Projekt pre predmet Rozpoznávanie obrazcov
Juraj Mašlej

Obsah:

1. Úvod

Rozhodli sme sa pre dataset s údajmi o obyvateľstve USA, nakoľko mal požadované množstvo príznakov a pôsobil zaujímavý na analýzu. Dáta sa skladali z 30 príznakov ako štát, región, množstvo obyvateľov, príjem, rasové zloženie obyvateľstva a podobne.
Predpovedať sme chceli mieru chudoby pre daný región (pozn. región = county).

2. Dataset

Dataset sme získali z kaggle.com. Jedná sa o prieskum obyvateľstva v USA v roku 2015.

2.1. Predspracovanie dát

Načítavanie prebieha numpy funkciou genfromtxt(). Z dát sú 2 príznaky textové (štát, región), tie sme previedli na číselné hodnoty. Následne je celý numpy array prevedený na typ float.

Dáta sa pred rozdelením na trénovací a testovací set náhodne premiešajú. Tento krok je významný pretože dáta sú zoradené podľa polohy (štát a región), teda bez náhodnej permutácie dát sa istá oblasť alebo štát nedostane do tréningových dát.

Pre rozdelenie dát na tréningový a testovací set sme použili pomer 80/20.

Pre formulovanie problému ako klasifikačný sme hodnoty predpovedaného parametera (miera chudoby) rozdelili do 5 kategórií.

3.Redukcia príznakov

Rozhodli sme sa použiť 2 metódy. Pca, Principal component analysis a Lda, Latent Dirichlet allocation. Pôvodný počet príznakov bol 37, pri obidvoch metódach sme najlepšie výsledky dostávali pri redukovaní na 7 príznakov. K tomuto počtu príznakov sme sa dostali postupným odoberaním príznakov, pokiaľ sa výsledky neprestali zlepšovať a nezačali zhoršovať. Lepšie výsledky sme dosahovali pri použití Pca.

4. Klasifikácia

Rozhodli sme sa pre využitie dvoch metód, Svm, Support vector machine a Rfc, Random forest classifier.

4.1. Svm

Testovali sme kernely lineárny, rbf kernel a sigmoid. S rbf a lineárnym kernelom sme dosahovali podobné výsledky, pre sigmoid boli podstatne horšie.

Výraznú prevahu 1. a 2. triedy sme chceli vyriešiť nastavením hyperparametra class_weight, ktorý upravuje váhu chyby v príslušnej triede. Málo zastúpeným triedam sme zvýšili váhu chyby, no to spôsobilo len zvýšenie chyby modelu, nie zameranie sa na správne klasifikovanie aj menej častých tried. Takto upravený model bol najlepší len v malom prípade behov programu, do 10%. Ďalej sme upravovali parameter gamma, no ten na naše výsledky nemal veľký vplyv.

4.2. Náhodný les

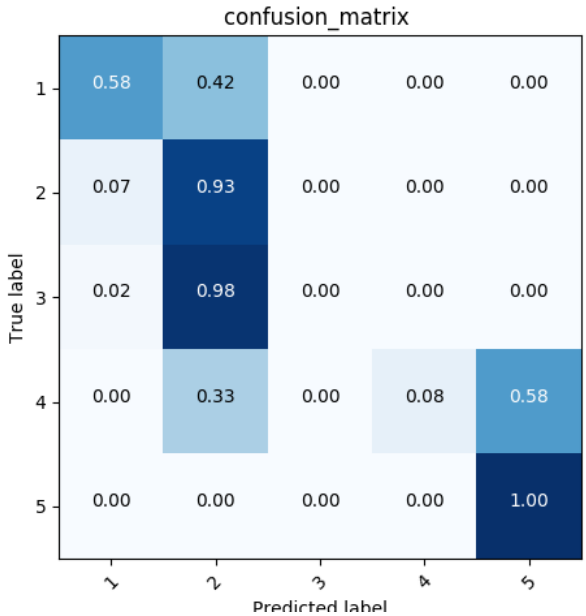
Testovali sme kombináciu parametrov maximálnej hĺbky a počtu stromov v lese. Najlepšie výsledky sme dosahovali s maximálnou hĺbkou 20 a rovnakým počtom stromov.

