MACHINE LEARNING

EXERCISES

Elements of parametric techniques

All the course material is available on the web site

Course web site: http://pralab.diee.unica.it/MachineLearning

PART 4: Elements of parametric techniques

Required knowledge to solve these exercises

1. Discriminant functions g(x) for Gaussian classes
   a. General case (different priors, different covariance matrix per class)
   $$g(x;\ \mu, \Sigma) = -\frac{1}{2}x^T\Sigma^{-1}x + \ \mu^T\Sigma^{-1}x - \frac{1}{2}\mu^T\Sigma^{-1}\mu + \ln p(\omega) - \frac{1}{2}\ln|\Sigma|$$
   b. Degenerate cases
      i. Same priors, same isotropic covariance matrix $\Sigma = \sigma^2 \mathrm{I}$ for each class (Nearest Mean Centroid classifier)
      $g(x) = \mu^T x - \frac{1}{2}\mu^T\mu$ or $g(x) = -||x-\mu||^2$
      ii. Different priors, same isotropic covariance matrix $\Sigma = \sigma^2 \mathrm{I}$ for each class
      $g(x) = \frac{1}{\sigma^2}\ \mu^T x - \frac{1}{2\sigma^2}\mu^T\mu + \ln p(\omega)$ or $g(x) = -\frac{||x-\mu||^2}{2\sigma^2} + \ln p(\omega)$
      iii. Same priors, (arbitrary) covariance matrix equal for each class
      $g(x) = \ \mu^T\Sigma^{-1}x - \frac{1}{2}\mu^T\Sigma^{-1}\mu + \ln p(\omega)$

   Note that, if Gaussians are isotropic but different for each class, g(x) remains quadratic. The degenerate cases listed above hold only <u>when the covariance matrix is the same for each class</u>!

2. Classify a point using the MAP criterion:
   $\mathrm{argmax}_k\ g_k(x)$

3. Plot decision boundaries among classes in feature space
   a. Determine the values of x^* for which it holds that $g_1(x) = g_2(x)$
   b. Find the subset of points x^* for which g_1 and g_2 are the dominant classes, i.e., $g_1(x^*) > g_3(x^*)$, $g_1(x^*) > g_4(x^*)$, etc.
   This subset of points will identify the *active* boundary between class 1 and class 2
   c. Repeat for any other pair of classes

4. Estimate parameter values (mean and covariance matrices) from data (using MLE)
   a. Mean $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$
   b. Covariance $\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T$
   Note that the product inside the summation is an *outer* product between a column and a row vector of *d* elements, being *d* the number of features, and that the sum is computed over the training samples. Accordingly, the covariance matrix is a *d* x *d* matrix.

   Depending on the assumptions made on the covariance matrix (e.g., equal for all classes, isotropic, etc.), one should adjust the above estimate accordingly (see Exercise 3 for some examples).

Exercise 1

Let us consider a three-class problem, within a bi-dimensional feature space.

a. Classify the pattern $x_t = (1,1/3)^T$ under the following assumptions:

- the probability density function of each class is Gaussian, with the same covariance matrix $\mathbf{\Sigma}=\sigma^2\mathbf{I}$, and the parameter σ (variance) is unknown
- all the classes have the same prior probability
- the mean vectors of the three classes are: $\mu_1=(0,0)^T$; $\mu_2=(0,2)^T$; $\mu_3=(2,1)^T$.

b. Classify the same pattern $x_t =(1,1/3)^T$ under the different assumptions about priors:

$P(\omega_1) = P(\omega_2) = 1/4;\ P(\omega_3) = 1/2$

Solution

The general form of a multivariate normal pdf is

$$p(x|\omega_i) = \frac{1}{(2\,\pi)^{\frac{d}{2}}\,|\Sigma_i|^{\frac{1}{2}}} \exp[-\frac{1}{2}\,(x-\mu_i)^T\,\Sigma_i^{-1}\,(x-\mu_i)]$$

The discriminant function $g_i(x)$ is

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t\,\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_i\right| + \ln P(\omega_i)$$

Considering a Gaussian model with $\sum_i=\sigma^2\mathrm{I}$ we obtain

$$\left|\boldsymbol{\Sigma}_i\right| = \sigma^{2d};\ \boldsymbol{\Sigma}_i^{-1} = \left(1/\sigma^2\right)\mathbf{I}$$

and finally

$$g_i(\mathbf{x}) = -\frac{\left\|\mathbf{x}-\boldsymbol{\mu}_i\right\|^2}{2\sigma^2} + \ln\left(P(\omega_i)\right)$$

a) Equal priors

For this particular case, the discriminant function becomes:

$$g_i(\mathbf{x}) = -\left\|\mathbf{x}-\boldsymbol{\mu}_i\right\|^2$$

for the pattern $x_t =(1,\ 1/3)^T$

$$g_1(x)|_{x=x_t} = -\|x-\mu_1\|^2 = -10/9$$

$g_2(x)|_{x=x_t} = -\|x - \mu_2\|^2 = -34/9$

$g_3(x)|_{x=x_t} = -\|x - \mu_3\|^2 = -13/9$

Comparing the values of the above discriminant functions g_1, g_2 and g_3 for the pattern xt we can see that the value of the posterior probability is maximum for class 1. Therefore, the pattern $(1,1/3)^T$ is assigned to class 1, **regardless of the unknown value of σ.**

b) **Different priors**

$$P(\omega_1) = P(\omega_2) = 1/4;\ P(\omega_3) = 1/2$$

For this particular case, the linear discriminant function becomes

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x} - \boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln(P(\omega_i))$$

for the pattern $x_t = (1, 1/3)^T$

$$g_1(\mathbf{x}) = \frac{-\|\mathbf{x} - \boldsymbol{\mu}_1\|^2}{2\sigma^2} + \ln(1/4)$$

$$g_2(\mathbf{x}) = \frac{-\|\mathbf{x} - \boldsymbol{\mu}_2\|^2}{2\sigma^2} + \ln(1/4)$$

$$g_3(\mathbf{x}) = \frac{-\|\mathbf{x} - \boldsymbol{\mu}_3\|^2}{2\sigma^2} + \ln(1/2)$$

Note that the inequality $g_1(x)>g_2(x)$ does not depend on the value of $\ln(1/4)$, whereas the other two inequalities depend on the difference between ln(1/4) and ln(1/2).

Let's calculate the value of the discriminant functions **for the considered pattern $x=x_t$:**

$$g_1(\mathbf{x}) = \frac{-\|\mathbf{x} - \boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln(1/4) = -\frac{10/9}{2\sigma^2} + \ln(1/4)$$

$$g_2(\mathbf{x}) = \frac{-\|\mathbf{x} - \boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln(1/4) = -\frac{34/9}{2\sigma^2} + \ln(1/4)$$

$$g_3(\mathbf{x}) = \frac{-\|\mathbf{x} - \boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln(1/2) = -\frac{13/9}{2\sigma^2} + \ln(1/2)$$

Class 1 vs class 2

For the considered pattern, the inequality $g_1(x_t)>g_2(x_t)$ is always true.

Class 1 vs class 3

From the inequality $g_1(x_t)>g_3(x_t)$, we obtain

$$-\frac{10/9}{2\sigma^2}+\ln(1/4)>-\frac{13/9}{2\sigma^2}+\ln(1/2)$$

$$\sigma^2<\frac{3}{18}\left(\frac{1}{\ln(4)-\ln(2)}\right)=0.2404$$

Pattern $(1,1/3)^T$ is assigned to class 1 only if $\sigma^2 < 0.2404$

Pattern $(1,1/3)^T$ will be assigned to class 3 if $\sigma^2 > 0.2404$

Class 2 vs. class 3

From the inequality $g_2(x_t)>g_3(x_t)$ we obtain

$$\log(2)-\log(4)>\frac{7}{6}\frac{1}{\sigma^2}$$

This inequality never holds, accordingly:

Pattern $(1,1/3)^T$ is assigned to class 1 only if $\sigma^2 < 0.2404$

Pattern $(1,1/3)^T$ will be assigned to class 3 if $\sigma^2 > 0.2404$

Homework:

Plot boundaries for

a) equal priors
b) P1=P2=1/4, $\sigma = 0.1$
c) P1=P2=1/4, $\sigma = 0.5$

Exercise 2

Let us consider an hypothetical detection problem of network intrusions. We want to classify network traffic using two characteristic measures (two features). We have identified three data classes of interest (denial-of-service attack, probing attack, normal traffic). Traffic patterns follow a normal distribution with different, known parameters (e.g., parameters have been previously estimated using the Maximum Likelihood method).

