Statistics and Sensor Data Fusion

3. Pattern Recognition and Bayes Optimal Classifier

Pattern Recognition and Bayes Optimal Classifier

Statistical Pattern Recognition:

Statistical pattern recognition is concerned with the **classification of objects** on the basis of **quantitative features**.

The term pattern indicates that the classification is facilitated by **characteristic structures** of the objects under consideration:

> "One of the most interesting aspects of the world is that it can be considered to be made up of **patterns**. A pattern is essentially an **arrangement**. It is characterized by the **order of the elements** of which it is made rather than by the intrinsic nature of these elements."
>
> *Norbert Wiener*

Pattern Recognition and Bayes Optimal Classifier

Mathematical Description of Patterns:

- Objects or patterns to be classified are described by features which constitute the different dimensions of the feature space.
- Within the framework of statistical pattern recognition, features are interpreted as either **discrete** or **continuous random variables** which are combined in feature vectors.
- The observed feature values of an object to be classified are interpreted as **realization** of the feature vector.
- Based on this realization of the feature vector, the object is classified, i.e. a **decision** is made about the underlying class.

Pattern Recognition and Bayes Optimal Classifier

Feature Extraction and Feature Selection:

- The term feature extraction refers to the **definition** and **computation** of suitable features, which is problem-specific and depends on the objects to be classified:
  - Signals: Spectral analysis, Fourier transform, ...
  - Images: Segmentation, shape features, 2-dim. Fourier transform, ...
- The term feature selection refers to the choice of a **maximally informative subset** of features with respect to a specific classification task.

  In the case of feature selection, the set of all features to select from has to be determined already.

Pattern Recognition and Bayes Optimal Classifier

Example of a Classification Problem:

In a fish cannery, the distinction between salmon and sea bass should be done automatically on the basis of **grayscale images**.

Therefore, the following **processing steps** are necessary:

1. Camera shot (original image)
2. Preprocessing of the image (noise filtering, calculation of grayscale values, segmentation, ...)
3. Feature extraction and selection (e.g. lightness and width)
4. Classification

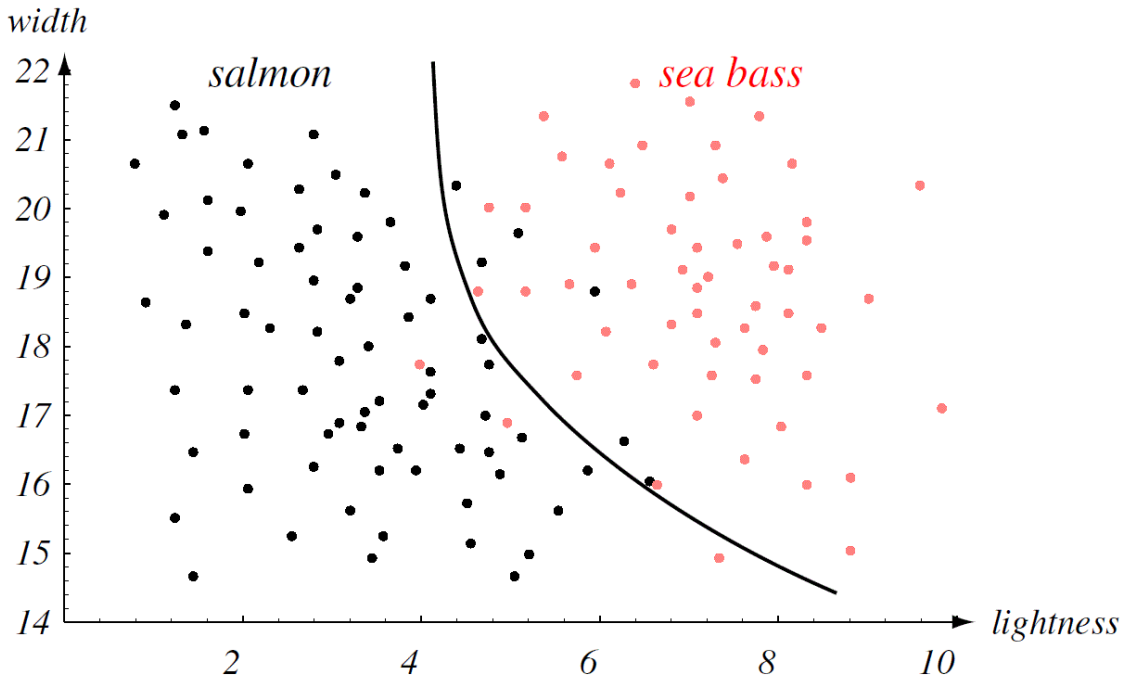
Pattern Recognition and Bayes Optimal Classifier

