

Семинарски рад из Статистичког софтвера 3

Данијел Алексић, Лука Перовић

15. децембар, 2024.

2021. година, фебруар месец. Спољнополитички кореспондент Политике за војно-безбедносна питања и амбасадор Републике Србије у Руској Федерацији, Његова екселенција Мирослав Лазански, у званичној је посети опитно-техничком центру Капустин Јр. Заједно са својим дугогодишњим пријатељем, министром одбране Руске Федерације, генералом армије Сергејем Шојгуом, као и осталим представницима Генералштаба Оружаних снага Руске Федерације, присуствовао је пробним гађањима последње генерације руског балистичког арсенала.

Од стране шефа катедре за системе наоружања Машинског факултета Универзитета у Београду проф. др Момчила Милиновића, а потенцијално и одређених безбедносних структура, замољен је да у истраживачко-информативне сврхе прикупи тактичко-техничке податке о кинеским балситичким системима наоружања, а за које се сматра да се могу обезбедити у таквим условима.

У неформалном разговору са другом министром, уз пиршке и чашицу вотке у ресторану "Главпир-торг", Мирослав је покушао да обезбеди жељене податке. Увеселини друг генерал обавестио је његову екселенцију да информације наводно није могуће обезбедити с обзиром на активно присуство средстава присмотре Федералне службе безбедности. Генерал је предложио алтернативни начин прикупљања података.

Након непроспаване ноћи проведене са пуковником ракетно-космичких снага Даријом Н, у комуналки поред парка Сокољники, Мирослав је успео да обезбеди детаљне податке о броју и капацитетима кинеских балистичких ракета у оперативној употреби.

Прикупљени подаци били су достављени на микрофилму сакривеном у корицама књиге "Пять смертей академика Королёва", коју је понео на дар професору Милиновићу. Непажњом запослених Службе специјалних веза и информација Руске Федерације (Спецсвязь России), са друге стране микрофилма нашли су се детаљни подаци о руским балистичким ракетама у оперативној употреби.

Због познатог дејства сиденафил-цитрата на микрофилм, а који су случајно дошли у контакт у путној торби Његове екселенције, подаци су у Београд стигли са малим оштећењима. Такође, микрофилм није погодан начин записа података у сврху рачунарске обраде.

Шодно томе, професор Милиновић је ручно прекуцао податке на рачунару *Commodore 64*, те их предао својим докторандима на даљу обраду. Они су те податке превели у формат погодан за обраду на савременим рачунарима.

Прераном смрћу Његове екселенције, обрада ових података постала је од секундарног интереса господину професору, те је одлучено да се за њихову обраду ангажују стручњаци за статистику са Математичког факултета Универзитета у Београду. Због редукованих средстава за научноистраживачки рад, одлучено је да се безбедносно мање осетљив део обраде тих података пренесе на студенте. Како је техника израде ракета постала униформна широм планете, одлучено је да се и руске и кинеске ракете анализирају заједно.

Подаци су доступни у фајлу `rakete.ods`.

СЕМИНАРСКИ РАД - ЗАДАЦИ

ЗАДАТАК 1 [5%]. Пронаћи и инсталирати одговарајући пакет који служи за учитавање датотека у .ods формату, те учитати податке као data.frame.

ЗАДАТАК 2 [15%]. Недостајања у подацима нису великих пропорција, али ипак постоје. Пронаћи и на што смисленији начин импутирати недостајућа поља. Брисање опсервација носиће 0%, а импутација средњом вредношћу и сл. 1%. Користити потенцијалне везе недостајућег поља са доступним пољем. Може и "од ока", ако је смислено и има оправдање. [10%]

Укратко објаснити који начин импутације је коришћен, како и зашто баш тај.[5%]

Напомена. Иако недостајања има јако мало, те би овде сваки приступ био довољно добар, на студенту је да се "претвара" да су недостајања озбиљна, те да озбиљно приступи теми.

