Pattern Recognition and Image Analysis

Dr. Manal Helal – Fall 2014

Lecture 3

BAYES DECISION THEORY

In Action 2

Recap Example

Let blue, green, and red be three classes with prior probabilities given by

$$P(\text{ blue}) = \frac{1}{4} \quad (4.4)$$

$$P(\text{ green}) = \frac{1}{2} \quad (4.5)$$

$$P(\text{ red}) = \frac{1}{4} \quad (4.6)$$

Example (cont.)

These three classes correspond to sets of objects coloured blue, green and red respectively. Let there be three types of objects—"pencils", "pens", and "paper". Let the class-conditional probabilities of these objects be

$$P(\text{ pencil} \mid \text{ green}) = \frac{1}{3}; P(\text{ pen} \mid \text{ green}) = \frac{1}{2}; P(\text{ paper} \mid \text{ green}) = \frac{1}{6} \quad (4.7)$$

$$P(\text{ pencil} \mid \text{ blue}) = \frac{1}{2}; P(\text{ pen} \mid \text{ blue}) = \frac{1}{6}; P(\text{ paper} \mid \text{ blue}) = \frac{1}{3} \quad (4.8)$$

$$P(\text{pencil} \mid \text{red}) = \frac{1}{6}; P(\text{pen} \mid \text{red}) = \frac{1}{3}; P(\text{paper} \mid \text{red}) = \frac{1}{2} \quad (4.9)$$

Example (cont.)

Assign colours to objects.

Example (cont.)

Consider a collection of pencil, pen, and paper with equal probabilities. We can decide the corresponding class labels, using Bayes classifier, as follows:

$$P(\text{green} \mid \text{pencil}) =$$

$$\frac{P(\text{ pencil} \mid \text{green})P(\text{green})}{P(\text{pencil} \mid \text{green})P(\text{green}) + P(\text{pencil} \mid \text{blue})P(\text{blue}) + P(\text{pencil} \mid \text{red})P(\text{red})} \quad (4.10)$$

which is given by

$$P(\text{green} \mid \text{pencil}) =$$

$$\frac{\frac{1}{3}.\frac{1}{2}}{\frac{1}{3}.\frac{1}{2} + \frac{1}{2}.\frac{1}{4} + \frac{1}{6}.\frac{1}{4}} = \frac{1}{2}$$

Example (cont.)

Similarly, it is possible to compute $P(\text{blue} \mid \text{pencil})$ as

$$P(\text{ blue} \mid \text{ pencil}) = \frac{\frac{1}{2} \cdot \frac{1}{4}}{\frac{1}{3} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{4} + \frac{1}{6} \cdot \frac{1}{4}} = \frac{3}{8}$$

$$P(\text{red} \mid \text{pencil}) = \frac{\frac{1}{6} \cdot \frac{1}{4}}{\frac{1}{3} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{4} + \frac{1}{6} \cdot \frac{1}{4}} = \frac{1}{8}$$

Example (cont.)

This would mean that we decide that pencil is a member of class "green" because the posterior probability is $\frac{1}{2}$, which is greater than the posterior probabilities of the other classes ("red" and "blue"). The posterior probabilities for "blue" and "red" classes are $\frac{3}{8}$ and $\frac{1}{8}$ respectively. So, the corresponding probability of error, $P(\text{error} \mid \text{pencil}) = \frac{1}{2}$.

$$P(\text{red} \mid \text{pencil}) = \frac{1}{8}$$

$$P(\text{green} \mid \text{pencil}) = \frac{1}{2}$$

$$P(\text{blue} \mid \text{pencil}) = \frac{3}{8}$$

Example (cont.)

Assign colour to pen objects.

$$P(\text{blue}) = \frac{1}{4}$$
$$P(\text{green}) = \frac{1}{2}$$
$$P(\text{red}) = \frac{1}{4}$$

$$P(\text{pencil} \mid \text{green}) = \frac{1}{3}; P(\text{pen} \mid \text{green}) = \frac{1}{2}; P(\text{paper} \mid \text{green}) = \frac{1}{6}$$

$$P(\text{pencil} \mid \text{blue}) = \frac{1}{2}; P(\text{pen} \mid \text{blue}) = \frac{1}{6}; P(\text{paper} \mid \text{blue}) = \frac{1}{3}$$

$$P(\text{pencil} \mid \text{red}) = \frac{1}{6}; P(\text{pen} \mid \text{red}) = \frac{1}{3}; P(\text{paper} \mid \text{red}) = \frac{1}{2}$$

Example (cont.)

In a similar manner, for pen, the posterior probabilities are

$$P(\text{green} \mid \text{pen}) = \frac{2}{3}; P(\text{blue} \mid \text{pen}) = \frac{1}{9}; P(\text{red} \mid \text{pen}) = \frac{2}{9} \tag{4.14}$$

This enables us to decide that pen belongs to class "green" and $P(\text{error} \mid \text{pen}) = \frac{1}{3}$.

Example (cont.)

Assign colour to paper objects.

$$P(\text{blue}) = \frac{1}{4}$$
$$P(\text{green}) = \frac{1}{2}$$
$$P(\text{red}) = \frac{1}{4}$$

$$P(\text{pencil} \mid \text{green}) = \frac{1}{3};\ P(\text{pen} \mid \text{green}) = \frac{1}{2};\ P(\text{paper} \mid \text{green}) = \frac{1}{6}$$

$$P(\text{pencil} \mid \text{blue}) = \frac{1}{2};\ P(\text{pen} \mid \text{blue}) = \frac{1}{6};\ P(\text{paper} \mid \text{blue}) = \frac{1}{3}$$

$$P(\text{pencil} \mid \text{red}) = \frac{1}{6};\ P(\text{pen} \mid \text{red}) = \frac{1}{3};\ P(\text{paper} \mid \text{red}) = \frac{1}{2}$$

Example (cont.)

