# Clasificare

## Obiecte şi caracteristici

- Pentru a desena graficul unei funcţii în scară se utilizează funcţia stairs (x,y) unde x şi y sunt vectori de aceeaşi dimensiune;
  [xb,yb] = stairs (x,y) returnează vectorii xb, yb astfel încât
  plot(xb,yb) să deseneze graficul funcţiei în scară.

```
» x=2:26;
»  y1=[0 4 4 6 2 2 8 6 4 4 8 8 10 10 18 20 20 22 22 10 8 8 8 8 0];
»  y2=[6 8 18 16 12 10 12 20 18 4 2 3 6 4 6 2 4 2 2 8 6 8 6 4 2 ]
»  [xb,yb1]=stairs(x,y1);
» [xb,yb2]=stairs(x,y2);
» plot(xb,yb1,'r--',xb,yb2,'k')
```

- *A* este un vector,
  m = mean(A) returnează media componentelor vectorului *A*;

  să calculăm media lungimilor bibanilor de mare:

```
»A = [3 3 3 3 4 4 4 4 5 5 5 5 5 5 6 6 7 7 8 8 8 8 8 8 8 8 9 9 9 9 9 9 10 10 10 10 11 11 11 11 12 12
12 12 12 12 12 12 13 13 13 13 13 13 13 13 14 14 14 14 14 14 14 14 14 14 15 15 15 15 15 15 15
15 15 15 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 16 17 17 17 17 17 17 17 17 17 17
17 17 17 17 17 17 17 17 17 17 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18 18
19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 20 20 20 20 20 20 20 20 20
20 20 20 20 20 20 20 20 20 20 20 20 21 21 21 21 21 21 21 21 21 21 22 22 22 22 22 22 22 22
23 23 23 23 23 23 23 23 24 24 24 24 24 24 24 24 25 25 25 25 25 25 25 25]
      »mean(A)
```

  *mediana* este numărul real care împarte în două efective egale seria statistică dată, în cazul nostru mulţimea de antrenament, datele fiind ordonate crescător.

med = median(A) returnează mediana componentelor vectorului *A*:

```
»med = median(A)
```

- Matricea clasificării greşite
  - Cum arată matricea clasificării greşite în cazul unei clasificări perfecte?
  - Care este suma elementelor matricei clasificării greşite ?

- Exemple:
  - date nominale: răspunsurile la o anchetă de opinie (foarte mulţumit, mulţumit, nemulţumit, foarte nemulţumit)
  - date nominale numerice: codurile telefonice specifice ale anumitor regiuni.
  - date nominale binare - masculin / feminin
  - date numerice discrete: numărul de consultaţii pe an la care a fost supus un pacient, codul PIN.
  - date numerice continue: presiunea atmosferică

## Clasificare Bayesiană

- Sunteti o persoana importantă in Ministerul Sănătăţii şi o anumita boală este prezentă în rândurile populaţiei în proporţia de o persoană bolnavă la 10 000. Un mare laborator farmaceutic oferă spre vânzare un test de depistare a bolii, nou creat:

- daca o persoană este bolnavă testul este 99% pozitiv
- dacă o persoană este sănătoasă testul este 0.1% pozitiv

Cifrele par a fi excelente, dar pentru comercializare suntem interesați de probabilitatea ca o persoana să fie bolnavă dacă testul este pozitiv

B evenimentul "persoana este bolnavă"

T evenimentul "testul este pozitiv"

$P(B) = 0.0001;\ P(\overline{B}) = 0.9999;\ P(T \mid B) = 0.99;\ P(T \mid \overline{B}) = 0.001$

$$P(B \mid T) = \frac{P(T \mid B) \times P(B)}{P(T \mid B) \times P(B) + P(T \mid \overline{B}) \times P(\overline{B})} = 0.09$$