For the sake of simplicity, let us assume that covariance matrices are equal to each other:

$$\Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{bmatrix} 4 & -3 \\ -3 & 4 \end{bmatrix},$$

Whereas Gaussian's centers are as follows

$$\mathbf{m}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{m}_2 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \mathbf{m}_3 = \begin{bmatrix} -2 \\ -2 \end{bmatrix},$$

Moreover, let us suppose that prior class probabilities are as follows (e.g. they have been estimated by the amount of traffic of each class)

$$P(\omega_1) = P(\omega_2) = \frac{1}{4}$$

A)

Find the discriminant functions using the MAP decision rule

Classify the network traffic pattern $x_c=(1,1)^t$.

B)

Find the decision boundaries

SOLUTION

The general form of a multivariate normal pdf is

$$p(x|\omega_i) = \frac{1}{(2\,\pi)^{\frac{d}{2}}\,|\Sigma_i|^{\frac{1}{2}}} \exp[-\frac{1}{2}\,(x-\mu_i)^T\,\Sigma_i^{-1}\,(x-\mu_i)]$$

The discriminant function $g_i(x)$ is in the form

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t\,\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

A)

Find the discriminant functions with the MAP decision rule

Classify the network traffic pattern $x_c=(1,1)^t$.

From Part 4 of the course, we know that if $\Sigma_i=\Sigma$, patterns generate hyper-ellipsoidal "clusters" having the same dimension and shape, centered in μ_i , whose discriminant functions become

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln P(\omega_i)$$

Alternatively, dropping the term $-\frac{1}{2}x^t \boldsymbol{\Sigma}^{-1}\mathbf{x}$ we can write

$$g_i(\mathbf{x}) = \mathbf{w}_i^t \mathbf{x} + w_{i0}$$

where

$$\mathbf{w}_i = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i \; ; \;\; w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln P(\omega_1)$$

Therefore, we can substitute the values of Σ_i e μ_i, compute $g_i(x)$ for $x=(1,1)^t$ and assign this pattern to the class i such that $g_i(x) > g_j(x)$, $j \neq i$.

The MAP decision criterion is

$$x \in \omega_i \Leftrightarrow p(\omega_i \mid x) > p(\omega_j \mid x)$$

and, using Bayes,

$$x \in \omega_i \Leftrightarrow p(x \mid \omega_i)\cdot P(\omega_i) > p(x \mid \omega_j)\cdot P(\omega_j) \;\; \forall j$$

From this equation, we obtain the discriminant functions related to $p(x \mid \omega_i)\cdot P(\omega_i)$

The covariance matrices are as follows:

$$\Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{bmatrix} 4 & -3 \\ -3 & 4 \end{bmatrix}, |\Sigma| = 7 \qquad \Sigma^{-1} = \tfrac{1}{7}\begin{pmatrix} 4 & 3 \\ 3 & 4 \end{pmatrix}$$

In the logarithmic space, we can skip common additive constants

$$g_i(x) = \left[-\frac{1}{2} \cdot (x - m_i)^T \cdot \Sigma^{-1} \cdot (x - m_i) \right] + \ln(p(\omega_i))$$

$$= -\frac{1}{2}\left[\underbrace{\boldsymbol{x}^T \Sigma^{-1} \boldsymbol{x}} + \underbrace{(-\boldsymbol{x}^T \Sigma^{-1} \boldsymbol{\mu_i} - \boldsymbol{\mu_i}^T \Sigma^{-1} \boldsymbol{x})} + \boldsymbol{\mu_i}^T \Sigma^{-1} \boldsymbol{\mu_i}\right] + \ln p(\omega_i) \Rightarrow$$

$$\Rightarrow g_i(x) = -\frac{1}{2}\left[-2\,\boldsymbol{\mu_i}^T \Sigma^{-1} \boldsymbol{x} + \boldsymbol{\mu_i}^T \Sigma^{-1} \boldsymbol{\mu_i}\right] + \ln p(\omega_i) =$$

$$= \boldsymbol{\mu_i}^T \Sigma^{-1} \boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu_i}^T \Sigma^{-1} \boldsymbol{\mu_i} + \ln p(\omega_i)$$

So, for each class we have (remember that $P_1=P_2=1/4$):

$$g_1(\boldsymbol{x}) = \ln p(\omega_1) = \ln \frac{1}{4}$$

$$g_2(\boldsymbol{x}) = (2\ \ 2)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - 4\ + \ln p(\omega_2)\ = 2(x_1 + x_2) - 4\ + \ln \frac{1}{4}$$

$$g_3(\boldsymbol{x}) = -(2\ \ 2)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - 4\ + \ln p(\omega_3)\ = -2(x_1 + x_2) - 4\ + \ln \frac{1}{2}$$

As expected, being all covariance matrices equal to each other, we obtain **linear discriminant functions**. Decision boundaries are defined by planes placed between each "neighboring" class.

Let us classify pattern $x_c=(1,1)^t$.

$g_1(x_c)$=-1.3863
$g_2(x_c)$=-1.3863
$g_3(x_c)$=-8.6931

Thus, such a point can be assigned to class 1 or 2.

B) Find the decision boundaries.

Graphs shown below are not strictly necessary for the resolution of the problem; they are shown for clarity.

The eigenvectors of Σ are as follows: $\begin{pmatrix} -\sqrt{2}/2 \\ -\sqrt{2}/2 \end{pmatrix}$ and $\begin{pmatrix} -\sqrt{2}/2 \\ \sqrt{2}/2 \end{pmatrix}$, with eigenvalues 1 and 7, respectively;

Gaussians show this shape:

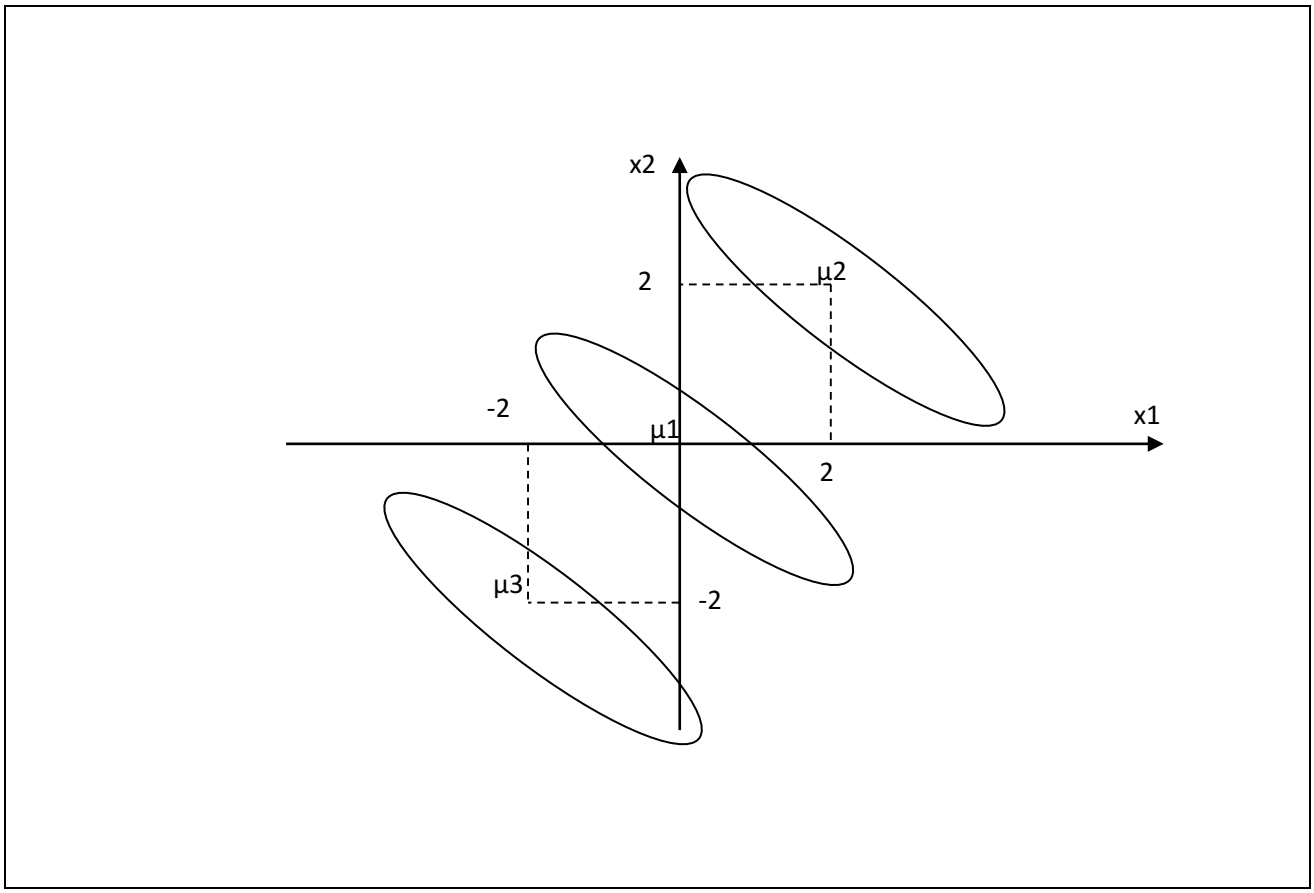


Fig.1 Probability densities of classes (schematic representation of contour lines)- prior probabilities are not considered

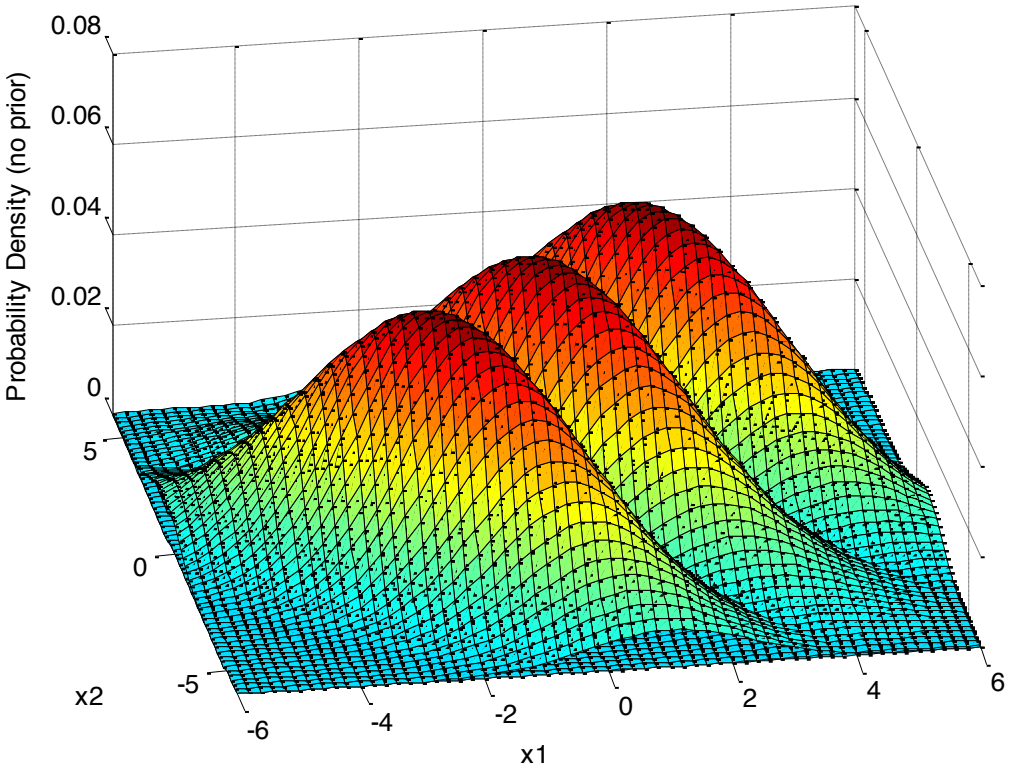


Fig.2 Probability densities of classes - prior probabilities are not considered

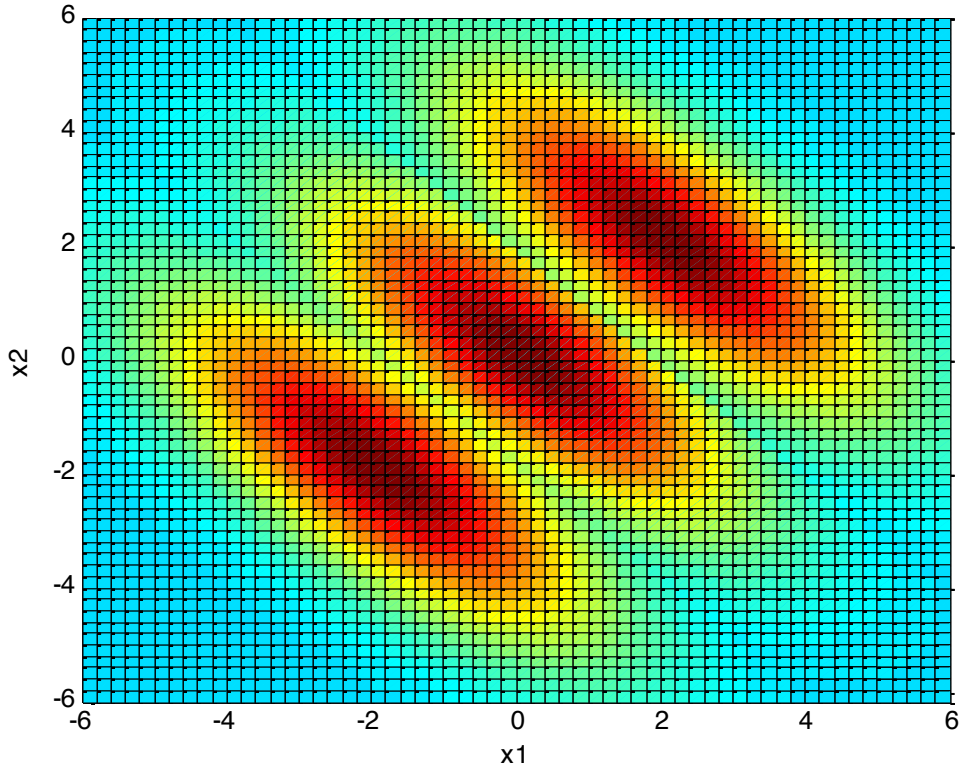


Fig.3 Probability densities of classes - prior probabilities are not considered

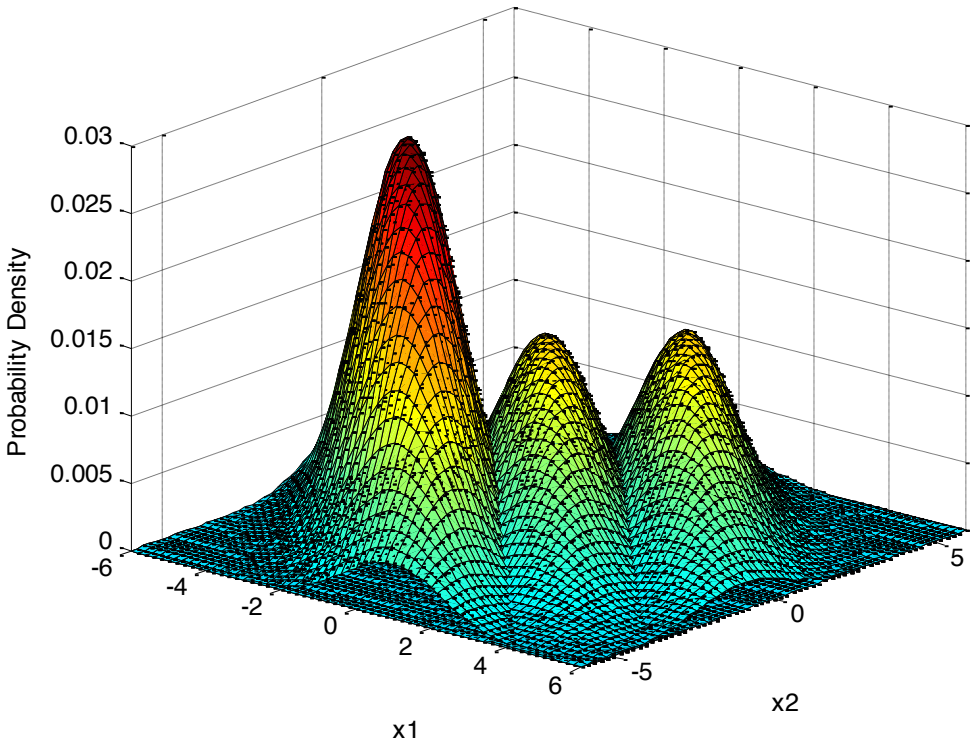


Fig.4 Probability densities of classes - prior probabilities are 1/4 (class 1), 1/4 (class 2), 1/2 (class 3)