Finally, for paper, the posterior probabilities are

$$P(\text{green} \mid \text{paper}) = \frac{2}{7};\ P(\text{blue} \mid \text{paper}) = \frac{2}{7};\ P(\text{red} \mid \text{paper}) = \frac{3}{7} \tag{4.15}$$

Based on these probabilities, we decide to assign paper to "red" which has the maximum posterior probability.

So, $P(\text{error} \mid \text{paper}) = \frac{4}{7}$

Example (cont.)

Average probability of error =

$$P(\text{error} \mid \text{pencil}) \times \frac{1}{3} + P(\text{error} \mid \text{pen}) \times \frac{1}{3} + P(\text{error} \mid \text{paper}) \times \frac{1}{3} \tag{4.16}$$

As a consequence, its value is

$$\text{Average probability of error} = \frac{1}{2}.\frac{1}{3} + \frac{1}{3}.\frac{1}{3} + \frac{1}{3}.\frac{4}{7} = \frac{59}{126} \tag{4.17}$$

Solving Posteriors (Deterministic)

Logistic Regression

Logistic Regression

Logistic Regression is a discriminative model, because it models the posterior probabilities $p(y|\boldsymbol{x})$ directly.

Posteriors and the Logistic Function

For two classes $y \in \{0, 1\}$ we get:

$$
\begin{aligned}
p(y = 0|\boldsymbol{x}) &= \frac{p(y = 0) \cdot p(\boldsymbol{x}|y = 0)}{p(\boldsymbol{x})} \\
&= \frac{p(y = 0) \cdot p(\boldsymbol{x}|y = 0)}{p(y = 0)p(\boldsymbol{x}|y = 0) + p(y = 1)p(\boldsymbol{x}|y = 1)} \\
&= \frac{1}{1 + \frac{p(y=1)p(\boldsymbol{x}|y=1)}{p(y=0)p(\boldsymbol{x}|y=0)}}
\end{aligned}
$$

Posteriors and the Logistic Function (cont.)

$$p(y=0|\boldsymbol{x}) = \frac{1}{1+\frac{p(y=1)p(\boldsymbol{x}|y=1)}{p(y=0)p(\boldsymbol{x}|y=0)}}$$

(Trick: extend with exponential and logarithm)

$$= \frac{1}{1+e^{\log\frac{p(y=1)p(\boldsymbol{x}|y=1)}{p(y=0)p(\boldsymbol{x}|y=0)}}}$$

$$= \frac{1}{1+e^{-\log\frac{p(y=0)}{p(y=1)}-\log\frac{p(\boldsymbol{x}|y=0)}{p(\boldsymbol{x}|y=1)}}}$$

Posteriors and the Logistic Function (cont.)

We see that the posterior for class $y = 0$ can be written in terms of a logistic function:

$$p(y = 0|\boldsymbol{x}) = \frac{1}{1 + e^{-F(\boldsymbol{x})}}$$

And thus the posterior for the other class $y = 1$:

$$\begin{aligned} p(y = 1|\boldsymbol{x}) &= 1 - p(y = 0|\boldsymbol{x}) \\ &= \frac{e^{-F(\boldsymbol{x})}}{1 + e^{-F(\boldsymbol{x})}} \\ &= \frac{1}{1 + e^{F(\boldsymbol{x})}} \end{aligned}$$

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

Posteriors and the Logistic Function (cont.)

Definition

The *logistic function* (also called *sigmoid function*) is defined by

$$g(x) = \frac{1}{1 + e^{-x}}$$

where $x \in \mathbb{R}$.

Posteriors and the Logistic Function (cont.)

The derivative of the sigmoid function fulfills the nice property:

$$
\begin{aligned}
g'(x) &= \left(\frac{1}{1+e^{-x}}\right)' = \frac{1}{(1+e^{-x})^2} \cdot e^{-x} \\
&= \frac{1}{(1+e^{-x})} \cdot \frac{e^{-x}}{(1+e^{-x})} \\
&= \frac{1}{(1+e^{-x})} \cdot \frac{1}{(1+e^{x})} \\
&= g(x)g(-x) \\
&= g(x)(1-g(x)) \quad .
\end{aligned}
$$

Posteriors and the Logistic Function (cont.)

