7. Clasificatori statistici

7.1. Clasificatorul bazat pe metrica Mahalanobis

Parte din limitările specifice clasificatorului de tip distanţă minimă, ce utilizează metrica euclidiană, pot fi înlăturate utilizându-se pentru aceasta o altă metrică, respectiv, metrica Mahalanobis. În acest mod noul clasificator, bazat pe metrica Mahalanobis, va depăşi problemele generate de scalarea setului de date şi/sau cele generate de existenţa trăsăturilor corelate.

7.1.1. Distanţa standardizată

1. Caracterizarea statistică a unei clase

Să considerăm procesul aleator x caracterizat de un set de N realizări particulare ale acestuia: a^1, a^2, …, a^N. În continuare vom considera că toate aceste realizări particulare aparţin aceleiaşi clase. După cum s-a arătat anterior variabila aleatoare x poate fi caracterizată de doi parametri statistici fundamentali: media şi varianţa. Media realizărilor particulare ale variabilei aleatoare x se estimează cu ajutorul relaţiei (5.229) sau a relaţiei (5.230) şi poate fi scrisă:

$$\hat{m}_x = \frac{1}{N}\left(a^1 + a^2 + \cdots + a^N\right) \quad (7.1)$$

Dacă setul de date, a^1, a^2, …, a^N, aparţine aceleiaşi clase (condiţie satisfăcută în cazul nostru) valoarea medie este aproximativ centrul clasei respective, vezi **Figura 7.1**. În această situaţie spunem că media clasei este o valoare tipică pentru clasa respectivă. În cazul clasificatorului de tip minimă distanţă vectorul mediu era considerat vectorul prototip al clasei (pentru x, variabilă aleatoare, media reprezenta valoarea prototip).

Varianţa (momentul centrat de ordin doi al variabilei aleatoare x) este şi ea o măsură ce caracterizează clasa respectivă însă din punct de vedere al mărimii clasei respective, al extinderii ei spaţiale; cu alte cuvinte, varianţa

reflectă dispersia clasei sau cât de departe poate fi o valoare particulară a lui *x* faţă de cea tipică, dată de media clasei.

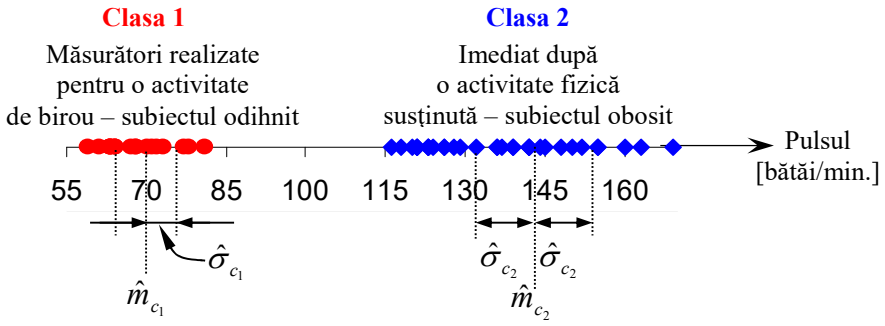


Figura 7.1. Media şi deviaţia standard pentru cazul unidimensional a două clase

Valoarea estimată a varianţei este, aşa după cum ştim, dată de media aritmetică, calculată la nivel de eşantion, a pătratelor deviaţiilor individuale de la medie:

$$\hat{\sigma}_x^2 = \frac{1}{N}\left[\left(a^1 - m_x\right)^2 + \left(a^2 - m_x\right)^2 + \cdots + \left(a^N - m_x\right)^2\right] \quad (7.2)$$

Rădăcina pătratică a varianţei este deviaţia standard. În **Figura 5.31** se prezintă corelaţia ce există între repartiţia numărului de elemente şi deviaţia standard a distribuţiei acestor elemente, dacă variabila aleatoare *x* este caracterizată de o funcţie densitate de probabilitate de tip *gauss*-iană. În acest caz 68% dintre exemplare vor fi la o distanţă de maximum o deviaţie standard faţă de medie în timp ce 95% dintre exemplare vor fi la o distanţă maximă de două deviaţii standard.

În reprezentarea grafică din **Figura 7.1** se prezintă, în mod intuitiv, două clase şi, respectiv, parametrii statistici de ordin unu şi doi ce le caracterizează. Din această figură se observă că media (sau generalizând vectorul mediu) ne furnizează informaţii ce privesc poziţionarea clasei (a centrului ei) în spaţiul trăsăturilor, în timp ce deviaţia standard ne furnizează informaţii despre "raza" sau dimensiunea clasei.

Observaţie 7.1: Acelaşi tip de analiză, folosind statistica descriptivă, poate fi făcută şi la nivelul fiecărei variabile aleatoare – componentă a vectorului aleator multidimensional *x*. Astfel, dacă x_i reprezintă variabila aleatoare aferentă trăsăturii *i* din vectorul aleator de trăsături *x*, atunci putem vorbi de m_{ji} ca fiind media trăsăturii *i* pentru clasa *j*, şi,

respectiv, de σ_{ji} ca fiind deviaţia standard pentru aceeaşi trăsătură i şi aceeaşi clasă j.

2. Definiţia distanţei standardizate

În general, orice valoare numerică pentru o anumită trăsătură, componentă a unui vector aleator de trăstăuri, x, poate avea o unitate de măsură specifică. În acest caz vorbim, implicit, şi de o scală de lucru corespunzătoare. Dacă o astfel de trăsătură este multiplicată cu o cantitate q (factor de scalare) atunci, atât media cât şi deviaţia standard sunt, la rândul lor, multiplicate cu aceeaşi cantitate q – în cazul mediei – şi, respectiv, cu q^2 – în cazul varianţei; aceste rezultate reprezintă particularizări ale relaţiilor (6.12) şi (6.15) pentru cazul unidimensional.

În concluzie, atât poziţionarea dar şi dispersia clasei se modifică ca efect al modificării scalei oricăreia dintre componentele vectorului de trăsături – aspect care poate fi în unele aplicaţii foarte supărător.

Pentru uşurarea calculelor în unele situaţii este de dorit o scalare a setului de date astfel încât deviaţia standard a variabilei aleatoare studiate să fie egală cu unitatea. Această operaţie este foarte uşor de realizat prin împărţirea variabilei aleatoare x la deviaţia ei standard, σ_x. În acest mod, pe lângă obţinerea efectului dorit, noua variabilă aleatoare rezultantă este o variabilă aleatoare fără unitate de măsură. Această caracteristică este foarte importantă întrucât, pe de o parte, ea rezolvă neajunsul prezentat mai sus, iar pe de altă parte, ne permite definirea unei distanţe care este, de asemenea, una independentă de unităţile de măsură folosite la observarea diverselor trăsături analizate.

Pentru măsurarea distanţei de la un element a la centrul clasei m_x (ca în cazul clasificatorului de tip minimă distanţă) este, deci, util să măsurăm această distanţă în mod relativ, prin împărţirea la deviaţia standard. Această distanţă este numită în literatura de specialitate **distanţa standardizată** şi ea este dată de:

$$r = \left| \frac{a - m_x}{\sigma_x} \right| \quad (7.3)$$

Utilizând relaţia (7.3) se poate demonstra uşor că distanţa r este invariantă la operaţiile de translare şi scalare. Aceaste observaţii sugerează o importantă generalizare a clasificatorului de tip minimă distanţă bazat pe metrica Euclidiană.

Problemă 7.1: Demonstraţi faptul că distanţa standardizată este invariantă la operaţiile de translare şi scalare.

În măsurarea distanţei dintre un vector oarecare de trăsături a şi vectorul mediu m_j, reprezentând prototipul clasei j, putem utiliza, de asemeni, şi următoarea distanţa standardizată dată de relaţia:

$$r^2(a, m_j) = \left(\frac{a_1 - m_{j1}}{\sigma_{j1}}\right)^2 + \left(\frac{a_2 - m_{j2}}{\sigma_{j2}}\right)^2 + \cdots + \left(\frac{a_d - m_{jd}}{\sigma_{jd}}\right)^2 \quad (7.4)$$

Această distanţă are şi ea, la rândul ei, aceeaşi importantă proprietate şi anume aceea de a fi invariantă la operaţiile de scalare şi translare. Astfel, dacă utilizăm această distanţă în cadrul unui clasificator, unităţile de măsură pe care noi le utilizăm pentru exprimarea valorilor diferitelor trăsături nu vor mai afecta în nici un mod distanţa măsurată între vectori şi, în consecinţă, nu vor mai influenţa nici rezultatul final al clasificării.

O generalizare directă a acestei distanţe standard pentru cazul multidimensional este şi distanţa bazată pe metrica *Mahalanobis*, metrică pe care o prezentăm în cele ce urmează.

7.1.2. Metrica Mahalanobis

Să presupunem că avem un vector aleator d-dimensional x care este caracterizat de media m_x şi de matricea de covarianţa C_x. În continuare utilizăm o matrice A, d x d dimensională, pentru transformarea vectorului aleator x într-un vector aleator y prin intermediul relaţiei:

$$y = A \cdot x \quad (7.5)$$

În cadrul acestui subcapitol dorim să găsim o metrică capabilă de a generaliza conceptul de distanţă, astfel încât distanţa de la o realizare particulară a a vectorului aleator x la m_x (media clasei) cât şi distanţa de la b la m_y (atât b cât şi m_y au fost obţinuţi din a şi, respectiv, m_x prin intermediul transformatei A) să fie egale în ambele spaţii în care realizăm măsurătorile.

Cu alte cuvinte, având o astfel de metrică nu vom mai fi obligaţi, în cazul trăsăturilor corelate, să găsim o transformare care să ne ducă din spaţiul iniţial existent într-un alt spaţiu în care matricea de covarianţă să fie una diagonală. Cu ajutorul acestei metrici vom putea realiza clasificarea în chiar spaţiul iniţial, deoarece performanţele metricii vor fi aceleaşi indiferent de spaţiul în care lucrăm (astfel spus, ne este indiferent spaţiul unde facem clasificarea). În această nouă abordare putem economisi putere de calcul

întrucât nu mai suntem obligaţi să aflăm transformata care să determine decorelarea trăsăturilor şi pe care, ulterior, să o mai şi aplicăm întregului set de date.

Observaţia 7.2: În cazul utilizării metricii euclidiene distanţe egale pentru o realizare particulară *a* în cele două spaţii (în cel iniţial şi, respectiv, în cel obţinut în urma aplicării transformării) sunt garantate numai în acele cazuri particulare (de exemplu, în situaţia în care matricea *A* doar reflectă întregul spaţiu în raport cu un hiperplan de referinţă) Ceea ce dorim să facem în mod real este să normalizăm distanţele, similar cu situaţia monodimensională dată de relaţia (7.3) astfel încât noua distanţă obţinută să fie invariantă la aplicarea oricărui tip de operator liniar.

Scopul principal al acestui subcapitol este de a generaliza distanţa standard, dată de (7.3), la o distanţă capabilă să lucreze cu vectori multidimensionali de trăsături.

Dacă rescriem relaţia (7.3) în forma:

$$r^2 = \left(\frac{a - m_x}{\sigma_x}\right)^2 = (a - m_x)\frac{1}{\sigma_x^2}(a - m_x) \quad (7.6)$$

atunci, o generalizare imediată a distanţei unidimensionale standardizate pentru cazul unui vector aleator multidimensional va fi dată de relaţia:

$$r^2 = (a - m_x)^T \frac{1}{C_x}(a - m_x) \quad (7.7)$$

Distanţa anterioară (de la vectorul aleator *a* la vectorul m_x) este cunoscută sub numele de **distanţă Mahalanobis**.

Problemă 7.2: Să se demonstreze că distanţa Mahalanobis este invariantă la orice transformare liniară aplicată setului de date.

Rezolvare:

Să considerăm un vector aleatoriu real *x* (caracterizat de media m_x şi matricea de covarianţă C_x) căruia îi aplicăm transformarea dată de relaţia (7.5). În final va rezulta un vector aleator *y* pentru care avem:

$$m_y = A\, m_x \quad (7.8)$$

şi

$$C_y = A\, C_x\, A^T \quad (7.9)$$

Ştiind că:

$$C_y^{-1} = (A^{-1})^T C_x^{-1} A^{-1} \quad (7.10)$$

putem scrie:

$$\begin{aligned} r_y^2 &= (y-m_y)^T C_y^{-1} (y-m_y) = \\ &= (Ax-Am_x)^T \left(A^{-1}\right)^T C_x^{-1} A^{-1} (Ax-Am_x) = \\ &= (x-m_x)^T A^T \left(A^{-1}\right)^T C_x^{-1} A^{-1} A(x-m_x) = \\ &= (x-m_x)^T \left(A^{-1}A\right)^T C_x^{-1} \left(A^{-1}A\right)(x-m_x) = \\ &= (x-m_x)^T C_x^{-1} (x-m_x) = \\ &= r_x^2 \end{aligned} \quad (7.11)$$

Această demonstraţie dovedeşte invarianţa distanţei Mahalanobis la orice tip de transformare liniară.

Se poate arăta că suprafeţele de decizie date de norma Mahalanobis sunt pătratice iar pentru distanţe r constante faţă de m_x obţinem elipsoizi care sunt centraţi în acest vector.

7.1.3. Clasificatorul Mahalanobis

Un clasificator de minimă distanţă ce utilizează distanţa Mahalanobis se numeşte **clasificator Mahalanobis**. În continuare prezentăm modul de utilizare a distanţei Mahalanobis în cadrul unui clasificator de tip minimă distanţă.

