VIBE PROJECT

Virtual Biomedical and STEM/STEAM Education

2021-1-HU01-KA220-HED-000032251

Funded by the European Union

Erasmus+

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

PÉCSI TUDOMÁNYEGYETEM
UNIVERSITY OF PÉCS

U.PORTO

Politechnika Śląska

DEX innovation centre

VIBE PROJECT

PATTERN RECOGNITION IN BIOMEDICAL ENGINEERING

SUPERVISED LEARNING

Introduction

VIBE PROJECT

Erasmus+

PÉCSI TUDOMÁNYEGYETEM
UNIVERSITY OF PÉCS

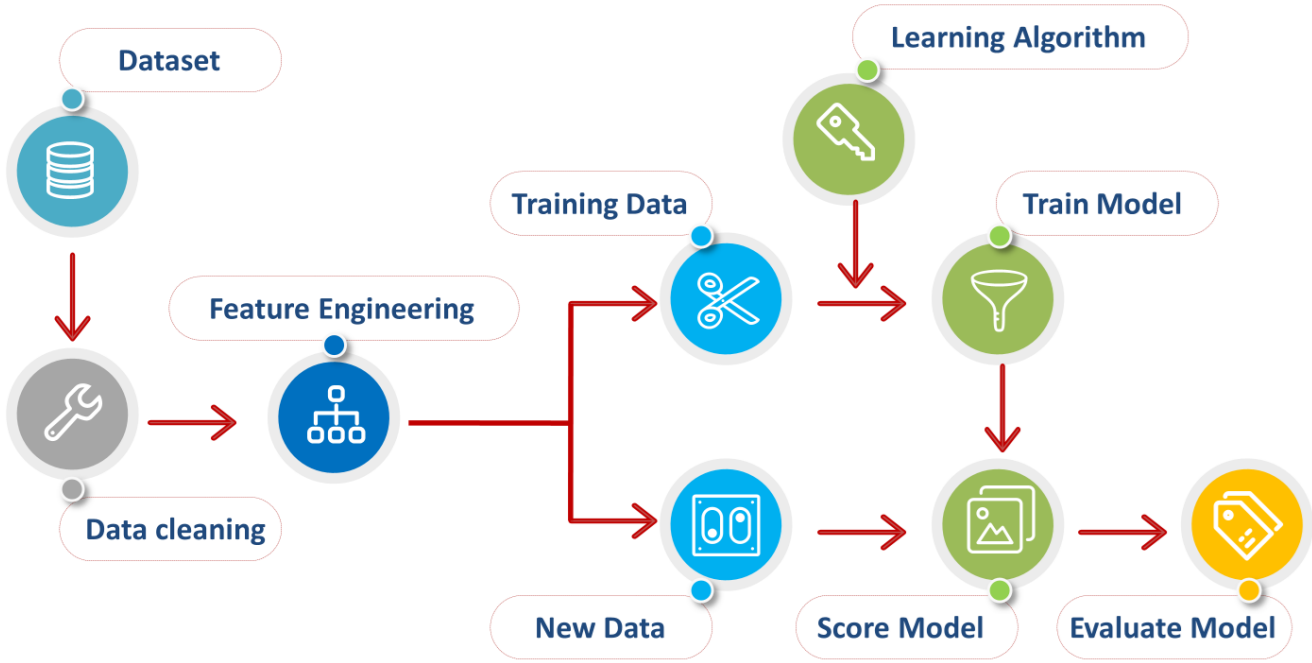
Silesian University of Technology

U.PORTO

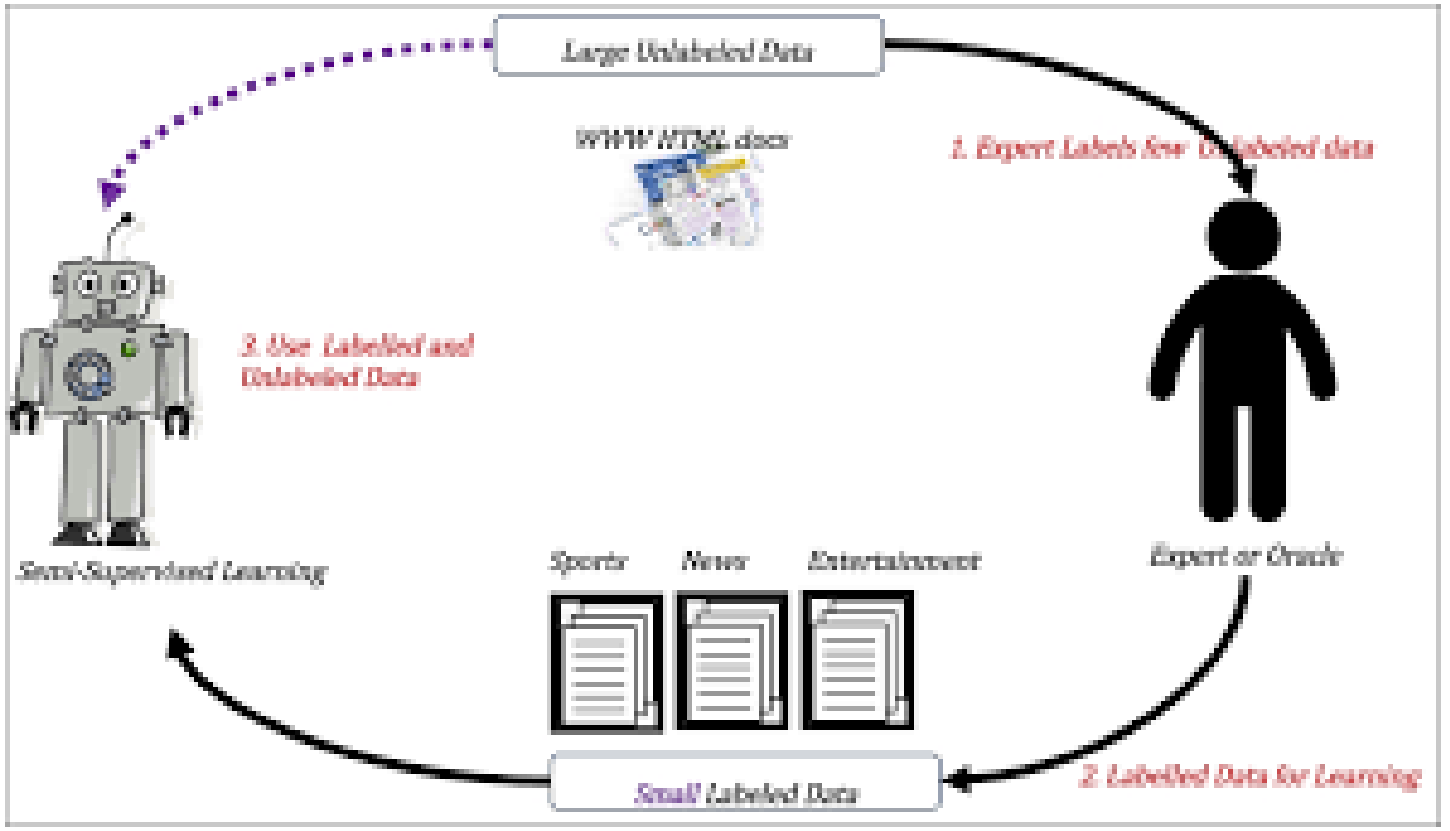
DEX innovation centre

Learning Strategies

- **Unsupervised Learning**
- clusters unlabelled training data described by feature vectors into similar groups.



Learning Strategies



- **Semi-Supervised Learning**
- applies both the labelled and unlabelled training for designing a classification system.

