

Jeśli zmienne losowe X, Y są niezależne, to

$$Cov(X, Y) = 0$$

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$$

gdzie a, b – stałe
Wniosek: jeśli zmienne losowe X i Y są niezależne, to

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$$

Para zmiennych losowych

Współczynnik korelacji ρ

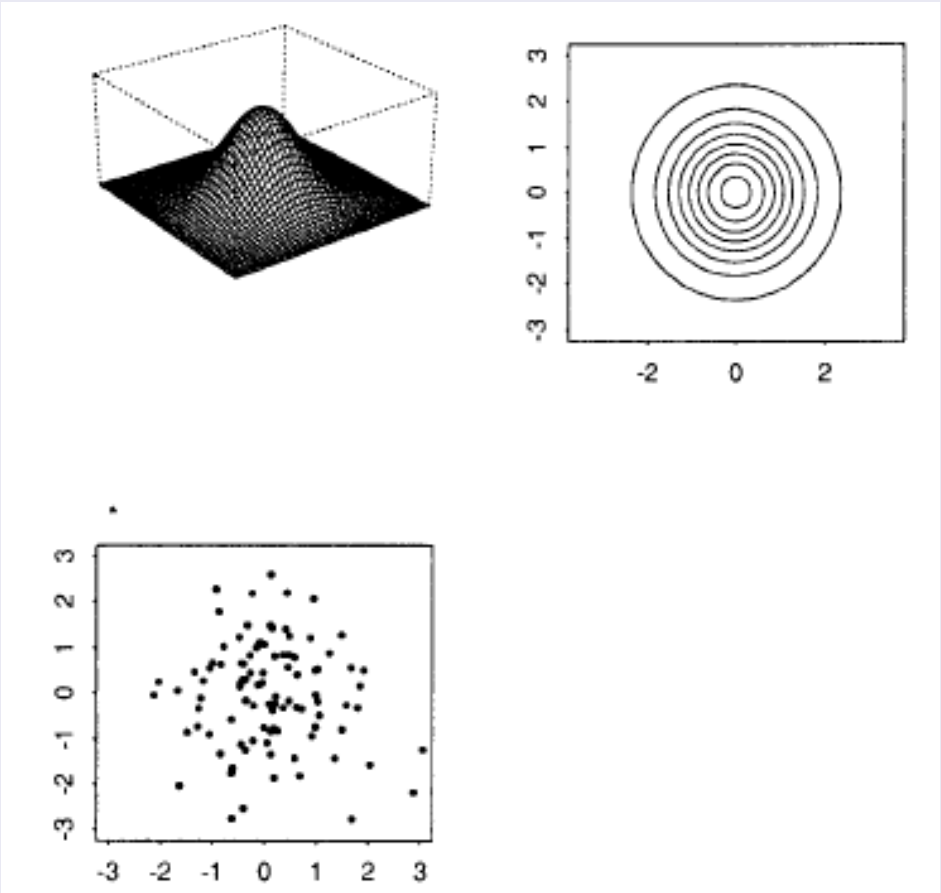
$$\rho = \frac{Cov(X, Y)}{\sqrt{Var(X)\, Var(Y)}}$$

Właściwości:

- $-1 \leq \rho \leq 1$
- $\rho = 1 \leftarrow$ jeżeli $Y = a + bX$, $b > 0$, a, b – stałe
- $\rho = -1 \leftarrow$ jeżeli $Y = a + bX$, < 0, a, b – stałe
- $\rho = 0 \leftarrow$ jeżeli zmienne losowe X i Y są niezależne

