Matematičke metode u marketingu. Klasifikacijski modeli. Binarni klasifikator

Lavoslav Čaklović
PMF-MO

2016

Primjer

Na temelju ispitivanja tržišta predvidjeti kupnju proizvoda:

	god	*spol*	*primanja*	vježba	*kupnja*
Učenje	35	*M*	$[2,3)$	*red*	*NE*
	45	*F*	$[4,5)$	*pov*	*DA*
	51	*F*	$[10,\infty)$	*ne*	*NE*

	god	*spol*	*primanja*	vježba	*kupnja*
Predikcija	38	*F*	5	*pov*	?

Problem spada u domenu strojnog učenja i danas je sastavni dio marketinških analiza. Varijabla *kupnja* je binarna (za početak).

Preživljavanje.

X	Y	$\hat{Y}$
x_1	0	1
x_2	1	1
x_3	0	0
$\vdots$	$\vdots$	$\vdots$
x_m	1	0

Zadano: $\Omega := \{1, \ldots, m\}$ – skup ideksa (subjekata), $X : \Omega \to \mathbb{R}$ i $Y : \Omega \to \{0, 1\}$ dihotomna ($\{ne, da\}$).
Za $x_0 \in range(X)$ definiramo $\hat{Y} : \Omega \to \{0, 1\}$:

$$x_i \geqslant x_0 \iff \hat{y}_i = 1,$$

Smisleno je tražiti x_0 ali tako da $\hat{Y}$ 'najbolje' aproksimira Y. Što znači najbolje, u kom smislu?

Preživljavanje.

X	Y	$\hat{Y}$
x_1	0	1
x_2	1	1
x_3	0	0
$\vdots$	$\vdots$	$\vdots$
x_m	1	0

Zadano: $\Omega := \{1, \ldots, m\}$ – skup ideksa (subjekata), $X : \Omega \to \mathbb{R}$ i $Y : \Omega \to \{0, 1\}$ dihotomna ($\{ne, da\}$). Za $x_0 \in range(X)$ definiramo $\hat{Y} : \Omega \to \{0, 1\}$:

$$x_i \geqslant x_0 \iff \hat{y}_i = 1,$$

Smisleno je tražiti x_0 ali tako da $\hat{Y}$ 'najbolje' aproksimira Y. Što znači najbolje, u kom smislu?

Zbunj-matrica[1].

	$Y \backslash \hat{Y}$	predikcija 1	predikcija 0
sada	1	n_{11}	n_{12}
	0	n_{21}	n_{22}

n_{11} . . . broj korektno prepoznatih 1.
n_{12} . . . broj nekorektno prepoznatih 1.
n_{21} . . . broj nekorektno prepoznatih 0.
n_{22} . . . broj korektno prepoznatih 0.

[1] eng. confusion matrix (CM)

Preživljavanje.

X	Y	$\hat{Y}$
x_1	0	1
x_2	1	1
x_3	0	0
$\vdots$	$\vdots$	$\vdots$
x_m	1	0

Zadano: $\Omega := \{1, \ldots, m\}$ – skup ideksa (subjekata), $X : \Omega \to \mathbb{R}$ i $Y : \Omega \to \{0, 1\}$ dihotomna ($\{ne, da\}$). Za $x_0 \in range(X)$ definiramo $\hat{Y} : \Omega \to \{0, 1\}$:

$$x_i \geqslant x_0 \iff \hat{y}_i = 1,$$

Smisleno je tražiti x_0 ali tako da $\hat{Y}$ 'najbolje' aproksimira Y. Što znači najbolje, u kom smislu?

Zbunj-matrica[1].

		predikcija $\hat{Y}$	
	$Y \setminus \hat{Y}$	1	0
sada	1	TP	FN
	0	FP	TN

TP ... True Positive
FN ... False Negative.
FP ... False Positive.
TN ... True Negative.

$$TPR = \frac{TP}{TP+FN} \text{ (osjetljivost)} \qquad TNR = \frac{TN}{FP+TN} \text{ (specifičnost)}$$

[1] eng. confusion matrix (CM)

