STATYSTYCZNE METODY REGRESJI PORZĄDKOWEJ

Marta Sommer

MiNI, Politechnika Warszawska

10 marca 2015

PRZYKŁAD

Załóżmy, że chcemy modelować zmienną odpowiedzi przyjmującą wartości ze zbioru:

$$\mathcal{Y} = \{ \textit{bardzo\_sie\_nie\_zgadzam}, \quad \textit{nie\_zgadzam\_sie}, \quad \textit{zgadzam\_sie}, \quad \textit{bardzo\_sie\_zgadzam} \}.$$

W naturalny sposób są to dane uporządkowane:

$$\textit{bardzo\_sie\_nie\_zgadzam} \prec \textit{nie\_zgadzam\_sie} \prec \textit{zgadzam\_sie} \prec \textit{bardzo\_sie\_zgadzam}.$$

Mamy też dane zmienne objaśniające - oznaczmy je standardowo przez $\mathcal{X}$.

MODEL PROPORCJONALNYCH SZANS

Interesują nas:

$$\Pi_j(\underline{x}) = \mathbb{P}(Y = j \mid \underline{x}), \qquad \text{dla} \quad j = 1, \ldots, J.$$

Rozważamy skumulowane prawdopodobieństwa:

$$\mathbb{P}(Y \leqslant j \mid \underline{x}) = \Pi_1(\underline{x}) + \ldots + \Pi_j(\underline{x}).$$

Oraz model:

$$\log \frac{\mathbb{P}(Y \leqslant j \mid \underline{x})}{\mathbb{P}(Y > j \mid \underline{x})} = \alpha_j + \underline{\beta}'\underline{x}.$$

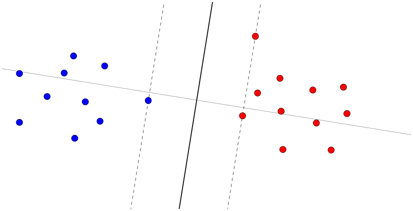
Współczynniki wyliczamy metodą Raphsona-Newtona, a szukane prawdopodobieństwa - po prostym przeliczeniu - dostaniemy ze wzoru:

$$\mathbb{P}(Y \leqslant j \mid \underline{x}) = \frac{e^{\alpha_j + \underline{\beta}'\underline{x}}}{1 + e^{\alpha_j + \underline{\beta}'\underline{x}}}.$$

WEKTORY MASZYN PODPIERAJĄCYCH (SVM)

Konstruujemy dwie równoległe i maksymalnie oddalone od siebie hiperpłaszczyzny, we wnętrzu których nie leży ani jeden element próby uczącej. Ich równania to:

$$\begin{cases} x_i'\mathbf{w} + b & = 1 \\ x_i'\mathbf{w} + b & = -1 \end{cases}$$



Odległość między hiperpłaszczynami wynosi: $\frac{2}{||\mathbf{w}||}$, więc celem jest minimalizacja $\frac{||\mathbf{w}||^2}{2}$ przy ograniczeniach:

$$\begin{cases} x_i'\mathbf{w} + b & \geqslant 1 \\ x_i'\mathbf{w} + b & \leqslant -1 \end{cases}$$

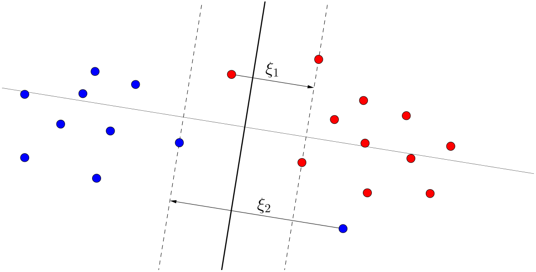
PRZYPADEK DWUKLASOWY, LINIOWO NIESEPAROWALNY

Dokładamy stałe $\xi_i \geqslant 0$ osłabiające warunek liniowej separowalności (kary za nieidealne rozdzielenie). W takiej sytuacji rozwiązujemy problem optymalizacyjny:

$$\min \left\{ \frac{1}{2} ||\mathbf{w}||^2 + C \sum_{i=1}^{n} \xi_i \right\}$$

przy ograniczeniach:

$$\begin{cases} x_i'\mathbf{w} + b & \geqslant 1 - \xi_i \\ x_i'\mathbf{w} + b & \leqslant -1 + \xi_i \end{cases}$$