Possible Original Images:

https://www.indiamart.com/proddetail/atlantic-salmon-fish-21055714788.html
https://www.aoseafood.co.uk/buy/sea-bass/

Pattern Recognition and Bayes Optimal Classifier

Possible Classification:

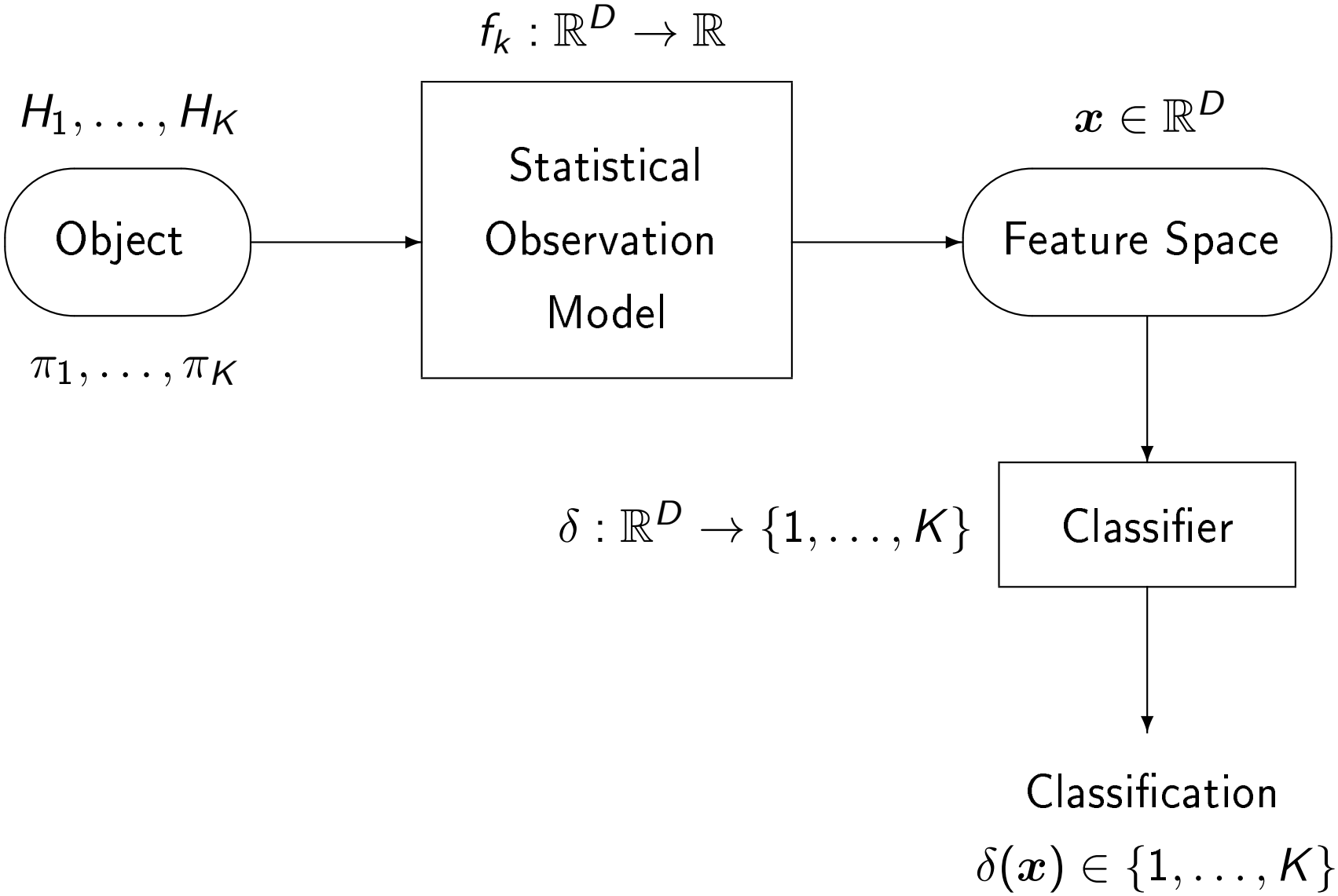


Based on the computed values of the two features **lightness** and **width**, a binary valued classifier might reach a decision on the type of the fish according to the plotted black line. Observe that the classifier will not be error-free in general.