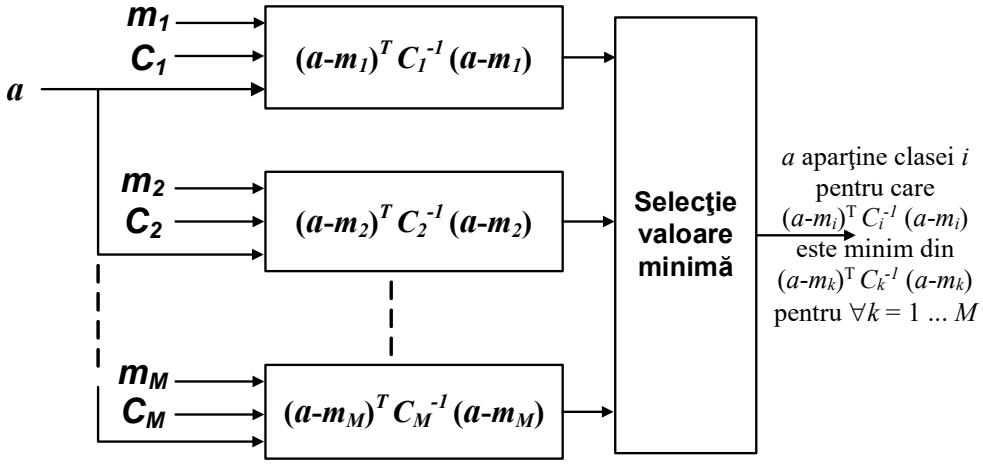


Figura 7.2. Schema bloc a clasificatorului bazat pe norma Mahalanobis

Fie m_1, m_2, ... , m_M centrele de masă (şabloanele, *template*-urile) pentru cele M clase considerate (implicit se presupune că acestea formează o partiţie a eşantionului), iar C_1, C_2, ... , C_M matricile de covarianţă corespunzătoare. Cu ajutorul normei Mahalanobis se măsoară distanţele de la vectorul a la fiecare element prototip aparţinând celor M clase. Vectorul a va fi, în final, atribuit acelei clase pentru care distanţa Mahalanobis de la el la media clasei este minimă. Schema bloc a acestui clasificator este prezentată în **Figura 7.2**.

Observaţia 7.3: În cazul particular când trăsăturile sunt necorelate (matricea de covarianţă, C_x, este diagonală) iar varianţele în toate direcţiile spaţiului de trăsături sunt aceleaşi (toate elementele de pe diagonala principală a lui C_x sunt egale între ele), suprafeţele de decizie devin hipersfere iar distanţa Mahalanobis, în această situaţie, este identică cu distanţa Euclidiană. În concluzie, un clasificator de minimă distanţă ce utilizează distanţa Mahalanobis **este, de fapt, o <u>generalizare</u> a unui clasificator de tip minimă distanţă ce utilizează distanţa Euclidiană**.

7.1.4. Avantaje şi dezavantaje ale clasificatorului Mahalanobis

Utilizarea metricii Mahalanobis înlătură o serie de limitări prezente în cazul utilizării metricii Euclidiene, asigurând totodată următoarele avantaje:

(1) independenţă faţă de scalarea uneia sau alteia dintre trăsături;
(2) capacitate de a lucra cu trăsături corelate;
(3) flexibilitate superioară a suprafeţelor de decizie, care de această dată sunt unele pătratice.

Toate aceste avantaje prezentate anterior au şi un preţ pe care trebuie să-l plătim. Astfel:

(1) Pentru utilizarea metricii Mahalanobis trebuie să estimăm (din setul de date) matricile de covarianţă pentru fiecare clasă în parte. Din păcate însă aceste matrici sunt extrem de <u>dificil de estimat cu acurateţe</u>. Mai mult, <u>timpul de calcul cât şi memoria de stocare creşte pătratic</u> cu numărul de trăsături pe care îl utilizăm.

Dacă avem N realizări particulare pentru vectorul aleator real x, a^1, a^2, ... , a^N, toate aparţinând aceleiaşi clase, matricea de covarianţă va fi calculată cu relaţia:

$$C_x = \frac{1}{N}\left[\left(a^1 - m_x\right)\left(a^1 - m_x\right)^T + \left(a^2 - m_x\right)\left(a^2 - m_x\right)^T + \cdots + \left(a^N - m_x\right)\left(a^N - m_x\right)^T \right] \tag{7.12}$$

Dacă ţinem cont că vectorul aleator *x* este unul *d*-dimensional rezultă că matricea de covarianţă C_x este o matrice *d* x *d* dimensională.

(2) Dacă *N*, numărul total de trăsături cu ajutorul cărora estimăm C_x, este mai mic decât $d + 1$ atunci ne aflăm în situaţia particulară în care matricea C_x este o matrice singulară[1]; acest fapt generează o problemă majoră, întrucât noi suntem obligaţi să inversăm matricea de covarianţă pentru a putea calcula mai departe distanţa Mahalanobis. Pentru ca o matrice de covarianţă să nu fie singulă, la limită, va trebui să avem un număr minim de vectori de trăsături care să fie mai mare de $d + 1$. Lucrurile se complică, însă, şi mai mult atunci când conştientizăm că numărul de vectori aleatori, cu ajutorul cărora se estimează matricea de covarianţă pentru fiecare clasă în parte, trebuie să fie mai mare decât acest prag. În consecinţă, pentru o aplicaţie cu *M* clase şi un vector de trăsături *d*-dimensional setul minim de date ar trebui să conţină cel puţin *M* x $(d + 1)$ vectori de trăsături, cu minim $d + 1$ vectori aleatori pentru fiecare clasă.

Observaţia 7.4: Chiar în situaţia în care numărul de vectori aleatori ai setului de date va fi puţin mai mare de $d + 1$, estimarea matricei de covarianţă nu va fi una foarte exactă. Prin utilizarea unei matrici de covarianţă a cărei estimare nu este foarte exactă vom obţine o acurateţe scăzută atât în cazul clasificatorului de tip minimă distanţă ce utilizează o normă Mahalanobis cât şi în cazul clasificatorului *Bayes*-ian, aşa cum de altfel vom prezenta şi în subcapitolul următor. În ceea ce priveşte existenţa unui număr limitat de vectori de trăsături aceasta reprezintă un impediment major în cazul ambelor clasificatoare (*Bayes*-ian şi, respectiv, a celui bazat pe norma *Mahalanobis*).

(3) Matricea de covarianţă (pentru vectori aleatori *d*-dimensionali) conţine d^2 elemente, din care – datorită simetriei acestei matrici – doar un număr de $d(d+1)/2$ elemente sunt independente. În acest caz putem spune că o estimare corectă a matricii C_x se obţine efectiv abia în momentul în care numărul de vectori de trăsături cu ajutorul cărora calculăm această matrice ajunge aproape de $d(d + 1)/2$ sau, ideal, depăşeşte această valoare. Dacă problema de rezolvat este una „mică”

[1] O matrice este singulară dacă determinantul ei este egal cu zero.

din punct de vedere al numărului trăsăturilor implicate, acest fapt nu ar ridica probleme deosebite. Dar, din păcate, nu este ieşit din comun să avem probleme de clasificare caracterizate de vectori de trăsături ce au, de exemplu, în jur de 100 de componente.

***Exemplu 7.1**:* În cazul unei probleme de clasificare a 5 *task*-uri mentale pe baza coeficienţilor de amplitudine ai modelului ANAPP (*Adaptive Nonlinear Amplitude and Phase Process*) – model prezentat în referinţa [Dobrea, 2007] – s-au utilizat vectori de trăsături cu un număr de 104 componente; aceste componente reprezentau, concatenat într-un singur vector de trăsături, diverşi parametri estimaţi pe un număr de 6 secvenţe de semnal EEG înregistrate simultan de la 6 electrozi plasaţi la nivelul scalpului unui subiect. În acest caz, o realizare particulară a vectorului aleator de trăsături a fost asociată unor înregistrări (simultane) ale semnalului EEG de la cei 6 electrozi în timp ce setul de date a fost generat de N astfel de înregistrări simultane. În această aplicaţie particulară d a fost egal cu 104 iar un calcul simplu ne relevă un necesar de cel puţin $d(d+1)/2$, adică de cel puţin 5460 de realizări particulare pentru o estimare corectă a matricei de covarianţă pentru doar o singură clasă. Ţinând cont că aplicaţia de clasificare viza 5 astfel de clase, rezultă că în total ar fi fost nevoie de minim 27300 de vectori de trăsături pentru o estimare corectă a celor cinci matrici de covarianţă. În realitate baza de date utilizată a avut doar 1670 de vectori de trăsături [Dobrea, 2007]. Dacă în această situaţie s-ar fi utilizat un clasificator de tipul Mahalanobis, rezultatele clasificării ar fi fost aproape cu siguranţă unele foarte slabe. Motivul unor astfel de performanţe scăzute nu s-ar fi datorat puterii scăzute de discriminare a acestei metode, aşa cum am fi fost poate tentaţi să credem la o primă analiză ci ele ar fi fost generate, în principal, de o estimare nesatisfăcătoare a parametrilor modelului.

În concluzie, se observă că o dată cu creşterea numărului de trăsături apar o serie de factori limitativi ce determină o acurateţe scăzută a clasificatorului bazat pe metrica Mahalanobis.

Aplicaţie 7.1: Pentru înţelegerea clasificatorului Mahalanobis s-a realizat un program al cărui cod se află în directorul „***Comparatie minDist-Mahalanobis-Bayes***” asociat acestui subcapitol. Programul este unul mai general şi el implementează trei clasificatori elementari (clasificatorul de minimă distanţă bazat pe metrica euclidiană,

clasificatorul *Mahalanobis* şi, respectiv, clasificatorul *Bayes*-ian, ce va fi prezentat mai târziu). Programul permite:

- încarcarea unui set de vectori bidimensionali de trăsături, set de date ce poate conţine un număr de maxim 7 clase – operaţia se realizează prin intermediul elementelor 1 de pe interfaţa grafică (vezi **Figura 7.3)**;
- plasarea în spaţiul trăsăturilor, cu ajutorul *mouse*-ului, a unui element necunoscut urmând ca, ulterior, să se determine aparteneţa acestuia la una din clasele posibile utilizând, pentru aceasta, unul din clasificatorii anterior menţionaţi; programul va prezenta simultan şi valorile funcţiilor discriminant calculate pentru elementul necunoscut – pentru aceasta se folosesc elementele din registrul 2;
- trasarea, pentru setul de date încărcat, a suprafeţelor de decizie obţinute folosind cei trei clasificatori – butoanele dedicate sunt cele din registrul 3;
- Afişarea, în valori absolute sau procentuale, a performanţelor clasificării obţinute folosind succesiv, pe acelaşi set de date, cei trei clasificatori; pentru afişarea performanţelor s-a folosit ca formă de prezentare matricea confuziilor.

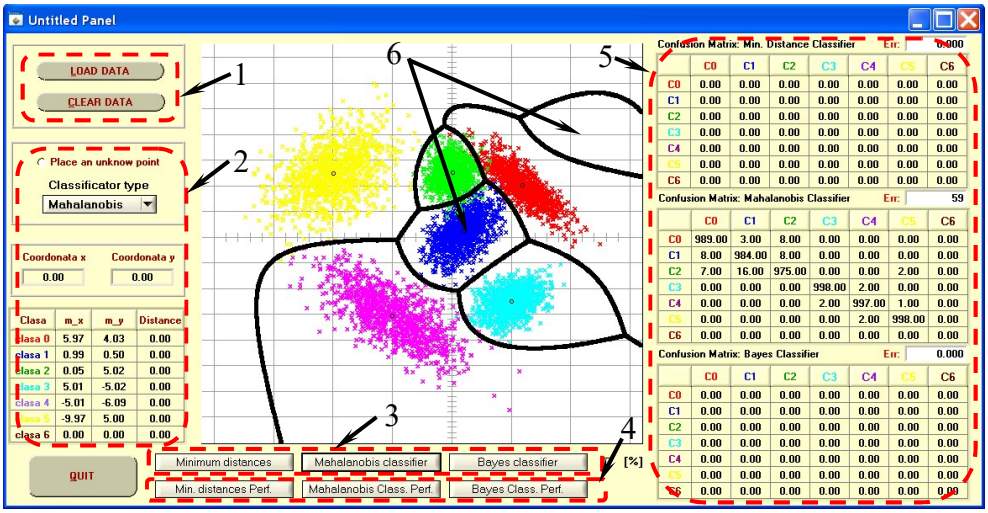


Figura 7.3. Interfaţa grafică a programului

Utilizând acest program se cere să se atingă următoarele obiective:

1. Încărcaţi unul din următoarele seturi de date: „***6 Classes.txt***”, „***7 Classes.txt***” (aceste două seturi de date conţin 6, respectiv, 7 clase,

cu număr similar de vectori de trăsături bidimensionali – 1000 pentru fiecare clasă în parte; aceste clase sunt caracterizate, fiecare în parte, de o matrice de covarianţă şi, respectiv, de o medie proprie), „***6 Classes - Different prob.txt***” (acest set de date este derivat din „***6 Classes.txt***”, de care diferă prin aceea că cea de a treia clasă este compusă din doar 10% din elementele clasei originale), „***Femei-Barb Greut-Inalt Bayes rnd.txt***”, „***Odihnit-Obosit Puls-PresSist rnd.txt***”.

2. Determinaţi generarea suprafeţelor de decizie pentru clasificatorul de minimă distanţă şi pentru cel *Mahalanobis*.
3. Determinaţi performanţele de clasificare ale acestor doi clasificatori (analizate prin intermediul matricilor confuziilor) şi corelaţi rezultatele obţinute cu poziţionarea suprafeţelor de decizie anterior obţinute.
4. Plasaţi elemente necunoscute în spaţiul de intrare şi analizaţi corecta lor atribuire la una din clasele existente prin intermediul clasificatorului de minimă distanţă şi a celui de tip *Mahalanobis*.
5. Prin trasarea suprafeţelor de decizie pentru clasificatorul *Mahalanobis* şi analizând rezultatele obţinute la punctul precedent se poate observa că una din clase este caracterizată de existenţa a două regiuni decizionale disjuncte (respectiv, regiunile 6); dintre aceste regiuni, una nici măcar nu conţine elemente. Explicaţi, în mod intuitiv, ce anume a determinat obţinerea acestui rezultat.

7.2. Clasificatorul *Bayes*-ian

7.2.1 Alegerea optimă a suprafeţei de decizie bazată pe modelul statistic al datelor

Clasificatorii prezentaţi până acum şi-au bazat regula de decizie pe rezultatul măsurării unor distanţe (distanţele dintre forma de intrare şi un set de vectori de referinţă sau puncte prototip din spaţiul caracteristicilor). O nouă abordare a problemelor de clasificare – bazată de această dată pe modelul statistic al setului de date (vezi definiţia modelului statistic dată în ***Subcapitolul 5.5.1***) – este prezentată în cele ce urmează.

Ca element de referinţă, *tehnicile statistice de clasificare* se bazează pe o presupunere fundamentală şi anume aceea că pentru fiecare clasă există o funcţie densitate de probabilitate ce ne ajută în determinarea probabilităţii ca o formă de intrare să aparţină unei clase sau alteia dintre clasele posibile.

