Лабораториска Вежба 1

Класификација со Naïve Bayes и Gaussian Discriminant Analysis модели (LDA и QDA)

- Целта на оваа лабораториска вежба е да се одбере некое податочно множество и да се запознаете со истото користејќи разни визуелизации и после тоа да направите класификација користејќи ги методите кои ги имаме учено: Naïve Bayes, Linear Discriminant Analysis (LDA) и Quadratic Discriminant Analysis (QDA).

- Во прилог ви испраќам и пример за што отприлика треба да направите. Во склоп на овој пример како што ќе видите, има упатство за како да читате од податочното множество, како да направите некои визуелизации (histograms, scatterplots, feature correlations) со цел да се запознаете со множеството и финално како да истренирате некој класификационен модел и да ја евалуирате точноста на истиот. Рокот на оваа лабораториска е 2 недели после поставување на истата (**17.11.2019**). Доколку го пропуштите овој рок, за да ги добиете сите поени има дополнително барање кое треба да го направите.

Што точно треба да направите?

1. **Избирање на податочно множество**. Во прилог ви испраќам линкови до некои податочни множества кои можете да ги искористите. Во склоп на линковите имате податоци за податочните множества за што означува секој атрибут (feature) и што означува класата. Во некои од потешките податочни множества нема "стандардна" класа која треба да ја одредите, па имате можност да експериментирате со што тоа класифицирате. Исто така, во некои од потешките податочни множества има колони кои се состојат од текстуални податоци кои не би можеле да ги користите така да треба да ги отстраните (или во самиот .csv фајл или во кодот). Воглавно за почетници ги препорачувам некои од првите 4 податочни множества, па понатаму можете да пробувате и на некое од останатите.
    a. https://www.kaggle.com/ronitf/heart-disease-uci/download - бинарна класификација на срцева болест. Податоци: 303, Атрибути: 13
    b. https://archive.ics.uci.edu/ml/datasets/Abalone - повеќекласна класификација на старост на Abalone (вид на гастропод). Податоци: 4177, Атрибути: 8
    c. https://archive.ics.uci.edu/ml/datasets/Glass+Identification – повеќекласна класификација на вид на стакло. Податоци: 214, Атрибути: 10

d. https://archive.ics.uci.edu/ml/datasets/MAGIC+Gamma+Telescope – бинарна класификација на гама честички во однос на шум. Податоци: 19020, Атрибути: 10
e. https://www.kaggle.com/abcsds/pokemon#Pokemon.csv – Во ова податочно множество нема стандардна класа, така да може да тестирате повеќе работи како одредување тип на покемон, одредување дали е легендарен, одредување на генерација итн. Податоци: 721, Атрибути: 13
f. https://www.kaggle.com/mylesoneill/game-of-thrones#character-predictions.csv – Тука би го користеле третото множество (characters-predictions.csv) каде што целта е да предвидите дали некој карактер од серијата Game of Thrones е жив или не. Има дел текстуални атрибути кои би требало да ги тргнете пред класификацијата и дел од атрибутите (колоните 2-5) треба да се тргнат поради тоа што претставуваат предикции на луѓето што го направиле податочното множество. Податоци: 1946, Атрибути: 32 (не сите се употребливи).
g. Доколку сакате да искористите некое различно множество може да најдете на https://www.kaggle.com/datasets или на https://archive.ics.uci.edu/ml/datasets.php

2. **Запознавање со податочното множество**. Во овој дел би требало да направите дел од визуелизациите (не мора сите работи што се правени во примерот, и не мора да правите визуелизации на сите атрибути туку тие што ви изгледаат побитни) на вашето податочно множество за да имате некоја претстава за атрибутите во однос на класите, колку се меѓусебно поврзани. (За тие кои не се запознаени со PCA, немора да го правите тој дел). Можете слободно да направите и други визуелизации кои би ги нашле на страниците на seaborn и matplotlib.

3. **Класификација**. Истренирајте три класификатори користејќи ја библиотеката Scikit-Learn (Naïve Bayes, LDA и QDA) на податочното множество што сте го избрале. Како што е направено во примерот, прво поделете го оригиналното податочно множество на тренинг и тест множество и истренете го класификаторот на тренинг множеството. После пресметајте ја точноста (accuracy) на тест множеството за секој класификатор. Пробајте да заклучите зошто еден метод е подобар од друг во однос на претпоставките кој ги користат методите.

4. **Документација:**

a. Доколку користевте jupyter, на курсот поставете .ipynb или .html фајл од истиот.
b. Доколку користевте стандардна работна околина, на курсот поставете .zip кој го содржи .py кодот и screenshots од резултатите и коментарите.

5. ***Дополнителна работа за поминат рок од 2 недели.*** Направете намалување на димензионалноста користејќи LDA dimensionality reduction на вашето податочно множество и одново истренирајте три класификатори. Која е промената во точност?