L5: Quadratic classifiers

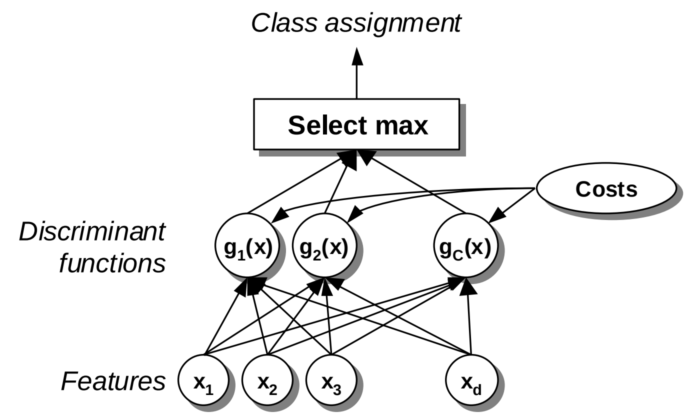
Bayes classifiers for Normally distributed classes

- Case 1: $\Sigma_i = \sigma^2 I$
- Case 2: $\Sigma_i = \Sigma$ (Σ diagonal)
- Case 3: $\Sigma_i = \Sigma$ (Σ non-diagonal)
- Case 4: $\Sigma_i = \sigma_i^2 I$
- Case 5: $\Sigma_i \neq \Sigma_j$ (general case)

Numerical example

Linear and quadratic classifiers: conclusions

Bayes classifiers for Gaussian classes



Recap

- On L4 we showed that the decision rule that minimized $P[error]$ could be formulated in terms of a family of discriminant functions

For normally Gaussian classes, these DFs reduce to simple expressions

- The multivariate Normal pdf is

$$f_X(x) = (2\pi)^{-N/2}|\Sigma|^{-1/2}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

- Using Bayes rule, the DFs become

$$g_i(x) = P(\omega_i|x) = \big(P(\omega_i)p(x|\omega_i)\big)/p(x)$$

$$= (2\pi)^{-N/2}|\Sigma_i|^{-1/2}e^{-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)}\, P(\omega_i)/p(x)$$

- Eliminating constant terms

$$g_i(x) = |\Sigma_i|^{-1/2}\, e^{-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)}\, P(\omega_i)$$

- And taking natural logs

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\log|\Sigma_i| + logP(\omega_i)$$

- This expression is called a **quadratic discriminant function**

Case 1: $\Sigma_i = \sigma^2 I$

This situation occurs when features are statistically independent with equal variance for all classes

– In this case, the quadratic DFs become

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T(\sigma^2 I\,)^{-1}(x-\mu_i) - \frac{1}{2}\log|\sigma^2 I| + logP_i \equiv -\frac{1}{2\sigma^2}(x-\mu_i)^T(x-\mu_i) + logP_i$$

– Expanding this expression

$$g_i(x) = -\frac{1}{2\sigma^2}\left(x^T x - 2\mu_i^T x + \mu_i^T \mu_i\right) + logP_i$$

– Eliminating the term $x^T x$, which is constant for all classes

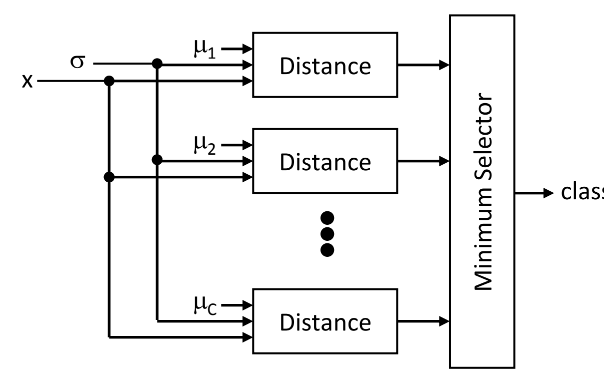
$$g_i(x) = -\frac{1}{2\sigma^2}\left(-2\mu_i^T x + \mu_i^T \mu_i\right) + logP_i = w_i^T x + w_0$$

