Posebnosti binarne (dvojiške) klasifikacije

Ljupčo Todorovski
Univerza v Ljubljani, Fakulteta za upravo

Marec 2018

Pregled predavanja

Dvojiško razvrščanje (binarna klasifikacija)

- Dve vrsti napak pri binarni klasifikaciji
- Tabela napačnih klasifikacij
- Slabovidnost napovedne napake in alternative

Prostor ROC

- Cenovno občutljiva klasifikacija in izohipse
- Krivulja ROC in ploščina pod krivuljo

Prostor PR (preciznost in priklic)

- Krivulja PR in ploščina pod krivuljo
- Mera F in alternative

Naloga klasifikacije z $D_Y = \{0, 1\}$

Klasifikacija (razvrščanje) v dva razreda

- $Y = 1$: pozitivni primeri
- $Y = 0$ (včasih tudi $Y = -1$): negativni primeri

Dve vrsti klasifikacijskih modelov

- Napoved razreda: $m : \bigtimes_{i=1}^{p} D_i \to D_Y$
- Napoved verjetnosti pozitivnega razreda: $m : \bigtimes_{i=1}^{p} D_i \to [0, 1]$

Dve vrsti napak

Običajna napaka za funkcijo izgube $L_{01}(y, \hat{y}) = 1 - I(y, \hat{y})$

$$Err(m, S) = \frac{1}{S} \sum_{(\boldsymbol{x}, y) \in S} L_{01}(y, m(\boldsymbol{x}))$$

izmerjena torej na modelu, ki napoveduje razred in ne verjetnosti.

L_{01} ne razlikuje med napakama

1. napačno razvrščeni negativni primer (*false positive*, *FP*): zavračanje veljavne hipoteze; zaznavanje pojava, ki ga ni; lažni alarm
2. napačno razvrščeni pozitivni primer (*false negative*, *FN*): sprejetje neveljavne hipoteze

Tabela napačnih klasifikacij (*confusion matrix*)

	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	pravilno razvrščeni pozitivni primeri *true positives*, TP	nepravilno razvrščeni negativni primeri *false positives*, FP
$\hat{Y} = 0$	nepravilno razvrščeni pozitivni primeri *false negatives*, FN	pravilno razvrščeni negativni primeri *true negatives*, TN
Σ	pozitivni primeri, P	negativni primeri, N

$$P = TP + FN, N = FP + TN, n = P + N = TP + FN + FP + TN$$

Tabela napačnih klasifikacij in napovedna napaka

Napovedna napaka $Err = (FP + FN)/n$

Idealni model $Err = 0$

Napovedna točnost (*accuracy*) $Acc = (TP + TN)/n$

$Acc = 1 - Err$, idealni model $Acc = 1$

Slabovidnost mer *Acc* in *Err*

Ne upoštevata napaki 1 in 2 ter njihovo razmerje,
zato veliko zelo različnih modelov imajo isto napako (točnost).

Trije enako točni modeli: $Err = 0.3, Acc = 0.7$

$|S| = n = 100,\ P = 50,\ N = 50$

	M_1		M_2		M_3	
$\hat{Y} = 1$	50	30	20	0	35	15
$\hat{Y} = 0$	0	20	30	50	15	35

Dva pristranska in nepristranski model

1. M_1 je "optimističen", napove 80% pozitivnih primerov
2. M_2 je "pesimističen", napove le 20% pozitivnih primerov
3. M_3 je nepristranski, napove 50% pozitivnih primerov; prav tak je delež pozitivnih primerov v množici S

Bolj občutljivi meri

Delež pravilno razvrščenih pozitivnih primerov $TPR = TP/P$

- *True Positive Rate*
- Občutljivost, senzitivnost (*sensitivity*) ali priklic (*recall*)
- Točnost v množici pozitivnih primerov

Delež napačno razvrščenih negativnih primerov $FPR = FP/N$

- *False Positive Rate*
- 1−specifičnost (*specificity*) ali izpad (*fall out*)
- Napaka v množici negativnih primerov

