Test de laborator la Inteligență Artificială, seria 35
- 1 februarie 2021 -

1. Problema de învățare automată: clasificarea de litere mici și mari (4.5 puncte)

Recunoașterea textului scris de mână (*handwritten text recognition*) este o problemă fundamentală în inteligența artificială. O componentă importantă a acestei probleme este recunoașterea literelor. În cele ce urmează considerăm problema discriminării între anumite clase de litere mici și mari. Codul Python de la care porniți încarcă literele *a* ("a mic") și *A* ("A mare") din fișierele text corespunzătoare și plotează primele cinci exemple de litere scrise de mână din fiecare clasă. Fiecare fișier conține 100 de exemple ale fiecărei litere, de mărime 28 × 28 pixeli, fiecare pixel cu valori între 0 (negru) și 255 (alb).

(a) (b)

Figura 1. (a) *Cinci exemple de litere din clasele a și A; (b) exemple prototip pentru clasele a și A*

Realizați următoarele:

a. Împărțiți cele 100 de litere din fiecare clasă (a și A) astfel: primele 80 de litere din fiecare clasă intră în mulțimea de antrenare, următoarele 20 intră în mulțimea de testare. Veți avea astfel o mulțime de antrenare cu 160 de exemple de litere din clasele a și A și o mulțime de testare cu 40 de exemple din cele două clase. (**0.5 puncte**)

b. Calculați și plotați pentru fiecare clasă (a și A) ***exemplul prototip*** al clasei de litere pe baza mulțimii de antrenare făcând media tututor literelor din acea clasă care se regăsesc în mulțimea de antrenare. Astfel, pentru a calcula exemplul protitip al clasei A trebuie să faceți media celor 80 de litere din mulțimea de antrenare a clasei A. Exemplele prototip obținute ar trebuie să semene cu cele din Figura 1b. (**1 punct**)

c. Clasificați cele 40 de exemple din mulțimea de testare folosind clasificatorul cel mai apropiat vecin (1-NearestNeighbor) folosind cele două exemple prototip ale claselor a și A calculate la punctul anterior. Practic, pentru fiecare exemplu de testare, decideți clasa acestuia pe baza distanței (Euclidiană sau Manhattan) dintre exemplul curent și cele două exemple prototip, transferând eticheta celui mai apropiat exemplu prototip. Afișați matricea de confuzie 2 × 2 și acuratețea clasificatorului vostru. (**1 punct**)

d. Modificăm problema inițială adăugînd clasele inițiale de litere c (c mic) și C (C mare). Scopul acum este de a discrimina între două clase noi: litere mici (a sau c) și litere mari (A sau C). Încarcați datele pentru clasele inițiale c și C din fișierele corespunzătoare ("c_small.txt" și "C_big.txt"). Împărțiți cele 400 de litere din fiecare clasă inițială (a, c, A, C) astfel: primele 80 de litere din fiecare clasă inițală intră în mulțimea de antrenare, următoarele 20 intră în mulțimea de testare. Veți avea astfel o mulțime de antrenare cu 320 de exemple de litere mici și mari, din care 160 vor fi litere mici (a sau c) și 160 vor fi litere mari (A sau C). Mulțimea de testare va avea 80 de exemple de testare, 40 de exemple vor fi litere mici (a sau c) și 40 de exemple vor fi litere mari (A sau C). (**0.5 puncte**)

e. Antrenați un clasificator SVM liniar pe mulțimea de antrenare și calculați acuratețea lui pe mulțimea de testare. Afișati matricea de confuzie 2 × 2 (aveți două clase, litere mici vs litere mari) și acuratețea lui. Comparați performanța obținută cu cea de la punctul c. (**1.5 puncte**)