Системи одлучивања у медицини (13E053COM)

Пројектни задатак 2023/2024

Пројекат се може радити индивидуално или у пару, може се бранити у било ком испитном року и носи 30% укупне оцене. Група (или појединац) се пријављује за неку од база наведених у табели у оквиру тима на MS Teams платформи. Може се бирати само између база које нису већ изабране од стране неке друге групе. Базу је потребно изабрати најкасније до почетка јунског испитног рока (5. јун), у супротном се сматра да студенти не ради пројекат.

Финални извештај треба припремити тако да садржи све релевантне информације о самој бази података, као и резултате и коментаре појединачних тачака анализе.

1. **[4]** Извршити анализу скупа података. Ако има недостајућих/неадекватних вредности заменити их очекиваним вредностима или их уклонити уз одговарајуће образложење. Уколико постоје номинални/категорички атрибути извршити адекватно кодирање и превођење у нумеричке атрибуте уз одговарајуће образложење. За класе које имају испод 3-5% (одабрати границу) укупне популације сматрати да су мале и формирати класу „Остало“ у оквиру које ће се налазити све класе бројности испод 3-5%.

2. **[5]** Одредити информациону добит за свако обележје и коментарисати резултате. На основу резултата изабрати 10 најкориснијих обележја (уколико база има мање обележја, изабрати сва) и испитати корелацију између обележја, образложити која метода је коришћена. Приказати расподелу коефицијента корелације за изабрана обележја. Приказати график расподеле одбирака по класама за два обележја која имају највећу корелацију са класом.

3. **[5]** Применом LDA методе за редукцију димензија над целим скупом обележја, испитати који је минималан број димензија на који можемо редуковати оригинални скуп тако да индекс информативности (проценат варијансе) буде већи од 80%. Коментарисати резултате.

4. **[5]** На подацима након редукције димензија применити:

- у случају две класе: параметарски класификатор по избору у складу са сепарабилношћу података;
- у случају више класа: тест више хипотеза.

Уколико редукција димензија из тачке 3 не даје задовољавајућу сепарабилност класа, размотрити могућност смањења броја обележја на основу њихове информативности, па пројектовати адекватан класификатор на бази тестирања хипотеза над таквим вишедимензионим подацима.

Резултате класификације приказати у виду конфузионе матрице.

5. **[5]** Изабрати једну од метода непараметарске класификације и над целим скупом података истренирати изабрани класификатор. Методом кросвалидације пронаћи оптималне вредности параметара за изабрани класификатор, и то:

- у случају KNN класификатора оптималан број суседа
- у случају стабла одлучивања максималну дубину стабла

Приложити график зависности тачности класификације од вредности параметара који се оптимизује као и конфузиону матрицу као резултат тестирања за оптималну вредност параметра.

6. **[6]** Над целим скупом података извршити обучавање и тестирање различитих структура неуралне мреже: са једним и више скривених слојева, као и са различитим бројем неурона у слојевима. Илустровати пример доброг и лошег (премало или превише) броја неурона. Правилан избор вредности ће се бодовати.

За превелик број неурона показати могућности заштите од преобучавања: а) раним заустављањем, б) регуларизацијом. За свако обучавање мреже приложити график перформансе као и конфузиону матрицу као резултат тестирања.

Списак база података

За сваки скуп података је дат .csv фајл у коме се налазе подаци. На датом линку се налази опис базе података као и потенцијално интересантне референце ка радовима. Ови радови могу да буду корисни са стране рангирања атрибута и додатних информација о бази.

1. Arrhythmia
   https://archive.ics.uci.edu/ml/datasets/Arrhythmia
2. Autistic Spectrum Disorder (Children)
   https://www.kaggle.com/datasets/fabdelja/autism-screening-for-toddlers
3. Bone Marrow Transplant
   https://archive.ics.uci.edu/ml/datasets/Bone+marrow+transplant%3A+children
4. Breast Cancer
   https://archive.ics.uci.edu/ml/datasets/Breast+Cancer
5. Breast Cancer Wisconsin
   https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29
6. Mammographic
   https://archive.ics.uci.edu/ml/datasets/Mammographic+Mass
7. Stroke Prediction
   https://www.kaggle.com/datasets/shashwatwork/cerebral-stroke-predictionimbalaced-dataset
8. Heart Disease
   https://archive.ics.uci.edu/ml/datasets/Heart+Disease
9. Heart Failure
   https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records
10. Statlog Heart
    https://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29
11. Hepatitis C
    https://www.kaggle.com/datasets/amritpal333/hepatitis-c-virus-blood-biomarkers
12. Contraceptive Method Choice
    https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice
13. Dermatology
    https://archive.ics.uci.edu/ml/datasets/Dermatology
14. Diabetic Retinopathy Debrecen
    https://archive.ics.uci.edu/ml/datasets/Diabetic+Retinopathy+Debrecen+Data+Set
15. Fertility
    https://archive.ics.uci.edu/ml/datasets/Fertility
16. HCV data
    https://archive.ics.uci.edu/ml/datasets/HCV+data)
17. Thoracic Surgery
    https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data

18. Estimation of obesity
    https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+
19. Chronic Kidney Disease
    https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease
20. COVID presence
    https://www.kaggle.com/datasets/hemanthhari/symptoms-and-covid-presence
21. Heart Disease Prediction
    https://www.kaggle.com/code/andls555/heart-disease-prediction
22. Cardiovascular Disease
    https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset
23. Adult Autism
    https://www.kaggle.com/datasets/faizunnabi/autism-screening
24. Cardiotocography (izlaz kolona NSP)
    https://archive.ics.uci.edu/ml/datasets/Cardiotocography
25. Cardiotocography (izlaz kolona CLASS)
    https://archive.ics.uci.edu/ml/datasets/Cardiotocography
26. Cervical cancer (izlaz kolona Hinselmann)
    https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29
27. Cervical cancer (izlaz kolona Schiller)
    https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29
28. Cervical cancer (izlaz kolona Citology)
    https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29
29. Cervical cancer (izlaz kolona Biopsy)
    https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29
30. Hepatitis
    https://archive.ics.uci.edu/ml/datasets/Hepatitis

Напомена: У случају тема 24-25 и 26-29 постоји више могућности избора излаза (назив за сваку тему је дефинисан у загради). Приликом класификације потребно је одбацити остале могућности излаза из скупа података.