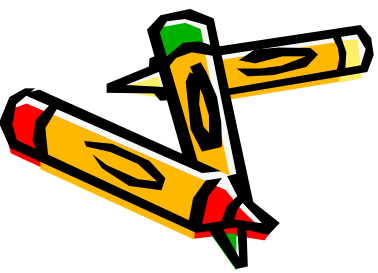
Clasificarea bayesiana

Marina Gorunescu
mgorun@inf.ucv.ro

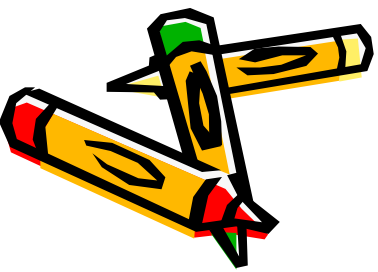
risc prognozat

Strategiile de decizie Bayes sunt folosite în scopul de a minimiza *„riscul prognozat"* .
Se aplică în probleme de clasificare cu număr mare de clase.



Termenul de *şansă* exprimă un grad subiectiv de încredere într-un rezultat particular.

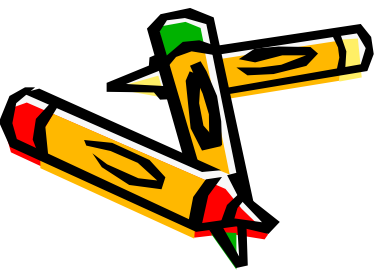
Abordarea matematică a noțiunii de şansă a dat naştere termenului de ***probabilitate***



probabilitatea conditionata

Probabilitatea condiționată este folosită pentru a măsura încrederea că un eveniment aleator va avea loc știind că alt eveniment aleator a avut loc.

Fie două evenimente A și B, probabilitatea condiționată ca evenimentul A să aibă loc știind ca evenimentul B a avut loc este $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$



- studenții promovează examenele A și B cu probabilitățile: $A \sim 70\%$, $B \sim 40\%$ şi $A \cap B \sim 35\%$.

- Probabilitatea condiționată ca studenții să promoveze cele două examene, dacă au promovat examenul A este

$$P(A \cap B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{1}{2}.$$

- Probabilitatea condiționată ca studenții să promoveze cele două examene, dacă au promovat examenul B este

$$P(A \cap B \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{7}{8}$$

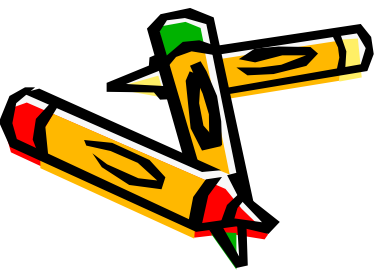
- Pentru a obține probabilitatea ca studenții să promoveze cele două examene, dacă au promovat cel puțin un examen calculăm

probabilitatea ca să promoveze cel puțin un examen:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.7 + 0.4 - 0.35 = 0.75$$

și astfel

$$P(A \cap B \mid A \cup B) = \frac{P(A \cap B)}{P(A \cup B)} = \frac{7}{15}$$



- Fie $\mathbf{X}$ o bază de date medicale pentru pacienții cu diferite boli ale ficatului.
  - $\mathbf{x}_n$ este vectorul caracteristicilor medicale corespunzător pacientului cu numărul *n*, pacient ce este diagnosticat cu cancer hepatic,
  - $\mathbf{x}$ este vectorul caracteristicilor medicale al unui pacient ce nu a fost încă diagnosticat.

  Probabilitatea ca pacientul căruia nu i s-a pus încă un diagnostic să aibă cancer hepatic dacă pacientul cu numărul *n* are este $P(\mathbf{x}|\mathbf{x}_n)$.

formula Bayes

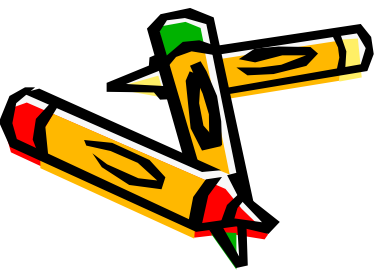
- Fie $(\Omega, \Sigma, \mathbf{P})$ un *spațiu de probabilitate*, B un eveniment arbitrar din Σ și $\{\mathbf{A}_1, \ldots, \mathbf{A}_n\}$ o partiție a spațiului Ω. Atunci:

$$P(A_i \mid B) = \frac{P(B \mid A_i) \cdot P(A_i)}{P(B)} = \frac{P(B \mid A_i) \cdot P(A_i)}{\sum_{i=1}^{n} P(B \mid A_i) \cdot P(A_i)},$$

$$P(B) > 0, \quad P(A_i) > 0, i = 1, .., n$$

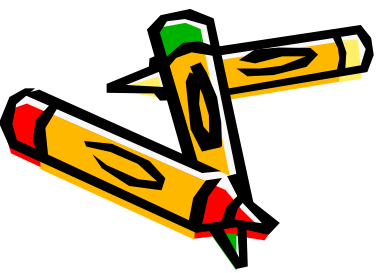
denumiri uzuale

- $P(A_i \mid B)$ - probabilitate *posterioară* (posterior probability),
- $P(A_i)$ - probabilitate *apriorică* (prior probability),
- $P(B \mid A_i)$ - *verosimilitate* (likelihood),
- iar $P(B)$ - *evidență/ dovadă* (evidence).