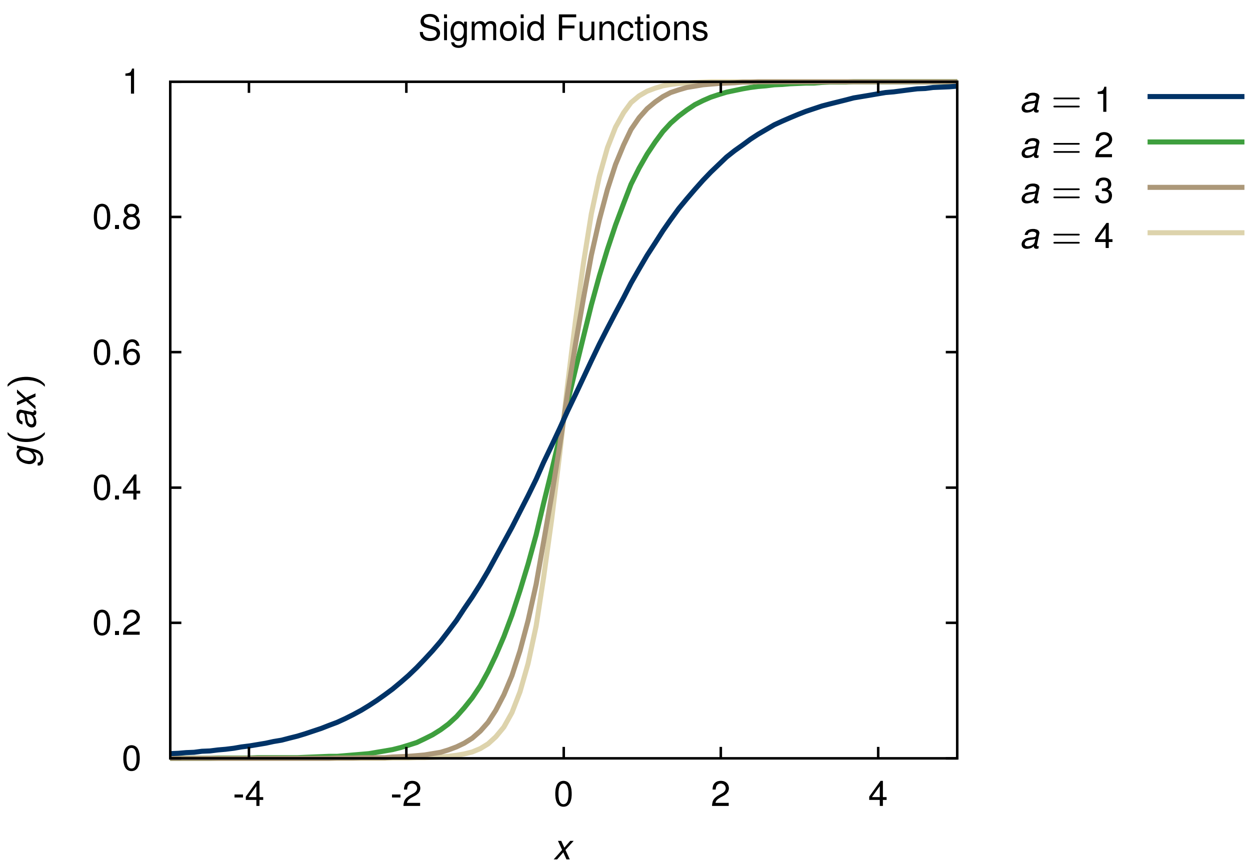


Fig.: Sigmoid function: $g(ax) = 1/(1 + e^{-ax})$ for $a = 1, 2, 3, 4$

Decision Boundary

The decision boundary $\delta(\boldsymbol{x}) = 0$ (zero level set) in feature space separates the two classes.

Points $\boldsymbol{x}$ on the decision boundary satisfy:

$$p(y = 0|\boldsymbol{x}) \quad = \quad p(y = 1|\boldsymbol{x})$$

and thus

$$\log \frac{p(y = 0|\boldsymbol{x})}{p(y = 1|\boldsymbol{x})} \quad = \quad \log 1 = 0 \quad .$$

Decision Boundary (cont.)

Lemma

The decision boundary is given by $F(\boldsymbol{x}) = 0$.

Proof:

$$\log \frac{p(y=0|\boldsymbol{x})}{p(y=1|\boldsymbol{x})} = F(\boldsymbol{x}) = 0$$

$$\frac{p(y=0|\boldsymbol{x})}{p(y=1|\boldsymbol{x})} = e^{F(\boldsymbol{x})}$$

$$p(y=0|\boldsymbol{x}) = e^{F(\boldsymbol{x})} p(y=1|\boldsymbol{x})$$

Decision Boundary (cont.)

