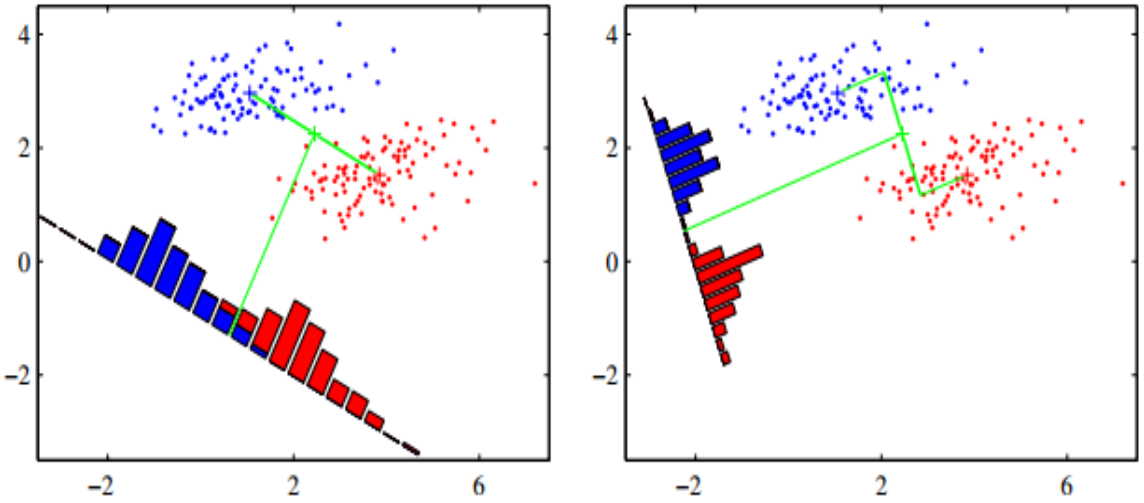
ΕΡΓΑΣΙΑ ΣΤΟ ΜΑΘΗΜΑ: ΕΥΦΥΗΣ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ – ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ

ΑΝΤΙΚΕΙΜΕΝΟ ΕΡΓΑΣΙΑΣ: ΠΑΡΟΥΣΙΑΣΗ ΠΛΗΘΥΣΜΩΝ ΣΕ 2 ΔΙΑΣΤΑΣΕΙΣ ΜΕ ΤΗΝ ΚΑΛΥΤΕΡΗ ΔΙΑΧΩΡΙΣΙΜΟΤΗΤΑ ΜΕ FISHER LINEAR DISCRIMINANT ANALYSIS (FLD)

ΦΟΙΤΗΤΕΣ : ΚΑΚΑΒΑΣ ΝΙΚΟΛΑΟΣ – ΚΑΤΣΙΜΑΡΔΟΣ ΔΗΜΟΣ

ΠΜΣ Η/Ε - ΔΠΜΣ ΗΕΠ



Πάτρα 2016, Μάιος

Εισαγωγή

Ενώ η PCA (Principal Components Analysis) αναζητεί διευθύνσεις που είναι αποδοτικές για αναπαράσταση των δεδομένων μας, η ανάλυση διαχωρισμού αναζητά διευθύνσεις οι οποίες είναι αποδοτικές για διαχωρισιμότητα – διάκριση των πληθυσμών μας.

Ο στόχος της ανάλυσης διαχωρισμού είναι η περιστροφή της ευθείας σε τέτοιο προσανατολισμό, ώστε τα προβαλλόμενα δεδομένα μας πάνω σε αυτή την ευθεία να παρουσιάζουν όσο το δυνατόν μικρότερη επικάλυψη, και κατά συνέπεια μεγάλη διαχωρισιμότητα.

Το Πρόβλημα

Δίνονται διανύσματα για 5 πρόσωπα, 10 διανύσματα για κάθε πρόσωπο.

Ο αρχικός χώρος είναι 5 – διαστάσεων.

