Istraživanje podataka 1

Vežbe 4

21. mart 2021

Outline

Outline

C5.0

- koristi informacionu dobit (mera nečistoće entropija)
- binarna podela kada se numerički atribut koristi za test
- za kategoričke atribute podrazumevana podela - jedna vrednost jedna grana, a vrednosti mogu i da se grupišu

Opis nekih opcija

- korišćenje podeljenog skupa (trening i test skup)
- grupisanje kategoričkih podataka
- *boosting* - pravljenje više modela u nizu radi povećanja preciznosti. Prvi model se pravi na uobičajen način, a svaki sledeći se fokusira na instance koje su pogrešno klasifikovane prethodnim modelom. Za klasifikaciju instance se primenjuju svi modeli i koristi se sistem glasanja.
- *unakrsna-validacija* - pravljenje modela nad podskupovima radi procene preciznosti modela napravljenim nad celim skupom

Opis nekih opcija

- opcija za naklonjenost ka preciznosti ili uopštenosti modela
- očekivan procenat instanci sa greškom u trening skupu
- *strogost pri potkresivanju* - povećanjem vrednosti dobija se manje stablo
- minimalan broj instanci koji mora da bude u dete-čvoru nakon podele da bi se izvršila podela
- *winnow atributes* - izračunavanje važnosti atributa pre pravljenja modela
- matrica cene pogrešne klasifikacije

Outline

Zadatak

Primeniti klasifikaciju nad skupom *iris.* korišćenjem C5.0.

Outline

Klasifikacija - mere za ocenu modela

Najčešće korišćene mere za ocenu modela:

- *preciznost* $= \frac{\text{Broj slogova čija klasa je dobro predviđena modelom}}{\text{Ukupan broj slogova}}$ (eng. accuracy)
- stopa greške $= \frac{\text{Broj slogova čija klasa nije dobro predviđena modelom}}{\text{Ukupan broj slogova}}$ (eng. error rate)

Nisu dovoljne za skupove podataka sa neuravnoteženim klasama.

Klasifikacija - mere za ocenu modela

Za bolji uvid kako se model ponaša za svaku klasu koristi se matrica konfuzije.

Tabela: Matrica konfuzije za 4 klase

		Dodeljena klasa			
		C_1	C_2	C_3	C_4
Stvarna klasa	C_1				
	C_2		x		
	C_3				
	C_4	y			

Klasifikacija - mere za ocenu modela

- Postoje mere koje daju bolji uvid kako se model ponaša za svaku klasu.
- Pri binarnoj klasifikaciji, retka klasa se označava kao pozitivna a većinska klasa kao negativna

Klasifikacija - matrica konfuzije

		Dodeljena klasa	
		+	-
Stvarna klasa	+	TP	FN
	-	FP	TN

- *TP* (eng. true positive) - broj instanci pozitivne klase koje je model pravilno klasifikovao
- *FN* (eng. false negative) - broj instanci pozitivne klase koje je model pogrešno klasifikovao
- *FP* (eng. false positive) - broj instanci negativne klase koje je model pogrešno klasifikovao
- *TN* (eng. true negative) - broj instanci negativne klase koje je model pravilno klasifikovao

Klasifikacija - mere za ocenu modela

- stopa stvarno pozitivnih ili osetljivost (eng. true positive rate, sensitivity) $TPR = \frac{TP}{TP+FN}$
- stopa stvarno negativnih ili specifičnost (eng. true negative rate, specificity) $TNR = \frac{TN}{FP+TN}$
- stopa lažno pozitivnih (eng. false positive rate) $FPR = \frac{FP}{FP+TN}$
- stopa lažno negativnih (eng. false negative rate) $FNR = \frac{FN}{TP+FN}$

Klasifikacija - mere za ocenu modela

- preciznost (eng. precision) $p = \frac{TP}{TP+FP}$
  Za skup sa više klasa preciznost se računa za svaku klasu C_i sa

$$p = \frac{\text{broj instanci klase } C_i \text{ kojima model dodeljuje klasu } C_i}{\text{broj instanci kojima model dodeljuje klasu } C_i}$$

