1. Obiecte si caracteristici

Marina Gorunescu
mgorun@inf.ucv.ro

clasificare

Prezentăm câteva abordări mai cunoscute a termenului de **clasificare**.

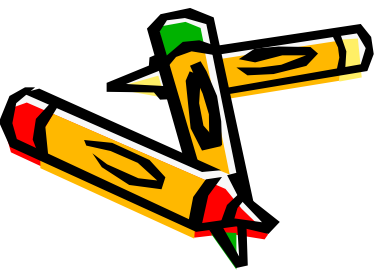
Clasificarea taxonomică reprezintă procesul plasării unui obiect sau concept într-o anumită categorie, pe baza proprietăților (caracteristicilor) acelui obiect.

Taxonomiştii sunt preocupați să construiască clasificări care însumează relații între unitățile taxonomice de diferite tipuri.

Aceste unități sunt incluse în clase care sunt disjuncte şi dispuse ierarhic, cum este spre exemplu sistemul lui Linne.

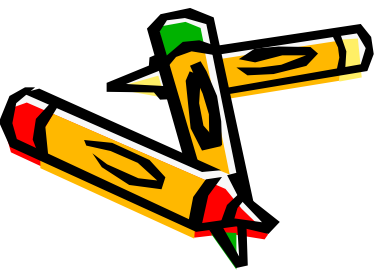
Clasificarea ştiinţifică (cunoscută şi ca *taxonomia ştiinţifică*), s-a bazat pe gruparea organismelor în diferite specii.

Clasificarea modernă are ca părinte pe omul de ştiința suedez **Carl von Linne** (Carolus Linnaeus, 1707-1778), care a grupat speciile în funcție de caracteristicile lor fizice.



probleme

- Arheologii sunt interesați de a găsi similarități în artefactele, cum ar fi ornamente sau unelte de piatră, găsite prin excavații, ceea ce le-ar permite să studieze distribuția spațială a tipurilor de artefacte.
(Hodson, Sneath, Doran,1966)

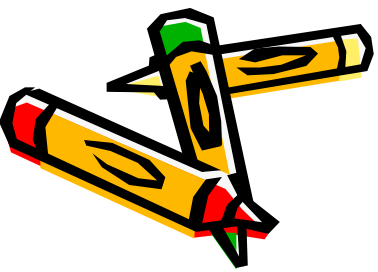


- Ecologiștii culeg informații asupra speciilor de plante ce se află într-o mulțime de parcele, inventariind speciile existente în fiecare parcelă și eventual înregistrând o măsură a bogăției fiecărei specii în parcelă.

Scop: a împărți aceste parcele în clase, astfel încât parcelele ce aparțin unei clase să aibă proprietăți distinctive față de celelalte clase.

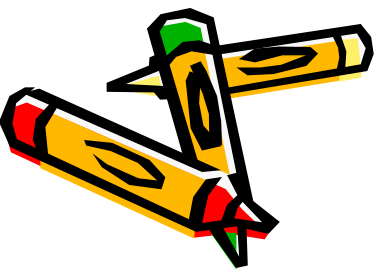
(Greig - Smith,1964)

- Analiştii social studiază interacţiunile existente într-o mulţime de indivizi şi sunt interesaţi de identificarea indivizilor ce au aceleaşi însuşiri (atribute).
(Arabie,Carrol,1989)



- Producători de whisky sunt interesați de o clasificare a distileriilor ce produc whisky, aceasta permițându-le să stabilească gama de proprietăți ale diverselor sortimente, cunoscând distileria care le produce.

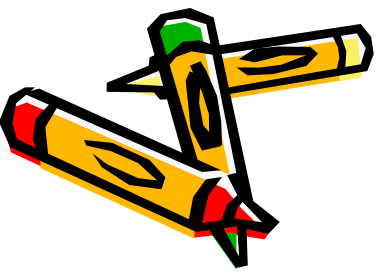
  Folosind aceste clasificări, producătorii își pot identifica concurența. (Lapointe, Legendre,1994)



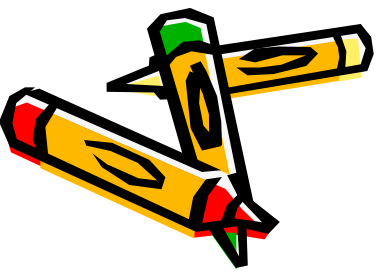
În aceste exemple un „obiect" este un artefact, o parcelă cu vegetație, un individ, o distilerie de whisky.

Obiectele sunt descrise printr-o mulțime de variabile, de exemplu:

- variabilă în cazul artefactelor este o proprietate fizică a acestora;
- în cazul parcelelor, o specie de plante este o variabilă.

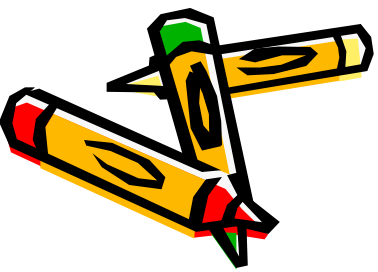


Rezultatul unei clasificări este o partiție a mulțimii obiectelor în clase disjuncte, obiectele aceleași clase fiind asemănătoare unele cu altele.



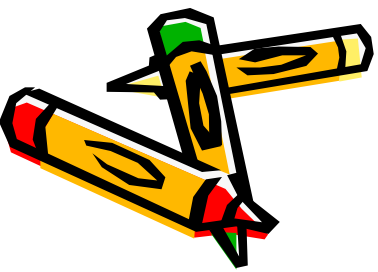
În sistemul lui Linne, o unitate taxonomică poate aparține unei specii, unui gen, unei familii, unui ordin.

Este interesant să obținem o clasificare ierarhică, care ar indica diferitele relații între clase.



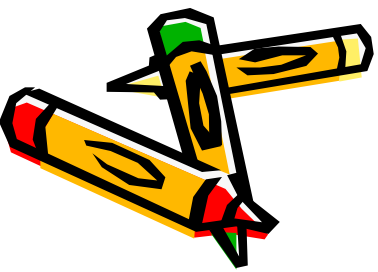
Clasificarea poate fi privită din puncte de vedere diferite:

- avem de determinat numărul de clase, trăsăturile caracteristice ale fiecărei clase şi obiectele ce constituie fiecare clasă;
- cunoscând clasele, determinăm apartenența fiecărui obiect la o anumită clasă;

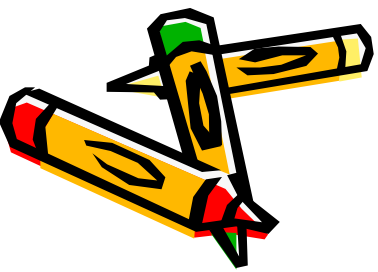


pattern recognition

- în „recunoaşterea formelor" (pattern recognition"), se presupune că fiecare obiect aparține unei anumite clase. Numărul claselor este cunoscut și cu ajutorul unei mulțimi de antrenament (**training set**) se determină proprietățile specifice fiecărei clase.
  Scopul este de a determina clasa căreia îi aparține un obiect .



Clasificarea statistică reprezintă o procedură statistică prin care obiecte individuale sunt plasate în diferite grupuri pe baza informației cantitative la dispoziție privind una sau mai multe caracteristici și utilizând o mulțime de antrenament la care se știe corespondența între fiecare obiect și categoria de care aparține.



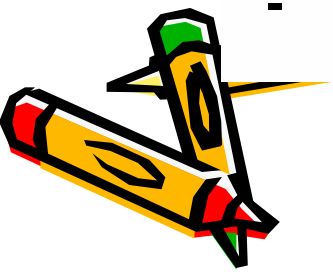
Formal, având la dispoziție o mulțime de antrenament

$$(X,Y)=\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$$

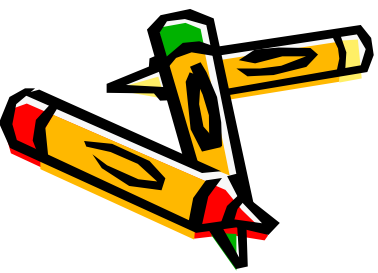
trebuie găsită o funcție de clasificare (***clasificator***)
$h: X \to Y$ pentru fiecare obiect x_i și clasă y_i
Odată funcția *h* estimată pe baza mulțimii de antrenament, utilizând corespondența $h(x_i)=y_i$, ea va fi utilizată pentru găsirea clasei corespunzătoare, adică $h(x)=?$ pentru un obiect nou, necunoscut x

aplicatii ale clasificarii

- Imagistica medicală;
- Recunoaşterea caracterelor optice;
- Geostatistica;
- Recunoaşterea vocală, a scrisului etc.;
- Biometria;
- Clasificarea documentelor;
- Căutarea pe Internet;
- Detectarea spam-urilor la poşta electronică.



Obiecte si caracteristici



clase (categorii)

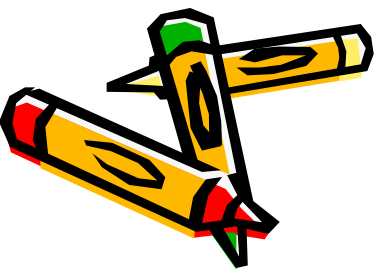
Obiectivul clasificării este de a grupa datele în *clase* (*categorii*):

- între elementele aceleiași clase avem un grad mare de **similaritate** (asemănare),
- între elementele din clase diferite gradul de similaritate este extrem de mic.

În raport cu scopul propus, se definește o măsură a similarității acestor elemente.