Ζητείται η παρουσίαση των πληθυσμών σε 2 – διαστάσεις με την καλύτερη διαχωρισιμότητα, μέσω της μεθόδου γραμμικού διαχωρισμού του Fisher.

Θεωρία Ανάλυσης Διαχωρισμού του Fisher

Ένας τρόπος για να δούμε ένα μοντέλο ταξινόμησης είναι μέσω της μείωσης των διαστάσεων. Εάν υποθέσουμε πως έχουμε δύο μόνο κλάσεις, και παίρνουμε το διάνυσμα εισόδου D – διαστάσεων, x και το προβάλλουμε σε 1 διάσταση μέσω του τύπου :

$$y = \mathbf{w}^{\mathrm{T}}\mathbf{x}. \qquad (1)$$

Εάν βάλουμε ένα όριο απόφασης y >= -w0 σαν κλάση C_1, αλλιώς κλάση C_2, τότε έχουμε έναν απλό γραμμικό ταξινομητή. Γενικά, η προβολή σε μία διάσταση οδηγεί σε σημαντική απώλεια πληροφορίας, και οι κλάσεις που είναι καλά διαχωρισμένες στον αρχικό χώρο των D – διαστάσεων μπορεί να εμφανίσουν μεγάλη επικάλυψη στην μία διάσταση. Ωστόσο, ρυθμίζοντας τα συστατικά του διανύσματος βάρους w, μπορούμε να επιλέξουμε μία προβολή που μεγιστοποιεί τον διαχωρισμό των κλάσεων.

Ας θεωρήσουμε ένα πρόβλημα 2-κλάσεων, όπου έχουμε N_1 σημεία στην 1η κλάση και N_2 σημεία στην κλάση C_2. Τα διανύσματα των μέσων τιμών δίνονται από :

$$\mathbf{m}_1 = \frac{1}{N_1}\sum_{n \in \mathcal{C}_1} \mathbf{x}_n, \qquad \mathbf{m}_2 = \frac{1}{N_2}\sum_{n \in \mathcal{C}_2} \mathbf{x}_n. \qquad (2)$$

Η πιο απλή μέτρηση του διαχωρισμού των κλάσεων όταν προβληθούν στην ευθεία w, είναι ο διαχωρισμός των προβαλλόμενων μέσων τιμών των κλάσεων. Αυτό

υποδεικνύει πως ίσως επιλέξουμε ένα w, τέτοιο ώστε να μεγιστοποιήσουμε την ποσότητα :

$$m_2 - m_1 = \mathbf{w}^{\mathrm{T}}(\mathbf{m}_2 - \mathbf{m}_1) \tag{3}$$

Όπου

$$m_k = \mathbf{w}^{\mathrm{T}}\mathbf{m}_k \tag{4}$$

Είναι η μέση τιμή των προβαλλόμενων δεδομένων από την κλάση C_k. Ωστόσο, αυτή η έκφραση μπορεί να γίνει μεγάλη, απλά αυξάνοντας το πλάτος του w. Για να λύσουμε αυτό το πρόβλημα, θα μπορούσαμε να κάνουμε το w μοναδιαίο, ώστε :

$$\sum_i w_i^2 = 1. \tag{5}$$

Χρησιμοποιώντας έναν πολλαπλασιαστή Lagrange για την μεγιστοποίηση του προβλήματος, βρίσκουμε πως το $w \sim (m_2 - m_1)$. Ωστόσο, υπάρχει ένα ακόμη πρόβλημα με αυτή την προσέγγιση. Έχουμε μεγάλη επικάλυψη των προβαλλόμενων δεδομένων. Αυτό το πρόβλημα προέρχεται από τις μη διαγώνιες συνδιακυμάνσεις των κατανομών των κλάσεων. Η ιδέα που προτάθηκε από τον Fisher ήταν η μεγιστοποίηση μιας συνάρτησης η οποία θα δώσει μεγάλη διαχωρισιμότητα μεταξύ των προβαλλόμενων μέσων τιμών των κλάσεων ενώ ταυτόχρονα θα δίνει μικρή διακύμανση μέσα σε κάθε κλάση, ελαχιστοποιώντας έτσι την επικάλυψη των κλάσεων.

