1η εργασία για το μάθημα «Αναγνώριση προτύπων»

Σημειώσεις:
1. Η παρούσα εργασία είναι η πρώτη από 2 συνολικά εργασίες, η κάθε μια από τις οποίες θα βαθμολογηθεί με 0.6 μονάδες του τελικού βαθμού του μαθήματος.
2. Οι απαντήσεις να όσο το δυνατόν συντομότερες.

1η άσκηση:
Να παράγετε $N = 500$ δισδιάστατα σημεία από μία κανονική κατανομή $\mathcal{N}(m, S)$, με μέση τιμή $m = [0,\ 0]^T$ και μήτρα συνδιασποράς $S = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$, για τις παρακάτω περιπτώσεις
(α) $\sigma_1^2 = \sigma_2^2 = 1,\ \sigma_{12} = 0$
(β) $\sigma_1^2 = \sigma_2^2 = 0.2,\ \sigma_{12} = 0$
(γ) $\sigma_1^2 = \sigma_2^2 = 2,\ \sigma_{12} = 0$
(δ) $\sigma_1^2 = 0.2,\ \sigma_2^2 = 2,\ \sigma_{12} = 0$
(ε) $\sigma_1^2 = 2,\ \sigma_2^2 = 0.2,\ \sigma_{12} = 0$
(στ) $\sigma_1^2 = \sigma_2^2 = 1,\ \sigma_{12} = 0.5$
(ζ) $\sigma_1^2 = 0.3,\ \sigma_2^2 = 2,\ \sigma_{12} = 0.5$
(η) $\sigma_1^2 = 0.3,\ \sigma_2^2 = 2,\ \sigma_{12} = -0.5$
Απεικονίστε γραφικά το παραπάνω σύνολο και σχολιάστε τα σχήματα των ομάδων που προκύπτουν.

***Υπόδειξη*:** Για την παραγωγή του πρώτου συνόλου δεδομένων χρησιμοποιούμε τις παρακάτω εντολές:

```
randn('seed',0)  %Initialization of the randn function
m=[0 0]';
S=[1 0; 0 1];
N=500;
X=mvnrnd(m,S,N)';
```

όπου X είναι το μητρώο που περιέχει τα διανύσματα στις στήλες του.

Για τη γραφική απεικόνιση των δεδομένων του πρώτου συνόλου χρησιμοποιούμε τις παρακάτω εντολές:

```
figure(1), plot(X(1,:), X(2,:),'.');
figure(1), axis equal
figure(1), axis([-7 7 -7 7])
```

Για τα υπόλοιπα σύνολα εργαζόμαστε με παρόμοιο τρόπο.

2η άσκηση:

Θεωρείστε ένα πρόβλημα δύο ισοπίθανων κλάσεων στο μονοδιάστατο χώρο, οι οποίες προκύπτουν από τις κανονικές κατανομές $\mathcal{N}(m1, s)$[1] και $\mathcal{N}(m2, s)$, αντίστοιχα, όπου $m1 = 1$, $m2 = 7$ και $s = 2$.

(α) Ταξινομήστε το σημείο $x = 3.5$ σε μία από τις δύο κλάσεις, με βάση τον κανόνα του Bayes (κάνετε τις πράξεις στο χαρτί).

(β) Να παράγετε ένα σύνολο Y με $n = 200$ σημεία, από τα οποία τα μισά θα προέρχονται από την πρώτη κλάση και τα υπόλοιπα από τη δεύτερη. Στη συνέχεια ταξινομήστε καθένα από τα στοιχεία αυτά στις δύο κλάσεις χρησιμοποιώντας τον ταξινομητή Bayes. Προτείνετε έναν απλό τρόπο εκτίμησης της πιθανότητας λάθους, P_e, με βάση τα αποτελέσματα της ταξινόμησης των σημείων του Y, και εφαρμόστε τον προκειμένου να υπολογίσετε την εκτίμηση της P_e (χρήση MATLAB).

