

Изпит – МСМО 2020

Част 1: Проект

Даден е csv файл (email_spam_classification.csv) с извадка данни, представяща в табличен вид броя срещания на предварително указани думи на английски език (по колоните) в конкретни имейл съобщения - всеки ред съответства на отделен и независим от другите обработен текст от имейл. Използвайки колоните от 1 до n-1, тренирайте класификационни алгоритми, с които да предвидите стойността в последната колона n (explained variable), която съдържа единствено стойности {1, 0}, указващи респективно дали съобщението е спам или не.

Очаква се да изпратите до 05/07/2020, 23:59 Jupyter notebook с решението. Вашите ноутбук и решение следва да съдържат:

- подходящо зареждане на данните и предварителна обработка (preprocessing), ако има нужда
- поне два различни класификационни модела/алгоритъма, различни от невронна мрежа
- бонус: невронна мрежа, в допълнение към останалите алгоритми
- подходящи мерки (метрики) за производителност/качество на бинарната класификация
- кръстосано валидиране (k-Fold cross validation, $k > 5$), на базата на което да се изчисли производителността
- коментари в кода
- базова визуализация
- сравнение резултатите на различните алгоритми

* Материалът от упражненията в ноутбуките е напълно достатъчен за успешното изпълнение на задачата

** Проектът е индивидуален

*** Уверете се, че имате инсталирани и функциониращи Anaconda, Jupyter, както и Python модулите, които ще използвате (тези разгледани в лекциите и упражненията са напълно достатъчни).

Част 2: Устно изпитване (опционално) - теоретични въпроси

Студентите, които желаят да повишат оценката си и/или смятат, че не са се справили добре с проекта ще имат възможност за устно изпитване в рамките на 15-20 минути в Дискорд - индивидуално определени начало и край за всеки студент, по азбучен ред. Ще получите допълнителна информация в края на Юни/Началото на Юли.

Въпросите ще включват:

- теоретични въпроси върху материала от всички лекции, включително за свойства и начина на функциониране на конкретни методи и модели за машинно обучение
- специфични въпроси върху имплементациите на методите и алгоритмите от упражненията/ноутбуците