

Statistička analiza klasifikatora i dijagnostičkih alata

Izv. prof. dr. sc. Ivan Dražić

Sveučilište u Rijeci, Tehnički fakultet

Predavanje za nastavnike matematike srednjih škola

PROBLEMATIKA ANALIZE DIJAGNOSTIČKOG TESTA

"SARS-CoV-2 antigen test brza je imunokromatografska metoda namijenjena za kvalitativnu detekciju specifičnog N antigena SARS-CoV-2.

Test ima osjetljivost 96,52% i specifičnost 99,68%, a navedene vrijednosti su dobivene analizom 426 uzoraka koji su testirani u dvije nezavisne institucije."

1. Što je to osjetljivost, a što specifičnost?
2. Kako se računaju osjetljivost i specifičnost?

PROBLEMATIKA KLASIFIKATORA U UMJETNOJ INTELIGENCIJI (AI)

U strojnom učenju, klasifikacija se odnosi na problem prediktivnog modeliranja gdje se oznaka klase dodjeljuje svakom ulaznom podatku.

Primjeri:

1. Je li poruka elektroničke pošte *spam*?
2. Pridružiti skeniranom slovu stvarno slovo.

Iz perspektive modeliranja, klasifikacija zahtijeva skup podataka za obuku s mnogo primjera ulaza i izlaza iz kojih se može učiti.

Algoritmi klasifikacije mogu se shvatiti kao dijagnostički testovi te se analizirati na isti način.

ZAŠTO BI SE TREBALI BAVITI OVAKVIM STVARIMA U NASTAVI?

Paradigma Education 4.0 - pandan četvrte industrijske revolucije

Nevjerojatna brzina razvoja znanosti i tehnologije zahtjeva sve obrazovanije i obrazovanije kadrove, koje od ranih dana treba upoznavati s mogućnostima suvremene znanosti.

Funkcioniranje društva sve je manje determinističko, a sve više stohastičko na što čovječanstvo nije u potpunosti spremno!

KRATKA POVIJEST TEORIJE VJEROJATNOSTI

- ▶ 1654. de Mereov problem, korespodencija Pascal-Fermat
- ▶ 1764. Thomas Bayes - definicija uvjetne vjerojatnosti (objavljeno posthumno)
- ▶ 1812. Pierre de Laplace: "Analitička teorija vjerojatnosti", primjena vjerojatnosti izvan kockanja
- ▶ 1933. Andrej Nikolajevič Kolmogorov: "Temelji teorije vjerojatnosti", aksiomatika teorije vjerojatnosti
- ▶ početak 21. stoljeća - široka primjena računala i mogućnost simulacija

UVJETNA VJEROJATNOST - MOTIVACIJA

Svi zaposlenici jednog poduzeća morali su pristupiti ispitu za obnovu certifikata. Poduzeće zapošljava 40 muškaraca i 60 žena, a certifikat je uspješno obnoviti 60 radnika. Rezultati ispita prikazani su u sljedećoj tablici:

	Muškarci	Žene	Ukupno
Prošli	24	36	60
Pali	16	24	40
Ukupno	40	60	100

Slučajnim odabirom bira se jedan radnik.

Odredite vjerojatnost da će izabrani radnik proći, **ako je poznato da je muškog spola**.

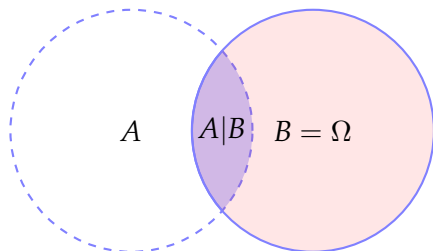
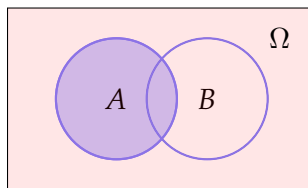
$$P(\text{"Radnik će proći na ispitu"} | \text{"Radnik je muškog spola"}) = ?$$

UVJETNA VJEROJATNOST

Neka je (Ω, \mathcal{F}, P) prostor vjerojatnosti i neka je $B \in \mathcal{F}$ događaj za koji je $P(B) > 0$. **Uvjetnu vjerojatnost** $P(\cdot | B)$ definiramo pravilom

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

za svaki $A \in \mathcal{F}$.