R. Duda, P. Hart, D. Stork, Pattern Classification, 2nd edition, Wiley, 2001

Pattern Recognition and Bayes Optimal Classifier

Statistical Model of Classification:



Pattern Recognition and Bayes Optimal Classifier

Statistical Model of Classification:

1. An object belongs to one out of K classes represented by the **hypotheses** $H_1, \ldots, H_K$, which are characterized by the prior probabilities $\pi_1, \ldots, \pi_K$.

2. The true class label cannot be observed directly, but instead a **feature vector** $\boldsymbol{x} \in \mathbb{R}^D$ is measured, which is interpreted as a realization of the random vector $\boldsymbol{X}$.

3. For an object from class k, the random vector $\boldsymbol{X}$ is distributed according to the class-conditional pdf $f_k : \mathbb{R}^D \to \mathbb{R}$.

4. A **classifier** $\delta : \mathbb{R}^D \to \{1, \ldots, K\}$, $\boldsymbol{x} \mapsto \delta(\boldsymbol{x})$, estimates the true class on the basis of the observed feature vector $\boldsymbol{x} \in \mathbb{R}^D$.

5. The goal is to construct an **optimal classifier** δ^* which minimizes the probability of error.

Pattern Recognition and Bayes Optimal Classifier

Classifier and Decision Regions:

A classifier

$$\delta : \mathbb{R}^D \to \{1, \ldots, K\}, \quad \boldsymbol{x} \mapsto \delta(\boldsymbol{x})$$

corresponds to a **partition** of the feature space $\mathbb{R}^D$ into mutually exclusive subsets or decision regions $\mathcal{R}_k \subseteq \mathbb{R}^D$, where

$$\bigcup_{k=1}^{K} \mathcal{R}_k = \mathbb{R}^D, \quad \mathcal{R}_i \cap \mathcal{R}_j = \varnothing \quad \text{for } i \neq j$$

and

$$\delta(\boldsymbol{x}) = \begin{cases} 1, \text{ if } \boldsymbol{x} \in \mathcal{R}_1 \\ 2, \text{ if } \boldsymbol{x} \in \mathcal{R}_2 \\ \vdots \\ K, \text{ if } \boldsymbol{x} \in \mathcal{R}_K \end{cases}$$

Pattern Recognition and Bayes Optimal Classifier

Classifier and Decision Regions:

A classifier $\delta : \mathbb{R}^D \to \{1, \ldots, K\}$ corresponds to a partition of the feature space $\mathbb{R}^D$ into decision regions $\mathcal{R}_k \subseteq \mathbb{R}^D$, where

$$\mathcal{R}_k = \{\boldsymbol{x} \in \mathbb{R}^D \mid \delta(\boldsymbol{x}) = k\}, \quad k = 1, \ldots, K$$

Error-free classification is only possible, if the sets

$$\mathcal{S}_k = \{\boldsymbol{x} \in \mathbb{R}^D \mid f_k(\boldsymbol{x}) > 0\}, \quad k = 1, \ldots, K$$

are mutually exclusive, that means if the class-conditional densities $f_k(\boldsymbol{x})$ **do not overlap**.

