Model ansambla

Tim adagrad

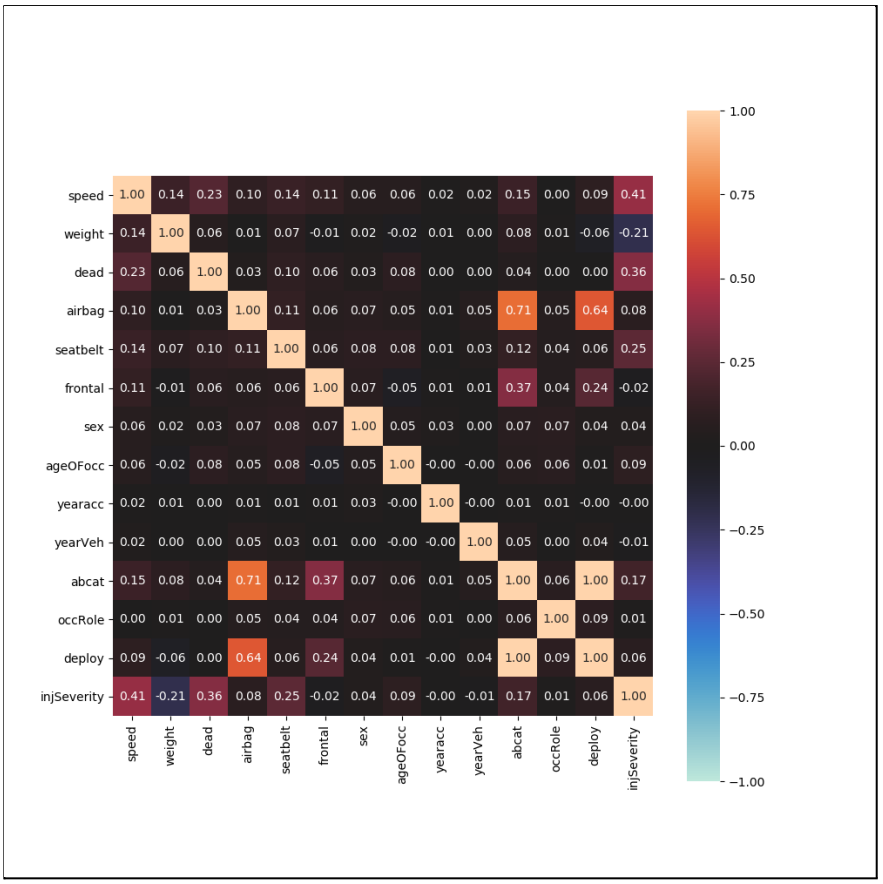
- SW-29/2018, Albert Makan
- SW-48/2018, Anastasija Đurić
- SW-52/2018, Dina Petrov

Zadatak

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). Zadatak je uspešno urađen ukoliko se na kompletnom testnom skupu podataka dobije makro f1 mera (eng. macro f1 score) veća od 0.30. Zadatak se rešava upotrebom ansambla klasifikatora.

Pristup rešavanju problema

Na samom početku iscrtali smo matricu korelacije (*slika 1*) na osnovu koje smo odlučili da iz skupa podataka izbacimo sledeća obeležja: *airbag*, *deploy*, *ageOFoc*, *sex*, *occRole*, *seatbelt*, *injSeverity*. Vrednosti kategoričkih obeležja *abcat* i *dead* zamenili smo inkrementalnim numeričkim vrednostima počevši od 0. Što se tiče rada sa nedostajućim vrednostima u trening skupu, najbolje rezultate ostvarili smo kada smo na mesto nedostajućih vrednosti postavili vrednost 0. Pored toga, pokušali smo da izbacimo date torke i da zamenimo nedostajuće vrednosti sa srednjom vrednošću datog atributa.



Slika 1. Matrica korelacije

Za određivanje parametara modela i verifikaciju rešenja korišćen je *K-fold cross-validation* metod sa podelom skupa u 5 grupa (k=5).

Isprobani algoritmi i ostvareni rezultati

- ### AdaBoost Classifier

  base estimator:
  - ➔ DecisionTreeClassifier(max_depth=15, min_samples_split=4, random_state=0)

  n estimators: 110
  learning rate: 0.7
  average f1 macro: 0.4276926044350208

- ### Extra Trees Classifier

  n estimators: 115
  criterion: 'entropy'
  min samples split: 3
  average f1 macro: 0.4001599821201475

- ### Bagging Classifier

  base estimator:
  - ➔ DecisionTreeClassifier(max_depth=15)

  n estimators: 100
  average f1 macro: 0.3540538232127017

- ### Gradient Boosting Classifier

  n estimators: 100
  max depth: 15
  average f1 macro: 0.4214860563860947

- ### Voting Classifier

  estimators:
  - ➔ KNeighborsClassifier(n_neighbors=3)
  - ➔ LogisticRegression
  - ➔ RandomForestClassifier(n_estimators=100)
  - ➔ DecisionTreeClassifier(max_depth=50)

  voting: 'hard'
  average f1 macro: 0.3652905086138661

Odabrano rešenje

Najbolji rezultat ostvaren je upotrebom **AdaBoost Classifier** modela stoga je on izabran kao konačno rešenje. Na celokupnom test skupu na *malepy* platformi dobijena macro f1 mera iznosi: 0.468865332971677.