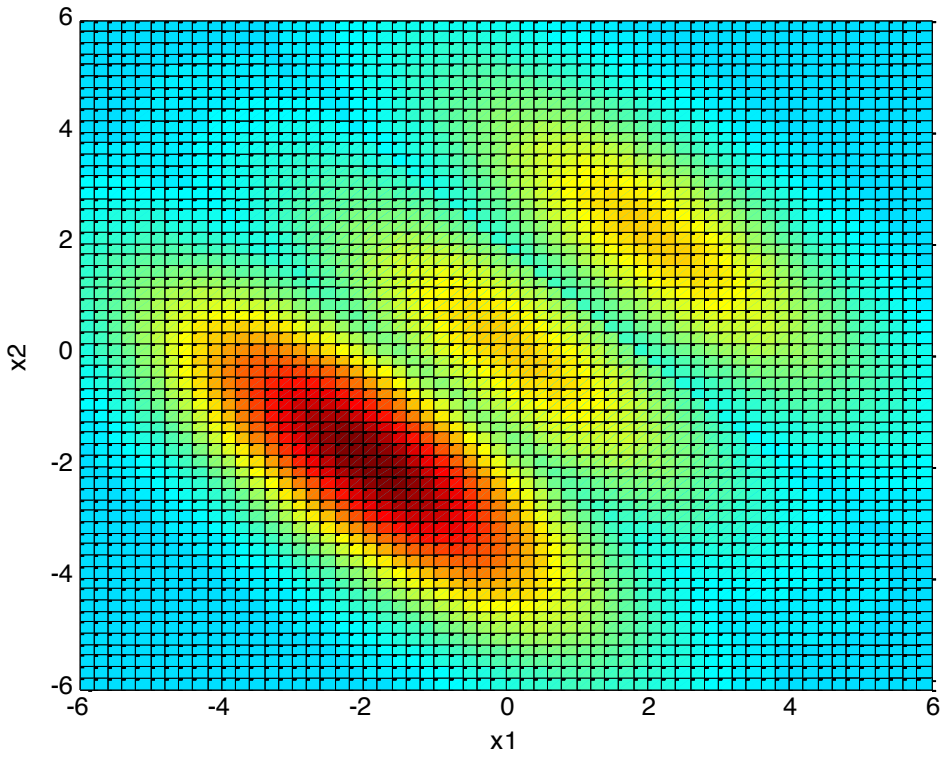


Fig.4 Probability densities of classes - prior probabilities are 1/4 (class 1), 1/4 (class 2), 1/2 (class 3)

Decision boundaries can be found considering that they correspond to $g_i(x) = g_j(x)$ - that is, to the set of points for which the posterior probability for class *i* and class *j* are identical - provided that no other class has higher probability.

Discriminant functions are:

$$g_1(\boldsymbol{x}) = \ln\frac{1}{4}$$

$$g_2(\boldsymbol{x}) = 2(x_1 + x_2) - 4 \ + \ln\frac{1}{4}$$

$$g_3(\boldsymbol{x}) = -2(x_1 + x_2) - 4 \ + \ln\frac{1}{2}$$

First decision boundary:

$$g_1(\boldsymbol{x}) = g_2(\boldsymbol{x}) \Longrightarrow 2(x_1 + x_2) - 4 = 0$$

$$(x_1 + x_2) = 2$$

$\hat{\mathbf{x}} = \begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \end{pmatrix}$ is defined by equation $x_1 + x_2 = 2$

Substituting this vector on the discriminant functions g1( ), g2 ( ) and g3 ( ) we obtain the value of these functions for the boundary defined by g1( )= g2 ( )

$g_1(\hat{\mathbf{x}}) = -1.3863;$

$g_2(\hat{\mathbf{x}}) = -1.3863;$

$g_3(\hat{\mathbf{x}}) = -8.6931;$

Obviously, we find the equality $g_1(\hat{\mathbf{x}}) = g_2(\hat{\mathbf{x}})$.

The inequality $g_3(\hat{\mathbf{x}}) < g_1(\hat{\mathbf{x}}) = g_2(\hat{\mathbf{x}})$ tells us that the first hyper-plane is a decision boundary between class 1 and 2, for all values of $\hat{x}$

Second decision boundary

$$g_2(\boldsymbol{x}) = g_3(\boldsymbol{x}) \Longrightarrow 2(x_1 + x_2) + \ln\frac{1}{4} = -2(x_1 + x_2) + \ln\frac{1}{2}$$

$$\Rightarrow (x_1 + x_2) = \frac{1}{4}(\ln 2)$$

$\hat{\mathbf{x}} = \begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \end{pmatrix}$ is defined by equation $x_1 + x_2 = \frac{1}{4}\ln 2$

Substituting this vector on the discriminant functions g1( ), g2 ( ) and g3 ( ) we obtain the value of these functions defined by g2( )= g3 ( )

$g_1(\hat{\mathbf{x}}) = -1.3863;$
$g_2(\hat{\mathbf{x}}) = -5.0397;$
$g_3(\hat{\mathbf{x}}) = -5.0397;$

Obviously, we find the equality $g_2(\hat{\mathbf{x}}) = g_3(\hat{\mathbf{x}})$.

The inequality $g_1(\hat{\mathbf{x}}) > g_2(\hat{\mathbf{x}}) = g_3(\hat{\mathbf{x}})$ tells us that the second hyper-plane is **not** a decision boundary between class 2 and 3, as for points corresponding to $\hat{\boldsymbol{x}}$ the probability for class 1 is greater than the probability for classes 2 and 3.

Therefore, this hyper-plane is NOT a decision boundary, since $g_1(\hat{\mathbf{x}}) > g_j(\hat{\mathbf{x}}), j = 2,3.$

Third decision boundary

$$g_1(\boldsymbol{x}) = g_3(\boldsymbol{x}) \Rightarrow \ln\frac{1}{4} = -2(x_1 + x_2) - 4 + \ln\frac{1}{2}$$

$$x_1 + x_2 = \frac{1}{2}\ln 2 - 2$$

$\hat{\mathbf{x}} = \begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \end{pmatrix}$ is defined by equation $x_1 + x_2 == \frac{1}{2}\ln 2 - 2$

Substituting this vector on the discriminant functions g1( ), g2 ( ) and g3 ( ) we obtain the value of these functions in the hyper-plane defined by g1( )= g3 ( )

$g_1(\hat{\mathbf{x}}) = -1.3863;$
$g_2(\hat{\mathbf{x}}) = \text{-}8.6931;$
$g_3(\hat{\mathbf{x}}) = \text{-}1.3863;$

Obviously, the value of g1 is constant, and we find the equality $g_1(\hat{\mathbf{x}}) = g_3(\hat{\mathbf{x}})$.

