NEAREST NEIGHBOUR

TRAISTARU VLAD-VIOREL
MOISE ANDREI
442A

Cuprins

1. State of the Art

Invatarea automata [1] (in engleza, “Machine Learning”) este subdomeniu al informaticii si o ramura a inteligentei artificiale, al carui obiectiv este de a dezvolta tehnici care permit calculatoarelor posibilitatea de a invata. Mai precis, se urmareste sa creeze programe capabile de generalizare a comportamentului pe baza informatiilor furnizate in formularul de exemple.

Clasificatorul 1-NN [2] poate fi conceput ca o generalizare a clasificatorului de distanta minima, in care fiecare clasa este reprezentata de toate esantioanele preclasificate disponibile, considerati vectori prototip ai clasei.

2. Principiu

Algoritmul traditional al celui mai apropiat vecin (Nearest Neighbour) este un clasificator neparametric, care aloca vectorul de intrare de clasificat X, acelei clase care corespunde celui mai apropiat vecin al lui X din lotul vectorilor de referinta (etichetati).

3. Algoritm

Fie setul de N vectori de referinta etichetati:

$$\chi = \{X_1 \ldots X_N\}$$

cu etichetele de apartenenta la una din cele M clase corespunzatoare:

$$\Omega(\chi) = \{\omega(X_1) \ldots \omega(X_N)\}$$

unde:

$$\omega(X_i) \in \{\omega_1 \ldots \omega_M\},\ i=1 \ldots N,$$

Daca se aplica la intrare un vector X care trebuie clasificat, se determina distanta minima de la X la vectorii de referinta etichetati. Presupunem ca:

$$d(X, X_j) = \min\{d(X,X_i),\ i=1 \ldots N\}$$

ceea ce inseamna ca X_j este cel mai apropiat vecin al lui X. Daca:

$$\omega(X) = \omega_h,$$

se aloca vectorului X clasa celui mai apropiat vecin din lotul de vectori etichetati.

4. Descrierea bazei de date

Baza de date WINE [3] contine pe fiecare linie clasa si 13 atribute care definesc apartenenta vinului la clasa respectiva. Fiecare vin din lista poate face parte dintr-una dintre clasele 1,2 sau 3. Sunt 59, 71 si 48 de vinuri care fac parte din clasele 1,2 si 3 respectiv. Cele 13 atribute in ordine sunt:

1) Alcohol

2) Malic acid

3) Ash

4) Alcalinity of ash

5) Magnesium

6) Total phenols

7) Flavanoids

8) Nonflavanoid phenols

9) Proanthocyanins

10)Color intensity

11)Hue

12)OD280/OD315 of diluted wines

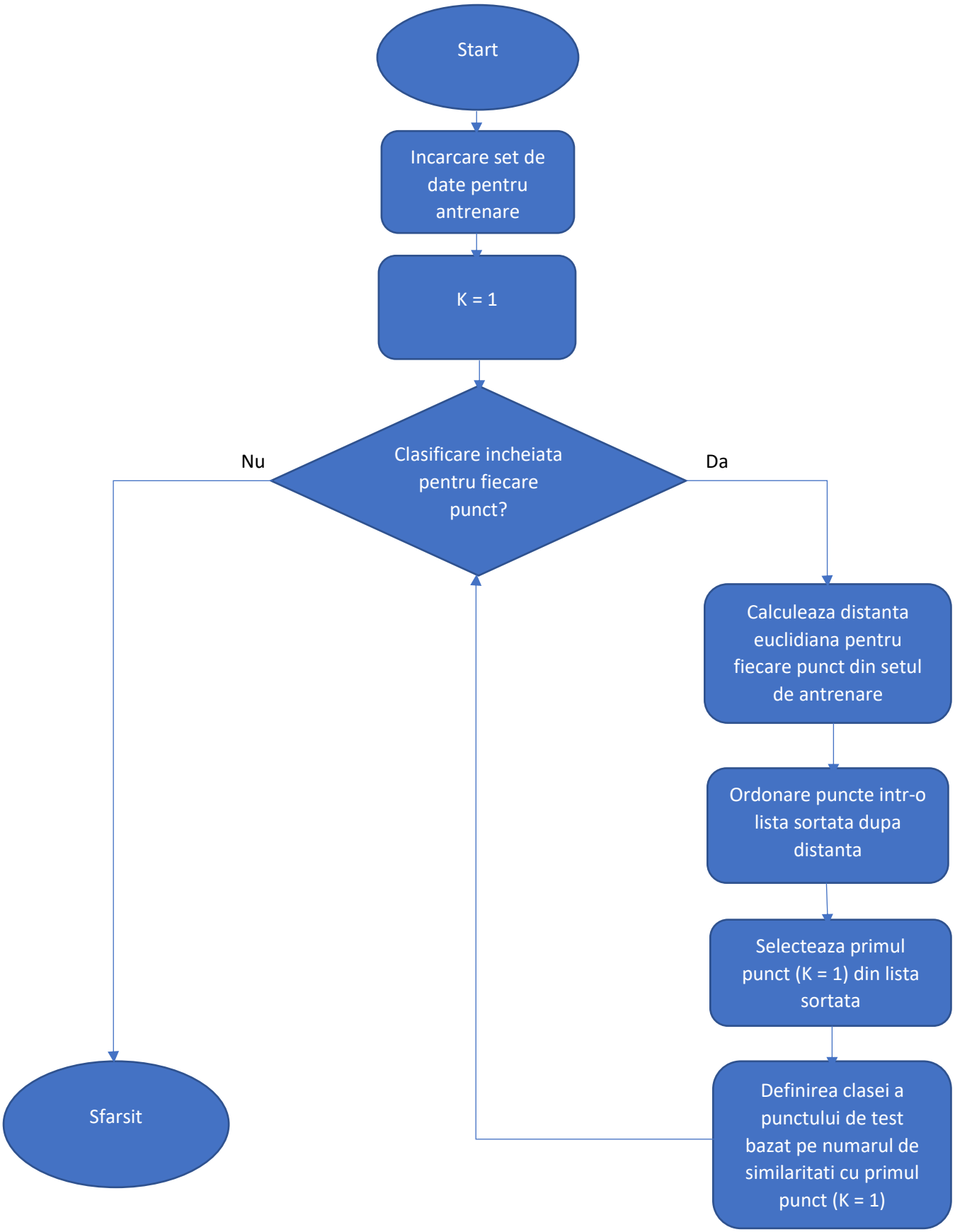
13)Proline

Un exemplu de linie din baza de date:

1,13.05,1.77,2.1,17,107,3,3,.28,2.03,5.04,.88,3.35,885

Primul numar defineste clasa vinului si urmatoarele 13 numere sunt atributele vinului in ordinea in care sunt definite.

5. Flowchart al algoritmului



6. Descrierea codului

```
%clear everything
close all; clc; clear;

%load database
load wine.data;

%select samples as baseline
sample_X = 1:5:length(wine);

%select different samples to recognize
resemble_Y = 2:5:length(wine);

%get the classes from the first column
X_classes = wine(sample_X,1);
Y_classes = wine(resemble_Y,1);

%get the relevant data from the other columns
X = wine(sample_X,2:size(wine,2));
Y = wine(resemble_Y,2:size(wine,2));

%get the indexes of the nearest neighbors
Idx = knnsearch(X, Y);

%get the classes of the nearest neighbors
nn_classes = X_classes(Idx);

%initialize classification status with zeros (false)
classified_status = zeros(1, size(nn_classes,1));

%if some classes have been correctly classified
for i = 1:size(nn_classes)
    if nn_classes(i) == Y_classes(i)
        %change the status to one (true)
        classified_status(i) = 1;
    end
end

%get all correctly classified wines
correctly_classified = length(classified_status(classified_status == 1));

%print out the classification result
fprintf('%d out of %d (%.2f%%) wines correctly classified.\n',
correctly_classified, length(nn_classes),
correctly_classified/length(nn_classes)*100);
```

