Važnije mjere statistike, teorije informacija i dubinske analize podataka

Matej Mihelčić

Prirodoslovno-matematički fakultet, Sveučilište u Zagrebu

matmih@math.hr

19. veljače, 2026.

Statističke mjere

Osoba	Visina	Težina	Broj cipele
Marko	178	82	42
Petar	192	88	46
Marija	165	63	38
Ivana	170	69	40

Statističke mjere

- **Aritmetička sredina**: za ulazni vektor $\vec{v} = (v_1, \ldots, v_n)$, se definira kao $\bar{v} = \frac{\sum_{i=1}^{n} v_i}{n}$. Centralna tendencija (snažno utjecana od strane ekstremnih vrijednosti). Npr. aritmetička sredina atributa *Visina* je 176.25.
- Ponekad se koriste:
  - **Geometrijska sredina**: $\bar{v} = (\prod_{i=1}^{n} x_i)^{\frac{1}{n}}$.
  - **Harmonijska sredina**: $\bar{v} = n \cdot (\sum_{i=1}^{n} \frac{1}{x_i})^{-1}$.
- **Minimalna/maksimalna** vrijednost. Daju nam informaciju o rasponu vrijednosti u skupu podataka ili atributu. Npr. minimalna vrijednost atributa *Visina* je 165 a maksimalna 192.
- **Median** - centralna vrijednost. Definiran kao $med(\vec{v}) = v_{(n+1)/2}$ za neparni n, $med(\vec{v}) = \frac{v_{(n)/2} + v_{(n+1)/2}}{2}$ za parni. U znatno manjoj mjeri na njega utječu ekstremne vrijednosti. Median atributa *Visina* je 174.

Statističke mjere

- **Raspon** - razlika maksimalne i minimalne vrijednosti u skupu podataka. Npr. raspon atributa *Visina* je 27.
- **Mod** - najčešća vrijednost u skupu podataka. U našem primjeru nema moda pošto su sve vrijednosti različite. Generalno možemo imati više od jednog moda u podacima.
- **Kvartili** (Q_1, Q_2, Q_3) - dijele skup podataka na 4 dijela. Q_1 je vrijednost iz skupa podataka takva da je 25% elemenata skupa manje ili jednako Q_1. Q_2 (median skupa) je vrijednost takva da je 50% elemenata skupa manje ili jednako Q_2. Q_3 je vrijednost iz skupa podataka takva da je 75% elemenata skupa manje ili jednako Q_3. Za atribut *Visina*, $Q_1 = \frac{165+170}{2} = 167.5$, $Q_2 = 174$, $Q_3 = \frac{178+192}{2} = 185$.
- **Interkvartilni raspon** - mjera disperzije. Računa se kao $IQR = Q_3 - Q_1$. IQR za atribut *Visina* je 12.75.

Statističke mjere

- **Očekivanje** - $\mu(\vec{v}) = \sum_{i=1}^{n} v_i \cdot P(v_i)$, gdje $P(v_i)$ predstavlja vjerojatnost pojavljivanja vrijednosti v_i u skupu. Pošto su sve vrijednosti atributa *Visina* različite, $P(v_i) = \frac{1}{4}$, $\forall i$, stoga je očekivanje u ovom slučaju jednako aritmetičkoj sredini, $\mu = 176.25$.
- **Varijanca** (uzoračka) - očekivano kvadratno odstupanje vrijednosti od očekivanja. $Var(\vec{v}) = \frac{1}{n-1} \sum_{i=1}^{n} (v_i - \mu(v_i))^2$. Uzoračka varijanca vrijednosti atributa *Visina* je 138.9167.
- **Standardna devijacija** - $\sigma(\vec{v}) = \sqrt{Var(\vec{v})} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (v_i - \mu(v_i))^2}$. Standardna devijacija vrijednosti atributa *Visina* je 11.786.

Statističke mjere

Koeficijenti korelacije

- **Pearsonov koeficijent korelacije** - mjeri linearne odnose između dva skupa realnih vrijednosti (npr. dva atributa, dva mjerenja itd.). Definiran je kao $\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \cdot \sigma_Y}$. Ovaj koeficijent korelacije je osjetljiv na **outliere**. Daje najtočnije procjene korelacije za normalno distribuirane podatke (iako to nije preduvjet za njegovu upotrebu). Pearsonov koeficijent korelacije između atributa *Visina* i *Težina* je 0.957.
- **Spearmanov koeficijent korelacije** - mjeri monotone odnose između dva skupa kategorijskih ili realnih vrijednosti. Računa se kao $r_{X,Y} = \rho_{R[X],R[Y]} = \frac{cov(R[X],R[Y])}{\sigma_{R[X]} \cdot \sigma_{R[Y]}}$. R označava rank podataka iz određenog skupa (slučajne varijable). Nije osjetljiv na outliere jer koristi rank, odnosi mogu biti i ne linearni, mjeri smjer odnosa prikazan preko rankova. Spearmanov koeficijent korelacije između atributa *Visina* i *Težina* je 1.0.

Statističke mjere

- **Kendall tau** koeficijent korelacije - baziran na rankovima (poretku) vrijednosti elemenata. Mjeri monotone odnose kao Spearman. Vrijednosti dva skupa prikazujemo kao parove oblika $(x_1, y_1), \ldots, (x_n, y_n)$. Pretpostavka je da nema jednakih vrijednosti, inače se postupak malo modificira. Za dva elementa (x_i, y_i) i (x_j, y_j) kažemo da se slažu ako $x_i > x_j$ i $y_i > y_j$ ili $x_i < x_j$ i $y_i < y_j$. $\tau_{X,Y} = \frac{S - N}{\binom{n}{2}}$, gdje S označava broj parova koji se slažu, a N broj parova koji se ne slažu. Robusniji test od Spearman-a i jednostavniji za interpretirati. Kendall tau između atributa *Visina* i *Težina* je 1.0.
- **Gudmanov i Kruskalov gamma** - također baziran na rankovima i mjeri monotone odnose. Računa se kao $G = \frac{S - N}{S + N}$. Potencijalno preferirana mjera na skupovima podataka s puno ponavljanja (identičnih vrijednosti). Goodman and Kruskal's gamma između atributa *Visina* i *Težina* je 1.0.

Statističke mjere

Statistički testovi - provjeravaju statističku značajnost hipoteze (ovisno o testu i zadanim vrijednostima). Mogu ispitivati:

- statističku značajnost odnosa aritmetičkih sredina dva uzroka (npr. t-test, Welch t-test, Z-test)
- statističku značajnost odnosa aritmetičkih sredina više uzoraka (npr. ANOVA)
- statističku značajnost odnosa mediana (npr. median test)
- nezavisnost, homogenost i slaganje uzorka s teorijski pretpostavljenom distribucijom (npr. Pearson-ov χ^2 test)
- usporedbu distribucija (npr. Wilcoxon signed-rank, Mann-Whitney U, Kruskal-Wallis, Kolmogorov-Smirnov test)
- usporedbu varijanci (npr. F-test)
- testirati normalnost uzoraka (npr. Lilliefors, Shapiro–Francia, Shapiro–Wilk)
- statističku značajnost odnosa u tablicama kontingencije (npr. Fisher egzaktni test, McNemar, Barnard, Boschloo test).

Mjere teorije informacija

Mjere i kvantificiraju **količinu informacija i nesigurnosti** vezanih uz moguća stanja ili ishode slučajnih varijabli.