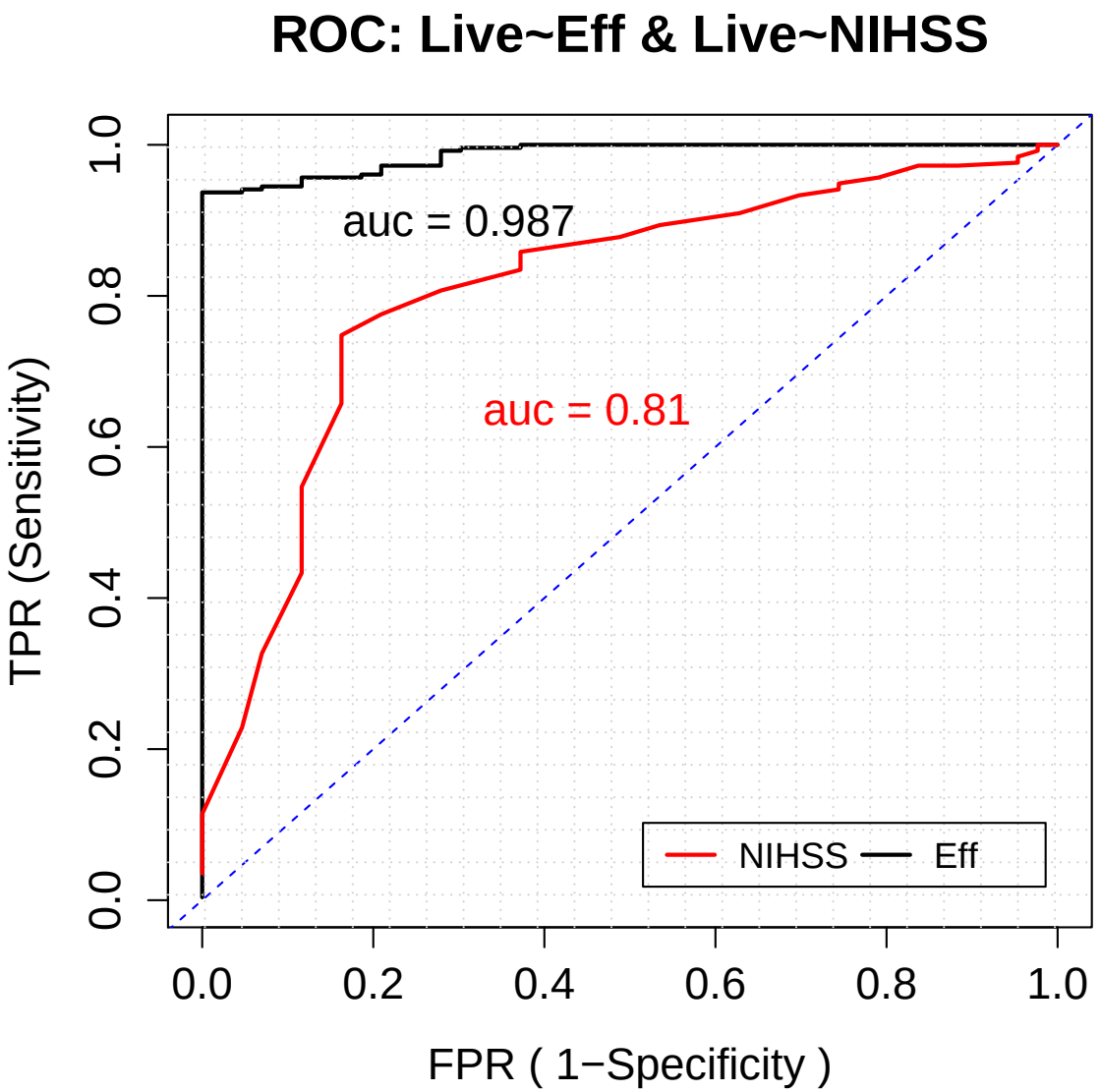
Confusion matrix[2] - bis

		Predicted condition		
	Total population	Predicted Condition positive	Predicted Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$
True condition	condition positive	**True positive**	**False Negative** (Type II error)	True positive rate (TPR), Sensitivity, Recall $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$
	condition negative	**False Positive** (Type I error)	**True negative**	False positive rate (FPR), Fall-out $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$
	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$
		False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$

False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$
True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$
Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$

[2] izvor: Wikipedia

ROC[3] krivulja



Eff	*NIHSS*	*Y*
89	40	1
46	25	1
12	5	0
38	13	0
49	17	1
⋮	⋮	⋮

$Eff \geqslant e \to CM_e \to T_e$

$T_e := (FPR_e, TPR_e)$

e raste $\implies$ T_e se giba po krivulji od $(1,1)$ do $(0,0)$.

Koji je najbolji prerez e?
Koja krivulja je bolja?

Zadatak. Interpretirajte dijagonalu ROC dijagrama.

Točka na dijagonali pripada matrici za koju je $TPR = FPR$, a takva matrica podjednako prepoznaje i dobre i loše kao dobre (slučajnost).

Kriterij optimalnosti prereza (preciznost). Prihvatljiv kriterij je

$$\frac{n_{11} + n_{22}}{n_{11} + n_{21} + n_{12} + n_{22}} \to \max . \qquad (*)$$

Pitanje. Što se dešava kad kriterij (*) primijenimo na

$$M_1 = \begin{bmatrix} 20 & 60 \\ 60 & 30 \end{bmatrix} \text{ ili } M_2 = \begin{bmatrix} 40 & 40 \\ 80 & 10 \end{bmatrix} ?$$

Ako $Y = 1$ predstavlja opasnost, onda instrument M_1 ne prepoznaje opasnost, a instrument M_2 je indiferentan na opasnost, a diže uzbunu kad slučaj nije opasan.

Definirajte sami svoj kriterij i testirajte ga.