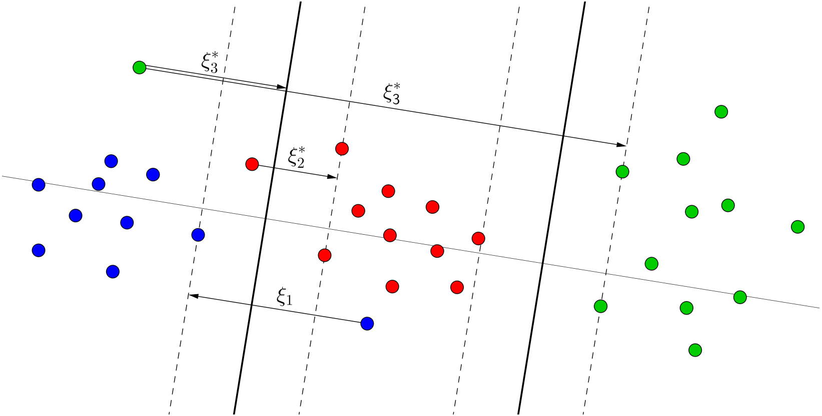


Optymalizujemy:

$$\min_{\mathbf{w},b_k}\left\{\frac{1}{2}||\mathbf{w}||^2 + C\sum_{j=1}^{J-1}\left(\sum_{k=1}^{j}\sum_{i=1}^{n_k}\xi_{ki}^{j} + \sum_{k=1}^{j}\sum_{i=1}^{n_k}\xi_{ki}^{*j}\right)\right\}$$

przy ograniczeniach:

$$\begin{cases}\mathbf{w}x_i^k - b_j & \leqslant -1 + \xi_{ki}^{j}, \qquad \text{dla} \quad k = 1, \ldots, j \quad \text{oraz} \quad i = 1, \ldots, n_k \\ \mathbf{w}x_i^k - b_j & \geqslant +1 - \xi_{ki}^{*j}, \qquad \text{dla} \quad k = j+1, \ldots, J \quad \text{oraz} \quad i = 1, \ldots, n_k\end{cases}$$



ξ_3^*
ξ_3^*
ξ_2^*
ξ_1

METODA ZAPROPONOWANA PRZEZ E. FRANKA I M. HALLA

Podejście Franka i Halla do zagadnienia regresji porządkowej opiera się nie na stworzeniu nowego modelu, ale na odpowiednim przedefiniowaniu zbioru danych, a następnie na sprowadzeniu zadania do problemu zwykłej klasyfikacji z dwoma klasami.

$J-$klasowy problem regresji porządkowej $\longrightarrow$ $J-1$ dwuklasowych problemów klasyfikacji

Na czym więc polega transformacja danych? Przeanalizujmy to na przykładzie:

$$\mathcal{Y} = \{ \quad \textit{bardzo\_sie\_nie\_zgadzam}, \quad \textit{nie\_zgadzam\_sie}, \quad \textit{zgadzam\_sie}, \quad \textit{bardzo\_sie\_zgadzam} \quad \}.$$

Dzielimy 4−klasowy problem regresji porządkowej na 3 dwuklasowe problemy klasyfikacji w następujący sposób:

$$\begin{array}{lll} \text{Zbiór nr 1:} & \textit{Cel} > & \textit{bardzo\_sie\_nie\_zgadzam} \\ \text{Zbiór nr 2:} & \textit{Cel} > & \textit{nie\_zgadzam\_sie} \\ \text{Zbiór nr 3:} & \textit{Cel} > & \textit{zgadzam\_sie} \end{array}$$

Przy czym modyfikacji ulega jedynie wektor $\mathcal{Y}$, zbiór $\mathcal{X}$ pozostaje bez zmian.

$\mathcal{X}$	$\mathcal{Y}$
1 3 5 A	*bardzo_sie_nie_zgadzam*
3 3 3 A	*bardzo_sie_nie_zgadzam*
1 4 4 F	*nie_zgadzam_sie*
2 3 3 A	*bardzo_sie_zgadzam*
5 9 1 A	*zgadzam_sie*
1 6 5 B	*bardzo_sie_nie_zgadzam*

Zbiór nr 1:

$\mathcal{X}$	*Cel*
1 3 5 A	0
3 3 3 A	0
1 4 4 F	1
2 3 3 A	1
5 9 1 A	1
1 6 5 B	0

Zbiór nr 2:

$\mathcal{X}$	*Cel*
1 3 5 A	0
3 3 3 A	0
1 4 4 F	0
2 3 3 A	1
5 9 1 A	1
1 6 5 B	0

Zbiór nr 3:

$\mathcal{X}$	*Cel*
1 3 5 A	0
3 3 3 A	0
1 4 4 F	0
2 3 3 A	1
5 9 1 A	0
1 6 5 B	0

$\mathbb{P}(Cel > \textit{bardzo\_sie\_nie\_zgadzam} | \underline{x})$

$\mathbb{P}(Cel > \textit{nie\_zgadzam\_sie} | \underline{x})$

$\mathbb{P}(Cel > \textit{zgadzam\_sie} | \underline{x})$

Następnie dostajemy nową obserwację o zmiennych objaśnianych $\underline{x}$ i chcemy dowiedzieć się, do jakiej klasy z $\mathcal{Y}$ będzie ona należała. Co robimy? Korzystając z poprzednich modeli, liczymy po kolei:

$$\begin{aligned}
\mathbb{P}(\mathit{Cel} &> \mathit{bardzo_sie_nie_zgadzam} \mid \underline{x}) \\
\mathbb{P}(\mathit{Cel} &> \mathit{nie_zgadzam_sie} \mid \underline{x}) \\
\mathbb{P}(\mathit{Cel} &> \mathit{zgadzam_sie} \mid \underline{x})
\end{aligned}$$