Fie o mulțime de elemente $X = \{x_1, \ldots, x_n\}$, pentru care este definită o măsură a similarității acestora;

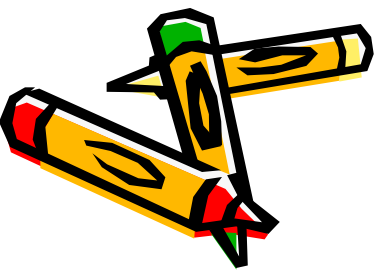
să determinăm clasele $\Omega_1, \ldots, \Omega_r$, clase ce formează o partiție a mulțimii X, astfel încât obiectele x_i care aparțin clasei Ω_j să fie cât mai asemănătoare între ele.



caracteristici (atribute)

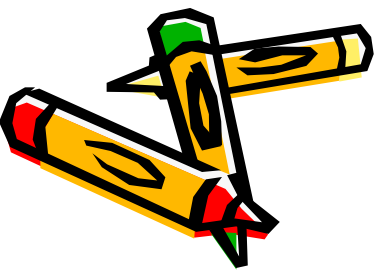
Pentru a putea decide într-o clasificare cărei categorii (clase) îi aparțin inputurile (obiectele $\mathbf{x_i}, i \in \{1, 2, ..., n\}$) este necesară cunoașterea mai multor caracteristici (*atribute*), provenite de obicei din măsurători.

Aceste caracteristici care pot fi considerate a fi componentele unui vector.

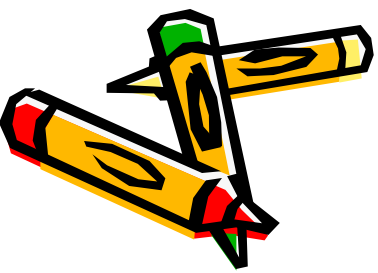


exemple actuale de clasificari automate

- recunoaşterea automată a vocii,
- recunoaşterea automată a amprentelor,
- recunoaşterea automată a lanţurilor ADN.

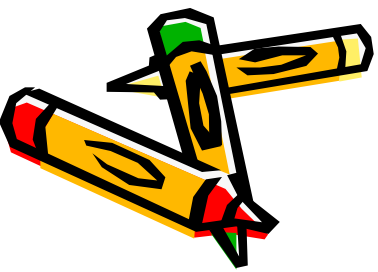


Alegerea caracteristicilor



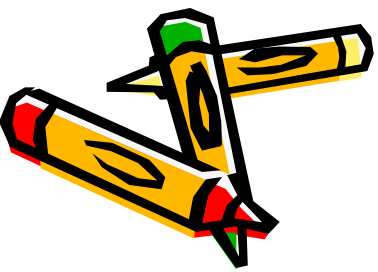
exemple clasice in literatura de specialitate

- Separarea a două specii de peşte: somon şi biban de mare, folosind senzori optici. (într-o fabrica de conserve de peşte)

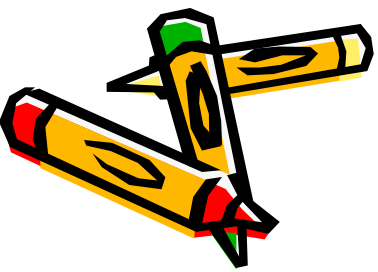


Prima etapă:
observarea diferențelor dintre cele două specii: lungime, luminozitate, lățime, numărul de configurații ale aripioarelor, etc.

Apar anumite variații de luminozitate, respectiv variații ale poziționării peștelui pe banda transportoare.

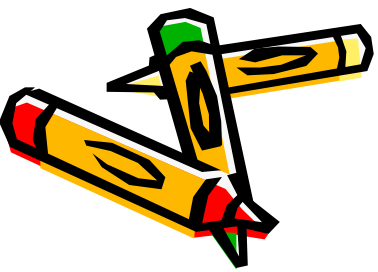


- *segmentarea*: imaginea unui peşte este izolată de a celorlalți;
- *extractorul de caracteristici* (se ocupă de reducerea bazei de date): păstrează doar o serie de caracteristici semnificative ale imaginii unui peşte.
- *clasificator*: ia decizia finală asupra speciei pe baza valorilor caracteristicilor



Tehnicile de clasificare, bazate în principal pe modele matematice, determină fiecare o anumită interpretare a structurii datelor.

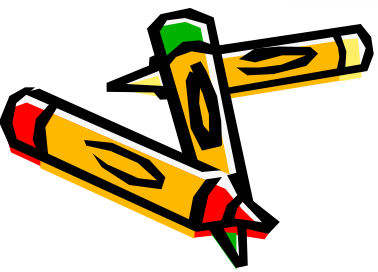
Se porneşte de obicei de la o metodă oarecare, care dă o prima clasificare ce ne furnizează o informaţie, care va fi ulterior îmbunătăţită prin aplicarea altor tehnici de clasificare.



alegerea caracteristicilor

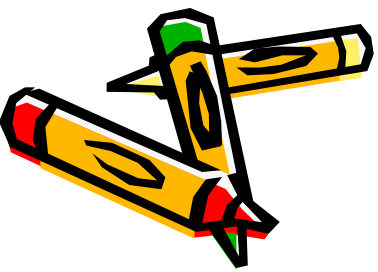
1. bibanul de mare este mai lung decât somonul;
   lungimea peștelui este o caracteristică care merită utilizată:
   dacă lungimea este mai mare decât o anumită valoare l ,
   peștele este biban de mare.

   Pentru alegerea acestui l vom face mai multe măsurători folosind o mulțime de antrenament.



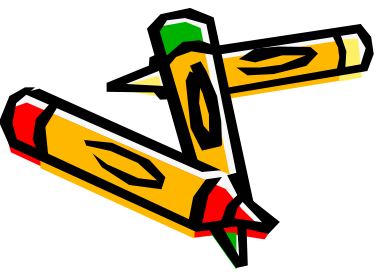
Mulțimea de antrenament este formată din 220 bibani de mare și 189 somoni.

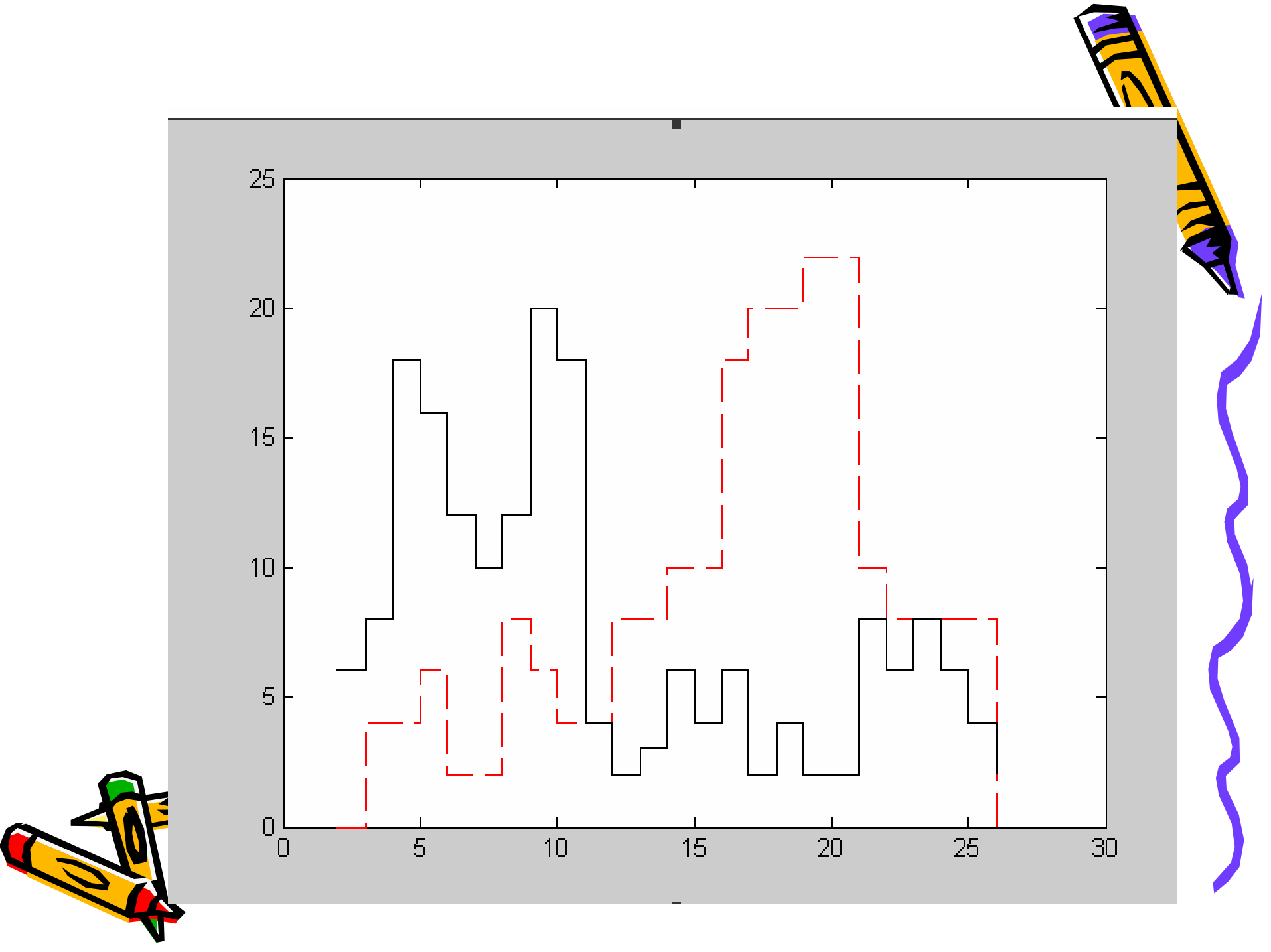
Vom reprezenta apoi grafic rezultatele obținute:



Lungime (unități)	Număr bibani de mare	Număr somoni		Lungime (unități)	Număr bibani de mare	Număr somoni
2	0	6		14	10	6
3	4	8		15	10	4
4	4	18		16	18	6
5	6	16		17	20	2
6	2	12		18	20	4
7	2	10		19	22	2
8	8	12		20	22	2
9	6	20		21	10	8
10	4	18		22	8	6
11	4	4		23	8	8
12	8	2		24	8	6
13	8	3		25	8	4

Vom desena histogramele corespunzătoare celor două specii de peşte, folosind MATLAB:

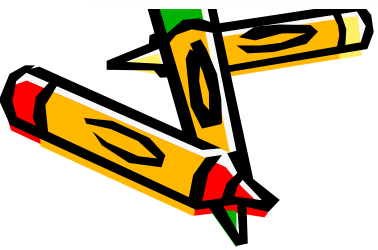



25
20
15
10
5
0
0
5
10
15
20
25
30

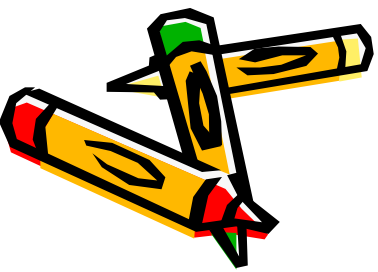
În acest caz este însă preferabil să folosim mediana lungimilor bibanului $med_B = 17$, respectiv mediana lungimilor somonului $med_S = 12$ (apar destul de multe valori extreme, care ar influența vădit valoarea mediei) .

$$l^* = \frac{med_B + med_S}{2} = 15 .$$

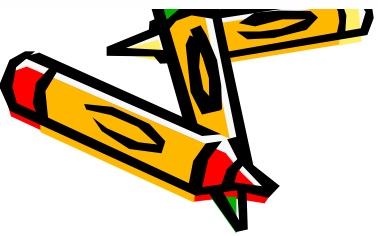
Criteriul de clasificare ales doar pe baza analizei lungimii peștilor este nepotrivit.

