

Mašinsko učenje 1

MAS Informatike – Nauka o podacima

Evaluacija modela klasifikacije

- Evaluacija modela vrši se na osnovu mere njegove sposobnosti predviđanja.
- Mere kvaliteta klasifikacije zasnivaju se na matrici konfuzije (eng. confusion matrix)

Matrica konfuzije

- Element c_{ij} matrice konfuzije predstavlja broj primera klase i koji su klasifikovani u klasu j .
- Klasifikacija je najbolja kada je ova matrica dijagonalna.
- Nedijagonalni elementi označavaju greške.

		Klasa određena modelom		
		C1	C2	C3
Stvarna klasa	C1	5	1	0
	C2	1	4	1
	C3	2	0	4

Matrica konfuzije - binarna klasifikacija

		Klasa određena modelom	
		DA	NE
Stvarna klasa	DA	Stvarno Pozitivni TP	Lažno Negativni FN
	NE	Lažno Pozitivni FP	Stvarno Negativni TN

- U slučaju binarne klasifikacije jedna klasa se naziva pozitivnom, a druga negativnom.
- Stvarno pozitivni (**true positive - TP**) instance su pozitivne instance koje su od strane modela prepoznate kao pozitivne.
- Stvarno negativne (**true negative - TN**) instance su negativne instance koje su od strane modela prepoznate kao negativne.
- Lažno pozitivne (**false positive - FP**) instance su negativne instance koje su od strane modela proglašene pozitivnim.
- Lažno negativne (**false negative - FN**) instance su pozitivne instance koje su od strane modela proglašene negativnim.

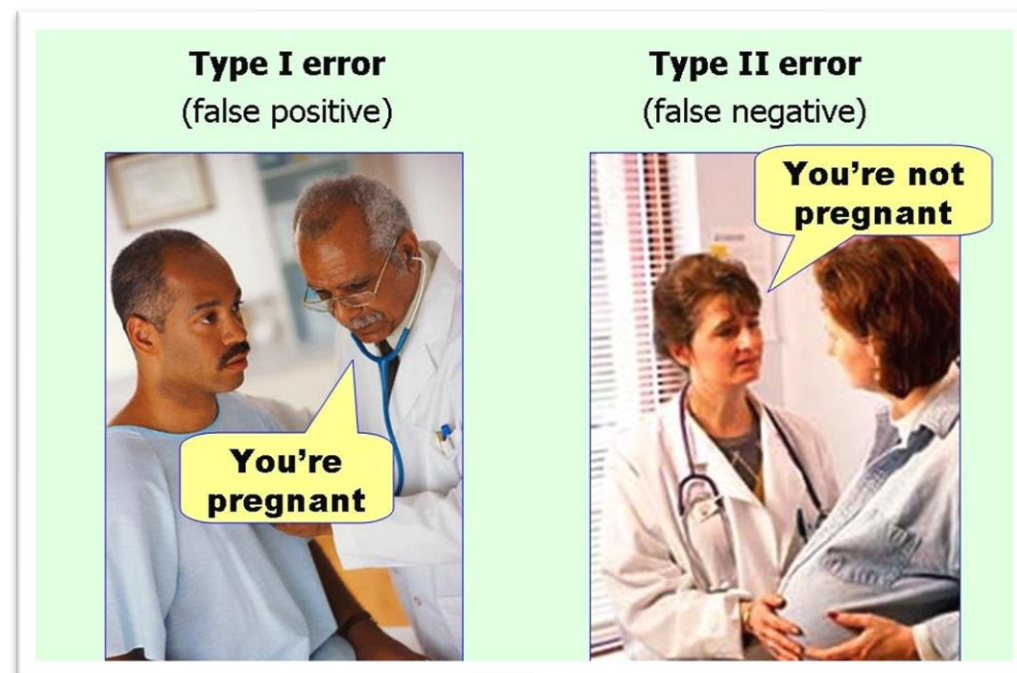
Konfuziona matrica - Primer

		Klasa određena modelom		
		DA	NE	
Stvarna klasa	DA	TP = 40	FN = 20	Pozitivni = 60
	NE	FP = 15	TN = 35	Negativni = 50

Example confusion matrix for a binary classifier

- There are two possible predicted classes: "yes" - have the disease and "no" - don't have the disease.
- The classifier made a total of 165 predictions
 - 165 patients were being tested for the presence of that disease.
- Out of those 165 cases, the classifier predicted "yes" 110 times, and "no" 55 times.
- In reality, 105 patients in the sample have the disease, and 60 patients do not.
- **true positives (TP)**: These are cases in which we predicted yes (they have the disease), and they do have the disease.
- **true negatives (TN)**: We predicted no, and they don't have the disease.
- **false positives (FP)**: We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")
- **false negatives (FN)**: We predicted no, but they actually do have the disease. (Also known as a "Type II error.")

		Klasa određena modelom	
		DA	NE
Stvarna klasa	DA	100	5
	NE	10	50



Tačnost klasifikacije - Accuracy

- Tačnost klasifikacije predstavlja udeo tačno klasifikovanih instanci u ukupnom broju instanci.
- U slučaju binarne klasifikacije:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Za prethodni primer: $(TP+TN)/total = (100+50)/165 = 0.91$

Klasa određena modelom		
	DA	NE
DA	100	5
NE	10	50

Konfuziona matrica

		Klasa određena modelom	
		DA	NE
Stvarna klasa	DA	TP = 40	FN = 20
	NE	FP = 15	TN = 35

Accuracy

$$\frac{TP + TN}{TP + FP + TN + FN}$$

Preciznost (Precision)

Od ukupnog broja onih koji su klasifikovani kao pozitivni, koliko njih je zapravo pozitivno?

$$Precision = \frac{TP}{TP + FP}$$

- Za prethodni primer: $100/110 = 0.91$

Klasa određena modelom		
	DA	NE
DA	100	5
NE	10	50

- Precision je dobra mera u slučajevima kada je visoka cena lažno pozitivnih klasifikacija
- In email spam detection, a false positive means that an email that is non-spam (actual negative) has been identified as spam (predicted spam). The email user might lose important emails if the precision is not high for the spam detection model.

