Технически Университет - София

Курсов проект

МСМО/ИУЗ – Януари, 2025

Част 1: Задание на проекта

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

Очаква се да изпратите до 02/02/2025, 23:59 Jupyter notebook с решението. Вашите ноутбук и решение се предполага, че ще съдържат:

- подходящо зареждане на данните и предварителна обработка (pre-processing), ако има нужда

- поне два различни класификационни модела/алгоритъма, различни от невронна мрежа

- бонус: нервонна 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

- релевантна визуализация

- сравнение на резултатите на различните алгоритми, коментар и заключение на базата на сравненията

* Материалът от упражненията в ноутбуците 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, по желание) - 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 10-15 минути в Дискорд - индивидуално определени начало и край за всеки студент пожелал такова (напр. по азбучен ред). Ще получите допълнителна информация в края на януари.

Въпросите ще включват:

- теоретични въпроси върху материала от всички лекции, включително за свойствата и начина на функциониране на конкретни методи и модели за машинно обучение

- специфични въпроси върху имплементациите на методите и алгоритмите от упражненията/ ноутбуците