Teoretická časť

Na riešenie tohto problému bol použitý Quadratic Discriminant Analysis. Tento klasifikátor vykázal najlepší výsledok spomedzi ostatných použitých klasifikátorov. Klasifikátor s kvadratickou rozhodovacou hranicou, vytvorený fitovaním podmienených hustôt tried na údaje a použitím Bayesovho pravidla. Tento operátor vykonáva kvadratickú diskriminačnú analýzu (QDA). Diskriminačná analýza sa používa na určenie, ktoré premenné rozlišujú medzi dvoma alebo viacerými prirodzenými skupinami. Základnou myšlienkou diskriminačnej analýzy je určiť, či sa skupiny líšia v priemere premennej, a potom použiť túto premennú na predpovedanie príslušnosti k skupine.

Navrh riešenia

Na začiatku bol problém s numpy.load. Na vyriešenie tohto problému bol použitý agrument allow_pickle(True), ktorý povoľuje alebo zakazuje načítanie vybraných objektov zo súboru .npy.

Pri analýze dátumov sme našli reťazcové premenné, ktoré je potrebné nahradiť číslami. Na konverziu týchto údajov bol vyvinutý nástroj LabelEncoder. V tomto prípade bolo 14 tried. Používal sa aj na predspracovanie údajov StandardScaler.

Neskôr sa vyskytol aj problém s NaN, ktorý je tiež potrebné nahradiť. Chcel som použiť SimpleImputer, ale narazil som na problém. SimpleImputer pracuje s dataframe a np.load vytvoril naše súbory ako objekty. Preto som sa rozhodol napísať malý algoritmus. Tento algoritmus nájde priemer v stĺpci a nahradí prázdne dátumy týmto číslom.

Po spracovaní údajov bol použitý Splitter train_test_split, ktorý oddeľuje dáta pre tréning a testovanie. Potom sa použil klasifikátor QuadraticDiscriminantAnalysis. Najprv trénujeme údaje pomocou metódy fit a potom môžeme predikovať nase Testovaci a evaluation údaje.

Vysledky

Najlepší výsledok vykázal klasifikátor QuadraticDiscriminantAnalysis(reg_param = 0.1) = 97%
Na druhom mieste sa umiestnil klasifikátor NuSVC (kernel = "poly", degree = 2, nu = 0,9) = 95%
Na treťom mieste je SVC(kernel="poly", degree=2, C=1) = 94 %

Ďalšie použité klasifikátory, ako napr: RandomForestClassifier, SGDClassifier, DecisionTreeClassifier, LinearDiscriminantAnalysis - vykazovali výsledok nižší ako 90 %.