(γ) Επαναλάβετε το (β) για $n = 2000$ και για $n = 20000$.

(δ) Προσδιορίστε θεωρητικά την πιθανότητα λάθους του ταξινομητή Bayes για τις δύο κλάσεις.

(ε) Συγκρίνετε τα αποτελέσματα που προέκυψαν από τα (β), (γ), (δ) και εξάγετε συμπεράσματα.

Υπόδειξη: Χρησιμοποιήστε τον παρακάτω κώδικα MATLAB για το βήμα (β) (αλλάζοντας την τιμή του N, ο κώδικας μπορεί να χρησιμοποιηθεί και για τα (γ), (δ)).

```
randn('seed',0) %Initialization of the random number generator for
normal distr.

P1=0.5;  %a priori class probabilities
P2=0.5;

m1=1;    % mean of normal distributions
m2=7;
s=2;     % variance of the normal distributions

N=200;   %Total number of points (Give only even numbers)

% The vector containing the set of points (1st half from 1st class)
Y=[randn(1,N/2)+m1 randn(1,N/2)+m2];
% The vector containing the true class labels for the points (if
t(i)=1(2), the
% i-th point comes from class 1(2))
t=[ones(1, N/2) 2*ones(1, N/2)];

output=[]; % Vector containing the class labels of
for i=1:N
    %Computation of the pdfs for both classes on the specific data
points
    p1=(1/(sqrt(2*pi)*s))*exp(-(Y(i)-m1)^2/(2*s));
    p2=(1/(sqrt(2*pi)*s))*exp(-(Y(i)-m2)^2/(2*s));
```

[1] Προσοχή: Το s εδώ συμβολίζει τη διασπορά

```
    % Application of the Bayes rule
    if(P1*p1>P2*p2)
        output=[output 1];
    else
        output=[output 2];
    end
end

% Check for the points that are not classified correctly.
bayes_res=(t~=output)  %if  bayes_res(i)=1  then  the  i-th  point  is
correctly classified
```

3η άσκηση:

Θεωρείστε ένα πρόβλημα δύο ισοπίθανων κλάσεων στον τρισδιάστατο χώρο, όπου οι δύο κλάσεις μοντελοποιούνται από κανονικές κατανομές με μέσες τιμές $m_1 = [0.2,\ 0.2,\ 0.2]^T$ και $m_2 = [0.7,\ 0.7,\ 0.7]^T$, αντίστοιχα. Η μήτρα συνδιασποράς για τις δύο κατανομές είναι

$$S = \begin{bmatrix} 0.8 & 0.01 & 0.01 \\ 0.01 & 0.2 & 0.01 \\ 0.01 & 0.01 & 0.2 \end{bmatrix}$$

(α) Παράγετε $N = 200$ διανύσματα (τα μισά από την πρώτη και τα υπόλοιπα από τη δεύτερη κατανομή) και ταξινομήστε τα χρησιμοποιώντας (i) τον ταξινομητή ελάχιστης Ευκλείδειας απόστασης και (ii) τον ταξινομητή ελάχιστης Mahalanobis απόστασης. Εκτιμήστε την πιθανότητα λάθους σε κάθε περίπτωση.

(β) Επαναλάβετε το (α) για $N = 2000$ και $N = 20000$.

(γ) Σχολιάστε τα αποτελέσματα που προέκυψαν από τα (α) και (β).

