Mašinsko učenje 2020

Zadatak 4

Sadržaj

Zadatak 3 - Rekapitulacija

Zadatak 3 - Rekapitulacija

- Procenat uspešnosti: **81.82%** (27/33).
- Sva rešenja sa **Accuracy ≥ 0.961** se smatraju odličnim.
- Najveće preklapanje izvornih kodova prema alatu za detekciju plagijata: **25%.**

- Najbolji rezultati po terminima:

Termin	Tim	Accuracy
Ponedeljak	**tim1_20**	**0.99**
Utorak 1	**tim8_20**	**0.968**
Utorak 2	**demo**	**0.97**
Četvrtak	**sljub_duo** i **sljub_trio**	**0.99**
Petak	**masinski_ucenjaci**	**0.99**

Zadatak 3 - Rekapitulacija

- Dobre stvari (na nivou generacije):
  - Pretprocesiranje i vektorizacija;
  - Dodatno istraživanje;
  - Propratni izveštaji.

- Stvari koje mogu biti bolje (na nivou generacije):
  - Rad sa trening skupom podataka.

Zadatak 4

Zadatak 4

- Klasifikacija - Model ansambla:
  - Dostupan je deo policijskih izveštaja o saobraćajnim nesrećama u SAD u periodu 1997 - 2002. Na osnovu dostupnih podataka izvršiti procenu brzine vozila u trenutku sudara (kolona **speed**):
    - 1-9km/h
    - 10-24
    - 25-39
    - 40-54
    - 55+

Zadatak 4

- Klasifikacija - Model ansambla:
  - Zadatak je uspešno urađen ukoliko se na kompletnom testnom skupu podataka dobije mikro f1 mera (eng. *micro f1 score*) veća od 0.51.
  - Zadatak se rešava upotrebom ansambla klasifikatora.
  - Rok: **24.05.2020. u 12:59h.**
  - Trening skup podataka sadrži nedostajuće vrednosti (prazne ćelije).
  - Instalirane biblioteke za Zadatak 4 (verzije date u Uputstvu):
    - **NumPy**
    - **SciPy**
    - **Pandas**
    - **scikit-learn.**

Zadatak 4

- Dostupni atributi (kolone):
  - **weight** - procenjena masa učesnika udesa (sadrži neprecizne procene)
  - **dead** - da li je učesnik preživeo udes:
    - **alive** - preživeo
    - **dead** - nije preživeo
  - **airbag** - da li je učesnik imao airbag:
    - **none** - ne
    - **airbag** - da
  - **seatbelt** - da li je učesnik bio vezan:
    - **none** - ne
    - **belted** - da

Zadatak 4

- Dostupni atributi (kolone):
  - **frontal** - da li je u pitanju bio čeoni sudar:
    - **0** - ne
    - **1** - da
  - **sex** - pol učesnika:
    - **f** - ženski
    - **m** - muški
  - **ageOFocc** - starost učesnika
  - **yearacc** - godina kada se dogodila saobraćajna nesreća
  - **yearVeh** - godina proizvodnje vozila

Zadatak 4

- Dostupni atributi (kolone):
  - **abcat** - da li se aktivirao airbag:
    - **unavail** - vozilo nije imalo airbag za tog učesnika
    - **nodeploy** - airbag se nije aktivirao
    - **deploy** - airbag se aktivirao
  - **occRole** - tip učesnika:
    - **driver** - vozač
    - **pass** - suvozač
  - **deploy** - označava da li se airbag aktivirao:
    - **0** - airbag nije dostupan za tog učesnika ili se nije aktivirao
    - **1** - airbag se aktivirao

Zadatak 4

- Dostupni atributi (kolone):
  - **injSeverity** - stepen povreda učesnika:
    - **0** - bez povreda
    - **1** - lakše telesne povrede
    - **2** - teže telesne povrede, bez invaliditeta
    - **3** - teže telesne povrede, sa invaliditetom
    - **4** - smrt
    - **5** - nepoznato
    - **6** - teške telesne povrede sa smrtnim ishodom (smrt nastupila kasnije).

Zadatak 4

- Trening skup podataka sadrži nedostajuće vrednosti (u pitanju su prazne ćelije).
- Testni skup podataka **ne** sadrži nedostajuće vrednosti.

- Zadatak se **mora** rešiti upotrebom neke od metoda ansambla.
- Metode ansambla u scikit-learn.