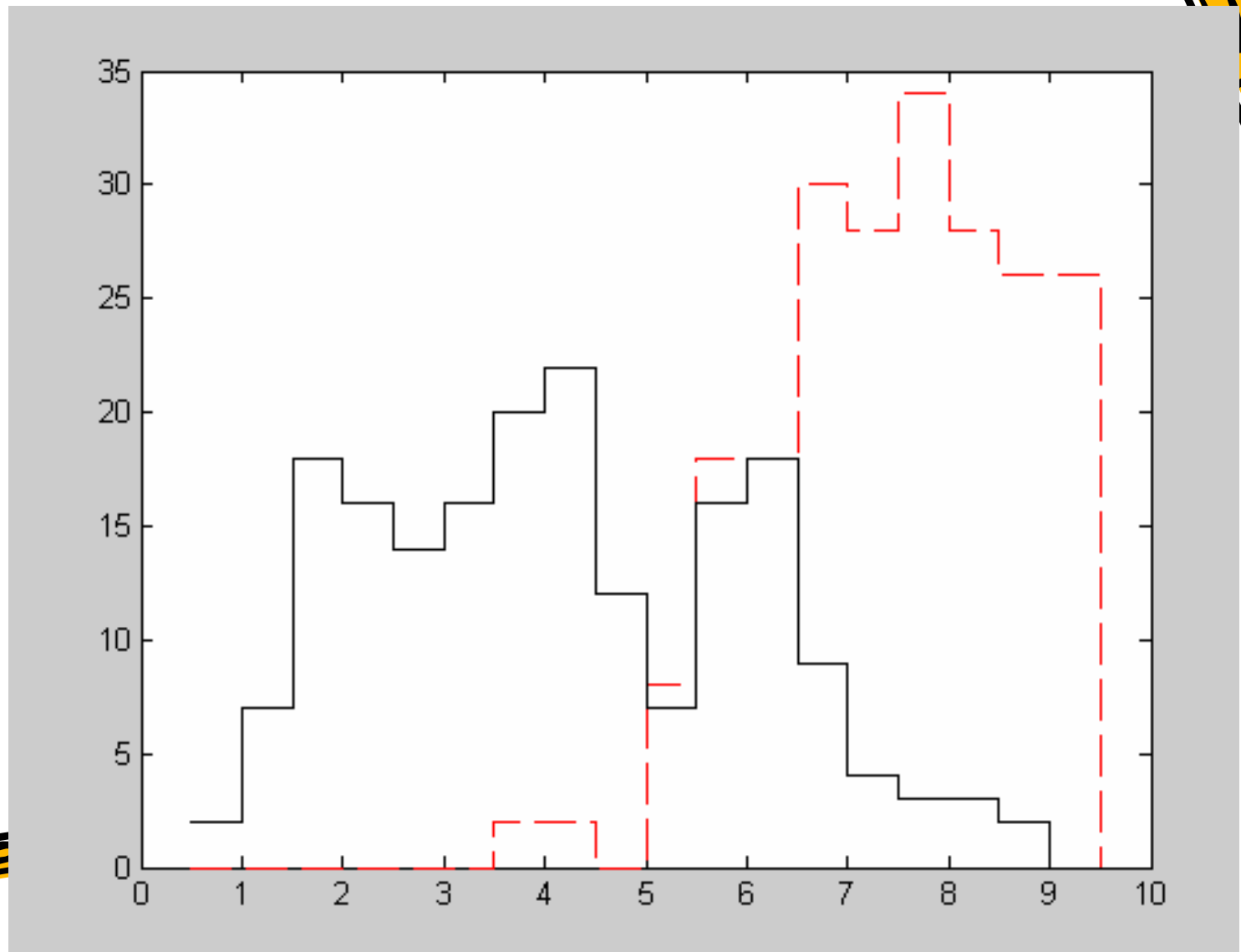


2. Pentru o clasificare mai precisă, vom alege drept caracteristică ***luminozitatea*** peștelui, după eliminarea „zgomotelor", adică a variațiilor de iluminare. Dacă luminozitatea va fi mai mică decât o anumită valoare x, peștele va fi somon, altfel este biban de mare.



Luminozitate (unități)	Număr bibani de mare	Număr somoni		Luminozitate (unități)	Număr bibani de mare	Număr somoni
0.5	0	2		5.5	18	16
1.0	0	7		6.0	18	18
1.5	0	18		6.5	30	9
2.0	0	16		7.0	28	4
2.5	0	14		7.5	34	3
3.0	0	16		8.0	28	3
3.5	2	20		8.5	26	2
4.0	2	22		9.0	26	0
4.5	0	12		9.5	0	0
5.0	8	7				



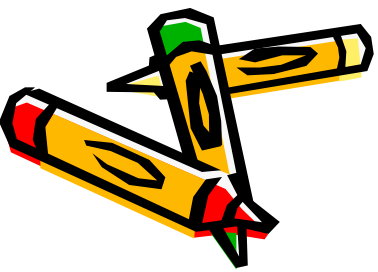

35
30
25
20
15
10
5
0
0
1
2
3
4
5
6
7
8
9
10

Mediile unităților ce măsoară luminozitatea sunt $\overline{m_B} = 7.21$, $\overline{m_S} = 3.93$ şi astfel $x = 5.57$.

Dacă lucrăm cu medianele, avem

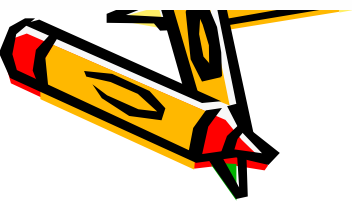
$med_B = 7.5$, $med_S = 4$ şi astfel $x^* = 5.75$.

Folosind luminozitatea peşteluio clasificare nesatisfăcătoare, reprezentând totuşi o abordare mai bună decât prima.



O clasificare greșită are evident, un anumit cost:

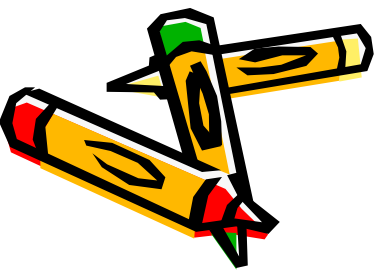
- dacă un somon este vândut ca biban de mare, fabrica este în pierdere;
- din punctul de vedere al cumpărătorului, dacă găsește într-o conservă etichetată biban de mare o bucată de somon, nu e deranjat, dar dacă într-o conservă etichetată somon găsește o bucată de biban de mare, sigur nu va mai cumpără acest tip de produs.



frontiera de decizie

Pentru a nu-și pierde clienții, fabrica va prefera prima variantă.

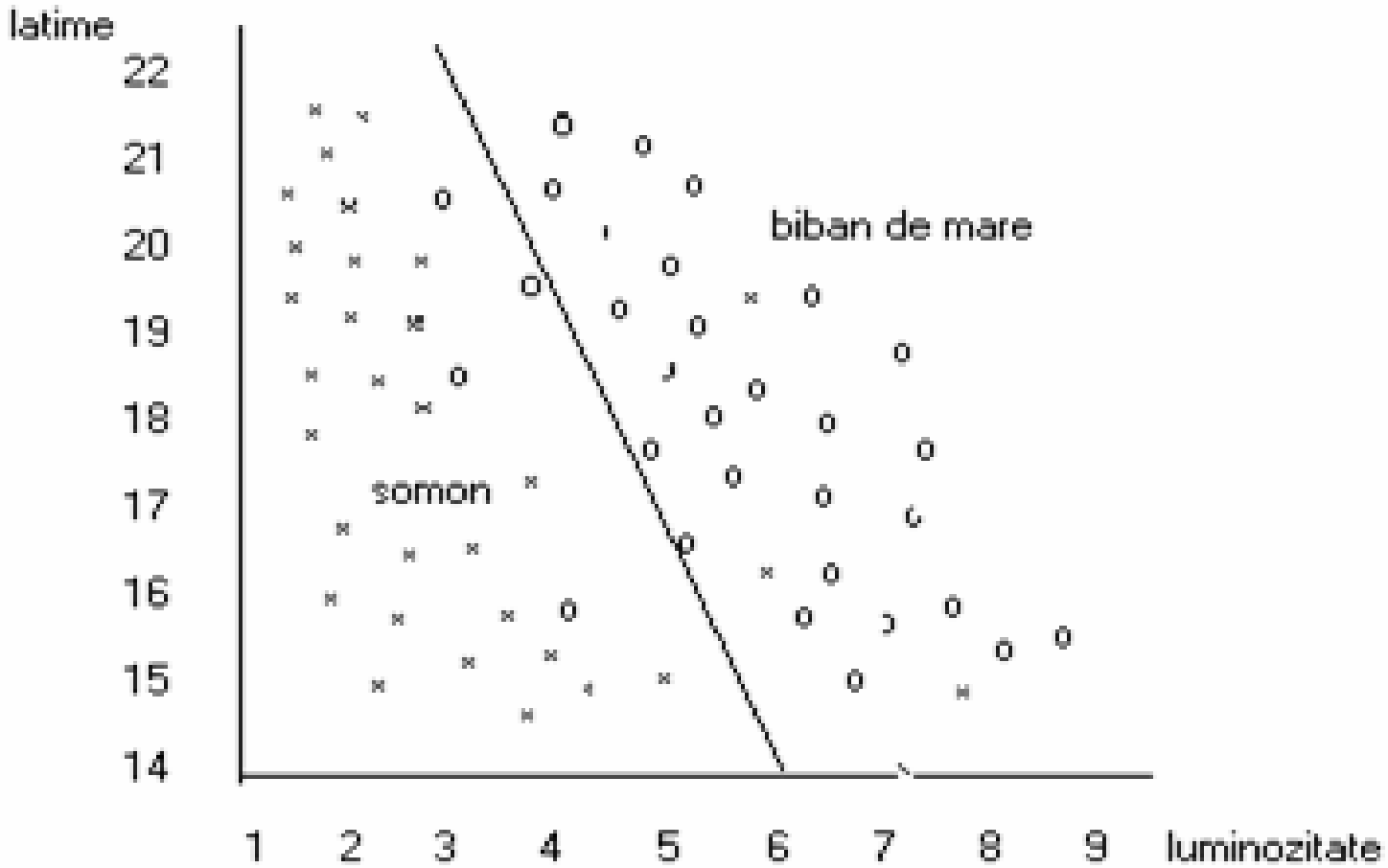
Obiectivul urmărit este alegerea ***frontierei*** (*suprafeței*) ***de decizie***, ceea ce înseamnă de fapt obținerea unei reguli de decizie, care să minimizeze costurile clasificării greșite (***misclassification cost***).