Trije enako točni modeli: $Err = 0.3, Acc = 0.7$

$|S| = n = 100,\ P = 50,\ N = 50$

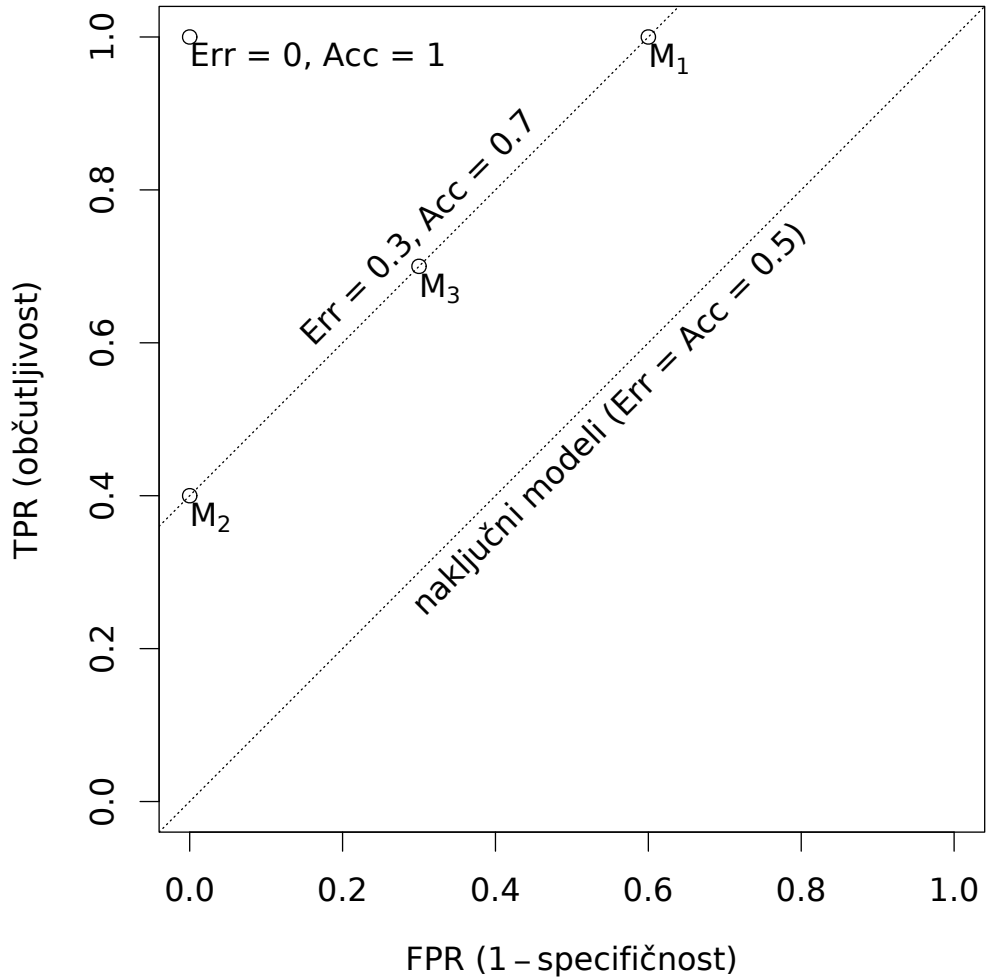
	M_1		M_2		M_3	
$\hat{Y} = 1$	50	30	20	0	35	15
$\hat{Y} = 0$	0	20	30	50	15	35

Bolj občutljivi meri

1. $TPR(M_1) = 50/50 = 1,\ FPR(M_1) = 30/50 = 0.6$
2. $TPR(M_2) = 20/50 = 0.4,\ FPR(M_2) = 0/50 = 0$
3. $TPR(M_3) = 35/50 = 0.7,\ FPR(M_3) = 15/50 = 0.3$

FPR (x os) in *TPR* (y os) razpenjata prostor ROC

ROC = *Receiver Operating Characteristic*



Izbira modela in cenovno-občutljiva klasifikacija

Kako izbrati med M_1, M_2 in M_3?

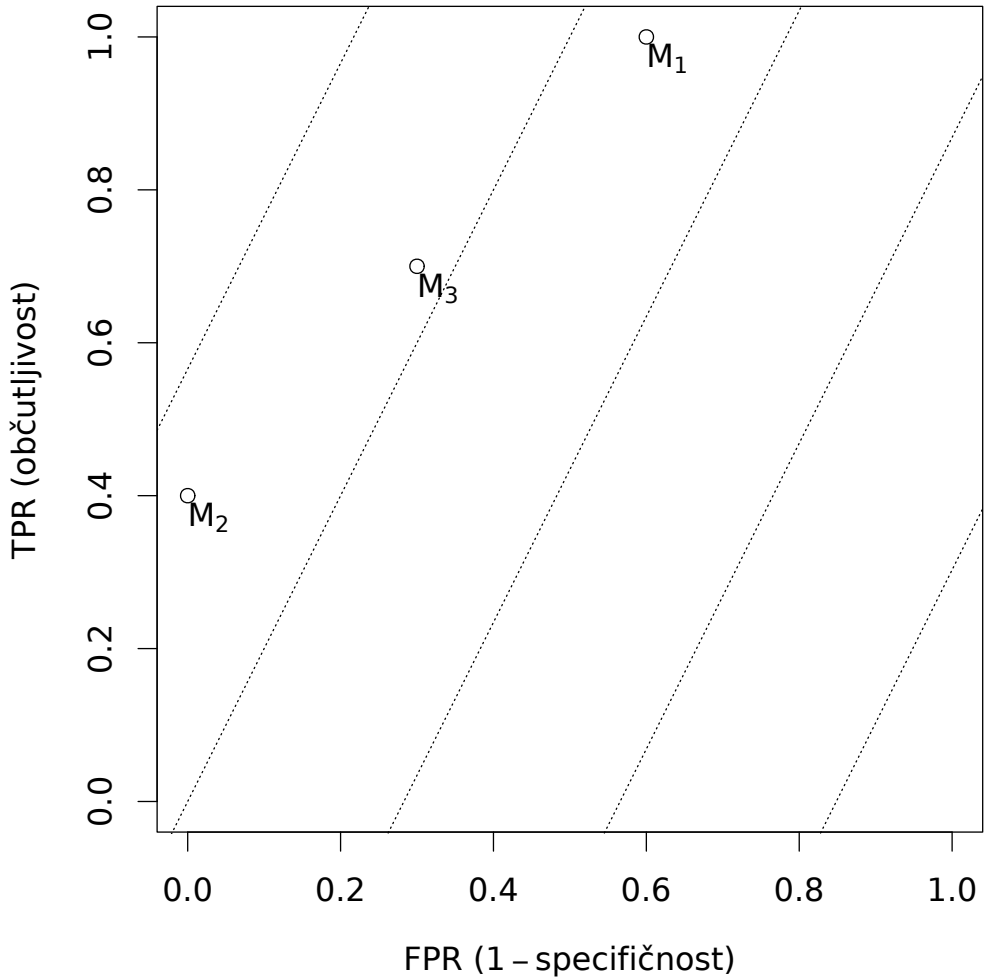
- Če sta napaki tipa 1 in 2 enakovredni, so tudi modeli enakovredni.
- Kaj pa če nista?