(δ) Ποια εκτιμάτε ότι θα ήταν η απόδοση του ταξινομητή Bayes; (απαντήστε χωρίς να εφαρμόσετε τον κανόνα του Bayes)

***Υπόδειξη**:* Χρησιμοποιήστε για το (α) τον ακόλουθο κώδικα

```
randn('seed',0) %initialization of the random number generator

m1=[0.2 0.2 0.2]';  %mean vectors
m2=[0.7 0.7 0.7]';
m=[m1 m2];
S=[0.8 0.01 0.01; 0.01 0.2 0.01; 0.01 0.01 0.2]; %Covariance matrix

N=200; %Number of data vectors (only even numbers)

%Generation of the data set
Y=[mvnrnd(m1',S,N/2); mvnrnd(m2',S,N/2)]'; %Y is an 3xN matrix

%Class labels (if (t(i)=1(2), the i-th vector is from class 1(2))
t=[ones(1,N/2) 2*ones(1,N/2)];

%Application of the Euclidean classifier
out_eucl=euclidean_classifier(m,Y);
```

```
%Application of the Mahalanobis classifier
out_maha=mahalanobis_classifier(m,S,Y);
```

```
function [z]=euclidean_classifier(m,X)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% FUNCTION
%   [z]=euclidean_classifier(m,X)
% This function classifies a set of data vectors in one out of c
possible
% classes, according to the Euclidean classifier.
%
% INPUT ARGUMENTS:
%   m:       an lxc dimensional matrix, whose i-th column corresponds
to the
%           mean of the i-th class.
%    X:        an lxN dimensional matrix whose columns are the data
vectors to
%           be classified.
%
% OUTPUT ARGUMENTS
%   z:      an N-dimensional vector whose i-th component contains the
label
%                of the class where the i-th data vector has been
assigned.
%
% (c) 2008 A. Pikrakis, S. Theodoridis, K. Koutroumbas, D. Cavouras
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

    [l,c]=size(m);
    [l,N]=size(X);

    for i=1:N
       for j=1:c
          de(j)=sqrt((X(:,i)-m(:,j))'*(X(:,i)-m(:,j)));
       end
       [num,z(i)]=min(de);
    end
```

```
function z=mahalanobis_classifier(m,S,X)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% FUNCTION
%   [z]=mahalanobis_classifier(m,S,X)
% This function classifies a set of data vectors in one out of c
possible
% classes, according to the Mahalanobis classifier.
%
% INPUT ARGUMENTS:
%   m:       an lxc dimensional matrix, whose i-th column corresponds
to the
%           mean of the i-th class
%   S:      an lxl dimensional matrix which corresponds to the matrix
```

```
%             involved in the Mahalanobis distance (when the classes
have
%              the same covariance matrix, S equals to this common
covariance
%            matrix).
%    X:        an lxN dimensional matrix, whose columns are the data
vectors
%            to be classified.
%
% OUTPUT ARGUMENTS
%    z:      an N-dimensional vector whose i-th component contains the
label
%                 of the class where the i-th data vector has been
assigned.
%
% (c) 2008 A. Pikrakis, S. Theodoridis, K. Koutroumbas, D. Cavouras
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

    [l,c]=size(m);
    [l,N]=size(X);

    for i=1:N
       for j=1:c
          dm(j)=sqrt((X(:,i)-m(:,j))'*inv(S)*(X(:,i)-m(:,j)));
       end
       [num,z(i)]=min(dm);
    end
```

4η άσκηση:

Θεωρείστε ένα πρόβλημα δύο κλάσεων στο μονοδιάστατο χώρο, όπου οι δύο κλάσεις μοντελοποιούνται από κανονικές κατανομές $\mathcal{N}(m1,s)$ και $\mathcal{N}(m2,s)$, αντίστοιχα, όπου $m1=1, m2=6$ και $s=2$.

(α) Να παράγετε $N1=15$ σημεία από την πρώτη κατανομή και $N2=15$ σημεία από τη δεύτερη κατανομή και έστω X_1 και X_2 τα αντίστοιχα σύνολα δεδομένων για κάθε κλάση.

(β) Να παράγετε ένα σύνολο Y με $N=1000$ σημεία, από τα οποία τα πρώτα 500 θα προέρχονται από την πρώτη κατανομή ενώ τα υπόλοιπα από τη δεύτερη.