The inequality $g_2(\hat{\mathbf{x}}) < g_1(\hat{\mathbf{x}}) = g_3(\hat{\mathbf{x}})$ tell us that the third hyper-plane is a decision boundary between class 1 and 3

Decision boundaries:

Between classes 1 and 2:

$$x_1 + x_2 = 2$$

Between classes 1 and 3:

$$x_1 + x_2 = \frac{1}{2}\ln 2 - 2 = -1.6534$$

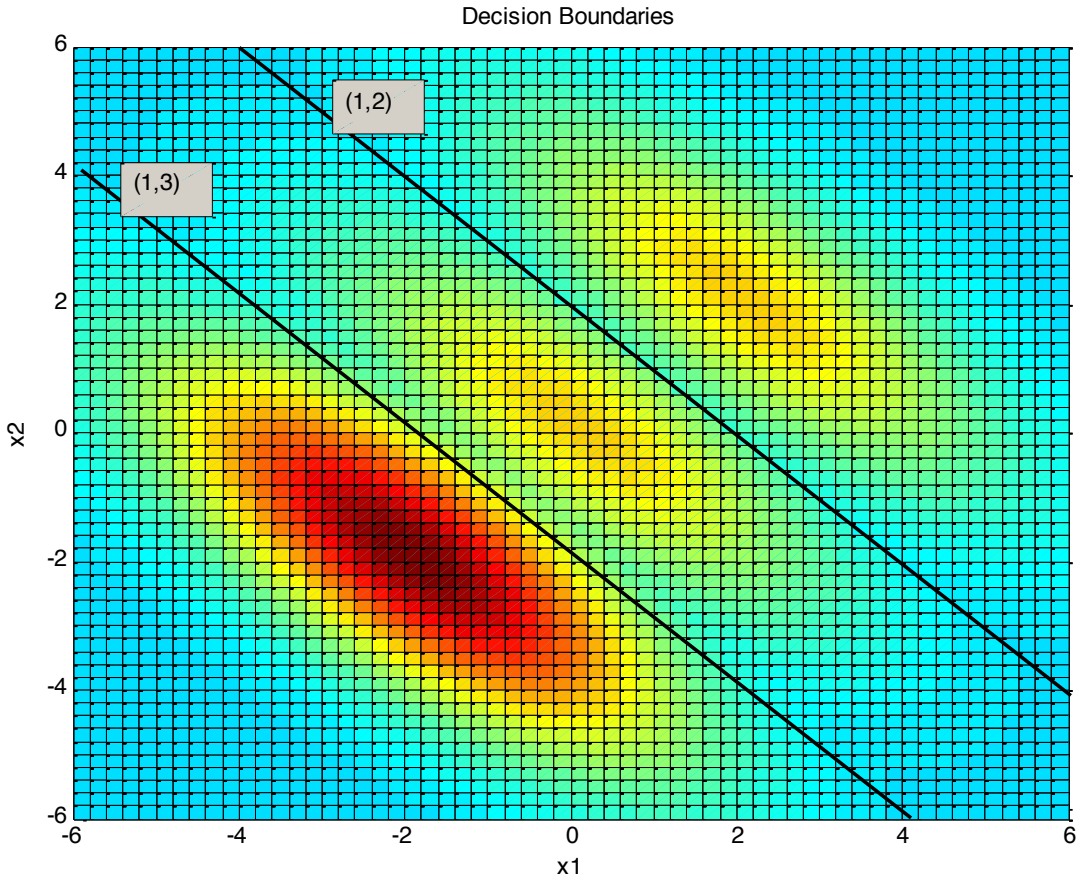


Fig.5 Probability densities and decision boundary. The decision boundary between class 1 and class 3 is identify by (1,3). The decision boundary between class 1 and class 2 is identify by (1,2).

It is easy to note that the point

$$x_c = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

lies exactly on the hyper-plane (1,2), in fact this point can be assigned to class 1 or 2 (see Exercise 2-A)

From the graphic above, it is easy to see that separation lines are not symmetric with respect to Gaussian centers, due to different class prior probabilities.

Exercise 3

Let us consider the following training samples for class ω_A and ω_B, respectively.

A	B
(-1, 0)^T	(5, 4)^T
(1, 0)^T	(3, 3)^T
(-0.1, 2)^T	(4, 7)^T
(0.3, -1.8)^T	(6.5, 2)^T

Class-conditional distributions are as follows:

$p(\mathbf{x}|\omega_A) = N(\mu_A, \Sigma_A);\ p(\mathbf{x}|\omega_B) = N(\mu_B, \Sigma_B);$

$\mu_A, \Sigma_A, \mu_B, \Sigma_B$ **are not given but they should be estimated.**

Find the optimal discriminant function and classify the pattern $\hat{\mathbf{x}} =$ (2.2, 2.2)' under the following hypotheses:

1) $\Sigma_A = \Sigma_B = \sigma^2\mathbf{I}$ and $P(\omega_1) = P(\omega_2)$,

2) $\Sigma_A = \Sigma_B = \sigma^2\mathbf{I}$ and $P(\omega_1) = 2P(\omega_2)$

3) $\Sigma_A \neq \Sigma_B$ and $P(\omega_1) = 2P(\omega_2)$

4) $\Sigma_A = \Sigma_B$ **arbitrary and** $P(\omega_1) = 2P(\omega_2)$

Note that by changing the hypotheses, classification changes as well.

SOLUTION

The general form of a multivariate normal pdf is

$$p(x|\omega_i) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp[-\frac{1}{2}\ (x-\mu_i)^T\ \Sigma_i^{-1}\ (x-\mu_i)]$$

Taking the logarithm of the above expression, one yields the discriminant function $g_i(x)$ as

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t\,\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

1) Under the hypothesis $\Sigma_A = \Sigma_B = \sigma^2 \mathbf{I}$ and $P(\omega_1) = P(\omega_2)$, find the optimal discriminant function and classify the pattern $\hat{\mathbf{x}} =$ (2.2, 2.2)'

1.a classify the pattern

Under this hypothesis, we don't need to estimate the covariance matrix. We just need to exploit the general expression of g_i (see Part 4 of the course)

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

By removing the terms common to g_A and g_B, the discriminant functions are as follows:

$$g_A(\mathbf{x}) = -\left\|\mathbf{x}-\boldsymbol{\mu}_A\right\|^2$$

$$g_B(\mathbf{x}) = -\left\|\mathbf{x}-\boldsymbol{\mu}_B\right\|^2$$

And the optimal discriminant plane is:

$$g_A(\mathbf{x}) = g_B(\mathbf{x})$$

I can estimate the mean vectors using:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$$

$$\boldsymbol{\mu}_A = \begin{pmatrix} 0.05 \\ 0.05 \end{pmatrix};\ \boldsymbol{\mu}_B = \begin{pmatrix} 4.625 \\ 4 \end{pmatrix}$$

Squared distances of X from the above means:

$d_A^2 = 9.2450$; $d_B^2 = 9.1206$

$g_A(X) = -9.2450$; $g_B(X) = -9.1206$;

Pattern is assigned to class B

1.b Find the boundary

The boundary is defined by the equation

$$\left||\boldsymbol{x}-\boldsymbol{\mu_A}|\right|^2 = \left||\boldsymbol{x}-\boldsymbol{\mu_B}|\right|^2$$

where the generic 2-dimensional vector **x** is $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, that is,

$$\left\| \begin{pmatrix} x_1 - \mu_{A1} \\ x_2 - \mu_{A2} \end{pmatrix} \right\|^2 = \left\| \begin{pmatrix} x_1 - \mu_{B1} \\ x_2 - \mu_{B2} \end{pmatrix} \right\|^2$$

$$(x_1 - \mu_{A1})^2 + (x_2 - \mu_{A2})^2 = (x_1 - \mu_{B1})^2 + (x_2 - \mu_{B2})^2$$

…

after a simple calculation, we obtain

$$9.15\, x_1 + 7.9\, x_2 = 37.3856$$

An alternative approach is

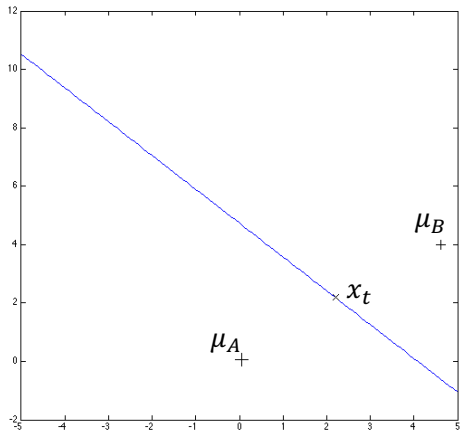
$$\left||\boldsymbol{x}-\boldsymbol{\mu_A}|\right|^2 = \left||\boldsymbol{x}-\boldsymbol{\mu_B}|\right|^2$$

$$\left(\mathbf{x}-\boldsymbol{\mu}_A\right)^T\left(\mathbf{x}-\boldsymbol{\mu}_A\right)=\left(\mathbf{x}-\boldsymbol{\mu}_B\right)^T\left(\mathbf{x}-\boldsymbol{\mu}_B\right)$$

$$-\mathbf{x}^T\boldsymbol{\mu}_A-\boldsymbol{\mu}_A{}^T\mathbf{x}\ +\boldsymbol{\mu}_A{}^T\boldsymbol{\mu}_A=-\mathbf{x}^T\boldsymbol{\mu}_B-\boldsymbol{\mu}_B{}^T\mathbf{x}\ +\boldsymbol{\mu}_B{}^T\boldsymbol{\mu}_B$$

$$-2\boldsymbol{\mu}_A{}^T\mathbf{x}\ +\boldsymbol{\mu}_A{}^T\boldsymbol{\mu}_A=-2\boldsymbol{\mu}_B{}^T\mathbf{x}\ +\boldsymbol{\mu}_B{}^T\boldsymbol{\mu}_B$$

$$9.15\, x_1 + 7.9\, x_2 = 37.3856$$



The point (2.2, 2.2) is slightly above the boundary, and thus classified as class B (as discussed before).