5. Výsledky

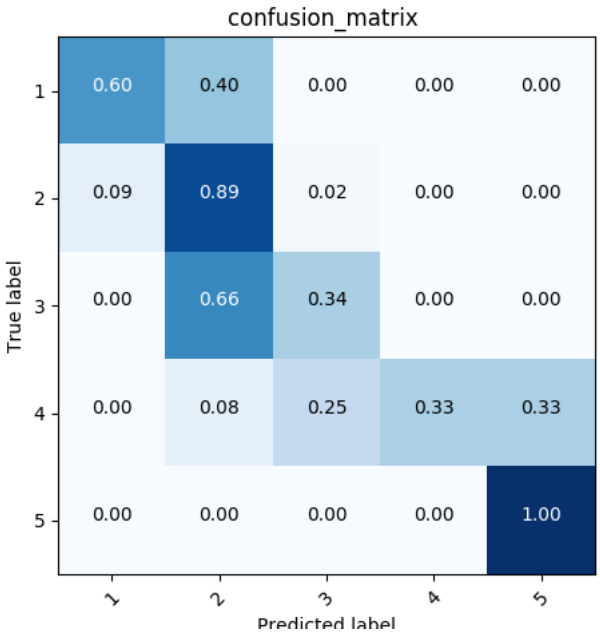
Priemerná presnosť sa pri obidvoch modeloch pri použití pca pohybovala nad 70%. Náhodný les dosahoval lepšie výsledky s hodnotou medzi 72% a 74%, pca medzi 70% a 72%. Pri použití Lda boli presnosti medzi 59% a 61% pri oboch modeloch.

5.1. Matica zámen

Matica zámen pre svm.



Matica zámen pre náhodný les



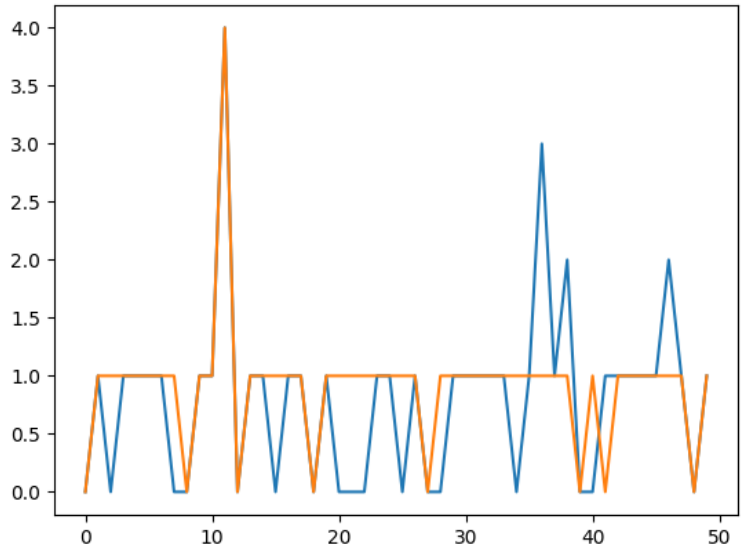
5.3. Klasifikácia do tried

Pre každú z 5 tried je vytvorený graf v ktorom je vyčíslený počet správne priradených príklad do triedy, správne nepriradených do danej triedy, nesprávne priradených príklad do triedy a nesprávne nepriradených príklad do danej triedy.

Grafy prikladáme ako samostatné súbory. Na osi y je zobrazený počet klasikácií, na osi x príslušná kategória.

5.4. Klasifikácia na dátach

Nasledovné grafy zobrazujú klasifikáciu pre prvých 50 dát pre náhodný les. Modrá línia zobrazuje pôvodné dáta, oranžová predpovedané. Ak sa prekrývajú, je predpoveď správna. Z grafu sa dá odčítať, že klasifikátor ma problémy s vyššími triedami.



6. Poznámky ku kódu

Kód je rozdelený do 2 súborov. data_loader.py obsahuje načítanie dát, ich rozdelenie na testovací a trénovací set a následne prevedenie všetkých dát na typ float.

Súbor main.py obsahuje triedu Main, ktorá riadi redukciu príznakov aj samotnú klasifikáciu. Takisto normalizuje dáta a rozdelí hodnoty atribútu „poverty“ do 5 kategórií pre potreby klasifikácie.

Dataset sme získali z adresy https://www.kaggle.com/muonneutrino/us-census-demographic-data/data .

6.1. Výpis do konzoly

Program vypíše rozmery testovacieho a validačného datasetu, priemernú úspešnosť svm, najlepší model svm, priemernú úspešnosť náh. lesu a najlepší model náh. lesu.

7. Záver

Použitie pca zlepšilo výsledky oboch klasifikátorov približne o 10%, a to z 60% na 70%. Nedosiahnutie lepších výsledkov odôvodňujeme pravdepodobne použitím nie najvhodnejších klasifikátorov pre daný problém.