Klasifikace metodou logistické regrese

Jiří Vomlel
Ústav teorie informace a automatizace AV ČR

Tato prezentace je k dispozici na webu
http://www.utia.cas.cz/vomlel/

a vychází z knihy

T. Hastie, R. Tibshirani, and J. Friedman: The elements of statistical learning: Data Mining, Inference, and Prediction. Springer-Verlag, 2003.

Značení

Předpokládáme vektor vstupů (atributů) $X = (X_1, \ldots, X_p)$.

Snažíme se predikovat výstup Y nabývajících hodnot $1, \ldots, K$.
Náš prediktor označíme G.

Jinými slovy:
naším cílem je klasifikovat objekt popsaný atributy do jedné z K tříd.

$P(G = k|X = x)$ značí pravděpodobnost, že objekt popsaný vektorem atributů x patří do třídy k.

Model logistické regrese

Logistická funkce

$$\text{logit } P(G = k|X = x) = \log \frac{P(G = k|X = x)}{1 - \sum_{\ell=1}^{K-1} P(G = \ell|X = x)}$$

Logistická regrese

$$\begin{aligned}
\text{logit } P(G = 1|X = x) &= \beta_{1,0} + \beta_1^T x \\
\text{logit } P(G = 2|X = x) &= \beta_{2,0} + \beta_2^T x \\
&\dots \\
\text{logit } P(G = K-1|X = x) &= \beta_{K-1,0} + \beta_{K-1}^T x
\end{aligned}$$

kde $\theta = (\beta_{1,0}, \beta_1, \dots, \beta_{K-1,0}, \beta_{K-1})$ jsou parametry modelu.

Z předchozího plyne, že

$$
\begin{aligned}
P(G=1|X=x) &= \frac{\exp(\beta_{1,0}+\beta_1^T x)}{1+\sum_{\ell=1}^{K-1}\exp(\beta_{\ell,0}+\beta_\ell^T x)} \\
P(G=2|X=x) &= \frac{\exp(\beta_{2,0}+\beta_2^T x)}{1+\sum_{\ell=1}^{K-1}\exp(\beta_{\ell,0}+\beta_\ell^T x)} \\
&\cdots \\
P(G=K-1|X=x) &= \frac{\exp(\beta_{K-1,0}+\beta_{K-1}^T x)}{1+\sum_{\ell=1}^{K-1}\exp(\beta_{\ell,0}+\beta_\ell^T x)} \\
P(G=K|X=x) &= \frac{1}{1+\sum_{\ell=1}^{K-1}\exp(\beta_{\ell,0}+\beta_\ell^T x)}
\end{aligned}
$$

Označíme

$$
P(G=k|X=x) = p_k(x,\theta)
$$

Maximálně věrohodný odhad

Předpokládejme data:

$$\begin{aligned}(x_1, g_1) &= ((x_{1,1}, \ldots, x_{1,p}), g_1)\\ (x_2, g_2) &= ((x_{2,1}, \ldots, x_{2,p}), g_2)\\ &\ldots\\ (x_N, g_N) &= ((x_{N,1}, \ldots, x_{N,p}), g_N)\end{aligned}$$

maximálně věrohodný odhad parametů modelu θ pro pozorovaná data maximalizuje

$$\ell(\theta) = \sum_{i=1}^{N} \log p_{g_i}(x_i, \theta) = \sum_{i=1}^{N} \log P(G = g_i | X = x_i)$$

Model logistické regrese pro $K = 2$

Logistická regrese je pak definována jednou rovnicí

$$\text{logit } p_1(x_i, \theta) \quad = \quad \beta_{1,0} + \beta_1^T x_i$$

což zjednodušíme zavedením $\beta = (\beta_{1,0}, \beta_1^T)$ a $x_i = (1, x_{i,1}, \ldots, x_{i,p})$ a $p_1(x_i, \theta) = p(x_i, \beta)$ na

$$\text{logit } p(x_i, \beta) \quad = \quad \beta^T x_i$$

Zakódujeme dvě hodnoty G pomocí y nabývajících hodnot 0 (pro $G = 1$) a 1 (pro $G = 2$) pak pro logaritmickou věrohodnostní funkci můžeme psát

$$\ell(\beta) \quad = \quad \sum_{i=1}^{N} y_i \log p(x_i, \beta) + (1 - y_i) \log(1 - p(x_i, \beta))$$