spatiul bidimensional al caracteristicilor

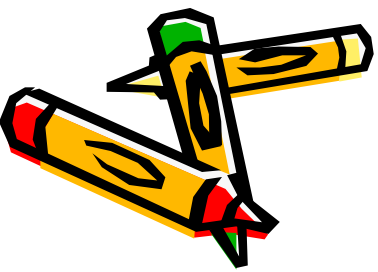
3. Pentru a îmbunătăți clasificarea, vom utiliza două caracteristici: luminozitatea peștelui x_1 și lățimea peștelui x_2, considerând vectorul $\mathbf{x} = (x_1, x_2)$, vector care reprezintă fiecare pește.

Exemplele de pești cunoscute vor fi reprezentate în spațiul bidimensional al caracteristicilor.
Construim o dreaptă, *frontiera de decizie*, care va separa planul în două regiuni de decizie, corespunzătoare celor două tipuri (*clase*) de pește.


latime
22
21
20
19
18
17
16
15
14
biban de mare
somon
1
2
3
4
5
6
7
8
9
luminozitate

merită să mărim dimensiunea spațiului caracteristicilor, luând în considerare și alte atribute, cum ar fi lungimea, culoarea sau poziționarea ochilor?

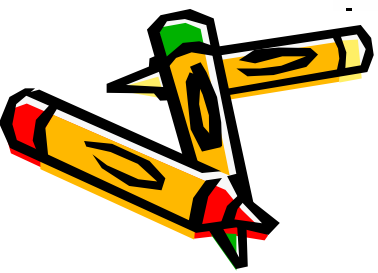
Unele caracteristici pot fi redundante, de exemplu culoarea ochilor peștilor este corelată cu lățimea lor, și astfel nu merită introdus acest nou atribut.



overfitting

Prea multe caracteristici pot influența negativ clasificarea, rezultând o suprafață de decizie particulară, specifică doar mulțimii de antrenament utilizate -fenomenul de ***overfitting***.

Obiectivul propus este de a determina, utilizând mulțimea de antrenament, o suprafață de decizie care să poată fi folosită cu succes în cazul unor exemple (inputuri) noi.



- Clasificarea celor trei tipuri de flori de Iris: *Iris Setosa*, *Virginica* şi *Versicolor* (există o bază de date foarte bună, datorată lui R.A. Fisher, 1936).

  Una dintre metode constă în măsurarea lungimii şi lăţimii petalelor. Notând

  x_1 = lungimea în cm a petalei şi

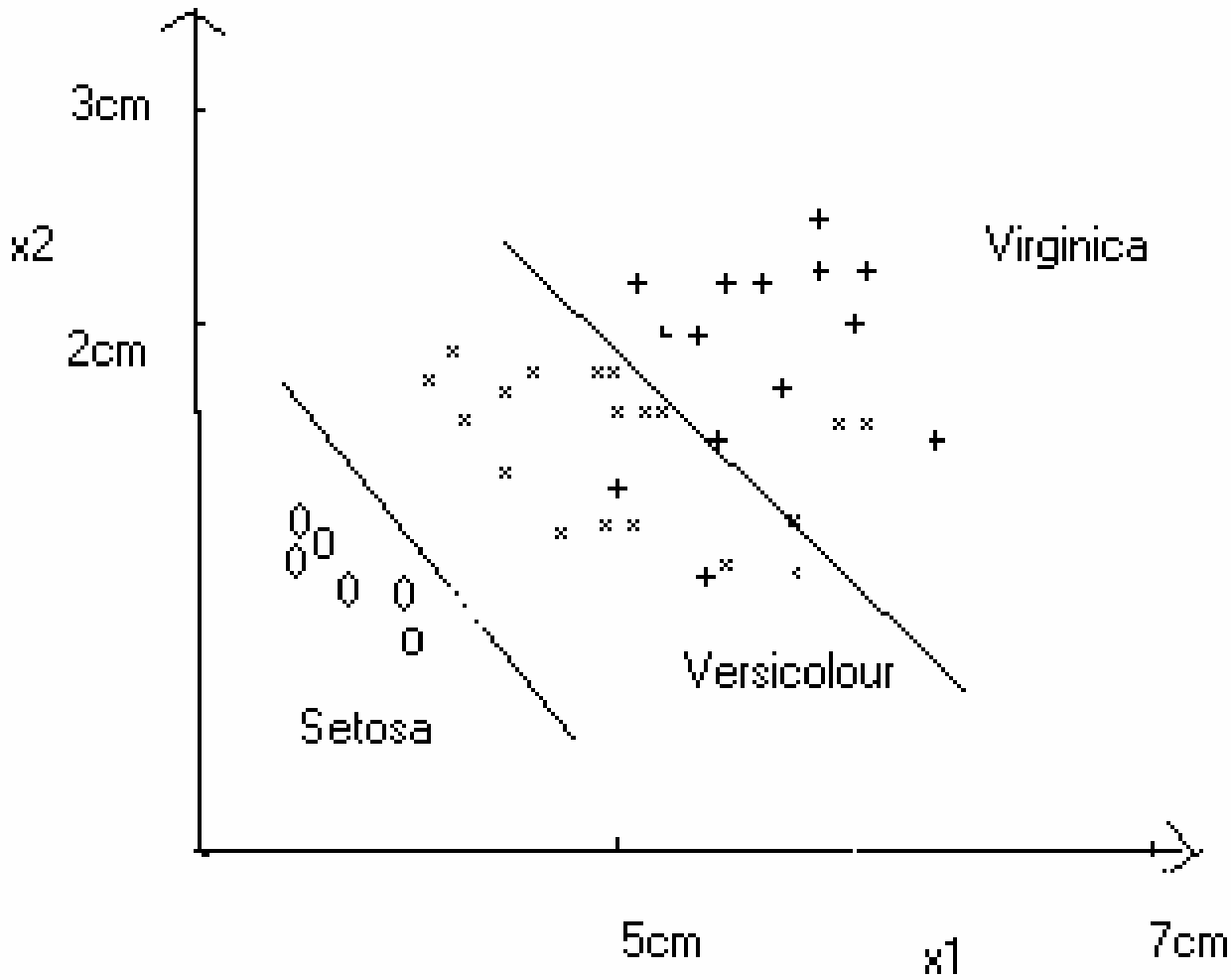
  x_2 = lăţimea în cm a petalei,

  construim vectorul $\mathbf{x} = (x_1, x_2)$, corespunzător unui iris.

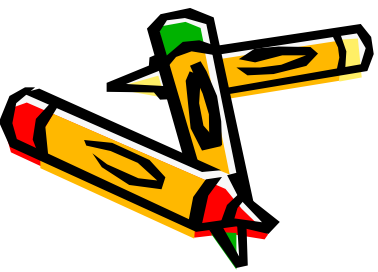
Iris setosa

Iris versicolor

Iris verginica

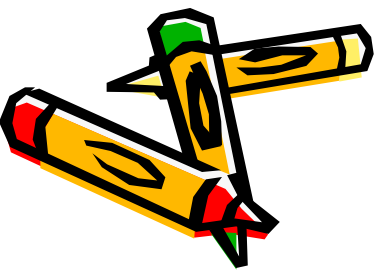

3cm
x2
2cm
Virginica
Versicolour
Setosa
5cm
x1
7cm

Fiecare punct din planul euclidian al caracteristicilor reprezintă un exemplu real de floare.
Cele două drepte, numite *suprafețe de decizie*, separă planul în trei regiuni, corespunzătoare celor trei clase (tipuri) de Iris.



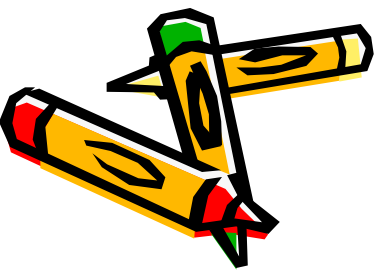
alegerea optima a caracteristicilor

În cazul unui număr mai mare de caracteristici (*curse of dimensionality*), este preferabilă reducerea dimensiunii vectorilor, studiind atent corelațiile existente între componente, folosind diverse metode statistice



În general, încercăm să lucrăm cu un număr mic de caracteristici, care sunt mai ușor de utilizat și cu ajutorul cărora vom obține regiuni de decizie mai simple.