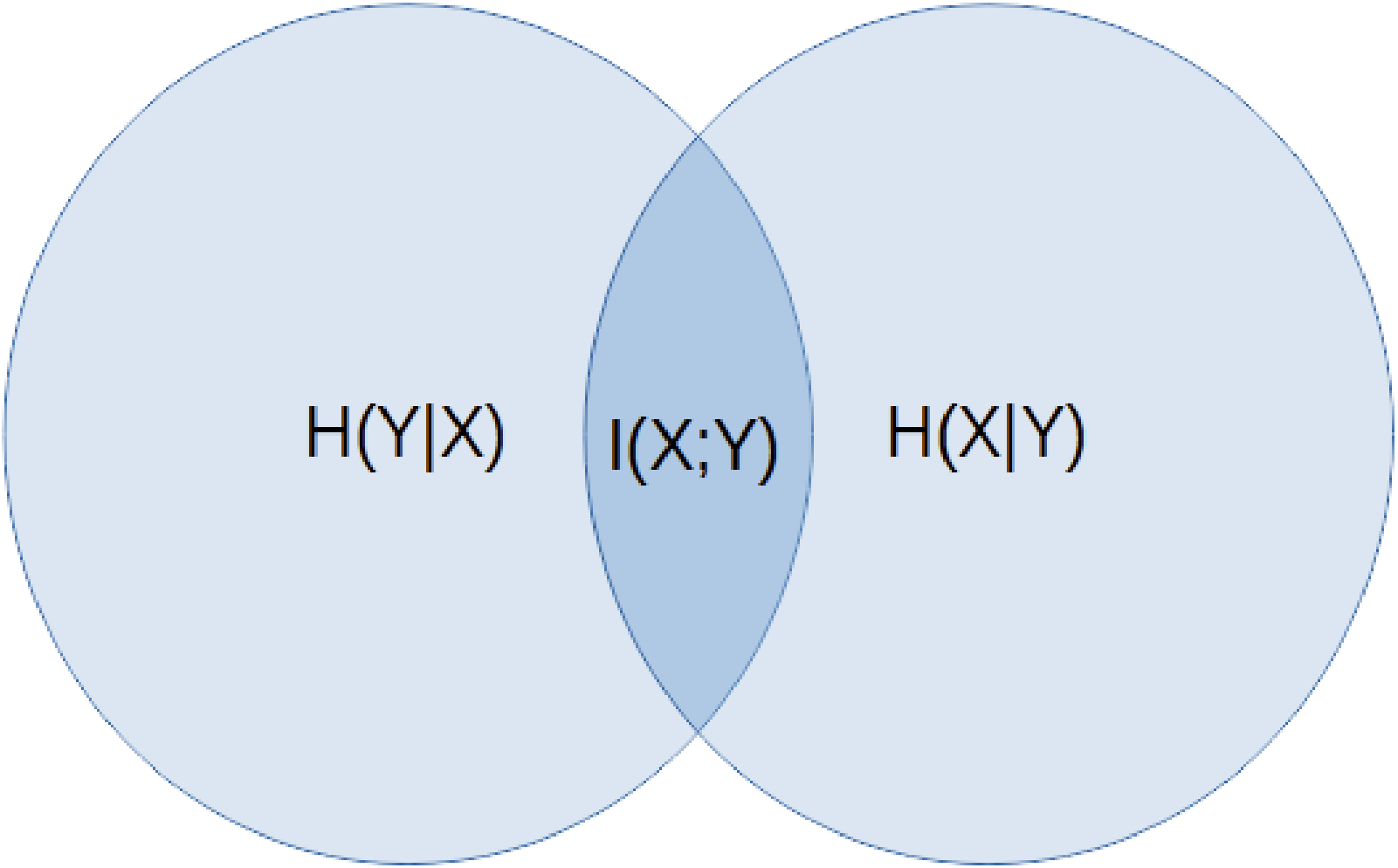
- **Entropija** - očekivana količina informacija potrebna za opisati stanja varijable. Definira se kao: $H(X) = -\sum_{x \in \mathcal{X}} p(x) \cdot \log p(x)$. Entropija atributa koji ima sve identične vrijednosti je 0. Zbog toga entropiju možemo smatrati i mjerom **uređenosti** podataka.
  Za kontinuiranu slučajnu varijablu s funkcijom gustoće vjerojatnosti $f(x)$ s konačnom ili beskonačnom potporom $\mathcal{X}$ definiramo:
  $H(X) = \mathbb{E}[-\log f(X)] = -\int_{\mathcal{X}} f(x) \log(f(x)) dx.$
- **Uvjetna entropija** - izražava količinu nesigurnosti varijable X, ukoliko znamo informacije o varijabli Y. Računa se kao:
  $H(X|Y = y_j) = -\sum_{x \in \mathcal{X}} p(x|y_j) \cdot \log p(x|y_j).$
  $H(X|Y) = \sum_{y_j \in \mathcal{Y}} p(y_j) H(X|Y = y_j).$

Mjere teorije informacija

- **Uvjetna diferencijalna entropija** (nastavak) - za kontinuirane slučajne varijable X, Y sa zajedničkom funkcijom gustoće vjerojatnosti $f(x,y)$ definiramo: $H(X|Y) = -\int_{\mathcal{X},\mathcal{Y}} f(x,y)\log f(x|y)dxdy$.
- **Uzajamna informacija** - količina informacije koju dobijemo o varijabli X poznavajući vrijednost varijable Y. Definiramo je kao: $I(X;Y) = H(X) - H(X|Y)$. U našem primjeru je količina dobivene informacije o atributu *Veličina*, ukoliko znamo vrijednosti varijable *Težina* jednaka entropiji atributa *Veličina*.

$$I(X;Y) = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\cdot\log\frac{p(x,y)}{p(x)p(y)}.$$

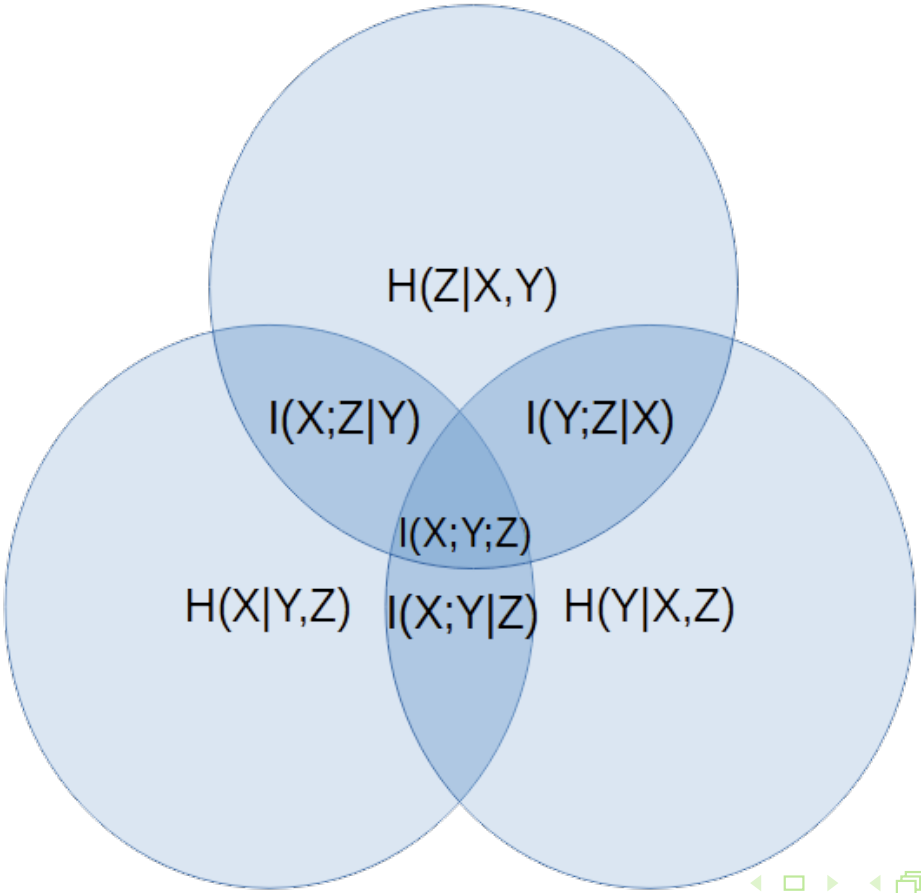
Mjere teorije informacija



Mjere teorije informacija

- **Uzajamna informacija** (nastavak) - formula se može generalizirati i na više od dvije varijable:

$$I(X;Y;Z) = \sum_{x\in\mathcal{X},y\in\mathcal{Y},z\in\mathcal{Z}} p(x,y,z)\cdot\log\frac{p(x,y)p(x,z)p(y,z)}{p(x,y,z)p(x)p(y)p(z)}.$$



Mjere teorije informacija

- **Uzajamna informacija** (nastavak) - uzajamna informacija $I(X;Y)$ dvije slučajne varijable X i Y definiramo:
$$I(X;Y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dydx.$$
- **Relativna entropija** ili **Kullback-Leibler divergencija** - daje mjeru razlike između distribucija P i Q. Definirana je kao:
$$D(P\|Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}.$$
U slučaju kontinuiranih varijabli s funkcijama gustoće vjerojatnosti f i g definiramo: $D(f\|g) = \int_{\mathcal{X}} f(x) \log \frac{f(x)}{g(x)} dx.$
- **Unakrsna entropija** - mjeri učinkovitost kodiranja događaja iz distribucije P, pretpostavljenom distribucijom Q. Definira se kao: $H(P,Q) = \sum_x -P(x) log Q(x)$. U slučaju $P = Q$ dobijemo formulu za entropiju.