Konfuziona matrica

		Klasa određena modelom	
		DA	NE
Stvarna klasa	DA	TP = 40	FN = 20
	NE	FP = 10	TN = 40
		Precision $\frac{TP}{TP + FP}$	

Odziv – Recall (Sensitivity, True positive rate)

- Udeo pronađenih pozitivnih instanci u svim pozitivnim instancama

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

- Za prethodni primer: $100/105 = 0.95$

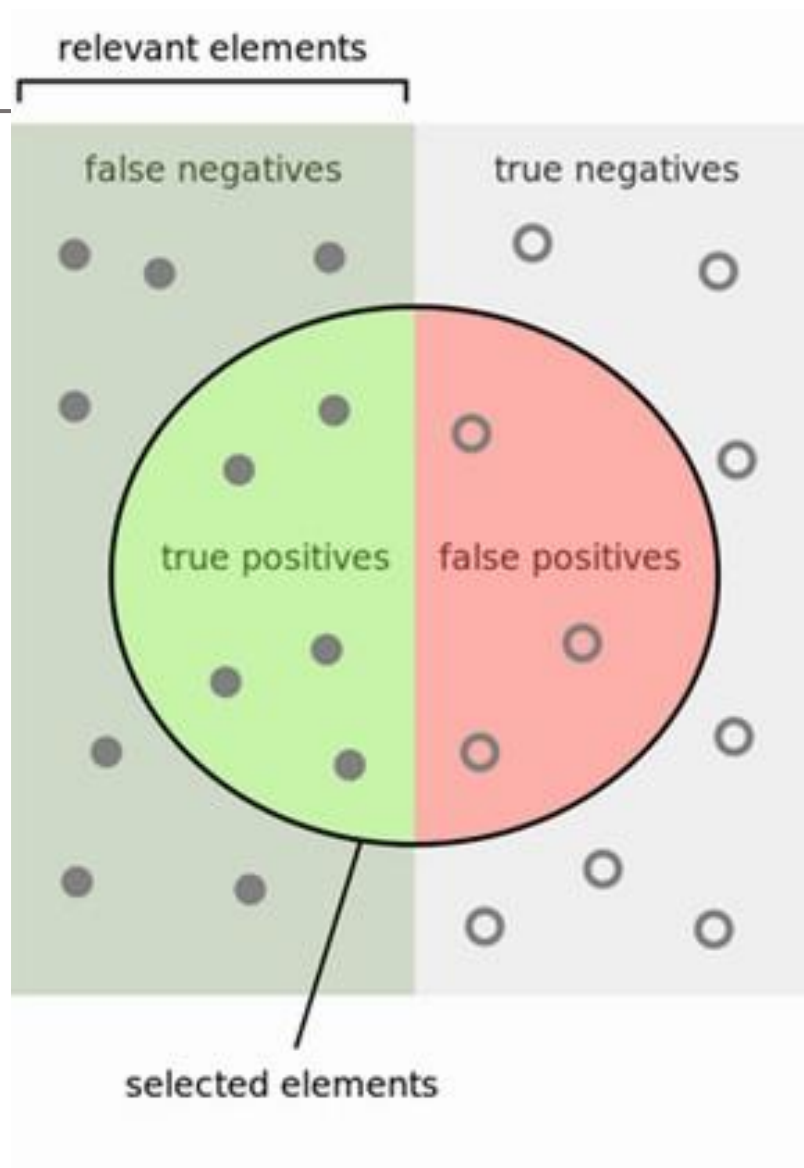
Klasa određena modelom		
	DA	NE
DA	100	5
NE	10	50

- If a fraudulent transaction (Actual Positive) is predicted as non-fraudulent (Predicted Negative), the consequence can be very bad for the bank.
- If a sick patient (Actual Positive) goes through the test and predicted as not sick (Predicted Negative). The cost associated with False Negative will be extremely high if the sickness is contagious.

Konfuziona matrica

		Klasa određena modelom		
		DA	NE	
Stvarna klasa	DA	TP = 40	FN = 20	Sensitivity $\frac{TP}{TP + FN}$
	NE	FP = 10	TN = 40	

Vizuelizacija Preciznosti i Odziva



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

F1 score

- In most problems, you could either give a higher priority to maximizing precision, or recall, depending upon the problem you are trying to solve.
- A simpler metric which takes into account both precision and recall, and therefore, you can aim to maximize this number to make your model better.
- This metric is known as F1 score, which is simply the harmonic mean of precision and recall.

$$F_1 = 2 \frac{\textit{Precision} \cdot \textit{Recall}}{\textit{Precision} + \textit{Recall}}$$

True Negative Rate - Specificity

- **True Negative Rate:** When it's actually no, how often does it predict no?

$$\text{Specificity} = \frac{TN}{TN + FP}$$

- $TN/\text{actual no} = 50/60 = 0.83$
- equivalent to 1 minus False Positive Rate
- also known as "Specificity"

Klasa određena modelom		
	DA	NE
DA	100	5
NE	10	50

Konfuziona matrica

		Klasa određena modelom	
		DA	NE
Stvarna klasa	DA	TP = 40	FN = 20
	NE	FP = 10	TN = 40
		Specificity $\frac{TN}{TN + FP}$	

False Positive Rate

- **False Positive Rate:** When it's actually no, how often does it predict yes?

$$\text{False Positive Rate} = \frac{FP}{TN + FP}$$

- FP/actual no = 10/60 = 0.17

Klasa određena modelom		
	DA	NE
DA	100	5
NE	10	50

Konfuzionna matrica

		Klasa određena modelom	
		DA	NE
Stvarna klasa	DA	TP = 40	FN = 20
	NE	FP = 10	TN = 40
		Negative Predictive Value TN $\frac{TN}{TN + FN}$	