(γ) Ας προσποιηθούμε τώρα ότι για το πρόβλημά μας γνωρίζουμε τα ακόλουθα: (i) τα σύνολα δεδομένων X_1 και X_2 για τις δύο κλάσεις και (ii) το γεγονός ότι οι δύο κλάσεις μοντελοποιούνται από κανονικές κατανομές γνωστής (κοινής) διασποράς $s=2$ και άγνωστων μέσων τιμών. Εκτιμήστε τις μέσες τιμές των κατανομών των δύο κλάσεων με τη μέθοδο της μέγιστης πιθανοφάνειας (ML).

(δ) Αφού εκτιμήσετε τις a priori πιθανότητες των δύο κλάσεων εφαρμόστε τον ταξινομητή Bayes προκειμένου να ταξινομήσετε τα στοιχεία του συνόλου Y χρησιμοποιώντας (i) τις πραγματικές τιμές των μέσων τιμών των κατανομών και (ii) τις εκτιμηθείσες τιμές αυτών. Εκτιμήστε την πιθανότητα λάθους στις δύο περιπτώσεις.

(ε) Επαναλάβετε το (α) για $N1 = 1000$ και $N2 = 1000$ σημεία και στη συνέχεια επαναλάβετε τα (γ) και (δ).

(στ) Σχολιάστε τα αποτελέσματα.

(ζ) Με βάση τα δεδομένα που υποτίθεται ότι είναι γνωστά στο (γ) θα είχε νόημα η χρήση της μεθόδου της μεγίστης εκ των υστέρων πιθανότητας (MAP) για την εκτίμηση των μέσων τιμών των κατανομών;

***Υπόδειξη**:* Χρησιμοποιήστε τον παρακάτω κώδικα για τη διεξαγωγή των πειραμάτων

```
randn('seed',0)  %Initialization of the random number generator for
normal

m1=1; %The parameters of the distributions
m2=6;
s=2;

% Genereation of X1
N1=15;
X1=randn(1,N1)+m1;

%Generation of X2
N2=15;
X2=randn(1,N2)+m2;

%Generation of Y
randn('seed',100) %Initialization that guarantees that the same set Y
will be produced.
N=1000;
Y=[randn(1,N/2)+m1 randn(1,N/2)+m2];
t=[ones(1,N/2) 2*ones(1,N/2)];

%Maximum likelihood estimates of the means
m1_ML=sum(X1)/N1;
m2_ML=sum(X2)/N2;

% Estimation of the a priori probabilities
P1=N1/(N1+N2);
P2=N2/(N1+N2);

%Bayes rule for the case where the true means are used
output=[];
for i=1:N
    p1=(1/(sqrt(2*pi)*s))*exp(-(Y(i)-m1)^2/(2*s));
    p2=(1/(sqrt(2*pi)*s))*exp(-(Y(i)-m2)^2/(2*s));
    % Application of the Bayes rule
    if(P1*p1>P2*p2)
        output=[output 1];
    else
        output=[output 2];
    end
end

bayes_res=(t~=output);  %if bayes_res(i)=1 then the i-th point is
correctly classified
```

```
%Bayes rule for the case where the estimated means are used
output_ML=[];
for i=1:N
    p1=(1/(sqrt(2*pi)*s))*exp(-(Y(i)-m1_ML)^2/(2*s));
    p2=(1/(sqrt(2*pi)*s))*exp(-(Y(i)-m2_ML)^2/(2*s));
    % Application of the Bayes rule
    if(P1*p1>P2*p2)
        output_ML=[output_ML 1];
    else
        output_ML=[output_ML 2];
    end
end

bayes_res_ML=(t~=output_ML);  %if  bayes_res_ML(i)=1  then  the  i-th
point is correctly classified
```

5η άσκηση:

Θεωρείστε ένα πρόβλημα δύο κλάσεων στο μονοδιάστατο χώρο, όπου οι δύο κλάσεις μοντελοποιούνται από κανονικές κατανομές $\mathcal{N}(m1, s)$ και $\mathcal{N}(m2, s)$, αντίστοιχα, όπου $m1 = 1, m2 = 5$ και $s = 1.75$.