Mjere teorije informacija

- **Unakrsna entropija** (nastavak) - za kontinuirane slučajne varijable s funkcijama gustoće vjerojatnosti P i Q, unakrsnu entropiju definiramo kao: $H(P, Q) = -\int_{\mathcal{X}} P(x) \log Q(x) dx$.

Mjere teorije informacija

Entitet	Boja	Okus	Voće
Jagoda	Crvena	Sladak	DA
Prokulica	Zelena	Gorak	NE
Kruška	Žuta	Sladak	DA
Radič	Zelena	Gorak	NE

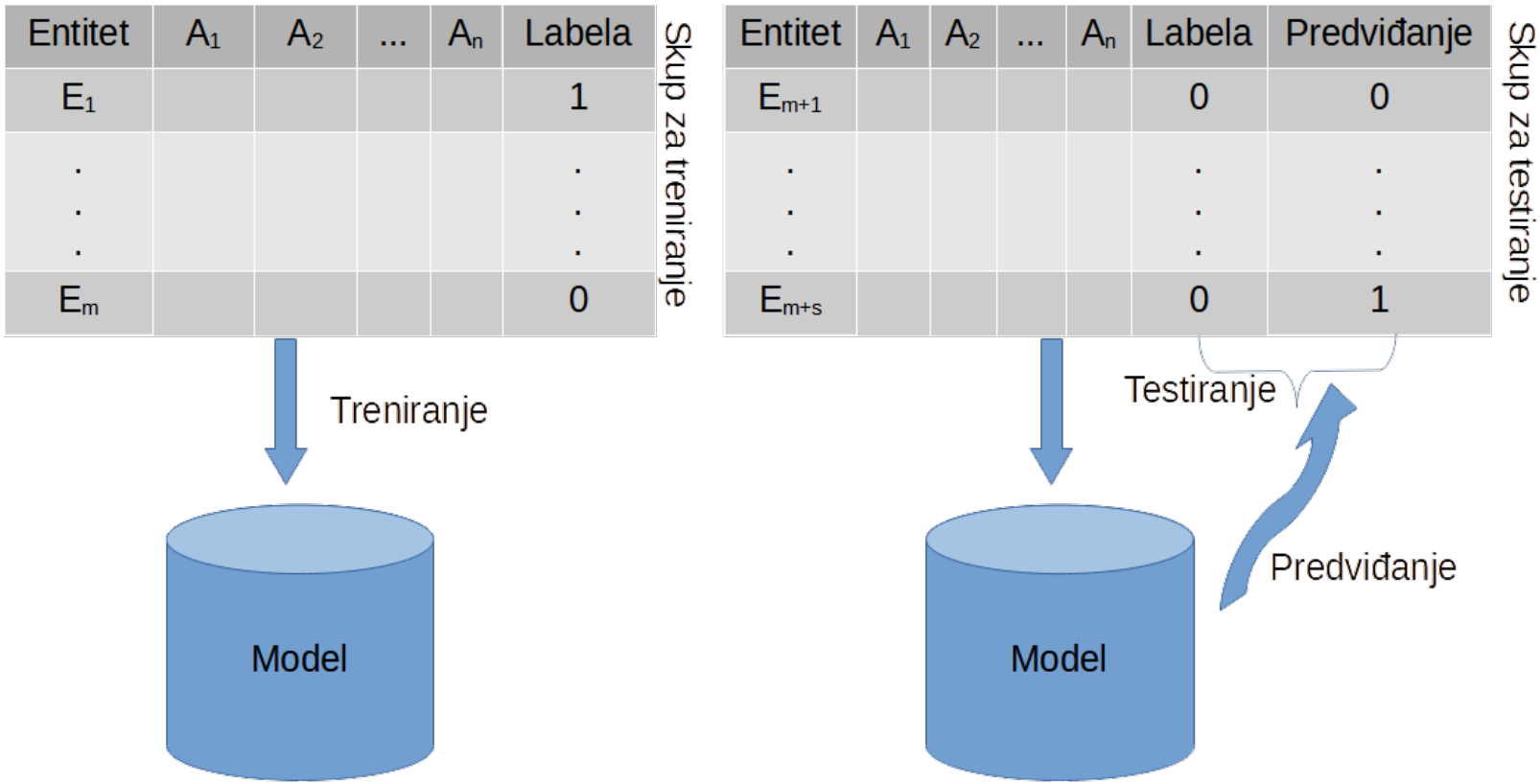
Mjere teorije informacija

- **Porast informacije** - mjeri porast informacije nakon podjele skupa podataka vrijednostima izabranog atributa. Označimo sa x_a vrijednost attributa a za entitet x. Skup $S_a(v) = \{x \in T |\ x_a = v\}$. Porast informacije definiramo kao: $PI(D, a) = H(D) - H(D|a)$. Dakle, $PI(D, a) = H(D) - \sum_{v \in vals(a)} P_a(v) H(S_a(v))$.

  U našem primjeru $H(D) = -0.5 \cdot log(0.5) - 0.5 \cdot log(0.5) = 1$, zato što u skupu imamo dva primjera s vrijednosti *DA* atributa *Voće* i dva primjera s vrijednosti *NE* (stoga je vjerojatnost svake vrijednosti 0.5). Pretpostavimo da podijelimo podatke koristeći vrijednosti atributa *Okus*. Dobijemo dva podskupa skupa podataka, jedan sadrži sve podatke za koje vrijedi Okus = Sladak, a drugi sadrži sve podatke za koje vrijedi Okus = Gorak. $H(S_{\text{Okus}}(\text{Sladak})) = H(S_{\text{Okus}}(\text{Gorak})) = 0$, zato što svi primjeri u skupu imaju identičnu vrijednost atributa *Voće*. Stoga je $PI(D, \text{Okus}) = 1 - 0.5 \cdot 0 - 0.5 \cdot 0 = 1$.