UVJETNA VJEROJATNOST - PRIMJER

	Muškarci	Žene	Ukupno
Prošli	24	36	60
Pali	16	24	40
Ukupno	40	60	100

$$\begin{aligned} & P(\text{"Radnik će proći na ispitu"} | \text{"Radnik je muškog spola"}) \\ = & \frac{P(\text{"Radnik će proći na ispitu"} \text{ i } \text{"Radnik je muškog spola"})}{P(\text{"Radnik je muškog spola"})} \\ & = \frac{24}{40} = \frac{24}{40} = 60\% \end{aligned}$$

ANALIZA KLASIFIKATORA / DIJAGNOSTIČKOG TESTA

Problemi:

1. Kolika je vjerojatnost da će test "uloviti" sve bolesne?
2. Kolika je vjerojatnost da će test zdravu osobu okarakterizirati kao bolesnu?
3. ...

A - POZITIVAN TEST, A^C - NEGATIVAN TEST

B - PACIJENT JE BOLESTAN, B^C - PACIJENT JE ZDRAV

Rješenje:

1. $P(A|B)$
2. $P(A|B^C)$
3. ...

MATRICA KONFUZIJE

		PREDIKTIRANO	
		POZITIVNO	NEGATIVNO
IZMJERENO	POZITIVNO	TP	FN
	NEGATIVNO	FP	TN

TP - stvarno pozitivni

TN - stvarno negativni

FP - lažno pozitivni

FN - lažno negativni

PREDIKTIRANO = NOVA DIJAGNOSTIČKA METODA

IZMJERENO = STARA DIJAGNOSTIČKA METODA (zlatni standard)

MATRICA KONFUZIJE - PRIMJER

Nova metoda dijagnostike inzulinske rezistencije uspoređivana je sa zlatnim standardom na uzorku od 100 pacijenata.

U uzorku je jednak broj pacijenata kod kojih je zlatnim standardom utvrđena inzulinska rezistencija i onih kod kojih je ta dijagnoza isključena.

Kod 40 pacijenata inzulinska rezistencija potvrđena je na oba testa, a kod njih 45 isključena je na oba testa.

Formirati matricu konfuzije.

		NOVA METODA	
		POZITIVNO	NEGATIVNO
STARA METODA	POZITIVNO	40	10
	NEGATIVNO	5	45

STOPA POGREŠKE I TOČNOST

Stopa pogreške (error rate (ERR)) - vjerojatnost netočnog predviđanja:

$$ERR = P(A^C|B) + P(A|B^C) = \frac{FN + FP}{TP + FN + FP + TN} = \frac{15}{100} = 15\%$$

0% ← dobar; loš → 100%

Točnost (accuracy (ACC)) - vjerojatnost točnog predviđanja:

$$ACC = P(A|B) + P(A^C|B^C) = \frac{TP + TN}{TP + FN + FP + TN} = \frac{85}{100} = 85\%$$

