CLASIFICARE

COLECTIE DE DATE (BAZA DE DATE)

- Colectie de obiecte pentru care se cunosc un set de caracteristici
- Se poate reprezenta ca un tabel in care pe linii avem obiectele, iar pe coloane caracteristicile obiectelor
- Caracteristicile le numim atribute
- Obiectele le numim inregistrari sau exemple sau entitati

ID	Venit lunar	Credit	Clasa
1	4800	Excelent	h1
2	2800	Bun	h1
3	1200	Excelent	h1
4	2400	Bun	h1
5	4400	Bun	h1
6	1600	Excelent	h1
7	3200	Nesatisfacator	h2
8	1600	Nesatisfacator	h3
9	2400	Nesatisfacator	h2
10	200	Nesatisfacator	h3

Outlook	Temp	Humidity	Windy	Play
sunny	hot	high	F	N
sunny	hot	high	T	N
overcast	hot	high	F	Y
rain	mild	high	F	Y
rain	cool	normal	F	Y
rain	cool	normal	T	N
overcast	cool	normal	T	Y
sunny	mild	high	F	N
sunny	cool	normal	F	Y
rain	mild	normal	F	Y
sunny	mild	normal	T	Y
overcast	mild	high	T	Y
overcast	hot	normal	F	Y
rain	mild	high	T	N

TIPURI DE ATRIBUTE

- CONTINUI – pot lua valori numerice intr-un interval dat

  Ex:

  - cantitatea vanduta,
  - temperatura (in grade),
  - greutatea (in kg)

- DISCRETE – iau un numar finit de valori

  Ex:

  - culoarea(albastru, rosu, galben, alb, etc),
  - Temperatura (ridicata, normala, scazuta)
  - Tip ciuperca(otravitoare, comestibila) –atribut binar

OPERATII DATA MINING DIRECTIONATE

- Atributele bazei de date se impart in:
  - Atribut tinta
  - Atribute de intrare

  Scopul este de a estima valorile atributului tinta.
- Tipuri de operatii directionate:
  - Clasificarea
  - Regresia (estimarea)

- Operatiile data mining pot fi directionate sau nedirectionate. In cazul celor directionate unul din atributele bazei de date este considerat atribut tinta pe care, de cele mai multe ori, dorim sa il estimam. O parte sau chiar toate celelalte atribute sunt considerate atribute de intrare in functie de care se determina valoarea atributului tinta.
- Se pare ca oamenilor le place mult sa faca clasificari. Clasificam animalele in specii, materia in elemente, oamenii in rase.

CLASIFICAREA

- A clasifica = a examina trasaturile si caracteristicile unui obiect si a-l repartiza unui set de clase predefinite.
- Data o baza de date, clasificarea inseamna a adauga o noua coloana bazei de date (un nou atribut, numit de multe ori Clasa) si a determina pentru fiecare inregistrare care este clasa careia ii apartine.
- Atributul tinta Clasa este discret.

Outlook	Temp	Humidity	Windy	Play
sunny	hot	high	F	N
sunny	hot	high	T	N
overcast	hot	high	F	Y
rain	mild	high	F	Y
rain	cool	normal	F	Y
rain	cool	normal	T	N
overcast	cool	normal	T	Y
sunny	mild	high	F	N
sunny	cool	normal	F	Y
rain	mild	normal	F	Y
sunny	mild	normal	T	Y
overcast	mild	high	T	Y
overcast	hot	normal	F	Y
rain	mild	high	T	N

Se pare ca oamenilor le place mult sa faca clasificari. Clasificam animalele in specii, materia in elemente, oamenii in rase.

- A clasifica inseamna a examina trasaturile si caracteristicile unui obiect si a-l repartiza unui set de clase predefinite. Obiectele care sunt caracterizate in general sunt reprezentate de inregistrarile unei baze de date sau fisier, iar clasificarea inseamna a adauga o noua coloana (un nou atribut) cu un cod al unei clase de un anumit tip si a determina pentru fiecare inregistrare care este clasa careia ii apartine.