Cenovno-občutljiva klasifikacija: cenovna matrika

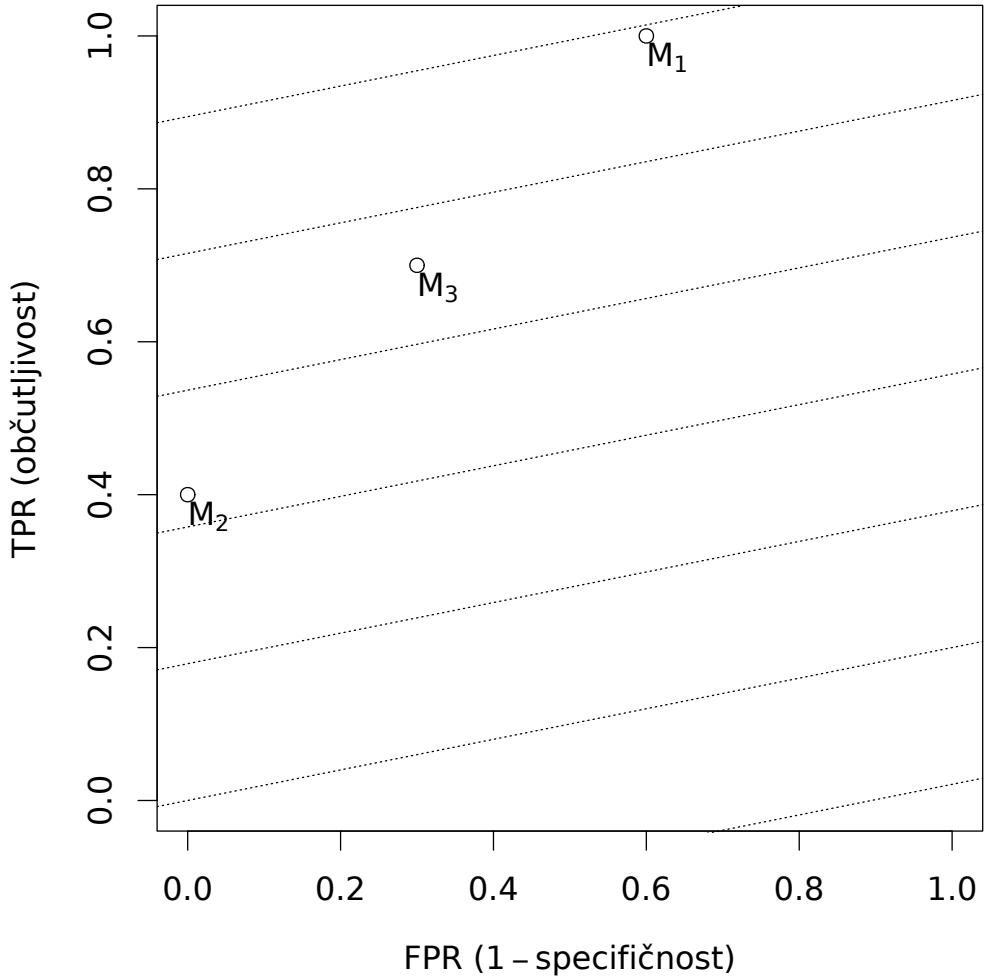
	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	0	2
$\hat{Y} = 0$	1	0

- Napaka tipa 1 (FP) je 2 krat hujša kot napaka tipa 2 (FN).
- Kateri model izberemo v tem primeru?