Para zmiennych losowych

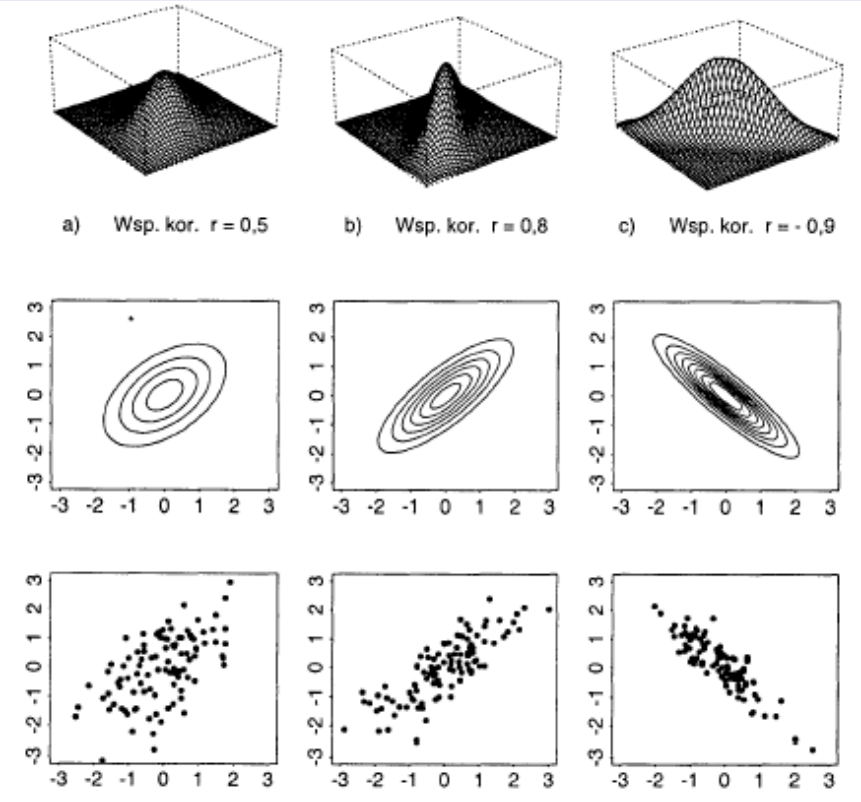
Współczynnik korelacji ρ



Rys. 1: Gęstość dwuwymiarowego rozkładu normalnego $N(0, 0, 1, 1, 0)$ oraz warstwice i przykładowa próba losowa

Para zmiennych losowych

Współczynnik korelacji ρ



Rys. 2: Gęstości dwuwymiarowego rozkładu normalnego oraz warstwice i przykładowe próby losowe dla różnych wartości współczynnika korelacji r

Para zmiennych losowych

Współczynnik korelacji ρ

Przykład: **dwuwymiarowy rozkład normalny** $N(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$, dany łączną gęstością

$$f(x,y) = \frac{e^{-q/2}}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}$$

$\infty < x, y < \infty$, $\sigma_X, \sigma_Y > 0$, $-1 < \rho < 1$

$$q = \frac{1}{1-\rho^2}\left[\left(\frac{x-m_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-m_X}{\sigma_X}\right)\left(\frac{y-m_Y}{\sigma_Y}\right) + \left(\frac{y-m_Y}{\sigma_Y}\right)^2\right]$$

Para zmiennych losowych

Dwuwymiarowy rozkład normalny

1. Para zmiennych losowych Xi Y ma dwuwymiarowy rozkład normalny wtedy i tylko wtedy, gdy każda kombinacja liniowa tych zmiennych, $aX + bY$, gdzie a i b – dowolne stałe, ma rozkład normalny.
2. Zmienne losowe X i Y są niezależne wtedy i tylko wtedy, gdy ich współczynnik korelacjirhi jest równy 0.
3. Jeśli X i Y są niezależne i mają rozkłady normalne odpowiednio $N(\mu_1, \sigma_1)$ i $N(\mu_2, \sigma_2)$, to dla dowolnych liczb a i b, nie będących równocześnie równe 0, $aX + bY$ ma rozkład normalny $N(\mu, \sigma)$, gdzie $\mu = a\mu_1 + b\mu_2$ i $\sigma = \sqrt{a^2\sigma_1^2 + b^2\sigma_2^2}$.

Para zmiennych losowych

Współczynnik korelacji próbkowej

Wprowadzimy obecnie pojęcie współczynnika korelacji próbkowej będącego estymatorem współczynnika korelacji. Jego wartość obliczona dla konkretnych wartości próby ułatwia w wielu przypadkach określenie siły zależności. Współczynnik korelacji zmiennych losowych X i Y został zdefiniowany jako wartość średnia standaryzowanych zmiennych $(X - \mu_X)/\sigma_X$ i $(Y - \mu_Y)/\sigma_Y$. Współczynnik korelacji próbkowej jest odpowiednikiem tej definicji dla próby $(X_1, Y_1), \ldots, (X_n, Y_n)$.

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{S_X} \right) \left(\frac{Y_i - \bar{Y}}{S_Y} \right)$$

$\bar{X}$ i S_X – średnia i odchylenie próby $X_1, X_2, \ldots, X_n$; $\bar{Y}$ i S_Y – średnia i odchylenie próby $Y_1, Y_2, \ldots, Y_n$