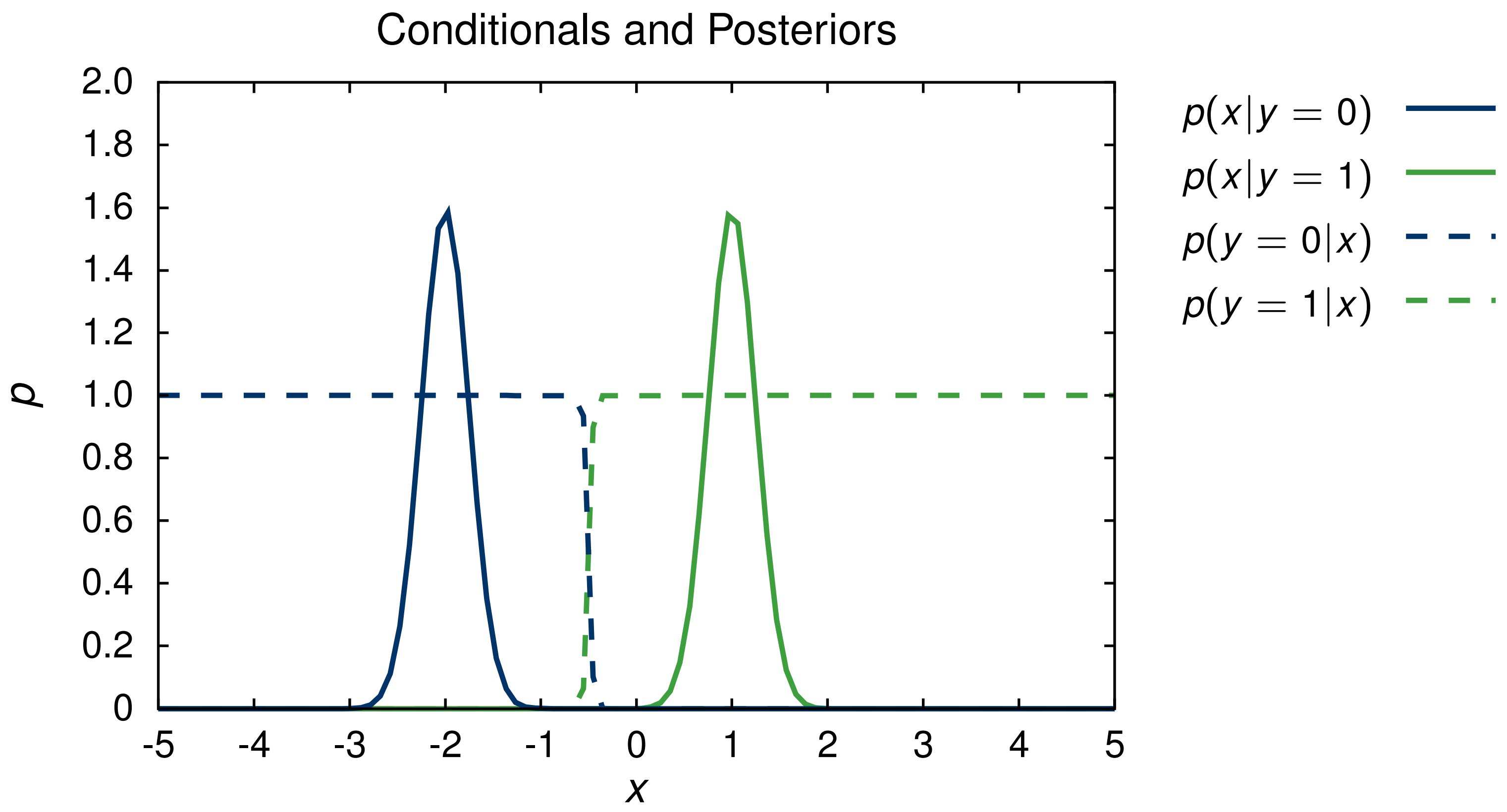


Fig.: Two Gaussians and its posteriors: $\sigma_0 = \sigma_1 = 0.25$, $\mu_0 = -2$, $\mu_1 = 1$

Decision Boundary (cont.)

Example

Let us assume both classes have normally distributed d-dimensional feature vectors:

$$p(\boldsymbol{x}|y) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma}_y)}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_y)^T\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_y)}$$

Then we can write the posterior of $y = 0$ in terms of a logistic function:

$$p(y=0|\boldsymbol{x}) = \frac{1}{1+e^{-F(\boldsymbol{x})}} = \frac{1}{1+e^{-(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}+\boldsymbol{\alpha}^T\boldsymbol{x}+\alpha_0)}}$$

$$F(\boldsymbol{x}) = \log\frac{p(y=0|\boldsymbol{x})}{p(y=1|\boldsymbol{x})} = \log\frac{p(y=0)p(\boldsymbol{x}|y=0)}{p(y=1)p(\boldsymbol{x}|y=1)}$$

Decision Boundary (cont.)

Example (cont.)

$$F(\boldsymbol{x}) = \log \frac{p(y=0)}{p(y=1)} + \log \frac{\frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma}_0)}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)}}{\frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma}_1)}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}}$$

This function has the constant component:

$$c = \log \frac{p(y=0)}{p(y=1)} + \frac{1}{2} \log \frac{\det(2\pi\boldsymbol{\Sigma}_1)}{\det(2\pi\boldsymbol{\Sigma}_0)}$$

We observe:

- Priors imply a constant offset of the decision boundary.
- If priors and covariance matrices of both classes are identical, this offset is $c = 0$.

Decision Boundary (cont.)

Example (cont.)

$$F(\boldsymbol{x}) \quad = \quad \log \frac{p(y=0)}{p(y=1)} + \log \frac{\frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma}_0)}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)}}{\frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma}_1)}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}}$$

This function has the constant component:

$$c \quad = \quad \log \frac{p(y=0)}{p(y=1)} + \frac{1}{2} \log \frac{\det(2\pi\boldsymbol{\Sigma}_1)}{\det(2\pi\boldsymbol{\Sigma}_0)}$$

We observe:

- Priors imply a constant offset of the decision boundary.
- If priors and covariance matrices of both classes are identical, this offset is $c = 0$.

Decision Boundary (cont.)

Example (cont.)

Furthermore we have:

$$\log \frac{e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)}}{e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}} =$$

$$= \frac{1}{2}\left((\boldsymbol{x}-\boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) - (\boldsymbol{x}-\boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)\right)$$

$$= \frac{1}{2}\Big(\boldsymbol{x}^T(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_0^{-1})\boldsymbol{x} - 2(\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_0^T \boldsymbol{\Sigma}_0^{-1})\boldsymbol{x} +$$

$$+ \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^T \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0\Big)$$

Decision Boundary (cont.)

Example (cont.)