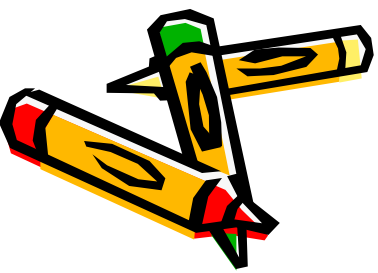
formula Bayes

$$probabilitate\ posterioara = \\ = \frac{verosimilitate \times probabilitate\ apriorica}{evidenta}$$



exemple

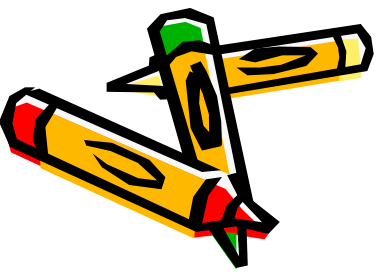
- In secția de gastroenterologie dintr-un spital sunt internați 49% bărbați și 51% femei.
  Din experiența anterioară estimăm că 2.5% dintre bărbați și 1.9% din femei au cancer hepatic.
  Notații:
  M- bărbați, F-femei
  A evenimentul ca diagnosticul să fie cancer hepatic



probabilitatea ca un individ, arbitrar ales, să fie diagnosticat cu cancer hepatic se calculează cu formula probabilității totale:

$$P(A) = P(A|M) \cdot P(M) + P(A|F) \cdot P(F) =$$
$$= 0.025 \cdot 0.49 + 0.09 \cdot 0.51 = 0.022 = 2.2\% ,$$

2.2% dintre bolnavii internați în secție au acest diagnostic



unde

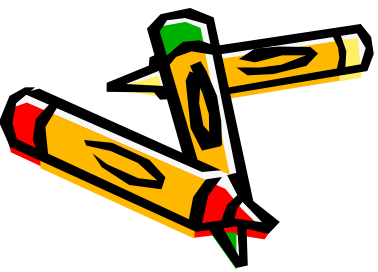
- $P(A|M)$ - probabilitatea ca o persoană să fie diagnosticată cancer hepatic, condiționată de faptul că este bărbat;
- $P(A|F)$ - probabilitatea ca o persoană să fie diagnosticată cancer hepatic, condiționată de faptul că este femeie;
- $P(M)$ proporția pacienților bărbați din totalul pacienților;
- $P(F)$ proporția pacienților femei din totalul pacienților.

- Folosind formula Bayes, vom calcula probabilitatea ca un bolnav diagnosticat cu cancer hepatic să fie bărbat, respectiv femeie:

$$P(M \mid A) = \frac{P(A \mid M) \cdot P(M)}{P(A)} = 0.5583 = 55.83\%$$

$$P(F \mid A) = \frac{P(A \mid F) \cdot P(F)}{P(A)} = 0.4417 = 44.17\%$$

Dintre bolnavii cu acest diagnostic 55.83% sunt bărbați, respectiv 44.17% sunt femei.

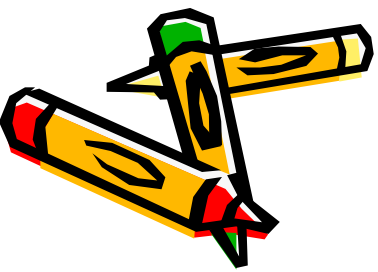


- Considerăm că într-o companie se produc aceleași produse în trei unități distincte U_1, U_2, U_3, ce au capacitățile de producție 60%, 30%, 10%, procente ce reprezintă probabilitățile ca un produs să provină de la una dintre cele trei unități.
  Fiecare unitate are rata de a produce obiecte cu defecțiuni de 6%, 3%, 5%.
  Probabilitatea ca un produs defect, arbitrar ales, să provină de unitatea U_1, respectiv U_2 sau U_3.

Notăm cu *A* evenimentul ca un produs ales la întâmplare să fie defect. Calculăm care este probabilitatea ca un produs arbitrar ales să fie defect.

$$P(A) = P(A|U_1) \cdot P(U_1) + P(A|U_2) \cdot P(U_2) +$$
$$+ P(A|U_3) \cdot P(U_3) = 0.06 \cdot .6 + 0.03 \cdot 0.3 + 0.05 \cdot 0.1 = 0.05$$

(formula probabilității totale)



calculăm probabilitatea ca un produs defect să provină din U_i:

$$P(U_1 \mid A) = \frac{P(A \mid U_1) \cdot P(U_1)}{P(A)} = \frac{0.06 \cdot 0.6}{0.05} = 0.72 = 72\%.$$

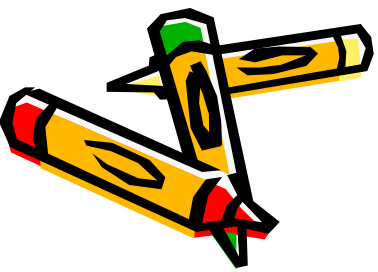
$$P(U_2 \mid A) = \frac{P(A \mid U_2) \cdot P(U_2)}{P(A)} = 0.18 = 18\%$$

$$P(U_3 \mid A) = \frac{P(A \mid U_3) \cdot P(U_3)}{P(A)} = 0.1 = 10\%$$

(formula Bayes)