CLASIFICAREA

- Clasificarea consta in construirea unui model care sa poata fi aplicat unor date neclasificate inca tocmai pentru a putea fi clasificate.
- Procesul construirii unui clasificator (model de clasificare) are doua etape:
  - Construirea modelului plecand de la o multime de training = multime de exemple preclasificate
  - Aplicarea modelului unor date neclasificate

EXEMPLE DE APLICATII

Anumite persoane pot fi clasificate

- dupa modul in care se estimeaza ca vor raspunde unei anumite promotii trimise prin email
- dupa modul in care se estimeaza ca isi vor schimba compania de telefonie
- Dupa cum ar fi buni candidati pentru o interventie chirurgicala

MULTIMEA DE TRAINING

Datele pot fi obtinute pentru ca

- Se cunosc anumite informatii anterioare despre obiecte de acelasi fel

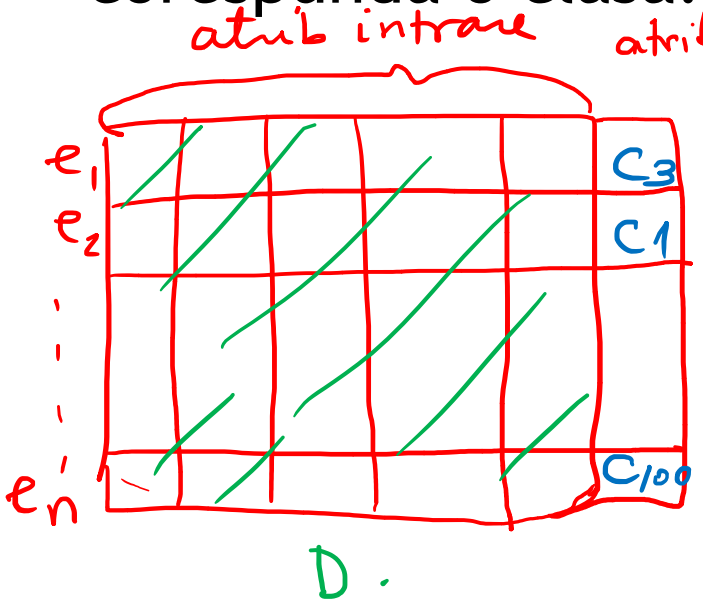
  Ex:
  - Pacienti care au urmat un anume tratament
  - Clienti care au renuntat la serviciile companiei
- Se fac experimente asupra unui esantion din baza de date in urma carora se obtin informatii

Ex:

- se trimit anumite oferte promotionale doar unui esantion din totalul clientilor unei companii si in urma raspunsului acestora la oferta se construieste mt de training
- Un expert realizeaza o preclasificare a unui esantion din baza de date

FORMULAREA PROBLEMEI

- Data o baza de date D={e_1, e_2, …, e_n} si o multime de clase C={c_1, c_2, …, c_m}, a clasifica inseamna a gasi o functie f:D->C astfel incat fiecarui exemplu din baza de date sa ii corespunda o clasa: $f(e_i)=c_j$.



CLASIFICAREA BAYESIANA

- Se presupune ca toate atributele sunt independente unele de altele si la fel de importante pentru realizarea clasificarii.
- Se bazeaza pe **regula lui Bayes:**

- **Date o multime de evenimente observabile :** $X = x_1, x_2, ..., x_n$

si o multime de ipoteze $H = h_1, h_2, ..., h_m$

presupunand ca o singura ipoteza poate sa apara in acelasi timp atunci probab ca o ipoteza *h* sa fie adevarata dat fiind un eveniment observabil *x* este:

$$P(h/x) = \frac{P(x/h)P(h)}{P(x)}$$

unde

P(x) = probab de aparitie a evenimentului x,

P(x/h) = probab ca data ipoteza h, evenimentul x sa apara.

***P(h/x)* se numeste probab aposteriori,**

***P(h)* se numeste probab apriori.**

APLICATIE

- Problema: Pentru acordarea unui credit pentru cumpararea unui produs se cere clientilor sa furnizeze informatiile legate de venitul lunar si indicele de creditare urmand ca institutia care acorda creditul sa decida daca un client dat poate beneficia de acest credit.
- Baza de date:

  atribute:

  ID (numarul de identificare a clientului,

  Venit lunar, Credit(Indice de creditare),

  Clasa - cu valori posibile

  h_1= se acorda creditul,

  h_2 = se acorda creditul dar cu anumite restrictii,

  h_3=nu se acorda creditul.

MULTIMEA DE TRAINING

ID	Venit lunar	Credit	Clasa
1	4800	Excelent	h1
2	2800	Bun	h1
3	1200	Excelent	h1
4	2400	Bun	h1
5	4400	Bun	h1
6	1600	Excelent	h1
7	3200	Nesatisfacator	h2
8	1600	Nesatisfacator	h3
9	2400	Nesatisfacator	h2
10	200	Nesatisfacator	h3

DISCRETIZAREA ATRIBUTULUI VENIT LUNAR

Interval_venit:

- 1 corespunde intervalului [0, 400)
- 2 corespunde intervalului [400, 2000)
- 3 corespunde intervalului [2000, 4000]
- 4 corespunde intervalului [4000, infinit).