Η φόρμουλα προβολής μετασχηματίζει το σετ των δεδομένων μας στο x σε ένα άλλο σετ δεδομένων μίας διάστασης με ετικέτα y. Η within-class variance των μετασχηματισμένων δεδομένων από την κλάση C_k δίνεται από τη σχέση :

$$s_k^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2 \tag{6}$$

Όπου

$$y_n = \mathbf{w}^{\mathrm{T}}\mathbf{x}_n. \tag{7}$$

Μπορούμε να ορίσουμε την ολική within – class variance για όλο σύνολο των δεδομένων να είναι απλά S1^2 + S2^2. Το κριτήριο του Fisher ορίζεται ως η αναλογία της between-class variance προς την within-class variance :

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}. \tag{8}$$

Μπορούμε να κάνουμε την εξάρτηση από το w πιο εμφανή, εάν εφαρμόσουμε τις προηγούμενες σχέσεις, οπότε και καταλήγουμε στην επόμενη σχέση :

$$J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}} \tag{9}$$

Όπου η S_B είναι ο between-class covariance πίνακας και δίνεται από :

$$\mathbf{S}_{\mathrm{B}} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^{\mathrm{T}} \tag{10}$$

Και S_W είναι ο total within-class covariance πίνακας και δίνεται από :

$$\mathbf{S}_{\mathrm{W}} = \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^{\mathrm{T}} + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^{\mathrm{T}}. \tag{11}$$

Η σχέση (9) είναι γνωστή στην Μαθηματική Φυσική σαν γενικευμένη εξίσωση του Rayleigh. Είναι αρκετά εύκολο να δειχθεί πως ένα διάνυσμα w που μεγιστοποιεί την J(.) , πρέπει να ικανοποιεί τη σχέση :

$$S_B\, w = \lambda\, S_W\, w\,, \tag{12}$$

Για κάποια σταθερά λ, το οποίο είναι ένα γενικευμένο πρόβλημα ιδιοτιμών. Εάν το S_W είναι μη μοναδιαίο, έχουμε ένα απλό πρόβλημα ιδιοτιμών απλά γράφοντας :

$$\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{w} = \lambda\mathbf{w}. \tag{13}$$

Στην περίπτωση των δυο κλάσεων δεν είναι απαραίτητη η επίλυση του γενικευμένου προβλήματος ιδιοτιμών καθώς το S_B w είναι πάντα στη διεύθυνση του m_2-m_1. Μιας και ο παράγοντας w είναι ασήμαντος, μπορούμε να γράψουμε την λύση για το w που βελτιστοποιεί τη συνάρτηση J(w) κατευθείαν ως :

$$\boxed{\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2).} \tag{14}$$

Έτσι, έχουμε εκτιμήσει το w για την Fisher's linear discriminant – την γραμμική συνάρτηση που επιτυγχάνει το μέγιστο λόγο του between-class scatter προς το within-class scatter. Δηλαδή, η ταξινόμηση έχει μετατραπεί από D – διαστάσεων πρόβλημα σε ένα πρόβλημα μίας διάστασης, εμφανώς πιο διαχειρίσιμο.