- odziv (eng. recall) $r = \frac{TP}{TP+FN}$
  Za skup sa više klasa odziv se računa za svaku klasu C_i sa

$$r = \frac{\text{broj instanci klase } C_i \text{ kojima model dodeljuje klasu } C_i}{\text{broj instanci klase } C_i}$$

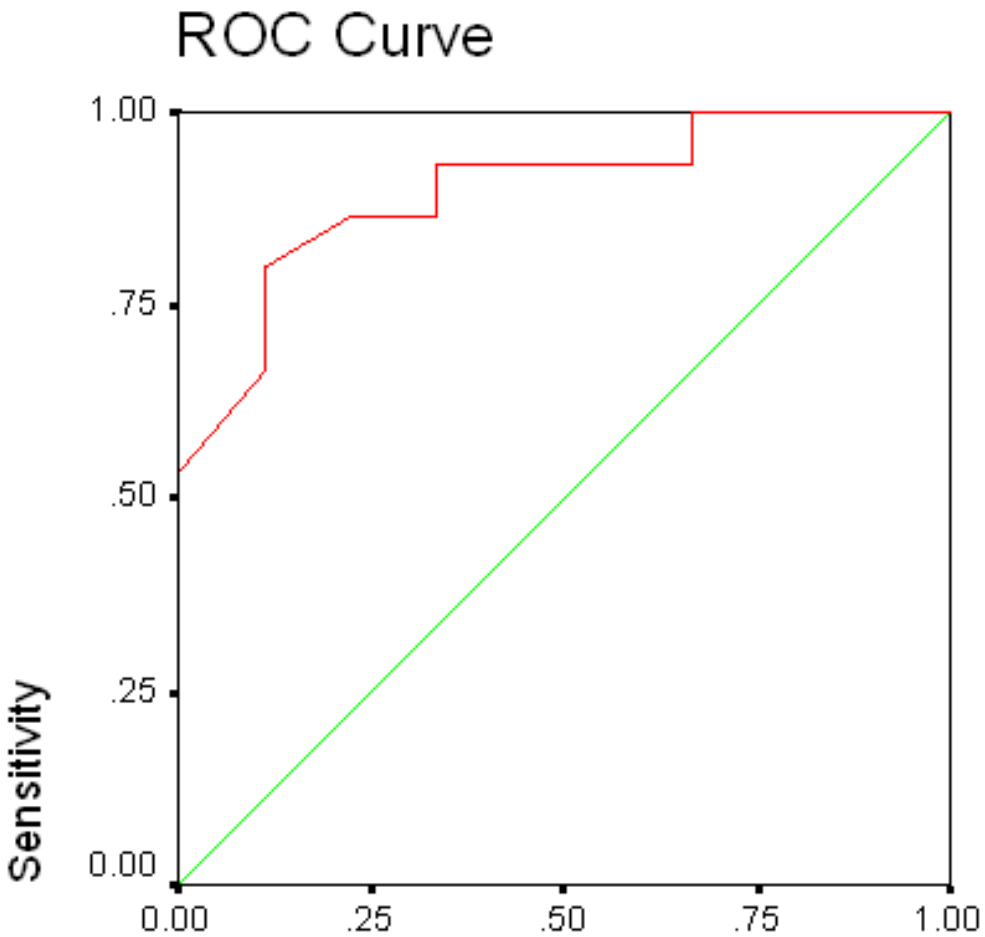
Klasifikacija - mere za ocenu modela

- F_1 uzima u obzir preciznost i odziv $F_1 = \frac{2rp}{r+p} = \frac{2}{\frac{1}{r}+\frac{1}{p}}$
  - harnomijska sredina preciznosti i odziva
  - bliža je manjoj vrednosti

Klasifikacija - mere za ocenu modela

- ROC kriva (receiver operating characteristic curve)
  - grafički prikaz kompromisa između *TPR* i *FPR*
  - *x* osa - *FPR* ili 1-specifičnost
  - *y* osa - *TPR* ili osetljivost

Klasifikacija - mere za ocenu modela



Klasifikacija - ROC kriva

- Interpretacija određenih tačaka
  - (TPR=0 i FPR=0)