– So the DFs are linear, and the boundaries $g_i(x) = g_j(x)$ are hyper-planes

– If we assume equal priors

$$g_i(x) = -\frac{1}{2\sigma^2}(x-\mu_i)^T(x-\mu_i)$$

- This is called a minimum-distance or nearest mean classifier
- The equiprobable contours are hyper-spheres
- For unit variance ($\sigma^2 = 1$), $g_i(x)$ is the Euclidean distance



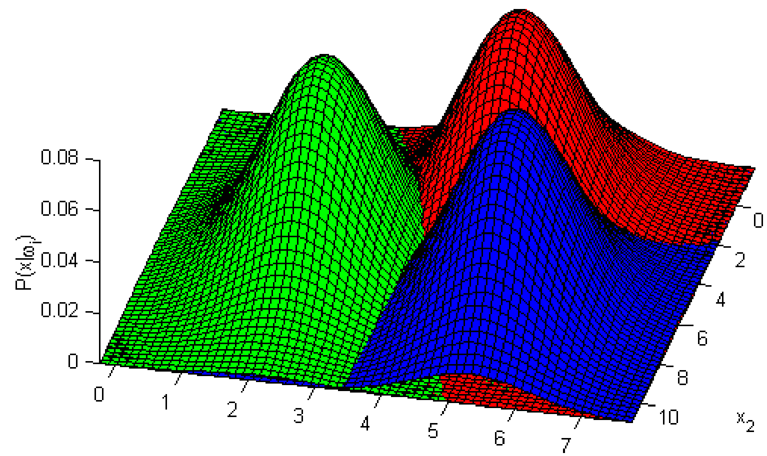
[Schalkoff, 1992]

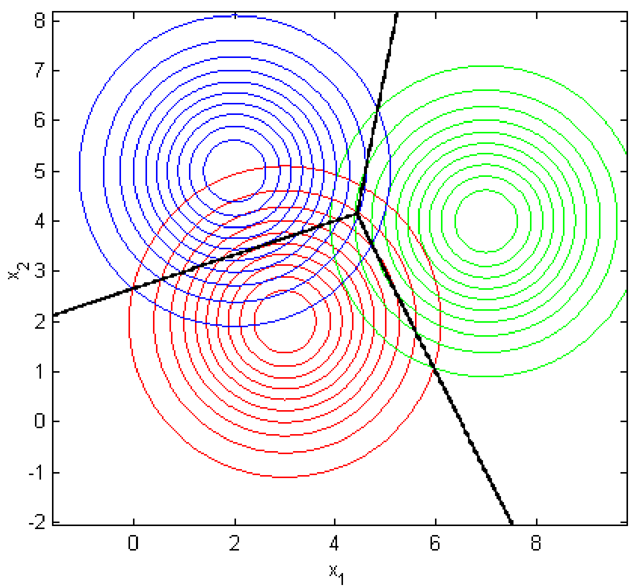
Example

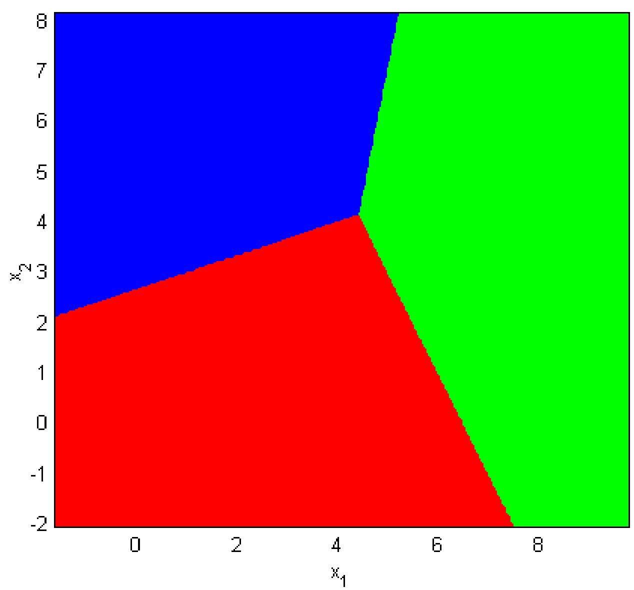
– Three-class 2D problem with equal priors

$$\mu_1 = [3\ 2]^T \qquad \mu_2 = [7\ 4]^T \qquad \mu_3 = [2\ 5]^T$$

$$\Sigma_1 = \begin{bmatrix} 2 & \\ & 2 \end{bmatrix} \qquad \Sigma_1 = \begin{bmatrix} 2 & \\ & 2 \end{bmatrix} \qquad \Sigma_1 = \begin{bmatrix} 2 & \\ & 2 \end{bmatrix}$$