Izbira modela M_2 pri $FP : TP = 2 : 1$



Izbira modela M_1 pri $FP : TP = 1 : 2$



Napovedovanje verjetnosti in prag odločitve

Napovedni model za verjetnost pozitivnega razreda

$$m(\boldsymbol{x}) = p(Y = 1 | X = \boldsymbol{x})$$

Za napoved razreda rabimo še prag θ

$$\hat{Y}_\theta = I(m(\boldsymbol{x}) \geq \theta)$$

Običajni izbor $\theta = 0.5$.

Kako izberemo optimalno vrednost θ?

Narišemo modele za vse možne vrednosti θ v prostoru ROC.

Izračun krivulje ROC za model M

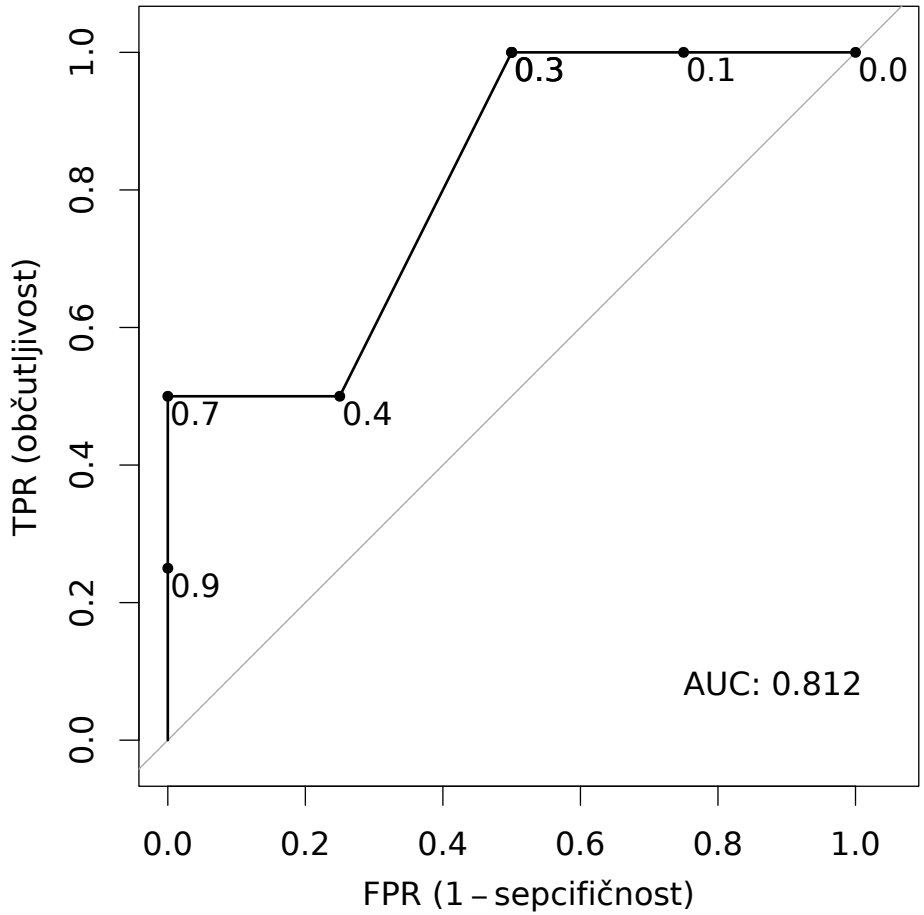
X	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	TPR	FPR
Y	0	0	1	1	0	0	1	1		
$m(X)$	0.0	0.1	0.3	0.3	0.3	0.4	0.7	0.9		
$\hat{Y}_{1.0}$	0	0	0	0	0	0	0	0	0.00	0.00
$\hat{Y}_{0.9}$	0	0	0	0	0	0	0	1	0.25	0.00
$\hat{Y}_{0.7}$	0	0	0	0	0	0	1	1	0.50	0.00
$\hat{Y}_{0.4}$	0	0	0	0	0	1	1	1	0.50	0.25
$\hat{Y}_{0.3}$	0	0	1	1	1	1	1	1	1.00	0.50
$\hat{Y}_{0.1}$	0	1	1	1	1	1	1	1	1.00	0.75
$\hat{Y}_{0.0}$	1	1	1	1	1	1	1	1	1.00	1.00

Izračun posameznih točk krivulje ROC

X	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Y	0	0	1	1	0	0	1	1
$m(X)$	0.0	0.1	0.3	0.3	0.3	0.4	0.7	0.9