Použijeme-li vztah $p(x_i, \beta) = \frac{\exp(\beta^T x_i)}{1+\exp(\beta^T x_i)}$ dostaneme

$$\begin{aligned}
\ell(\beta) &= \sum_{i=1}^{N} \left(y_i \log \frac{\exp(\beta^T x_i)}{1+\exp(\beta^T x_i)} + (1-y_i) \log \left(1 - \frac{\exp(\beta^T x_i)}{1+\exp(\beta^T x_i)} \right) \right) \\
&= \sum_{i=1}^{N} \left(y_i \log \frac{\exp(\beta^T x_i)}{1+\exp(\beta^T x_i)} + (1-y_i) \log \frac{1}{1+\exp(\beta^T x_i)} \right) \\
&= \sum_{i=1}^{N} \left(y_i \beta^T x_i - \log(1+\exp(\beta^T x_i)) \right)
\end{aligned}$$

Chceme-li maximalizovat logaritmickou věrohodnostní funkci, požadujeme aby její parciální derivace byly nulové

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{N} \left(y_i x_i - \frac{x_i \exp(\beta^T x_i)}{1+\exp(\beta^T x_i)} \right) = \sum_{i=1}^{N} x_i (y_i - p(x_i, \beta)) = 0$$

což je $p+1$ nelineárních rovnic. Pro vyřešení úlohy můžeme použít

Newtonovu numerickou metodu, která vyžaduje Hessovu matici

$$\begin{aligned}
\frac{\partial^2 \ell(\beta)}{\partial\beta\partial\beta^T} &= -\sum_{i=1}^{N} x_i \cdot \frac{\partial}{\partial\beta} p(x_i,\beta) = -\sum_{i=1}^{N} x_i \cdot \frac{\partial}{\partial\beta}\left(\frac{\exp(\beta^T x_i)}{1+\exp(\beta^T x_i)}\right) \\
&= -\sum_{i=1}^{N} x_i x_i^T \cdot \frac{\exp(\beta^T x_i)(1+\exp(\beta^T x_i)) - \exp(\beta^T x_i)^2}{(1+\exp(\beta^T x_i))^2} \\
&= -\sum_{i=1}^{N} x_i x_i^T \cdot p(x_i,\beta)(1-p(x_i,\beta))
\end{aligned}$$

Je-li β^{old} stará hodnota, novou hodnotu spočteme jako

$$\beta^{new} = \beta^{old} - \left(\frac{\partial^2 \ell(\beta)}{\partial\beta\partial\beta^T}\right)^{-1} \frac{\partial \ell(\beta)}{\partial\beta}$$

Použijeme maticový zápis:

$$
\begin{aligned}
\mathbf{y} &= (y_1, \ldots, y_N)^T \\
\mathbf{X} &= \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,p} \\ 1 & x_{2,1} & \ldots & x_{2,p} \\ \ldots & & & \\ 1 & x_{N,1} & \ldots & x_{N,p} \end{pmatrix} \\
\mathbf{p} &= (p(x_1, \beta^{old}), \ldots, p(x_N, \beta^{old}))^T \\
\mathbf{W} &= \begin{pmatrix} p(x_1, \beta^{old})(1 - p(x_1, \beta^{old})) & 0 & \ldots & 0 \\ 0 & p(x_2, \beta^{old})(1 - p(x_2, \beta^{old})) & 0 & \ldots \\ \ldots & & & \\ 0 & \ldots & 0 & p(x_N, \beta^{old})(1 - p(x_N, \beta^{old})) \end{pmatrix}
\end{aligned}
$$