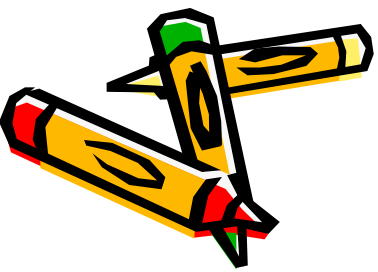
Suntem interesați de date „robuste", care nu sunt influențate de factorii externi (zgomot).



În cazul a *p* caracteristici, acestea vor fi componentele unui vector *p*-dimensional $\mathbf{x}_k = (x_1^k, \ldots, x_p^k)$.

Considerând *n* vectori de dimensiune *p*, notația x_i^k utilizată de obicei se referă la a *i*-a variabilă (caracteristică observată) a vectorului $\mathbf{x}_k$ (obiectul numărul *k* din mulțimea de antrenament).

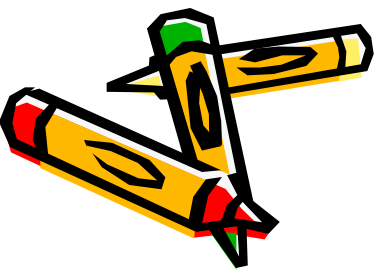
	Variabila 1	. .	Variabila *i*	. .	Variabila *p*
Obiectul 1	x_1^1		x_i^1		x_p^1
.....					
Obiectul *k*	x_1^k		x_i^k		x_p^k
.....					
Obiectul *n*	x_1^n		x_i^n		x_p^n



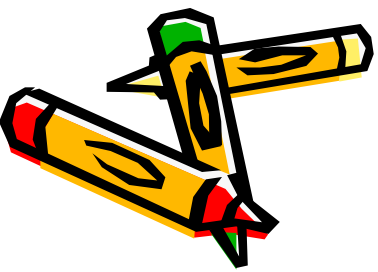
O alternativă mai utilă este aceea a reprezentării acestor date sub forma unei matrice **X**, cu *n* linii și *p* coloane:

$$\mathbf{X} = \begin{pmatrix} x_1^1 & \dots & x_i^1 & \dots & x_p^1 \\ \dots & \dots & \dots & \dots & \dots \\ x_1^k & \dots & x_i^k & \dots & x_p^k \\ \dots & \dots & \dots & \dots & \dots \\ x_1^n & \dots & x_i^n & \dots & x_p^n \end{pmatrix}$$

Asemenea date vectoriale multidimensionale pot fi vizualizate până la un anumit punct, folosind software de vizualizare (de exemplu XGOBI). Aceste instrumente de vizualizare sunt utilizate pentru a vedea legăturile în spațiu dintre vectori, ca un ghid în alegerea caracteristicilor distinctive.



Recunoaşterea statistică a formelor se ocupă nu numai de optimizarea suprafeței de decizie, ci face prognoze asupra modului cum va reacționa clasificatorul construit la noile exemple.



- construirea clasificatorului folosind mulțimea de antrenament,
- verificarea performanței utilizând o mulțime de testare, în cazul căreia se cunoaşte cărei clase îi corespunde fiecare element.
- tabel cu numărul de clasificări corecte, respectiv incorecte, din fiecare clasă.
- din procentajul de clasificări corecte pentru fiecare clasă, se calculează procentajul clasificării corecte pentru toată mulțimea.

procentaj clasificare corecta

Presupunem că mulțimea **X** are *n* elemente, care aparțin la *r* clase $\Omega_1,\ldots,\Omega_r$.

Numărul elementelor din fiecare clasă Ω_i este n_i, în timp ce numărul elementelor corect clasificate în clasa Ω_i este $m_i \leq n_i$.

Procentaj clasificare corectă

$$\sum_{i=1}^{r} \frac{n_i}{n} \cdot \frac{100 m_i}{n_i} = \frac{\sum_{i=1}^{r} 100 m_i}{n}$$

misclassification matrix

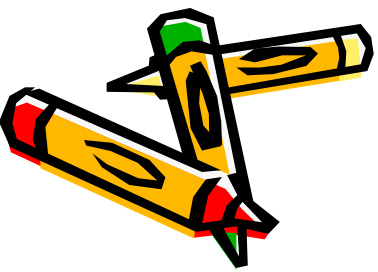
Pentru a vizualiza calitatea clasificării se foloseşte matricea clasificărilor greşite (*misclassification matrix*):

$$M_C = \begin{pmatrix} c_{11} & c_{12} & \dots & \dots & \dots & c_{1r} \\ c_{21} & c_{22} & \dots & \dots & \dots & c_{2r} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & c_{ij} & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ c_{r1} & c_{r2} & \dots & \dots & \dots & c_{rr} \end{pmatrix}$$

c_{ij} reprezintă numărul vectorilor de testare care sunt din clasa Ω_i, dar au fost clasificați greşit ca aparținând clasei Ω_j:

etapele unei probleme de clasificare

- pre-procesarea datelor;
- alegerea caracteristicilor;
- stabilirea funcției de decizie (funcția discriminant);
- clasificarea.



pre-procesarea datelor asigura

- reducerea dimensiunii datelor;
- filtrarea factorilor externi (zgomot);
- suprimarea detaliilor nerelevante;
- sublinierea caracteristicilor importante;
- invarianța în raport cu translațiile și rotațiile;
- pregătirea datelor pentru procedeul de decizie prin scalarea sau normalizarea lor.

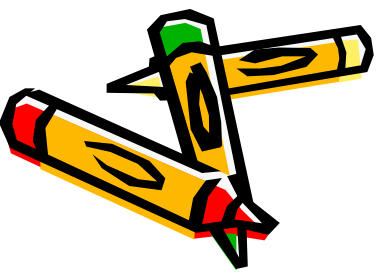
clasificarea datelor

Datele obținute prin măsurare pot fi clasificate în funcție de tipul de informație conținut.

- ***Datele categoriale*** sunt acele date care împart obiectele în diferite categorii.
- ***Datele numerice***

date categoriale

- date ***nominale***, ca de exemplu grupa sanguină (A/B/AB/O), culoarea ochilor, specia de Iris (Iris Setosa, Virginica și Versicolour).
- datele ***ordinale*** sunt date enumerative ordonate ca, de exemplu: gradul fumatului (nefumător, fost fumător, fumător ‚amator', fumător ‚înrăit'), ierarhizarea durerii (mică, medie, mare),



date nominale

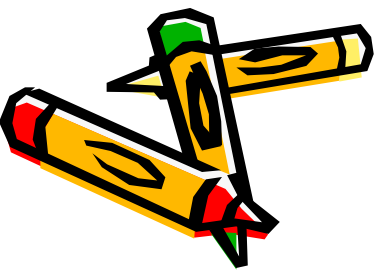
Datele nominale pot fi şi numerice, de exemplu codurile poştale

Datele nominale pot fi:

- binare (0 sau 1, da / nu, adevărat / fals)
- enumerative (date discrete pentru care nu este definită o ordine, cum ar fi categoriile socio-profesionale sau culoarea ochilor).

date numerice

- Datele ***discrete*** apar atunci când este vorba de observații numerice întregi, privitoare la un anumit proces de numărare ; de exemplu: numărul de copii ai unei familii, pulsul, codul numeric.
- Datele numerice ***continue*** se obțin de obicei în urma unor măsurători, de exemplu înălțimea, greutatea, tensiunea arterială, colesterolul unei anumite persoane, temperatura, viteza vântului, valoarea contului din bancă sau valoarea acțiunilor tranzacționate la Bursă etc.



remarca 1

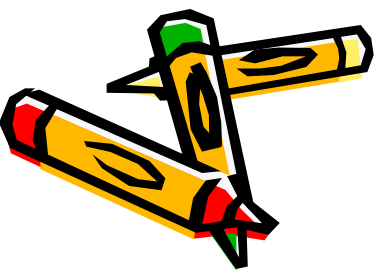
Din date continue se pot obține date discrete:

evaluarea venitului lunar:

venit lunar < 1000 RON,

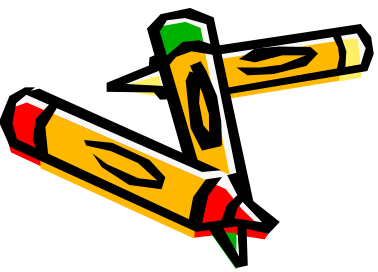
1000 RON < venit lunar < 2000 RON .