Klasifikacija - ROC kriva

- Interpretacija određenih tačaka
  - (TPR=0 i FPR=0) - model svakoj instanci dodeljuje negativnu klasu

Klasifikacija - ROC kriva

- Interpretacija određenih tačaka
  - (TPR=0 i FPR=0) - model svakoj instanci dodeljuje negativnu klasu
  - (TPR=1 i FPR=1)

Klasifikacija - ROC kriva

- Interpretacija određenih tačaka
  - (TPR=0 i FPR=0) - model svakoj instanci dodeljuje negativnu klasu
  - (TPR=1 i FPR=1) - model svakoj instanci dodeljuje pozitivnu klasu

Klasifikacija - ROC kriva

- Interpretacija određenih tačaka
  - (TPR=0 i FPR=0) - model svakoj instanci dodeljuje negativnu klasu
  - (TPR=1 i FPR=1) - model svakoj instanci dodeljuje pozitivnu klasu
  - (TPR=1 i FPR=0)

Klasifikacija - ROC kriva

- Interpretacija određenih tačaka
  - (TPR=0 i FPR=0) - model svakoj instanci dodeljuje negativnu klasu
  - (TPR=1 i FPR=1) - model svakoj instanci dodeljuje pozitivnu klasu
  - (TPR=1 i FPR=0) - idealan model

Klasifikacija - ROC kriva

- Interpretacija određenih tačaka
  - (TPR=0 i FPR=0) - model svakoj instanci dodeljuje negativnu klasu
  - (TPR=1 i FPR=1) - model svakoj instanci dodeljuje pozitivnu klasu
  - (TPR=1 i FPR=0) - idealan model
- *AUC* (eng. area under the ROC curve) - površina ispod ROC krive

Zadatak

Proceniti performanse dva klasifikaciona modela, M_1 i M_2. Skup podataka sadrži 26 binarnih atributa, označenih od A do Z. Tabela prikazuje posteriorne verovatnoće za pozitivnu klasu dobijene primenom modela na skup podataka. Nacrtati ROC krive.

Zadatak

Instance	True Class	$P(+\|A,\ldots,Z,M_1)$	$P(+\|A,\ldots,Z,M_2)$
1	+	0.73	0.61
2	+	0.69	0.03
3	−	0.44	0.68
4	−	0.55	0.31
5	+	0.67	0.45
6	+	0.47	0.09
7	−	0.08	0.38
8	−	0.15	0.05
9	+	0.45	0.01
10	−	0.35	0.04

Rešenje zadatka - korak 1

Verovatnoća	**0,08**	0,15	0,35	0,44	0,45	0,47	0,55	0,67	0,69	0,73	
Prava klasa	-	-	-	-	+	+	-	+	+	+	
Dodeljena klasa	+	+	+	+	+	+	+	+	+	+	
TP	5										
FP	5										
TN	0										
FN	0										
TPR	**1**										
FPR	**1**										

Rešenje zadatka - korak 2

Verovatnoća	0,08	**0,15**	0,35	0,44	0,45	0,47	0,55	0,67	0,69	0,73	
Prava klasa	-	-	-	-	+	+	-	+	+	+	
Dodeljena klasa	-	+	+	+	+	+	+	+	+	+	
TP	5	5									
FP	5	4									
TN	0	1									
FN	0	0									
TPR	1	**1**									
FPR	1	**0,8**									

Rešenje zadatka - korak 3

Verovatnoća	0,08	0,15	**0,35**	0,44	0,45	0,47	0,55	0,67	0,69	0,73	
Prava klasa	-	-	-	-	+	+	-	+	+	+	
Dodeljena klasa	-	-	+	+	+	+	+	+	+	+	
TP	5	5	5								
FP	5	4	3								
TN	0	1	2								
FN	0	0	0								
TPR	1	1	**1**								
FPR	1	0,8	**0,6**								

Rešenje zadatka - korak 4

Verovatnoća	0,08	0,15	0,35	**0,44**	0,45	0,47	0,55	0,67	0,69	0,73	
Prava klasa	-	-	-	-	+	+	-	+	+	+	
Dodeljena klasa	-	-	-	+	+	+	+	+	+	+	
TP	5	5	5	5							
FP	5	4	3	2							
TN	0	1	2	3							
FN	0	0	0	0							
TPR	1	1	1	**1**							
FPR	1	0,8	0,6	**0,4**							