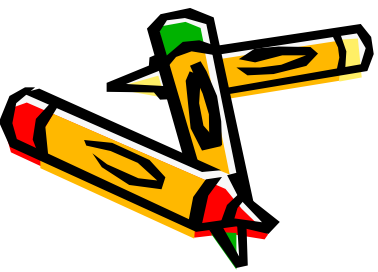
Prezentăm o problemă medicală simplificată:

- Testul unei boli rare este corect în proporție de 99%.
- O persoana din 100 000 prezintă această boală.
- Efectuați testul și răspunsul este pozitiv



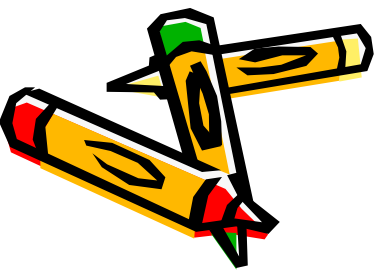
Se afirmă că din 1 000 000 de persoane, 10 000 (1%) vor fi considerate a fi posibili bolnavi, în timp ce doar 10 persoane (1 la 10 000) au intr-adevăr aceasta boală Așadar acest test, cu fiabilitatea de 99%, in cazul în care este pozitiv, dă 999 alerte false din 1000.

Nici un test nu este perfect și o problema serioasă o constituie rezultatele fals pozitive.



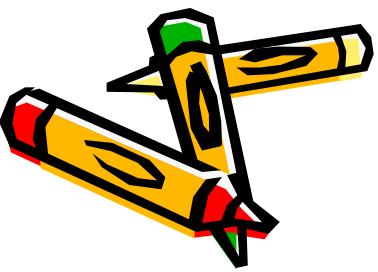
În general, în cazul in care o persoană este testată dacă suferă de o anumită boală, riscul ca rezultatul să fie pozitiv, dacă persoana este sănătoasă, este infim.

Problema este să determinăm în cazul unei boli rare probabilitatea ca un test pozitiv să fie greşit.



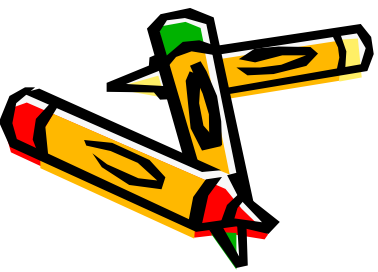
Considerăm cazul unei boli rare și aplicăm un test foarte fiabil:

- dacă pacientul a contractat boala, testul este pozitiv în 99% din cazuri, adică cu o probabilitate de 0.99.
- dacă pacientul este sănătos, testul este corect, adică negativ în 95% din cazuri, (cu o probabilitate de 0.95).



Boala atinge o persoană din 1000, deci are probabilitatea de 0.001 (pare mică dar în cazul unei boli mortale este considerabilă).

Avem toate informațiile pentru a determina probabilitatea ca testul să fie fals pozitiv

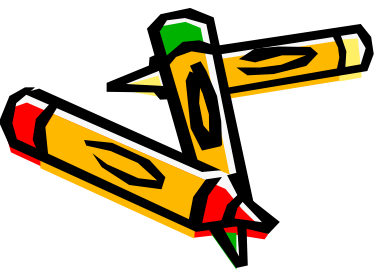


A evenimentul: "pacientul a contractat boala"

B evenimentul " testul este pozitiv"

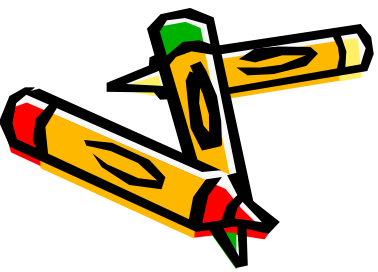
$$P(A \mid B) = \frac{0.99 \times 0.001}{0.99 \times 0.001 + 0.05 \times 0.999} = 0.019$$

Ştiind că testul este pozitiv, probabilitatea ca pacientul să fie sănătos este $1 - 0.019 = 0.981$



Dacă tratamentul este complicat, costisitor sau periculos, pentru un pacient sănătos este nevoie de un test complementar, care va fi sigur mai precis şi mai costisitor.

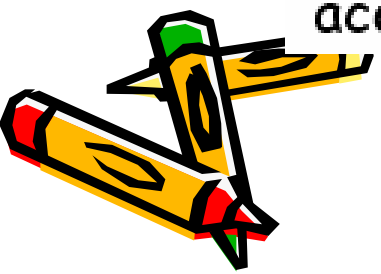
Primul test a eliminat cazurile cele mai evidente.



