Vežbe br. 8

MAŠINSKO UČENJE

Logistička regresija

- Algoritam nadgledanog učenja, koristi se za rešavanje klasifikacionih problema
- Najčešće se koristi za binarnu klasifikaciju, tj. za odgovaranje na „da – ne“ pitanja
- Predstavlja vid nadgradnje linearne regresije, jer uzima njen izlaz i pretvara ga u binarni oblik
- Može se koristiti i za procenjivanje verovatnoće da određeni uzorak pripada određenoj klasi

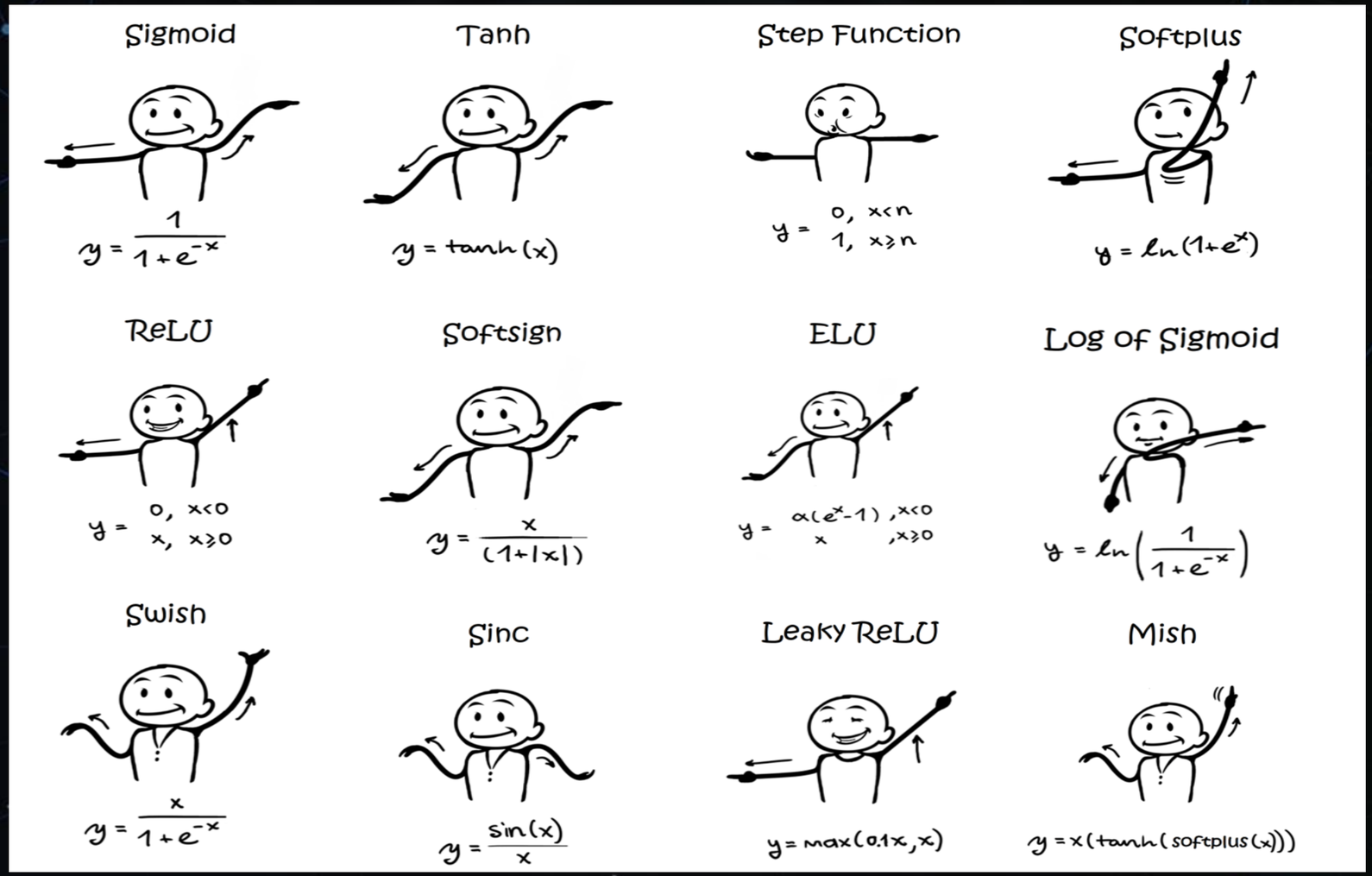
Logistička regresija

- Da bi logistička regresija davala dobre rezultate za neki skup podataka, potrebno je da budu ispunjeni sledeći preduslovi:
  - Dovoljno veliki skup podataka (što veći)
  - Bez anomalija
  - Bez velikih korelacija između nezavisnih promenljivih
  - Postoji linearna zavisnost između nezavisnih promenljivih i „log odds“ („logaritma šanse“) zavisne promenljive