Command Window

```
26 out of 36 (72.22%) wines correctly classified.
```

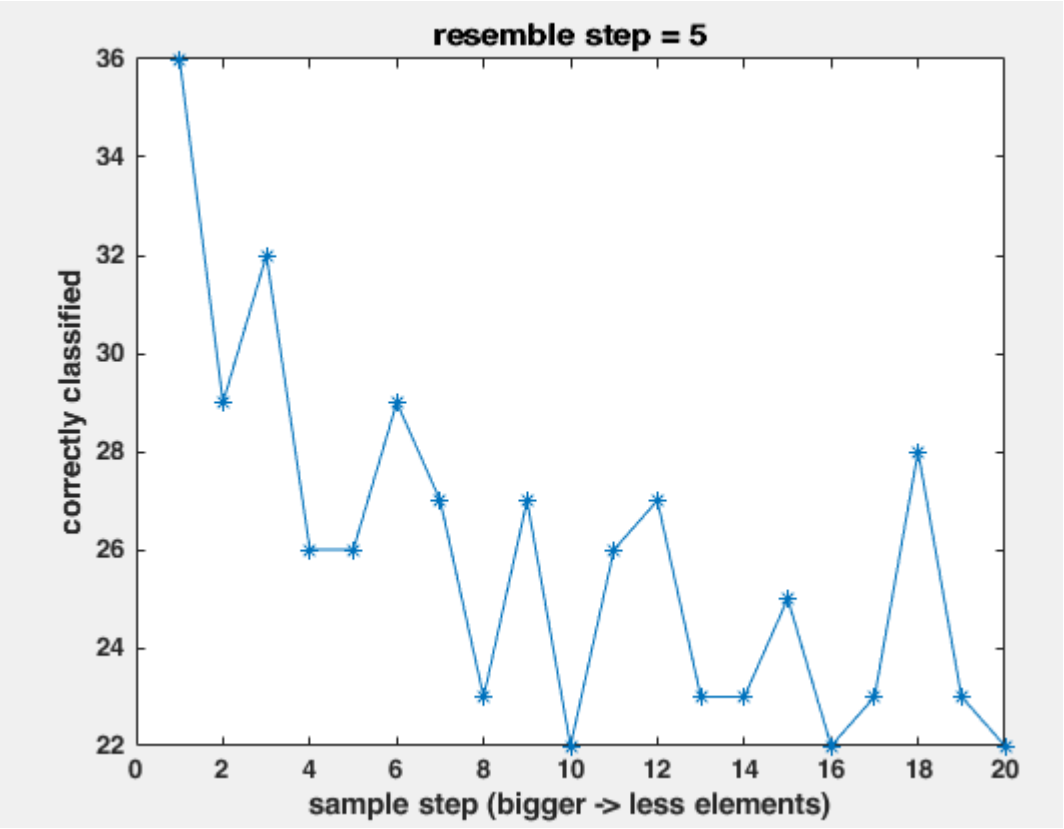
In momentul executiei programului se desfasoara urmatoarele procese: se elibereaza si se curata spatiul de lucru, se incarca baza de date(WINE), se selecteaza esantioanele ca baza de pornire si se selecteaza diferitele esantioane de recunoscut.

Apoi se iau clasele din prima coloana, se iau datele relevante din celelalte clase, se iau indexurile celor mai apropiati vecini, se iau clasele celor mai apropiati vecini, se initializeaza status-ul clasificarii cu 0 (fals) si se verifica daca clasele au fost clasificate corect(daca da, atunci se schimba status-ul la 1 (adevarat)), se iau toate vinurile clasificate corect si se afiseaza rezultatul clasificarii.

Clasificarea este realizata cu ajutorul functiei knnsearch [4] din Matlab. Idx = knnsearch(X, Y) cauta cel mai apropiat vecin in X pentru fiecare punct din Y si returneaza indecsii celor mai apropiati vecini, un vector de coloane. Idx are acelasi numar de linii ca Y.

Pentru calcularea distantei dintre 2 puncte, implicit se foloseste distanta euclidiana.