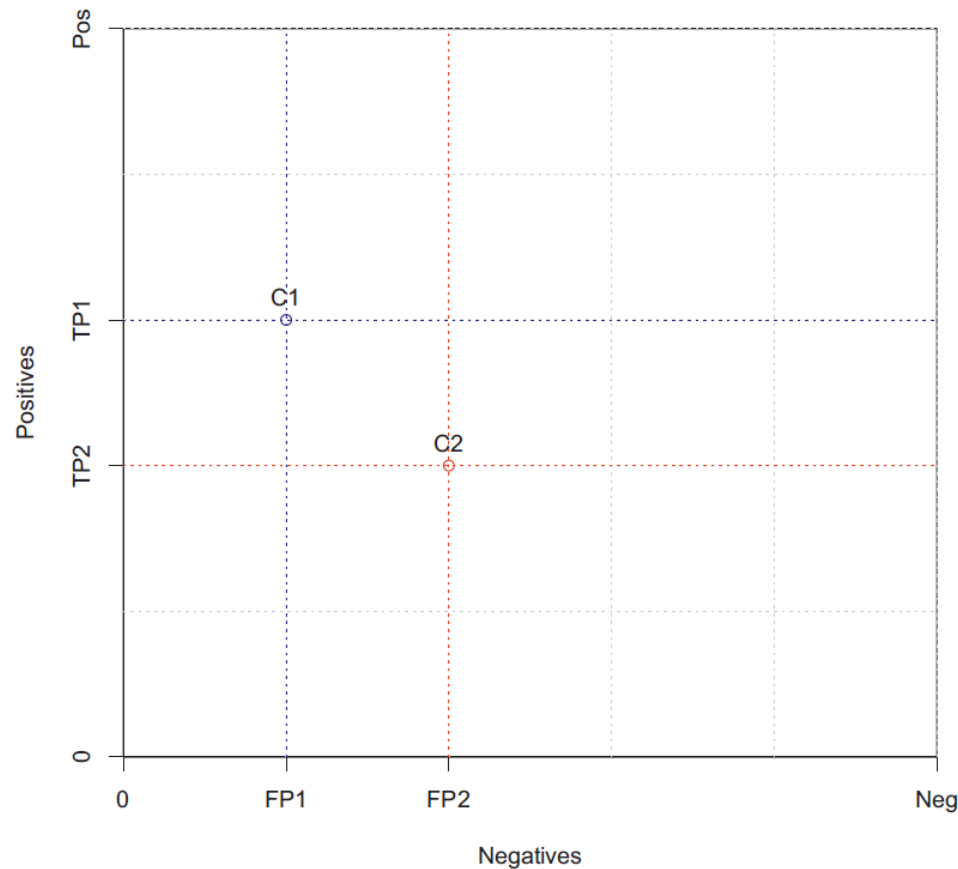
Konfuzionna matrica

		Klasa određena modelom		
		DA	NE	
Stvarna klasa	DA	TP = 40	FN = 20	Sensitivity $\frac{TP}{TP + FN}$
	NE	FP = 15	TN = 35	Specificity $\frac{TN}{TN + FP}$
		Precision $\frac{TP}{TP + FP}$	Negative Predictive Value $\frac{TN}{TN + FN}$	Accuracy $\frac{TP + TN}{TP + FP + TN + FN}$

Konfuzionna matrica

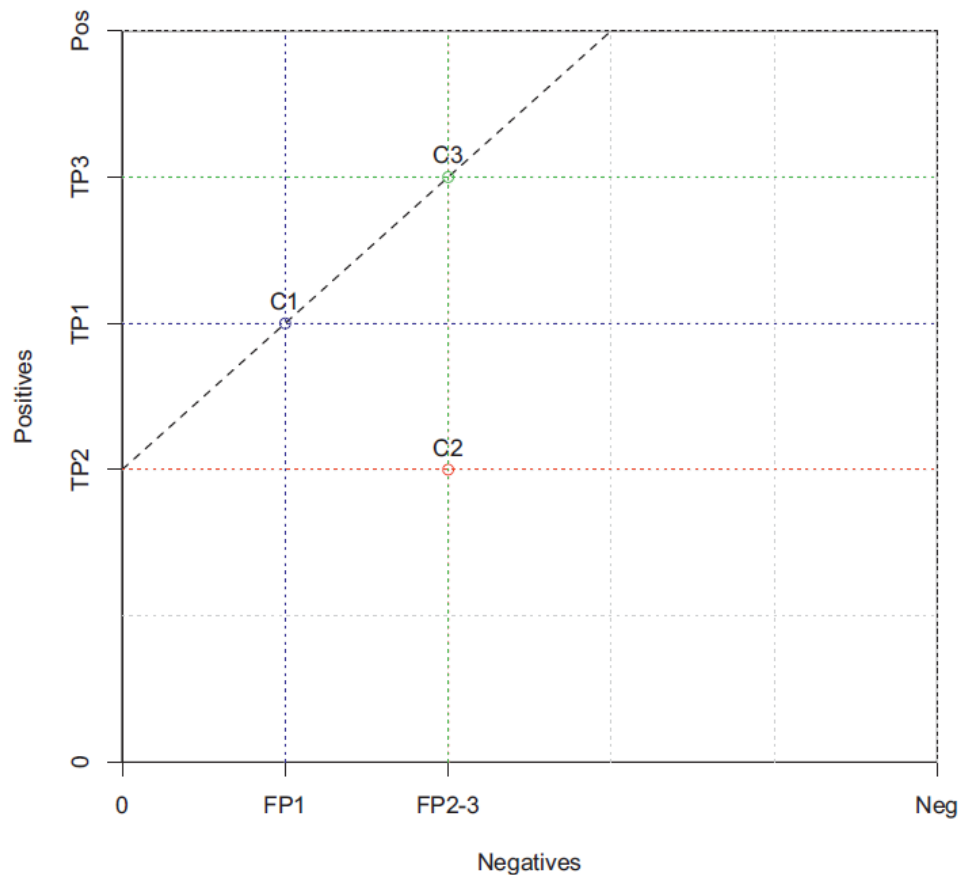
		Klasa određena modelom			
		DA	NE		
Stvarna klasa	DA	TP = 40	FN = 20	Sensitivity $\frac{TP}{TP + FN}$	
	NE	FP = 15	TN = 35	Specificity $\frac{TN}{TN + FP}$	False Positive Rate $\frac{FP}{TN + FP} = 1 - \text{Specificity}$
		Precision $\frac{TP}{TP + FP}$	Negative Predictive Value $\frac{TN}{TN + FN}$	Accuracy $\frac{TP + TN}{TP + FP + TN + FN}$	

Coverage plot



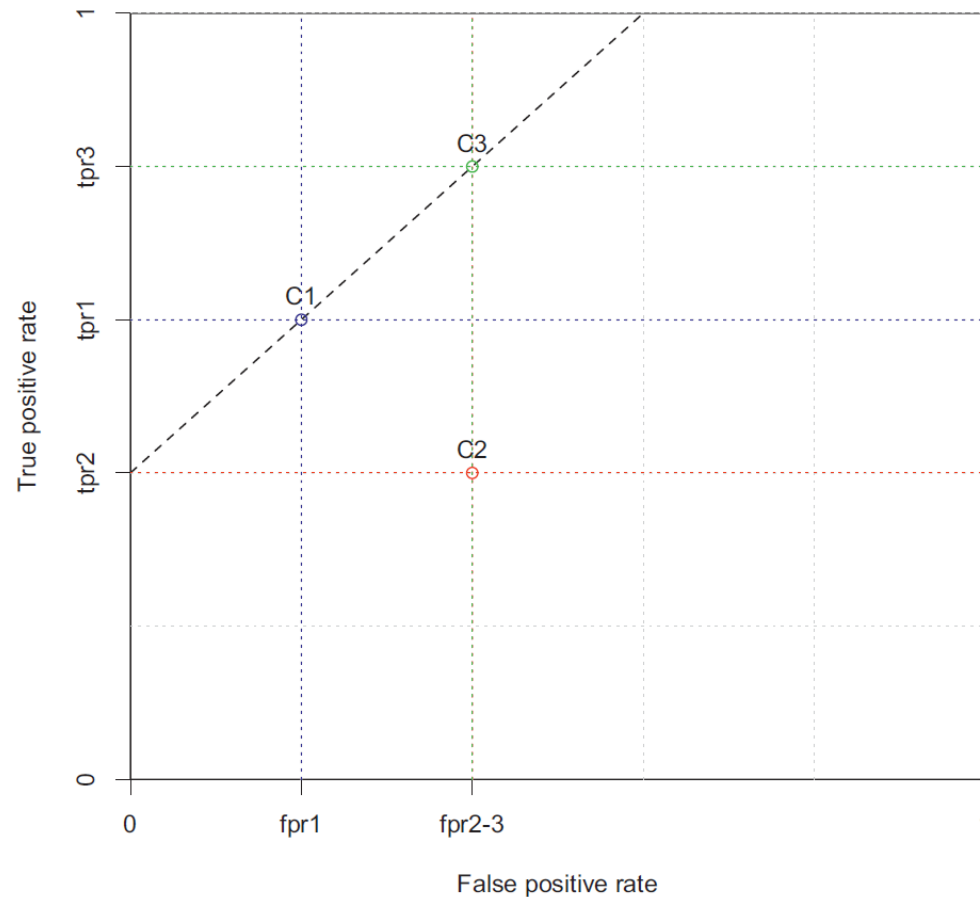
- Four numbers completely determine the confusion matrix: the number of positives *Pos*, the number of negatives *Neg*, the number of true positives *TP* and the number of false positives *FP*.
- A coverage plot visualizes these four numbers.
- Klasifikator C_1 je bolji od klasifikatora C_2
- Klasifikator C_1 dominira klasifikator C_2