NOUA BAZA DE DATE

ID	Venit lunar	Credit	Clasa	Interval_venit
1	4800	Excelent	h1	4
2	2800	Bun	h1	3
3	1200	Excelent	h1	2
4	2400	Bun	h1	3
5	4400	Bun	h1	4
6	1600	Excelent	h1	2
7	3200	Nesatisfacator	h2	3
8	1600	Nesatisfacator	h3	2
9	2400	Nesatisfacator	h2	3
10	200	Nesatisfacator	h3	1

ATRIBUTE

- Atribute de intrare:
  - Interval_venit
  - Credit
- Atribut tinta:
  - Clasa

Data de intrare noua:

$$\{Venit = 5200, Credit = Excelent\}$$

Se calculeaza:

$$P(h_i / \{Interval\_venit = 4, Credit = Excelent\})$$

Si se alege valoarea cu probab cea mai mare.

- Se calculeaza:

$$\max\{P(\{Interval\_venit = 4\}/h_i) \cdot P(\{Credit = Excelent\}/h_i) \cdot P(h_i)/1 \le i \le 3\}$$

- Calculam probabilitatile:

H	P(H)
h_1	6/10=0.6
h_2	2/10=0.2
h_3	2/10=0.2

- Pentru fiecare atribut:

H / Interval_venit	h_1	h_2	h_3
1	0/6=0	0/2=0	1/2
2	2/6	0/2=0	1/2
3	2/6	2/2=1	0/2=0
4	2/6	0/2=0	0/2=0

H / Credit	h_1	h_2	h_3
Excelent	3/6=1/2	0/2=0	0/2=0
Bun	3/6=1/2	0/2=0	0/2=0
Nesatisfacator	0/6=0	2/2=1	2/2=1

- P({int_venit=4, Credit=Exc}/h1}=(2/6)*(3/6)*(6/10)
- P({int_venit=4, Credit=Exc}/h2}=0
- P({int_venit=4, Credit=Exc}/h3}=0
- Rezulta ca clasa care se atribuie datei noi este h_1.

DISCUTIE

- O problema care poate sa apara atunci cand se foloseste clasificarea bayesiana este aceea ca daca multimea de training nu contine exemple cu destul de multe valori posibile pentru atributul tinta atunci probabilitatea ca acele valori sa apara este zero. Acest lucru implica existenta valorii 0 in matricea probabilitatilor conditionate.
- De exemplu, pentru atributul interval_venit, pentru ca nu exista nici un exemplu care sa aiba interval_venit =1 si clasa h1 atunci in matrice avem intrarea 0. La fel si pentru clasa h2.

- Acest lucru inseamna ca din start eliminam posibilitatea ca un client sa poata avea interval_venit = 1 si sa fie in clasele h1 si h2.
- Uneori aceste zerouri sunt justificate dar multimea de training nu ar trebui sa impuna astfel de reguli.
- Pentru a elimina aceasta problema (existenta de 0 in matricea probabilitatilor conditionate) se foloseste estimatorul Laplace.

ESTIMATORUL LAPLACE

- Pentru fiecare atribut, fiecare coloana a matricii probabilitatilor conditionate care contine valoarea 0 se modifica astfel:
  - se adauga 1 la numarator si
  - se adauga k la numitor unde k =numarul de valori posibile ale atributului.

  - Obs. Acest estimator nu este obligatoriu sa fie 1 ci poate fi orice $\lambda > 0$ si atunci se adauga λ la numarator si se adauga $k\lambda$ la numitor. De obicei se foloseste $\lambda=1$.

EXEMPLU

- Pentru atributul Credit, coloana corespunzatoare lui h1, se modifica din:

3/6
3/6
0/6

in

4/9
4/9
1/9

- Pentru atributul Interval_venit, coloana corespunzatoare lui h1, se modifica din:

0/6
2/6
2/6
2/6

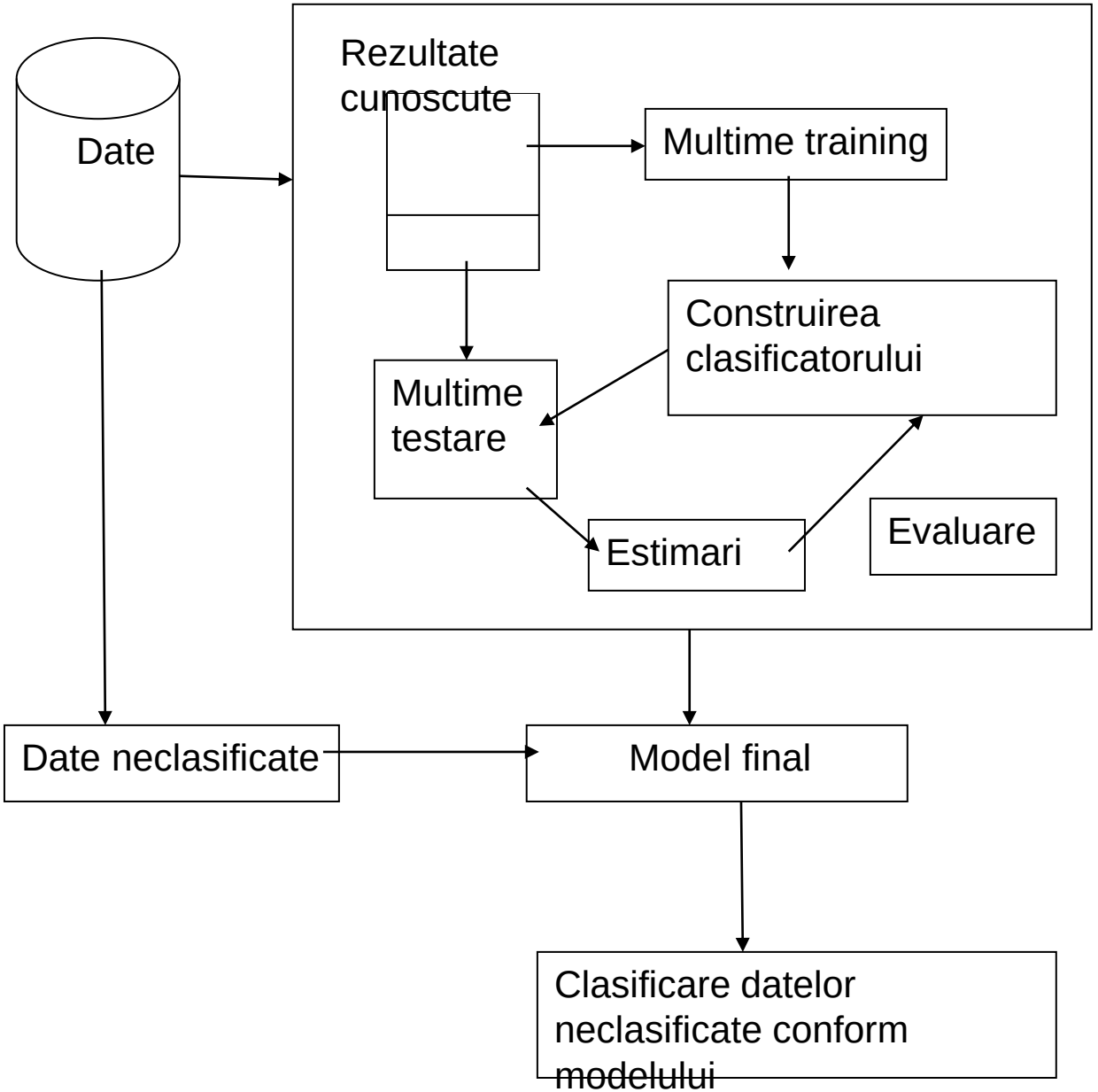
in

1/10
3/10
3/10
3/10

EVALUAREA MODELULUI DE CLASIFICARE

- Cat de bine estimeaza un clasificator atributul tinta?
- Acuratetea unui clasificator= $\frac{nr.exemple\ clasificate\ corect}{nr.total\ exemple}$
- rata erorii= $\frac{nr.exemple\ clasificate\ incorect}{nr.total\ exemple}$
- Acuratetea = 1-rata erorii

PROCESUL CLASIFICARII



- Multimea de training este independenta de multimea de testare.
- De obicei, alegerea multimii de training se face prin alegerea unui esantion din multimea datelor cunoscute
- Acuratetea si rata erorii se determina pe multimea de testare.

MATRICEA DE CONFUZIE (VEZI CONFUSION MATRIX-TANAGRA)

- Daca atributul tinta poate lua valorile

  $\{c_1, c_2, \ldots, c_m\}$, matricea de confuzie este o matrice mxm cu:

  C[i][j]=nr exemple care au sunt clasificate ca avand valoarea c_i iar clasificatorul le-a estimat ca avand fiind in clasa c_j .

  Suma pe diagonala principala = nr de exemple clasificate corect

Classifier performances

Error rate	0,0714
Values prediction	Confusion matrix

Value	Recall	1-Precision
N	0,8000	0,0000
Y	1,0000	0,1000

	N	Y	Sum
N	4	1	5
Y	0	9	9
Sum	4	10	14