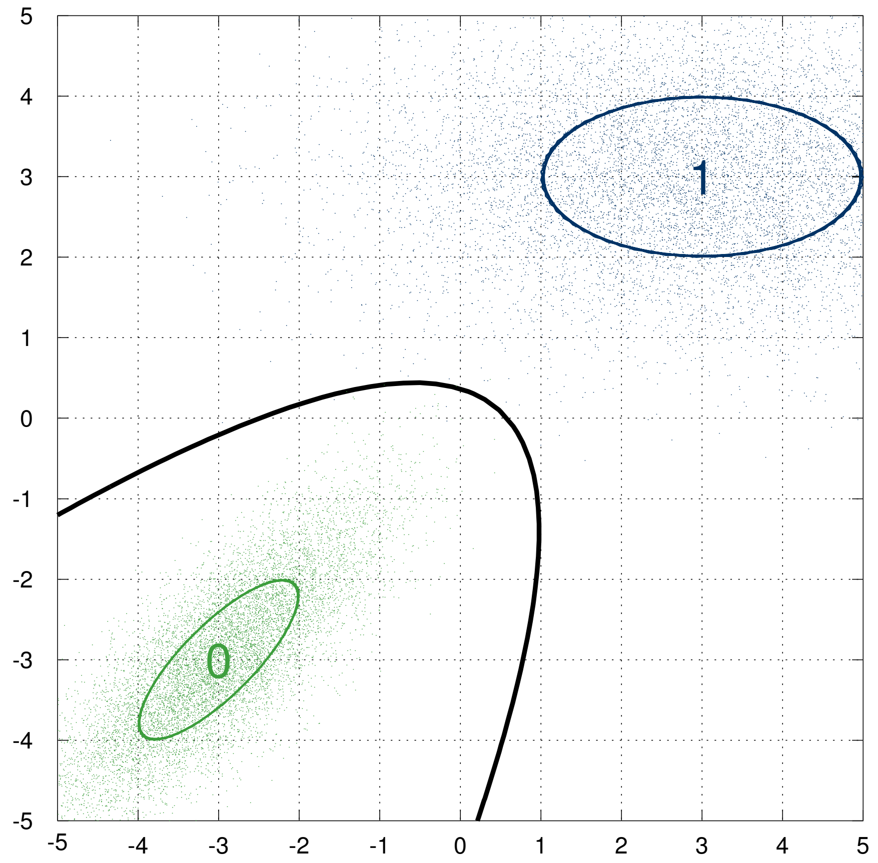
Now we have:

$$\boldsymbol{A} = \frac{1}{2}(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_0^{-1})$$

$$\boldsymbol{\alpha}^T = \boldsymbol{\mu}_0^T \boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_1^{-1}$$

$$\alpha_0 = \log \frac{p(y=0)}{p(y=1)} + \frac{1}{2}\left(\log \frac{\det(2\pi\boldsymbol{\Sigma}_1)}{\det(2\pi\boldsymbol{\Sigma}_0)} + \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^T \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \right)$$

Decision Boundary (cont.)

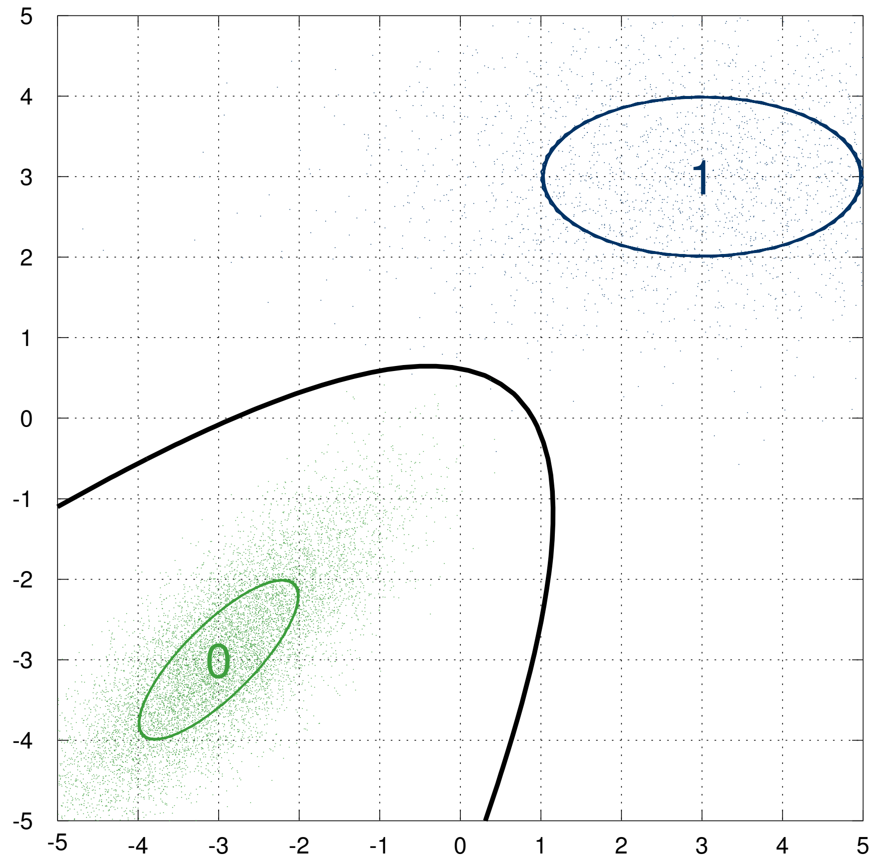


$p(y = 0) = 0.5$

$p(y = 1) = 0.5$

Fig.: Two Gaussian sample sets and the decision boundary

Decision Boundary (cont.)