Zadatak 4

- Kao meru performansi modela u ovom zadatku imamo mikro f1 meru (eng. *micro f1 score*).
- Ova metrika se, kao i većina metrika klasifikacije, "izvodi" iz matrice konfuzije (eng. *confusion matrix*):

		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

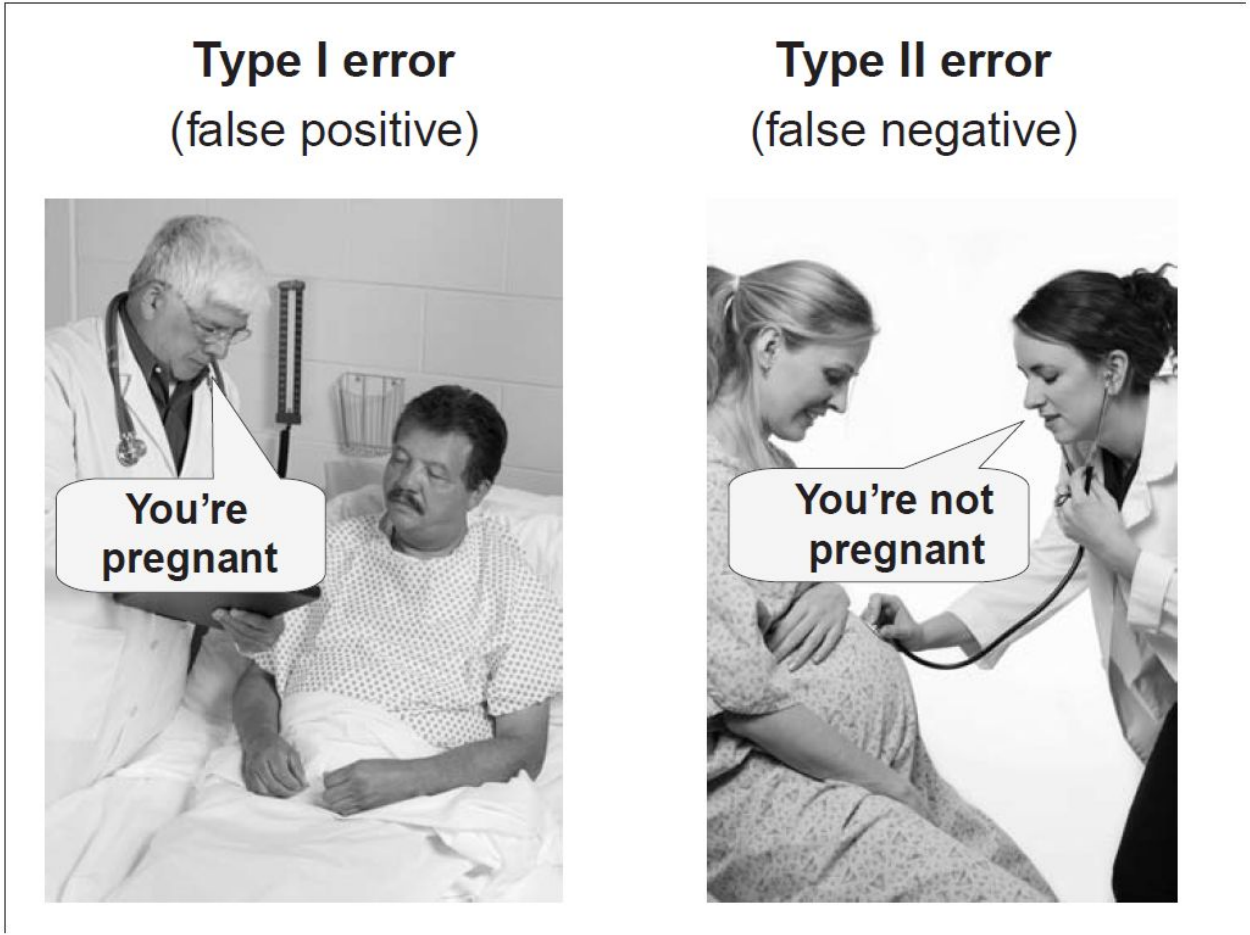

Figure 3.1 Type I and Type II errors

Zadatak 4

- **Precision** - procenat relevantnih (tačnih) među prediktovanim - **P = TP / (TP + FP)**
- **Recall** - procenat relevantnih (tačnih) koje su prediktovane - **R = TP / (TP + FN)**
- **F1 score (**aka ***F- measure*****)** - harmonijska sredina **P** i **R** - **F1 = 2*P*R / (P + R)**
- **Micro F1 score** - računa globalne **TP**, **FN** i **FP**:
  - *sklearn.metrics.f1_score(y_true, y_pred, average='micro')*

- Prilikom treninga, od pomoći može biti i classification_report.