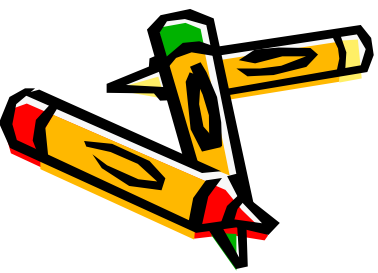
problema juridica si sociala

Dacă probabilitatea unui anumit tip de comportament, să zicem **delicvența** depinde de anumiți factori sociali, culturali sau ereditari, atunci:

- Aceasta presupune o reducere parțială a responsabilității morale şi juridice a delincventului, ceea ce antrenează o creştere a responsabilității societății, care nu a ştiut sau nu a reuşit să neutralizeze aceşti factori.

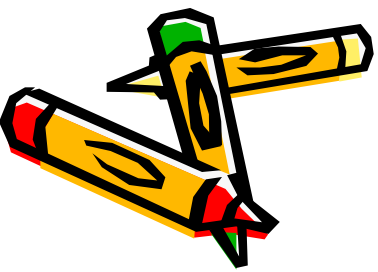


- Pe de altă parte această informație poate fi utilizată pentru ca politica de prevenție să fie orientată corespunzător şi trebuie văzut dacă interesul public sau morala se va acomoda la aceasta discriminare de facto a cetățenilor, chiar dacă este pozitivă.



riscul asteptat

Scopul, în teoria deciziilor, este de a minimiza probabilitatea de a greși sau *riscul așteptat*.



exemplu

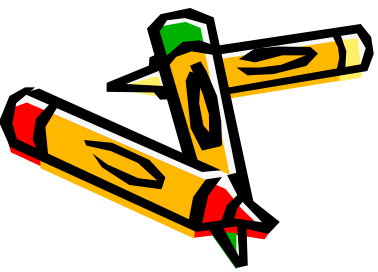
- Să considerăm un set de date medicale, corespunzătoare unei mulțimi $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ de pacienți.
  În urma analizelor avem două rezultate posibile: benign sau malign, ceea ce corespunde la două clase Ω_1 și Ω_2. Dacă mulțimea $\mathbf{X}$ este suficient de mare, definim probabilitățile apriorice $P(\Omega_1)$ și $P(\Omega_2)$.
  Dacă afirmația „pacientul aparține clasei Ω_1 (*benign*)" apare în 9 cazuri din 10, avem $P(\Omega_1) = 0.9$ și $P(\Omega_2) = 0.1$.

Încercând să clasificăm pacienții cunoscând doar probabilitățile apriorice, conform regulii:

$\mathbf{x} \in \Omega_1$ dacă $P(\Omega_1) > P(\Omega_2)$,

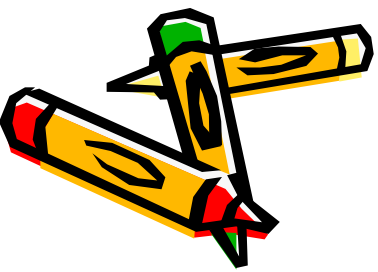
rezultatul nu este mulțumitor.

Conform regulii de mai sus orice nou pacient va fi clasificat în principiu ca fiind benign, cu toate că știm că un caz din 10 este malign.



regula de decizie bayesiana

- Fie D_i regula de decizie referitoare la clasa Ω_i .
- Fiind dat un vector $\mathbf{x}$, eroarea relativă la clasa Ω_i este definită de $P\{eroare/\mathbf{x}\} = 1 - P(\Omega_i|\mathbf{x})$.
- Se *minimizează* probabilitatea de a greşi.



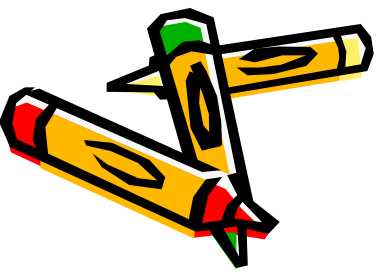
- Regula bayesiană de decizie este:

  Alege D_j dacă

  $$P(\Omega_j|\mathbf{x}) > P(\Omega_i|\mathbf{x}),\ i \in \{1,\ldots,j-1,j+1,\ldots r\}$$

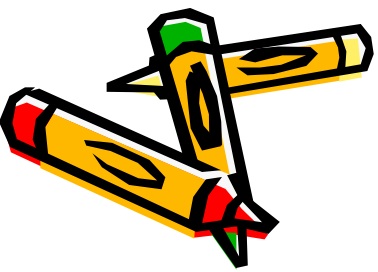
  sau echivalent

  $$P(\mathbf{x}|\Omega_j)\cdot P(\Omega_j) > P(\mathbf{x}|\Omega_i)\cdot P(\Omega_i)\ ,i \in \{1,\ldots,j-1,j+1,\ldots r\}.$$



Să considerăm un set de date care urmează a fi clasificate utilizând un clasificator bayesian; presupunem că fiecare atribut (inclusiv atributul corespunzător etichetei de clasă) este o variabilă aleatoare.

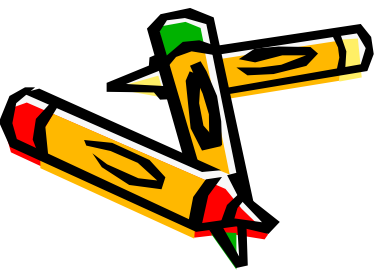
Fiind dat un obiect cu atributele $\{A_1, A_2, \ldots, A_p\}$ ne propunem clasificarea sa în clasa Ω_i.



