Zadanie č. 1

Inteligentné systémy v informatike 2021

1 Klasifikátor

Klasifikátor SVC je založený na algoritme Support Vector Machines (SVM).

Maximalizačný problém, ktorý SVM rieši spočíva v nájdení hranice rozdeľujúcej dáta, ktorá spĺňa podmienku maximálnej vzdialenosti (marginu) medzi podpornými vektormi (Support vectors).

2 Preprocessing

2.1 Preprocessing ordinálnych dát

Dataset v zadaní obsahuje číselné dáta (float), v ktorých niektoré údaje chýbajú. Na doplnenie chýbajúcich dát som použil metódu imputácie, v ktorej sa chýbajúce hodnoty nahrádzajú priemernou hodnotou daného stĺpca. Imputáciu som implementoval pomocou funkcie *SimpleImputer()* z knižnice scikit-learn.

Aby klasifikátor nemal problém s naučením sa správnej hranice, je potrebné číselné dáta, ktoré mu poskytneme na naučenie škálovať. Na škálovanie som použil funkciu *StandardScaler()* z knižnice scikit-learn.

2.2 Preprocessing nominálnych dát

Druhá časť datasetu pozostáva z nominálnych hodnôt, konkrétne reťazcov. Kedže klasifikátor na učenie vyžaduje na vstupe číselné hodnoty, je potrebné tieto reťazce transformovať. Na túto úlohu som použil funkciu *OneHotEncoder()* (OHE) z knižnice scikit-learn. OHE pre každý unikátny reťazec vytvorí samostatný stĺpec, v ktorom pre každú vzorku sa nachádza 1 pre daný string, 0 ináč.

3 Výsledky

3.1 Voľba klasifikátora

Ako stratégiu na nájdenie vhodného klasifikátora pre daný dataset som zvolil metódu *Gridsearch*, ktorá spočíva v brute-force skúšaní všetkých kombinácií zadaných parametrov pre zadaný klasifikátor. Túto metódu som implementoval pomocou slučky, v ktorej som volal funkciu *GridSearchCV()* z knižnice scikit-learn v každej iterácii na iný klasifikátor s inou množinou parametrov.

Klasifikátory, ktoré som touto metódou vyskúšal boli: *DecisionTreeClassifier()*, *RandomForestClassifier()*, *LinearSVC()*, *NuSVC()*, *SVC()*.

Víťazom v GridSearchCV() sa stal klasifikátor SVC(), ktorý pre konkrétne parametre *SVC(C=0.0005, coef0=0, kernel='poly', degree='2')* dosiahol presnosť 0.985.

3.2 Voľba parametrov pre klasifikátor

Po natrénovaní SVC modelu s parametrami z GridSearchCV() mi model predikoval dáta v testovacej zložke s chabou priemernou presnosťou 0.5 pri použití metriky roc_auc_score na 100 rôznych rozdeleniach.

Došiel som k záveru, že parametre ktoré som dostal z GridSearchCV silno korelovali s test/train rozdelením datasetu, na ktorom bol GridSearchCV volaný a mohlo dôjsť k pretrénovaniu.

Po zmene parametra *C* model *SVC(C=1, coef0=0, kernel='poly', degree=2)* pri metrike roc_auc_score dosahoval na 100 rôznych rozdeleniach priemernú presnosť 0.976.

Tento model som nakoniec zvolil aj na finálnu predikciu X_eval dát.