20 00 RON < venit lunar



remarca 2

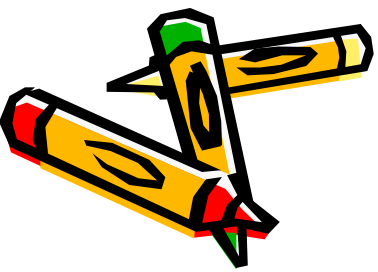
Datele numerice discrete sunt câteodată tratate ca date categoriale, de exemplu numărul de copii născuți de o femeie, 0, 1, 2, 3, 4, împart mamele în categoriile corespunzătoare numărului de copii.



remarca 3

nu este corect să interpretăm datele categoriale ordonate ca date numerice:

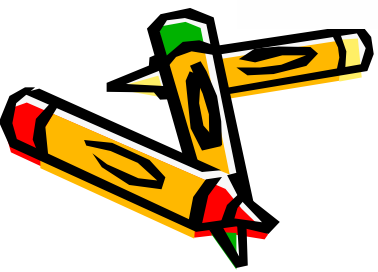
la stadiile în anumite boli, stadiul IV nu este de două ori mai rău decât stadiul II, ş.a.m.d.



alegerea tipului de pre-procesare

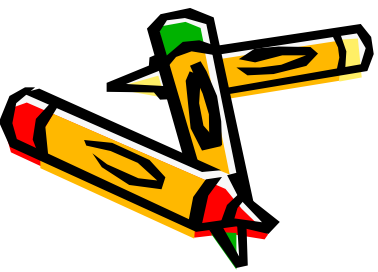
Dacă vectorii din baza de date au componente numerice, ***scalarea*** este absolut necesară:

- m este variabila obținută prin măsurare;
- x este variabila scalată
- $(m_{\min}, m_{\max})$ intervalul inițial al valorilor obținute prin măsurare
- $(x_{\min}, x_{\max})$ noul interval



$$x = \frac{x_{\max} - x_{\min}}{m_{\max} - m_{\min}}(m - m_{\min}) + x_{\min} =$$

$$\frac{x_{\max} - x_{\min}}{m_{\max} - m_{\min}} \cdot m + \left(x_{\min} - \frac{x_{\max} - x_{\min}}{m_{\max} - m_{\min}} \cdot m_{\min} \right)$$

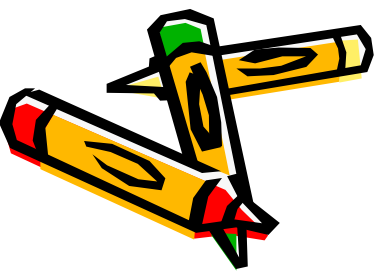


O posibilitate de a ***normaliza*** datele este folosirea formulei:

$$x_{norm} = \frac{x - x_{\min} + \varepsilon_1}{x_{\max} - x_{\min} + \varepsilon_2};$$

alegem constantele arbitrare ε_1 și ε_2, astfel încât $\varepsilon_1 < \varepsilon_2$, variabila normalizată x_{norm} aparține intervalului deschis (0,1).

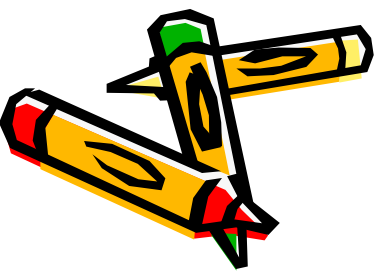
Constantele ε_1 și ε_2 trebuie judicios alese pentru a evita efectul de clusterizare (îngrămădire).



metoda de normalizare (Logan & Corwin)

Se folosește o funcție sigmoidă asimetrică care garantează că variabilele normalizate vor fi în intervalul (0, 1), în timp ce valorile periferice vor fi comprimate.

Funcția sigmoidă este de forma $f(x) = \frac{1}{1 + e^{-x}}$.



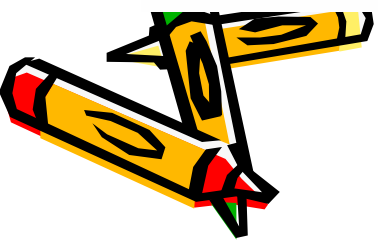
aplicarea metodei

1. Transformăm mulțimea de date $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ astfel încât seria statistică corespunzătoare să aibă media zero.

   Considerăm un atribut *A* având seria statistică $a_1, a_2, \ldots, a_n$.

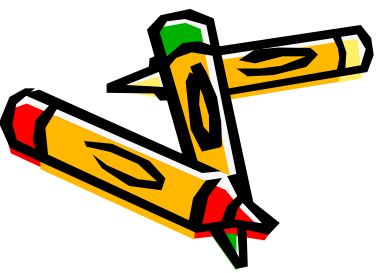
   Dacă media este $\overline{\mu} = \frac{1}{n} \sum_{j=1}^{n} a_j$, vom considera datele

   $a_j^* = a_j - \overline{\mu}$ și astfel $\overline{\mu^*} = \frac{1}{n} \sum_{j=1}^{n} (a_j - \overline{\mu}) = 0$.



2. Scalăm mulțimea X astfel încât marea majoritate a datelor (99.7%) să se afle în intervalul $(-3\sigma, 3\sigma)$, unde σ este deviația standard și media este 0.

Reamintim că, teoretic, *dispersia* corespunzătoare unei serii statistice $a_1, a_2, \ldots, a_n$ este definită de formula $\sigma^2 = \frac{1}{n} \sum_{j=1}^{n} (a_j - \overline{\mu})^2$, unde $\overline{\mu}$ reprezintă media (teoretică) a populației.



deviatia standard

Seria statistică cu care lucrăm reprezintă doar un eşantion mai mic al populaţiei. Cunoscând doar media acestui eşantion, notată $\bar{a}$, dăm o formulă de aproximare (o estimaţie) a dispersiei, înlocuind media $\bar{\mu}$ cu $\bar{a}$ şi împărţind la $n-1$ în loc de n:

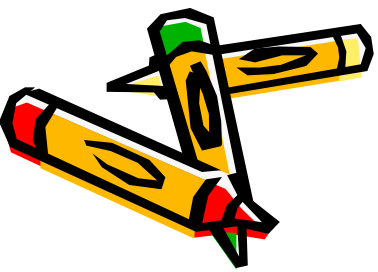
$$\sigma^2 = \frac{1}{n-1} \sum_{j=1}^{n} (a_j - \bar{a})^2$$

Deviaţia standard (*abaterea medie pătratică*) este o mărime folosită în locul dispersiei, dată de

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{j=1}^{n} (a_j - \bar{a})^2} \ ,$$

Normalizarea se face folosind ecuația:

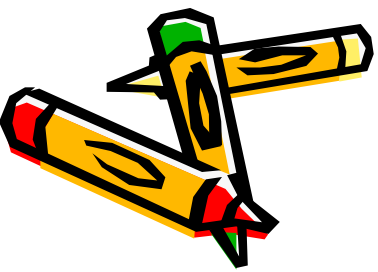
$$x_{norm} = \frac{1}{1 + e^{\lambda x}}.$$



Determinarea valorii lui λ:

- datele sunt în proporție de 99.7% în intervalul $(-3\sigma,3\sigma)$.
- pentru $\varphi(x)=\frac{1}{1+e^{\lambda x}}$, $\lambda<0$ avem

$$\varphi(-3\sigma,3\sigma)=\left(\frac{1}{1+e^{-3\lambda\sigma}},\frac{1}{1+e^{3\lambda\sigma}}\right)\subset(0,1)$$

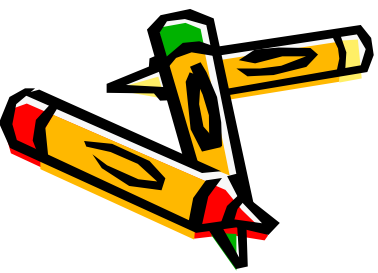


Să impunem, spre exemplu ca

$$\frac{1}{1+e^{3\lambda\sigma}} = 0.999$$

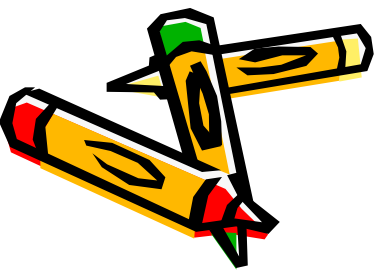
și astfel obținem:

$$\lambda = \frac{\ln\left(\frac{1}{0.999} - 1\right)}{3\sigma} = \frac{-6.9068}{3\sigma}.$$



paradoxul lui Cover

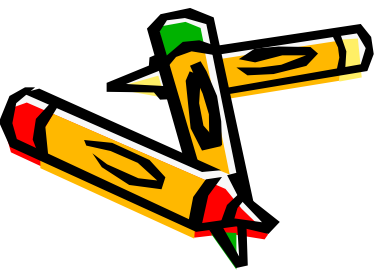
„mulțimea caracteristicilor alese și evaluate una câte una nu este la fel de bună ca mulțimea unor caracteristici combinate, adică cea mai bună alegere poate fi o combinație de caracteristici, independentă de semnificația fiecărei caracteristici în parte".



clasificator

Un ***clasificator*** poate fi considerat a fi o aplicație între mulțimea de caracteristici și mulțimea claselor.

Aceasta se realizează folosind distanțe (metrici), definite pe spațiul caracteristicilor și elemente de teoria probabilităților.

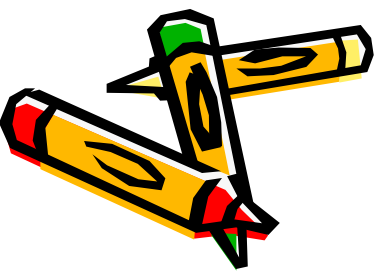


- *distanța euclidiană* în $\mathbb{R}^p$ între doi vectori $\mathbf{x}=(x_1,\ldots,x_p)$ și $\mathbf{y}=(y_1,\ldots,y_p)$ este definită prin:

$$d(\mathbf{x},\mathbf{y})=\sqrt{\sum_{k=1}^{p}(x_k-y_k)^2}$$

- *distanța Manhattan* $d_1(\mathbf{x},\mathbf{y})=\sum_{k=1}^{p}|x_k-y_k|$

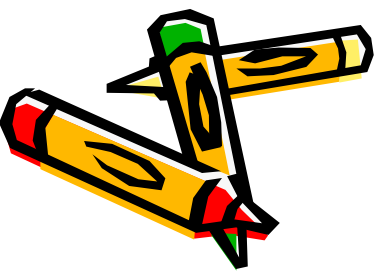
- distanța *Cebâșev* $d_\infty(\mathbf{x},\mathbf{y})=\max_{1\le k\le p}|x_k-y_k|$



Merita retinut!

- pentru *datele binare*, alegem distanța definită prin:
  $d(0,0) = d(1,1) = 0$ și $d(1,0) = d(0,1) = 1$;

- pentru *datele enumerative*, distanța cea mai utilizată este $d(x,y) = \begin{cases} 0, & x = y \\ 1, & x \neq y \end{cases}$

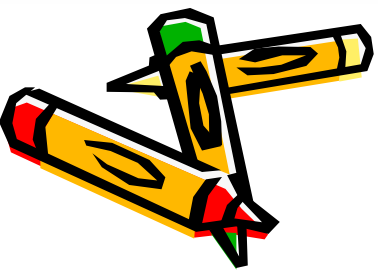


functii de decizie

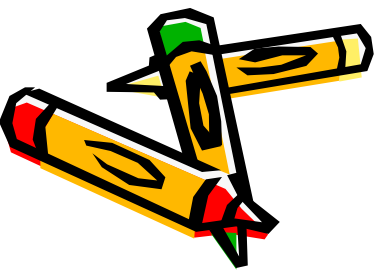
Marea majoritate a tehnicilor de recunoaştere a formelor (pattern-uri) folosesc *funcţii de decizie* (funcţii *discriminant*), care reprezintă frontiere în spaţiul caracteristicilor.