Aktivacione funkcije

- Aktivaciona funkcija je funkcija čija je svrha da neku ulaznu vrednost (ili skup ulaznih vrednosti) mapira u izlaznu vrednost, koja će se nalaziti u opsegu [0, 1]
- Kod logističke regresije, ulazni podatak je ono što daje linearna regresija, a izlaz predstavlja verovatnoću pripadnosti nekoj klasi
- Aktivacione funkcije najčešće se pominju u radu sa neuronskim mrežama
- Sinonim za aktivacionu funkciju je ***funkcija prenosa***

Aktivacione funkcije



Sigmoid

Aktivaciona funkcija koju koristi logistička regresija

Neka je dat skup nezavisnih varijabli i očekivanih izlaza:

$$X = \begin{vmatrix} x_{11} & \dots & x_{1m} \\ x_{21} & \dots & x_{2m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nm} \end{vmatrix} \qquad Y = \begin{cases} 0 & \text{if } Class\ 1 \\ 1 & \text{if } Class\ 2 \end{cases}$$

Potrebno je primeniti multi-linearnu funkciju na skup nezavisnih varijabli:

$$z = (\sum_{i=1}^{n} w_i x_i) + b$$

Nakon toga, koristi se jednačina sigmoid funkcije za određivanje verovatnoće:

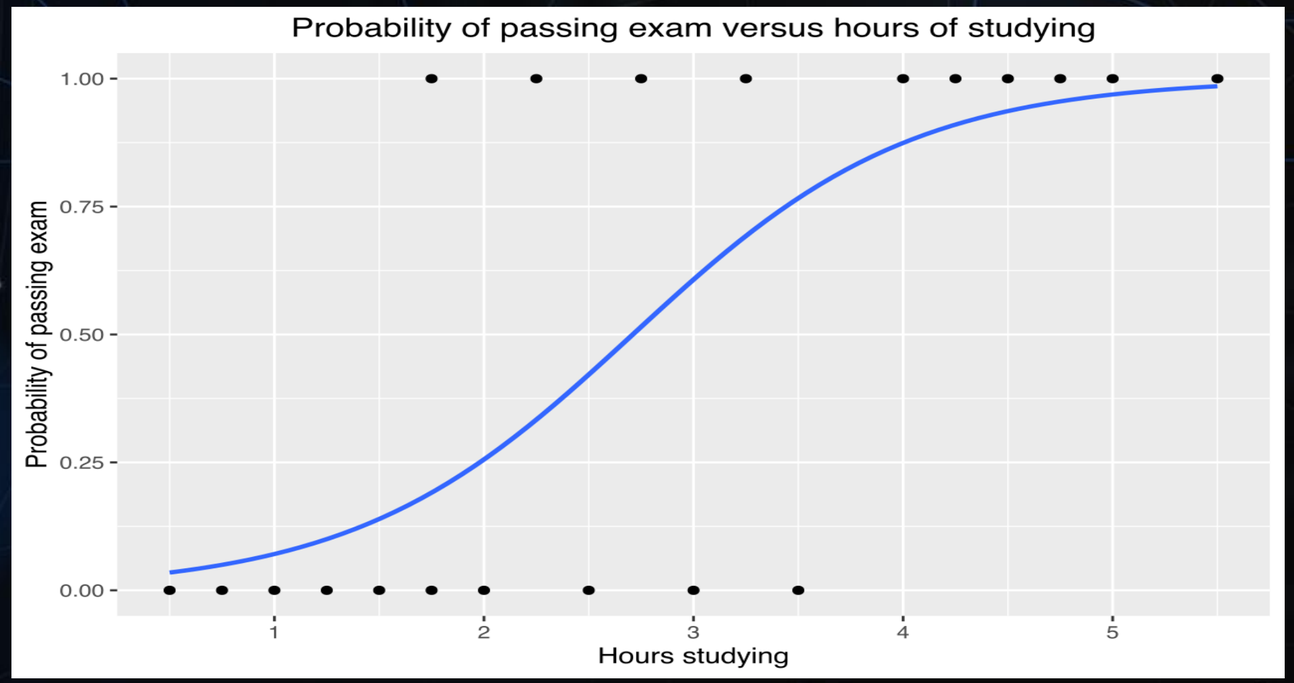
$$\sigma(z) = \frac{1}{1-e^{-z}}$$

Sigmoid

Za dati skup podataka:

Hours (x_k)	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass (y_k)	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1

Primena sigmoid aktivacione funkcije daje sledeću krivu:

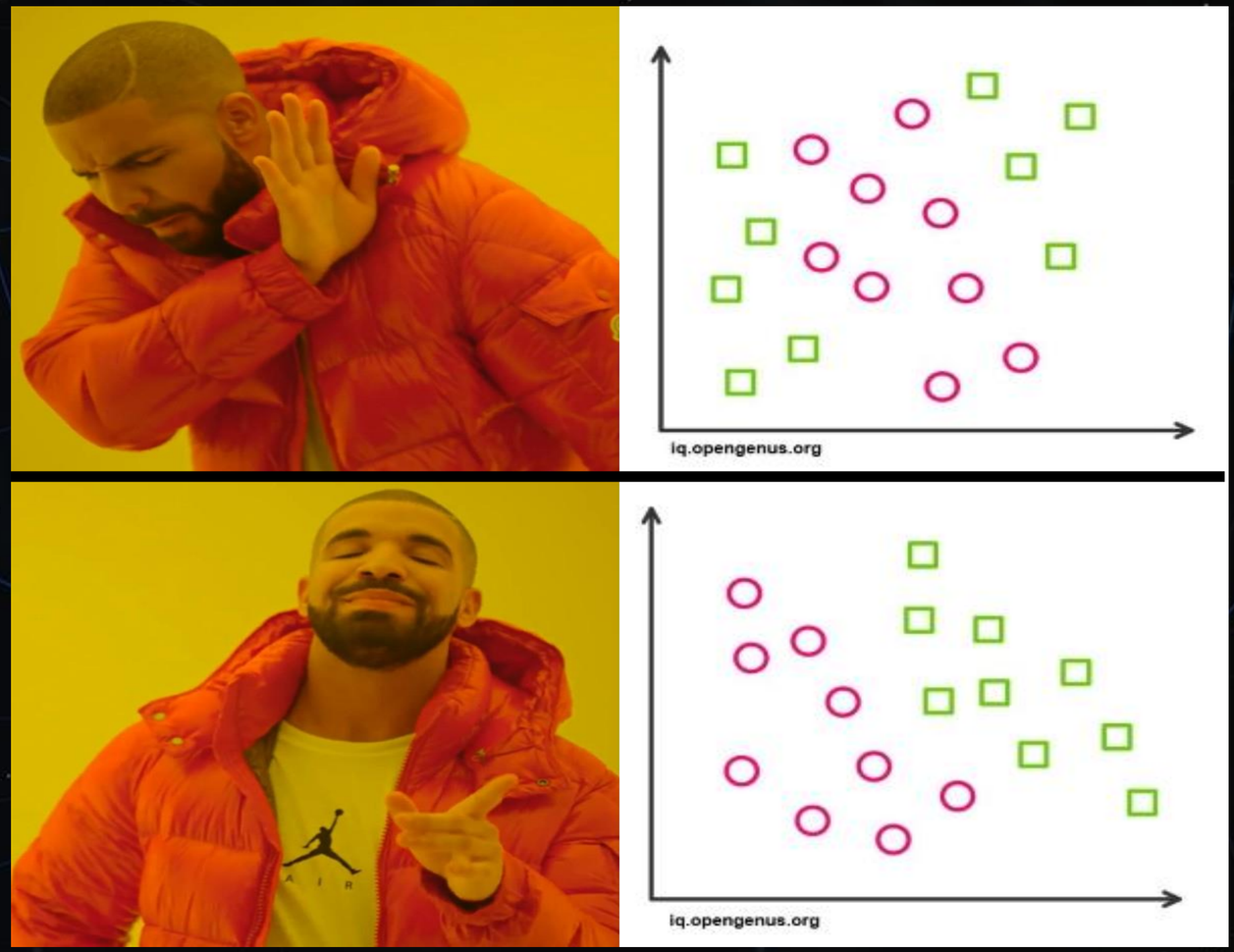


Logistička regresija

- Prednosti:
  - Model jednostavan za interpretaciju
  - Brzo računanje
  - Daje informaciju o značaju svakog atributa

- Mane:
  - Nekada previše jednostavna
  - Podložna overfitting-u (u nekim slučajevima)
  - Zahteva linearne odnose, zbog čega se retko koristi u stvarnim problemima

Logistička regresija



Zadatak za vežbu

- Kreirati model, zasnovan na logističkoj regresiji, koji će predviđati da li osoba ima ili nema kancer. Koristiti ugrađeni skup podataka „load_breast_cancer“. Prikazati uspešnost kreiranog klasifikatora.