Εφαρμογή στο Matlab

```
%% Fisher Linear Discriminant (FLD)
% 5 Faces - 10 vectors for each face - 5D space
clear ; close all ; clc;

load vvoice.mat;

%mean values , mass centers of 5-D space
m1=mean(v1');
m2=mean(v2');
m3=mean(v3');
m4=mean(v4');
m5=mean(v5');


%within scatter matrix

cov1=cov(v1');
cov2=cov(v2');
cov3=cov(v3');
cov4=cov(v4');
cov5=cov(v5');

Sw = cov1 + cov2 + cov3 + cov4 + cov5;


%between scatter matrix
a=zeros(5,5);

a(1,:)=m1;
a(2,:)=m2;
a(3,:)=m3;
a(4,:)=m4;
a(5,:)=m5;

Sb = cov(a);


%Generalized eigenvalues equation
[v, d] = eig(Sw,Sb);


% v:eigenvectors/ w:transformation matrix
w = abs(v(:,1:2));


% transformation from 5_D space to  2-D space
yv1=w'*v1;
yv2=w'*v2;
yv3=w'*v3;
yv4=w'*v4;
yv5=w'*v5;
```

Αρχικά, φορτώνουμε τα 5 πρόσωπα και υπολογίζουμε τις μέσες τιμές αυτών όπως και τους πίνακες συνδιακύμανσης για κάθε πρόσωπο. Ο within-scatter matrix (wsm) είναι το άθροισμα όλων των πινάκων συνδιακύμανσης για τα 5 πρόσωπα.

Στη συνέχεια υπολογίζουμε τον between scatter matrix που χρειάζεται για την επίλυση του γενικευμένου προβλήματος ιδιοτιμών → [v,d] = eig(Sw,Sb). Η τελευταία εντολή επιστρέφει έναν διαγώνιο πίνακα d των ιδιοτιμών, και έναν πίνακα v, του οποίου οι στήλες είναι τα αντίστοιχα ιδιοδιανύσματα στις ιδιοτιμές. Τελικά, κρατάμε τις δύο πρώτες στήλες οι οποίες αντιστοιχούν στην μεγαλύτερη διασπορά (δηλαδή οι δύο μεγαλύτερες ιδιοτιμές), και τις αντιστοιχούμε στον πίνακα μετασχηματισμού w.

Για την μετάβαση στον νέο 2-διαστάσεων χώρο χρησιμοποιούμε τον πίνακα w , όπου size(w) = 5x2, και πολλαπλασιάζοντας τα αρχικά πρόσωπα με αυτό τον πίνακα μεταβαίνουμε στο νέο χώρο.

Με το παρακάτω απόσπασμα κώδικα γίνεται η απεικόνιση των προσώπων στον νέο χώρο :

```
%% REPRESENTATION
%mass centers of the 2-D space
m2_1=mean(yv1');
m2_2=mean(yv2');
m2_3=mean(yv3');
m2_4=mean(yv4');
m2_5=mean(yv5');

%coordinates of the 2_D space mass centers
x1=(m2_1(1,1));y1=(m2_1(1,2));
x2=(m2_2(1,1));y2=(m2_2(1,2));
x3=(m2_3(1,1));y3=(m2_3(1,2));
x4=(m2_4(1,1));y4=(m2_4(1,2));
x5=(m2_5(1,1));y5=(m2_5(1,2));

% Represantation of the vectors in the 2-D space

figure(1),plot(yv1(1,:),yv1(2,:),'r*') %voice 1
hold on
plot(yv2(1,:),yv2(2,:),'g*') %voice 2
hold on
plot(yv3(1,:),yv3(2,:),'b*') %voice 3
hold on
plot(yv4(1,:),yv4(2,:),'m*') %voice 4
hold on
plot(yv5(1,:),yv5(2,:),'c*') %voice 5
title('Representation of the vectors in the 2-D space');
```