7. Interpretare rezultate



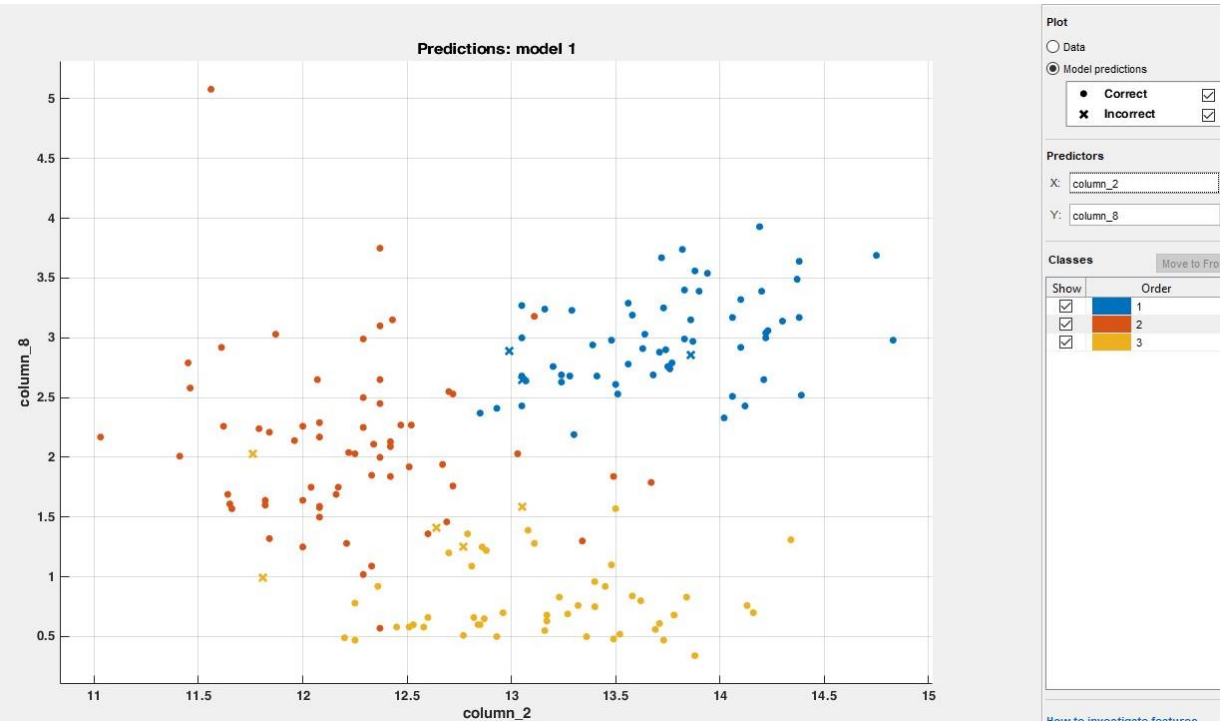
sample step – variabila care determina din cat in cat sunt luate elemente din baza de date care sunt folosite pentru “antrenare”

resemble step – variablia care determina din cat in cat sunt luate elemente din baza de date care sunt folosite pentru clasificarea corecta a acestora