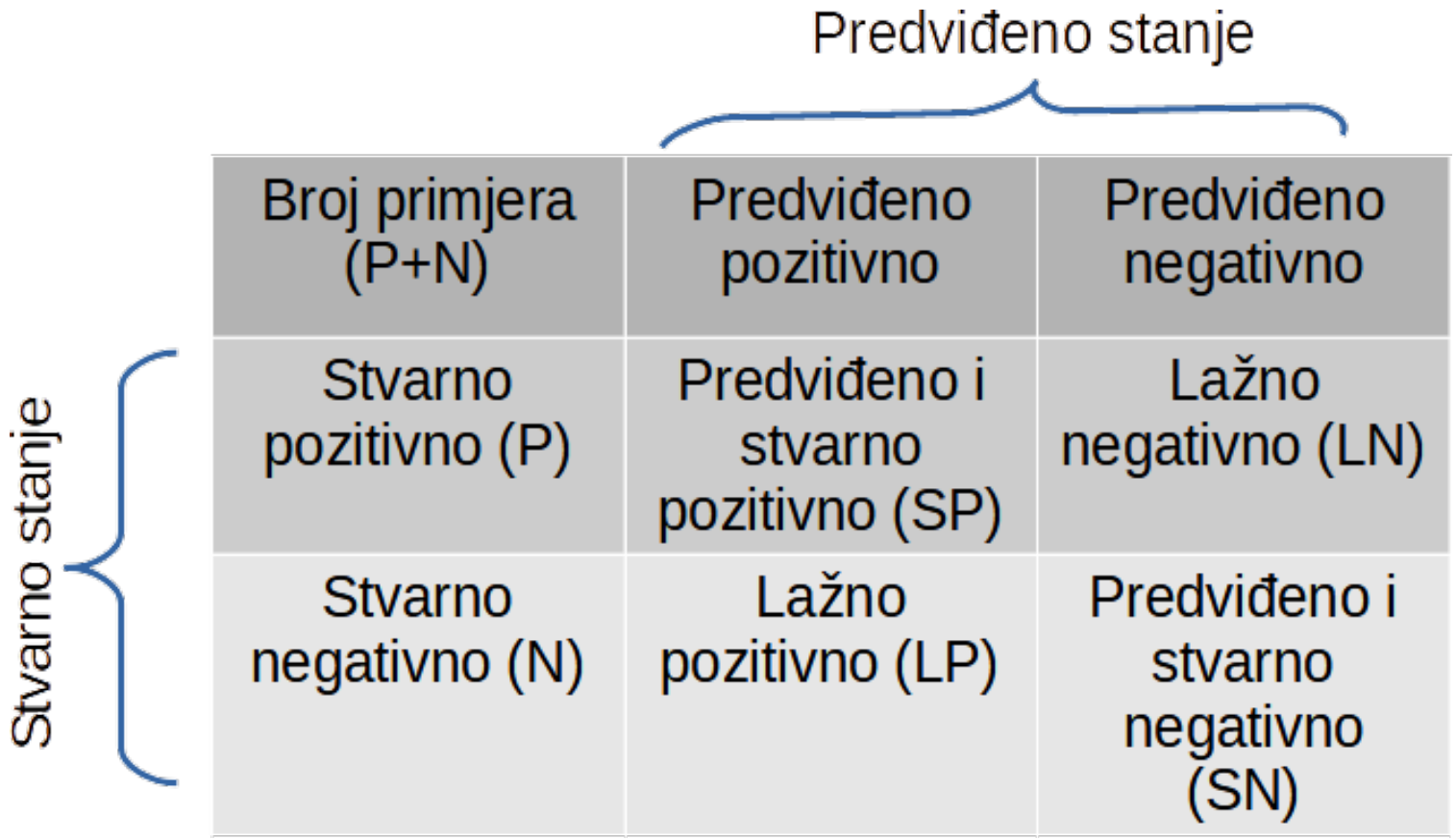
Mjere uz prisutnost ciljne labele

Postavke zadatka za koji definiramo mjere:



Mjere uz prisutnost ciljne labele

Predviđanja nekog prediktivnog, klasifikacijskog modela sistematiziramo **matricom konfuzije**.
Na donjoj slici je generalna matrica konfuzije za klasifikacijske probleme kod kojih ciljna klasa ima dvije različite vrijednosti *DA* ili *NE*, 1 ili 0 itd.



Mjere uz prisutnost ciljne labele

Matricu konfuzije možemo generalizirati i za klasifikacijske probleme čija ciljna labela ima više od dvije vrijednosti.

Broj primjera (A_1+A_2+...+A_n)	Predviđena vrijednost: A_1	Predviđena vrijednost: A_2	. . .	Predviđena vrijednost: A_n
Stvarna vrijednost klase: A_1	Predviđeno i stvarno A_1			
Stvarna vrijednost klase: A_2		Predviđeno i stvarno A_2		
. . .			. . .	
Stvarna vrijednost klase: A_n				Predviđeno i stvarno A_n

Mjere uz prisutnost ciljne labele

Pretpostavimo da je za neki problem model predvidio ciljnu binarnu klasu kao na primjeru:

Entitet.	A_1	A_2	...	Labela (klasa)	Predviđanje
E_1	.	.	.	1	1
E_2	.	.	.	0	1
E_3	.	.	.	0	0
E_4	.	.	.	1	1

Mjere uz prisutnost ciljne labele

Matrica konfuzije za gornji primjer je:

Broj primjera: 4	Predviđeno pozitivno	Predviđeno negativno
Stvarno pozitivno (2)	2	0
Stvarno negativno (2)	1	1

Dakle: $SP = 2$, $SN = 1$, $LN = 0$ i $LP = 1$.

Mjere uz prisutnost ciljne labele

- **Preciznost** - definira se kao: $P(\mathcal{M}, D_{test}) = \frac{SP}{SP + LP}$. $\mathcal{M}$ označava klasifikacijski model, a D_{test} skup podataka za testiranje. Udio točno predviđenih pozitivnih primjera od svih primjera koje je model klasificirao kao pozitivne. U gornjem primjeru, $P(\mathcal{M}, D_{test}) = \frac{2}{2+1} = 0.66\dot{6}$.
- **Odziv** (eng. *recall* ili *sensitivity* ili *true positive rate*) - definira se kao: $R(\mathcal{M}, D_{test}) = \frac{SP}{SP + LN}$. Udio točno pozitivno predviđenih primjera od svih pozitivnih primjera u skupu podataka za testiranje. $R(\mathcal{M}, D_{test}) = \frac{2}{2+0} = 1$.
- **Točnost** (eng. *accuracy*) - definira se kao: $Toc(\mathcal{M}, D_{test}) = \frac{SP + SN}{SP + SN + LP + LN}$. Udio točno predviđenih primjera od svih primjera u skupu podataka. U gornjem primjeru, $Acc(\mathcal{M}, D_{test}) = \frac{2+1}{2+1+1} = 0.75$.

Mjere uz prisutnost ciljne labele

- **Specifičnost** (eng. *specificity* ili *true negative rate*) - definira se kao: $Spec(\mathcal{M}, D_{test}) = \frac{SN}{SN + LP}$. Udio točno predviđenih primjera klase 0 (negativ) od ukupnog broja primjera koji pripadaju klasi 0. U gornjem primjeru, $Spec(\mathcal{M}, D_{test}) = \frac{1}{1+1} = 0.5$.
- **Ispadanje** (eng. *fall-out* ili *false positive rate*) - definira se kao: $Isp(\mathcal{M}, D_{test}) = \frac{LP}{LP + SN}$. Udio primjera koje je klasifikator predvidio kao pozitivne (npr. klasa 1), a koji imaju stvarnu ciljnu labelu 0, od svih negativnih primjera (onih s klasom 0) u skupu podataka. U gornjem primjeru, $Isp(\mathcal{M}, D_{test}) = \frac{1}{1+1} = 0.5$.
- **Omjer promašaja** (eng. *miss rate* ili *false negative rate*) - definira se kao: $Prom(\mathcal{M}, D_{test}) = \frac{LN}{LN + SP}$. Omjer primjera koje je klasifikator predvidio kao negativne (npr. klasa 0), a zapravo su pozitivne (pripadaju npr. klasi 1) od svih primjera u skupu za testiranje koji pripadaju klasi 1. U gornjem primjeru, $Prom(\mathcal{M}, D_{test}) = \frac{0}{0+2} = 0$.

Mjere uz prisutnost ciljne labele

- **Omjer lažnih otkrića** (eng. *false discovery rate*) - definira se kao: $LO(\mathcal{M}, D_{test}) = \frac{LP}{SP + LP}$. Udio lažno pozitivno predviđenih primjera kroz broj ukupno pozitivno predviđenih primjera od strane modela. U gornjem primjeru, $LO(\mathcal{M}, D_{test}) = \frac{1}{2+1} = 0.33\dot{3}$.
- **Balansirana točnost** - definira se kao: $BToc(\mathcal{M}, D_{test}) = \frac{R(\mathcal{M}, D_{test}) + Spec(\mathcal{M}, D_{test})}{2}$. Prosjek udjela točno predviđenih pozitivnih primjera od svih pozitivnih i udjela točno predviđenih negativnih primjera od svih negativnih u skupu za testiranje. U gornjem primjeru, $BToc(\mathcal{M}, D_{test}) = \frac{1+0.5}{2} = 0.75$.
- F_1 - definira se kao: $F_1(\mathcal{M}, D_{test}) = 2 \cdot \frac{P(\mathcal{M}, D_{test}) \cdot R(\mathcal{M}, D_{test})}{P(\mathcal{M}, D_{test}) + R(\mathcal{M}, D_{test})}$. Ovo je ekvivalentno s: $F_1(\mathcal{M}, D_{test}) = \frac{2SP}{2SP + LP + LN}$. U gornjem primjeru, $F_1(\mathcal{M}, D_{test}) = \frac{2 \cdot 2}{2 \cdot 2 + 1 + 0} = 0.8$.