Rešenje zadatka - korak 5

Verovatnoća	0,08	0,15	0,35	0,44	**0,45**	0,47	0,55	0,67	0,69	0,73	
Prava klasa	-	-	-	-	+	+	-	+	+	+	
Dodeljena klasa	-	-	-	-	+	+	+	+	+	+	
TP	5	5	5	5	5						
FP	5	4	3	2	1						
TN	0	1	2	3	4						
FN	0	0	0	0	0						
TPR	1	1	1	1	**1**						
FPR	1	0,8	0,6	0,4	**0,2**						

Rešenje zadatka - korak 6

Verovatnoća	0,08	0,15	0,35	0,44	0,45	**0,47**	0,55	0,67	0,69	0,73	
Prava klasa	-	-	-	-	+	+	-	+	+	+	
Dodeljena klasa	-	-	-	-	-	+	+	+	+	+	
TP	5	5	5	5	5	4					
FP	5	4	3	2	1	1					
TN	0	1	2	3	4	4					
FN	0	0	0	0	0	1					
TPR	1	1	1	1	1	**0,8**					
FPR	1	0,8	0,6	0,4	0,2	**0,2**					

Rešenje zadatka - korak 7

Verovatnoća	0,08	0,15	0,35	0,44	0,45	0,47	**0,55**	0,67	0,69	0,73	
Prava klasa	-	-	-	-	+	+	-	+	+	+	
Dodeljena klasa	-	-	-	-	-	-	+	+	+	+	
TP	5	5	5	5	5	4	3				
FP	5	4	3	2	1	1	1				
TN	0	1	2	3	4	4	4				
FN	0	0	0	0	0	1	2				
TPR	1	1	1	1	1	0,8	**0,6**				
FPR	1	0,8	0,6	0,4	0,2	0,2	**0,2**				

Rešenje zadatka - korak 8

Verovatnoća	0,08	0,15	0,35	0,44	0,45	0,47	0,55	**0,67**	0,69	0,73	
Prava klasa	-	-	-	-	+	+	-	+	+	+	
Dodeljena klasa	-	-	-	-	-	-	-	+	+	+	
TP	5	5	5	5	5	4	3	3			
FP	5	4	3	2	1	1	1	0			
TN	0	1	2	3	4	4	4	5			
FN	0	0	0	0	0	1	2	2			
TPR	1	1	1	1	1	0,8	0,6	**0,6**			
FPR	1	0,8	0,6	0,4	0,2	0,2	0,2	**0**			

Rešenje zadatka - korak 9

Verovatnoća	0,08	0,15	0,35	0,44	0,45	0,47	0,55	0,67	**0,69**	0,73	
Prava klasa	-	-	-	-	+	+	-	+	+	+	
Dodeljena klasa	-	-	-	-	-	-	-	-	+	+	
TP	5	5	5	5	5	4	3	3	2		
FP	5	4	3	2	1	1	1	0	0		
TN	0	1	2	3	4	4	4	5	5		
FN	0	0	0	0	0	1	2	2	3		
TPR	1	1	1	1	1	0,8	0,6	0,6	**0,4**		
FPR	1	0,8	0,6	0,4	0,2	0,2	0,2	0	**0**		

