Домаћи задатак из Статистичког софтвера 3

Стефан Малбашић

19. децембар, 2022.

1 Напомене и правила

Домаћи задатак представља предиспитну обавезу на курсу Статистички софтвер 3 и на њему је могуће освојити 15 поена. **Рок за предају домаћег је 31. децембар 2022. године у 23:59 часова.** Домаћи задатак се ради самостално.

Преписивање је строго забрањено и лако за детектовати, те ће као такво бити адекватно санкционисано сходно важећим правним актима Математичког факултета и Универзитета. Рјешење које је смислено и ради биће вредновано максималним бројем поена, докле год има прецизност бољу од насумичног погађања. Квалитетна рјешења биће награђена додатним поенима, сходно квалитету.

Формат у којем предајете рјешење није строго одређен, докле год је могуће покренути ваш код. Дакле то може бити Rmarkdown фајл, може бити обични .R фајл, може бити .txt фајл итд. То значи да слика кода не долази у обзир, а све остало је прихватљиво. **Обавезно је да предати кодови буду искоментарисани!** Неискоментарисано рјешење се неће сматрати смисленим.

Рјешења је потребно послати на мејл адресу stefan.malbasic@matf.bg.ac.rs. Након што буду прегледани сви радови, број поена ће бити истакнут на сајту.

2 Задаци

Задатак 1. [3 поена] Учитати базу Carseats из пакета ISLR. Подијелити базу на тренинт и тест скуп у размјери по избору. На тренинг скупу направити линеарни модел гдје је Sales зависна промјенљива, а остале предиктори. Манипулисати предикторима (трансформисати, избацити непотребне) у циљу што бољег модела. Упоредити средњеквадратне грешке на тренинг и на тест скупу. Кратко продискутовати резултате.

Задатак 2. [3.5 поена] Из пакета mlbench учитати базу Glass. Подијелити је на тренинг и на тест скуп. На тренинг скупу направити логистички модел који класификује колону Type на основу осталих. Манипулисати предикторима у циљу што веће прецизности. Провјерити прецизност на тест скупу и упоредити је са оном на тренинг скупу. Приказати матрицу конфузије за класификацију. Кратко продискутовати резултате.

Задатак 3. [4 поена] Из пакета MASS учитати базу biopsy. Спровести анализу главних компоненти у циљу смањења димензионалности података, те извршити предвиђање на основу постављеног модела. Кратко продискутовати резултате.

Задатак 4. [4.5 поена] Поставити у R-у set.seed(737) и генерисати узорке обима $n = 20$ случајних величина $X_1 \sim \mathcal{B}\left(3, \frac{1}{4}\right)$, $X_2 \sim \mathcal{G}\left(\frac{1}{4}\right)$, $X_3 = [T]$ гдје је $T \sim \mathcal{U}[0, 6]$ и $Y = \begin{pmatrix} 1 & 2 \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}$.

(а) Спровести одговарајућу класификацију примјеном k најближих сусједа. Којој класи припадају опсервације $(1, 1, 0)$ и $(4, 8, 5)$? За које k модел даје најбоље резултате?

(б) Поставити set.seed(833) и покренути добијање горе дефинисаних промјенљивих само за $n = 200$. Спровести одговарајућу класификацију примјеном LDA и QDA метода.