Case 2: $\Sigma_i = \Sigma$ (diagonal)

Classes still have the same covariance, but features are allowed to have different variances

– In this case, the quadratic DFs becomes

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\log|\Sigma_i| + logP_i =$$

$$-\frac{1}{2}\sum_{k=1}^{N}\frac{\left(x_k-\mu_{i,k}\right)^2}{\sigma_k^2} - \frac{1}{2}log\prod_{k=1}^{N}\sigma_k^2 + logP_i$$

– Eliminating the term x_k^2, which is constant for all classes

$$g_i(x) = -\frac{1}{2}\sum_{k=1}^{N}\frac{-2x_k\mu_{i,k}+\mu_{i,k}^2}{\sigma_k^2} - \frac{1}{2}log\Pi_{k=1}^{N}\sigma_k^2 + logP_i$$

– This discriminant is also linear, so the decision boundaries $g_i(x) = g_j(x)$ will also be hyper-planes

– The equiprobable contours are hyper-ellipses aligned with the reference frame

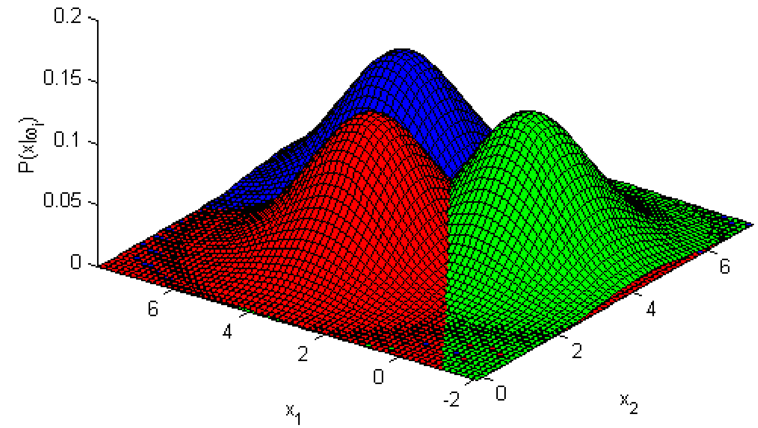
– Note that the only difference with the previous classifier is that the distance of each axis is normalized by its variance

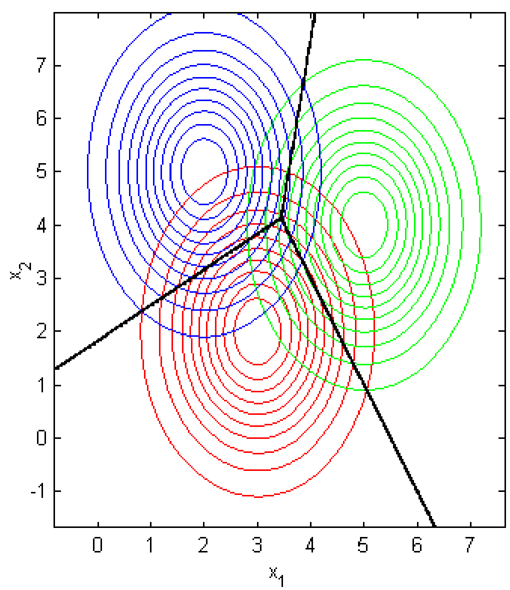
Example

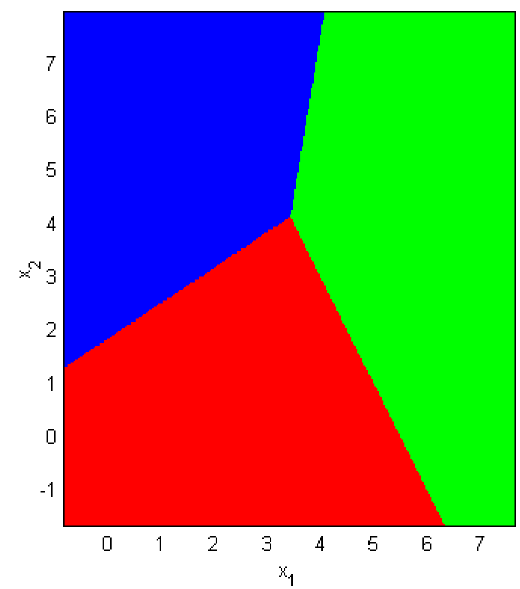
– Three-class 2D problem with equal priors