Dacă ne situăm în $\mathbf{R}^p$, funcţiile de decizie sunt funcţii reale definite pe $\mathbf{R}^p$.

Considerând că mulţimea datelor $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ este partiţionată în r clase $\Omega_1, \ldots, \Omega_r$ avem următoarele cazuri:



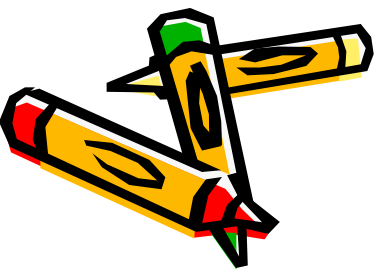
1. Fiecare clasă Ω_i este separată de celelalte printr-o funcție de decizie, existând astfel r funcții de decizie. Dacă $g_i : \mathbf{R}^p \rightarrow \mathbf{R}$ este funcția de decizie ce corespunde clasei Ω_i, atunci ecuația suprafeței de decizie ce separă clasa Ω_i de celelalte clase este $g_i(\mathbf{x}) = 0$



Avem regula: dacă $\mathbf{x} \in \Omega_i$ atunci $g_i(\mathbf{x}) > 0$

Astfel, dacă pentru un element nou $\mathbf{x} \in \mathbf{X}$ avem:

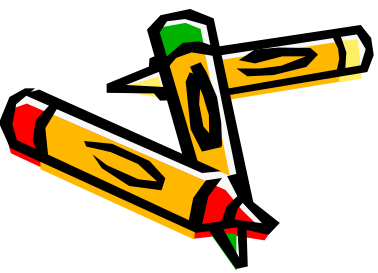
$g_i(\mathbf{x}) > 0$ şi $g_j(\mathbf{x}) > 0, \forall j \neq i$, atunci $\mathbf{x} \in \Omega_i$

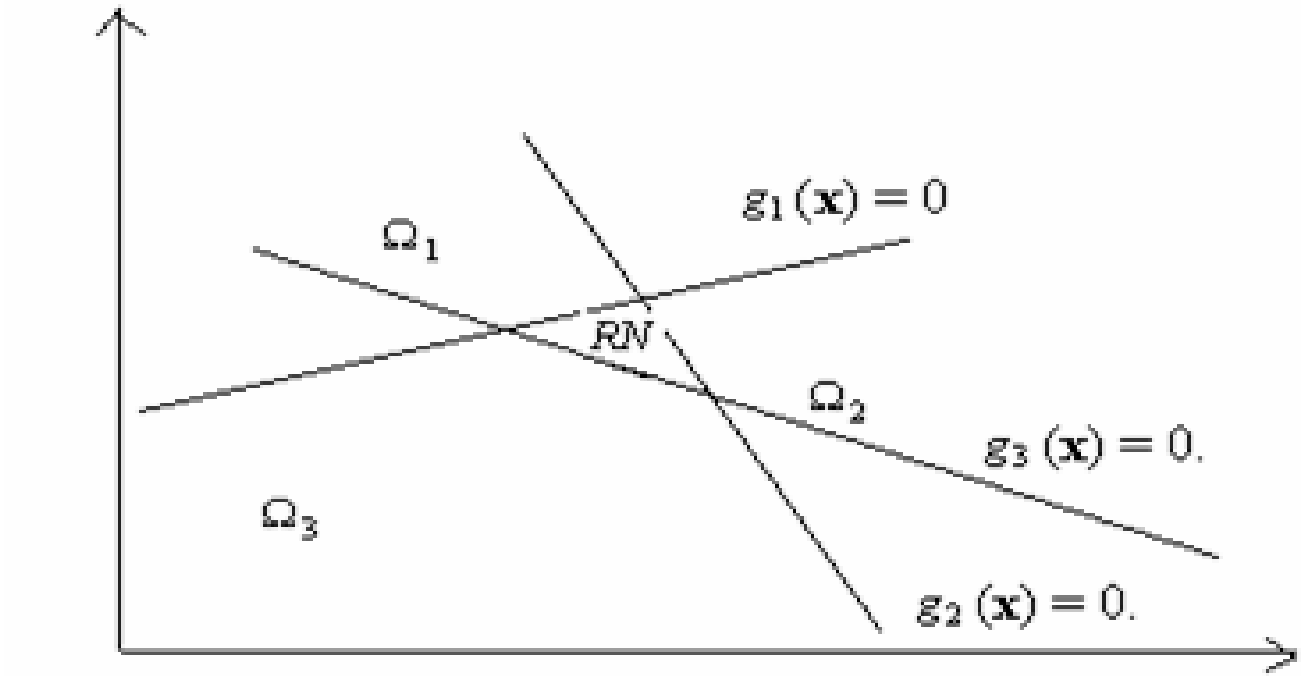


Putem defini regiunea de decizie corespunzătoare clasei Ω_i, ca fiind

$$\mathbf{R}_i = \{ \mathbf{x} \in \mathbf{X} \mid g_i(\mathbf{x}) > 0 \text{ şi } g_j(\mathbf{x}) < 0, \ \forall j \neq i \}.$$

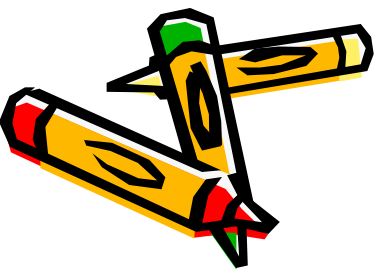
Există elemente ce nu aparţin nici unei regiuni de decizie, acestea formând ***regiunea de nedeterminare***.
Punctele suprafeţelor de decizie aparţin acestei regiuni.




$g_1(\mathbf{x}) = 0$
Ω_1
RJN
Ω_2
$g_3(\mathbf{x}) = 0.$
Ω_3
$g_2(\mathbf{x}) = 0.$

2. Clasele sunt două câte două separabile, adică fiecare clasă este separată de oricare alta printr-o suprafață de decizie distinctă.

   Suprafața de decizie corespunzătoare claselor Ω_i și Ω_j este definită de ecuația $g_{ij}(\mathbf{x}) = 0$.

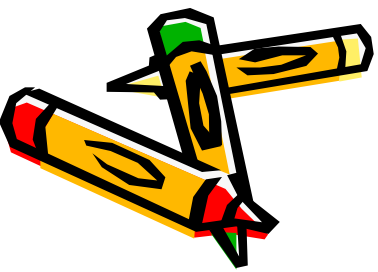


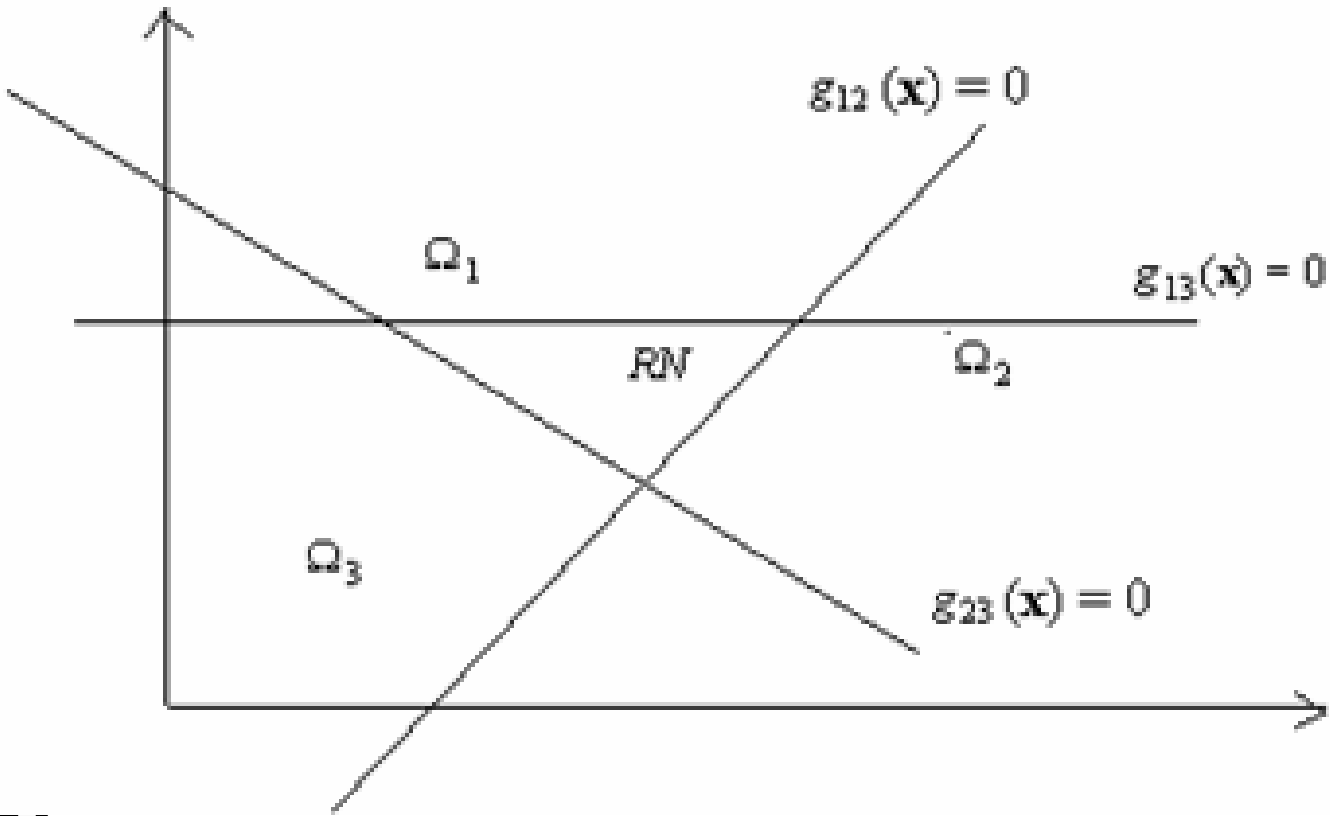
Regula de decizie este

$\mathbf{x} \in \Omega_i$ dacă și numai dacă $g_{ij}(\mathbf{x}) > 0, \ \forall j \neq i$.