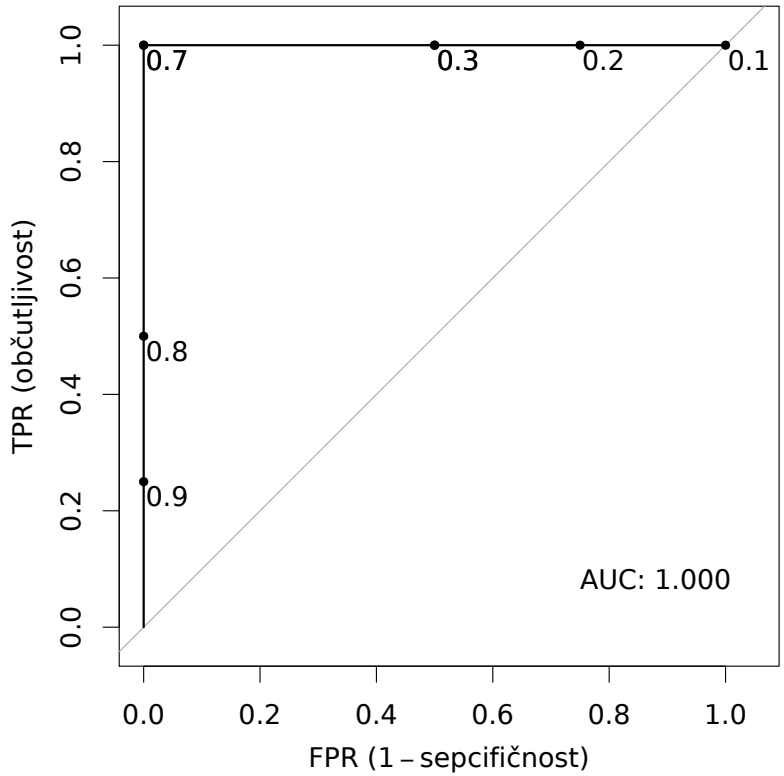
	$\theta = 1.0$		$\theta = 0.9$		$\theta = 0.4$		$\theta = 0.3$	
$\hat{Y} = 1$	0	0	1	0	2	1	4	2
$\hat{Y} = 0$	4	4	3	4	2	3	0	2
	0.00	0.00	0.25	0.00	0.50	0.25	1.00	0.50

Krivulja ROC za model *M*



Krivulja ROC za idealni model

X	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Y	0	0	0	0	1	1	1	1
$m(X)$	0.1	0.2	0.3	0.3	0.7	0.7	0.8	0.9



Krivulja ROC je monotono (in ne strogo) naraščajoča

Če zmanjšujemo vrednost θ

Narašča število napovedi $\hat{Y} = 1$.

Narašča torej $TP + FP$

In posledično (N in P sta konstantni) tudi $TPR + FPR = TP/P + FP/N$.

Ker sta TPR in FRP nenegativni se lahko zgodi troje

1. Vertikalni segment: poveča se le TPR
2. Horizontalni segment: poveča se le FPR
3. Poševni segment: povečata se hkrati TPR in FPR

AUC: Ploščina pod krivuljo ROC

$$AUC(m) = \int_{\theta} FPR(m_\theta)\, TPR(m_\theta)\, d\theta$$

Mera točnosti modela, ki napoveduje verjetnost

- Izračun ne zahteva odločitve o vrednosti θ
- Omogoča primerjavo točnosti modelov, ki napovedujejo verjetnost

Zaloga vrednosti $[0, 1]$; pomembne vrednosti

1. 1: model, ki idealno loči pozitivne od negativnih primerov
2. 0.5: naključni model
3. 0: model, ki tudi idealno loči pozitivne od negativnih primerov, a vse pozitivne primere razglasi za negativne in obratno

Interpretacija *AUC* in krivulj ROC

Kaj pove *AUC*?

$$AUC(m) = p(m(\boldsymbol{x}^+) > m(\boldsymbol{x}^-))$$

$\boldsymbol{x}^+$ (oz. $\boldsymbol{x}^-$) je naključno izbran pozitiven (oz. negativen) primer: je ocena razdalje med napovedmi za pozitivne in napovedmi za negativne primere.

Česa *AUC* ne pove?

- Ne ponuja jasne interpretacije ekspertu s področja uporabe
- Veliki del ploščine "nezanimiv" zaradi previsokega *FPR*
- Zato raje krivulje ROC primerjamo vizualno

Nekaj praktičnih napotkov

Kdaj boste uporabljali krivulje ROC in AUC?

- AUC lahko zamenja napovedno točnost pri izbiri modela
- Caret podpira uporabo AUC za primerjavo modelov
- Pri podatkovnih množicah z neenakomerno porazdelitvijo Y: poiščemo optimalni model oz. optimalni prag odločanja

Preciznost (*precission, positive predictive value, PPV*)

$$PPV = TP/(TP + FP)$$

- Delež pozitivno razvrščenih primerov ($\hat{Y} = 1$), ki so dejansko pozitivni ($Y = 1$)
- Tudi: verjetnost, da je pozitivna napoved pravilna

Pogosto jo opazujemo skupaj s priklicem $TPR = TP/P$

- *True Positive Rate*
- Občutljivost, senzitivnost (*sensitivity*) ali priklic (*recall*)
- Točnost v množici pozitivnih primerov