Mjere uz prisutnost ciljne labele

- **Omjer lažnih otkrića negativne vrijednosti klase** (eng. *false omission rate*) - definira se kao: $LON(\mathcal{M}, D_{test}) = \frac{LN}{SN + LN}$. Udio lažno negativno predviđenih primjera kroz ukupan broj negativno (npr. klasa 0) predviđenih primjera od strane modela. U gornjem primjeru, $LON(\mathcal{M}, D_{test}) = \frac{0}{1+0} = 0$.
- **Preciznost predviđanja negativne klase** (eng. *negative predictive value*) - definira se kao: $PN(\mathcal{M}, D_{test}) = \frac{SN}{SN + LN}$. Udio točno predviđenih negativnih primjer (npr. ciljne klase vrijednosti 0) od svih primjera koje je model klasificirao kao negativne. U gornjem primjeru, $PN(\mathcal{M}, D_{test}) = \frac{1}{1+0} = 1$.
- **Obilježenost** (eng. *markedness*) - definira se kao: $Ob(\mathcal{M}, D_{test}) = P(\mathcal{M}, D_{test}) + PN(\mathcal{M}, D_{test}) - 1$. Označava u kojoj mjeri klasifikator daje točna predviđanja točnih vrijednosti ciljnih labela. Na njega manje utječu nebalansirani podaci ili neravnomjernosti u distribucijama vrijednosti ciljnih labela.

Mjere uz prisutnost ciljne labele

- **Obilježenost** (nastavak) - u gornjem primjeru, $Ob(\mathcal{M}, D_{test}) = 0.66\dot{6} + 1 - 1 = 0.66\dot{6}$.
- **Matthews koeficijent korelacije** - definira se kao: $MKK(\mathcal{M}, D_{test}) = \frac{}{\sqrt{R(\mathcal{M}, D_{test}) \cdot Spec(\mathcal{M}, D_{test}) \cdot P(\mathcal{M}, D_{test}) \cdot PN(\mathcal{M}, D_{test})} - \sqrt{Prom(\mathcal{M}, D_{test}) \cdot Isp(\mathcal{M}, D_{test}) \cdot LON(\mathcal{M}, D_{test}) \cdot LO(\mathcal{M}, D_{test})}}$. Mjeri korelaciju između stvarne vrijednosti ciljne klase i vrijednosti predviđene od strane klasifikatora. U gornjem primjeru, $MKK(\mathcal{M}, D_{test}) = \sqrt{1 \cdot 0.5 \cdot 0.66\dot{6} \cdot 1} - \sqrt{0 \cdot 0.5 \cdot 0 \cdot 0.33\dot{3}} = 0.57735$.
- **Indeks kritičnog uspjeha** (eng. *critical success index ili Jaccard index*) - definira se kao: $IKU(\mathcal{M}, D_{test}) = \frac{SP}{SP + LN + LP}$. Zanemaruje točno predviđene negativne primjere. Koristi se kada je puno bitnije točno predvidjeti primjere s pozitivnom vrijednosti ciljne labele (npr. 1). U gornjem primjeru, $IKU(\mathcal{M}, D_{test}) = \frac{2}{2+0+1} = 0.66\dot{6}$.

Mjere uz prisutnost ciljne labele

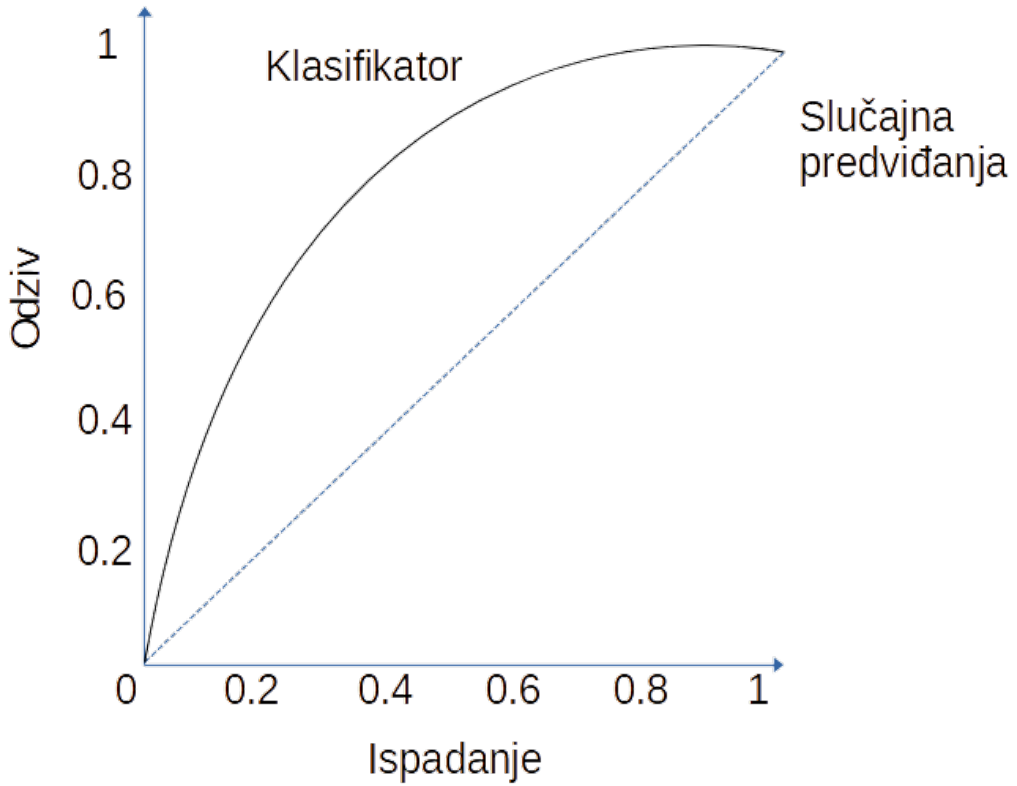
- **Udio vjerodostojnosti pozitivne vrijednosti klase** (eng. *positive likelihood ratio*) - definira se kao: $PUV(\mathcal{M}, D_{test}) = \frac{R(\mathcal{M}, D_{test})}{Isp(\mathcal{M}, D_{test})}$. Govori koliko činjenica da klasifikator predviđa pozitivnu vrijednost ciljne labele nekog entiteta zapravo povećava vjerodostojnost da je to zaista tako. Korisno npr. kod predviđanja bolesti, $PUV > 1$ označava pozitivan pomak, a $PUV < 1$ negativan pomak. U gornjem primjeru, $PUV(\mathcal{M}, D_{test}) = \frac{1}{0.5} = 2$.
- **Udio vjerodostojnosti negativne vrijednosti klase** (eng. *negative likelihood ratio*) - definira se kao: $NUV(\mathcal{M}, D_{test}) = \frac{Prom(\mathcal{M}, D_{test})}{Spec(\mathcal{M}, D_{test})}$. Govori koliko činjenica da klasifikator predviđa negativnu vrijednost ciljne labele nekog entiteta zapravo povećava vjerodostojnost da je to zaista tako. U gornjem primjeru, $NUV(\mathcal{M}, D_{test}) = \frac{0}{0.5} = 0$. NUV blizu ili jednak 0 daje visoku pouzdanost da klasifikator točno predviđa da je stvarna vrijednost ciljne labele entiteta negativna (npr. 0).

Mjere uz prisutnost ciljne labele

- **Udio vjerodostojnosti negativne vrijednosti klase** (nastavak) - može se i dalje dogoditi da klasifikator predvidi vrijednost ciljne labele 1, a da stvarna vrijednost bude 0.
- **Dijagnostički omjer izgleda** (eng. *diagnostic odds ratio*) - definira se kao: $DOI(\mathcal{M}, D_{test}) = \frac{PUV(\mathcal{M}, D_{test})}{NUV(\mathcal{M}, D_{test})} = \frac{SP \cdot SN}{LP \cdot LN}$. Mjeri učinkovitost klasifikatora (dijagnostičkog testa). U gornjem primjeru, $DOI(\mathcal{M}, D_{test}) = \frac{2}{0} = +\infty$. U idealnom slučaju, ova vrijednost bi predstavljala dijagnostički jako učinkovit test (false negative rate je 0), međutim u našem slučaju se radi o vrlo malom uzorku podataka.