Jeden krok Newtonova algoritmu je

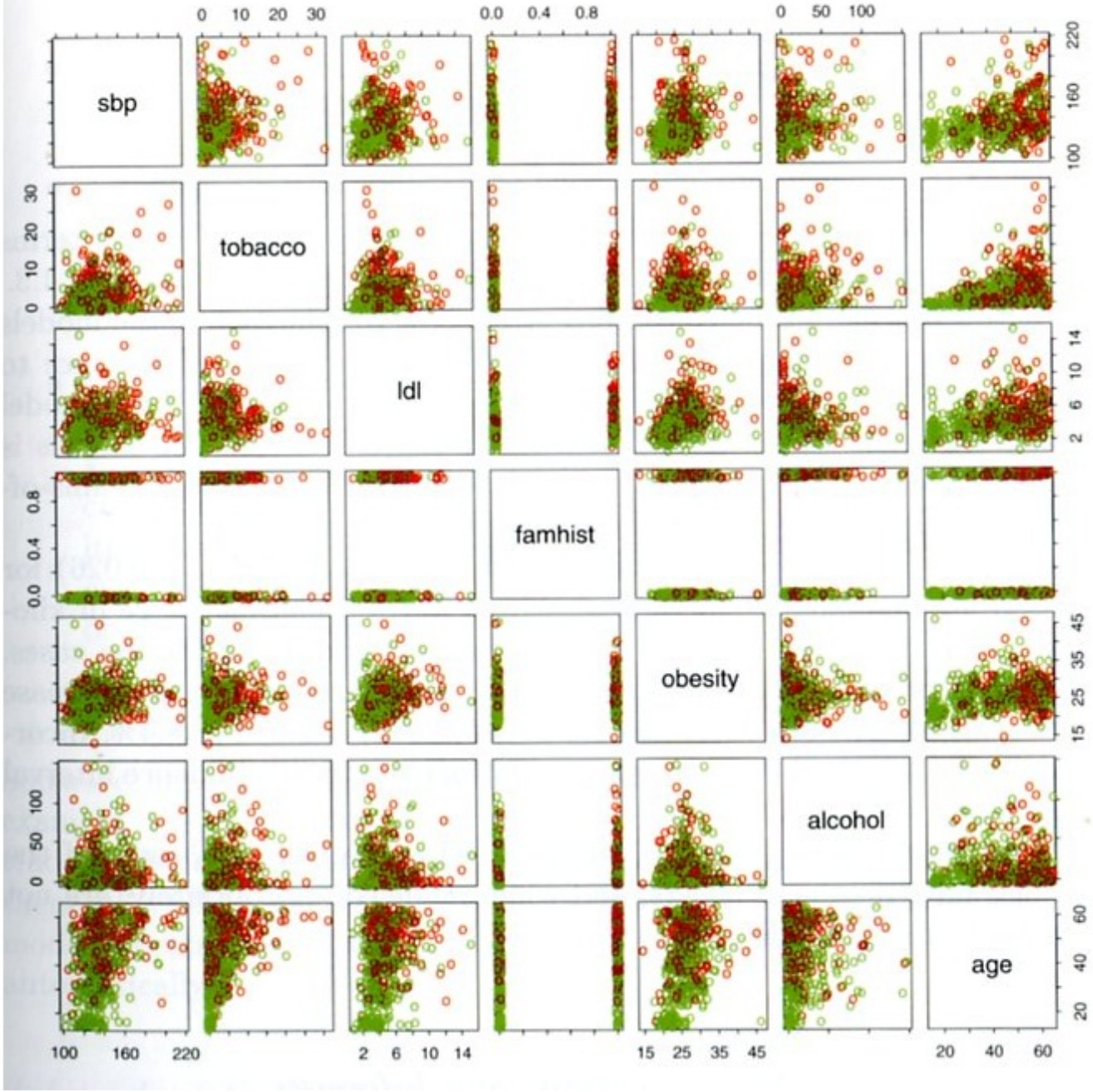
$$\begin{aligned}\beta^{new} &= \beta^{old} - (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y}-\mathbf{p}) \\ &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(\mathbf{X}\beta^{old} + \mathbf{W}^{-1}(\mathbf{y}-\mathbf{p})) \\ &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z}\end{aligned}$$

kde $\mathbf{z} = \mathbf{X}\beta^{old} + \mathbf{W}^{-1}(\mathbf{y}-\mathbf{p})$.

Tato formulace jednoho kroku algoritmu odpovídá úloze *vážených nejmenších čtverců*, neboť jeden krok algoritmu odpovídá

$$\beta^{new} = \arg\min_{\beta}(\mathbf{z}-\mathbf{X}\beta)^T\mathbf{W}(\mathbf{z}-\mathbf{X}\beta)$$

Celý algoritmus tak odpovídá metodě iterativně vážených nejmenších čtverců - *iteratively reweighted least squares (IRLS)*.


sbp
tobacco
ldl
famhist
obesity
alcohol
age
0 10 20 30
0.0 0.4 0.8
0 50 100
100 160 220
0 10 20 30
2 6 10 14
0.0 0.4 0.8
15 25 35 45
0 50 100
20 40 60

	Coefficient	Std. Error	Z Score
(Intercept)	−4.130	0.964	−4.285
sbp	0.006	0.006	1.023
tobacco	0.080	0.026	3.034
ldl	0.185	0.057	3.219
famhist	0.939	0.225	4.178
obesity	-0.035	0.029	−1.187
alcohol	0.001	0.004	0.136
age	0.043	0.010	4.184

	Coefficient	Std. Error	Z score
(Intercept)	−4.204	0.498	−8.45
tobacco	0.081	0.026	3.16
ldl	0.168	0.054	3.09
famhist	0.924	0.223	4.14
age	0.044	0.010	4.52