Trije enako točni modeli: $Err = 0.3, Acc = 0.7$

$|S| = n = 100,\ P = 50,\ N = 50$

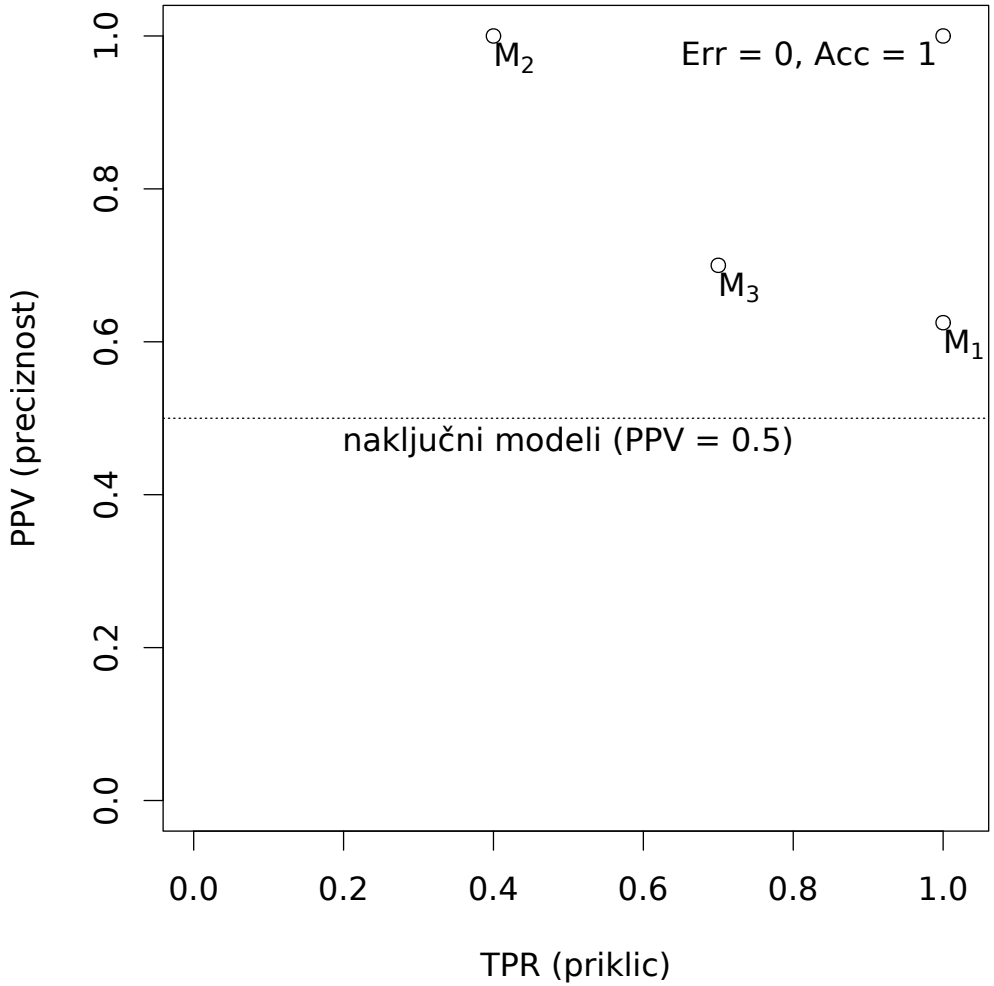
	M_1		M_2		M_3	
$\hat{Y} = 1$	50	30	20	0	35	15
$\hat{Y} = 0$	0	20	30	50	15	35

Preciznost in priklic

1. $PPV(M_1) = 50/80 = 0.625,\ TPR(M_1) = 50/50 = 1$
2. $PPV(M_2) = 20/20 = 1,\ TPR(M_2) = 20/50 = 0.4$
3. $PPV(M_3) = 35/50 = 0.7,\ TPR(M_3) = 35/50 = 0.7$