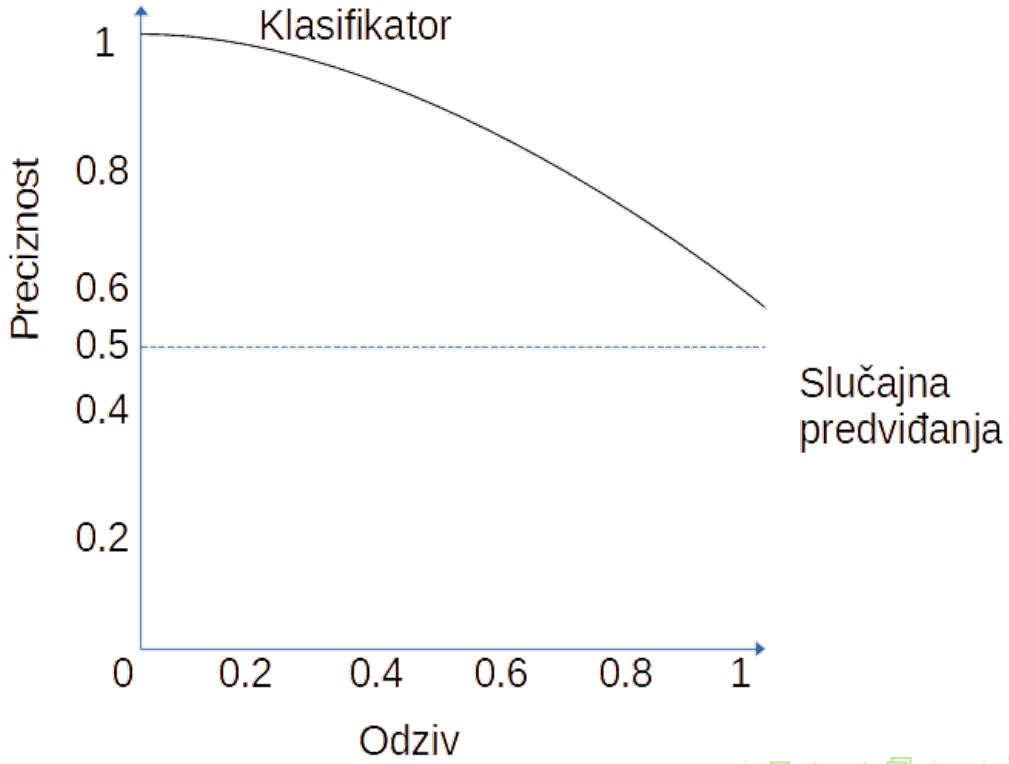
Mjere uz prisutnost ciljne labele

- **Površina ispod ROC krivulje** (eng. *area under the ROC curve* - AUC) - dobiva se tako da se prag mjere, koja mjeri pouzdanost klasifikatora u svoje predviđanje, lagano povećava. Za svaki prag računamo *Ispadanje* i *Odziv* klasifikatora, koje crtamo na ko-ordinatnom sustavu, gdje x-os čini *Ispadanje*, a y-os *Odziv*.



Mjere uz prisutnost ciljne labele

- **Površina ispod krivulje preciznosti i odziva** (eng. *area under the precision - recall curve* - AUPRC) - dobiva se tako da se prag mjere, koja mjeri pouzdanost klasifikatora u svoje predviđanje, lagano povećava. Za svaki prag računamo *Odziv* i *Preciznost* klasifikatora, koje crtamo na ko-ordinatnom sustavu, gdje x-os čini *Odziv*, a y-os *Preciznost*.



Mjere uz prisutnost cilja

Pretpostavimo da je za neki problem model predvidio numerički cilj kao na primjeru:

Entitet.	A_1	A_2	...	Labela (cilj)	Predviđanje
E_1	.	.	.	0.5	0.6
E_2	.	.	.	1.2	0.9
E_3	.	.	.	2.6	2.5
E_4	.	.	.	0.7	1.2

Mjere uz prisutnost cilja

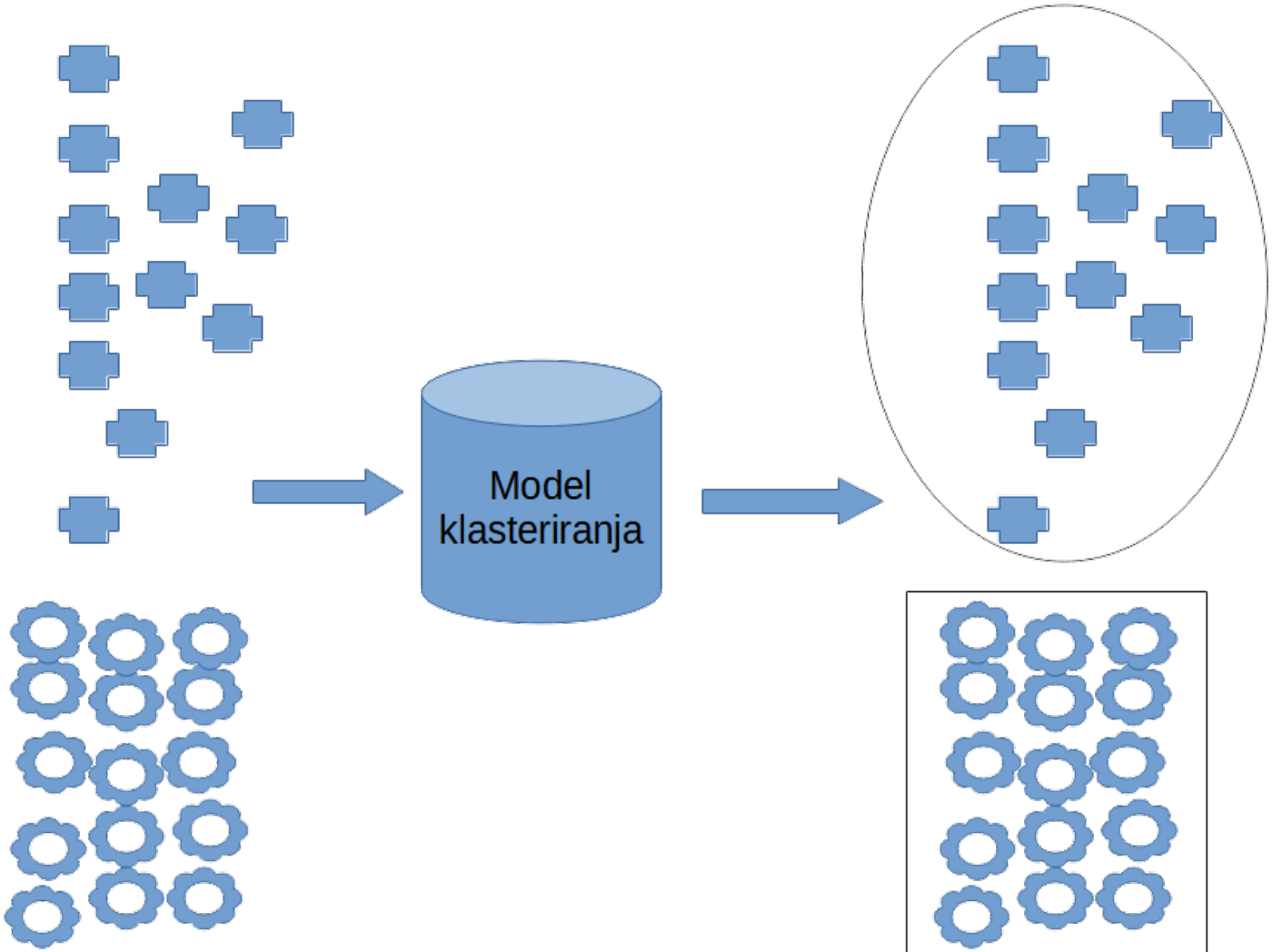
- **Srednja apsolutna pogreška** (eng. *mean absolute error* - MAE), definira se kao: $SAP(\mathcal{M}, D_{test}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$. U našem primjeru $SAP(\mathcal{M}, D_{test}) = \frac{1}{4}$.
- **Srednja kvadratna pogreška** (eng. *mean squared error* - MSE), definira se kao: $SKP(\mathcal{M}, D_{test}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. Snažnije penalizira veće pogreške. U našem primjeru $SKP(\mathcal{M}, D_{test}) = \frac{1}{4} \cdot 0.36 = 0.09$.
- **Korijen srednje kvadratne pogreške** (eng. *root mean squared error* - RMSE), definira se kao $KSKP(\mathcal{M}, D_{test}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$. Snažnije penalizira veće pogreške, ali rezultat prijavljuje u numeričkom rasponu bliskom originalnim vrijednostima cilja. U našem primjeru $KSKP(\mathcal{M}, D_{test}) = \sqrt{0.09} = 0.3$.