Rešenje

```
from sklearn.datasets import load_breast_cancer
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

import warnings
from pandas.core.common import SettingWithCopyWarning
warnings.simplefilter(action='ignore', category=FutureWarning)
warnings.simplefilter(action='ignore', category=UserWarning)
warnings.simplefilter(action='ignore', category=SettingWithCopyWarning)

X, y = load_breast_cancer(return_X_y=True)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, stratify=y)

clf = LogisticRegression()
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)

acc = accuracy_score(y_test, y_pred)
print("Tačnost modela logističke regresije (u %):", "%.2f" % (acc * 100))
```

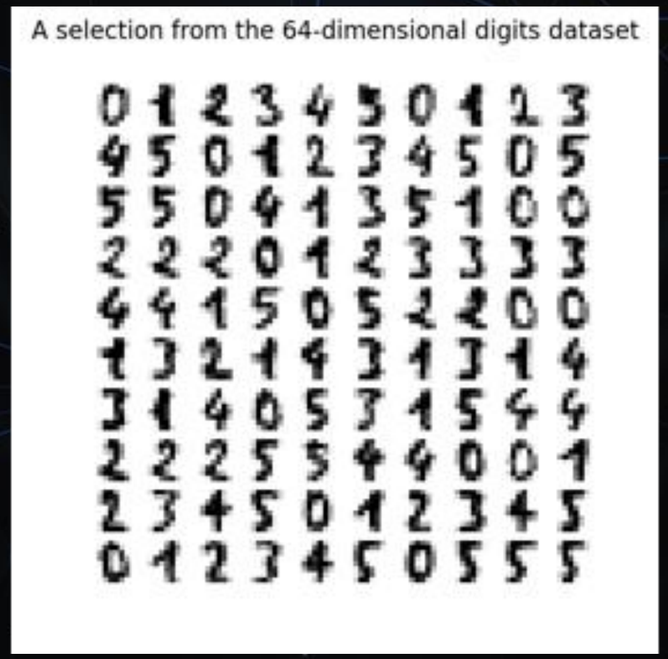
Multinomijalna logistička regresija

- Logistička regresija može se koristiti i u slučajevima kada zavisna varijabla može da ima više od 2 vrednosti
- Tada se koristi poseban oblik logističke regresije koji se zove ***multinomijalna logistička regresija***
- Umesto sigmoid aktivacione funkcije, kod multinomijalne logističke regresije koristi se ***softmax***

$$Pr(Y = c|\overrightarrow{X} = x) = \frac{e^{w \cdot x + b}}{\sum_{k=1}^{K} e^{w \cdot x + b}}$$

Zadatak za vežbu

- Kreirati model, zasnovan na logističkoj regresiji, koji će prepoznavati koja cifra se nalazi na slici. Koristiti ugrađeni skup podataka „digits“. Prikazati uspešnost kreiranog klasifikatora. Analizirati razliku u odnosu na prethodni primer.



Rešenje

```
from sklearn.datasets import load_digits
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split

import warnings
from pandas.core.common import SettingWithCopyWarning
warnings.simplefilter(action='ignore', category=FutureWarning)
warnings.simplefilter(action='ignore', category=UserWarning)
warnings.simplefilter(action='ignore', category=SettingWithCopyWarning)

X, y = load_digits(return_X_y=True)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, stratify=y)

clf = LogisticRegression()
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)

acc = accuracy_score(y_test, y_pred)
print("Tačnost modela logističke regresije (u %):", "%.2f" % (acc * 100))
```