ЗАДАТАК 3 [20%]. Направити модел који процењује максималан домет ракете на основу осталих предиктора. У обзир долази линеарна регресија, као и други видови процене нумеричке колоне који могу, али и не морају бити рађени на курсу. Ако студент буде користио ове друге, и заиста добије бољи модел неголи са првима, може очекивати додатне поене.

Подешавати модел на начине по избору тако да буде што бољи. Модел се сматра бољим ако мање греши у процени.

ЗАДАТАК 4 [30%]. Направити класификатор по избору који погађа да ли ракета има фиксан или покретан лансер. Испитати да ли је дошло до преприлагођавања и модификовати модел по потреби.

Напомена. Припазити при дељењу на тренинг и тест, јако је мало непокретних ракета. Уколико се студент потруди и научи неку технику за рад са небалансираним подацима (није рађено на курсу), може очекивати да буде награђен додатним поенима.

ЗАДАТАК 5 [20%]. Из базе избацити све ракете средњег домета. Након тога, избацити предикторе max_domest_km и klasa_po_dometu. Дакле, у овом задатку има се сматрати да су домет ракете и његова класа непознати, а да је познато да нема ракета средњег домета.

Неком методом невођеног учења извршити кластеризацију ракета по домету: кратки и дуги. Јасно, ништа неће гарантовати да је то подела баш по домету, јер домет формално немамо. Образложити зашто можемо закључити да јесте (не нужно математичко образложење). Не сме се користити домет ракете ни за тренинг ни за проверу, он је непознат! Као и раније, исказање из оквира курса носи додатне поене.

ЗАДАТАК 6 [10%]. Сваки озбиљан аналитичар података труди се да из њих извуче максимум. Стога, он се не сме зауставити само на њиховој математичкој анализи, већ, када год је то могуће, треба да се труди и да разуме појмове који се у њима појављују.

У бази нема много руских ракета средњег домета. Такође, оне су или су млађе од 2018. (нпр. Kinzhal), или су старије од 1988. (нпр. Zub). Код кинеских ракета таква се правилност не уочава. Објаснити зашто.

Хинт 1. Реаган-Горбачёв, INF.

Хинт 2. Болтон-Шойгу, разлог за састанак у октобру 2018.

СЕМИНАРСКИ РАД - НАПОМЕНЕ

НАПОМЕНА 1. Рок за израду је крај школске године. У зависности од рока у којем желите да браните семинарски рад, накнадно ћемо се договарати о термину када пошаљете рад. Радове слати на мејл luka.perovic@matf.bg.ac.rs

НАПОМЕНА 2. Није довољно послати само код, већ је неопходно детаљно исписати шта је урађено. Опције су: (1) RMarkdown документ, у коме би били интегрисани текст и код заједно; (2) LaTeX документ, у коме би било описано шта је рађено, а искоментарисан код да се пошаље посебно као обичан .R фајл; (3) Исто као (2), само је уместо LaTeX-а коришћен Word или нешто слично. Правите се да пишете извештај послодавцу. Дакле, нек буде шарено и лепо, нико не воли да чита досадан текст.

НАПОМЕНА 3. Ликови из уводне приче су прави, догађаји су апсолутно измишљени и не одговарају стварности. Не постоје никакве скривене инсинуације нити сублиминалне поруке. У питању је само омаж легенди српског и свецког новинарства. Подаци су прави, ручно су скупљани по сајтовима.

НАПОМЕНА 4. Дозвољено је коришћење свих средстава, осим преписивања од колега. Дакле гуглање, књиге... Дозвољено је питати и колегу са курса, уколико сте у стању да испродукујете два различита решења.

НАПОМЕНА 5. Дозвољено је писање семинарског рада и на енглеском.

НАПОМЕНА 6. Путем мејла можете постављати питања у вези са задацима. Уколико прогуглам ваше питање и нађем одговор међу првим резултатима, нећете добити одговор. У свим другим ситуацијама бићу више него рад да помогнем.