Exemplul 7.2: Determinarea stării de oboseală fizică a unui subiect funcţie de activitatea cardiacă a acestuia – cuantizată prin intermediul pulsului persoanei analizate (problemă prezentată anterior, în ***Subcapitolul 3.5.1***) –, poate fi modelată, de exemplu, şi tratată statistic astfel: dacă considerăm pulsul, măsurat în bătăi/minut, o cantitate guvernată de legile statisticii şi generată de două "procese" diferite – oameni odihniţi şi, respectiv, oameni obosiţi[2] fizic – observăm oportunitatea utilizării unui model statistic în cadrul acestei probleme. Cele două clase sunt date, în acest caz, de clasa {odihnit} şi, respectiv, clasa {obosit}.

Din măsurări repetate ale pulsului pentru cele două clase de subiecţi putem extrage **parametrii statistici** ce determină în mod unic **funcţiile densitate de probabilitate teoretice** propuse pentru a modela distribuţiile celor două clase.

În cazul ipotezei unor distribuţii *gauss*-iene monodimensionale – ipoteză folosită şi de noi în cazul problemei de faţă – avem nevoie să estimăm doar doi parametri, respectiv, media şi varianţa, în timp ce, în cazul distribuţiilor *gauss*-iene caracterizate de vectori *d*-dimensionali de trăsături ar fi trebuit să calculăm vectorul mediu şi matricea de covarianţă. O dată estimaţi aceşti parametri, pentru ambele clase,

[2] În această analiză starea de oboseală fizică a fost indusă de o activitate fizică intensă – urcarea unui deal într-un timp minim posibil fiecărui subiect

putem ulterior **determina complet forma funcţională a densităţilor de probabilitate** ce le caracterizează. Folosind, în plus, şi informaţiile *a priori* privind probabilitatea claselor putem deduce mai departe, cu ajutorul regulei lui Bayes, care sunt **probabilităţile posterioare ale claselor** (probabilităţile condiţionate revizuite ale claselor, reprezentând probabilitatea ca o realizare particulară să aparţină unei anumite clase). În final, folosind aceste probabilităţi posterioare ale claselor putem trece la implementarea clasificatorului statistic.

***Observaţia 7.5*:** Problema de mai sus poate fi una generalizată astfel: *avem la dispoziţie o realizare particulară, a^0, a unui vector aleator x, d-dimensional, de trăsături şi ne punem problema cărei clase din cele M posibile să îl atribuim*.

Soluţia cea mai logică şi directă ar fi să **atribuim această realizare particulară a^0 acelei clase ce are probabilitatea maximă de a-l conţine.** În concluzie, observăm că din nou teoria probabilităţilor şi statistica ne poate da o mână de ajutor în rezolvarea acestor clase de probleme.

Teoria probabilităţilor şi statistica ne propun un set de reguli foarte generale şi precise (aplicabile la o clasă foarte largă de probleme) pentru construcţia unui clasificator. Această teorie ne arată că un clasificator optimal va alege clasa c_i de apartenenţă a elementului a^0 (o realizare particulară a unui vector aleator x) drept acea clasă ce maximizează probabilitatea condiţionată $P(c_i \mid a^0)$. Acest ultim termen, $P(c_i \mid a^0)$, numit şi probabilitate posterioară (vezi ***Subcapitolul 5.4.5***, „Probabilitatea condiţionată. Regula produsului, a sumei şi teorema lui Bayes"), reprezintă probabilitatea clasei c_i de a include elementul a^0. Din această perspectivă vom avea:

$$a^0 \text{ va fi asignat clasei } c_i \text{ dacă: } P\left(c_i \middle| a^0\right) = \max_{j=1\ldots M} P\left(c_j \middle| a^0\right) \quad (7.13)$$

În relaţia (7.13) M este numărul total de clase cărora elementul a^0 poate să le aparţină virtual; de exemplu pentru problema practică prezentată la începutul acestui subcapitol $M = 2$ (obosit versus odihnit).

Observaţia 7.6*:** Din punctul de vedere al teoriei clasificatorilor (vezi ***Subcapitolul 3.5.2) se constată că în relaţia (7.13) termenii $P(c_i \mid a^0)$ sunt chiar funcţiile discriminant caracteristice fiecărei clase.

Deoarece probabilitatea posterioară nu poate fi determinată în mod direct din măsurători asupra setului de date, vom utiliza relaţia lui *Bayes* pentru a o deduce. Astfel, putem scrie:

$$P(c_i \mid a^0) = \frac{f_x(a^0 \mid c_i)P(c_i)}{f_x(a^0)} \quad (7.14)$$

unde:

- $P(c_i|a^0)$ este probabilitatea[3] clasei c_i de a include elementul a^0 (este probabilitatea ca a^0 chiar să aparţină clasei c_i);
- $f_x(a^0|c_i)$ probabilitatea realizării lui a^0 dată de modelul statistic al vectorului aleator x ce descrie clasa c_i (valoarea calculată în a^0 a funcţiei densitate de probabilitate care descrie distribuţia vectorului de trăsături pentru clasa c_i), mai simplu este probabilitatea ca elementul a^0 să fie generat de clasa c_i;
- $P(c_i)$ este probabilitatea apriorică a clasei, calculată în cazul problemei noastre de clasificare ca fiind probabilitatea de realizare a evenimentelor din clasa c_i;
- $f_x(a^0)$ este probabilitatea ca evenimentul a^0 să se întâmple indiferent de clasa căreia îi aparţine; în general, această probabilitate se calculează cu relaţia:

$$f_x(a^0) = \sum_{i=1}^{M} f_x(a^0 \mid c_i)P(c_i) \quad (7.15)$$

După cum se poate observa din relaţia (7.14), probabilitatea posterioară poate fi calculată ca produsul dintre probabilitatea apriorică a clasei, $P(c_i)$, şi probabilitatea ca elementul a^0 să fie generat de un proces caracterizat de clasa c_i, totul normalizat la $f_x(a^0)$.

Determinarea numerică a probabilităţii posterioare $P(c_i \mid a^0)$ şi asignarea unui element la o a anumită clasă este foarte simplă, ea presupunând parcurgerea următorilor paşi:

(1) ţinând cont de setul de date şi de presupunerile iniţiale făcute asupra formei funcţiilor densitate de probabilitate condiţionată, $f_x(a|c_i)$, se estimează aceste funcţii pentru fiecare clasă în parte;
(2) se estimează $P(c_i)$, probabilităţile apriorice, din datele existente, drept probabilitatea de realizare a evenimentelor din clasa c_i;

[3] Ca o observaţie, $P(c_i|a^0) = p(c_i|a^0)$, unde $p(\cdot)$ este funcţia masă de probabilitate condiţionată a clasei c_i dată de $\{x = a\}$.

(3) ulterior, din estimarea funcţiilor densitate de probabilitate ce descriu distribuţiile claselor (realizată în prima etapă), se determină probabilitatea $f_x(a^0|c_i)$, reprezentând valoarea funcţiei $f_x(a|c_i)$ calculată în punctul a^0;

(4) ultimul factor, $f_x(a^0)$, este un factor de normalizare ce este, de obicei, eliminat în cadrul aplicaţiilor de clasificare datorită:

   a. lipsei de informaţii pe care le aduce în procesul de decizie finală a clasificatorului, şi
   b. încărcării computaţionale inutile.

(5) În ultimul pas, folosindu-ne de relaţia (7.14), condiţia (7.13) devine:

a^0 va fi asignat clasei c_i dacă:

$$P(c_i) f_x\left(a^0 \middle| c_i\right) = \max_{j=1...M} P(c_j) f_x\left(a^0 \middle| c_j\right) \quad (7.16)$$

După cum am prezentat anterior, în relaţia (7.16) termenul $f_x(a^0)$, termen independent de clasa c_j, a fost omis deoarece este acelaşi factor normalizator, comun pentru toate clasele.

Problema 7.3: Să se determine suprafaţa optimă de decizie ce separă două clase de subiecţi umani (odihniţi şi obosiţi din punct de vedere fizic), funcţie de activitatea cardiacă proprie, utilizând pulsul drept trăsătură discriminatorie între cele două clase. În cadrul acestei probleme se presupune că înregistrările activităţii cardiace au fost realizate pe un grup de 100 de subiecţi (angajaţi ai aceleiaşi secţii, din aceeaşi întreprindere) la ora 8.00 AM, imediat după sosirea acestora la serviciu. Aşa după cum li s-a explicat şi cerut anterior, jumătate din subiecţi au parcurs ultima parte a drumului către serviciu pe jos, într-un ritm ceva mai alert iar restul au venit la serviciu cu ajutorul mijloacelor de transport în comun (astfel încât starea de oboseală fizică să nu se instaleze).

Seturile de date pentru un eşantion statistic de 100 respectiv 10000 de înregistrări se găsesc în directorul asociat acestui capitol în fişierele „***Puls Obosit-Odihnit 100.txt*** ”şi, respectiv, „***Puls Obosit-Odihnit 10K.txt***”.

Rezolvare: Această problemă va fi rezolvată pentru un eşantion de 100 de înregistrări (o singură zi) şi, respectiv, un eşantion de 10000 de înregistrări (efectuate în 100 de zile consecutive), scopul urmărit fiind

acela de a pune în evidenţă influenţa mărimii eşantionului statistic[4] asupra rezultatelor obţinute.

Dacă particularizăm relaţiile anterioare ce descriu clasificatorul *Bayes*-ian, la problema de clasificare a stării de oboseală funcţie de activitatea cardiacă a subiecților, în relația (7.16) avem $i = 1, 2$ şi, respectiv $M = 2$ (două clase de subiecţi: odihniţi şi obosiţi).

În ipoteza că funcţiile densitate de probabilitate condiţionată, $f_x(a|c_i)$, pentru cele două clase sunt densităţi *gauss*-iene, *estimăm parametrii* acestora – media şi, respectiv, varianţa – cu ajutorul relaţiilor:

$$\hat{m}_i = \frac{1}{K_i}\sum_{j=1}^{K_i} a_i^j \qquad (7.17)$$

$$\hat{\sigma}_i^2 = \frac{1}{K_i}\sum_{j=1}^{K_i} (a_i^j - \hat{m}_i)^2 \qquad (7.18)$$

unde K_i este numărul de realizări particulare ale variabilei aleatoare x (pulsul subiectului) pentru fiecare clasă i iar a_i^j este a j-a realizare particulară a variabilei aleatoare x pentru clasa i.

Parametrii statistici estimaţi pentru cele două eşantioane sunt cei prezentaţi în tabelul ataşat **Figurii 7.4**, în timp ce o reprezentare grafică a distribuţiilor celor două clase este dată în **Figura 7.4** pentru cazul eşantionului de 10.000 de înregistrări.

O dată cu estimarea parametrilor media şi varianţa pentru distribuţiile condiţionate $f_x(a|c_i)$ ale celor două clase de indivizi, obţinem, practic şi o determinare completă a acestor funcţii:

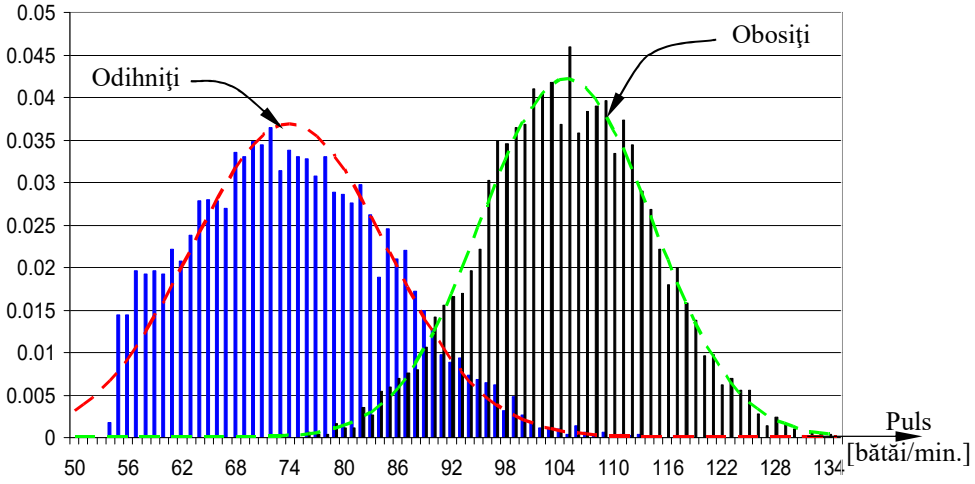
$$f_x(a\,|\,c_1) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_1} e^{-\frac{1}{2}\left[\frac{(a-\hat{m}_1)^2}{\hat{\sigma}_1^2}\right]} \qquad (7.20)$$

$$f_x(a\,|\,c_2) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_2} e^{-\frac{1}{2}\left[\frac{(a-\hat{m}_2)^2}{\hat{\sigma}_2^2}\right]} \qquad (7.21)$$

Probabilitatea $f_x(a^0|c_i)$, ca valoarea a^0 să fi fost generată de un subiect aparţinând uneia din cele două clase ({odihnit} sau {obosit}),

[4] În cadrul acestei probleme, prin eşantion statistic înţelegem numărul de înregistrări şi nu numărul de subiecţi care au participat la obţinerea setului de date.

se determină prin particularizarea funcţiilor densitate $f_x(a|c_i)$ în valoarea a^0.



	10000 de înregistrări	**100 de înregistrări**
Odihniţi	Media = 73.92 Deviaţia standard = 10.82	Media = 74.32 Deviaţia standard = 10.88
Obosiţi	Media = 104.08 Deviaţia standard = 9.46	Media = 105.15 Deviaţia standard = 9.56

Figura 7.4. Distribuţia seturilor de date: cu linie punctată sunt reprezentate funcţiile densitate de probabilitate ce caracterizează cele 2 clase iar prin segmente verticale sunt reprezentate histogramele celor două clase.

Termenii $P(c_i)$, reprezentând probabilităţile apriorice ale claselor, ar putea fi estimaţi folosind pentru aceasta, pe lângă informaţiile furnizate de setul de date, şi informaţii cu caracter mai subiectiv legate de regimul de viaţă al subiecţilor, predispoziţia acestora de a efectua, în mod uzual, activităţi fizice, starea de sănătate[5], corectitudinea şi conştiinţa angajaţilor (de exemplu, în a respecta un orar fix), sexul[6],

[5] Numărul de bătăi pe minut pentru o persoană sănătoasă în stare de repaus este în intervalul [60, 100]. Ritmuri cardiace de peste 100 bătăi/minut (90 după alţi autori) se consideră drept patogene – tahicardie. Valori ale activităţii cardiace sub o valoare de 60 poartă numele de bradicardie. Pentru anumite clase de subiecţi (sportivi, adulţi tineri şi sănătoşi) bradicardia este, în anumite condiţii, considerată normală.