2) Under the hypothesis $\Sigma_A = \Sigma_B = \sigma^2 \mathbf{I}$ and $P(\omega_1) = 2P(\omega_2)$, find the optimal discriminant function and classify pattern $\hat{\mathbf{x}} =$ (2.2, 2.2)'

As in previous exercise, we can start our calculation from the general form of the discriminant function

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

by removing the terms common to both members:

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x}-\boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln(P(\omega_i))$$

(the quadratic term is the same in each $g_i()$ and it can thus be ignored)

We need to estimate the term σ. The estimation of covariance matrices is

$$\hat{\boldsymbol{\Sigma}}_i = \frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i)(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i)^t$$

$$\hat{\Sigma}_A = \begin{pmatrix} 0.6967 & -0.25 \\ -0.25 & 2.41 \end{pmatrix}$$

$$\hat{\Sigma}_B = \begin{pmatrix} 2.2292 & -1.3333 \\ -1.3333 & 4.6667 \end{pmatrix}$$

But the hypothesis is $\Sigma_A = \Sigma_B = \sigma^2 \mathbf{I}$

When all classes show the same covariance matrix, it can be proven that the maximum likelihood estimate for such a covariance matrix is (see slides of Part 4 of the course):

$$\hat{\Sigma} = \sum_{i=1}^{c}\frac{n_i}{n}\hat{\Sigma}_i$$

or, if we know the priors,

$$\hat{\Sigma} = \sum_{i=1}^{c} P(\omega_i)\hat{\Sigma}_i$$

In order to enforce compliance between the obtained data and the hypothesis $\Sigma_A = \Sigma_B = \sigma^2 \boldsymbol{I}$, we may impose that these matrices are diagonal and proportional to the identity matrix (only zero values outside the diagonal). Essentially, we have to ignore out-of-diagonal terms, and average terms on the main diagonal. This corresponds to computing the variance of each feature, and then average the corresponding values (separately for each class). Finally, the average variance is computed using the above equation (a weighted mean), namely, weighing the value of each average class variance with its prior probability.

$$\Sigma_A = \begin{pmatrix} 0.6967 & 0 \\ 0 & 2.41 \end{pmatrix};$$

$$\hat{\Sigma}_A = \begin{pmatrix} \frac{0.6967 + 2.41}{2} & 0 \\ 0 & \frac{0.6967 + 2.41}{2} \end{pmatrix} = 1.5533\, \boldsymbol{I}$$

$$\Sigma_B = \begin{pmatrix} 2.2292 & 0 \\ 0 & 4.6667 \end{pmatrix}$$

$$\hat{\Sigma}_B = \begin{pmatrix} \frac{2.2292 + 4.6667}{2} & 0 \\ 0 & \frac{2.2292 + 4.6667}{2} \end{pmatrix} = 3.4479\, \boldsymbol{I}$$

The above diagonal terms of the covariance matrix have been obtained by averaging the previous variance values.

Finally, $\hat{\Sigma} = \sum_{i=1}^{c} P(\omega_i)\hat{\Sigma}_i = P_A\hat{\Sigma}_A + P_B\hat{\Sigma}_B = 2.1849\, \boldsymbol{I}$

so, $\sigma^2 = 2.1849$

2.a classify the pattern $\hat{\mathbf{x}} = (2.2, 2.2)'$

$$g_A(x_t) = -\frac{||x_t - \mu_A||^2}{2\,\sigma^2} + \log(P(\omega_A)) = -2.5212$$

$$g_B(x_t) = -\frac{||x_t - \mu_B||^2}{2\,\sigma^2} + \log(P(\omega_B)) = -3.1858$$

$g_B(\mathbf{x}) < g_A(\mathbf{x})$; **The pattern is assigned to class A**

Alternative approach

The discriminant functions can be expressed as

$g_i(x) = w_i^T \mathbf{x} + w_{i0}$, where

$$w_i = \frac{1}{\sigma^2}\mu_i; \;\; w_{i0} = -\frac{1}{2\sigma^2}\mu_i^T \mu_i + \log P(\omega_i)$$

(see SLIDES, chapt. 4)