TPR (x os) in *PPV* (y os) razpenjata prostor PR

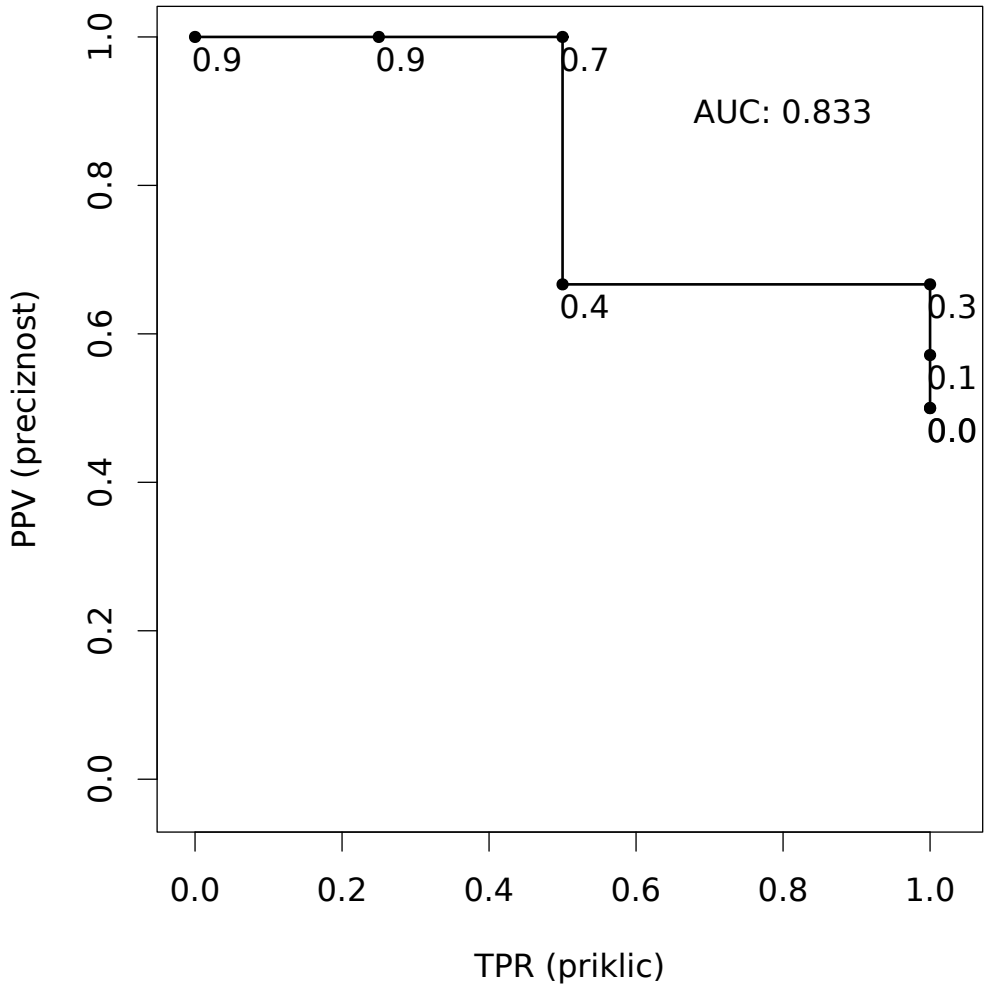
PR = *Precission-Recall*



Izračun krivulje PR za model m

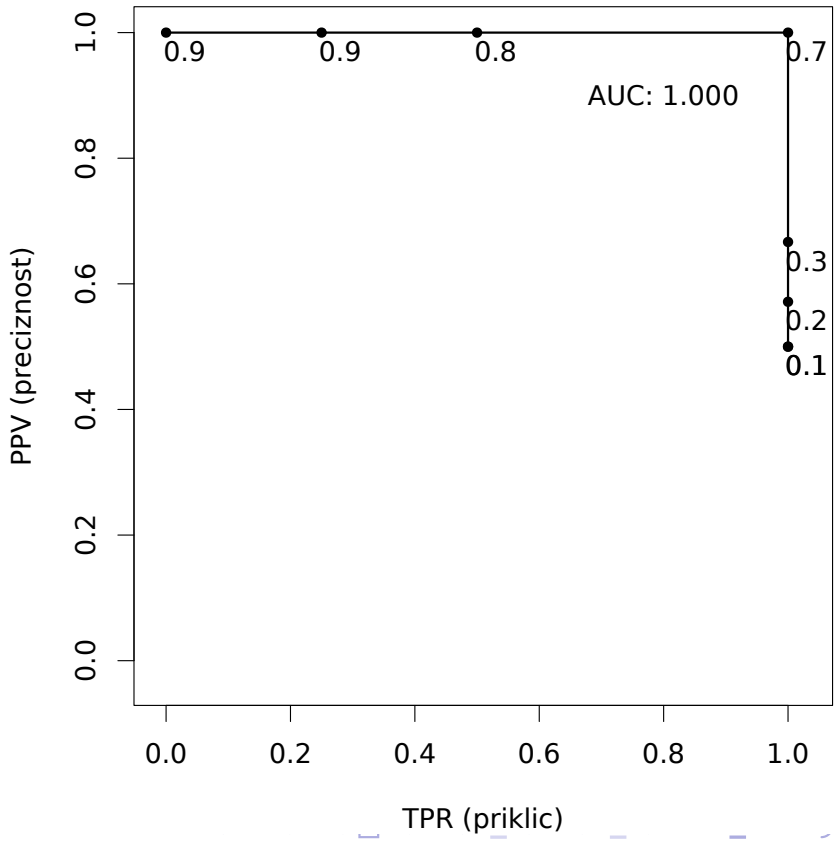
X	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	TPR	PPV
Y	0	0	1	1	0	0	1	1		
$m(X)$	0.0	0.1	0.3	0.3	0.3	0.4	0.7	0.9		
$\hat{Y}_{1.0}$	0	0	0	0	0	0	0	0	0.00	NaN
$\hat{Y}_{0.9}$	0	0	0	0	0	0	0	1	0.25	1.00
$\hat{Y}_{0.7}$	0	0	0	0	0	0	1	1	0.50	1.00
$\hat{Y}_{0.4}$	0	0	0	0	0	1	1	1	0.50	0.67
$\hat{Y}_{0.3}$	0	0	1	1	1	1	1	1	1.00	0.67
$\hat{Y}_{0.1}$	0	1	1	1	1	1	1	1	1.00	0.57
$\hat{Y}_{0.0}$	1	1	1	1	1	1	1	1	1.00	0.50