[6] În medie inima unui bărbat bate în jur de 73 de ori pe minut, iar pentru o femeie ritmul cardiac este în jur de 80 bătăi/minut.

vârsta[7] etc. În general însă, în problemele de clasificare şi de învăţare, probabilităţile apriorice ale claselor se calculează într-un mod mai obiectiv, ca frecvenţă relativă a evenimentelor din clasa c_i:

$$P(c_i) = \frac{K_i}{\sum_{j=1}^{M} K_j} \qquad (7.22)$$

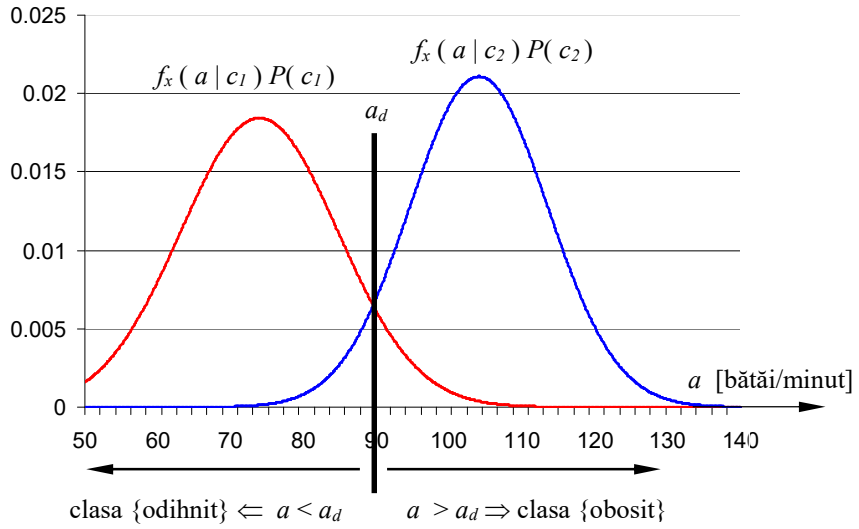


Figura 7.5. Determinarea suprafeţei de decizie optime, a_d, utilizând un clasificator *Bayes*-ian

Alegerea poziţiei suprafeţei de decizie $a = a_d$ (**Figura 7.5**) este dată de punctul în care probabilităţile posterioare calculate mai sus sunt egale, fiind deci soluţia ecuaţiei:

$$P(c_1 \mid a) = P(c_2 \mid a) \qquad (7.23)$$

de unde rezultă mai departe:

$$P(c_1)\frac{1}{\sqrt{2\pi}\hat{\sigma}_1}e^{-\frac{1}{2}\left(\frac{a-\hat{m}_1}{\hat{\sigma}_1}\right)^2} = P(c_2)\frac{1}{\sqrt{2\pi}\hat{\sigma}_2}e^{-\frac{1}{2}\left(\frac{a-\hat{m}_2}{\hat{\sigma}_2}\right)^2} \qquad (7.24)$$

[7] Inima unui nou născut bate de 120 de ori pe minut. Acest ritm scade o dată cu înaintarea în vârstă, la 10 ani ajungând în jur de 90 de bătăi pe minut. Pentru o persoană matură ritmul cardiac mediu este prezentat în observaţia anterioară.

$$P(c_1)\hat{\sigma}_2 e^{-\frac{1}{2}\left(\frac{a-\hat{m}_1}{\hat{\sigma}_1}\right)^2} = P(c_2)\hat{\sigma}_1 e^{-\frac{1}{2}\left(\frac{a-\hat{m}_2}{\hat{\sigma}_2}\right)^2} \quad (7.25)$$

Logaritmând, obţinem:

$$\ln(P(c_1)) + \ln(\hat{\sigma}_2) - \frac{1}{2}\left(\frac{a-\hat{m}_1}{\hat{\sigma}_1}\right)^2 = \ln(P(c_2)) + \ln(\hat{\sigma}_1) - \frac{1}{2}\left(\frac{a-\hat{m}_2}{\hat{\sigma}_2}\right)^2 \quad (7.26)$$

În final rezultă următoarea ecuaţie de gradul 2:

$$\frac{a^2}{2}\left(\frac{1}{\hat{\sigma}_2^2} - \frac{1}{\hat{\sigma}_1^2}\right) + a\left(\frac{\hat{m}_1}{\hat{\sigma}_1^2} - \frac{\hat{m}_2}{\hat{\sigma}_2^2}\right) + \left[\ln\left(\frac{P(c_1)}{P(c_2)}\right) + \ln\left(\frac{\hat{\sigma}_2}{\hat{\sigma}_1}\right) - \frac{1}{2}\left(\frac{\hat{m}_1^2}{\hat{\sigma}_1^2} - \frac{\hat{m}_2^2}{\hat{\sigma}_2^2}\right)\right] = 0 \quad (7.27)$$

Discuţii:

Pentru varianţe diferite ($\hat{\sigma}_1 \neq \hat{\sigma}_2$) relaţia (7.27) este o ecuaţie de gradul doi, cu două soluţii. Dacă varianţele $\hat{\sigma}_1$ *şi* $\hat{\sigma}_2$ sunt egale atunci relaţia (7.27) devine o ecuaţie de gradul întâi, având o singură soluţie.

a) În cazul aplicaţiei noastre, pentru care $\hat{\sigma}_1 \neq \hat{\sigma}_2$, ecuaţia are două rădăcini însă dintre acestea numai una este soluţia căutată, respectiv, soluţia plauzibilă din punct de vedere biomedical. Soluţiile găsite sunt:
   - pentru eşantionul cu 10000 de determinări: a_{d1} = 89.55 şi a_{d2} = 314.31 iar
   - pentru eşantionul cu 100 de determinări: a_{d1} = 90.29 şi a_{d2} = 328.85.

   Întrucât a doua soluţie obţinută nu este nici într-un caz, nici în altul, plauzibilă din punct de vedere biomedical, rezultă că suprafaţa de decizie (care pentru această problemă este un punct[8]) trebuie să fie aleasă în valorile 89.55 pentru primul şi, respectiv, 90.29 pentru cel de-al doilea eşantion de date. Pentru aceste valori ale pragului de decizie clasificatorul obţinut este unul optimal.

   Se observă că prin utilizarea unui set de date mult diminuat (doar 1% din primul eşantion de date) poziţia suprafeţei de decizie obţintă pentru această aplicaţie se modifică, însă nu în mod substanţial (90.29 versus 89.55), eroarea fiind una acceptabilă.

b) Dacă particularizăm problema şi presupunem că $\hat{\sigma}_1 = \hat{\sigma}_2 = \sigma$ atunci ecuaţia de gradul doi se transformă într-o ecuaţie de gradul întâi a cărei soluţie este:

[8] Suprafeţele de decizie sunt întotdeauna (d-1) dimensionale, vezi şi ***Subcapitolul 3.5.3***. Unde d este dimensiunea spaţiului de intrare, a spaţiului trăsăturilor.

$$a_d = \frac{\hat{m}_1 + \hat{m}_2}{2} + k \quad (7.26)$$

În relaţia (7.26), k depinde de raportul probabilităţilor apriorice ale claselor precum şi de mediile acestora, fiind dat de:

$$k = \frac{\sigma^2}{\hat{m}_2 - \hat{m}_1} \ln \frac{P(c_1)}{P(c_2)} \quad (7.27)$$

c) În momentul în care avem $\hat{\sigma}_1 = \hat{\sigma}_2 = \sigma$ şi, în plus, $P(c_1) = P(c_2)$ – cu alte cuvinte, varianţele ambelor clase sunt aceleaşi iar clasele sunt echiprobabile (se particularizează relaţia (7.27) pentru $P(c_1) = P(c_2)$, rezultând $k = 0$) – suprafaţa de decizie va fi poziţionată la jumătatea distanţei dintre centrele celor două clase. Altfel spus, suprafaţa de decizie este dată în această situaţie doar de medile claselor. Acest rezultat este identic cu cel ce se obţine în cazul utilizării clasificatorului de minimă distanţă bazat pe norma Euclidiană. În consecinţă, ***în aceste condiţii particulare clasificatorul Bayes-ian este identic cu un clasificator de tip minimă distanţă.*** Reamintim că modul de funcţionare a clasificatorului de tip minimă distanţă (vezi ***Subcapitolul 4.2***) presupune utilizarea în procesul de clasificare numai a informaţiei privind distanţele către mediile claselor.

d) Pentru clasificatorul *Bayes*-ian, soluţia ecuaţiei suprafeţei de decizie (vezi şi relaţia (7.27) obţinută pentru aplicaţia particulară, unidimensională, cu două clase, de mai sus), este una ce depinde, pe lângă informaţia dată de mediile claselor, şi de informaţiile furnizate de varianţele (mai general, matricile de covarianţă) şi, respectiv, probabilităţile apriorice ale claselor. Pentu a înţelege modul cum influenţează aceste informaţii poziţionarea suprafeţei de decizie facem următoarele discuţii pe cazul unidimensional al problemei noastre, extrapolarea acestor concluzii la cazul multidimensional fiind una directă:

(i) Existenţa a două varianţe diferite pentru cele două clase ($\hat{\sigma}_1 \neq \hat{\sigma}_2$) determină *mutarea pragului de decizie spre clasa cu varianţă minimă* (respectiv, spre dreapta în cazul problemei în discuţie; vezi **Figura 7.4.**).

Acest fenomen este şi unul intuitiv deoarece varianţă minimă înseamnă, în principal, o concentrare mai mare a datelor în jurul valorii medii. În consecinţă, alegerea suprafeţei de decizie

depinde şi de varianţa fiecărei aglomerări de puncte aparţinând unei anumite clase nu numai de poziţionarea centrului, a mediei clasei respective. Prin această observaţie evidenţiem practic faptul că în procesul de clasificare *avem nevoie de o metrică care să depindă nu numai de distanţa dintre elementul necunoscut la centrele claselor, ci şi de varianţa acestora* (de modul de distribuţie al claselor). O astfel de metrică este, aşa după cum ştim, metrica Mahalanobis.

În contextul de mai sus, în cazul în care clasele sunt echiprobabile ***clasificatorul Bayes-ian este identic***, din punct de vedere conceptual, cu un ***clasificator de tipul Mahalanobis***.

(ii) Din relaţia (7.27) se observă că raportul probabilităţilor apriorice ale claselor mută şi el suprafaţa de decizie spre dreapta sau spre stânga, funcţie de valoarea acestui raport. Intuitiv, ca regulă generală, pragul de decizie se mută întotdeauna mai aproape de clasa cu probabilitate mai mică. În consecinţă, clasa cu probabilitate mai mare va avea o regiunea de decizie mai mare.

Dacă luăm în discuţie complexitatea diferiţilor clasificatori prezentaţi până acum observăm că:

(a) clasificatorul de minimă distanţă ce utilizează norma Euclidiană are complexitatea cea mai redusă, folosind în procesul decizional numai informaţia furnizată de ***mediile claselor***;

(b) clasificatorul bazat pe norma Mahalanobis ***foloseşte în plus*** în procesul de decizie, faţă de clasificatorul anterior, şi informaţii ce ţin de ***dispunerea spaţială a claselor***; aceste informaţii sunt valorificate prin intermediul matricii de covarianţă (pentru spaţii de trăsături cel puţin bidimensionale), respectiv, prin intermediul varianţei (pentru cazul unidimensional). Din acest punct de vedere complexitatea acestui clasificator este şi ea una mai mare comparativ cu complexitatea clasificatorului de tip minimă distanţă, studiat în Capitolul 4.

(c) clasificatorul Bayes-ian ***utilizează în plus*** faţă de clasificatorul Mahalanobis – în determinarea funcţiilor discriminant şi, deci, a poziţiei suprafeţelor de decizie –, informaţii ce privesc ***probabilităţile apriorice ale claselor***. În concluzie, acest din urmă clasificator utilizează în procesul decizional, simultan, informaţii ce privesc centrele claselor, dispunerea spaţială a lor, precum şi probabilitatea apriorică a acestora. Din acest punct de vedere, clasificatorul *Bayes*-ian este cel mai complex din cei prezentaţi aici şi, mai mult, atât clasificatorul de minimă distanţă cât şi clasificatorul Mahalanobis reprezintă particularizări, în anumite

condiţii bine definite (vezi discuţiile anterioare), ale clasificatorului *Bayes*-ian.

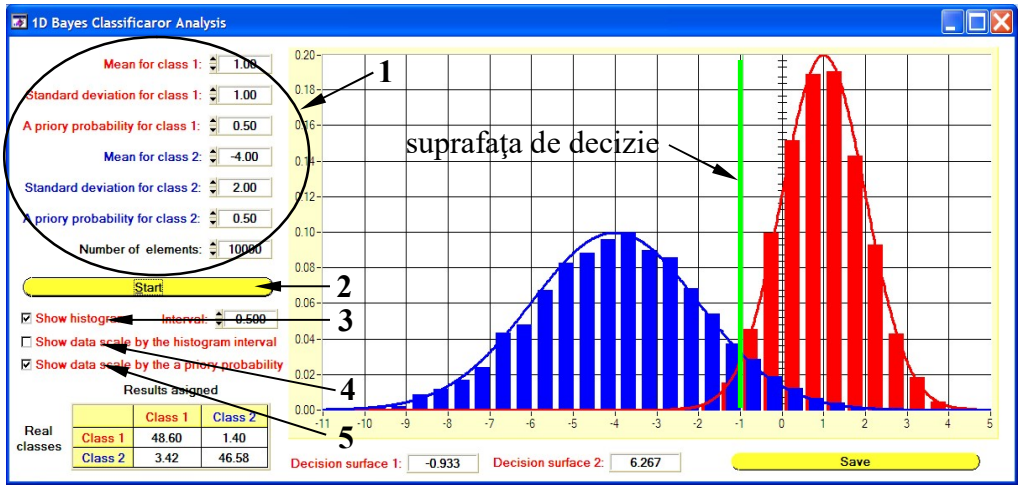


Figura 7.6. Interfaţa grafică a programului de analiză a clasificatorului *Bayes*-ian unidimensional

Aplicaţie 7.2: În directorul „***Bayes 1D***”, asociat acestui capitol, se găseşte codul sursă al unei aplicaţii *software* dezvoltată în mediul LabWindows CVI capabilă:

1. Să genereze două seturi de date distribuite conform a două funcţii densitate de probabilitate unidimensionale, $f(x|C_1)$ şi $f(x|C_1)$, ale căror parametri pot fi selectaţi de către utilizator (prin intermediul elementelor „1” de pe interfaţa grafică). Generarea unui nou set de date, cu noi parametri, se realizează prin apăsarea butonului „2”.
2. Să afişeze cele două funcţii densitate de probabilitate: $f(x|c_1)$ şi $f(x|c_2)$.
3. Să afişeze histogramele celor două seturi de date prin bifarea controlului „3”.
4. Să afişeze cele două funcţii densitate de probabilitate multiplicate cu intervalul pe care s-au determinat cele două histograme: $f(x|C_1)\cdot\Delta x$ şi $f(x|C_1)\cdot\Delta x$ – similar cu reprezentarea grafică din **Figura 7.4**. Această afişare se realizează prin selectarea elementului de control „4”.
5. Să prezinte suprafaţa/suprafeţele decizionale.
6. Să afişeze cele două funcţii densitate de probabilitate multiplicate cu probabilitatea apriorică a celor două clase: $f(x|c_1)\cdot P(c_1)$ şi $f(x|c_1)$

· $P(c_2)$. Acest mod de afişare este similar cu cel din **Figura 7.5** şi se realizează prin selectarea elementului „5”.