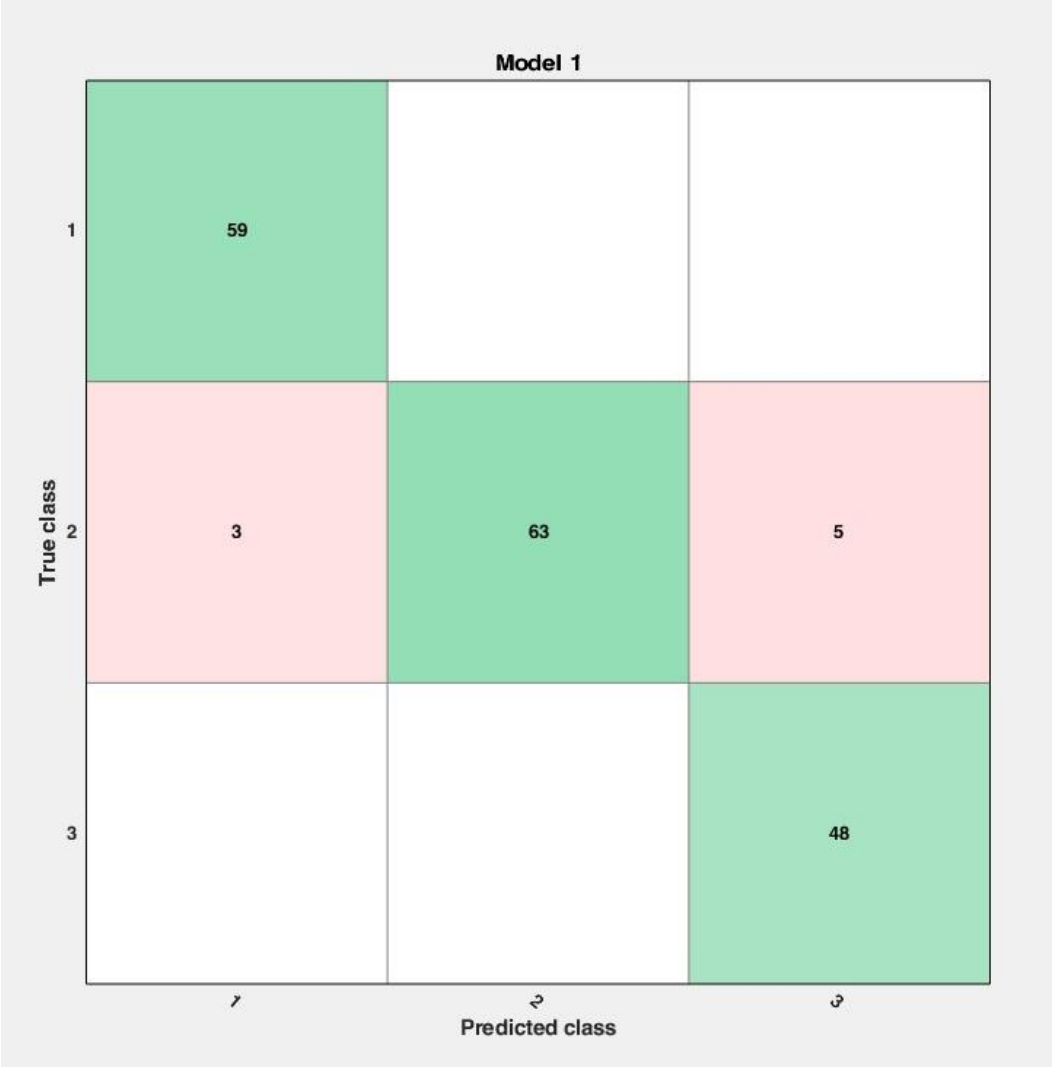
correctly classified – cate elemente au fost identificate corect din tot lotul de clasificare

In mod normal, cu cat creste sample step cu atat scade numarul de elemente identificate corect.

Dar din graficul de mai sus, se poate observa ca numarul de elemente clasficate corect alterneaza dar tinde sa scada. Aceasta consecinta este datorita faptului ca odata cu schimbarea lui sample step nu se garanteaza o distributie uniforma a elementelor/claselor din vectorul de “antrenare”.



Cele 3 clase din baza de date au fost afisate cu culori diferite in graficul de mai sus(1 – albastru, 2 – rosu, 3 – galben). Fiecare predictie corecta este afisata cu un cerc iar cele incorecte sunt reprezentate printr-un x. Se poate observa cum majoritatea vinurilor de acelasi tip sunt foarte apropiate unul de celelalte ca distanta in acest grafic.



In graficul de mai sus (matricea de confuzie) se prezinta cum au fost clasificate esantioanele din baza de date cu vinuri. Celulele care se afla pe diagonala principala sunt esantioanele care au fost clasificate correct. Celulele care nu se afla pe diagonala principala reprezinta esantioanele care au fost clasificate incorect. De exemplu, 3 vinuri de tip 2 au fost clasificate ca tip 1 iar 5 vinuri de tip 2 au fost clasificate ca tip 3. Modelul este bine antrenat deoarece sunt foarte putine esantioane care au fost clasificate incorect.

8. Bibliografie

[1] Wikipedia, "Învățare automată," [Online]. Available: https://ro.wikipedia.org/wiki/%C3%8Env%C4%83%C8%9Bare_automat%C4%83. [Accessed 2019].

[2] V. Neagoe, "Clasificatorii Nearest Prototype (NP) si k-Nearest Neighbor (k-NN)," [Online]. Available: http://www.victorneagoe.com/university/prai/lab1a.pdf. [Accessed 2019].

[3] Irvine, CA: University of California, School of Information and Computer Science., "UCI Machine Learning Repository: Wine Data Set," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/wine. [Accessed 2019].

[4] MathWorks, "Find k-nearest neighbors using input data - MATLAB knnsearch," [Online]. Available: https://www.mathworks.com/help/stats/knnsearch.html. [Accessed 2019].