Mjere uz prisutnost cilja

- R^2, definira se kao $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$, gdje $\bar{y}$ označava prosjek stvarnih vrijednosti cilja. Mjeri proporciju varijance cilja koju objašnjava regresijski model. U našem primjeru $R^2 = 1 - \frac{0.36}{2.69} = 0.8662$.
- **Srednja apsolutna postotna pogreška** (eng. *mean absolute percentage error* - MAPE), definira se kao $SAPP(\mathcal{M}, D_{test}) = \frac{100}{n}\sum_{i=1}^{n}|\frac{y_i - \hat{y}_i}{y_i}|$. Iskazuje pogrešku kao postotak stvarnih vrijednosti cilja. Mjera je nepouzdana kada su stvarne vrijednosti cilja blizu 0. U našem primjeru $SAPP(\mathcal{M}, D_{test}) = 25 \cdot 1.202747 = 30.068681$.

Mjere uz prisutnost ciljne labele

Sljedeće mjere se koriste za evaluaciju klasteriranja uz prisutnost ciljnih labela.



Mjere uz prisutnost ciljne labele

Entitet	A_1	. . .	A_k	Ciljna labela	Predviđanje
E_1	.	.	.	C_1	C_1
E_2	.	.	.	C_2	C_2
E_3	.	.	.	C_2	C_3
E_4	.	.	.	C_1	C_1

Mjere uz prisutnost ciljne labele

Neka je $S = \{E_1, \ldots, E_n\}$ neki skup entiteta, a $X = \{X_1, \ldots, X_l\}$ i $Y = \{Y_1, \ldots, Y_s\}$ dvije particije elemenata iz S. Ukoliko sa $p_{X,Y}$ označimo broj parova elemenata koji se nalaze u istoj particiji i u X i u Y, a sa $o_{X,Y}$ broj parova elemenata koji se nalaze u različitim particijama i u X i u Y, tada definiramo:

- **Randov indeks** - $RI(X, Y, D) = \frac{p_{X,Y} + o_{X,Y}}{\binom{n}{2}}$. U našem primjeru:

  $p_{X,Y} = 1$, $o_{X,Y} = 4$, $RI(X, Y, D) = \frac{4+1}{6} = 0.83\dot{3}$.
- **Prilagođen Randov index** - umanjuje vrijednost Randovog indeksa uzimajući u obzir očekivani broj parova elemenata koji zadovoljavaju tražena svojstva zbog slučajnosti. Npr. za veliki broj parova elemenata koji se **ne nalaze** u istoj particiji i u X i u Y je to posljedica slučajnosti, a ne točnosti klasteriranja.

Mjere uz prisutnost ciljne labele

- **Prilagođen Randov index** (nastavak) - za gore definirane particije prvo definiramo matricu kontingencije kao:

X\Y	Y_1	Y_2	. . .	Y_s	sume
X_1	$n_{1,1}$	$n_{1,2}$	. . .	$n_{1,s}$	A_1
X_2	$n_{2,1}$	$n_{2,2}$	. . .	$n_{2,s}$	A_2
. . .	. . .	. . .	. . .	. . .	. . .
X_I	$n_{I,1}$	$n_{I,2}$	. . .	$n_{I,s}$	A_I
sume	B_1	B_2	. . .	B_s	

Mjere uz prisutnost ciljne labele

- **Prilagođen Randov index** (nastavak) - $n_{i,j} = |X_i \cap Y_j|$. Definiramo:
$$PRI(X, Y, D) = \frac{\sum_{i=1}^{I}\sum_{j=1}^{s}\binom{n_{i,j}}{2} - [\sum_{i=1}^{I}\binom{a_i}{2}\sum_{j=1}^{s}\binom{b_j}{2}]/\binom{n}{2}}{0.5\cdot[\sum_{i=1}^{I}\binom{a_i}{2} + \sum_{j=1}^{s}\binom{b_j}{2}] - [\sum_{i=1}^{I}\binom{a_i}{2}\sum_{j=1}^{s}\binom{b_j}{2}]/\binom{n}{2}}.$$

Tablica kontingencije za gornji primjer je:

X/Y	Y_1	Y_2	Y_3	sume
X_1	2	0	0	2
X_2	0	1	1	2
sume	2	1	1	

Mjere uz prisutnost ciljne labele

- **Prilagođen Randov index** (nastavak) - prilagođen randov index u našem primjeru je:
  $PRI(X, Y, D) = \frac{1-(2\cdot 1)/6}{0.5\cdot 3-(2\cdot 1)/6} = \frac{2/3}{7/6} = \frac{4}{7} = 0.57142857.$
- **Uzajamna informacija** - mjeri slaganje predviđenog klasteriranja i stvarnog. Označimo s $P_s(i)$ vjerojatnost da je neki entitet u stvarnoj grupi i, a s $P_p(j)$ vjerojatnost da je neki entitet u predviđenoj grupi j. $P(i,j)$ označava vjerojatnost da je entitet u stvarnoj grupi i i u predviđenoj grupi j. Uzajamnu informaciju definiramo kao:
  $MI(X, Y, D) = \sum_{i=1}^{|X|}\sum_{j=1}^{|Y|} P(i,j)\log\frac{P(i,j)}{P_s(i)P_p(j)}$. U našem primjeru
  $P_s(C_1) = P_s(C_2) = \frac{1}{2}$, $P_p(C_1) = \frac{1}{2}$, $P_p(C_2) = P_p(C_3) = \frac{1}{4}$,
  $P(C_1, C_1) = 1$, $P(C_1, C_2) = P(C_1, C_3) = P(C_2, C_1) = 0$,
  $P(C_2, C_2) = P(C_2, C_3) = \frac{1}{2}$. Zbog malog skupa podataka, postoje parovi kategorija za koje $P(C_i, C_j) = 0$, stoga MI ne možemo izračunati koristeći standardnu formulu. Kod praktičnih izračuna, u tom slučaju možemo ispustiti sve članove MI kod kojih $P(C_i, C_j) = 0$.

Mjere uz prisutnost ciljne labele

- **Normalizirana uzajamna informacija** - uzajamna informacija normalizirana s prosjekom entropije pravog i predviđenog klasteriranja. Definira se kao: $NMI(X, Y, D) = \frac{MI(X, Y, D)}{\frac{1}{2}(H(X) + H(Y))}$. Ponekad se normalizirana uzajamna informacija dodatno uštimava na način da se brojniku i nazivniku oduzme očekivana uzajamna informacija između dva slučajna klasteriranja u određeni (zadani) broj grupa. Koristi se ažuriranje vrijednošću dobivene iz hipergeometrijske distribucije.
- **V mjera** - računa odnos **homogenosti** (svaki klaster sadrži entitete samo jedne klase) i **potpunosti** (klaster sadrži sve entitete zadane klase) klasteriranja. Označimo sa X pravo klasteriranje, a s Y predviđeno klasteriranje. Homogenost $hom(X, Y, D) = 1 - \frac{H(X|Y)}{H(X)}$, potpunost $pot(X, Y, D) = 1 - \frac{H(Y|X)}{H(Y)}$.

Mjere uz prisutnost ciljne labele

- **V mjera** (nastavak) - V-mjera se definira kao: $V(X, Y, D) = \frac{(1+\beta) \cdot hom(X, Y, D) \cdot pot(X, Y, D)}{\beta \cdot hom(X, Y, D) + pot(X, Y, D)}$. β određuje odnos homogenosti i potpunosti. $\beta > 1$ daje veću težinu potpunosti, a $\beta < 1$ veću težinu homogenosti. Često se uzima $\beta = 1$ (harmonijska sredina).
- **Fowlkes-Mallows indeks** - definira se kao: $FMI(\mathcal{M}, D) = \sqrt{P(\mathcal{M}, D) \cdot R(\mathcal{M}, D)}$ - geometrijska sredina između preciznosti i odziva. Indeks računamo usporedbom stvarnih i predviđenih klasteriranja za sve parove entiteta. SP - par entiteta se nalazi u istom pravom i predviđenom klasteru, LP - par entiteta se nalazi u različitom pravom klasteru, ali u istom predviđenom klasteru, LN - par entiteta se nalazi u istom pravom klasteru, ali različitom predviđenom klasteru i SN - par entiteta se nalazi u različitom pravom i predviđenom klasteru. U našem primjeru: $FMI(\mathcal{M}, D) = \sqrt{1 \cdot \frac{1}{2}} = 0.7071$.