Krivulja PR za model *M*



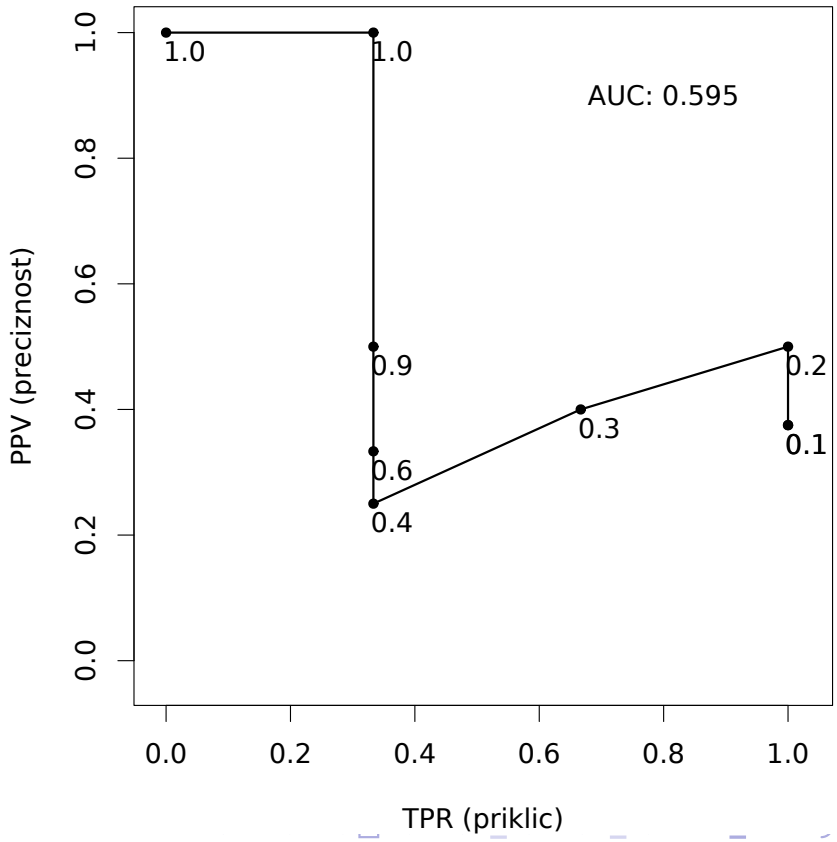
Krivulja PR za idealni model

X	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Y	0	0	0	0	1	1	1	1
$m(X)$	0.1	0.2	0.3	0.3	0.7	0.7	0.8	0.9



Krivulja PR ni nujno monotona: proti-primer

X	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Y	0	0	1	1	0	0	0	1
$m(X)$	0.1	0.1	0.2	0.3	0.4	0.6	0.9	1.0



AUC_{PR}: Ploščina pod krivuljo PR

$$AUC_{PR}(m) = \int_{\theta} PPV(m_\theta)\, TPR(m_\theta)\, d\theta$$

Alternativna mera točnosti modela

- Izračun ne zahteva odločitve o vrednosti θ
- Omogoča primerjavo točnosti modelov, ki napovedujejo verjetnost

Zaloga vrednosti $[0, 1]$; pomembne vrednosti

1. 1: model, ki idealno loči pozitivne od negativnih primerov
2. 0.5: naključni model

Mere F

$$F_1 = \frac{2}{\frac{1}{PPV} + \frac{1}{TPR}}$$

Alternativna formula za F_1

$$F_1 = 2\,\frac{PPV \cdot TPR}{PPV + TPR}$$

Zelo pogost način kombinacije preciznosti in priklica v enotno mero točnosti modela: maksimalno vrednost 1 doseže idealni model

Posplošitev F_β

$$F_\beta = (1 + \beta^2)\,\frac{PPV \cdot TPR}{\beta^2 \cdot PPV + TPR}$$

Slabosti prostora PR in mere F ter alternative

Kaj povesta AUC_{PR} in F?

Bore malo, pa še ne upošteva pravilno razvrščene negativne primere (TN).

Alternativa: Matthewsov korelacijski koeficient za klasifikacijo

$$MCC = \frac{TP \cdot TN + FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$