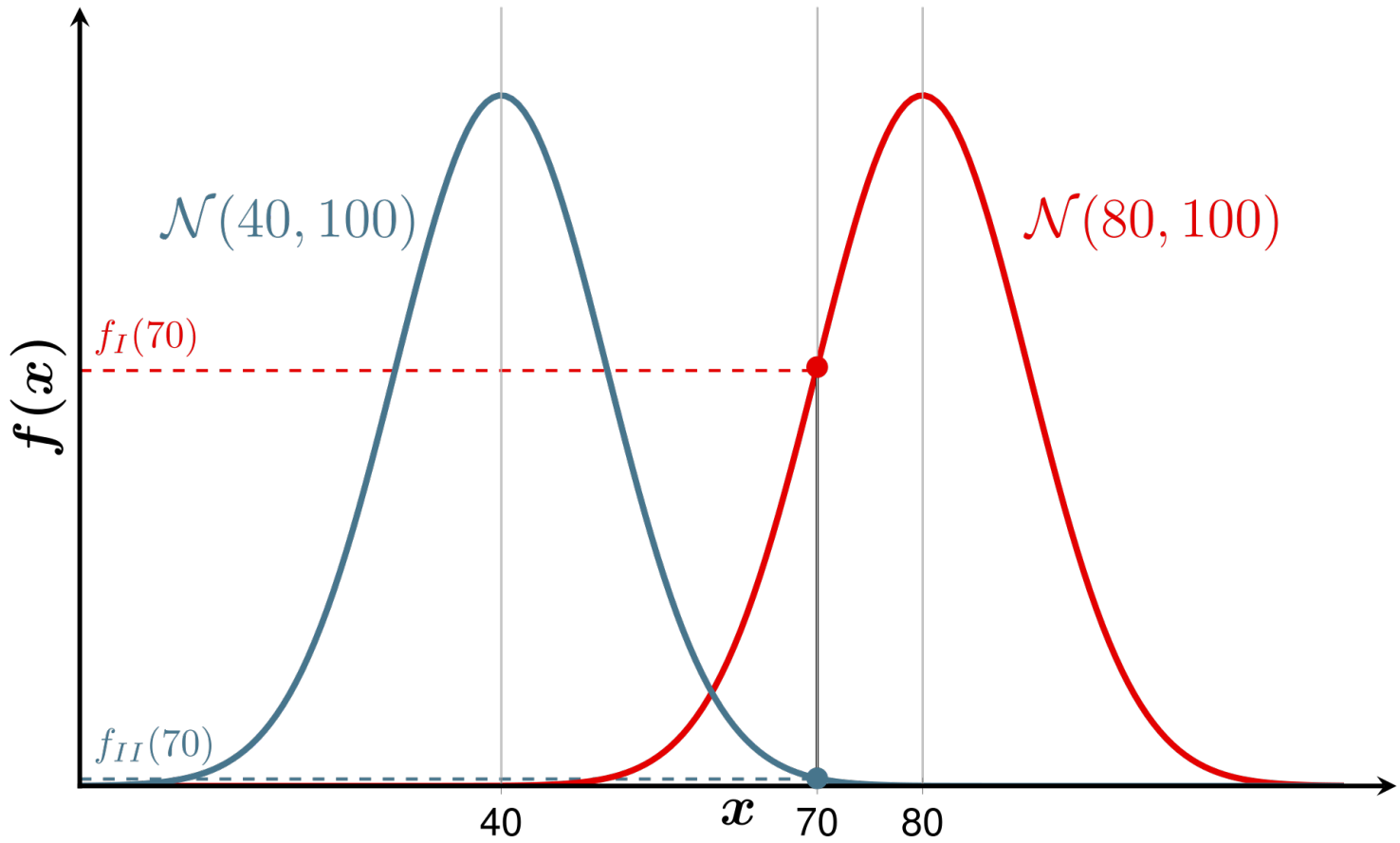
In this case, an error-free classifier is obtained via $\mathcal{R}_k = \mathcal{S}_k$, i.e.

$$\delta(\boldsymbol{x}) = k \iff \boldsymbol{x} \in \mathcal{S}_k$$

In general however, a classifier δ is not error-free.

Pattern Recognition and Bayes Optimal Classifier

Example: Overlapping Densities



Von MartinThoma - Eigenes Werk, CC0, https://commons.wikimedia.org/w/index.php?curid=34672879

Pattern Recognition and Bayes Optimal Classifier

Summary of the Different Probability Distributions:

1. Prior probabilities of the classes: $\pi_1, \ldots, \pi_K$
2. Class-conditional densities of the feature vectors: $f_k(\boldsymbol{x})$
3. Joint distribution over pairs: $f(\boldsymbol{x}, k) = \pi_k \cdot f_k(\boldsymbol{x})$

In the sequel, it is assumed that these probability distributions are **completely known** and can be used for the construction of the optimal classifier δ^*.

In Chapter 4, it will be shown how to estimate the parameters of the involved probability distributions based on a **data sample** or **training set**.