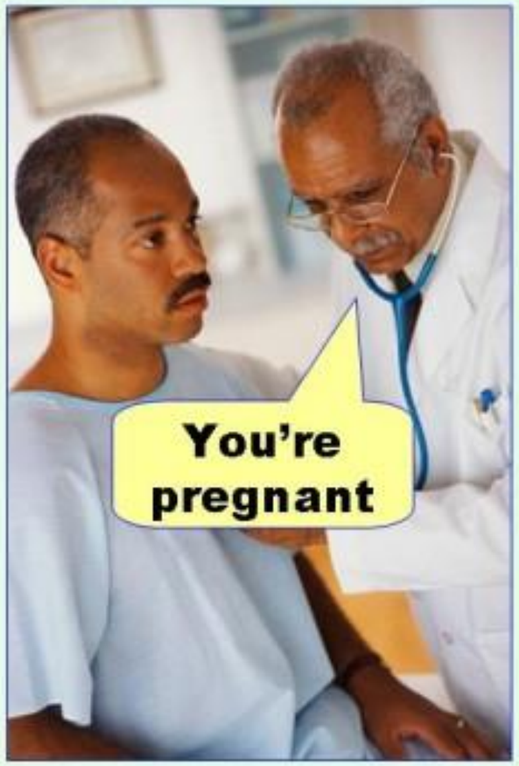
Merenje performansi klasifikatora

- Ako se priča o binarnoj klasifikaciji, postoje 4 moguća ishoda klasifikacije:
  - True Positive (pozitivni primerci klasifikovani kao pozitivni)
  - True Negative (negativni primerci klasifikovani kao negativni)
  - False Positive (negativni primerci klasifikovani kao pozitivni)
  - False Negative (pozitivni primerci klasifikovani kao negativni)

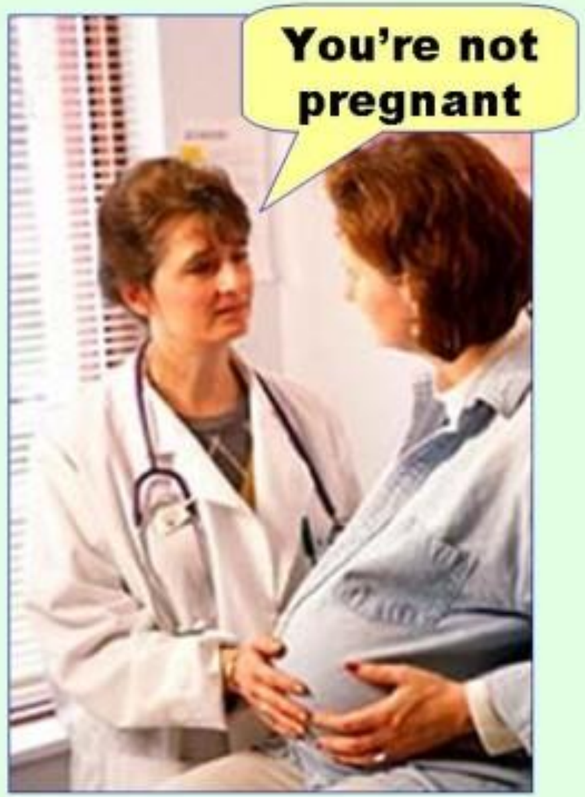
- False Positive greška naziva se još i „Tip 1“, dok se False Negative greška naziva i „Tip 2“

Merenje performansi klasifikatora

Type I error
(false positive)



Type II error
(false negative)



Merenje performansi klasifikatora

- Prikaz rezultata binarne klasifikacije, u cilju analize rezultata predikcije, može se uraditi pomoću ***matrice konfuzije***

		Tačna klasa	
		A	B
Rezultat klasifikatora	A	TP	FP
	B	FN	TN

Merenje performansi klasifikatora

- Tačnost (Accuracy)

$$A = \frac{TP+TN}{TP+TN+FP+FN} = \frac{TP+TN}{N}$$

- Prikladna mera kada su:
  - Klase balansirane
  - Greške podjednako bitne

Merenje performansi klasifikatora

- Tačnost (*Accuracy*)

Predikcija uvek = ¬Fraud

		Tačna klasa	
		Fraud	¬Fraud
Rezultat klasifikatora	Fraud	0	0
	¬Fraud	10	90

$A = 90/100 = 0.90$

- Da li je ovo dobar klasifikator?