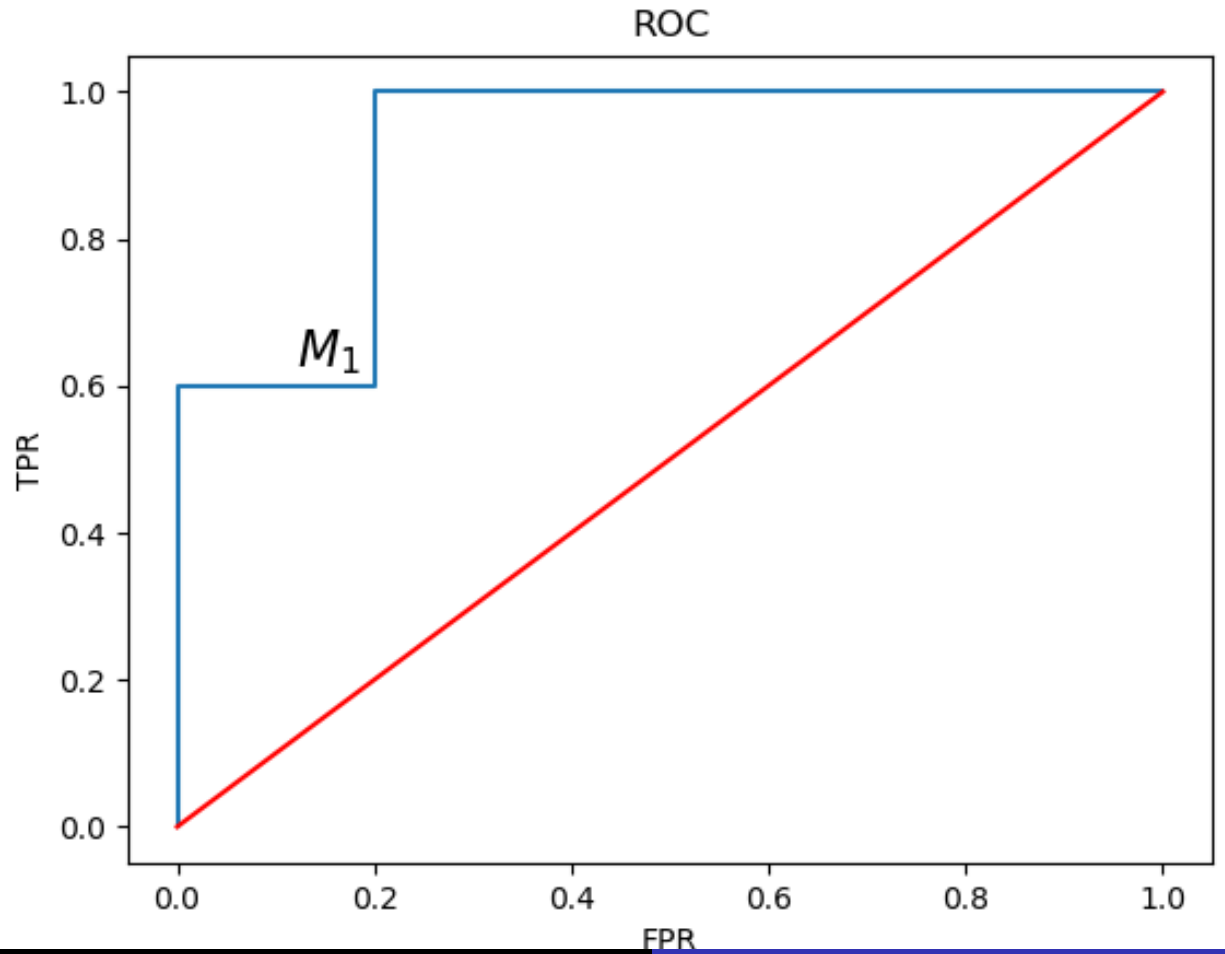
Rešenje zadatka - korak 10

Verovatnoća	0,08	0,15	0,35	0,44	0,45	0,47	0,55	0,67	0,69	**0,73**	
Prava klasa	-	-	-	-	+	+	-	+	+	+	
Dodeljena klasa	-	-	-	-	-	-	-	-	-	+	
TP	5	5	5	5	5	4	3	3	2	1	
FP	5	4	3	2	1	1	1	0	0	0	
TN	0	1	2	3	4	4	4	5	5	5	
FN	0	0	0	0	0	1	2	2	3	4	
TPR	1	1	1	1	1	0,8	0,6	0,6	0,4	**0,2**	
FPR	1	0,8	0,6	0,4	0,2	0,2	0,2	0	0	**0**	