Coverage plot



- Klasifikatori C_1 i C_3 oba dominiraju klasifikator C_2 .
- Klasifikatori C_1 i C_3 se međusobno ne dominiraju
- Kad god dobijemo TP, gubimo i TN. To ne utiče na zbir TP + TN, i stoga je tačnost (accuracy) ista gde god da se nalazimo.
- C_1 i C_3 imaju istu tačnost.

Normalised coverage plot



Isti crtež ali:

- Vrednosti na x osi podeljene sa ukupnim brojem negativnih:

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

- Vrednosti na y osi podeljene sa ukupnim brojem pozitivnih:

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

- C_1 i C_3 imaju istu prosečnu tačnost.

ROC grafikoni

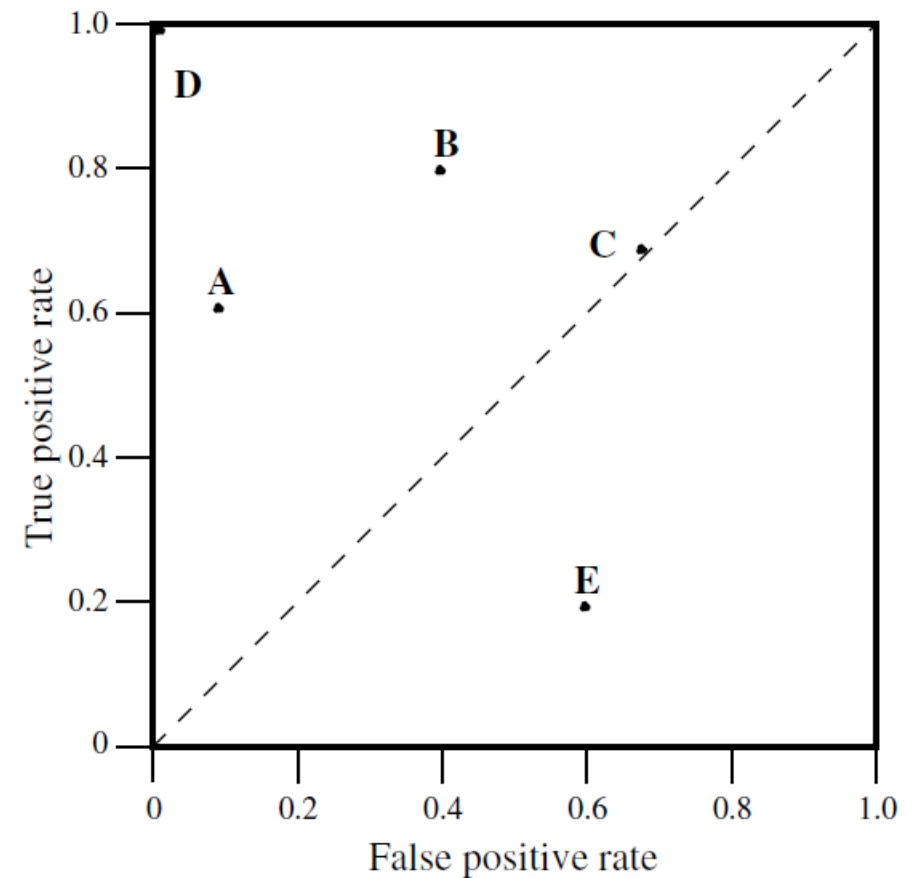
- Normalizovani coverage plot se naziva ROC grafikon.
- Grafikon karakteristika prijemnika (Receiver Operating Characteristics ROC) je tehnika za vizuelizaciju, organizovanje i izbor klasifikatora na osnovu njihovih performansi.
- ROC grafikoni se dugo koriste u teoriji detekcije signala za prikaz odnosa između pogodaka i promašaja klasifikatora signala.
- Široka upotreba u medicini za analizu rezultata statističkih obrada podataka.

ROC grafikoni

- ROC grafikoni su dvodimenzioni grafikoni kod kojih je:
 - na x osi je prikazan False Positive Rate $\frac{FP}{TN+FP}$
 - na y osi prikazan Recall (Sensitivity, True positive rate) $\frac{TP}{TP+FN}$
- False Positive Rate answers the question: "When the actual classification is negative, how often does the classifier incorrectly predict positive?,"
- True Positive Rate: "When the actual classification is positive, how often does the classifier predict positive?"
- ROC grafikoni oslikavaju relativni odnos benefita (TP) i cene (FP) primene klasifikatora.