$$\mu_1 = [3\ 2]^T \qquad \mu_2 = [5\ 4]^T \qquad \mu_3 = [2\ 5]^T$$

$$\Sigma_1 = \begin{bmatrix} 1 & \\ & 2 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 1 & \\ & 2 \end{bmatrix} \qquad \Sigma_3 = \begin{bmatrix} 1 & \\ & 2 \end{bmatrix}$$







Case 3: $\Sigma_i = \Sigma$ (non-diagonal)

Classes have equal covariance matrix, but no longer diagonal

- The quadratic discriminant becomes

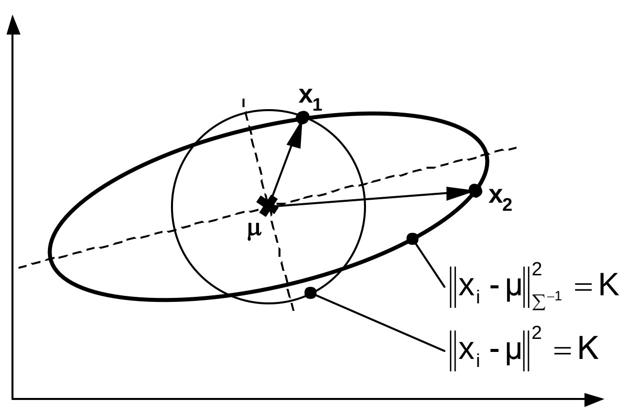
$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma^{-1} (x-\mu_i) - \frac{1}{2}\log|\Sigma| + logP_i$$

- Eliminating the term $\log|\Sigma|$, which is constant for all classes, and assuming equal priors

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma^{-1} (x-\mu_i)$$

- The quadratic term is called the **Mahalanobis distance**, a very important concept in statistical pattern recognition

- The Mahalanobis distance is a vector distance that uses a Σ^{-1}norm,
- Σ^{-1} acts as a stretching factor on the space
- Note that when $\Sigma = I$, the Mahalanobis distance becomes the familiar Euclidean distance



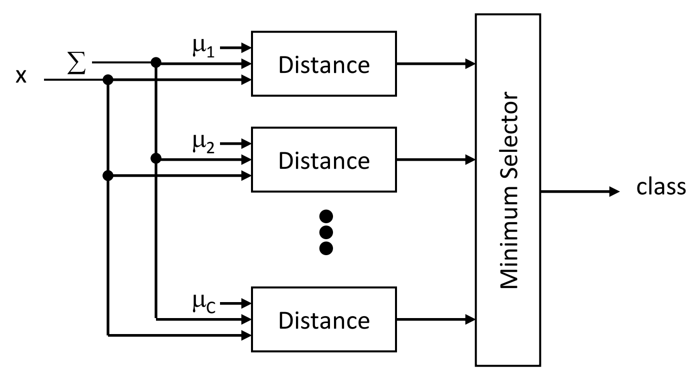
- Expanding the quadratic term

$$g_i(x) = -\frac{1}{2}\left(x^T\Sigma^{-1}x - 2\mu_i^T\Sigma^{-1}x + \mu_i^T\Sigma^{-1}\mu_i\right)$$

- Removing the term $x^T\Sigma^{-1}x$, which is constant for all classes

$$g_i(x) = -\frac{1}{2}\left(-2\mu_i^T\Sigma^{-1}x + \mu_i^T\Sigma^{-1}\mu_i\right) = w_1^T x + w_0$$

- So the DFs are still linear, and the decision boundaries will also be hyper-planes
- The equiprobable contours are hyper-ellipses aligned with the eigenvectors of Σ
- This is known as a minimum (Mahalanobis) distance classifier

