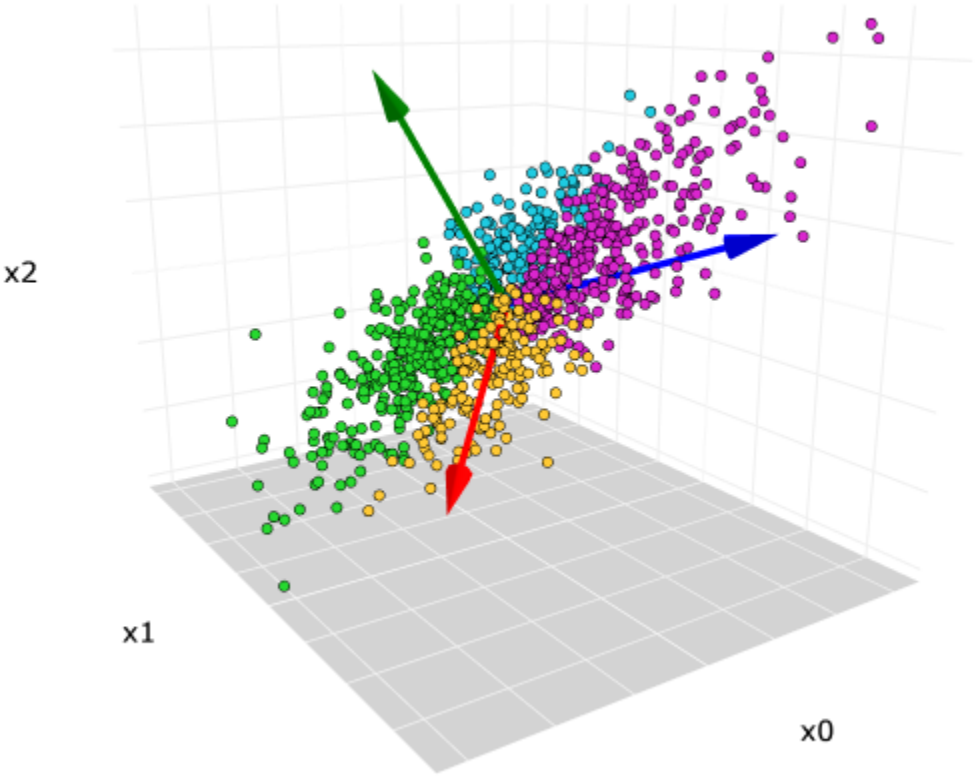
ИЗВЕШТАЈ ЗАДАТКА 6

Редукција димензионалности



Душан Бркић, Филип Живанац

12. липња лета Господњег 2022.

Софтверско инжењерство и информационе технологије

Факултет техничких наука

Универзитет у Новом Саду

ЗАДАТАК

На основу доступних информација о запосленима на источној обали САД, извршити предикцију њихове расе (race): 1.White, 2.Black, 3. Asian, 4. Other. Опис свих атрибута је доступан на пратећој презентацији за овај задатак. Задатак је успешно урађен уколико се на комплетном тестном скупу података добије макро ф1 мера (енг. macro ф1 score) > 0.22. Приликом израде задатка, обавезна је употреба PCA.

ПРИСТУП ПРОБЛЕМУ

У скупу података смо за свако обележје радили мапирање у бројевне вредности осим за обележја брачног статуса и нивоа образовања. За та обележја користили смо One hot encoding методу. Након тога уклонили смо све редове који садрже бар једно празно поље.

ИСПРОБАНЕ МЕТОДЕ

KernelPCA

Редукцију димензионалности први пут смо покушали са KernelPCA. Као најуспешнији кернел овог алгоритма био је косинусни кернел међутим и даље нисмо добијали довољно добре резултате.

PCA

Коришћењем PCA са бројем компоненти 4 постигли смо много боље резулатате редукције димензионалности.

РЕЗУЛТАТИ

AdaBoosting

За решење овог класификационог проблем користили смо AdaBoosting алгоритам. Да би

унапредили решење било је потребно да одаберемо параметре: max_depth, learning_rate, n_estimators и random_state. За learning_rate пробали смо вредности 1, 0.1 , 0.01 , 0.001 где се 0.01 испоставила као најбоља. Такође смо за n_estimators радили са вредностима 20, 50, 100 и 200 где се 50 испоставила као најбоља јер је уједно решење довољно добро и довољно робусно, код вредности 100 и 200 настао је пробле оверфитинга. За орпмизацију параметара max_depth и random_state корисили смо ПСО алгоритам као и у прошлим задацима.

ОДАБРАНО РЕШЕЊЕ

На основу резулатата које нам је дао ПСО најбољи random_state је 3427 а max_depth 15 који дају резултат 0.314664206261795.