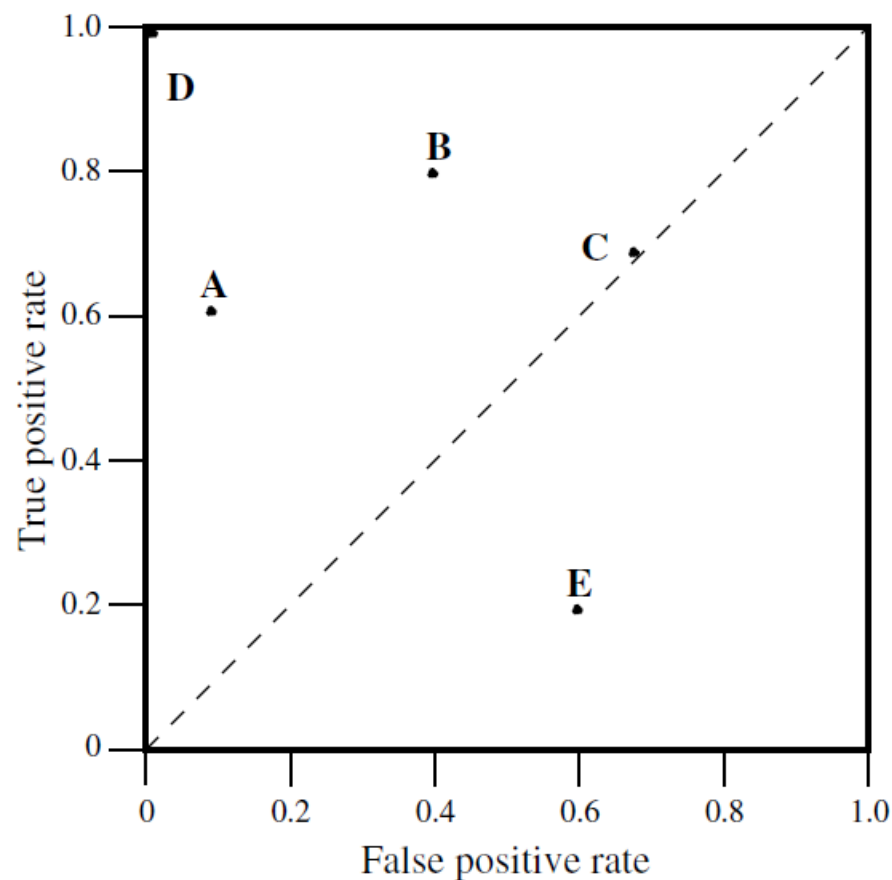
ROC grafikon

- ROC grafikon na kome je prikazano 5 diskretnih klasifikatora.
- Tačka D(0, 1) predstavlja idealni klasifikator.
- Jedna tačka je bolja od druge ako ima viši *TP rate* ili niži *FP rate* (ili oba) - nalazi se severozapadno u odnosu na drugu tačku.



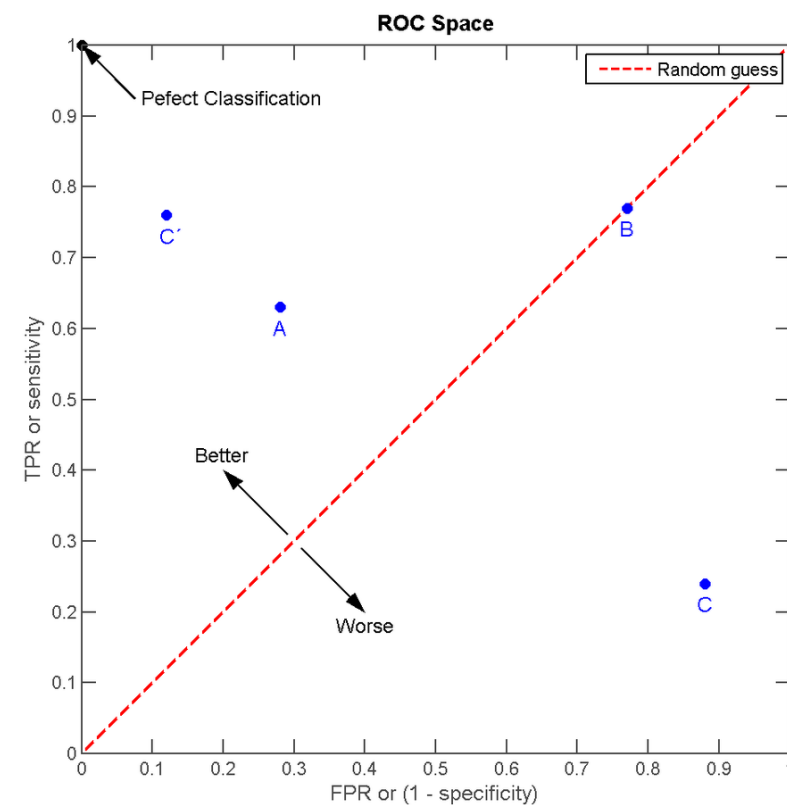
ROC grafikon

- Dijagonala $y=x$ predstavlja strategiju slučajnog pogađanja klase. Klasifikatori koji se na njoj nalaze su dobri koliko i obično bacanje novčića.
- Klasifikatori koji se nalaze u donjem trouglu imaju gore performanse i od slučajnog pogađanja. Negacija takvog klasifikatora nalaziće se u gornjem trouglu.



Primer

A			B			C			C'		
TP=63	FP=28	91	TP=77	FP=77	154	TP=24	FP=88	112	TP=76	FP=12	88
FN=37	TN=72	109	FN=23	TN=23	46	FN=76	TN=12	88	FN=24	TN=88	112
100	100	200	100	100	200	100	100	200	100	100	200
TPR = 0.63			TPR = 0.77			TPR = 0.24			TPR = 0.76		
FPR = 0.28			FPR = 0.77			FPR = 0.88			FPR = 0.12		
PPV = 0.69			PPV = 0.50			PPV = 0.21			PPV = 0.86		
F1 = 0.66			F1 = 0.61			F1 = 0.23			F1 = 0.81		
ACC = 0.68			ACC = 0.50			ACC = 0.18			ACC = 0.82		



Vrste klasifikatora

- Diskretni - Klasifikator čiji je izlaz naziv klase.
- Verovatnosni - Klasifikator čiji je izlaz broj ili rang koji govori o stepenu pripadnosti instance nekoj klasi (primer sa veći rangom pripada određenoj klasi sa većom verovatnoćom)
- U slučaju logističke regresije to je verovatnoća da primer pripada nekoj klasi.
- Verovatnosni klasifikator se može svesti na diskretni binarni klasifikator krišćenjem vrednosti praga: ako je izlaz iznad (ispod) praga onda je primer u jednoj (druvoj) klasi.

Primer crtanja ROC krive za verovatnosni klasifikator

Threshold	TP	FP	TN	FN
0.0	50	50	0	0
0.1	48	47	3	2
0.2	47	40	9	4
0.3	45	31	16	8
0.4	44	23	22	11
0.5	42	16	29	13
0.6	36	12	34	18
0.7	30	11	38	21
0.8	20	4	43	33
0.9	12	3	45	40
1.0	0	0	50	50

- Task is to diagnose 100 patients with a disease present in 50% of the general population.
- Let a model output a score between 0 and 1.
- We can alter the threshold for labeling a patient as positive (has the disease) to maximize the classifier performance.
- We will evaluate thresholds from 0.0 to 1.0 in increments of 0.1, at each step calculating the precision, recall, F1, and location on the ROC curve.