$$w_A = \begin{pmatrix} 0.0229 \\ 0.0229 \end{pmatrix}; w_{A0} = -0.4066$$

$$w_B = \begin{pmatrix} 2.1168 \\ 1.8308 \end{pmatrix}; w_{B0} = -9.6554$$

$g_A(\mathbf{x}) = w_A^T \mathbf{x} + w_{A0} = -0.3059$

$g_B(\mathbf{x}) = w_B^T \mathbf{x} + w_{B0} = -0.9706$

$g_B(\mathbf{x}) < g_A(\mathbf{x})$; **The pattern is assigned to class A**

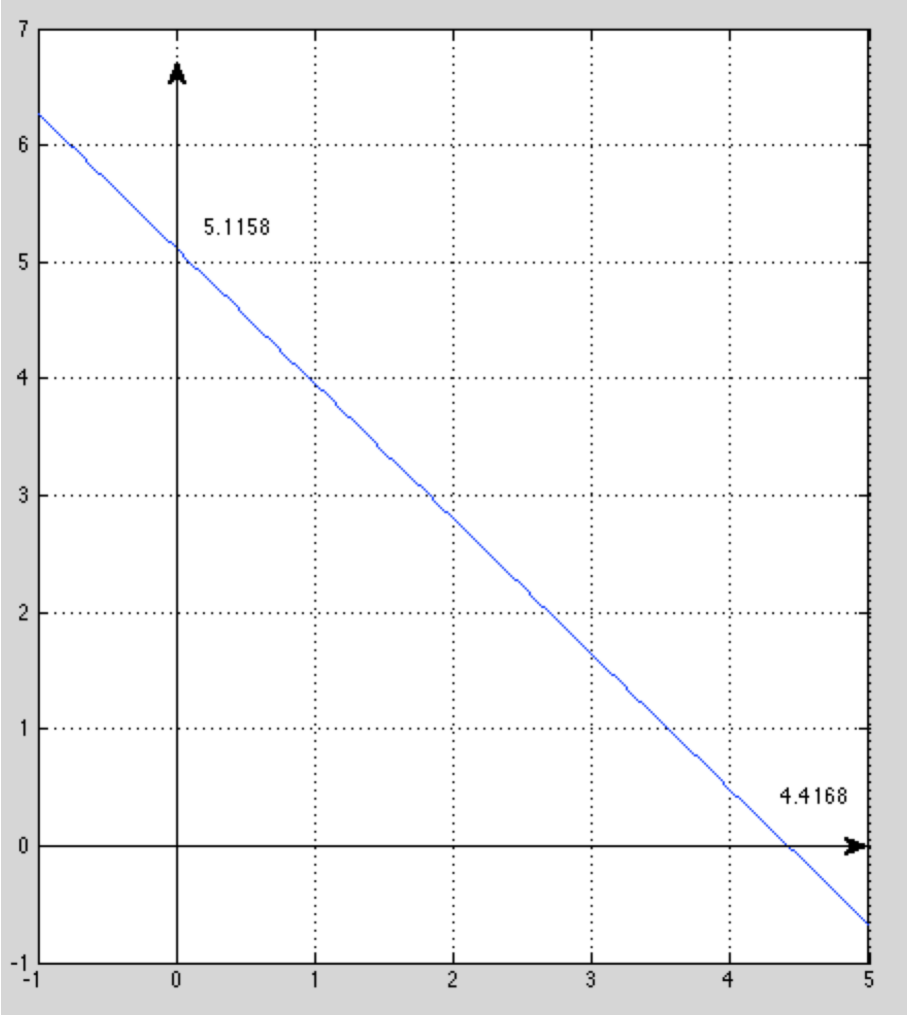
2.b find boundary

$$g_A(x) = -\frac{||x-\mu_A||^2}{2\,\sigma^2} + \log(P(\omega_A))$$

$$g_B(x) = -\frac{||x-\mu_B||^2}{2\,\sigma^2} + \log(P(\omega_B))$$

The optimal discriminant plane is the hyperplane $g_A=g_B$, that is,

$$2.094\,x_1 + 1.808\,x_2 = 9.2488$$



It is easy to see that (compare with point 1.b in the solution of this exercise) the boundary has been shifted towards the upper-right corner of the plot, due to the increase of the prior of the class A. Accordingly, the point (2.2, 2.2) is now below the decision boundary (thus in the decision region of class A).

(alternative approach)

The optimal discriminant plane is the hyperplane $g_A=g_B$, that is,

$\boldsymbol{w}_A^T\,\boldsymbol{x} + w_{A0} = \boldsymbol{w}_B^T\,\boldsymbol{x} + w_{B0}$

$(\boldsymbol{w}_A^T - \boldsymbol{w}_B^T)\boldsymbol{x} + w_{A0} - w_{B0} = 0$

The hyperplane can be written as

$$\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0$$

that is,

$2.094\,x_1 + 1.808\,x_2 = 9.2488$

3) By assuming that $\Sigma_A \neq \Sigma_B$ e $P(\omega_1) = 2P(\omega_2)$, **find the optimal discriminant boundary and classify pattern** $\hat{\mathbf{x}} =$ (2.2, 2.2)'

We can estimate the covariance matrices (see previous exercise):

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t$$

$$\Sigma_A = \begin{pmatrix} 0.6967 & -0.25 \\ -0.25 & 2.41 \end{pmatrix}$$

$$\Sigma_B = \begin{pmatrix} 2.2292 & -1.3333 \\ -1.3333 & 4.6667 \end{pmatrix}$$

$g_i(x)$ is a quadratic function which can be written as

$$g_i(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{W}_i\boldsymbol{x} + \boldsymbol{w}_i'\,\boldsymbol{x} + w_{i0}$$

where

$$\boldsymbol{W}_i = -\tfrac{1}{2}\,\boldsymbol{\Sigma}_i^{-1};\; w_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$$

$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

Substituting:

$$\boldsymbol{W}_A = \begin{pmatrix} -0.7454 & -0.0773 \\ -0.0773 & -0.2155 \end{pmatrix}$$

wa = [0.0823, 0.0293]'

wa0 = -0.6484

$$\boldsymbol{W}_B = \begin{pmatrix} -0.2705 & -0.0773 \\ -0.0773 & -0.1292 \end{pmatrix}$$

wb =[3.1207 , 1.7487]'

wb0 = -12.8893

we can compute the discriminant functions and classify the pattern $\boldsymbol{x}_t = \begin{pmatrix} 2.2 \\ 2.2 \end{pmatrix}$

Discriminant function:

$$g_a(\boldsymbol{x}) = -0.7454x_1^2 - 0.2155\,x_2^2 - 0.1547\,x_1x_2 + 0.08227\,x_1 + 0.02928\,x_2 - 0.6484$$

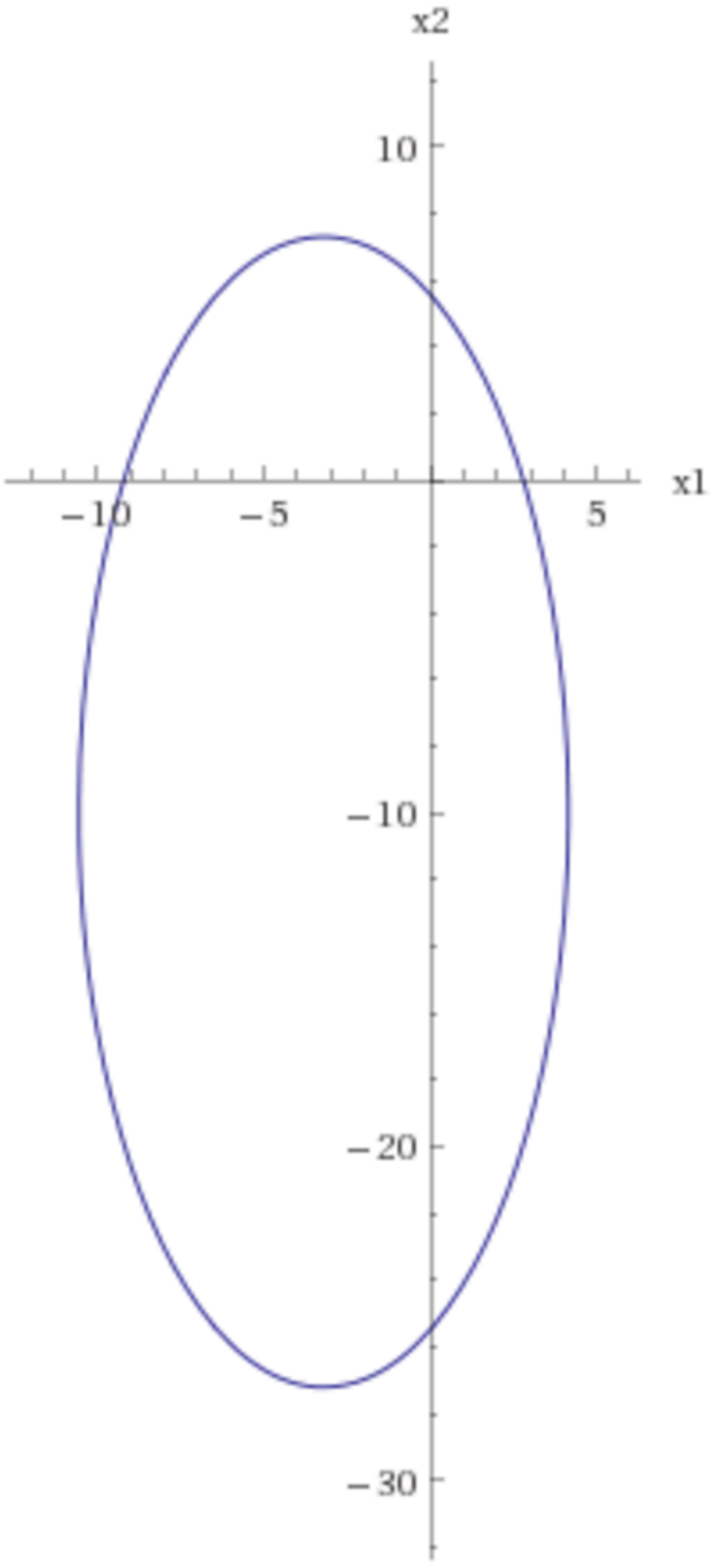
$$g_b(\boldsymbol{x}) = -0.2705\,x_1^2 - 0.1546\,x_1x_2 + 3.12\,x_1 - 0.1292\,x_2^2 + 1.749\,x_2 - 12.89$$

Boundary:

As we know from SLIDES, chapt. 4, we obtain a quadratic discriminant function

$g_a(\boldsymbol{x}) - g_b(\boldsymbol{x}) = 0 \Rightarrow$

$$\Rightarrow x_1^2 + 0.0001937\,x_1x_2 + 6.397\,x_1 + 0.1816\,x_2^2 + 3.621\,x_2 - 25.78 = 0$$



and finally

ga(**x_t**)= -5.825

gb(**x_t**)=-4 .8610

g_b(**x_t**)> g_A(**x_t**); **The pattern is assigned to class B**

4) $\Sigma_A = \Sigma_B$ **unknown and** $P(\omega_1) = 2P(\omega_2)$

Hint: we need to estimate the covariance matrix as a weighted mean of the two covariances and apply the Gaussian model which can satisfy the hypothesis $\Sigma_A = \Sigma_B$, as in Exercise 2.