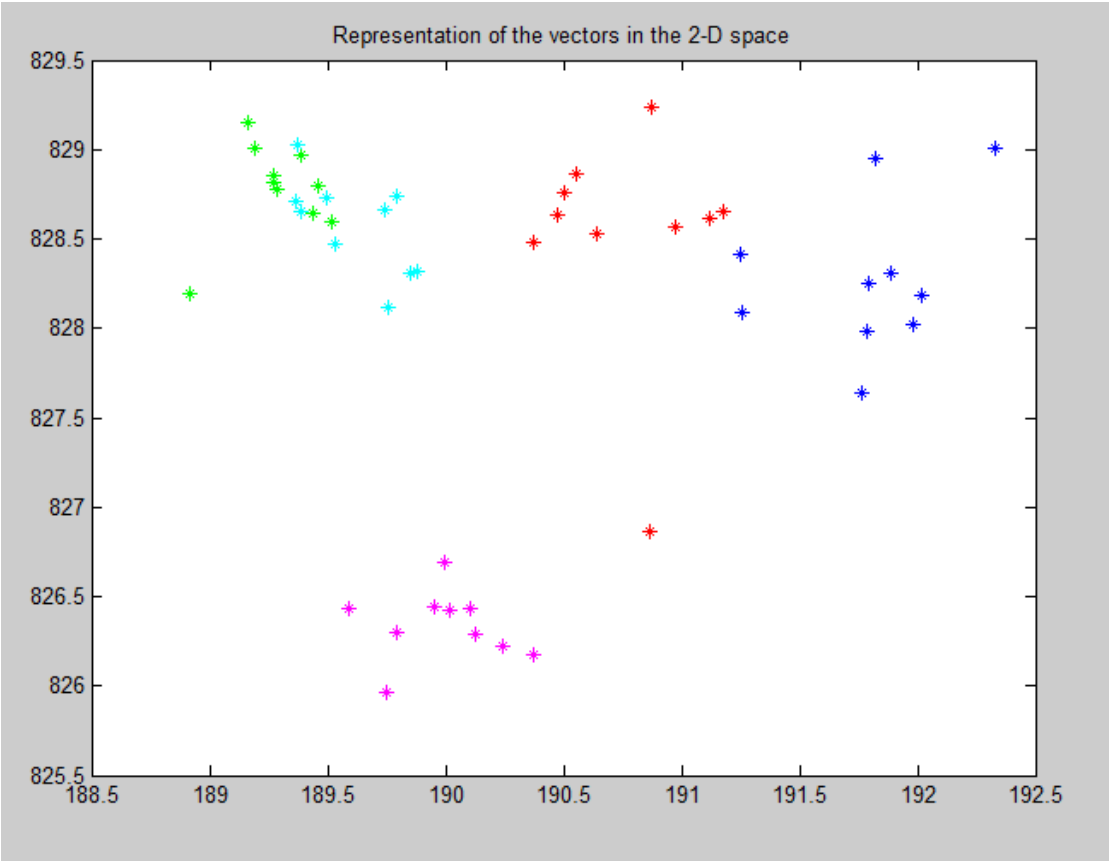


Figure 1: Vectors Representation in the 2-D space

Στη συνέχεια πραγματοποιείται τμηματοποίηση του χώρου με βάση την Ευκλείδια απόσταση και την απόσταση Mahalanobis.

```
%% Space Segmentation using Euclidian distance

% step 4
% euclidian distance
figure(2)
title('Space Segmentation using Euclidian distance')
hold on

for i=188.5:.05:192.5
    for j=825.5:.05:829.5
        d1=dist(m2_1,[i;j]);
        d2=dist(m2_2,[i;j]);
        d3=dist(m2_3,[i;j]);
        d4=dist(m2_4,[i;j]);
        d5=dist(m2_5,[i;j]);

        dd(1,1)=d1;
        dd(2,1)=d2;
        dd(3,1)=d3;
        dd(4,1)=d4;
        dd(5,1)=d5;

        [y,s]=min(dd);

        if s==1
            plot(i,j,'r+');
```

```
        elseif s==2
            plot(i,j,'g+');
        elseif s==3
            plot(i,j,'b+');
        elseif s==4
            plot(i,j,'m+');
        else
            plot(i,j,'c+');
        end
    end
end



hold on

plot(x1,y1,'black*');
hold on
plot(x2,y2,'black*');
hold on
plot(x3,y3,'black*');
hold on
plot(x4,y4,'black*');
hold on
plot(x5,y5,'black*');

axis off
```

```
%% Space Segmantation using Mahalanobis Distance



%using Mahalanobis distance

s1=cov(yv1');
s2=cov(yv2');