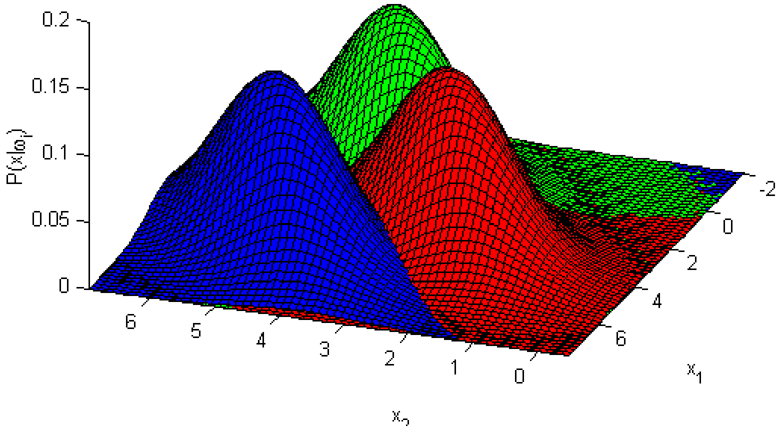


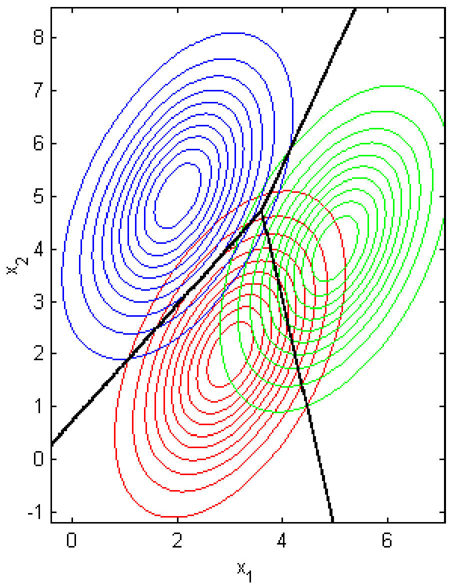
Example

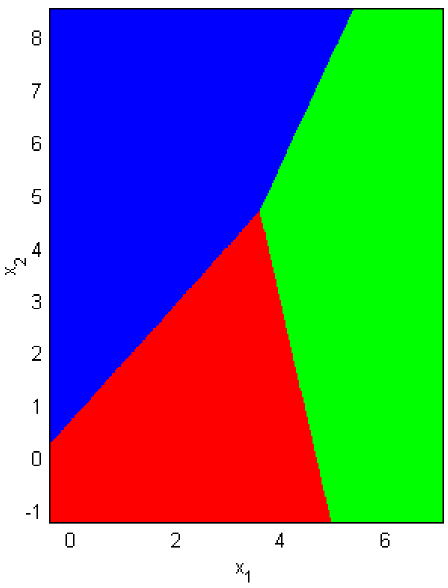
– Three-class 2D problem with equal priors

$$\mu_1 = [3\ 2]^T \qquad \mu_2 = [5\ 4]^T \qquad \mu_3 = [2\ 5]^T$$

$$\Sigma_1 = \begin{bmatrix} 1 & .7 \\ .7 & 2 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 1 & .7 \\ .7 & 2 \end{bmatrix} \qquad \Sigma_3 = \begin{bmatrix} 1 & .7 \\ .7 & 2 \end{bmatrix}$$







Case 4: $\Sigma_i = \sigma_i^2 I$

In this case, each class has a different covariance matrix, which is proportional to the identity matrix

- The quadratic discriminant becomes

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \sigma_i^{-2}(x-\mu_i) - \frac{1}{2} N \log\left|\sigma_i^2\right| + logP_i$$