Primer - nastavak

$$\text{recall (TPR)} = \frac{TP}{TP + FN} = \frac{50}{50 + 0} = 1$$

$$\text{precision} = \frac{TP}{TP + FP} = \frac{50}{50 + 50} = 0.5$$

$$F1 \text{ score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = 0.66$$

$$\text{false positive rate} = \frac{FP}{FP + TN} = \frac{50}{50 + 0} = 1$$

Klasa određena modelom			
Prag = 0.0	DA		NE
Stvarna klasa	DA	TP = 50	FN = 0
	NE	FP = 50	TN = 0

Primer - nastavak

$$\text{recall (TPR)} = \frac{TP}{TP + FN} = \frac{42}{42 + 13} = 0.76$$

$$\text{precision} = \frac{TP}{TP + FP} = \frac{42}{42 + 16} = 0.724$$

$$F1 \text{ score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = 0.74$$

$$\text{false positive rate} = \frac{FP}{FP + TN} = \frac{16}{16 + 29} = 0.36$$

Klasa određena modelom			
Prag = 0.5	DA		NE
Stvarna klasa	DA	TP = 42	FN = 13
	NE	FP = 16	TN = 29

Primer - nastavak

$$\text{recall (TPR)} = \frac{TP}{TP + FN} = \frac{20}{20 + 33} = 0.37$$

$$\text{precision} = \frac{TP}{TP + FP} = \frac{20}{20 + 4} = 0.83$$

$$F1 \text{ score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = 0.52$$

$$\text{false positive rate} = \frac{FP}{FP + TN} = \frac{4}{4 + 43} = 0.085$$

Klasa određena modelom			
Prag = 0.8	DA		NE
Stvarna klasa	DA	TP = 20	FN = 33
	NE	FP = 4	TN = 43

Primer - nastavak

$$\text{recall (TPR)} = \frac{TP}{TP + FN} = 0$$

$$\text{precision} = \frac{TP}{TP + FP} = 0$$

$$F1 \text{ score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = 0$$

$$\text{false positive rate} = \frac{FP}{FP + TN} = \frac{4}{4 + 43} = 0$$

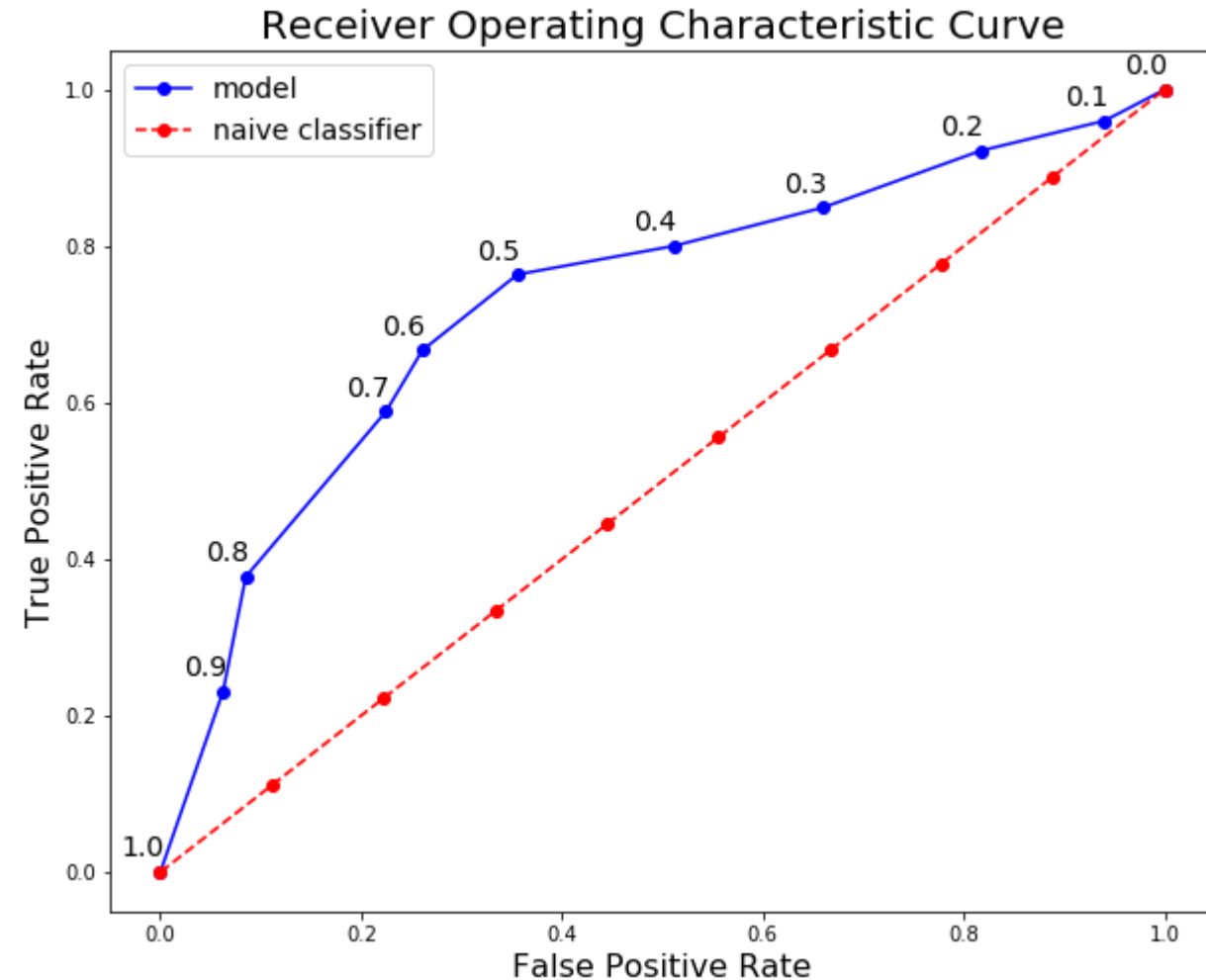
Klasa određena modelom			
Prag = 1.0		DA	NE
Stvarna klasa	DA	TP = 0	FN = 50
	NE	FP = 0	TN = 50

Primer - nastavak

threshold	recall	precision	f1	tpr	fpr
0.0	1	0.5	0.666667	1	1
0.1	0.96	0.505263	0.662069	0.96	0.94
0.2	0.921569	0.54023	0.681159	0.921569	0.816327
0.3	0.849057	0.592105	0.697674	0.849057	0.659574
0.4	0.8	0.656716	0.721311	0.8	0.511111
0.5	0.763636	0.724138	0.743363	0.763636	0.355556
0.6	0.666667	0.75	0.705882	0.666667	0.26087
0.7	0.588235	0.731707	0.652174	0.588235	0.22449
0.8	0.377358	0.833333	0.519481	0.377358	0.0851064
0.9	0.230769	0.8	0.358209	0.230769	0.0625
1.0	0	0	0	0	0

Primer - nastavak

- At a threshold of 1.0, we classify no patients as having the disease and hence have a recall and precision of 0.0.
- As the threshold decreases, the recall increases because we identify more patients that have the disease.
- As the recall increases, the precision decreases because in addition to increasing the *true* positives, we increase the *false* positives.
- At a threshold of 0.0, the recall is perfect — we find all patients with the disease — but our precision is low because we have many false positives.
- We can move along the curve for a given model by changing the threshold and select the threshold that maximizes the F1 score.