Pattern Recognition and Bayes Optimal Classifier

Derived Probability Distributions:

From the joint distribution $f(\boldsymbol{x}, k)$, the marginal distribution of the feature vectors (independent of the class) is obtained as

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} f(\boldsymbol{x}, k) = \sum_{k=1}^{K} \pi_k \cdot f_k(\boldsymbol{x})$$

Accordingly, the posterior distribution of the classes is given by the **conditional probabilities**

$$p(k\,|\,\boldsymbol{x}) = \frac{f(\boldsymbol{x}, k)}{f(\boldsymbol{x})} = \frac{\pi_k \cdot f_k(\boldsymbol{x})}{\sum_{j=1}^{K} \pi_j \cdot f_j(\boldsymbol{x})}$$

The posterior distribution is **normalized**, i.e.

$$\sum_{k=1}^{K} p(k\,|\,\boldsymbol{x}) = 1$$

Pattern Recognition and Bayes Optimal Classifier

Bayes Optimal Classifier:

The classifier δ^* which minimizes the **probability of error** is called Bayes optimal classifier and can be obtained as

$$\delta^*(\boldsymbol{x}) = \underset{k=1,\ldots,K}{\operatorname{argmax}}\ p(k\,|\,\boldsymbol{x}) = \underset{k=1,\ldots,K}{\operatorname{argmax}}\ \frac{\pi_k \cdot f_k(\boldsymbol{x})}{\sum_{j=1}^{K} \pi_j \cdot f_j(\boldsymbol{x})}$$

The Bayes optimal classifier maximizes the **posterior probability** of the class k given the realization of the feature vector $\boldsymbol{x} \in \mathbb{R}^D$.

However, for its implementation we need to know explicitly both

- the prior probabilities $\pi_1, \ldots, \pi_K$
- the class-conditional densities $f_1(\boldsymbol{x}), \ldots, f_K(\boldsymbol{x})$

Pattern Recognition and Bayes Optimal Classifier

Classifiers and Discriminant Functions:

In many cases, a classifier $\delta = \delta(\boldsymbol{x})$ is expressed in terms of a so-called discriminant function $g = g(\boldsymbol{x}, k)$:

$$\delta(\boldsymbol{x}) = \underset{k=1,\ldots,K}{\text{argmax}}\ g(\boldsymbol{x}, k)$$

For example, the **Bayes optimal classifier** can be expressed in terms of a discriminant function which is obtained via suitable transformations of the **posterior probability** $p(k|\boldsymbol{x})$:

1. $g(\boldsymbol{x}, k) = f(\boldsymbol{x}) \cdot p(k|\boldsymbol{x}) = f(\boldsymbol{x}, k) = \pi_k \cdot f_k(\boldsymbol{x})$
2. $g(\boldsymbol{x}, k) = \log(f(\boldsymbol{x}, k)) = \log(\pi_k) + \log(f_k(\boldsymbol{x}))$
3. $g(\boldsymbol{x}, k) = \log(p(k|\boldsymbol{x})) = \log(\pi_k \cdot f_k(\boldsymbol{x})) - \log\left(\sum_{j=1}^{K} \pi_j \cdot f_j(\boldsymbol{x})\right)$

Pattern Recognition and Bayes Optimal Classifier

Example: Multivariate Normal Distributions $\mathcal{N}_D(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

The classes should be characterized by the **prior probabilities** $\pi_1, \ldots, \pi_K$, the feature vectors should be characterized by the **class-conditional densities**

$$f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp\Big(-\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k) \Big)$$

where $\boldsymbol{\mu}_k \in \mathbb{R}^D$ and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{D \times D}$ symmetric and positive definite for $k = 1, \ldots, K$.

For the sequel, it is assumed that the prior probabilities $\pi_1, \ldots, \pi_K$, the mean vectors $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K \in \mathbb{R}^D$ and the covariance matrices $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K \in \mathbb{R}^{D \times D}$ are completely known.