7. Să prezinte funcţiile densitate de probabilitate scalate prin ambele informaţii prezentate la subpunctele 4 şi 6 anterior prezentate.

Utilizând acest program se cere să se parcurgă următoarele cerinţe:

1. Pentru două clase echiprobabile şi având deviaţii standard egale, să se determine numărul soluţiilor obţinute. De ce au fost obţinute atâtea soluţii? Unde este poziţionată suprafaţa de decizie? De ce?
2. Păstrând parametrii de la punctul 1 constanţi, variaţi doar valoarea probabilităţii apriorice a uneia dintre clase. În ce mod se modifică poziţia suprafeţei de decizie? Către ce clasă? (*Notă*: după fiecare modificare a valorii probabilităţii apriorice a uneia dintre clase apăsaţi pe butonul „2” pentru vizualizarea rezultatelor).
3. Folosind aceleaşi valori ale parametrilor de a punctul 1, variaţi acum doar deviaţia standard a uneia sau alteia dintre clase şi observaţi simultan modificarea poziţiei suprafeţei/suprafeţelor decizionale. În ce condiţii/condiţie clasificatorul *Bayes*-ian va genera două suprafeţe de decizie?
4. Păstrând constante valorile parametrilor de la punctul 1, variaţi acum media uneia dintre clase. Observaţi variaţia performanţelor de clasificare obţinute de clasificatorul *Bayes*-ian pentru valori mai mari sau mai mici ale mediei clasei modificate comparativ cu situaţia iniţială.
5. Corelaţi modificarea suprafeţei/suprafeţelor decizionale generată de modificările de la subpunctele 2 şi 3 cu performanţele de clasificare vizualizate cu ajutorul matricii confuziilor prezentă pe interfaţa grafică a programului.

7.2.2 Regula de decizie

Utilizând principiile generale, precum şi relaţiile prezentate anterior, ne este foarte uşor acum să construim un clasificator. Pentru a selecta clasa optimă putem alege una din următoarele două abordări:

(a) Având un vector de trăsături a^0, calculăm mai întâi setul de probabilităţi posterioare:

$$P(c_i \mid a^0) = \frac{f_x(a^0 \mid c_i)P(c_i)}{P(a^0)} \quad (7.28)$$

pentru toate clasele posibile şi, ulterior, realizăm asignarea elementului a^0 la acea clasă pentru care valoarea probabilităţii posterioare este maximă, conform regulei:

a^0 aparţine clasei c_i dacă $P(c_i \mid a^0) \geq P(c_j \mid a^0)$ pentru $\forall j \neq i$ (7.29)

Schema bloc generală de implementare a clasificatorului *Bayes*-ian pentru M clase şi un vector de intrare, a^0, multidimensional este prezentată în **Figura 7.7.**

(b) În mod alternativ, putem determina mai întâi suprafaţa/suprafeţele de decizie ce separă cele M clase şi, ulterior, funcţie de poziţionarea vectorului de trăsături a^0 faţă de aceste suprafeţe de decizie, luăm decizia privind clasa optimală de apartenenţă a lui.

Astfel, spre exemplu, pentru ***Problema 7.3***, se calculează mai întâi pragul a_d şi uterior, tragem concluzia de „subiect odihnit” ($a^0 < a_d$) sau de „subiect obosit” ($a^0 > a_d$).

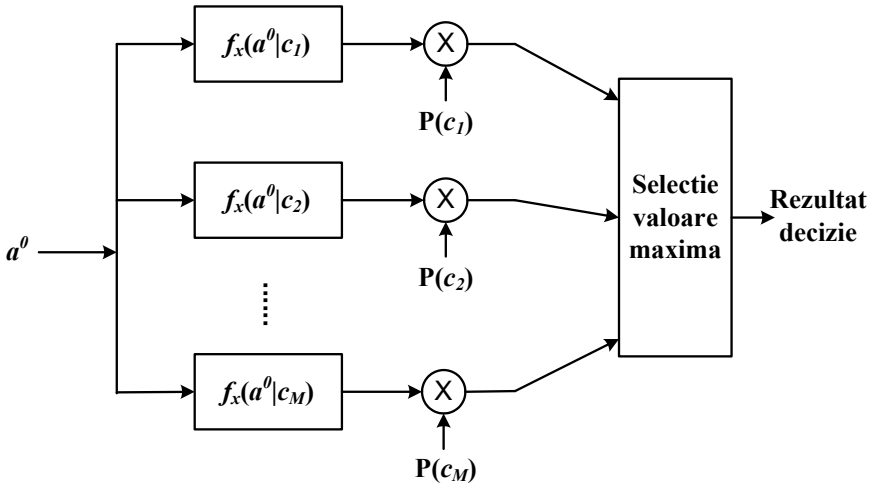


Figura 7.7. Schema bloc a clasificatorului *Bayes*-ian

7.2.3 Eroarea de clasificare

În general, asignarea elementelor la una din cele două sau mai multe clase posibile nu se poate realiza chiar fără nici un fel de eroare.

Exemplificând pe ***Problema 7.3*** (vezi şi **Figura 7.5**), se poate vedea cum "coada" distribuţiei de probabilitate pentru clasa indivizilor odihniţi, ce se extinde şi în dreapta punctului de decizie, ne dă eroarea de clasificare pentru clasa {*odihnit*} în timp ce "coada" distribuţiei de probabilitate pentru clasa indivizilor obosiţi, ce se extinde în stânga punctului de decizie, ne dă eroarea de clasificare pentru clasa {*obosit*}. În aceste condiţii, *eroarea de clasificare globală* este dată de suma celor două erori, respectiv, de *suma suprafeţelor mărginite de capetele funcţiilor de distribuţie*, $f_x(a|c_i)$, *axă şi pragul ales scalate la probabilitatea clasei repective*.

Observaţie 7.7: Cu cât suprapunerea este mai redusă cu atât eroarea de clasificare este şi ea mai mică. Întrucât regiunile de decizie depind de pragul a_d ales, rezultă că şi erorile de clasificare depind de poziţionarea acestui prag.

Ţinând cont de faptul că în majoritatea aplicaţiilor practice, reale, eroarea de clasificare obţinută cu diverşi clasificatori propuşi teoretic este una diferită de zero rezultă că rezolvarea unei probleme de clasificare revine la a găsi practic acel clasificator care să fie optimal.

Prin **clasificator optimal se înţelege acel clasificator pentru care probabilitatea de eroare de clasificare este minimă**.

Având în vedere faptul că modul de determinare a clasei dat de relaţia (7.13) minimizează probabilitatea de eroare, rezultă imediat de aici şi faptul că ***clasificatorul Bayes-ian este clasificator optimal***.

Observaţie 7.8: Atenţie! Reamintim aici că, prin clasificator optimal nu înţelegem un clasificator pentru care nu vom obţine nici o eroare de clasificare ci vom înţelege un clasificator cu care obţinem numărul minim de erori de clasificare posibile.

În cele ce urmează vom face o discuţie privind eroarea de clasificare pentru cazul unidimensional, cu două clase posibile, şi cu:

(i) o singură soluţie pentru ecuaţia suprafeţei de decizie şi, respectiv,
(ii) două soluţii pentru ecuaţia suprafeţei de decizie.

Discuţii:

(i) Pentru a determina eroarea de clasificare pentru ***Problema 7.3*** – pentru care avem o singură soluţie pentru ecuaţia suprafeţei de decizie – va trebui să integrăm ariile, R_1 şi R_2, care ne dau tocmai această eroare (vezi **Figura 7.8**). Acest lucru este simplu într-un spaţiu monodimensional dar devine dificil într-un spaţiu multidimensional.

În cazul particular al problemei prezentate anterior probabilitatea de eroare este:

$$P_{eroare} = \int_{R_2} f_x(a \mid c_1)P(c_1)da + \int_{R_1} f_x(a \mid c_2)P(c_2)da \quad (7.30)$$

sau

$$P_{eroare} = \int_{a_d}^{+\infty} f_x(a \mid c_1)P(c_1)da + \int_{-\infty}^{a_d} f_x(a \mid c_2)P(c_2)da \quad (7.31)$$

Eroarea este cu atât mai mică (respectiv, acurateţea clasificatorului este mai mare) cu cât supunerea dintre clase este mai mică. În mod intuitiv, pentru varianţe egale ale claselor, cu cât va fi mai mare distanţa dintre centrii clusterelor cu atât va fi mai mică suprapunerea între clase şi, respectiv, cu atât va fi mai mică şi eroarea de clasificare. În mod similar, pentru aceeaşi distanţă între centrele claselor eroarea va fi cu atât mai mică cu cât varianţa distribuţiei fiecărei clase, în parte, va fi mai mică.

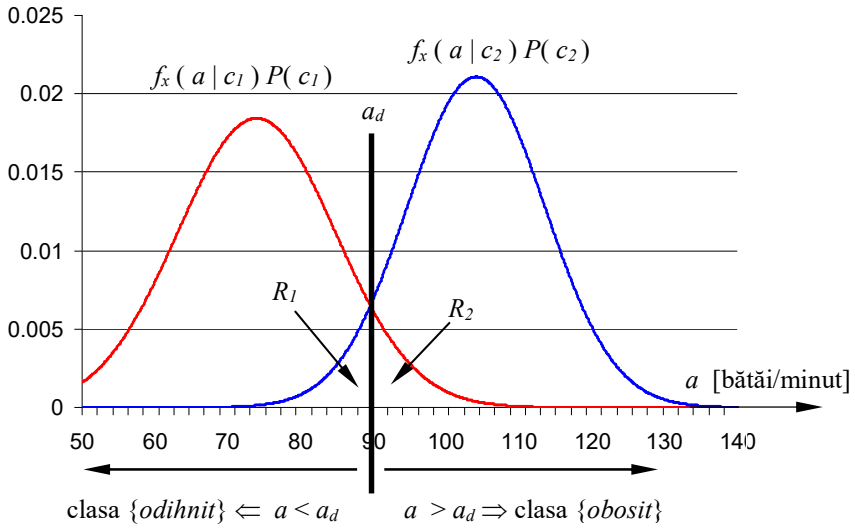


Figura 7.8. Probabilitatea de eroare în cazul clasificatorului *Bayess*-ian

(ii) În **Figura 7.9** este dată o reprezentare grafică pentru două clase arbitrare, caracterizate de două funcţii densitate de probabilitate g*auss*-iene. În acest caz se poate observa că ecuaţia suprafeţei de decizie are

două soluţii, a_{d1} şi a_{d2}. Pentru această situaţie eroarea de clasificare se calculează cu relaţia:

$$P_{eroare} = \int_{-\infty}^{a_{d1}} f_x(a \mid c_1)P(c_1)da + \int_{a_{d1}}^{a_{d2}} f_x(a \mid c_2)P(c_2)da + \int_{a_{d2}}^{+\infty} f_x(a \mid c_1)P(c_1)da \tag{7.32}$$

Mai general, probabilitatea medie a erorii de clasificare a clasificatorului Bayes-ian pentru cazul unei probleme cu două clase se scrie asfel:

$$P(eroare) = \int_{-\infty}^{+\infty} P(eroare, a)da = \int_{-\infty}^{+\infty} P(eroare \mid a)P(a)da \tag{7.33}$$

unde,

$$P(eroare \mid a) = \begin{cases} P(c_1 \mid a) & \text{dacă decidem clasa } c_2 \\ P(c_2 \mid a) & \text{dacă decidem clasa } c_1 \end{cases} \tag{7.34}$$

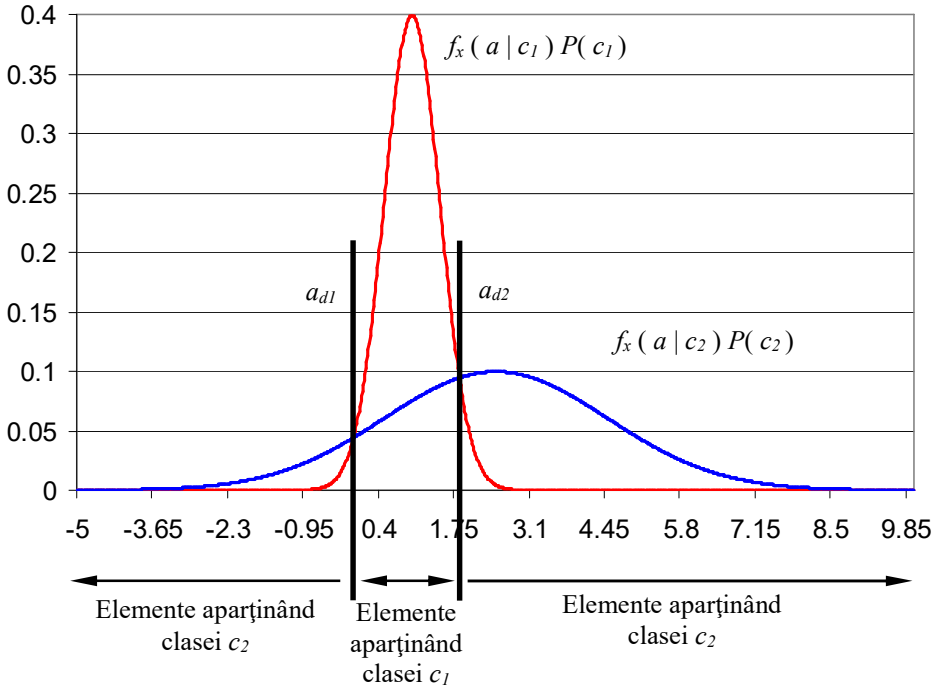


Figura 7.9. Probabilitatea de eroare în cazul clasificatorului *Bayes*-ian pentru două distribuţii arbitrare şi pentru două suprafeţe de decizie

În concluzie, factorii ce afectează eroarea de clasificare sunt, de fapt, o

combinaţie între diferenţa mediilor claselor, varianţele claselor şi probabilităţile apriorice ale acestora.