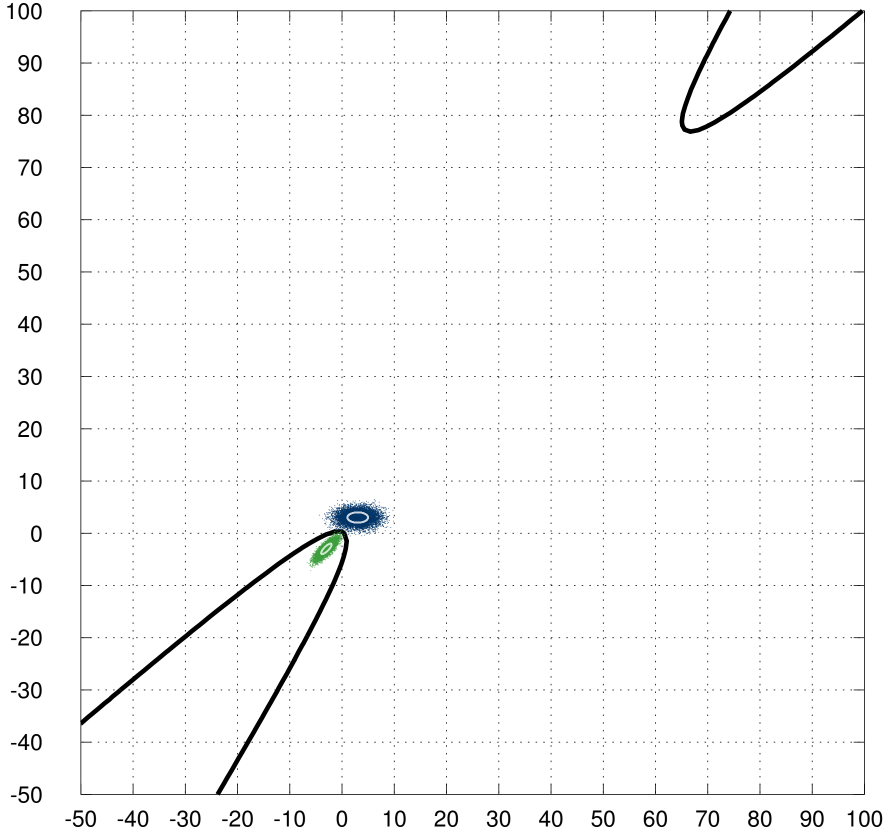


$p(y = 0) = 0.8$

$p(y = 1) = 0.2$

Fig.: Two Gaussian sample sets and the decision boundary

Decision Boundary (cont.)



$p(y = 0) = 0.5$

$p(y = 1) = 0.5$

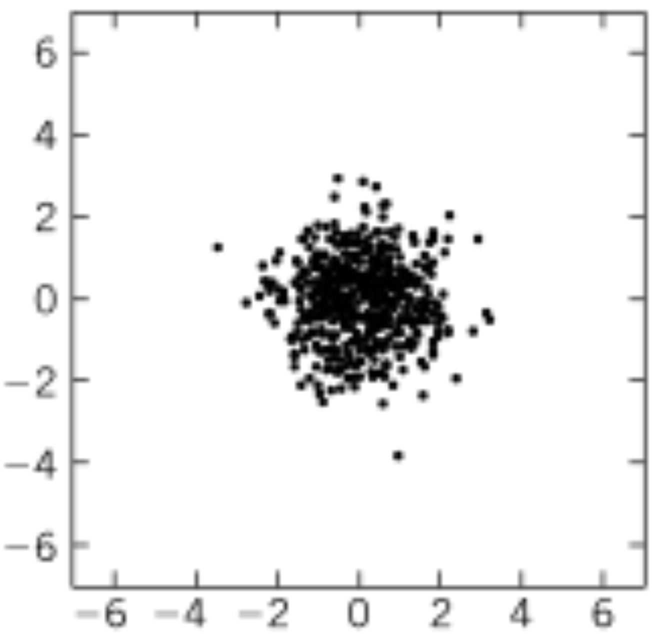
Fig.: Two Gaussian sample sets and the decision boundary

Matlab Exercise

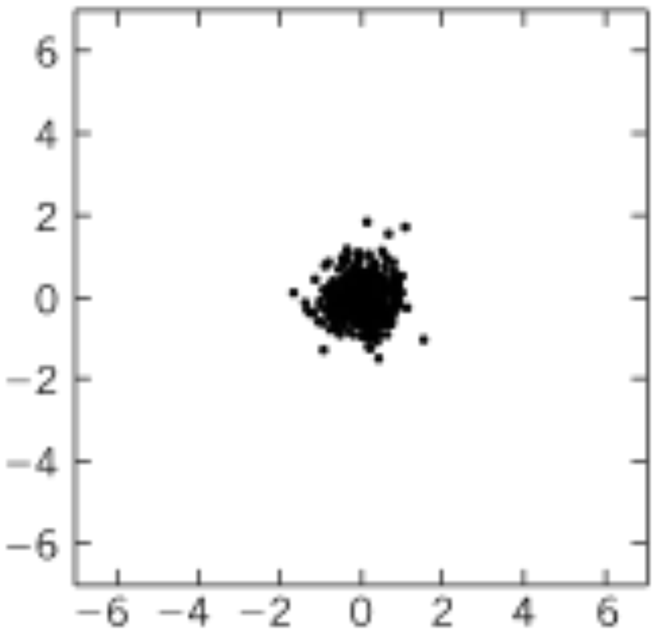
Gaussian Distributions for $S = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$ and $m = [0, 0]^T$

Spherically Shaped Data:

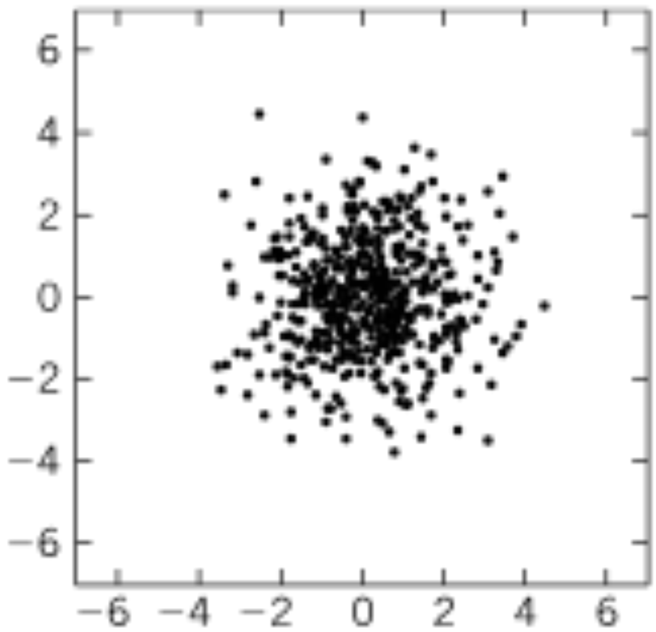
When the two coordinates of *x* are uncorrelated ($\sigma_{12} = 0$) and their variances are equal,