Avem 9% şanse ca o persoana cu test pozitiv să fie intr-adevăr bolnavă

**Clasificarea Bayesiană naivă**

- Realizați profilul persoanei care ia masa in oraş într-un restaurant sau un fast-food, folosind ca atribute vârsta, dacă este sau nu supraponderal, timpul disponibil (puțin/suficient) şi venitul lunar, folosind următoarea mulțime de antrenament

| | Vârsta | Supraponderal (sau cu tendințe) | Timp disponibil | Venit lunar (lei) | Unde mănâncă |
|---|---|---|---|---|---|
| 1 | 39 | nu | suficient | 1800 | Restaurant |
| 2 | 26 | nu | puțin | 900 | Fast-food |
| 3 | 30 | da | puțin | 1100 | Fast-food |
| 4 | 60 | nu | suficient | 2400 | Restaurant |
| 5 | 28 | nu | suficient | 1700 | Restaurant |
| 6 | 24 | nu | suficient | 800 | Fast-food |
| 7 | 45 | da | puțin | 1900 | Restaurant |
| 8 | 56 | nu | puțin | 2600 | Restaurant |
| 9 | 27 | da | suficient | 2100 | Restaurant |
| 10 | 35 | nu | puțin | 1000 | Fast-food |
| 11 | 32 | nu | suficient | 800 | Fast-food |
| 12 | 59 | da | suficient | 1500 | Restaurant |
| 13 | 42 | nu | puțin | 3600 | Restaurant |
| 14 | 23 | nu | suficient | 700 | Fast-food |
| 15 | 38 | da | suficient | 2700 | Restaurant |
| 16 | 51 | nu | puțin | 1300 | Restaurant |
| 17 | 64 | nu | suficient | 2100 | Restaurant |
| 18 | 27 | da | suficient | 2500 | Restaurant |
| 19 | 23 | da | puțin | 800 | Fast-food |
| 20 | 43 | nu | puțin | 3700 | Restaurant |

Unde va mânca o persoană de 25 ani, care nu este supraponderală şi nu are tendințe, care are timp suficient un un venit lunar de 1900 RON?

$$P(A_k \mid \Omega_i) = \frac{1}{\sqrt{2\pi} \cdot \sigma_{ki}} \cdot \exp\left( -\frac{(A_k - \mu_{ki})^2}{2 \cdot \sigma_{ki}^{\ 2}} \right)$$

- În domeniul bancar avem problema estimării riscului acordării unui credit unei anumite persoane care are următoarele atribute:(datorii puține, garanții adecvate,venit lunar 2000 lei), utilizând urmăoarea mulțime deantrenament

| client | **RISC** | datorii | garanții | venit lunar (lei) |
|---|---|---|---|---|
| 1 | înalt | multe | nu există | 850 |
| 2 | înalt | multe | nu există | 1000 |
| 3 | înalt | puține | nu există | 600 |
| 4 | înalt | puține | nu există | 500 |
| 5 | scăzut | puține | nu există | 1800 |
| 6 | înalt | puține | adecvate | 500 |
| 7 | înalt | puține | nu există | 700 |
| 8 | scăzut | puține | nu există | 1600 |
| 9 | scăzut | puține | nu există | 2800 |
| 10 | scăzut | multe | adecvate | 1100 |
| 11 | înalt | multe | nu există | 500 |
| 12 | înalt | multe | nu există | 600 |
| 13 | scăzut | multe | nu există | 1600 |
| 14 | înalt | multe | nu există | 1400 |
| 15 | înalt | multe | adecvate | 450 |
| 16 | înalt | puține | nu există | 700 |
| 17 | scăzut | puține | adecvate | 1200 |
| 18 | scăzut | puține | adecvate | 3200 |
| 19 | scăzut | puține | adecvate | 1100 |
| 20 | înalt | multe | nu există | 400 |

- O companie doreşte efectuarea unui studiu asupra timpului petrecut de telespectatori privind reclamele televizate la diverse produse şi servicii. Se urmăreşte realizarea profilului telespectatorului care urmăreşte cel puţin jumătate din timpul acordat publicităţii, in cadrul unui program. Desenaţi arborele de clasificare şi decizie corespunzător folosind trei masuri studiate. Avem la dispoziţie următoarea baza de date.