Pattern Recognition Chain

VIBE PROJECT

Erasmus+

PÉCSI TUDOMÁNYEGYETEM
UNIVERSITY OF PÉCS

Silesian University of Technology

U.PORTO

DEX

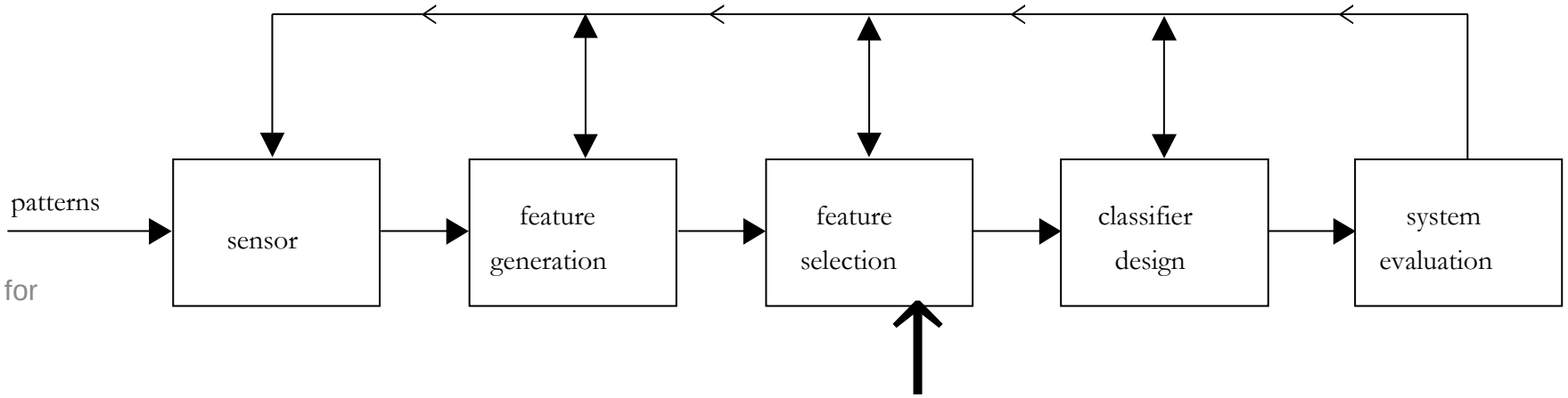
Pattern Recognition Chain

Bayes Decision Theory

Discriminant Functions and Decision Surfaces

Bayesian Classification for Normal Distributions

Estimation of Unknown Probability Density Functions



Grzegorzek Marcin & Doniec Rafał ,. (2024). *Pattern Recognition*. University: Universität zu Lübeck

Overview

- Introduction
- Bayes Decision Theory
- Discriminant Functions and Decision Surfaces
- Bayesian Classification for Normal Distributions
- Estimation of Unknown Probability Density Functions

Statistical Classification - Problem Statement

Classification of an unknown pattern in the most probable of the classes!

- Set of classes: $\{\omega_1,\omega_2,\ldots,\omega_M\}$
- Unknown pattern represented by its feature vector x
- Conditional probabilities: $P(\omega_i|x)$, $i = 1,2,\ldots,M$
- Classification result: the class with the maximum conditional probability

But how to compute the conditional probability for a particular class?

Using the Bayes Rule

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{p(\mathbf{x})} \qquad i = 1, 2 \qquad (1)$$

$p(\mathbf{x})$ – density function for $\mathbf{x}$

Higher a posteriori probability wins

If $P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x})$, $\mathbf{x}$ is classified to ω_1

If $P(\omega_1|\mathbf{x}) < P(\omega_2|\mathbf{x})$, $\mathbf{x}$ is classified to ω_2

Considering the Bayes Rule (Eq. 1)

$$\text{If} \quad \frac{p(\mathbf{x}|\omega_1)P(\omega_1)}{p(\mathbf{x})} > \frac{p(\mathbf{x}|\omega_2)P(\omega_2)}{p(\mathbf{x})}, \qquad \mathbf{x} \text{ is classified to } \omega_1$$

$$\text{If} \quad \frac{p(\mathbf{x}|\omega_1)P(\omega_1)}{p(\mathbf{x})} < \frac{p(\mathbf{x}|\omega_2)P(\omega_2)}{p(\mathbf{x})}, \qquad \mathbf{x} \text{ is classified to } \omega_2$$

Grzegorzek Marcin & Doniec Rafał ,. (2024). *Pattern Recognition*. University: Universität zu Lübeck

$p(\mathbf{x})$ can be disregarded, because it is the same for all classes

If $p(\mathbf{x}|\omega_1)P(\omega_1) > p(\mathbf{x}|\omega_2)P(\omega_2)$, $\mathbf{x}$ is classified to ω_1

If $p(\mathbf{x}|\omega_1)P(\omega_1) < p(\mathbf{x}|\omega_2)P(\omega_2)$, $\mathbf{x}$ is classified to ω_2