$\sigma_1^2 = \sigma_2^2 = 1, \sigma_{12} = 0$



$\sigma_1^2 = \sigma_2^2 = 0.2, \sigma_{12} = 0$



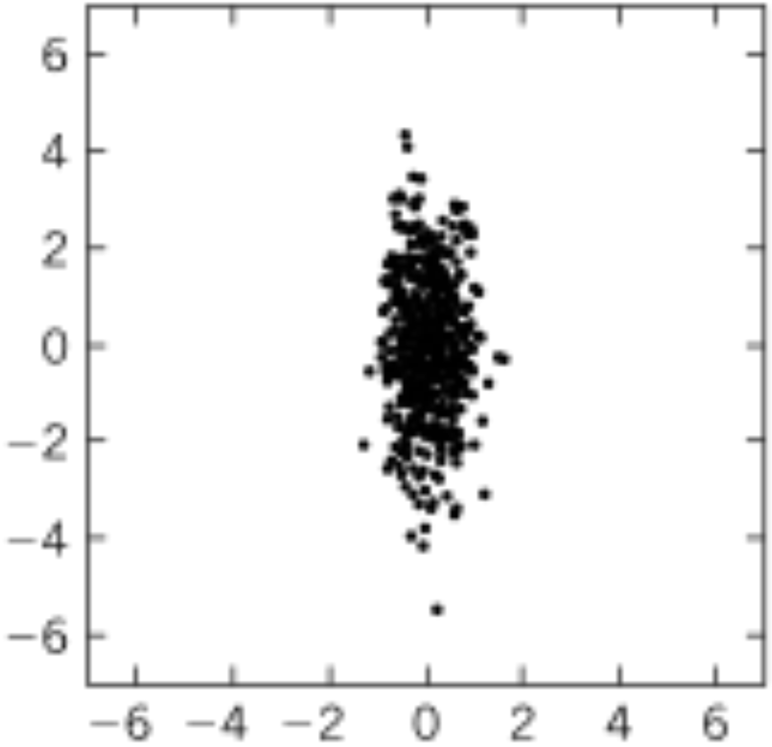
$\sigma_1^2 = \sigma_2^2 = 2, \sigma_{12} = 0$

Run Example 1.3.3

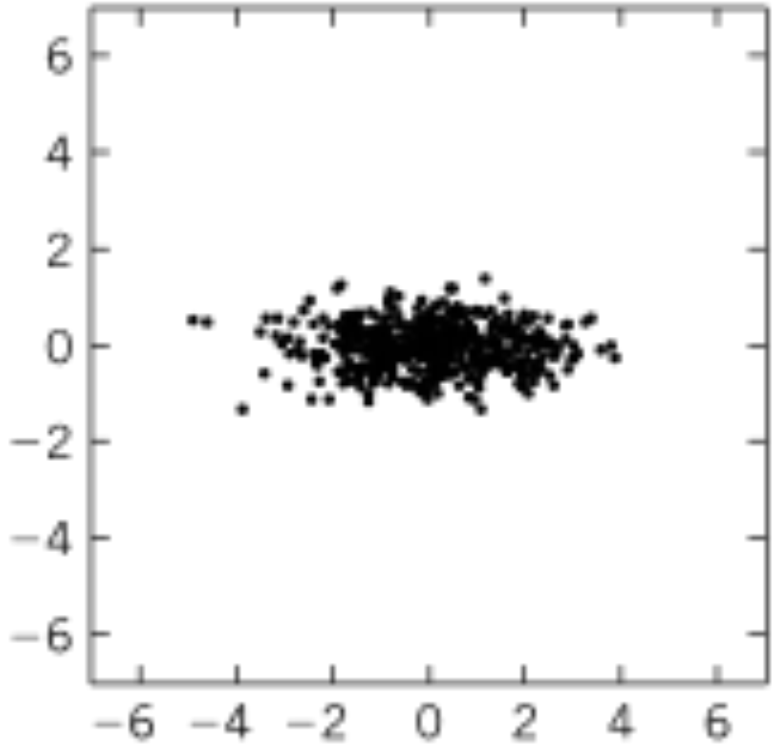
Gaussian Distributions for $S = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$ and $m = [0, 0]^T$

Ellipsoidally Shaped Data:

When the two coordinates of x are uncorrelated ($\sigma_{12} = 0$) and their variances are unequal,



$\sigma_1^2 = 0.2, \sigma_2^2 = 2, \sigma_{12} = 0$



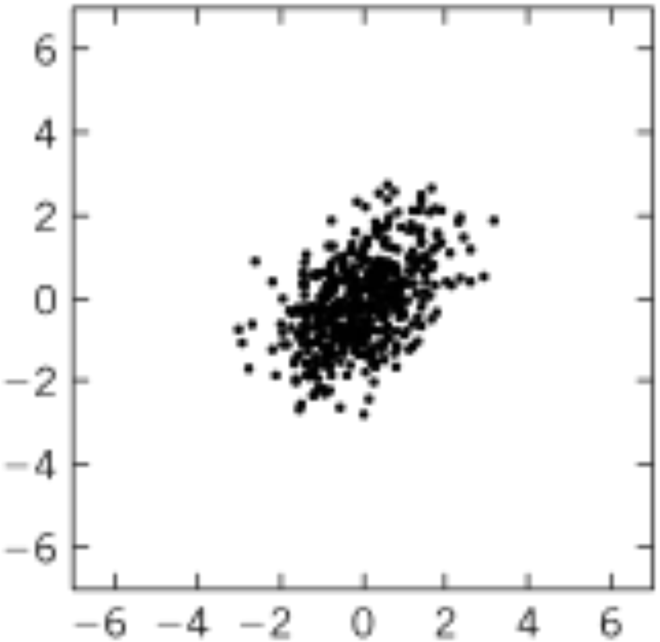
$\sigma_1^2 = 2, \sigma_2^2 = 0.2, \sigma_{12} = 0$