7.2.4. Un exemplu bidimensional de recunoaştere de *pattern*-uri

Exemplul prezentat anterior (unidimensional) este prea simplu pentru a pune în evidenţă:

- întreaga metodologie utilizată în determinarea poziţiei suprafeţelor de decizie – bazată pe modelarea statistică a setului de date;
- varietatea şi caracteristicile suprafeţei de decizie;
- anumite detalii şi dificultăţi ale proiectării clasificatorului.

Din aceste motive, în acest subcapitol vom trata cazul unui clasificator *Bayes*-ian bidimensional. Acest caz bidimensional a fost ales în principal datorită posibilităţii facile de vizualizare a seturilor de date şi de înţelegere intuitivă a fenomenelor implicate în procesul clasificării. Desigur, această metodă poate fi ulterior generalizată pentru orice dimensiune arbitrară, *d*, a vectorului de trăsături.

În cele ce urmează considerăm aceeași problemă ***Problema 7.3*** (clasificarea unei populații de subiecți în două clase, {*odihnit*} versus {*obosit*}, funcţie de activitatea cardiacă) doar că de această dată activitatea cardiacă a inimii va fi reprezentată prin doi parametri: numărul de bătăi pe unitatea de timp (minut) şi, respectiv, presiunea sistolică. Utilizăm aceste variabile aleatoare pentru construirea unui vector de trăsături aleator bidimensional ce va fi prezentat în continuare la intrarea clasificatorului *Bayes*-ian.

În continuare presupunem că funcţiile de distribuţie ale celor două clase sunt *gauss*-iene iar probabilităţile apriorice ale claselor sunt ambele egale cu ½ (clasele sunt echiprobabile).

Scopul principal al acestui subcapitol este acela *de a determina poziţia suprafeţei de decizie astfel încât să obţinem un clasificator optimal*.

Notă: Setul de date pe care îl vom prezenta şi analiza în acest subcapitol se găseşte în fişierul „***Odihnit-Obosit Puls-PresSist.txt***” din directorul asociat acestui subcapitol. O reprezentare grafică a unui eşantion de date de 10000 exemplare este redată, pentru această problemă, în **Figura 7.10**.

Conceptual metodologia de determinare a suprafeţelor de decizie este similară ca şi în cazul uni-dimensional al problemei anterior prezentate, unde suprafaţa de decizie s-a determinat din estimările făcute asupra mediei şi deviaţiei standard pentru cele două clase. În cazul multidimensional

diferenţele sunt date de faptul că mediile claselor sunt vectori *d*-dimensionali iar varianţele claselor sunt matrici de covarianţă, *d* x *d* dimensionale.

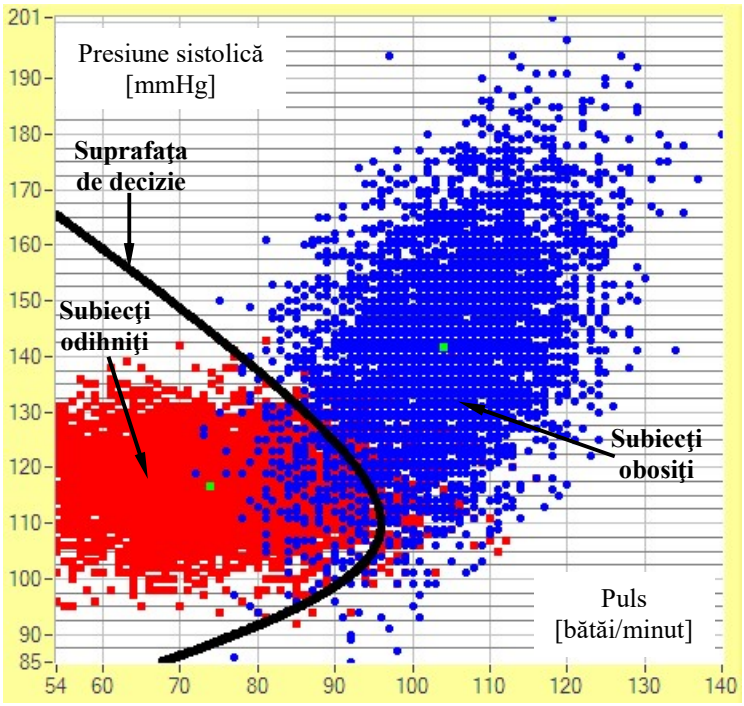


Figura 7.10. Distribuţia indivizilor funcţie de activitatea cardiacă (puls şi presiune sistolică) precum şi suprafaţa optimală de decizie

În modul cel mai general suprafaţa de decizie dintre două clase *i* şi *j* se găseşte prin egalarea funcţiilor discriminant ce caracterizează clasele (vedeţi ***Subcapitolele 3.5.2*** şi ***3.5.3***):

$$g_i(a) = g_j(a) \quad (7.35)$$

unde prin *g* s-au notat funcţiile discriminant ce caracterizează cele două clase.

În clasificatorul *Bayes*-ian regula de decizie este una exprimată, am văzut, în termenii probabilităţii posterioare a clasei (respectiv, probabilitatea clasei de a include un element particular analizat). Spre deosebitre de această formulare, o formă alternativă a regulei de decizie poate fi una exprimată în termenii funcţiilor discriminant, funcţii ce atribuie un scor maxim elementelor clasei pe care o caracterizează şi scoruri inferioare pentru vectorii de trăsături aparţinând celorlalte clase.

Făcând echivalenţa celor două forme de exprimare a regulei de decizie, obţinem:

$$g_i(a) = P(c_i \mid a) \quad \text{sau} \quad g_i(a) = f_x(a \mid c_i) \cdot P(c_i) \quad (7.36)$$

unde: $f_x(a|c_i)$ este probabilitatea elementului a condiţionată de clasa c_i.

Observaţie 7.9: În această ecuaţie am omis termenul $f_x(a)$ deoarece el este un factor comun tuturor discriminanţilor şi nu afectează, în esenţă, forma sau plasarea suprafeţei de decizie; din acest motiv poate fi ignorat în definirea funcţiei discriminant.

Existenţa exponenţilor în definirea lui $f_x(a|c_i)$ ne sugerează o alternativă de redefinire a funcţiilor discriminant prin intermediul logaritmului natural al relaţiei (7.36). În acest context, noile relaţii vor fi date de:

$$g_i(a) = ln\, f_x(a \mid c_i) + ln\, P(c_i) \quad (7.37)$$

Relaţia de mai sus reprezintă forma cea mai generală pentru funcţia discriminant, ea depinzând de funcţia de distribuţie a clasei c_i şi de probabilitatea apriorică a clasei, $P(c_i)$.

Pentru vectorul aleator bidimensional real x ($x = [x_1, x_2]^T$, x_1 – fiind prima trăsătură – pulsul în cazul nostru particular –, în timp ce x_2 este presiunea sistolică) funcţia densitate de probabilitate pentru clasa i este:

$$f_x(a|c_i) = \frac{1}{(2\pi)^{D/2}|C_i|^{1/2}} e^{-\frac{1}{2}(a-m_i)^T C_i^{-1}(a-m_i)} \quad (7.38)$$

unde indicele $i \in \{ od, ob \}$, indexează clasa od – {*odihnit*}, respectiv, clasa ob – {*obosit*}.

După logaritmarea relaţiei (7.38), funcţia discriminant devine:

$$g_i(a) = -\frac{1}{2}(a-m_i)^T C_i^{-1}(a-m_i) - \frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|C_i| + \ln P(c_i) \quad (7.39)$$

Ţinând cont de faptul că cele două clase sunt echiprobabile, ultimul termen din ecuaţia (7.39) va fi acelaşi pentru fiecare dintre cele două funcţii discriminant în parte, el putând fi astfel eliminat. Acelaşi lucru este valabil şi pentru termenul (D/2) ln(2π).

Din relaţia (7.39) se observă că, *în mod similar problemei de clasificare uni-dimensionale*, *ceea ce contează în definirea funcţiei discriminant este distanţa dintre element şi media clasei*, *normalizată la matricea de covarianţă. Analizând ecuaţiile funcţiilor discriminant pentru fiecare clasă, în scopul găsirii suprafeţelor de decizie, se observă că soluţia căutată este*, de fapt, *o funcţie dată de o distanţă normalizată* (*distanţă numită distanţă*

Mahalanobis[9]*), de matricile de covarianţă şi de probabilităţile apriori ale celor două clase.*

Tabelul 7.1. Estimarea vectorilor medii şi a matricelor de covarianţă ale claselor pentru două seturi de date de dimensiuni diferite

	10000 de elemente	100 de elemente
Clasa {*odihnit*}	$m_{od}=\begin{bmatrix}73.9\\116.7\end{bmatrix}$ $C_{od}=\begin{bmatrix}117 & -4\\-4 & 53.2\end{bmatrix}$	$m_{od}=\begin{bmatrix}72.8\\114\end{bmatrix}$ $C_{od}=\begin{bmatrix}114.4 & -12.6\\-12.6 & 53.9\end{bmatrix}$
Clasa {*obosit*}	$m_{ob}=\begin{bmatrix}104\\141.7\end{bmatrix}$ $C_{ob}=\begin{bmatrix}89.5 & 70\\70 & 277.4\end{bmatrix}$	$m_{ob}=\begin{bmatrix}102.8\\142.8\end{bmatrix}$ $C_{ob}=\begin{bmatrix}108.9 & 86.3\\86.3 & 300.3\end{bmatrix}$

În tabelul de mai sus se prezintă valorile estimate pentru covarianţele şi mediile claselor atunci când sunt luate în considerare 100, respectiv, 10000 de realizări particulare ale vectorului aleator pentru fiecare clasă în parte.

Suprafaţa de decizie a clasificatorului optimal (clasificatorul *Bayes*-ian) se obţine în aceeaşi manieră ca şi în cazul unidimensional, prin înlocuirea mediilor şi a matricilor de covarianţă estimate din datele de intrare – informații prezentate în **Tabelul 7.1 –**, în relația (7.39) și egalarea funcțiilor discriminant pentru cele două clase.

Inversele matricilor de covarianţă pentru ambele clase sunt următoarele:

$$C_{od}^{-1}=\begin{bmatrix}8.5\cdot10^{-3} & 6.5\cdot10^{-4}\\6.5\cdot10^{-4} & 1.8\cdot10^{-2}\end{bmatrix} \quad (7.41)$$

9 Reamintim că această distanţă normalizată este dată de:

$$d^2=(a-\hat{m}_x)^T\,\hat{C}_x^{-1}\,(a-\hat{m}_x) \quad (7.40)$$

În relaţia (7.40) $\hat{m}_x$ reprezintă vectorul mediu estimat al clasei iar $\hat{C}_x$ reprezintă estimatul matricei de covarianţă.

$$C_{ob}^{-1} = \begin{bmatrix} 1.4 \cdot 10^{-2} & -3.5 \cdot 10^{-3} \\ -3.5 \cdot 10^{-3} & 4.5 \cdot 10^{-3} \end{bmatrix} \quad (7.42)$$

Determinanţii acestor matrici sunt următorii $det(C_{od}) = 6214.7$ şi $det(C_{ob}) = 19926$. Introducând toate aceste valori în (7.39) obţinem:

$$g_{od}(a) = -\frac{1}{2}[a_1 - 73.9 \cdot a_2 - 116.7] \cdot \begin{bmatrix} 8.5 \cdot 10^{-3} & 6.5 \cdot 10^{-4} \\ 6.5 \cdot 10^{-4} & 1.8 \cdot 10^{-2} \end{bmatrix} \cdot \begin{bmatrix} a_1 - 73.9 \\ a_2 - 116.7 \end{bmatrix} - \frac{1}{2}\ln(6214.7) \quad (7.43)$$

Pentru clasa subiecţilor obosiţi discriminantul este:

$$g_{ob}(a) = -\frac{1}{2}[a_1 - 104 \cdot a_2 - 141.7] \cdot \begin{bmatrix} 1.4 \cdot 10^{-2} & -3.5 \cdot 10^{-3} \\ -3.5 \cdot 10^{-3} & 4.5^{-3} \end{bmatrix} \cdot \begin{bmatrix} a_1 - 104 \\ a_2 - 141.7 \end{bmatrix} - \frac{1}{2}\ln(19926) \quad (7.44)$$

Suprafaţa de decizie se obţine egalând funcţiile discriminant ale celor două clase:

$$g_{od}(a) = g_{ob}(a) \quad (7.45)$$

Ecuaţia suprafeţei de decizie ce se obţine în urma calculelor este :

$$a_1^2 - 8.15\, a_2^2 + 25228.58\, a_1 + 20895.12\, a_2 + 177.6\, a_1\, a_2 - 2808973.26 = 0 \quad (7.46)$$

Această suprafaţă este o cuadratică în spaţiul bidimensional al trăsăturilor, suprafaţă ce determină obţinerea celei mai mici erori de clasificare pentru problema de clasificare abordată aici. O reprezentare grafică a celor două clase precum şi a suprafeţei de decizie optimale obţinute este dată în **Figura 7.10**.

Problemă 7.4: Pornind de la relaţia (7.39) să se determine, pentru cazul bidimensional al vectorului aleator de trăsături *x,* formula generală ce descrie ecuaţia suprafeţei de decizie (formulă similară relaţiei (7.46)). Implementaţi această relaţie într-un program scris în limbajul LabWindows CVI şi verificaţi corectitudinea ei prin reprezentarea grafică a suprafeţei de decizie.