Rešenje zadatka - korak 11

Verovatnoća	0,08	0,15	0,35	0,44	0,45	0,47	0,55	0,67	0,69	0,73	
Prava klasa	-	-	-	-	+	+	-	+	+	+	
Dodeljena klasa	-	-	-	-	-	-	-	-	-	-	
TP	5	5	5	5	5	4	3	3	2	1	0
FP	5	4	3	2	1	1	1	0	0	0	0
TN	0	1	2	3	4	4	4	5	5	5	5
FN	0	0	0	0	0	1	2	2	3	4	5
TPR	1	1	1	1	1	0,8	0,6	0,6	0,4	0,2	**0**
FPR	1	0,8	0,6	0,4	0,2	0,2	0,2	0	0	0	**0**

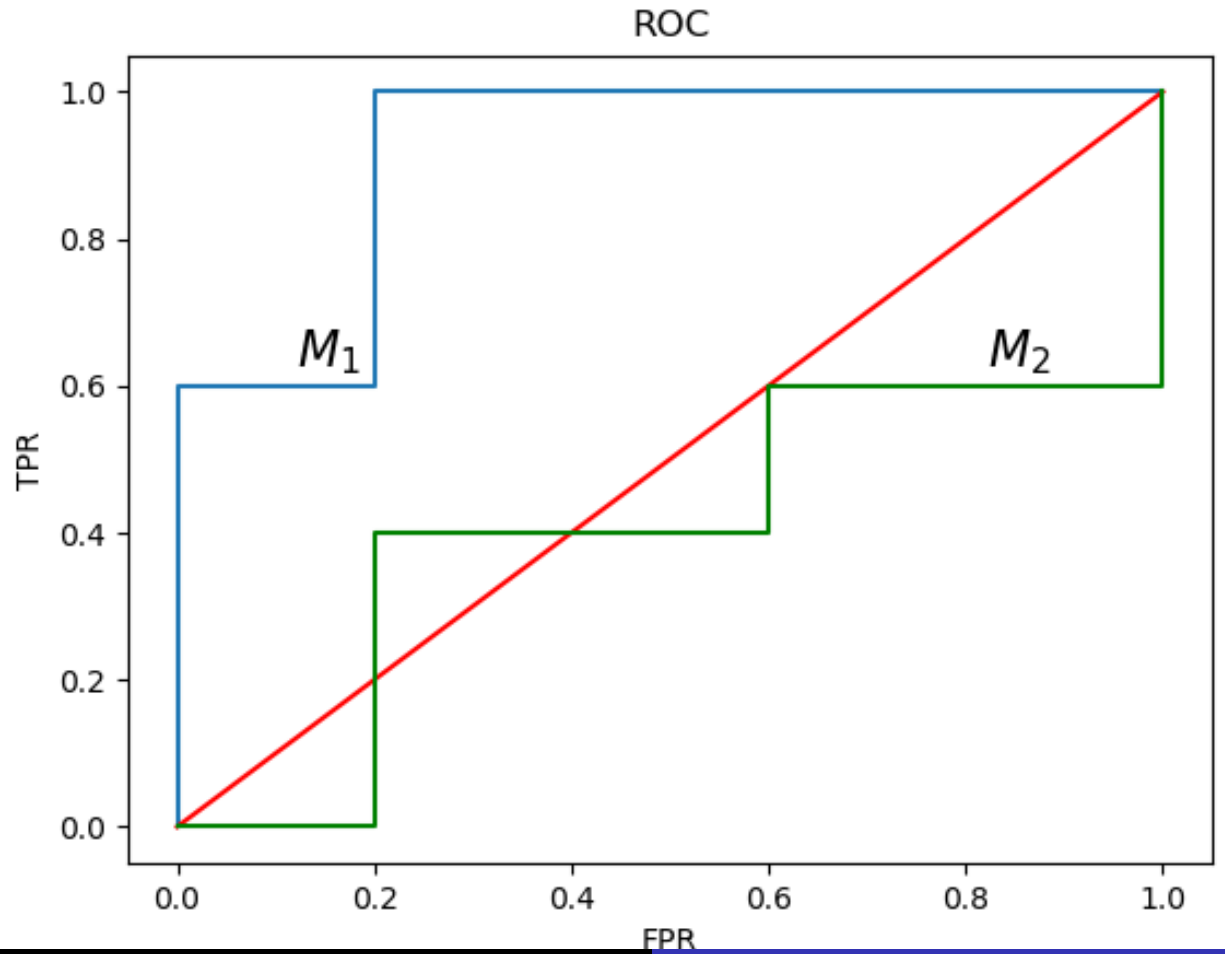
Rešenje zadatka - ROC kriva za model M_1



