Ekonomski fakultet
Sveučilište J. J. Strossmayera u Osijeku

Algoritmi i strukture podataka

Rudolf Scitovski, Ivan Vazler

1 Linearna regresija

Promatramo tzv. linearnu regresiju

$$x \mapsto f(x;\alpha,\beta) = \alpha x + \beta. \tag{1}$$

Za dane podatke mjerenja (točke u ravnini) (x_i, y_i), $i = 1, \ldots, m$, takve da je

$$y_i = f(x_i;\alpha,\beta) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}[0,\sigma^2],$$

treba odrediti optimalne parametre $\alpha^*, \beta^* \in \mathbb{R}$ linearne regresije, tako da odgovarajući pravac $y = \alpha^* x + \beta^*$ prolazi "što bliže" točkama (x_i, y_i), $i = 1, \ldots, m$.

1.1 l_2-pristup: metoda najmanjih kvadrata

Parametre α^*, β^* odredit ćemo tako da tražimo da suma kvadrata odstupanja stvarnih (y_i) od teoretskih $(\alpha^* x_i + \beta^*)$ vrijednosti bude minimalna, tj. minimizirat ćemo funkcional

$$F_2(\alpha,\beta) = \sum_{i=1}^{m} (y_i - \alpha x_i - \beta)^2 \to \min_{\alpha,\beta}.$$

Sjetimo se najprije da vrijedi (**??**)

$$\sum_{i=1}^{m} w_i (y_i - \lambda)^2 \geq \sum_{i=1}^{m} w_i (y_i - \bar{y})^2, \qquad \forall \lambda \in \mathbb{R},$$

pri čemu jednakost vrijedi za $\lambda = \bar{y}$. Neka su α^*, β^* vrijednosti parametara optimalnih u l_2 smislu. Tada vrijedi

$$\begin{aligned} F_2(\alpha^*,\beta^*) = \sum_{i=1}^{m} (y_i - \alpha^* x_i - \beta^*)^2 \quad &\geq \quad \sum_{i=1}^{m} (y_i - \alpha^* x_i - (\bar{y} - \alpha^* \bar{x}))^2 \\ &\quad \text{(pri čemu jednakost vrijedi za } \beta^* = \bar{y} - \alpha^* \bar{x}) \\ &= \quad \sum_{i=1}^{m} ((y_i - \bar{y}) - \alpha^* (x_i - \bar{x}))^2 =: \bar{F}_2(\alpha^*) \end{aligned} \tag{2}$$

Funkcional $\bar{F}(\alpha^*)$ možemo pisati

$$\bar{F}_2(\alpha^*) = \sum_{i=1}^{m} (x_i - \bar{x})^2 \left(\frac{y_i - \bar{y}}{x_i - \bar{x}} - \alpha^* \right)^2. \tag{3}$$

Ako minimizaciju funkcionala $\bar{F}$ shvatimo kao problem mjerenja za podatke $\frac{y_i-\bar{y}}{x_i-\bar{x}}$ s težinama $(x_i-\bar{x})^2$, onda je α^* težinska aritmetička sredina

$$\alpha^* = \frac{1}{\sum(x_i-\bar{x})^2}\sum(x_i-\bar{x})^2\frac{y_i-\bar{y}}{x_i-\bar{x}} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}.$$

Tako dobivamo formule za l_2-optimalne[1] vrijednosti parametara α, β linearne regresije (1)

$$\alpha^* = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}, \qquad \beta^* = \bar{y} - \alpha^*\bar{x}. \tag{4}$$

Zadatak 1. *Objasniti geometrijsko značenje formule (3). Nacrtati podatke i graf linearne regresije s l_2-optimalnim parametrima α^*, β^*.*

Zadatak 2. *Neka je $A(x) = \frac{1}{m}(x_1 + \cdots + x_m)$ aritmetička sredina niza podataka $x = (x_1, \ldots, x_m)$. Pokažite da vrijedi*

$$A(\alpha x + \beta y) = \alpha A(x) + \beta A(y),$$

gdje su x, y nizovi podataka, a α, β realni brojevi. Primijetite da smo ovo svojstvo aritmetičke sredine koristili u (2).

Zadatak 3. *Za zadane podatke (x_i, y_i), $i = 1, \ldots, m$ pokažite da je parametre α^*, β^* moguće dobiti rješavanjem sustava tzv.* normalnih jednadžbi.

$$\begin{aligned}\alpha\sum x_i^2 + \beta\sum x_i &= \sum x_i y_i\\ \alpha\sum x_i + \beta\sum 1 &= \sum y_i\end{aligned}$$

x_i	y_i	x_i^2	x_iy_i	$f(x_i)$	$y_i - f(x_i)$	$(y_i - f(x_i))^2$
-2	10	4	-20	9.2	+0.8	0.16
0	4	0	0	4.8	-0.8	0.16
1	1	1	1	2.6	-1.6	2.56
2	2	4	4	0.4	1.6	2.56
1	17	9	-15		0	5.44

$$\alpha^* = -\tfrac{77}{35} = -2.2, \qquad \beta^* = -15 - 9\alpha^* = \tfrac{168}{35} = 4.8.$$
$$V = \tfrac{1}{m}\sum(y_i - f(x_i))^2 = 1.36, \qquad \sigma = \sqrt{V} = 1.16619.$$

[1] Least Squares Deviations

1.2 l_1-pristup: najmanja apsolutna odstupanja

l_1-optimalne parametre[2] α^*, β^* dobivamo minimizacijom funkcionala

$$F_1(\alpha, \beta) = \sum_{i=1}^{m} |y_i - \alpha x_i - \beta|.$$

Zadatak 4. *Nacrtati podatke i graf linearne regresije s l_1-optimalnim parametrima α^*, β^*.*

Zadatak 5. *Neka je* $\operatorname{med}(x)$ *medijan niza podataka $x = (x_1, \ldots, x_m)$. Kontraprimjerom pokažite da ne vrijedi formula*

$$\operatorname{med}(x + y) = \operatorname{med}(x) + \operatorname{med}(y),$$

gdje su x, y nizovi podataka.

1.3 l_∞-pristup: najmanje maksimalno apsolutno odstupanje

l_∞-optimalne parametre α^*, β^* dobivamo minimizacijom funkcionala

$$F_\infty(\alpha, \beta) = \max_{i=1,m} |y_i - \alpha x_i - \beta|.$$

Zadatak 6. *Nacrtati podatke i graf linearne regresije s l_∞-optimalnim parametrima α^*, β^*.*

[2]Least Absolute Deviations