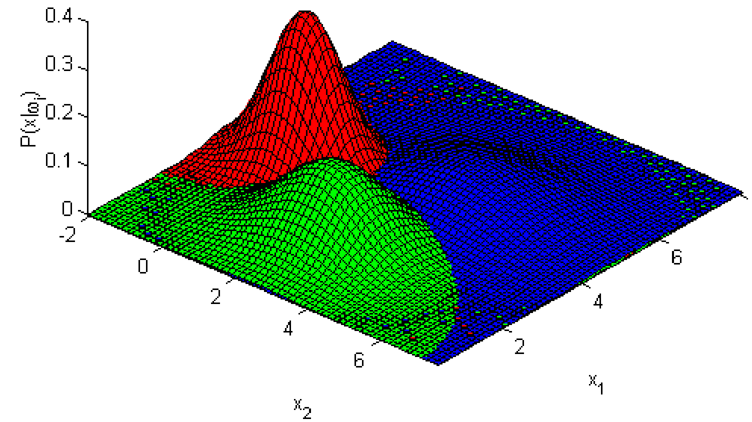
- This expression cannot be reduced further
- The decision boundaries are quadratic: hyper-ellipses
- The equiprobable contours are hyper-spheres aligned with the feature axis

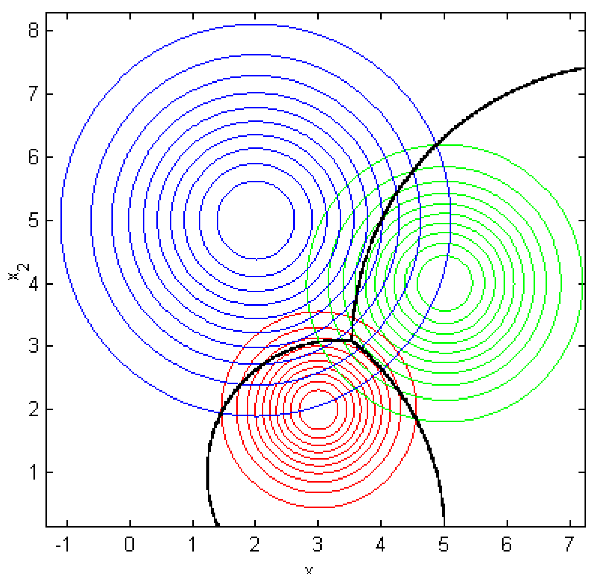
Example

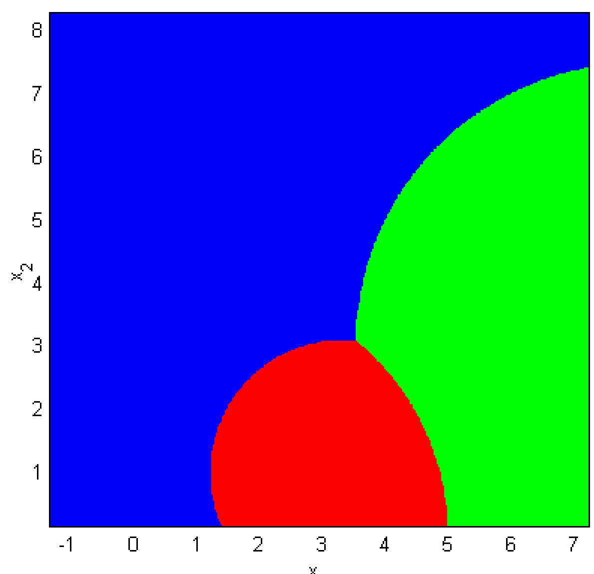
– Three-class 2D problem with equal priors

$$\mu_1 = [3\ 2]^T \qquad \mu_2 = [5\ 4]^T \qquad \mu_3 = [2\ 5]^T$$

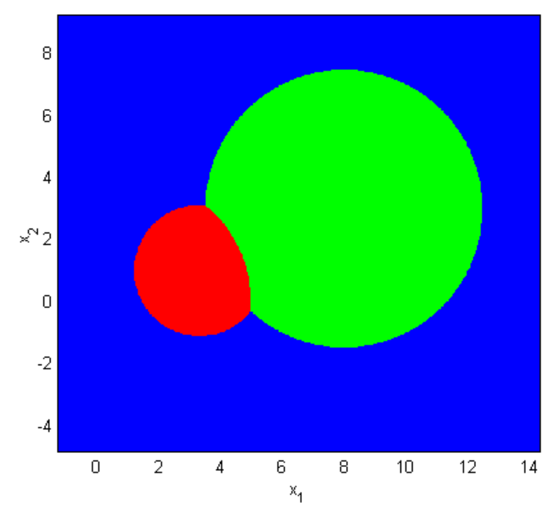
$$\Sigma_1 = \begin{bmatrix} .5 & \\ & .5 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 1 & \\ & 1 \end{bmatrix} \qquad \Sigma_3 = \begin{bmatrix} 2 & \\ & 2 \end{bmatrix}$$