Regiunea de decizie corespunzătoare clasei Ω_i este

$\mathbf{R}_i = \{ \mathbf{x} \in \mathbf{X} | \ g_{ij}(\mathbf{x}) > 0, \ \forall j \neq i \}$




$g_{12}(\mathbf{x}) = 0$
Ω_1
$g_{13}(\mathbf{x}) = 0$
RN
Ω_2
Ω_3
$g_{23}(\mathbf{x}) = 0$

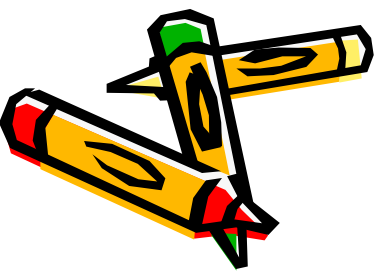
3. Există *n* funcții de decizie $g_i : \mathbf{R}^p \rightarrow \mathbf{R}$.

Regula de decizie este:

$\mathbf{x} \in \Omega_i$ dacă și numai dacă $g_i(\mathbf{x}) > g_j(\mathbf{x})$, $\forall j \neq i$,

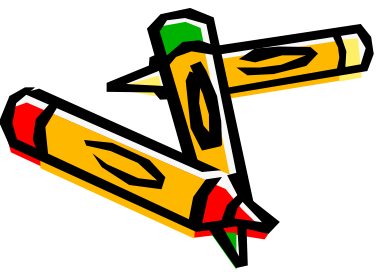
regiunea de decizie corespunzătoare clasei Ω_i fiind:

$\mathbf{R}_i = \{ \mathbf{x} \in \mathbf{X} | \; g_i(\mathbf{x}) > g_j(\mathbf{x}) , \; \forall j \neq i \}$.



Suprafața de decizie dintre clasele Ω_i şi Ω_j are ecuația: $g_i(\mathbf{x}) = g_j(\mathbf{x}), \mathbf{x} \in \mathbf{X}$

Elementele clasei Ω_i se află în partea pozitivă a suprafeței de separare.



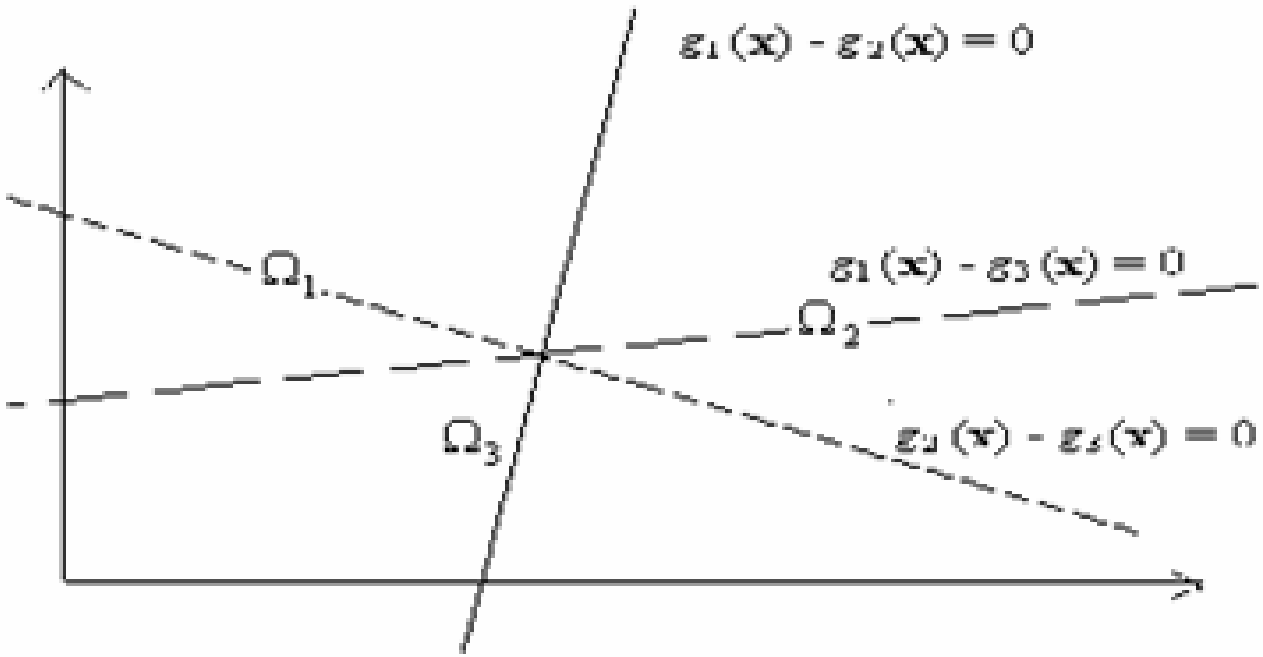
Presupunând separabilitatea claselor în acest caz, clasele vor fi separabile şi în cazul anterior (2):

notând $g_{ij}(\mathbf{x}) = g_i(\mathbf{x}) - g_j(\mathbf{x}),\ \mathbf{x} \in \mathbf{X},$

dacă $\mathbf{x} \in \Omega_i$, avem:

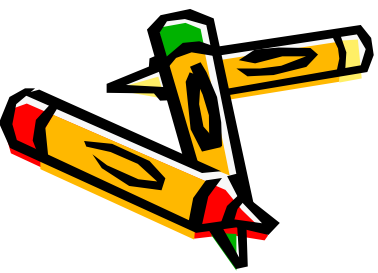
$$g_i(\mathbf{x}) > g_j(\mathbf{x}),\ \forall j \neq i\ ,\ \text{adică}\ g_{ij}(\mathbf{x}) > 0,\ \forall j \neq i\ .$$

Reciproca nu este în general valabilă.


$g_1(x) - g_2(x) = 0$
$g_1(x) - g_3(x) = 0$
$g_2(x) - g_3(x) = 0$
Ω_1
Ω_2
Ω_3

clasificatorul liniar

Cel mai simplu tip de funcție de decizie este *clasificatorul liniar*, caz în care suprafețele de decizie sunt *hiperplane*.
Un hiperplan unidimensional este o valoare prag (*treshhold*), un hiperplan bidimensional este o dreaptă în timp ce un hiperplan în $\mathbf{R}^3$ este un plan în sensul cunoscut.

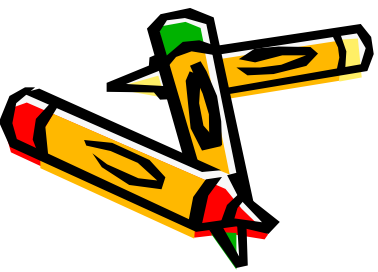


- Un exemplu simplu de funcție discriminant în cazul unidimensional este indicele de obezitate - IMC.

  Acesta se calculează raportând greutatea corporală (în kg) la pătratul înălțimii (în metri), conform formulei:

$$IMC = \frac{m}{h^2}$$

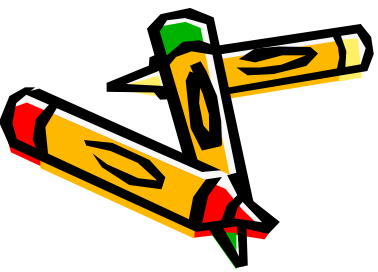
Un indice mai mare sau egal cu 30 clasifică persoana drept obeză, având nevoie de ajutor medical.



invatare

Orice metodă care colectează informație dintr-o mulțime de antrenament în scopul creării unui clasificator implică ***învațare***.

Învățarea se referă la estimarea parametrilor necunoscuți ai modelului şi la minimizarea erorii algoritmilor construiți, utilizând mulțimea de antrenament.

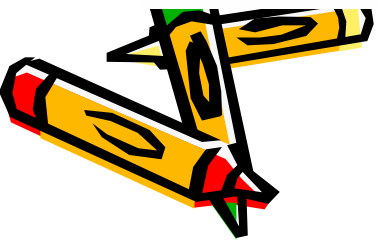


invatarea supervizata

În ***învățarea supervizată***, un „profesor" desemnează un cost pentru fiecare clasificare, scopul urmărit fiind reducerea sumei costurilor.

Ce se cere algoritmului creat:

- să fie stabil la variațiile parametrilor,
- să fie convergent în timp finit
- să dea soluții simple.

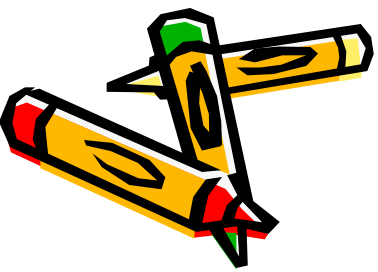


invatarea nesupervizata

În *învațarea nesupervizată*, nu există „profesor", sistemul fiind cel ce creează grupări naturale numite „*clustere*" din mulțimea de antrenament.

Problema de rezolvat:

evitarea reprezentărilor nepotrivite, deoarece algoritmi diferiți de clusterizare ne conduc la clustere diferite.



invatarea consolidata

Un tip particular de învățare nesupervizată este ***învățarea consolidată***, numită și învățare cu un „critic" . (agent).

Pentru a antrena și, respectiv, a îmbunătăți un clasificator, se determină eticheta cu ajutorul acestuia și se compară cu eticheta deja cunoscută.

Rezultatul obținut este corect / incorect.