Clasificarea este corectă atunci când probabilitatea condiționată:

$$P(\Omega_i \mid A_1, A_2, \ldots, A_p)$$

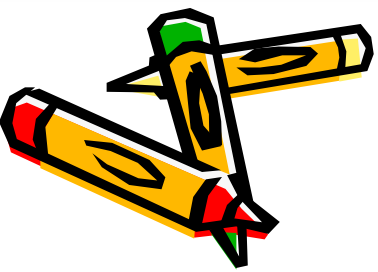
este maximă.



Problema concretă: a estima direct din date această probabilitate, în vederea maximizării sale.

- Se calculează probabilitățile posterioare $P(\Omega_i \mid A_1, A_2, \ldots, A_p)$ pentru toate clasele Ω_i, utilizând formula:

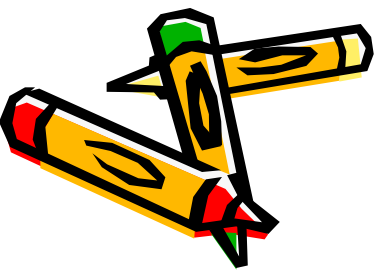
$$P(\Omega_i \mid A_1, A_2, \ldots, A_p) = \frac{P(A_1, A_2, \ldots A_p \mid \Omega_i) \cdot P(\Omega_i)}{P(A_1, A_2, \ldots A_p)}$$



Se alege apoi clasa Ω_j care maximizează

$P(\Omega_i \mid A_1, A_2, \ldots, A_p)$.

adică clasa Ω_j care maximizează $P(A_1 A_2 \ldots A_p \mid \Omega_i) \cdot P(\Omega_i)$.



clasificarea naiva Bayes

Naive Bayes presupune, de foarte multe ori fără niciun temei, independența evenimentelor.

În cazul de față, vom presupune independența reciprocă a atributelor. (ipoteză neadevărată de cele mai multe ori) pentru o anumită clasă Ω_i , adică:

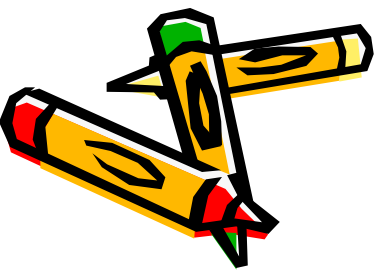
$$P(A_1, A_2, \ldots, A_p \mid \Omega_i) =$$

$$= P(A_1 \mid \Omega_i) \cdot P(A_2 \mid \Omega_i) \cdot \ldots \cdot P(A_p \mid \Omega_i)$$

Vom estima apoi probabilitățile $P(A_k \mid \Omega_i)$ pentru toate atributele A_k și clasele Ω_i, astfel încât un obiect nou, necunoscut, va fi clasificat în clasa Ω_j dacă probabilitatea corespunzătoare acestei clase:

$$P(\Omega_j) \cdot \prod_{k=1}^{p} P(A_k \mid \Omega_j),$$

este maximă față de celelalte.



exemple

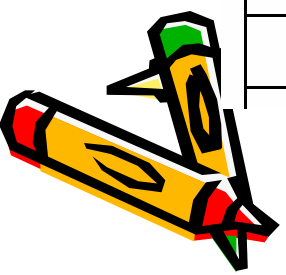
- Domeniul bancar, problema estimării riscului acordării unui credit unei anumite persoane

Folosim clasificarea bayesiană, având în vedere următoarele atribute:

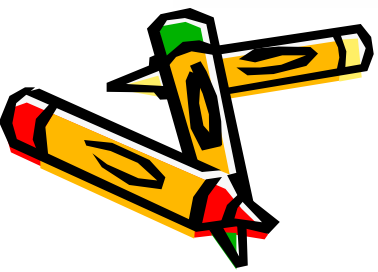
- datoria curentă,
- venit lunar
- garanții.

multime de antrenament

client	RISC	datorii	garanții	venit lunar
1	înalt	multe	nu există	850 RON
2	înalt	multe	nu există	1000 RON
3	înalt	puține	nu există	600 RON
4	înalt	puține	nu există	500 RON
5	scăzut	puține	nu există	1800 RON
6	înalt	puține	adecvate	500 RON
7	înalt	puține	nu există	700 RON
8	scăzut	puține	nu există	1600 RON
9	scăzut	puține	nu există	2800 RON



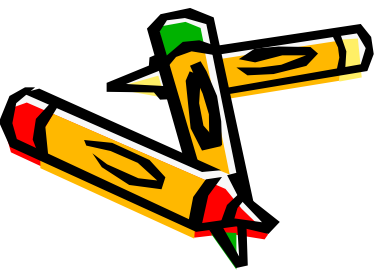
10	scăzut	multe	adecvate	1100 RON
11	înalt	multe	nu există	500 RON
12	înalt	multe	nu există	600 RON
13	scăzut	multe	nu există	1600 RON
14	înalt	multe	nu există	1400 RON
15	înalt	multe	adecvate	450 RON
16	înalt	puține	nu există	700 RON
17	scăzut	puține	adecvate	1200 RON
18	scăzut	puține	adecvate	3200 RON
19	scăzut	puține	adecvate	1100 RON
20	înalt	multe	nu există	400 RON