Pattern Recognition and Bayes Optimal Classifier

Calculation of the Discriminant Function:

For the discriminant function $g(\boldsymbol{x}, k)$, we choose

$$g(\boldsymbol{x}, k) = \log(\pi_k \cdot f_k(\boldsymbol{x})) = \log(\pi_k) + \log(f_k(\boldsymbol{x}))$$

$$= \log(\pi_k) - \tfrac{1}{2}\log((2\pi)^D|\boldsymbol{\Sigma}_k|) - \tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)$$

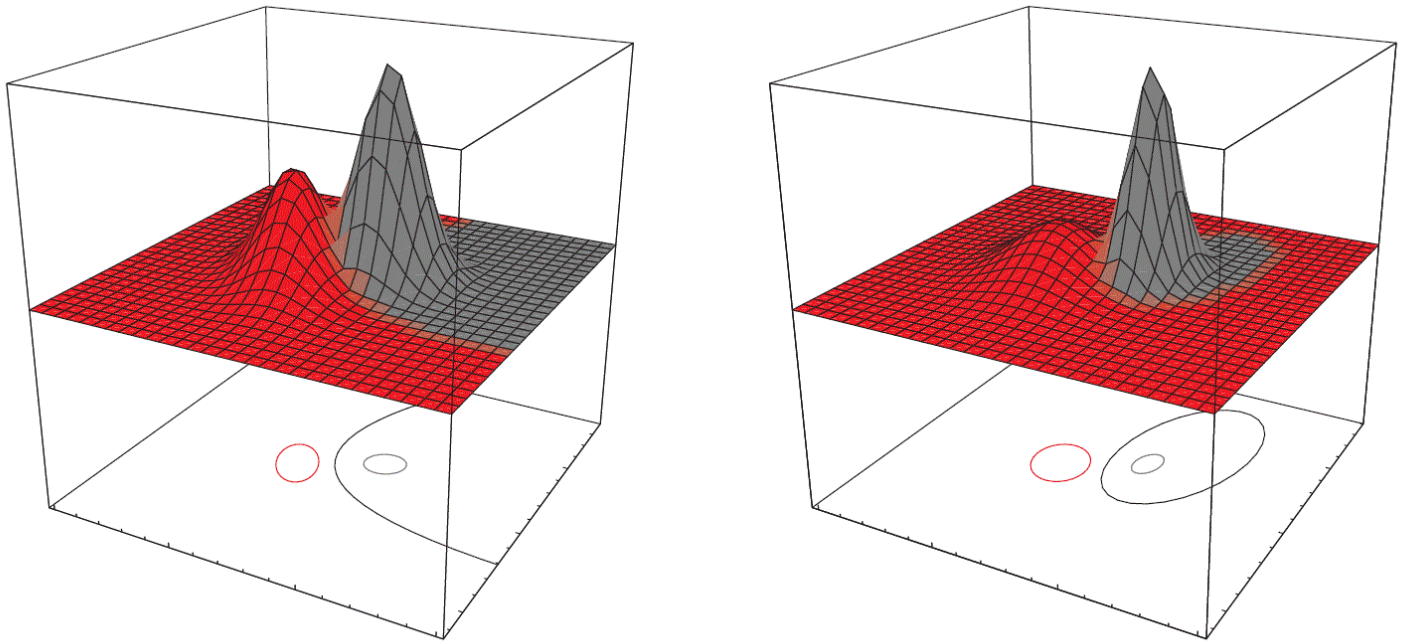
$$= \log(\pi_k) - \tfrac{1}{2}\log((2\pi)^D|\boldsymbol{\Sigma}_k|) - \tfrac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{x} + \tfrac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k$$

$$+\tfrac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{x} - \tfrac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k$$

In the general case of **different covariance matrices** $\boldsymbol{\Sigma}_k$, the discriminant function $g(\boldsymbol{x}, k)$ is quadratic in $\boldsymbol{x}$.

Pattern Recognition and Bayes Optimal Classifier

Example: Two Bivariate Normal Distributions with $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$



The boundary between the decision regions $\mathcal{R}_1$ and $\mathcal{R}_2$ of the two classes is given by a quadric (parabola, ellipse).

R. Duda, P. Hart, D. Stork, Pattern Classification, 2nd edition, Wiley, 2001

Pattern Recognition and Bayes Optimal Classifier

Bayes Optimal Classifier for Identical Covariance Matrices:

In the case of **identical covariance matrices**, i.e. $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$ for $k = 1, \ldots, K$, the discriminant function corresponding to the Bayes optimal classifier δ^* is given by

$$g(\boldsymbol{x}, k) = \log(\pi_k) - \tfrac{1}{2}\log((2\pi)^D|\boldsymbol{\Sigma}|) - \tfrac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}$$

$$+\tfrac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \tfrac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \tfrac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k$$