Označimo s $T(x, y)$ točku na ROC krivulji. Tri su glavna kriterija optimalnosti te točke:

1. Youdenov index[4]
2. DDIC
3. Fisher, McNemar, Barnard

Yodenova metoda odabire $T(x, y)$ tako da je njena udaljenost od glavne dijagonale najveća. Tangenta na ROC krivulju u T paralelna je s glavnom dijagonalom. Ekvivalentno:

$$\begin{aligned} y - x &\to \max \\ \underbrace{1 - y + x}_{d := \|(0,1), T\|} &\to \min \\ sens + spec &\to \max \\ d &\to \min \end{aligned}$$

[4] Schisterman, Perkins, Liu, Bondel (2005)

Označimo s $T(x, y)$ točku na ROC krivulji. Tri su glavna kriterija optimalnosti te točke:

1. Youden
2. DDIC[4]
3. Fisher, McNemar, Barnard

DDIC metoda) odabire točku T na presjeku ROC krivulje i sporedne dijagonale.

[4] Descending Diagonal Intersection Criterion, (Leal, Oliviera, Sanchez (2012)

Označimo s $T(x, y)$ točku na ROC krivulji. Tri su glavna kriterija optimalnosti te točke:

1. Youden
2. DDIC
3. Fisher[4], McNemar, Barnard

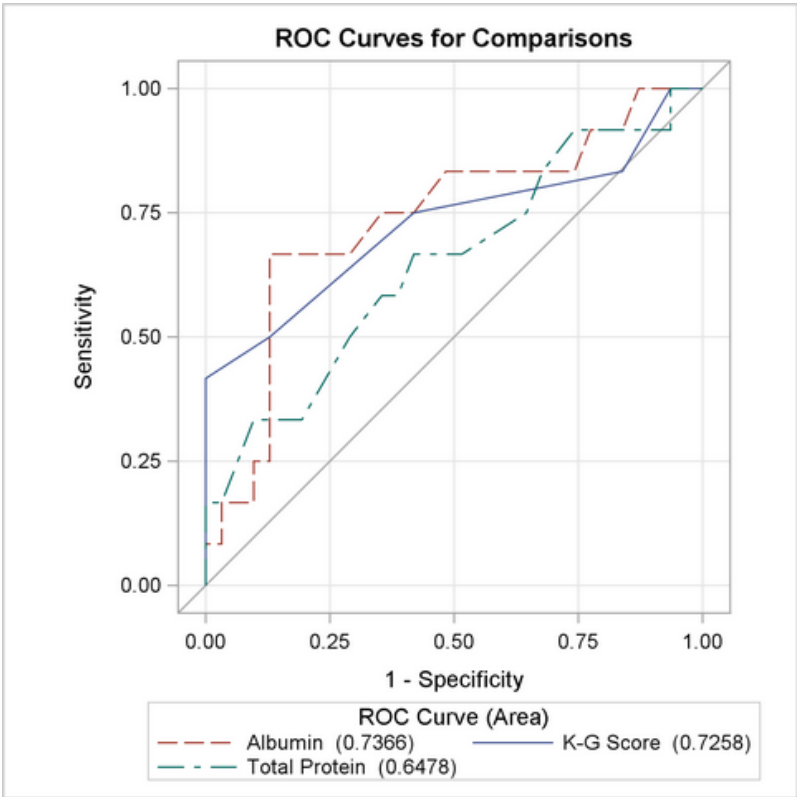
$Y \backslash \hat{Y}$	1	0	
1	a	b	r_1
0	c	d	r_2
	c_1	c_2	N

Vjerojatnost ove tablice je $p = \frac{\binom{r_1}{a}\binom{r_2}{c}}{\binom{N}{c_1}}$. **Fisher** testira 0-hipotezu: $p/(1-p) = 1$ nasuprot alternativnoj $p/(1-p) > 1$, r_i, c_i su fiksni. **McNemar** testira jednakost marginalnih frekvencija, tj. 0-hipotezu: $p_a + p_b = p_a + p_c$ $(p_b = p_c)$ nasuprot $p_b \neq p_c$. Statistika: $\chi^2 = \frac{(b-c)^2}{b+c}$. **Barnard** testira nezavisnost stupaca, tj. 0-hipotezu: $p_{c_1} = p_{c_2}$, a testna statistika (mjera odstupanja) je Waldova T-statistika za omjere (ratio).

Najbolja zbunj-matrica nalazi se maksimiziranjem testne statistike.

[4] Agresti A. Categorical data analysis (1992)

Uspoređivanje ROC krivulja



Najpoznatija mjera za uspoređivanje je AUC[5]. NIJE dobra u slučaju kad su krivulje nedominirane kao na slici.

AUC ima nedostatak što koristi različite 'misclassification cost' za različite klasifikatore. U tom slučaju H-mjera (Hand 2009) se pokazuje pogodnijom (R-ov paket `hmeasure`).

Ostale mjere su: Ginijev koeficijent, AUCH, KS.

[5] Area Under the Curve

Zadaci

– Učitati podatke `mu.data.csv` i za svaku varijablu A, B nacrtajte ROC krivulju. Odaberite najbolji prerez po Youdenovom kriteriju. Koja je varijabla bolja? Usporedite svoje rezultate s rezultatima nekog od `R`-ovih paketa.