Merenje performansi klasifikatora

- Tačnost (*Accuracy*)

Klasifikator 1		Tačna klasa A	Tačna klasa B
Rezultat klasifikatora	A	45	20
	B	5	30

A = 75/100 = 0.75

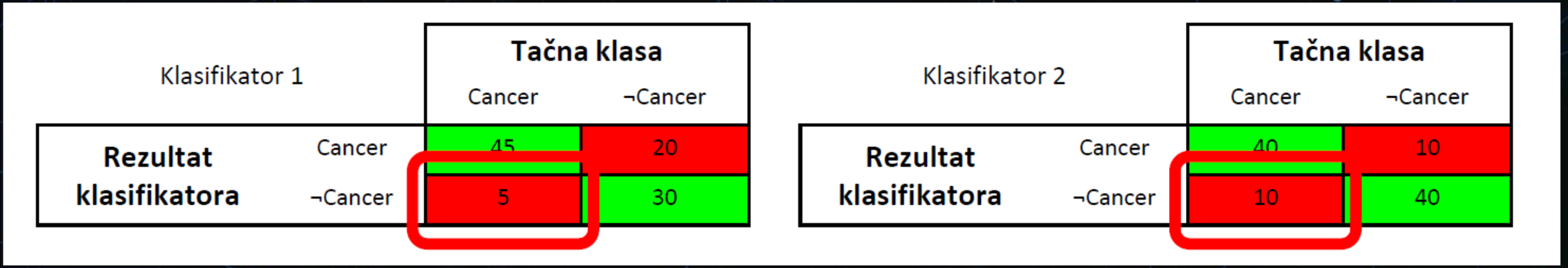
Klasifikator 2		Tačna klasa A	Tačna klasa B
Rezultat klasifikatora	A	40	10
	B	10	40

A = 80/100 = 0.80

- Koji klasifikator je bolji?

Merenje performansi klasifikatora

- Tačnost (Accuracy)



Klasifikator 1

		Tačna klasa: Cancer	Tačna klasa: ¬Cancer
Rezultat klasifikatora	Cancer	45	20
	¬Cancer	5	30

Klasifikator 2

		Tačna klasa: Cancer	Tačna klasa: ¬Cancer
Rezultat klasifikatora	Cancer	40	10
	¬Cancer	10	40

- Koji klasifikator je bolji?

Merenje performansi klasifikatora

- Preciznost (*Precision*)

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

- Odziv (*Recall)*

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

Merenje performansi klasifikatora

Klasifikator 1

		Tačna klasa: Cancer	Tačna klasa: ¬Cancer
Rezultat klasifikatora	Cancer	45	20
	¬Cancer	5	30

$P_1 = 45/65 = 0.69$

$R_1 = 45/50 = 0.9$

Klasifikator 2

		Tačna klasa: Cancer	Tačna klasa: ¬Cancer
Rezultat klasifikatora	Cancer	40	10
	¬Cancer	10	40

$P_2 = 40/50 = 0.8$

$R_2 = 40/50 = 0.8$

Model:“Svako ima rak“

		Class: Cancer	Class: ¬Cancer
Classified	Cancer	50	50
	¬Cancer	0	0

$P = 50/100 = 0.5$

$R = 50/50 = 1$

Merenje performansi klasifikatora

- F-mera (*F-score*)

$$F1 = 2 \cdot \frac{P \cdot R}{P + R}$$

- Predstavlja „spoj“ između preciznosti i odziva, tj. posmatranje obe metrike istovremeno

Merenje performansi klasifikatora

Klasifikator 1

		Tačna klasa: Cancer	Tačna klasa: ¬Cancer
Rezultat klasifikatora	Cancer	45	20
	¬Cancer	5	30

$F_1=2*(0.69*0.9)/(0.69+0.9)$

$= 0.78$

Klasifikator 2

		Tačna klasa: Cancer	Tačna klasa: ¬Cancer
Rezultat klasifikatora	Cancer	40	10
	¬Cancer	10	40

$F_2=2*(0.8*0.8)/(0.8+0.8)$

$=0.8$

Model:“Svako ima rak“

		Class: Cancer	Class: ¬Cancer
Classified	Cancer	50	50
	¬Cancer	0	0

$F=2*(0.5*1)/(0.5+1)=0.66$

Zadatak za vežbu

- Modifikovati prethodna rešenja, tako da se ispisuju i performanse klasifikatora (matrica konfuzije, preciznost, odziv, F-skor).

Rešenje

```
from sklearn.metrics import f1_score, confusion_matrix, precision_score, recall_score
print(confusion_matrix(y_test, y_pred))
print("Preciznost: %0.2f" % precision_score(y_test, y_pred))
print("Odziv: %0.2f" % recall_score(y_test, y_pred))
print("F-score: %0.2f" % f1_score(y_test, y_pred))
```