| | Timp vizionare | Venit | Domiciliu(urban/rural) | Vârsta | Salariat/şomer/pensionar |
|---|---|---|---|---|---|
| 1 | <1/2 | 1000 | urban | 25 | salariat |
| 2 | <1/2 | 1800 | urban | 30 | salariat |
| 3 | >1/2 | 800 | rural | 40 | şomer |
| 4 | <1/2 | 2500 | urban | 35 | salariat |
| 5 | <1/2 | 400 | rural | 45 | salariat |
| 6 | >1/2 | 700 | urban | 65 | pensionar |
| 7 | <1/2 | 3500 | urban | 50 | salariat |
| 8 | >1/2 | 300 | rural | 28 | şomer |
| 9 | <1/2 | 500 | urban | 51 | şomer |
| 10 | >1/2 | 1100 | rural | 62 | pensionar |
| 11 | >1/2 | 1500 | rural | 48 | salariat |
| 12 | >1/2 | 900 | urban | 69 | pensionar |
| 13 | <1/2 | 2000 | rural | 41 | salariat |
| 14 | >1/2 | 1000 | urban | 37 | salariat |
| 15 | <1/2 | 450 | rural | 60 | pensionar |
| 16 | <1/2 | 1700 | rural | 41 | salariat |
| 17 | >1/2 | 1300 | urban | 34 | salariat |
| 18 | <1/2 | 400 | urban | 32 | şomer |
| 19 | <1/2 | 1500 | rural | 46 | salariat |

| | | | | | |
|---|---|---|---|---|---|
| 20 | >1/2 | 900 | urban | 71 | pensionar |
| 21 | >1/2 | 600 | urban | 52 | şomer |

- O firmă de asigurări prezintă următoarele date referitoare la asiguraţii Casco; pentru o nouă politică în sensul aplicării unei ponderi clienţilor predispuşi la accidente, construiţi un arbore de clasificare şi decizie pe baza acestor date şi pe baza acestuia deduceţi regulile de clasificare.

| risc | vârsta | tipul maşinii | vechime | a avut accident |
|---|---|---|---|---|
| înalt | peste 30 | sport | sub 2 ani | da |
| moderat | peste 30 | familial | sub 2 ani | nu |
| scăzut | peste 30 | familial | peste 2 ani | nu |
| înalt | sub 30 | sport | peste 2 ani | da |
| înalt | sub 30 | familial | peste 2 ani | da |
| scăzut | peste 30 | familial | peste 2 ani | nu |
| moderat | sub 30 | sport | peste 2 ani | nu |
| moderat | peste 30 | familial | sub 2 ani | da |
| moderat | sub 30 | sport | peste 2 ani | nu |
| înalt | peste 30 | sport | sub 2 ani | da |

- Într-un supermarket, pe parcursul unei zile au loc 100 000 de tranzacţii; dintre acestea 7000 conţin cafea, 1000 conţin lapte condensat şi 10000 conţin biscuiţi. Cafea şi lapte condensat se regăsesc în 800 tranzacţii, biscuiţi şi lapte condensat în 300 tranzacţii, biscuiţi şi cafea în 2000 tranzacţii în timp ce cafea, lapte condensat şi bicuiţi se găsesc în 100 de tranzacţii.

  Calculaţi suportul, încrederea şi diferenţa de nivel pentru regulile cu trei articole, respectiv două articole, deduse din datele prezentate. Deduceţi care este regula cea mai bună.

  $$\text{Confidence} = \frac{frecventa\ \ aparitiilor}{frecventa\ \ antecedent};$$

  $$\text{Diferenţa de nivel} = \frac{confidence}{frecventa\ \ consecinta}$$

- Aplicaţi metoda k-nearest neighbor luând ca mulţime de antrenament aceasta bază de date restrânsă de irişi, în stabilirea tipului de iris ce are următoarele caracteristici (4,15,25,49)