Mjere uz prisutnost ciljne labele

Preciznost, odziv, točnost i F_1 se mogu koristiti i za evaluaciju nadzirane segmentacije, kod koje labeliramo piksele slike, te uspoređujemo stvarnu klasu piksela i predviđenu klasu piksela.

Mjere uz prisutnost ciljne labele

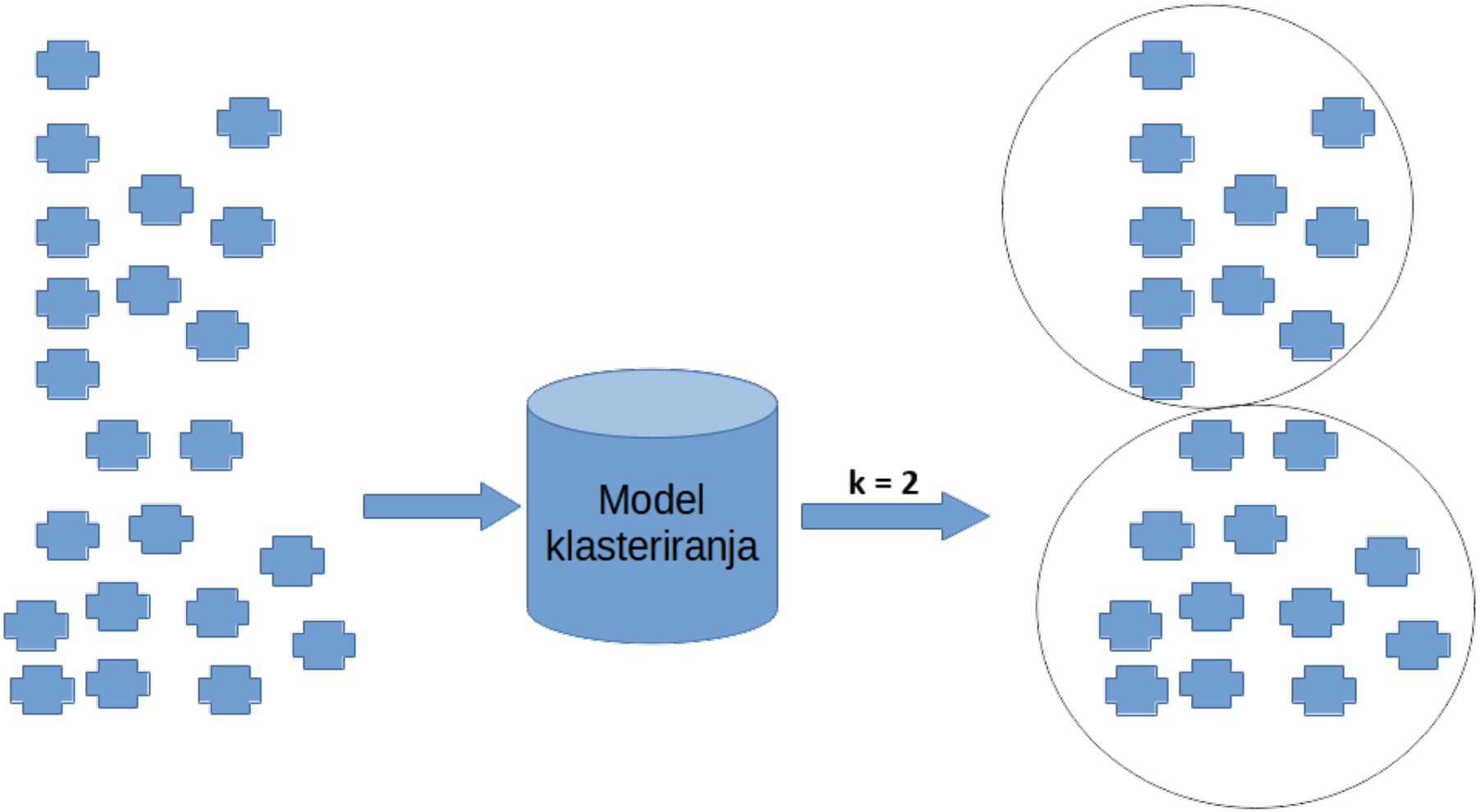
Uz navedene, kod nadzirane segmentacije još se koriste:

- **Presjek kroz unija** (eng. *intersection over union, IoU*) - definira se kao: $IoU(IM, IM_{pred}) = \frac{|IM \cap IM_{pred}|}{|IM \cup IM_{pred}|}$.
- **Dice koeficijent** (eng. *Dice coefficient*) - definira se kao: $DK(IM, IM_{pred}) = \frac{2 \cdot |IM \cap IM_{pred}|}{|IM| + |IM_{pred}|}$.
- **Srednja apsolutna pogreška** - definira se kao: $MAE(IM, IM_{pred}) = \frac{1}{n} \sum_{i=1}^{n} |p_i - \hat{p}_i|$, gdje p_i označava vrijednost i-tog piksela, a $\hat{p}_i$ predviđenu vrijednost i-tog piksela. U slučaju slika u boji (r, g, b) širine S i duljine D, definiramo $$MI(IM, IM_{pred}) = \frac{1}{S \cdot D \cdot 3} \sum_{i=1}^{S} \sum_{j=1}^{D} \sum_{b \in \{r,g,b\}} |p_{i,j,b} - p_{\hat{i},\hat{j},b}|.$$

Mjere uz prisutnost ciljne labele

- **Hausdorff-ova udaljenost** - mjeri koliko je rub predviđene segmentacije udaljen od stvarne segmentacije. Umjesto volumena preklapanja mjeri prostornu preciznost kontura. Definira se kao: $d_H(X, Y) = max\{sup_{x \in X} inf_{y \in Y}\ d(x, y),\ sup_{y \in Y} inf_{x \in X}\ d(x, y)\}$.

Mjere bez prisutnosti ciljne labele



Mjere bez prisutnosti ciljne labele

- **Koeficijent siluete** - računamo dvije vrijednosti, *kohezija* - prosječnu udaljenost točke $i \in C_i$ od svih ostalih točaka u C_i, $ad_k(i) = \frac{1}{|C_i| - 1} \sum_{l \in C_j,\ l \neq i} d(i, l)$, te *separaciju* - prosječnu udaljenost točke $i \in C_i$ od točaka u najbližem susjednom klasteru, $ad_s(i) = min_{j \neq i} \frac{1}{C_j} \sum_{l \in C_j} d(i, l))$. Koeficijent siluete entiteta i definiramo kao: $KS(i) = \frac{ad_s(i) - ad_k(i)}{max\{ad_s(i), ad_k(i)\}}$.
- **Calinski-Harabasz indeks** - mjeri omjer između *disperzije između klastera* i *disperzije unutar klastera*. Za dani broj klastera $k \in \mathbb{N}$, broj entiteta $N \in \mathbb{N}$, te klasteriranje $\mathcal{C}$, imamo: $CHI(\mathcal{C}) = \frac{DIK(\mathcal{C})}{DUK(\mathcal{C})} \cdot \frac{N - k}{k - 1}$. Disperziju između klastera definiramo kao: $DIK(\mathcal{C}) = \sum_{i=1}^{k} n_i \cdot \|c_i - c\|^2$, gdje c označava centroid skupa

Mjere bez prisutnosti ciljne labele

- **Calinski-Harabasz indeks** (nastavak) - centroid skupa podataka je prosjek svih točaka u skupu podataka. Disperzija unutar klastera definiramo kao: $DUK(\mathcal{C}) = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - c_i\|^2$.
- **Davies-Bouldin indeks** - označimo sa $d_c(i,j)$ udaljenost centroida c_i klastera C_i i centroida c_j klastera C_j. Definiramo $R_{i,j}(C_i, C_j) = \frac{\sum_{x \in C_i} \|x - c_i\|^2 + \sum_{x \in C_j} \|x - c_j\|^2}{d_c(i,j)}$. Definiramo $R_i(\mathcal{C}) = max_{j \neq i} R_{i,j}$. $DBI(\mathcal{C}) = \frac{1}{k} \sum_{i=1}^{k} R_i$. **Manja** vrijednost indeksa označava bolje klasteriranje. Niža vrijednost označava kompaktnije klasteriranje unutar klastera i veću udaljenost između različitih klastera.