Exercise 4

Let us consider a classification task with 3 data classes, a two-dimensional feature space, with Gaussian distributions for the 3 classes:

$\boldsymbol{\mu}_1 = \begin{pmatrix}0\\0\end{pmatrix};\ \boldsymbol{\mu}_2 = \begin{pmatrix}0\\1\end{pmatrix};\ \boldsymbol{\mu}_3 = \begin{pmatrix}1\\1\end{pmatrix};\ for\ each\ class\ \boldsymbol{\Sigma}_i = 2\boldsymbol{I}$

$P(\omega_1) = P(\omega_2) = \frac{1}{4}$

Find the optimal decision boundaries and the related decision regions.

Solution

Given that $\boldsymbol{\Sigma_i = \sigma^2 I}$ we have $\boldsymbol{g_i(x)}$:

$$g_i(x) = -\frac{\left\|x - \mu_i\right\|^2}{2\,\sigma^2} + \log P_i$$

$g_1(x) = -\frac{1}{4}(x_1^2 + x_2^2) + \log\frac{1}{4}$

$g_2(x) = -\frac{1}{4}(x_1^2 + x_2^2) - \frac{1}{4}(1 - 2\,x_2) + \log\frac{1}{4}$

$g_3(x) = -\frac{1}{4}(x_1^2 + x_2^2) - \frac{1}{4}(1 - 2\,x_1 + 1 - 2\,x_2) + \log\frac{1}{2}$

Equal terms can be deleted, so obtaining:

$g_1(x) = \log\frac{1}{4}$

$g_2(x) = -\frac{1}{4}(1 - 2\,x_2) + \log\frac{1}{4}$

$g_3(x) = -\frac{1}{4}(1 - 2\,x_1 + 1 - 2\,x_2) + \log\frac{1}{2}$

Decision boundary between class 1 and class 2

$g_1(x) = g_2(x) \Longrightarrow x_2 = \frac{1}{2}$

For $\widehat{\boldsymbol{x}} \equiv (x_2 = \frac{1}{2})$ we have

$g_1(\widehat{\boldsymbol{x}}) = g_2(\widehat{\boldsymbol{x}}) = \log\frac{1}{4}$;

$g_3(\widehat{\boldsymbol{x}}) = -\frac{1}{4}(1 - 2\,x_1) + \log\frac{1}{2}$

$x_2 = \frac{1}{2}$ is the boundary if and only if $g_1(\widehat{\boldsymbol{x}}) = g_2(\widehat{\boldsymbol{x}}) > g_3(\widehat{\boldsymbol{x}})$, therefore

Boundary between class 1 and 2

$$\begin{cases} x_2 = 1/2 \\ if\ x_1 < \frac{1}{2} - 2\log 2 \cong -0.8863 \end{cases}$$

Decision boundary between class 1 and class 3

$g_1(x) = g_3(x) \Longrightarrow x_1 + x_2 = 1 - 2\log 2$

For $\widehat{\boldsymbol{x}} \equiv (x_1 + x_2 = 1 - 2\ \log 2)$ we have

$g_1(\widehat{\boldsymbol{x}}) = g_3(\widehat{\boldsymbol{x}}) = \log\frac{1}{4}$;

$g_2(\widehat{\boldsymbol{x}}) = \frac{1}{2}x_2 - \frac{1}{4} + \log\frac{1}{4}$

$x_1 + x_2 = 1 - 2\ \log 2$ is the boundary only if $g_1(\widehat{\boldsymbol{x}}) = g_3(\widehat{\boldsymbol{x}}) > g_2(\widehat{\boldsymbol{x}})$, therefore

Boundary between class 1 and class 3

$$\begin{cases} x_1 + x_2 = 1 - 2\log 2 \cong -0.3863 \\ if\ x_2 < \frac{1}{2} \end{cases}$$

Decision boundary between class 2 and classe 3

$g_2(x) = g_3(x) \Longrightarrow x_1 = \frac{1}{2} - 2\ \log 2$

For $\widehat{\boldsymbol{x}} \equiv (x_1 = \frac{1}{2} - 2\ \log 2\ \ )$ we have

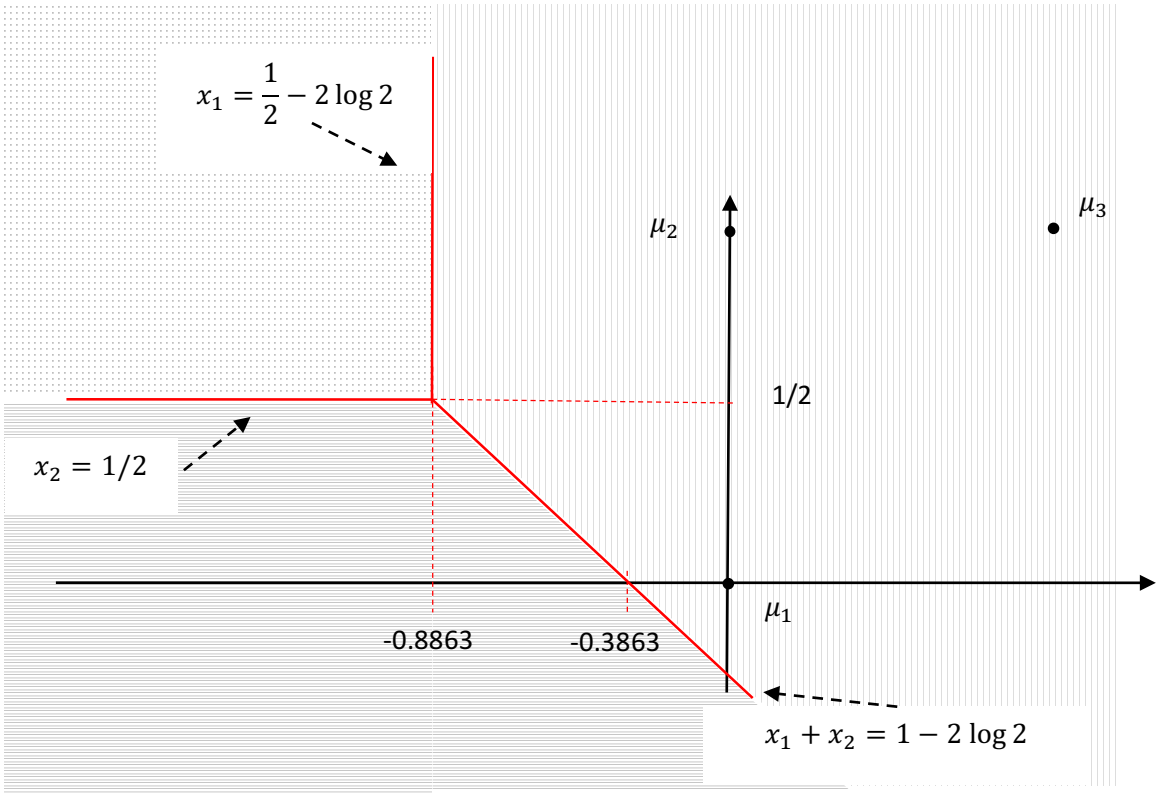
$g_2(\widehat{\boldsymbol{x}}) = g_3(\widehat{\boldsymbol{x}}) = -\frac{1}{4}(1 - 2\ x_2) + \log\frac{1}{4}$;

$g_1(\widehat{\boldsymbol{x}}) = \log\frac{1}{4}$

$-\frac{1}{4}(1 - 2\ x_2) + \log\frac{1}{4}$ is the boundary only if $g_2(\widehat{\boldsymbol{x}}) = g_3(\widehat{\boldsymbol{x}}) > g_1(\widehat{\boldsymbol{x}})$, therefore

Boundary class 2 and class 3

$$\begin{cases} x_1 = \frac{1}{2} - 2\log 2 \cong -0.8863 \\ \qquad se\ x_2 > \frac{1}{2} \end{cases}$$



Legenda

class 1 decision region

class 2 decision region

class 3 decision region

Due to the difference of priors, both μ_1 and μ_2 lie in the decision region of class 3.

All the course material is available on the web site

Course web site: http://pralab.diee.unica.it/MachineLearning