(α) Να παράγετε $N1 = 50$ σημεία από την πρώτη κατανομή και $N2 = 50$ σημεία από τη δεύτερη κατανομή και έστω X_1 και X_2 τα αντίστοιχα σύνολα δεδομένων για κάθε κλάση.

(β) Να παράγετε ένα σύνολο Y με $N = 2000$ σημεία, από τα οποία τα πρώτα 1000 θα προέρχονται από την πρώτη κατανομή ενώ τα υπόλοιπα από τη δεύτερη.

(γ) Ας προσποιηθούμε τώρα ότι για το πρόβλημά μας γνωρίζουμε **μόνο** τα σύνολα δεδομένων X_1 και X_2 για τις δύο κλάσεις. Εκτιμήστε τις a priori πυκνότητες πιθανότητας των δύο κλάσεων και, στη συνέχεια, εφαρμόστε τον ταξινομητή Bayes προκειμένου να ταξινομήσετε τα στοιχεία του συνόλου Y. Για την εκτίμηση της πυκνότητας πιθανότητας της κλάσης 1 σε ένα σημείο του Y χρησιμοποιήστε παράθυρα Parzen (με $h = 0.2$ και συνάρτηση βάσης (ϕ) την $\phi(x_i) = \begin{cases} 1 & |x_i| \leq h/2 \\ 0 & διαφορετικά \end{cases}$, η οποία ισούται με 1 για όλα τα σημεία που βρίσκονται στο διάστημα μεγέθους h που είναι κεντραρισμένο στο 0) και το σύνολο $X1$. Ομοίως και για την κλάση $X2$.

(δ) Χρησιμοποιήστε τον ταξινομητή Bayes για την ταξινόμηση των στοιχείων του Y, χρησιμοποιώντας τις πραγματικές τιμές των κατανομών και συγκρίνετε τα αποτελέσματα.

(ε) Επαναλάβετε για $N1 = 300$ και $N2 = 300$. Εξάγετε τα συμπεράσματά σας.

Υπόδειξη: Χρησιμοποιήστε τον παρακάτω κώδικα

```
randn('seed',0) %Initialization of the random number generator for
normal

m1=1; %The parameters of the distributions
m2=5;
s=1.75;

% Genereation of X1
N1=500;
X1=randn(1,N1)+m1;

%Generation of X2
N2=500;
X2=randn(1,N2)+m2;

%Generation of Y
randn('seed',100) %Initialization that guarantees that the same set Y
will be produced.
N=1000;
Y=[randn(1,N/2)+m1 randn(1,N/2)+m2];
t=[ones(1,N/2) 2*ones(1,N/2)];

% Estimation of the a priori probabilities
P1=N1/(N1+N2);
P2=N2/(N1+N2);

%Bayes rule for the case where the true means are used
output=[];
for i=1:N
    p1=(1/(sqrt(2*pi)*s))*exp(-(Y(i)-m1)^2/(2*s));
    p2=(1/(sqrt(2*pi)*s))*exp(-(Y(i)-m2)^2/(2*s));
    % Application of the Bayes rule
    if(P1*p1>P2*p2)
        output=[output 1];
    else
        output=[output 2];
    end
end

bayes_res=(t~=output); %if bayes_res(i)=1 then the i-th point is
correctly classified

%Bayes rule for the case where Parzen estimations are used
h=0.2;
output_Parzen=[];
for i=1:N
    p1=sum(abs((X1-Y(i))/h)<=1/2)/(N*h);   %   This   is   the   1-D
implementation of Parzen
    p2=sum(abs((X2-Y(i))/h)<=1/2)/(N*h);
    % Application of the Bayes rule
    if(P1*p1>P2*p2)
        output_Parzen=[output_Parzen 1];
    else
        output_Parzen=[output_Parzen 2];
    end
end

bayes_res_Parzen=(t~=output_Parzen); %if bayes_res_ML(i)=1 then the
i-th point is correctly classified
```