Din simpla vizualizare a claselor şi, respectiv, a poziţiei suprafeţei de decizie, vezi **Figura 7.10,** observăm faptul că, chiar şi în aceste condiţii, a utilizării unui clasificator optimal, se pot obţine, aşa cum am menţionat deja anterior, foarte multe erori de clasificare. Reamintim aici faptul că ***un clasificator optimal***, ai cărui parametri sunt corect estimaţi, ***va obţine – dintre toate familiile de clasificatori existenţi –, nu eroare zero ci cea mai mică eroare de clasificare pentru problema de clasificare abordată***.

În spaţii multidimensionale, estimarea corectă, cu o precizie adecvată, a tuturor elementelor matricelor de covarianţă aferente claselor este, în general, o sarcină dificilă. În practică, matricele de covarianţă, precum şi mediile claselor, sunt adesea estimate nesatisfăcător, aceasta şi datorită unor seturi de date insuficient de mari, ceea ce conduce, în mod direct, şi la obţinerea unor suprafeţe de decizie cu performanţe suboptimale.

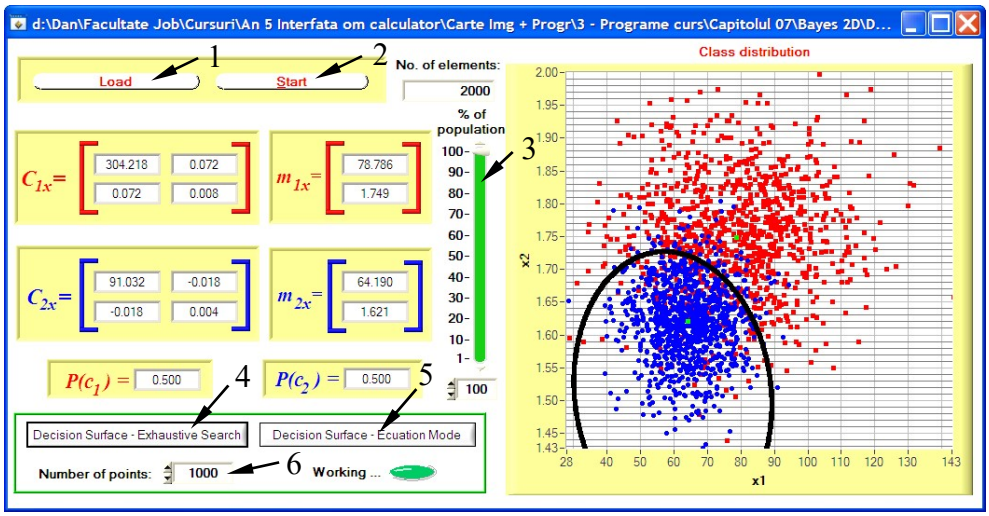


Figura 7.11. Interfaţa grafică a programului utilizat în clasificarea unor seturi de date caracterizate de vectori bidimensionali de trăsături.

Aplicaţie 7.3: În cadrul acestei exemplu se analizează, cu ajutorul unei aplicaţii practice, un clasificator *Bayes*-ian bidimensional. Codul programului şi *kit*-ul de instalare se găsesc în directorul „***Bayes 2D***” asociat acestui capitol. Acest program permite încărcarea oricărui set de date organizat pe trei coloane. Pe primele două coloane sunt valorile numerice ale celor două trăsături ce formează vectorul de trăsături, în timp ce elementele de pe ultima coloană pot lua doar valori egale cu 0 sau 1 şi furnizează informaţia de apartenenţă a elementului la una din cele două clase. Încărcarea setului de date se realizează prin apăsarea butonului „1”, vezi **Figura 7.11**. O dată cu încărcarea setului de date acesta se va afişa automat. Prin apăsarea butonului „2” se iniţiază procesul de calcul a parametrilor clasificatorului *Bayes*-ian, parametri ce sunt ulterior prezentaţi pe interfaţa grafică. Cu ajutorul acestui program se cere atingerea următoarele obiective:

1. Încărcaţi unul din următoarele seturi de date: "***Odihnit-Obosit Puls-PresSist rnd.txt***", "***Odihnit-Obosit Puls-PresSist.txt***", "***Femei-Barb Greut-Inalt Bayes rnd.txt***", "***Femei-Barb Greut-Inalt Bayes.txt***" sau "***Femei-Barb Inalt-Greut Bayes.txt***".
2. Apăsaţi bunonul „Start" pentru determinarea parametrilor statistici ai celor două clase. Verificaţi în mod intuitiv corelaţia existentă între dispunerea spaţială a celor două clase şi coeficienţii matricilor de covarianţă obţinute.
3. Verificaţi influenţa mărimii setului de date asupra estimării matricilor de covarianţă şi a vectorilor medii. Puteţi modifica mărimea setului de date utilizat de clasificator prin schimbarea valorii controlului „3" de pe interfaţa grafică. Este recomandat ca în cadrul acestui subpunct să utilizaţi unul din seturile de date ce au vectorii de trăsături aparţinând celor două clase amestecaţi într-un mod aleator (fişierele ce conţin aceste seturi de date au incluse în numele lor caractele „***rnd***").
4. Reprezentaţi grafic suprafaţa de decizie dintre cele două clase. În cadrul programului sunt implementate două metode de determinare a acestei suprafeţe de decizie. În prima din ele, se realizează o parcurgere exhaustivă a spaţiului de intrare într-un număr de paşi selectabili din controlul „6". În cea de a doua metodă se determină ecuația suprafeței de decizie, obținându-se o relație similară cu (7.46), care ulterior este reprezentată. Înţelegeţi codurile acestor două subrutine.
5. Observaţi variabilitatea acestor suprafeţe de decizie funcţie de mărimea setului de date.

Aplicaţie 7.4: Pentru programul prezentat în cadrul ***Aplicaţiei 7.1*** parcurgeţi paşii 1-5 încă o dată însă, de această dată, folosind şi clasificatorul *Bayes*-ian.

Utilizând setul de date „***Femei-Barb Greut-Inalt Bayes rnd.txt***" trasaţi suprafeţele de decizie generate de clasificatorii *Mahalanobis* şi *Bayes*-ian şi analizaţi, folosind matricea confuziilor, performanţele obţinute.

Trasaţi simultan suprafeţele de decizie pentru clasificatorul *Mahalanobis* şi pentru clasificatorul *Bayes*-ian utilizând, într-o primă fază setul de date „***6 Classes.txt***" iar, ulterior, setul de date „***6 Classes - Different prob.txt***". Puneţi în evidenţă modificările ce au loc în

cadrul clasificatorului *Bayes*-ian atunci când probabilităţile apriorice ale claselor nu mai sunt egale.

7.2.5. Sensibilitatea funcţiilor discriminant funcţie de mărimea setului de date

În subcapitolele precedente am prezentat o metodă cu ajutorul căreia am reuşit să determinăm suprafaţa optimă de decizie în ipoteza că funcţia de distribuţie de probabilitate pentru fiecare clasă este de tip densitate *gauss*-iană. Această metodă de clasificare este una foarte puternică însă ea are şi următoarele dezavantaje:

1. **se bazează pe anumite ipoteze în legătura cu distribuţia setului de date**, ipoteze asupra formei funcţiilor densitate de probabilitate ce caracterizează setul de date şi, în plus,
2. metoda **necesită un volum mare de date** pentru estimarea, cu erori mici, a parametrilor funcţiilor discriminant.

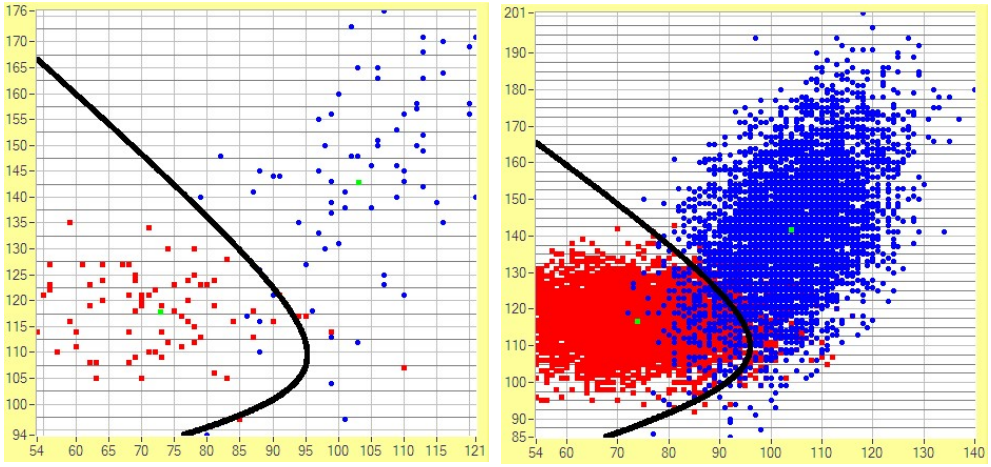


Figura 7.12. Poziţionarea suprafeţelor de decizie pentru mărimi diferite ale seturilor de date **(a).** 100 de elemente, şi **(b).** 10000 de elemente.

În consecinţă, şi *calitatea rezultatelor obţinute depinde*, pe de o parte, *de validitatea ipotezelor făcute*, iar pe de altă parte, de *calitatea valorilor estimate obţinute pentru parametrii populaţei* – calitate ce depinde direct de mărimea eşantionului de date.

În acest subcapitol vom ilustra efectul mărimii setului de date asupra corectitudinii estimării parametrilor funcţiilor discriminant.

Exemplul 7.3: Să presupunem că avem numai 100 de realizări particulare (vectori bidimensionali de trăsături: a = [puls, presiune sistolică]^T) din care 50 de elemente corespund unor subiecţi aleşi în mod aleator din clasa {*odihnit*} iar 50 de vectori de trăsături aparţin unor subiecţi aleşi în mod aleator din clasa {*obosit*}. Dacă calculăm media şi matricea de covarianţă pentru fiecare clasă în parte obţinem rezultatele din **Tabelul 7.1**. Din analiza acestui tabel se observă că valorile diferiţilor parametri se modifică faţă de cele calculate pe setul de 10000 de vectori de trăsături. În momentul când construim, pe baza acestor parametri, funcţia discriminant optimă pentru eşantionul de 100 de elemente observăm că forma ei este una similară cu cea obţinută pe eşantionul mai mare de date, vezi **Figura 7.12**, însă poziţia în spaţiul de intrare se modifică, ceea ce duce în mod corespunzător şi la modificarea probabilităţii de eroare (în particular, avem de-a face, aşa cum era de aşteptat, cu o creştere a acestei probabilităţi).

Tabelul 7.2. Estimarea parametrilor statistici pentru diferite mărimi ale seturilor de date

	100 de subiecţi	1000 de subiecţi
Femei	$m_f = \begin{bmatrix} 1.626 \\ 61.969 \end{bmatrix}$ $C_f = \begin{bmatrix} 0.004 & -0.182 \\ -0.182 & 104.9 \end{bmatrix}$	$m_f = \begin{bmatrix} 1.619 \\ 64.339 \end{bmatrix}$ $C_f = \begin{bmatrix} 0.003 & -0.023 \\ -0.023 & 93 \end{bmatrix}$
Bărbaţi	$m_{ob} = \begin{bmatrix} 1.731 \\ 77.469 \end{bmatrix}$ $C_{ob} = \begin{bmatrix} 0.009 & 0.117 \\ 0.117 & 249.795 \end{bmatrix}$	$m_{ob} = \begin{bmatrix} 1.744 \\ 78.64 \end{bmatrix}$ $C_{ob} = \begin{bmatrix} 0.007 & 0.152 \\ 0.152 & 313.256 \end{bmatrix}$

Clasa cea mai afectată în sensul modificării valorilor estimate ale parametrilor este clasa subiecţilor obosiţi (modificări cu aproape cca 20 de puncte ale elementelor de pe diagonala principală a matricii de covarinaţă, atunci când se trece de la o analiză a unui eşantion format din 10000 de elemente la unul format din doar 100 de elemente).

Exemplul 7.4: Într-un alt exemplu (problema discriminării în două clase, {*femei*} şi {*bărbaţi*}, a unui eşantion de subiecţi, caracterizat de vectorul aleator x = [înălţime, greutate]^T), se observă, conform **Tabelul 7.2**, că în cazul trecerii de la un eşantion iniţial de 1000 de

subiecţi (net mai mic faţă de cel prezentată în exemplul anterior, de 10000 de elemente) la unul de 100 de subiecţi, se obţin, de exemplu, variaţii de cca 60 de unităţi în determinarea elementelor matricii de corelaţie pentru clasa {bărbaţi}.

O primă concluzie ce s-ar putea desprinde din aceste două exemple ar fi aceea că „sensibilitatea" parametrilor clasificatorului Bayes-ian este puternic dependentă de tipul problemei analizate (de exemplu, de particularităţile trăsăturilor analizate) precum şi de specificităţile setului de date.

O analiză suplimentară a modului cum se modifică probabilitatea de eroare la o modificare a dimensiunii eşantionului ne aduce un surplus informaţional important şi anume: scăderea dimensiunii eşantionului se însoţeşte de o creştere a erorii de clasificare şi invers; o creştere a performanţelor clasificatorului atunci când estimarea parametrilor se face pe un eşantion mai mare de date se explică prin aceea că valoarea unui estimator, în general, tinde spre valoarea reală a parametrului populaţiei atunci când dimensiunea eşantionului tinde, la limită, spre infinit.

Problema de clasificare abordată în ***Subcapitolul 7.2.4*** a fost una relativ foarte simplă, în doar două dimensiuni, 6 parametri trebuind să fie estimaţi pentru fiecare clasă (3 parametri ai matricii de covarianţă, 2 parametri ai vectorului mediu şi probabilitatea apriorică a clasei) din cele 50 de elemente din cât este formată o clasă. Într-un spaţiu multidimensional se poate întâmpla ca numărul parametrilor ce trebuie să fie estimaţi să fie de acelaşi ordin de cardinalitate cu mărimea setului de date – în această situaţie suprafaţa de decizie va fi poziţionată cu certitudine departe de cea optimă datorită, în principal, estimării nesatisfăcătoare a parametrilor clasificatorului. Pentru o analiză şi descriere completă a acestei probleme citiţi şi subcapitolul în care se analizează avantajele şi dezavantajele clasificatorului Mahalanobis (vezi ***Subcapitolul 7.1.2***).

Găsirea, analiza şi implementarea unor clasificatori care să fie mai puţin sensibili la estimarea parametrilor necunoscuţi ai populaţiei este un obiectiv principal a teoriei clasificatorilor.