Primer crtanja ROC krive na osnovu podataka iz tabele

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

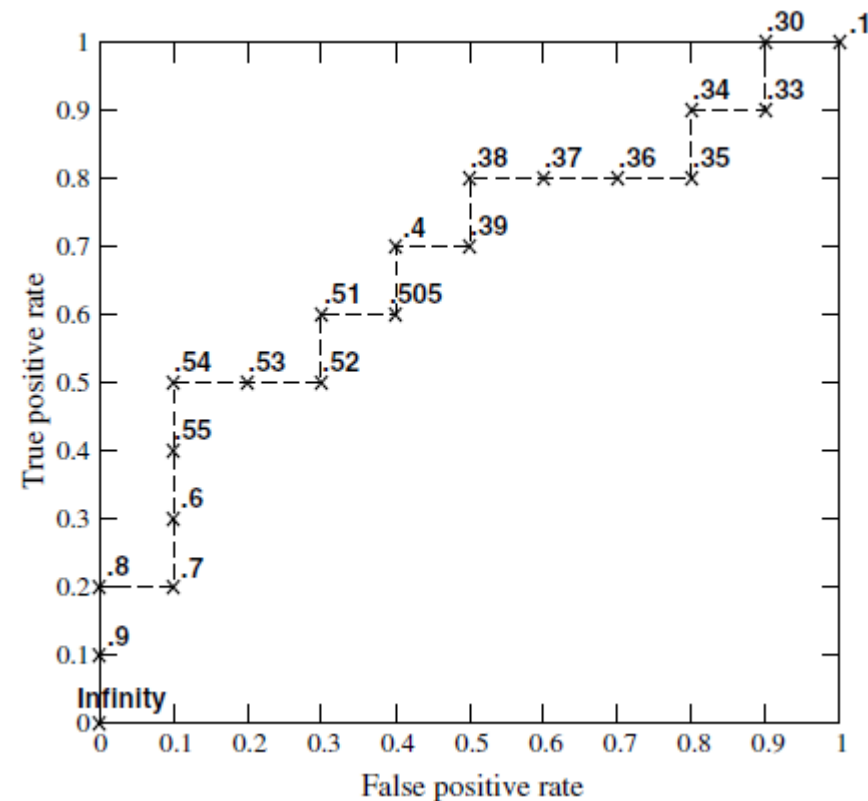
- Tabela sadrži rezultate klasifikacije 20 instanci.
- U koloni Class navedena je prava klasa (labela) instance (*p-positive, n-negative*).
- U koloni Score navedena je numerička vrednost koju je klasifikator dodelio svakoj instanci. Instance su navedene u opadajućem redosledu prema koloni Score.
- U zavisnosti od praga koji definišemo instanca će na osnovu tog skora biti klasifikovana kao pozitivna ili kao negativna.
- Na primer, ako je prag=0.6, onda će instance 1-4 biti klasifikovane kao pozitivne, a sve ostale kao negativne.

Primer crtanja ROC krive na osnovu podataka iz tabele

- ROC krivu crtamo tako što postepeno spuštamo nivo praga od 1 ka 0, a za svaki prag obeležavamo tačku na grafikonu koja odražava odnos FPR i TPR za taj prag.
- Na primer, kada je prag jednak 0.7 onda su pozitivno klasifikovani primeri 1,2 i 3, dok su svi ostali klasifikovani kao negativni. Što znači da je primer 3 false positive, a da među onim primerima koji su klasifikovani kao negativni imamo 9 true negatives i 8 false negatives.

$$FPR = \frac{FP}{FP+TN} = \frac{1}{1+9} = 0.1$$

$$TPR = \frac{TP}{TP+FN} = \frac{2}{2+8} = 0.2$$



Area under a ROC curve (AUC)

- Radi poređenja klasifikatora korišćenjem ROC krive može se koristiti jedna skalarna vrednost koja predstavlja očekivane performanse klasifikatora.
- Uobičajena metoda određivanja te skalarne vrednosti je izračunavanje površine ispod ROC krive, skraćeno AUC.
- AUC klasifikatora je ekvivalentan verovatnoći da će klasifikator rangirati slučajno izabranu pozitivnu instancu više od slučajno izabrane negativne instance.
- AUC je deo površine jediničnog kvadrata, pa ima vrednost između 0 i 1.
- Slučajno pogađanje ima $AUC \leq 0.5$

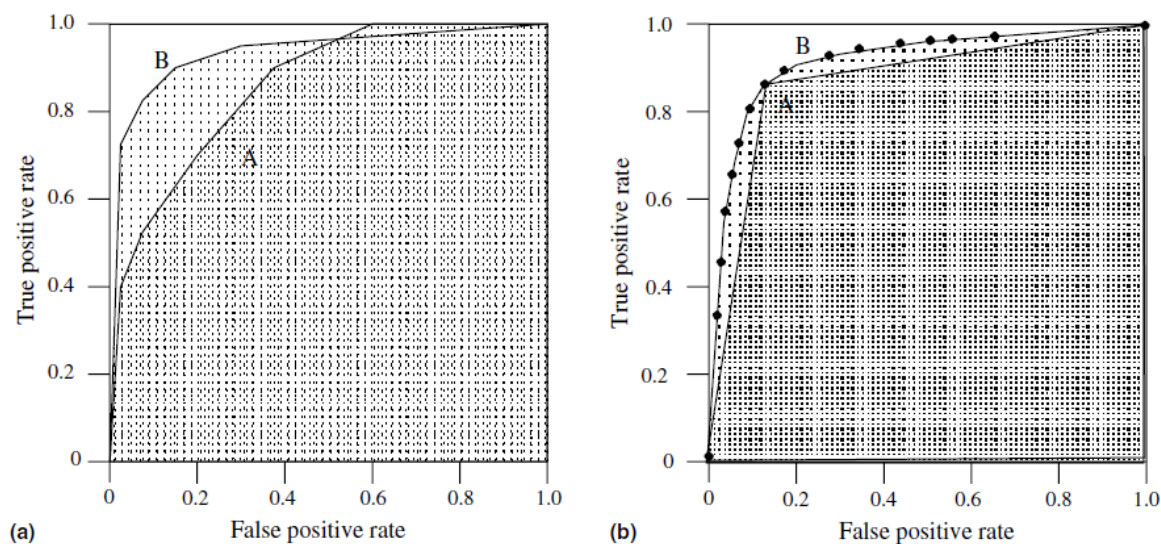
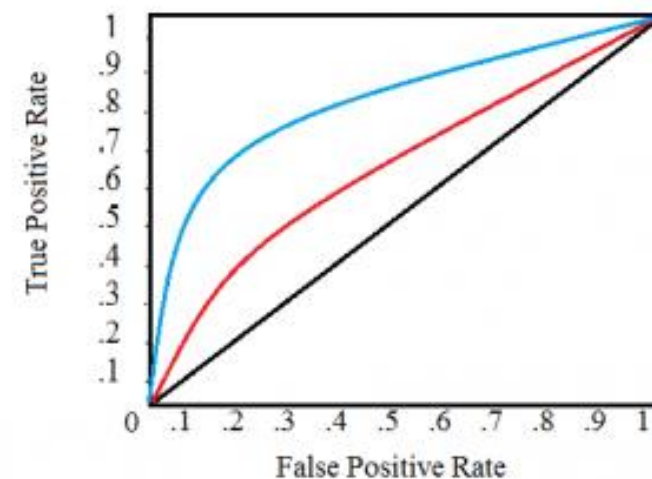