Run Example 1.3.3

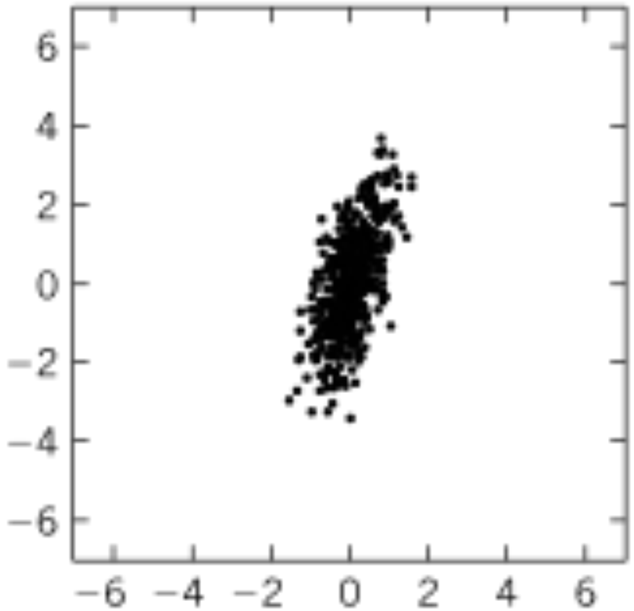
Gaussian Distributions for $S = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$ and $m = [0, 0]^T$

Spherically Shaped Data clustered unparalleled to the axes:

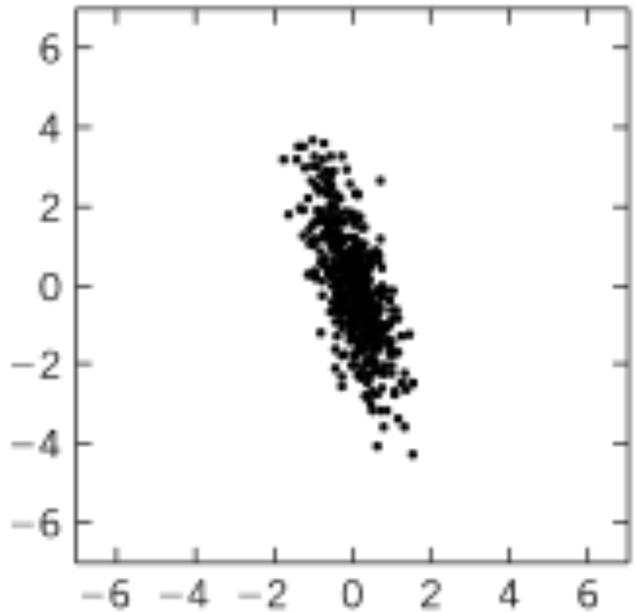
When the two coordinates of x are correlated ($\sigma_{12} \neq 0$) , The degree of rotation with respect to the axes depends on the value of σ_{12} ,



$\sigma_1^2 = \sigma_2^2 = 1, \sigma_{12} = 0.5$



$\sigma_1^2 = 0.3, \sigma_2^2 = 2, \sigma_{12} = 0.5$



$\sigma_1^2 = 0.3, \sigma_2^2 = 2, \sigma_{12} = -0.5$

Run Example 1.3.3

MINIMUM DISTANCE CLASSIFIERS

- The Euclidean Distance Classifier is the optimal Bayesian Classifier when:
  - The optimal Bayesian classifier is significantly simplified under the following assumptions:
  - The classes are equiprobable.
  - The data in all classes follow Gaussian distributions.
  - The covariance matrix is the same for all classes.
  - The covariance matrix is diagonal and all elements across the diagonal are equal. That is, $S = \sigma^2 I$, where I is the identity matrix.

$$||x - m_i|| \equiv \sqrt{(x - m_i)^T (x - m_i)} < ||x - m_j||, \quad \forall i \neq j$$

MINIMUM DISTANCE CLASSIFIERS

- The Mahalanobis Distance Classifier is the optimal Bayesian Classifier when the covairance matrix is not diagonal with equal elements:
  - The optimal Bayesian classifier is significantly simplified under the following assumptions:
  - The classes are equiprobable.
  - The data in all classes follow Gaussian distributions.
  - The covariance matrix is the same for all classes.

$$\sqrt{(x-m_i)^T S^{-1}(x-m_i)} < \sqrt{(x-m_j)^T S^{-1}(x-m_j)}, \quad \forall j \neq i$$

Run Example 1.4.1

Maximum Likelihood Parameter Estimation of Gaussian pdfs

- The maximum likelihood (ML) is a popular method for the estimation of an unknown mean value and the associated covariance matrix of a known pdf.

- Given N points, $x_i \in R^l$, $i = 1,2,\ldots,N$, which are known to be normally distributed, the ML estimates of the unknown mean value and the associated covariance matrix are given by:

$$m_{ML} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

and

$$S_{ML} = \frac{1}{N} \sum_{i=1}^{N} (x_i - m_{ML})(x_i - m_{ML})^T$$

Run Example 1.4.2

Practical Labs

- On Moodle you will find two Bayesian Classification examples:
  - Image Classification
  - Text Classification

Comprehensive Questions

- How can we model the posterior probabilities?
- Formulate the criterion for the decision boundary!
- Describe the shape of the decision boundary for a Gaussian with different and same class covariances!
- What effect does a change of the priors have on the decision boundary?