0% ← loš; dobar → 100%

OSJETLJIVOST I SPECIFIČNOST

Osjetljivost (sensitivity (SN)) - vjerojatnost točnog predviđanja kod bolesnih ("test lovi bolesne"):

$$SN = P(A|B) = \frac{TP}{TP + FN} = \frac{40}{40 + 10} = 80\%$$

0% ← loš; dobar → 100%

Specifičnost (specificity (SP)) - vjerojatnost točnog predviđanja kod zdravih ("test ne gomila bolesne"):

$$SP = P(A^C|B^C) = \frac{TN}{FP + TN} = \frac{45}{5 + 45} = 90\%$$

0% ← loš; dobar → 100%

POZITIVNA I NEGATIVNA PREDIKTIVNA VRIJEDNOST

Pozitivna prediktivna vrijednost (positive predictive value (PPV)) - vjerojatnost točnog predviđanja pozitivnog testa ("koliko se treba brinuti ako je test pozitivan?"):

$$PPV = P(B|A) = \frac{TP}{TP + FP} = \frac{40}{40 + 5} = 89\%$$

0% ← loš; dobar → 100%

Negativna prediktivna vrijednost (negative predictive value (NPV)) - vjerojatnost točnog predviđanja negativnog testa ("koliko se možemo opustiti ako je test negativan?"):

$$NPV = P(B^C|A^C) = \frac{TN}{FN + TN} = \frac{45}{10 + 45} = 82\%$$

0% ← loš; dobar → 100%

INTERPRETACIJA

Osjetljivost i specifičnost - orijentiranost na zdravstveni sustav

Želimo "uloviti" sve bolesne ljude (visoka osjetljivost), a da pritom ne bude previše lažno pozitivnih (visoka specifičnost).

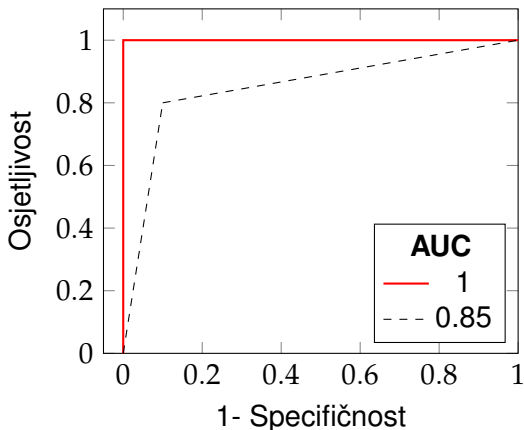
Česta praksa: Radimo dva testa, najprije visoko osjetljiv, a zatim visoko specifičan.

Pozitivna i negativna prediktivna vrijednost - orijentiranost na pacijenta (informiranje pacijenta).

Prevalencija (postotak bolesnih u populaciji) ne utječe na osjetljivost i specifičnost, ali utječe na pozitivnu i negativnu prediktivnu vrijednost.

Kod manje prevalencije je manja PPV, ali veća NPV.

ROC KRIVULJA



AUC - površina ispod ROC krivulje (area under curve) > 0.8

ANALIZA AI KLASIFIKATORA KOJI PREDIKTIRA PROLAZ STUDENATA

		PREDIKCIJA PUTEM AI	
		PROLAZ	PAD
STVARNO STANJE	PROLAZ	22	1
	PAD	2	13

$ERR = 8\%$

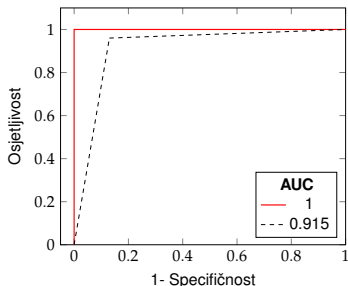
$ACC = 92\%$

$SN = 96\%$

$SP = 87\%$

$PPV = 92\%$

$NPV = 93\%$



MOGUĆE PRIMJENE U SREDNJOŠKOLSKOJ NASTAVI

Projektni zadaci sa stvarnim podacima:

1. Analiza kvalitete vremenske prognoze.
2. Određuje li ocjena prvog razreda ocjenu drugog razreda?
3. Analiza algoritama u suradnji s informatikom.
4. Analiza kvalitete različitih kemijskih i bioloških testova (međupredmetna suradnja).

Sadržaji unutar matematike:

1. Izračun površine kod određivanja AUC vrijednosti.