În literatură s-a ajuns la concluzia că pentru a atinge acest obiectiv trebuie ca ***forma funcţională a funcţiilor discriminant să fie cât mai simplă*** [Principe, 2000]. ***În principal este de dorit să utilizăm funcţii discriminant care au cât mai puţini parametri*** iar aceşti parametri să poată fi estimaţi într-un mod cât mai robust posibil din setul de date pe care îl avem la dispoziţie. Aparent, astfel de funcţii discriminant, mai simple, pot fi unele suboptimale pentru problema dată; experienţa însă ne arată că, frecvent, cu aceste funcţii suboptimale putem obţine performanţe mai bune

decât cele generate de un clasificator, teoretic, optimal. Acest lucru poate părea un paradox dar explicaţia este una ce ţine de estimarea, de cele mai multe ori, inexactă a parametrilor clasificatorului optimal. Astfel, chiar dacă folosim funcţii discriminant cuadratice (considerate optimale pentru clase cu distribuţie de probabilitate *gauss*-iană), similare cu cele din relația (7.46), acestea pot fi inexact poziționate, vezi **Figura 7.12(b)**, obținându-se astfel erori destul de mari de clasificare.

7.2.6. Selecţia trăsăturilor bazată pe funcţia densitate de probabilitate posterioară

Dintr-un proces oarecare se pot extrage un număr foarte mare de trăsături. O dată cu creşterea numărului de trăsături folosite, creşte însă şi complexitatea clasificatorului. De aceea se pune problema selecţiei acelor trăsături care sunt purtătoare de informaţie discriminantă maximă, necesară în procesul de clasificare.

Pentru un clasificator *Bayes*-ian se defineşte *riscul condiţionat* ca fiind dat de relaţia:

$$R(c_i \mid a) = 1 - f_x\left(a \mid c_i\right)P\left(c_i\right) \quad (7.47)$$

La relaţia de mai sus s-a ajuns astfel:

- fie $\{\alpha_1, ..., \alpha_M\}$ mulţimea finită a deciziilor posibile, cu $\{\alpha_1, ..., \alpha_M\} \subseteq \{ c_1, ..., c_M \}$;
- se defineşte o *funcţie de pierdere*, $\lambda(\alpha_i|c_j)$, ce stabileşte *pierderea* pe care o atrage după sine o decizie greşită a unei clase α_i în locul clasei c_j reale;
- presupunem că observăm o realizare particulară, a, a vectorului aleator multidimensional, x, pentru care luăm decizia de apartenenţă la clasa α_i; dacă adevărata clasă de apartenenţă a lui a este c_j atunci pierderea pe care o obţinem în acest caz este $\lambda(\alpha_i|c_j)$;
- întrucât $P(c_j|a)$ este probabilitatea clasei adevărate atunci, pierderea medie (numită şi *risc condiţionat*, în teoria deciziei), generată de alegerea clasei α_i, este dată de relaţia:

$$R(\alpha_i \mid a) = \sum_{j=1}^{M} \lambda(\alpha_i \mid c_j)P(c_j \mid a) \quad (7.48)$$

- pentru cazul particular al funcţiei de pierdere de tip zero-unu, definită astfel:

$$\lambda(\alpha_i \mid c_j) = \begin{cases} 0 & , \textit{pentru } i = j \\ 1 & , \textit{pentru } i \neq j \end{cases}, \text{ cu } i, j = \overline{1, M} \quad (7.49)$$

riscul condiţionat definit prin relaţia (B041) devine chiar probabilitatea medie a erorii fiind dat de:

$$R(\alpha_i \mid a) = \sum_{j=1}^{M} \lambda(\alpha_i \mid c_j) P(c_j \mid a) = \sum_{j \neq i} P(c_j \mid a) = 1 - P(c_i \mid a) \quad (7.50)$$

în timp ce eroare globală de clasificare este dată de:

$$\varepsilon = \sum_{i=1}^{c} \sum_{j=1}^{c} \int_{R_j, j \neq i} \left(1 - R(c_i \mid a)\right) da \quad (7.51)$$

unde *a* este un vector de trăsături. Din relaţia de mai sus se poate scrie:

$$\varepsilon = \sum_{i=1}^{c} \sum_{j=1}^{c} P(c_i) \int_{R_j, j \neq i} f_x(a \mid c_i)\, da \quad (7.52)$$

O expresie asemănătoare se poate scrie şi pentru fiecare trăsătură în parte:

$$\varepsilon_k = \sum_{i=1}^{c} \sum_{j=1}^{c} P(c_i) \int_{R_{j(k)}, j \neq i} f_x(a_k \mid c_i)\, da_k \quad (7.53)$$

Pentru a selecta acele trăsături purtătoare de o cantitate maximă de informaţie discriminatorie se pot urma paşii: (1). pentru fiecare trăsătură în parte se calculează ε_k, cunoscând aprioric sau estimând $P(c_i)$ şi $f_x(a_k \mid c_i)$, iar apoi, (2). se reţin acele trăsături pentru care ε_k este minim (mai mic decât un anumit prag) şi se elimină celelalte trăsături, considerându-se că sunt improprii pentru o clasificare corectă a setului de date.

7.3. Probleme

1. (a) Care este ideea de bază, fundamentală, care stă în spatele clasificatorului *Bayes*-ian (exprimaţi această idee prin cuvinte)? Ce reprezintă fiecare termen din relaţia lui *Bayes* – relaţie utilizată în cadrul clasificatorului *Bayes*-ian?
   (b) Clasificatorul *Bayes*-ian este considerat un clasificator optimal. Ce înţelegeţi prin clasificator optimal?
   (c) Explicaţi influenţa varianţei trăsăturilor asupra poziţionării suprafeţei de decizie.
   (d) Explicaţi influenţa probabilităţii apriorice a clasei asupra poziţionării suprafeţei de decizie.
   (e) Explicaţi influenţa setului de date asupra performanţelor clasificatorului *Bayes*-ian.
   (f) Desenaţi schema bloc a clasificatorului *Bayes*-ian.

2. Fie două clase distincte caracterizate de două funcţii de densitate de probabilitate *gauss*-iene unidimensionale. Prima clasă este de medie 1.5 şi varianţă 0.04 în timp ce cea de a doua clasă are media 2 şi varianţa 0.64. Probabilitatea apriorică a primei clase este 1/5 iar a celei de a doua clase 4/5.
   (a) Desenaţi în acelaşi grafic ambele funcţii densitate de probabilitate. Reprezentarea grafică va fi una calitativă dar va ţine cont în reprezentare de parametrii celor două funcţii de densitate.
   (b) Scrieţi relaţia matematică ce caracterizează funcţia discriminant pentru elementele primei clase.
   (c) Desenaţi schema bloc a clasificatorului *Bayes*-ian ce utilizează două funcţii discriminant caracteristice fiecărei clase în parte.
   (d) Determinaţi, în mod numeric, poziţia exactă a suprafeţelor de decizie.
   (e) Realizarea particulară 0.5, a variabilei aleatoare x, aparţine primei clase sau a celei de a doua ? Dar 1.4 ? Dar 1.8 ? Pentru fiecare valoare justificaţi, în mod numeric, decizia luată.
   (f) Dacă cea de a doua clasa va avea o probabilitate apriorică mai mică ce se va întâmpla cu fiecare suprafaţă de decizie în parte? Justificaţi-vă răspunsul în mod analitic sau printr-o analiză conceptuală.

3. Să presupunem că un vector de trăsături mono-dimensional x este utilizat pentru a decide între clasa ω_1 şi ω_2. Probabilitatea aproirică a clasei ω_1 este 2/7. Funcţiile densitate de probabilitate pentru x în ipoteza apartenenţei la clasele ω_1 şi ω_2 sunt date de relaţiile:

$$f_x(a|\omega_1) = \begin{cases} \frac{1}{4}, & -3 \le a \le 1 \\ 0, & \text{în rest} \end{cases} \quad (7.54)$$

$$f_x(a \mid \omega_2) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(a-m_x)^2}{2\sigma_x^2}} \quad (7.55)$$

   (a). Demonstraţi ca (7.54) este o funcţie de tip densitate de probabilitate legitimă (verificaţi proprietăţile acesteia).
   (b). Care este probabilitatea apriorică pentru clasa ω_2. Sunt clasele ω_1 şi ω_2 echiprobabile?
   (c). Determinaţi pragul selecţiei optimale a clasei ω_2 versus ω_1 (pentru situaţia în care σ_x = ½ iar m_x = 2).
   (d). Reprezentaţi grafic modalitatea de obţinere a suprafeţei de decizie în cazul utilizării unui clasificator *Bayes*-ian. Cum ar arăta această reprezentare grafică dacă m_x = 1.5?

4. O variabilă aleatoare x este utilizată pentru a discrimina două clase, ω_1 şi ω_2. Pentru prima clasă ştim că $P(\omega_1)$ = 1/6 iar pentru cea de a doua clasă $P(\omega_2)$ = 5/6. Funcţiile densitate de probabilitate pentru variabila aleatoare x în ipoteza apartenenţei la clasele ω_1 şi ω_2 sunt date de:

$$f_x(a|\omega_1) = \begin{cases} \frac{1}{2}, & 2 \le a \le 4 \\ 0, & \text{în rest} \end{cases} \quad (7.56)$$

$$f_x(a \mid \omega_2) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(a-m_x)^2}{2\sigma_x^2}} \quad (7.57)$$

   (a). Demonstraţi că relaţia (1) este o funcţie densitate de probabilitate legitimă (verificaţi proprietăţile acesteia).
   (b). Care este semnificaţia termenului echiprobabil? Sunt clase ω_1 şi ω_2 echiprobabile?
   (c). Determinaţi pragul selecţiei optimale a clasei ω_1 versus ω_2 pentru situaţia în care σ_x = ½ iar m_x = 1.5).

(d). Reprezentaţi grafic funcţiile densitate de probabilitate ale celor două distribuţii. Această reprezentare grafică va fi corelată cu rezultatele obţinute la punctele anterioare.

Valori ce pot fi utilizate în determinările numerice:

$\ln 1 = 0$;	$\ln 2 \approx 0.7$;
$\ln 3 \approx 1$;	$\ln 4 \approx 1.4$;
$\ln 5 \approx 1.6$;	$\ln 6 \approx 1.8$;
$\ln 7 \approx 1.9$;	$\ln 8 \approx 2$;
$\ln 9 \approx 2.2$;	$\ln 10 \approx 2.3$.
$\sqrt{2\pi} \approx 2.5$	

5. O variabilă aleatoare x este utilizată pentru a discrimina două clase echiprobabile, ω_1 şi ω_2. Funcţiile densitate de probabilitate pentru variabila aleatoare x în ipoteza apartenenţei la clasele ω_1 şi ω_2 sunt date de:

$$fx(a \mid \omega_1) = \begin{cases} \frac{1}{2}, & 0 \le a \le 2 \\ 0, & \text{în rest} \end{cases} \quad (7.58)$$

$$f_x(a|\omega_2) = \begin{cases} 0 & a < 1 \\ 0.4 \cdot a - 0.4 & 1 \le a \le 3 \\ 5.6 - 1.6 \cdot a & 3 < a \le 3.5 \\ 0 & a > 3.5 \end{cases} \quad (7.59)$$

(a) Demonstraţi că relaţia (7.58) este o funcţie densitate de probabilitate legitimă (verificaţi proprietăţile acesteia).
(b) Determinaţi pragul selecţiei optimale a clasei ω_1 versus ω_2.
(c) Generaţi o reprezentare grafică în care să se pună în evidenţă existenţa pragului de selecţie ce separă cele 2 clase. Notaţi pe această prezentare grafică funcţiile pe care le reprezentaţi. Această reprezentare grafică va fi corelată cu rezultatul obţinut la punctul anterior.
(d) Determinaţi probabilitatea de eroare a acestui clasificator.
(e) Determinaţi probabilitatea ca o valoare din eşantion să fie corect clasificată.
(f) Demonstraţi că relaţia (7.59) este o funcţie densitate de probabilitate legitimă (verificaţi proprietăţile acesteia).
(g) Explicaţi în mod intuitiv ce se va întâmpla cu poziţia suprafeţei de decizie şi de ce se va întâmpla acest lucru (comparativ cu situaţia

anterioară existentă în rezolvarea punctelor (a)-(f)) dacă probabilităţile apriorice ale claselor vor fi următoarele: $P(\omega_1) = 2/6$ şi, respectiv, $P(\omega_2) = 4/6$.

6. Două funcţii densitate de probabilitate date de:

$$f_x\left(a|\omega_1\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{a^2}{2}} \quad (7.60)$$

$$f_x\left(a|\,\omega_2\right) = \frac{4}{\sqrt{2\pi}} e^{-(2a-1)^2} \quad (7.61)$$

caracterizează două clase (ω_1 şi ω_2) ce au probabilităţile apriorice ale claselor $P(\omega_1) = \frac{4}{5}$ şi $P(\omega_2) = \frac{1}{5}$.

(a) Determinaţi poziţia optimă a suprafeţei de decizie folosindu-se pentru aceasta un clasificator *Bayes*-ian.

(b) Determinaţi probabilitatea de eroare a clasificatorului *Bayes*-ian pentru datele de la punctul anterior.

(c) Presupunând de această dată cele două clase (ω_1 şi ω_2) echiprobabile, definite de următoarele funcţii densitate de probabilitate:

$$f_x\left(a|\omega_1\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(a-4)^2}{2}} \quad (7.62)$$

şi, respectiv,

$$f_x\left(a|\omega_1\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(a-6)^2}{2}} \quad (7.63)$$

determinaţi poziţia suprafeţei de decizie optimale fără efectuarea nici unui calcul prealabil. Argumentaţi-vă decizia.

7. (a) Demonstraţi invarianţa distanţei Mahalonobis $r^2 = (x - m_x)^T C_x^{-1} (x - m_x)$ la orice transformare liniară de forma $\mathbf{y} = \mathbf{A}\,\mathbf{x}$.

(b) Cum utilizaţi distanţa Mahalanobis în cadrul unui clasificator de tip minimă distanţă ?

(c) Care sunt avantajele utilizării metricii Mahalanobis ?

(d) Care este dezavantajul major al utilizării metricii Mahalanobis ? Discutaţi această problemă în paralel şi în cazul clasificatorului *Bayes*-ian.

(e) Care este ideea de bază în cazul clasificatorului *Bayes*-ian ?

(f) În ce condiţii un element va fi asignat unei clase în cazul clasificatorului *Bayes*-ian ? (prezentaţi inclusiv relaţia care asigură această asignare şi explicaţi fiecare termen al acesteia).