Rešenje zadatka - tabela za model M_2

Verovatnoća	0,01	0,03	0,04	0,05	0,09	0,31	0,38	0,45	0,61	0,68	
Klasa	+	+	-	-	+	-	-	+	+	-	
TP	5	4	3	3	3	2	2	2	1	0	0
FP	5	5	5	4	3	3	2	1	1	1	0
TN	0	0	0	1	2	2	3	4	4	4	5
FN	0	1	2	2	2	3	3	3	4	5	5
TPR	1	0,8	0,6	0,6	0,6	0,4	0,4	0,4	0,2	0	0
FPR	1	1	1	0,8	0,6	0,6	0,4	0,2	0,2	0,2	0

Rešenje zadatka - ROC krive za modele M_1 i M_2



Outline

Zadatak

Primeniti klasifikaciju nad skupom *bank.csv* korišćenjem C5.0. Ciljni atribut je oročena štednja.

- Koji atributi su korišćeni pri pravljenju modela?
- Komentarisati dobijen model. Dati predlog za poboljšanje.

Outline

Matrica cene

- U praktičnoj primeni modela klasifikacije, pogrešna klasifikacija instanci jedne klase (ili jedne grupe klasa) može biti skuplja od druge.
- Matrica cena u klasifikaciji omogućava da korisnik definiše značaj za različite greške predviđanja i te vrednosti se uzimaju u obzir pri pravljenju modela.
- Matrica cena izgleda kao matrica konfuzije i prikazuje cenu za svaku moguću kombinaciju stvarne i dodeljene klase.

Matrica cene

Tabela: Podrazumevane vrednosti u matrici cena za 4 klase

		Dodeljena klasa			
		C_1	C_2	C_3	C_4
Stvarna klasa	C_1	0	1	1	1
	C_2	1	0	1	1
	C_3	1	1	0	1
	C_4	1	1	1	0

Matrica cene

Tabela: Primer matrice cena za 4 klase

		Dodeljena klasa			
		C_1	C_2	C_3	C_4
Stvarna klasa	C_1	0	1	1	1
	C_2	2	0	2	2
	C_3	1	1	0	1
	C_4	1	1	1	0

Cena se menja samo za pogrešna predviđanja, a za dobra predviđanja uvek ostaje 0.

Outline

Zadatak

Skup podataka *klasifikacija.txt* sadrži podatke o osobama koje su donirale krv. Koristeći alat IBM SPSS Modeler izvršiti klasifikaciju primenom drveta odlučivanja (algoritam C5.0) nad skupom. Ciljni atribut je *Class*. Model označiti sa *model1*.
U tekućem mesecu je nestašica krvi zbog čega će volonteri lično morati da zovu registrovane davaoce krvi kako bi što pre prikupili dovoljno krvi. Napraviti model koji bi volonterima bio od pomoći u odabiru koga da zovu. Dobijeni model označiti sa *model2*.

Zadatak

Podatke o dobijenim modelima (preciznost i matrice konfuzije na trening i test skupu) sačuvajte u html datotekama. U komentarima opisati dobijene modele i navesti koji su atributi najznačajniji za pravljenje modela. Navesti koje ste sve parametre postavili, koje vrednosti ste zadali i zašto.
Radni tok eksportovati i dodeliti mu ime u formatu **SPSS_klasifikacija_*vasBrojIndeksa***. Odgovore pišite u datoteku sa nazivom **SPSS_klasifikacija_*vasBrojIndeksa_odgovori***.

Zadatak

Opis atributa skupa:

- *Recency* : broj meseci od poslednje donacije krvi
- *Frequency* : ukupan broj donacija
- *Monetary* : ukupna krv donirana u cm^3
- *Time* : broj meseci od prve donacije
- *Class*: da li je donor krvi u tekućem mesecu (1 - da, 0 - ne)