Zoom out



Case 5: $\Sigma_i \neq \Sigma_j$ (general case)

We already derived the expression for the general case

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\log|\Sigma_i| + logP_i$$

– Reorganizing terms in a quadratic form yields

$$g_i(x) = x^T W_{2,i} x + w_{1,i}^T x + w_{0,i}$$

$$\text{where} \begin{cases} W_{2,i} = -\frac{1}{2}\Sigma_i^{-1} \\ w_{1,i} = \Sigma_i^{-1}\mu_i \\ w_{o,i} = -\frac{1}{2}\mu_i^T\Sigma_i^{-1}\mu_i - \frac{1}{2}\log|\Sigma_i| + logP_i \end{cases}$$

– The equiprobable contours are hyper-ellipses, oriented with the eigenvectors of Σ_i for that class

– The decision boundaries are again quadratic: hyper-ellipses or hyper-parabolloids

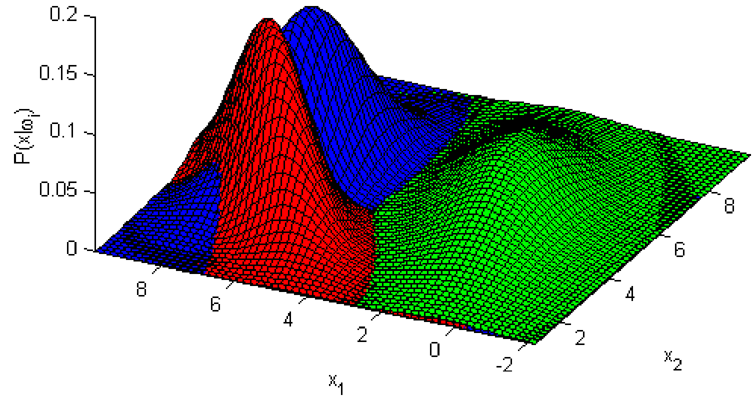
– Notice that the quadratic expression in the discriminant is proportional to the Mahalanobis distance for covariance Σ_i

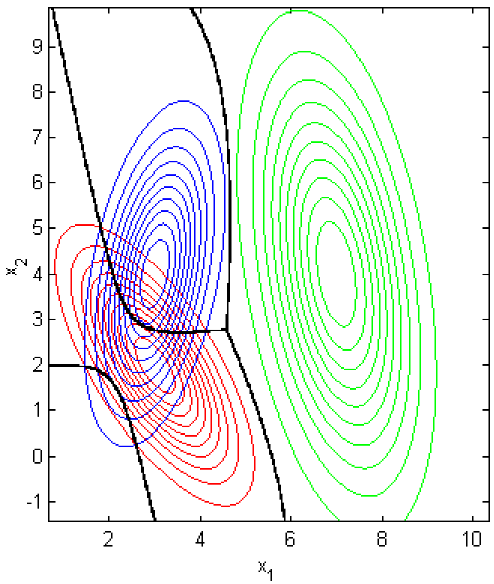
Example

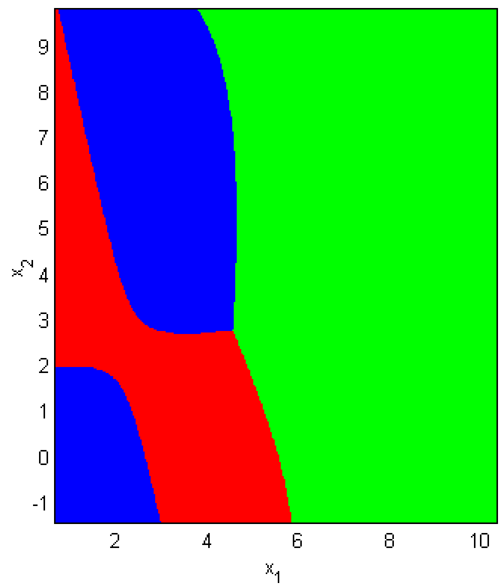
- Three-class 2D problem with equal priors