Există două clase distincte:
risc înalt și risc scăzut din punctul de vedere al riscului acordării unui credit
Probabilitățile celor două clase sunt:

$$P(\text{risc înalt}) = \frac{12}{20}$$

$$P(\text{risc scăzut}) = \frac{8}{20}.$$

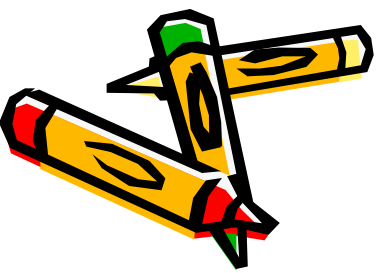


Probabilitățile condiționate de tipul $P(A_k \mid \Omega_i)$

- în cazul atributelor discrete, se vor calcula în mod natural după formula:

$$P(A_k \mid \Omega_i) = \frac{|A_{ki}|}{N_{\Omega_i}},$$

$|A_{ki}|$ reprezintă numărul instanțelor având atributul A_k și care aparțin clasei Ω_i.

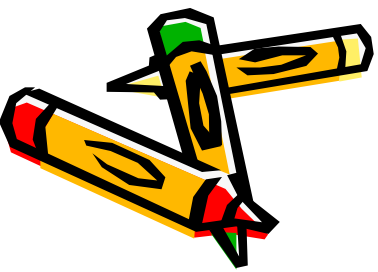


$$P(\text{datorii = puține } | \text{risc înalt}) = \frac{5}{12}$$

$$P(\text{datorii = puține } | \text{risc scăzut}) = \frac{6}{8}$$

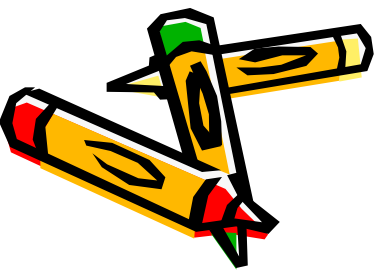
$$P(\text{garanții = adecvate} | \text{risc înalt}) = \frac{2}{12}$$

$$P(\text{garanții = adecvate} | \text{risc scăzut}) = \frac{4}{8}.$$



În cazul atributelor de tip continuu, pentru a evalua probabilitățile condiționate $P(A_k \mid \Omega_i)$, este nevoie de identificarea tipului de repartiție a atributului, privit ca variabilă aleatoare continuă.

De obicei, se presupune că toate atributele continue urmează legea normală, urmând ca din date să se estimeze parametrii acesteia (media și dispersia).



Odată densitatea de repartiție estimată, putem evalua probabilitatea condiționată $P(A_k \mid \Omega_i)$ pentru fiecare clasă în parte.

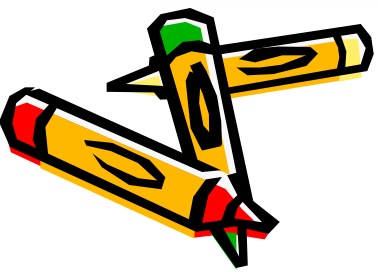
Atributul *Venit lunar* este considerat variabilă aleatoare continuă, de densitate:

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

Să analizăm acum modul de funcționare a clasificatorului astfel construit pe un caz nou:
unui individ care are următoarele atribute:

- datorii puține
- garanții adecvate
- venit lunar 2000 RON

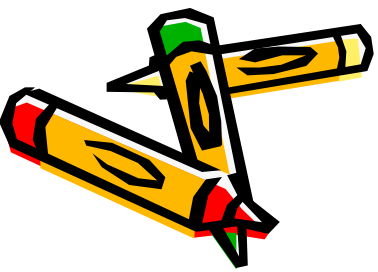
îi acordăm sau nu credit.



$$P(\text{venit lunar} = 2000 \mid \text{risc inalt})$$

$$= \frac{1}{\sqrt{2\pi} \cdot s1} \cdot \exp\left(- \frac{(2000 - m1)^2}{2 \cdot v1} \right) = 3.1803 \cdot 10^{-8}.$$

$$P(\text{venit anual} = 2000 \mid \text{risc scăzut}) = 4.8847 \cdot 10^{-4}$$



$$P(\text{datorii puține, garanții adecvate, venit} = 2000 | \text{risc înalt}) \times$$

$$\times P(\text{risc înalt}) = P(\text{datorii} = \text{puține} | \text{risc înalt}) \times$$

$$\times P(\text{garanții} = \text{adecvate} | \text{risc înalt}) \times$$

$$\times P(\text{venit anual} = 2000 | \text{risc inalt}) \times P(\text{risc înalt}) = 1.3251 \cdot 10^{-9}$$

$$P(\text{datorii puține, garanții adecvate, venit} = 2000 | \text{risc scăzut})$$

$$\times P(\text{risc scăzut}) = P(\text{datorii} = \text{puține} | \text{risc scăzut})$$

$$\times P(\text{garanții} = \text{adecvate} | \text{risc scăzut}) \times$$

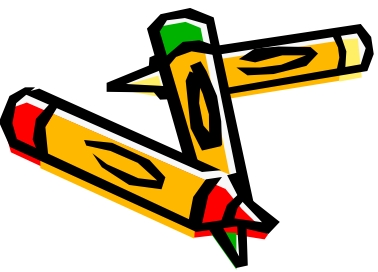
$$\times P(\text{venit anual} = 2000 | \text{risc scăzut}) \times P(\text{risc scăzut}) = 7.3271 \cdot 10^{-5}$$