Fig. 8. Two ROC graphs. The graph on the left shows the area under two ROC curves. The graph on the right shows the area under the curves of a discrete classifier (A) and a probabilistic classifier (B).

ROC curve

- Crna dijagonala označava slučajni klasifikator (bacanje novčića), dok crvena i plava kriva odgovaraju različitim klasifikacionim modelima.
- Za jedan model možemo se kretati samo po jednoj, njemu odgovarajućoj krivoj i to prilagođavanjem praga za klasifikaciju pozitivnih primera.
- Za prag 1.0, nalazili bismo se u koordinatnom početku jer nijedan primer ne bismo klasifikovali kao pozitivan, pa bi *true positives*=0 i *false positives*=0 (TPR = FPR = 0).
- Snižavanjem praga, više tačaka biva klasifikovano pozitivno, rezultirajući sa više true positives, ali i sa više false positives (TPR i FPR rastu). Konačno kada je prag 0.0 svi podaci su klasifikovani kao pozitivni, pa su i TPR i FPR maksimalni (TPR = FPR = 1.0).
- AUC plave krive je veći od AUC-a crvene krive, pa se može zaključiti da je klasifikator predstavljen plavom krivom bolji od klasifikatora datog crvenom krivom..

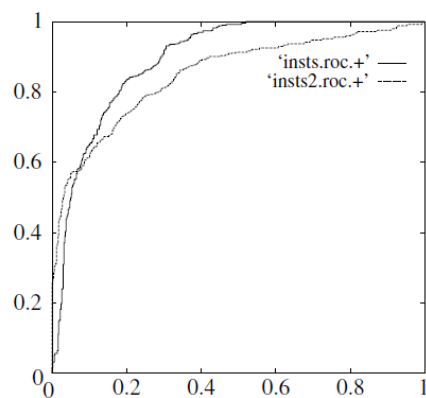


Osetljivost metrika na promene u raspodeli klasa

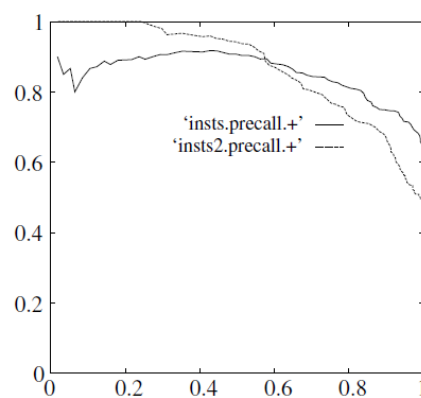
- ROC krive su neosetljive na promene u raspodeli klasa.
- Ako se odnos pozitivnih i negativnih instanci u testnom skupu promeni, ROC kriva se neće promeniti.
- Svaka mera performanse koja koristi vrednosti iz oba reda matrice će biti osetljiva na promene u odnosu klasa - class skews.
- Accuracy, precision, F1 score koriste vrednosti iz oba reda matrice .
- Kako se menja odnos broja pozitivnih i negativnih primera u testnom skupu, tako će se menjati i vrednosti ovih mera, iako će suštinski performanse klasifikatora ostati iste.

Klasa određena modelom				
		DA	NE	
Stvarna klasa	DA	Stvarno Pozitivni TP	Lažno Negativni FN	Pozitivni
	NE	Lažno Pozitivni FP	Stvarno Negativni TN	Negativni

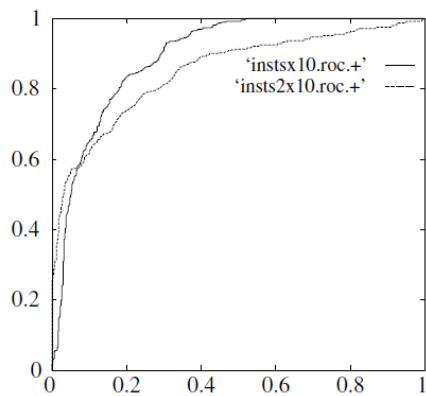
Osetljivost metrika na promene u raspodeli klasa



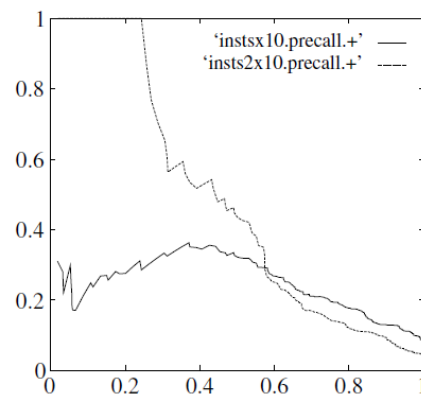
(a)



(b)



(c)



(d)

- Dva klasifikatora su evaluirana korišćenjem ROC krivih i precision-recall krivih.
- Na slikama a i b, testni skup ima balansirani odnos pozitivnih i negativnih primera (1:1).
- Krive c i d prikazuju ROC i precision-recall krive za iste klasifikatore na istom domenu, ali u slučaju kada je broj negativnih instanci povećan.
- Sami klasifikatori se nisu promenili, samo je distribucija klasa različita.
- ROC krive na slikama a i c su identične, dok se precision-recall krive na slikama b i d značajno razlikuju.