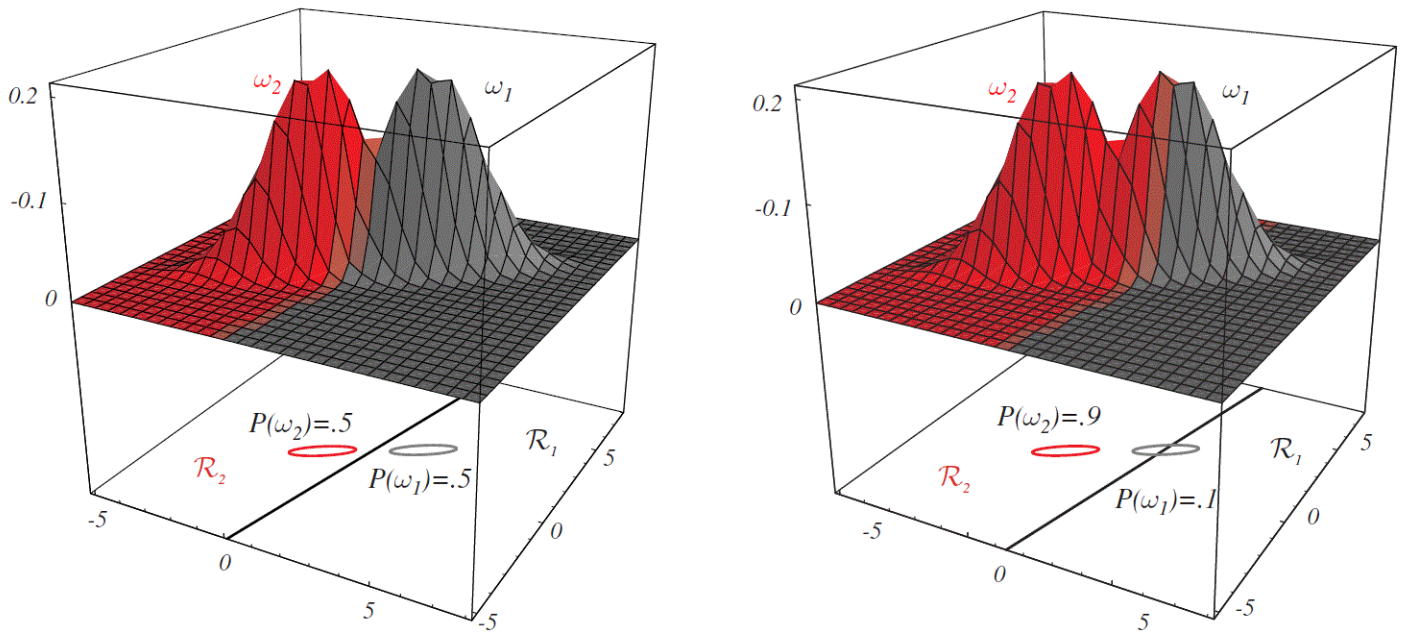
Since constant terms can be omitted, we eventually obtain

$$g(\boldsymbol{x}, k) = \log(\pi_k) + \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \tfrac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k$$

In the special case of **identical covariance matrices** $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_k$, the discriminant function $g(\boldsymbol{x}, k)$ is linear in $\boldsymbol{x}$.

Pattern Recognition and Bayes Optimal Classifier

Example: Two Bivariate Normal Distributions with $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$



The boundary between the decision regions $\mathcal{R}_1$ and $\mathcal{R}_2$ is given by a line. For $\pi_1 = \pi_2$, the line passes through the point $(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$, otherwise it is shifted in the direction of the less probable class.

R. Duda, P. Hart, D. Stork, Pattern Classification, 2nd edition, Wiley, 2001

Pattern Recognition and Bayes Optimal Classifier

Exercise:

Consider the following binary classification problem, where the class-conditional densities are given by

$$f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{3/2}|\boldsymbol{\Sigma}_k|^{1/2}} \cdot \exp\left(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right), \quad k = 1, 2$$

with

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}, \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 2 \\ 2 \\ 3 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 9 & 4 \\ 2 & 4 & 9 \end{pmatrix}$$

The observed feature vector should be $\boldsymbol{x} = (1, 2, 2)^T$. Determine the corresponding posterior probabilities $p(k|\boldsymbol{x})$ for the two cases

(i) $\pi_1 = \pi_2 = 0.5$ (ii) $\pi_1 = 0.4,\ \pi_2 = 0.6$