  **Tipul de Iris**:0 Setosa; 1 Virginica; 2 Versicolor
  PW laţimea petalei; PL lungimea petalei;
  SW lăţimea sepalei; SL lungimea sepalei

| Type | PW | PL | SW | SL |
|---|---|---|---|---|
| 0 | 2 | 14 | 33 | 50 |
| 1 | 24 | 56 | 31 | 67 |
| 1 | 23 | 51 | 31 | 69 |
| 0 | 2 | 10 | 36 | 46 |
| 1 | 20 | 52 | 30 | 65 |
| 1 | 19 | 51 | 27 | 58 |
| 2 | 13 | 45 | 28 | 57 |
| 2 | 16 | 47 | 33 | 63 |
| 1 | 17 | 45 | 25 | 49 |

| | | | | |
|---|---|---|---|---|
| 2 | 14 | 47 | 32 | 70 |
| 0 | 2 | 16 | 31 | 48 |
| 1 | 19 | 50 | 25 | 63 |
| 0 | 1 | 14 | 36 | 49 |
| 0 | 2 | 13 | 32 | 44 |
| 2 | 12 | 40 | 26 | 58 |
| 1 | 18 | 49 | 27 | 63 |
| 2 | 10 | 33 | 23 | 50 |
| 0 | 2 | 16 | 38 | 51 |
| 0 | 2 | 16 | 30 | 50 |
| 1 | 21 | 56 | 28 | 64 |
| 0 | 4 | 19 | 38 | 51 |
| 0 | 2 | 14 | 30 | 49 |
| 2 | 10 | 41 | 27 | 58 |
| 2 | 15 | 45 | 29 | 60 |
| 0 | 2 | 14 | 36 | 50 |
| 1 | 19 | 51 | 27 | 58 |
| 0 | 4 | 15 | 34 | 54 |
| 1 | 18 | 55 | 31 | 64 |
| 2 | 10 | 33 | 24 | 49 |
| 0 | 2 | 14 | 42 | 55 |
| 2 | 14 | 39 | 27 | 52 |
| 2 | 12 | 39 | 27 | 58 |
| 1 | 23 | 57 | 32 | 69 |
| 2 | 15 | 42 | 30 | 59 |
| 2 | 13 | 44 | 23 | 63 |

- Considerăm un lot de 14 pacienţi dintre care 6 au cancer hepatic (HCC). Pentru a decide dacă un pacient are sau nu cancer hepatic, studiile clinice arată că se iau în considerare anumite enzime serice (de obicei un număr de 15). Pentru simplificarea calculelor, le vom lua în considerare pe cele mai importante din punct de vedere clinic: alkaline phosphatase (FA), g-glutamyl transferase (gGT), leucine amino peptidase (LAP) şi colesterol (C)

| | FA | gCT | LAP | C | diagnostic |
|---|---|---|---|---|---|
| P1 | 162 | 255 | 74 | 258 | non HCC |
| P2 | 210 | 324 | 93 | 220 | non HCC |
| P3 | 259 | 208 | 115 | 266 | non HCC |
| P4 | 120 | 114 | 89 | 171 | non HCC |
| P5 | 246 | 173 | 98 | 210 | non HCC |
| P6 | 138 | 189 | 76 | 165 | non HCC |
| P7 | 132 | 177 | 48 | 138 | non HCC |
| P8 | 152 | 183 | 105 | 178 | non HCC |
| P9 | 422 | 488 | 183 | 292 | HCC |
| P10 | 607 | 259 | 65 | 275 | HCC |
| P11 | 446 | 283 | 176 | 309 | HCC |
| P12 | 460 | 1053 | 145 | 221 | HCC |
| P13 | 680 | 381 | 280 | 275 | HCC |
| P14 | 561 | 450 | 180 | 164 | HCC |

Utilizând metoda k-nearest neighbor stabiliţi diagnosticul următorilor pacienţi:
P’ (180,119,114,171) şi P’’ 236,270,88,150)