ML18-Z6-code10: Izveštaj

Pristup zadatku

Zadatak je urađen u Python-u uz biblioteke Pandas i Scikit-Learn.

Tačnost modela utvrđivali smo korišćenjem unakrsne validacije (*RepeatedStratifiedKFold)*, a parametri su optimizovani korišćenjem grid pretrage (*GridSearchCV*).

Isprobani algoritmi

Za enkodovanje kategoričkih podataka iskoristili smo *label encoding*.

Prilikom pretprocesiranja, podatke smo normalizovali korišćenjem *z-score* normalizacije s obzirom na to da je ona krucijalna za uspešnost redukcije dimenzionalnosti korišćenjem PCA.

Za nedostajuće vrednosti:

1. Dopunjavanje korišćenjem *mean* vrednosti
2. Dopunjavanje korišćenjem *median* vrednosti
3. Izbacivanje uzoraka sa nedostajućim vrednostima

Za redukciju dimenzionalnosti:

1. PCA
2. KernelPCA (rbf kernel)

Za klasifikaciju isprobali smo:

1. GradientBoostingClassifier
2. RandomForestClassifier

Konačno rešenje

Za najbolje rešenje uzorci sa nedostajućim vrednostima su izbačeni (ukupno 56 uzoraka) i korišćen je linearan PCA za redukciju dimenzionalnosti.
Oba isprobana klasifikatora na cross-validaciji davali su iste rezultate tako da smo se odlučili za GradientBoostingClassifier s obzirom da on važi za *state-of-the-art* klasifikator.
GBC i PCA primenili smo korišćenjem sklearn-ovog pipeline-a tako što smo redom primenili z-score normalizaciju, zatim PCA, i na kraju GBC.

Parametri PCA su sledeći:

parametar	Vrednost
svd_solver	full
whiten	True
n_components	6

Parametri GBC su sledeći:

parametar	Vrednost
n_estimators	200
max_depth	2
max_features	0.1
min_samples_leaf	0.066
min_samples_split	0.0001
subsample	0.6
warm_start	True

PCA sa 6 komponenti očuvava ~85% varijanse u podacima. Pokušali smo da sačuvamo ~93% s obzirom na to da se preporučuje čuvanje bar 90-95% varijanse, medjutim time smo dobili za nijansu lošiji rezultat prilikom klasifikacije.

Ostvareni rezultati

Konačan *Micro F1* score na test skupu iznosi 0.82.