Mašinsko učenje 2023

Zadatak 3

Sadržaj

Zadatak 2 - Rekapitulacija

Zadatak 2 - Rekapitulacija

- Procenat uspešnosti: **55.2%** (16/29).
- Najveće preklapanje izvornih kodova prema alatu za detekciju plagijata: **10%**.
- Najbolji rezultati po terminima:

Termin	Tim	RMSE
Ponedeljak - G4	**kupus**	**16041.39**
Utorak - G3	**tim14_23**	**22553.83**
Četvrtak - G5	**cokici**	**16213.23**
Petak - G2	**tim10_23**	**13704.42**
Petak - G1	**tim11_23**	**16888.31**

Zadatak 2 - Rekapitulacija

- Dobre stvari (na nivou generacije):
  - Vizuelizacija podataka
  - Rad sa outlier-ima
  - Rad sa trening skupom podataka
  - Implementacija algoritama
  - Računanje metrike
  - Propratni izveštaji.

- Stvari koje mogu biti bolje (na nivou generacije):
  - Normalizacija.

Zadatak 3

Zadatak 3

- Klasifikacija:
  - Klasifikovati utiske sa (pokojnog) sajta Donesi.com (kolona **Review**), na osnovu njihovog teksta u dve klase (kolona **Sentiment**):
    - 0 - negativan utisak
    - 1 - pozitivan utisak.
  - Zadatak je uspešno urađen ukoliko se na kompletnom testnom skupu podataka dobije **mikro f1 mera** (eng. *micro f1 score*) veća od 0.80.
  - Zadatak se rešava upotrebom SVM klasifikatora.
  - Rok za izradu zadatka je **25.04.2023. u 23:59h**.

Zadatak 3

- Klasifikacija:
  - Instalirane biblioteke za Zadatak 3:
    - **NumPy**
    - **Pandas**
    - **SciPy**
    - **scikit-learn**.

Zadatak 3

- Sledeći termin vežbi (odbrana Zadatka 3 i predstavljanje Zadatka 4):

Termin	Datum
Ponedeljak - G4	**08.05.2023.**
Utorak - G3	**09.05.2023.**
Četvrtak - G5	**04.05.2023.**
Petak - G2	**05.05.2023.**
Petak - G1	**05.05.2023.**

Zadatak 3

- **scikit-learn** biblioteka:
  - Instalacija
  - Docs.
- Izdvojeno:
  - Selekcija modela
  - Izdvajanje i selekcija obeležja
  - Metrike
  - Support Vector Machines.

Zadatak 3

- Koraci kod klasifikacije teksta:
  - Pretprocesiranje:
    - Transformacija ulaznog teksta:
      - Svođenje teksta na mala ili velika slova
      - Uklanjanje znakova interpunkcije
      - Uklanjanje reči bez značenja (eng. *stopwords*)
      - …
    - **Sav tekstualni ulaz (i trening i test) mora proći kroz isto pretprocesiranje.**
  - Vektorizacija:
    - Konverzija sirovih tekstualnih podataka u numeričke:
      - Bag of Words
      - TF-IDF
      - …
    - **Vektorizator obučen na trening skupu se primenjuje i na trening i na testni skup.**
  - Treniranje i evaluacija klasifikatora.

Zadatak 3

- 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)

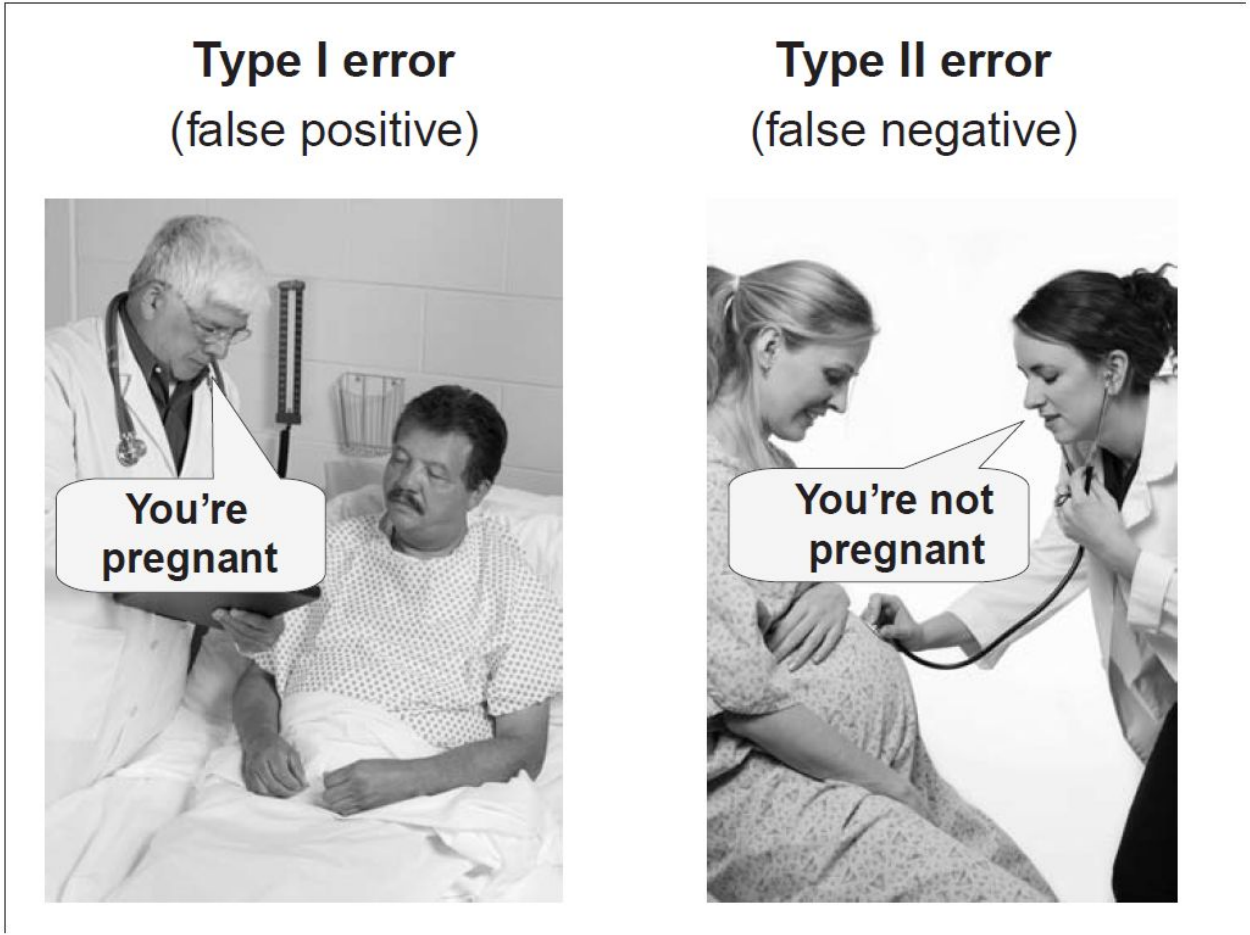

Figure 3.1 Type I and Type II errors

Zadatak 3

- **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 **Precision** i **Recall**:
  - **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*.