Kolejne prawdopodobieństwa wylicza się łańcuchowo. Mianowicie:

$$\begin{aligned}
\mathbb{P}(\mathit{Cel} = \mathit{bardzo_sie_nie_zgadzam} \,|\, \underline{x}) = &\quad 1 - \mathbb{P}(\mathit{Cel} > \mathit{bardzo_sie_nie_zgadzam} \,|\, \underline{x}) \\
\mathbb{P}(\mathit{Cel} = \mathit{nie_zgadzam_sie} \,|\, \underline{x}) = &\quad \mathbb{P}(\mathit{Cel} > \mathit{bardzo_sie_nie_zgadzam} \,|\, \underline{x}) - \\
&\quad \mathbb{P}(\mathit{Cel} > \mathit{nie_zgadzam_sie} \,|\, \underline{x}) \\
\mathbb{P}(\mathit{Cel} = \mathit{zgadzam_sie} \,|\, \underline{x}) = &\quad \mathbb{P}(\mathit{Cel} > \mathit{nie_zgadzam_sie} \,|\, \underline{x}) - \\
&\quad \mathbb{P}(\mathit{Cel} > \mathit{zgadzam_sie} \,|\, \underline{x}) \\
\mathbb{P}(\mathit{Cel} = \mathit{bardzo_sie_zgadzam} \,|\, \underline{x}) = &\quad \mathbb{P}(\mathit{Cel} > \mathit{bardzo_sie_zgadzam} \,|\, \underline{x})
\end{aligned}$$

Ostatecznie, naszej nowej obserwacji przypisujemy klasę
z maksymalnym prawdopodobieństwem.

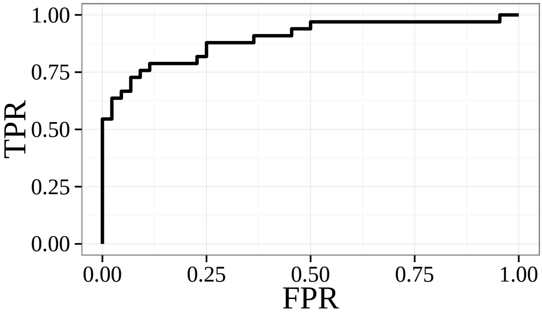
KRZYWA ROC I WSPÓŁCZYNNIK AUC

PRZYPADEK DWUKLASOWY

		Klasa	
		1	0
$\widehat{Klasa}$	1	TP	FP
	0	FN	TN

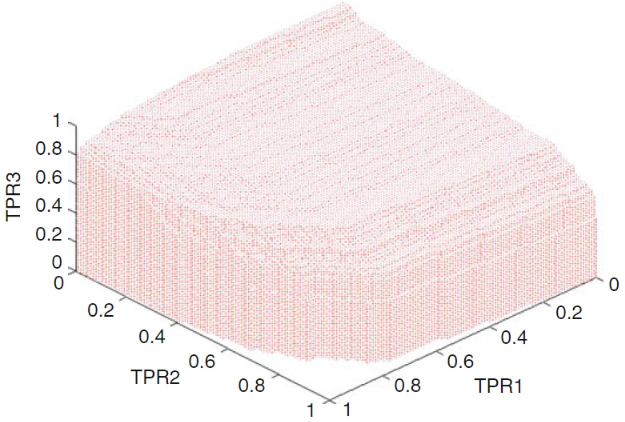
$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{TN}{FP + TN}$$



Wykres 1: Przykładowa krzywa ROC

PRZYPADEK KLASYFIKACJI PORZĄDKOWEJ



$$VUS = \frac{1}{\prod_{k=1}^{J} n_k} \sum_{y_{j_1} < \ldots < y_{j_J}} \mathbb{I}_{f(x_{j_1}) < \ldots < f(x_{j_J})}$$

Bibliografia

- Frank E., Hall M., A simple approach to ordinal classification, *Proceedings of the European Conference on Machine Learning*, Freibourg, Niemcy, 2001, str. 146–156.
- Waegman W., De Baets B., A survey on ROC-based ordinal regression, w: Fürnkranz J., Hüllermeier E. (Eds.), *Preference Learning*, Springer, 2010, str. 127-154.