s3=cov(yv3');
s4=cov(yv4');
s5=cov(yv5');

s1_2=inv(s1);
s2_2=inv(s2);
s3_2=inv(s3);
s4_2=inv(s4);
s5_2=inv(s5);

dets1=det(s1);
dets2=det(s2);
dets3=det(s3);
dets4=det(s4);
dets5=det(s5);

figure(3)
title('Space segmantation using Mahalanobis distance')
hold on

for i=188.5:.05:192.5
    for j=825.5:.05:829.5
        x=[i;j];
```

```
        t1=(x-m2_1');
        t2=(x-m2_2');
        t3=(x-m2_3');
        t4=(x-m2_4');
        t5=(x-m2_5');

        dm1=-(t1'*s1_2*t1)-log10(dets1);
        dm2=-(t2'*s2_2*t2)-log10(dets2);
        dm3=-(t3'*s3_2*t3)-log10(dets3);
        dm4=-(t4'*s4_2*t4)-log10(dets4);
        dm5=-(t5'*s5_2*t5)-log10(dets5);

        dmah(1,1)=dm1;
        dmah(2,1)=dm2;
        dmah(3,1)=dm3;
        dmah(4,1)=dm4;
        dmah(5,1)=dm5;

        [k,l]=max(dmah);

        if l==1
            plot(i,j,'r+');
        elseif l==2
            plot(i,j,'g+');
        elseif l==3
            plot(i,j,'b+');
        elseif l==4
            plot(i,j,'m+');
        else
            plot(i,j,'c+');
        end

    end
end

hold on

plot(x1,y1,'black*');
hold on
plot(x2,y2,'black*');
hold on
plot(x3,y3,'black*');
hold on
plot(x4,y4,'black*');
hold on
plot(x5,y5,'black*');

axis off
```