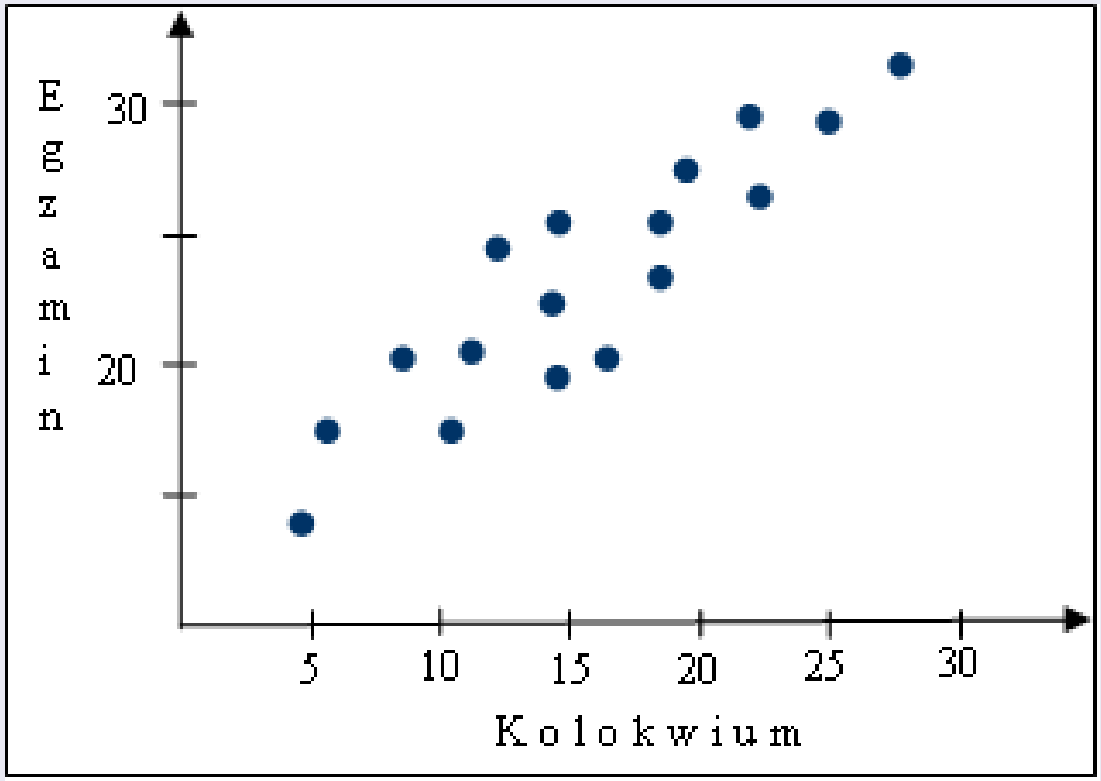
Para zmiennych losowych

Współczynnik korelacji próbkowej

(1) Próbkowy współczynnik korelacji jest ograniczoną $-1 \leq r \leq 1$. Wartości r bliskie -1lub 1 wskazują, że wykres rozproszenia jest skupiony wokół prostej.
(2) W przypadku liniowego charakteru wykresu rozproszenia próbkowy współczynnik korelacji mierzy siłę zależności między zmiennymi.

Para zmiennych losowych

Współczynnik korelacji próbkowej



Rys. 3: Przykładowy wykres rozproszenia wyników za kolokwium i egzamin (w punktach)

Para zmiennych losowych

Regresja liniowa

Nowe nazwy:

X – zmienna objaśniająca (zmienna niezależna)

Y – zmienna objaśniana (zmienna zależna)

Poszukujemy przybliżonej zależności funkcyjnej między tymi zmiennymi.

Założymy zależność liniową w postaci

$$y = b_0 + b_1 x$$

b_0 – wyraz wolny, b_1 – współczynnik kierunkowy

Mówimy, że $\hat{y}_i = b_0 + b_1 x_i$ to wartość y przewidywana na podstawie rozpatrywanej prostej dla wartości zmiennej objaśniającej x równej x_i. **Błąd oszacowania**, czyli tzw. **wartość resztowa** lub **residuum** wynosi $y_i - \hat{y}_i$.

Para zmiennych losowych

Regresja liniowa

Pytanie: jak przeprowadzić prostą przez chmurę wyników, by residua były jak najmniejsze?

Prostą regresji opartą na metodzie najmniejszych kwadratów nazywamy prostą $b_0 + b_1 x$, dla której wartość sumy

$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x))^2$$

traktowanej jako funkcja wszystkich możliwych wartości współczynnika kierunkowego i wyrazu wolnego, jest minimalna.

Para zmiennych losowych

Regresja liniowa

Prosta analiza daje nstp. wyniki:

$$b_0 = \frac{1}{n}(\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i) = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Wartość $y = b_0 + b_1 x$ nazywamy wartością przewidywaną zmiennej objaśnianej na podstawie prostej najmniejszych kwadratów (NMK) dla wartości zmiennej objaśniającej równej x.

Literatura

[1] J. Koronacki, J. Mielniczuk, Statystyka dla studentów kierunków technicznych i przyrodniczych, WNT, 2001

Koniec?

Koniec wykładu 4