$$\mu_1 = [3\ 2]^T \qquad \mu_2 = [5\ 4]^T \qquad \mu_3 = [3\ 4]^T$$

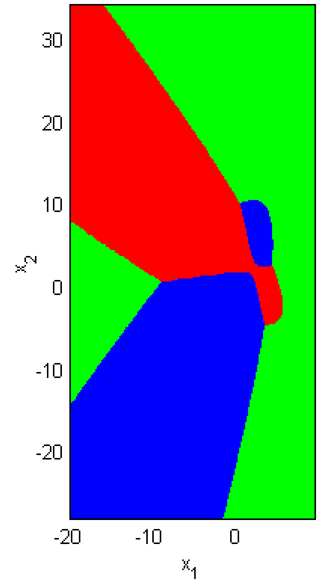
$$\Sigma_1 = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 1 & -1 \\ -1 & 7 \end{bmatrix} \qquad \Sigma_3 = \begin{bmatrix} .5 & .5 \\ .5 & 3 \end{bmatrix}$$







Zoom out



Numerical example

Derive a linear DF for the following 2-class 3D problem

$$\mu_1 = [0\ 0\ 0]^T; \mu_2 = [1\ 1\ 1]^T; \Sigma_1 = \Sigma_2 = \begin{bmatrix} .25 & & \\ & .25 & \\ & & .25 \end{bmatrix}; \mathrm{P}_2 = 2\mathrm{P}_1$$

– Solution

$$g_1(x) = -\frac{1}{2\sigma^2}(x-\mu_1)^T(x-\mu_1) + logP_1 = -\frac{1}{2}\begin{bmatrix} x-0 \\ y-0 \\ z-0 \end{bmatrix}^T \begin{bmatrix} 4 & & \\ & 4 & \\ & & 4 \end{bmatrix} \begin{bmatrix} x-0 \\ y-0 \\ z-0 \end{bmatrix} + log\,\frac{1}{3}$$

$$g_2(x) = -\frac{1}{2}\begin{bmatrix} x-1 \\ y-1 \\ z-1 \end{bmatrix}^T \begin{bmatrix} 4 & & \\ & 4 & \\ & & 4 \end{bmatrix} \begin{bmatrix} x-1 \\ y-1 \\ z-1 \end{bmatrix} + log\,\frac{2}{3}$$

$$g_1(x) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} g_2(x) \Rightarrow -2(x^2+y^2+z^2) + lg\,\frac{1}{3} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} -2\left((x-1)^2+(y-1)^2+(z-1)^2\right) + \mathrm{lg}\,\frac{2}{3}$$

$$\boxed{x+y+z \underset{\omega_1}{\overset{\omega_2}{\gtrless}} \frac{6-log2}{4} = 1.32}$$

– Classify the test example $x_u = [0.1\ 0.7\ 0.8]^T$

$$0.1 + 0.7 + 0.8 = 1.6 \underset{\omega_1}{\overset{\omega_2}{\gtrless}} 1.32 \Rightarrow x_u \in \omega_2$$

Conclusions

The examples in this lecture illustrate the following points

- The Bayes classifier for Gaussian classes (general case) is <u>quadratic</u>
- The Bayes classifier for Gaussian classes with equal covariance is <u>linear</u>
- The Mahalanobis distance classifier is Bayes-optimal for
  - normally distributed classes and
  - equal covariance matrices and
  - equal priors
- The Euclidean distance classifier is Bayes-optimal for
  - normally distributed classes and
  - equal covariance matrices proportional to the identity matrix and
  - equal priors
- Both Euclidean and Mahalanobis distance classifiers are linear classifiers

Thus, some of the simplest and most popular classifiers can be derived from decision-theoretic principles

- Using a specific (Euclidean or Mahalanobis) minimum distance classifier implicitly corresponds to certain statistical assumptions
- The question whether these assumptions hold or don't can rarely be answered in practice; in most cases we can only determine whether the classifier solves our problem