Deoarece

P(datorii puține, garanții adecvate, venit = 2000 |risc scăzut)>

> P(datorii puține, garanții adecvate, venit = 2000 |risc înalt)

rezultă că banca poate să-i acorde împrumutul solicitat.



Considerând un individ care are următoarele atribute:

- datorii puține
- garanții adecvate
- venit lunar 600 RON

vom avea:

$$P(\text{datorii puține, garanții adecvate, venit} = 600 \mid \text{risc înalt}) \times$$

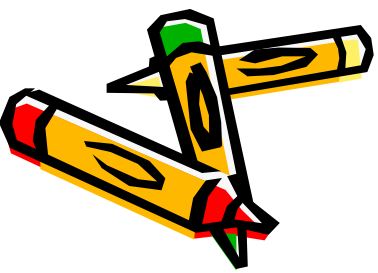
$$\times P(\text{risc înalt}) = 5.5936 \cdot 10^{-5}$$

$$P(\text{datorii puține, garanții adecvate, venit} = 600 \mid \text{risc scăzut}) \times$$

$$\times P(\text{risc scăzut}) = 2.3937 \cdot 10^{-5}$$

clasificarea naiva a textelor

Prin ipoteză, cuvintele ce apar în text se consideră a fi independente.



clasificarea documentelor

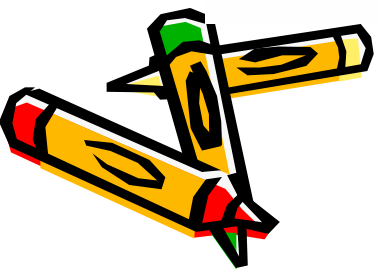
- ***Clasificarea ierarhică*** - folosită dacă se urmăreşte o analiză în detaliu a datelor. Această metodă corespunde aşa numitului „hard clustering", adică se acceptă o singură posibilitate de apartenenţă la o clasă (categorie).
- ***Clasificarea non- ierarhică*** - metodă mult mai rapidă, utilizată pentru baze mari de date. Această metodă este de tip „soft clustering", în sensul că în loc să accepte apartenenţa la o singură clasă, furnizează probabilitatea de apartenenţă a unui element la o anumită clasă (fiecare clasă va avea o anumită pondere de apartenenţă la ea).

clasificarea bayesiana naiva (soft clustering)

Să presupunem că avem r categorii (clase) de documente $\Omega = \{\Omega_1, \ldots, \Omega_r\}$.

A determina cărei categorii îi corespunde documentul D înseamnă a estima probabilitatea $P(\Omega_i \mid D)$ de apartenență a documentului D la clasa Ω_i, utilizând formula Bayes:

$$P(\Omega_i \mid D) = \frac{P(D \mid \Omega_i) \cdot P(\Omega_i)}{P(D)}.$$



$P(D \mid \Omega_i)$ este probabilitatea ca fiind dată clasa Ω_i, cuvintele din D să fie asociate cu această clasă.

Dacă documentul D este format din cuvintele $w_1,\ldots,w_m$, $P(D \mid \Omega_i)$ este probabilitatea apariției fiecărui cuvânt $w_j, 1 \le j \le m$ în clasa Ω_i, probabilitate ce poate fi calculată pe baza ipotezei de independență a variabilelor:

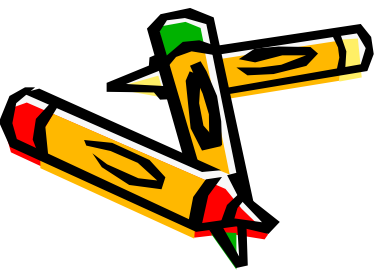
$$P(D \mid \Omega_i) = P(w_1 \mid \Omega_i) \cdot \ldots \cdot P(w_m \mid \Omega_i)$$

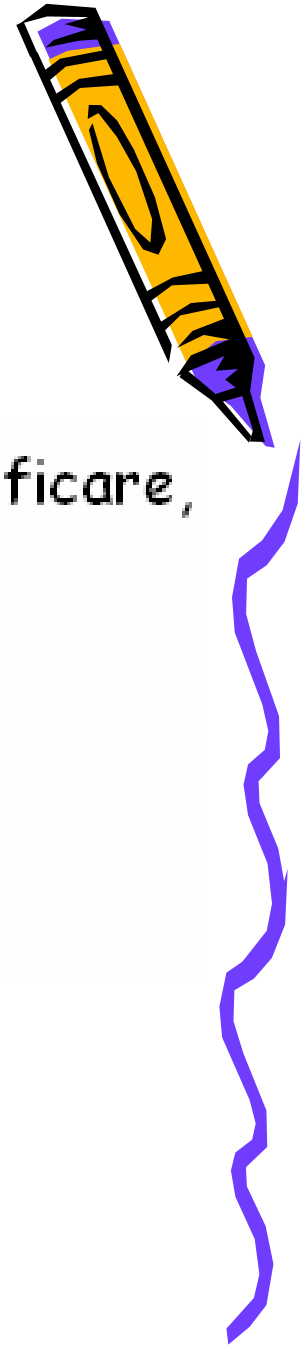
Notând cu n_{ij} numărul de apariții ale cuvântului ω_j în clasa Ω_i și cu n_i numărul de cuvinte din clasa Ω_i, calculăm:

$$P(\omega_j \mid \Omega_i) = \frac{n_{ij}}{n_i}$$

Notând cu n numărul total de cuvinte din Ω, avem:

$$P(\Omega_i) = \frac{n_i}{n}.$$



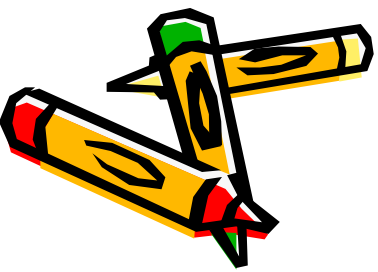


Există un program în MATLAB pentru această clasificare, program pentru se construiesc funcțiile:

parsefile,

addwords,

classify

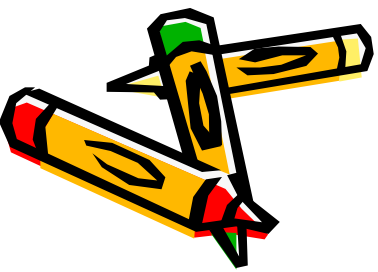


parsefile

- parsefile permite extragerea conținutului util dintr-un document.
  Funcția analizează textul, extrage cuvintele, ce vor fi puse în wordsfiles, însoțite de numărul lor de apariții în documentul *D*.
  Se obține un tabel cu cuvintele din document și cu numărul lor de apariții (valorile asociate lor).
  Funcția este folosită atât în timpul antrenamentului, cât și în clasificarea propriu-zisă.

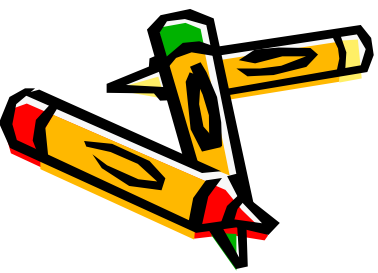
adwords

- addwords permite să adăugăm cuvinte şi valorile asociate lor. Este utilizată la antrenarea clasificatorului



classify

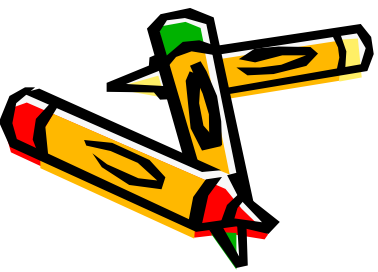
- classify face propriu-zis clasificarea, apelând la parsefile şi calculează toate probabilităţile ce apar în formula Bayes.
  Relaţia: $P(D \mid \Omega_i) = P(\omega_1 \mid \Omega_i) \cdot \ldots \cdot P(\omega_m \mid \Omega_i)$
  se logaritmează, transformând astfel produsul în sumă:
  $\log P(D \mid \Omega_i) = \log(P(\omega_1 \mid \Omega_i)) + \ldots + \log(P(\omega_m \mid \Omega_i))$
  Astfel se reduc calculele ce le va face clasificatorul şi se limitează erorile de "overflow".



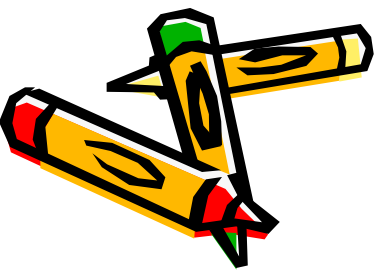
Într-o primă etapă cuvintele care nu apar în nicio clasă nu vor fi luate în considerare în calcule; unui asemenea cuvânt nu i se atribuie valoarea 0 (numărul de apariții) ci 0.1.

Antrenamentul este faza cea mai importantă, performanțele clasificatorului fiind determinate de gradul de învățare.

O temă interesantă ar fi adaptarea şi eventual îmbunătăţirea acestui program pentru texte în limba română, considerând clasele: politic, economic, administrativ, internaţional, sport, arte etc.



Această tehnică a fost utilizată pentru blocarea spam-urilor. Graham P., (http./www. paulgraham.com/better.html) folosind aceasta metodă, cu câteva modificări, a reuşit să stopeze 99.5% din spam-uri cu o eroare de clasificare mai mică de 0.03%.



avantaje

Clasificarea bayesienă (naivă) prezintă o serie de avantaje:

- Este robustă în ceea ce priveşte izolarea zgomotului din date;
- În cazul valorilor lipsă, ignoră obiectul respectiv în timpul estimării probabilităților;
- Este robustă la atributele irelevante.