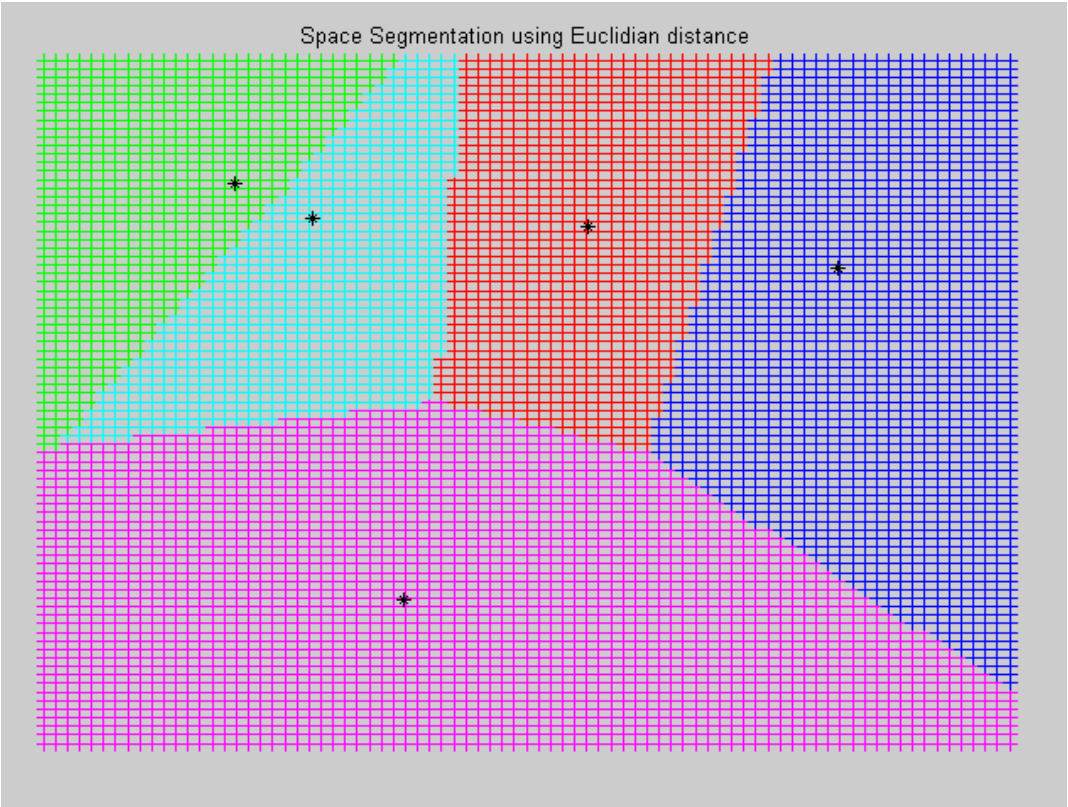


Figure 2: Space segmentation using Euclidean distance

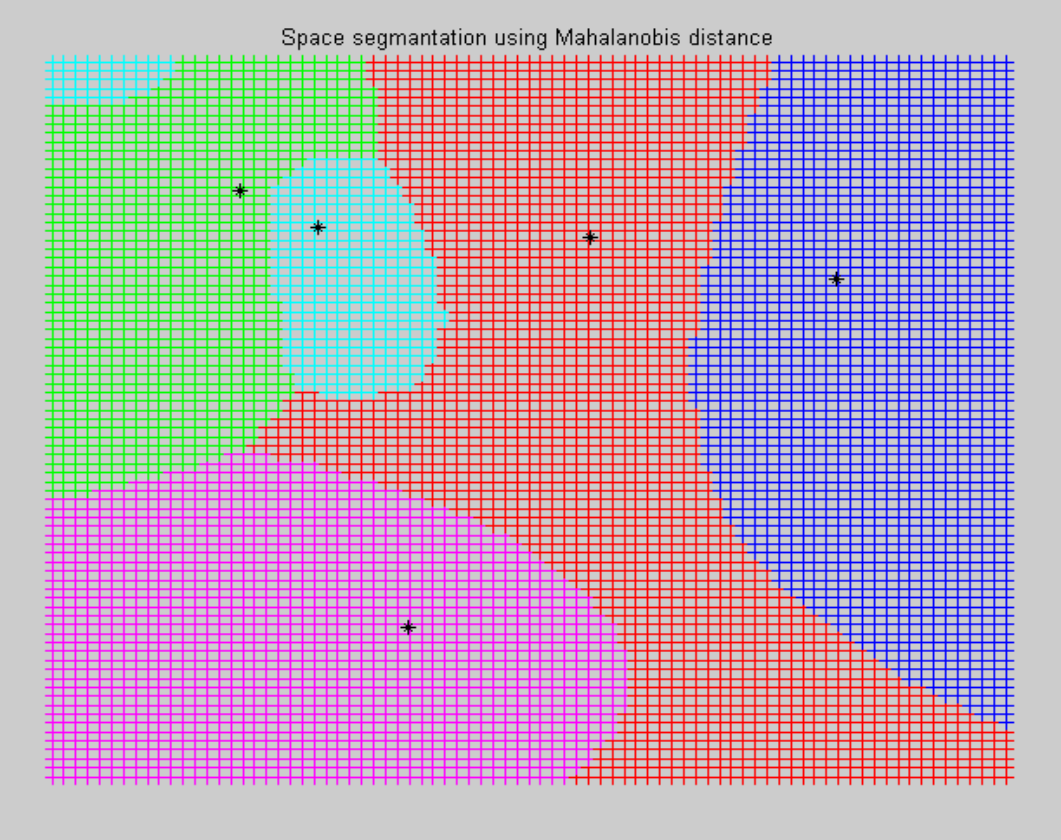


Figure 3: Space Segmentation using Mahalanobis distance.