Raport proiect

Problema pe care am primit-o este una de clasificare(binara) de tip SVM. Lucrarea de unde am obtinut setul de date si-a propus sa observe comportamentul (daca sunt profesionisti sau nu) unor bloggeri bazat pe tendintele lor pe intenet, intersele poitico-sociale si nivelul lor academic. Astfel, pe baza unor chestionare completate de 100 de utilizatori s-a obtinut un set de date care continue 6 coloane : Degree, caprice, topic, lmt, lpss, pb si 100 de linii. Coloana ‘pb’ desemneaza etichetele noastre (yes sau no), adica daca bloggerul in cauza sunt profesionisti sau nu.

Transformarile efectuate pe setul de date pentru a obtine o matrice numerica pe care aplicam algoritmul ML sunt:

* Pentru coloanele ‘pb’, ‘lpss’ si ‘lmt’ sunt doua posibilitati yes sau no, care au devenit 1 si 0
* Pentru coloana ‘topic’ : impression🡺0 , political🡺1, tourism🡺2, news🡺3, scientific🡺4
* Pentru coloana ‘caprice’: left🡺0, middle🡺1, right🡺2
* Pentru coloana ‘Degree’: low🡺0, medium🡺1, high🡺2

Am atasat doua poze pentru a se observa trecerea de la date de tip string la numeric:

(setul de date transformat complet este atasat in e-mail)

Table

Description automatically generated Table

Description automatically generated

Matricea obtinuta este [100x6]

Pentru antrenarea sistemului am ales 75% din date, mai exact primele 75 de linii si primele 5 coloane pentru matricea de date si ultima coloana pentru cele 75 de linii pentru etichete.

Iar 25% din date le-am folosit pentru testare: cele 25 de linii si primele 5 coloane ramase dupa primul slicing pentru matricea de test si ultima coloana cu cele 25 de linii pentru vectorul de etichete.

(In spyder matricea din excel este inversata(in oglinda), de aceea ‘pb’ este ultima coloana)

Pentru aplicarea algoritmului SVM am utilizat biblioteca sklearn. Tot din aceasta am important metrics(desi codul este comentat) pentru a face o comparatie intre acuratetea calculata manual si cea calculata de functia metrics.accuracy\_score(). De asemenea cu ajutorul bibliotecii pandas am citit documentul excel.

Parametrul variat este costul, care ia valorile 1/32, 1/8,1/2, 2, 8, 32 si 128. Iar pentru verificarea acuratetii am numarat cate valori de test din vectorul etichetelor coincide cu cele din din vectorul obtinut dupa predictie.