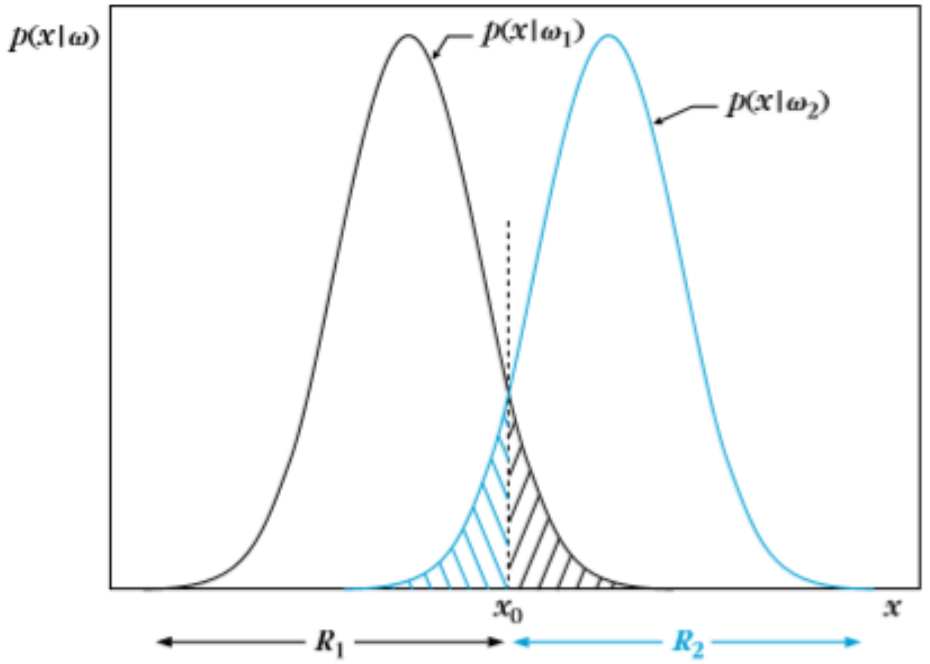
If the a priori probabilities are equal: $P(\omega_1) = P(\omega_2)$

If $p(\mathbf{x}|\omega_1) > p(\mathbf{x}|\omega_2)$, $\mathbf{x}$ is classified to ω_1

If $p(\mathbf{x}|\omega_1) < p(\mathbf{x}|\omega_2)$, $\mathbf{x}$ is classified to ω_2

We are done, since the likelihood density functions $p(\mathbf{x}|\omega_1)$ and $p(\mathbf{x}|\omega_2)$ are assumed to have been trained from examples!

Grzegorzek Marcin & Doniec Rafał ,. (2024). *Pattern Recognition*. University: Universität zu Lübeck



Error Probability: $P_e = \frac{1}{2}\int_{-\infty}^{x_0} p(x|\omega_2)dx + \frac{1}{2}\int_{x_0}^{\infty} p(x|\omega_1)dx$

Grzegorzek Marcin & Doniec Rafał ,. (2024). *Pattern Recognition*. University: Universität zu Lübeck

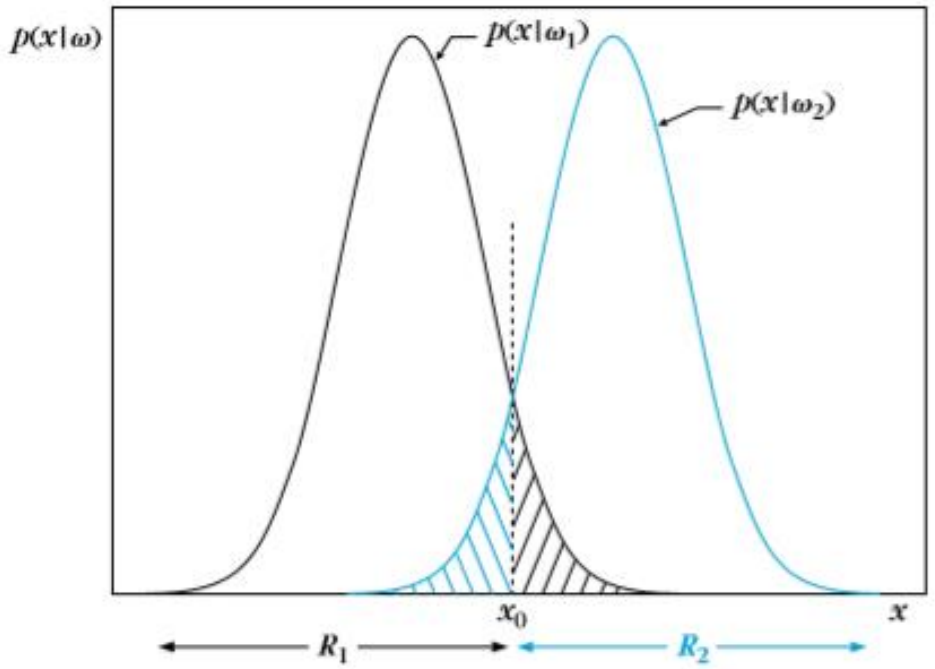
Source:https://medium.com/@thommaskevin/tinyml-gaussian-naive-bayes-classifier-31f8d241c67c

- A priori probabilities are not equal: $P(\omega_1) \neq P(\omega_2)$
- Feature vectors have more than one dimension: $l > 1$

$$\mathbf{x} = [x_1, x_2, \ldots, x_l]^{\mathrm{T}}$$

- General form:

$$P_e = P(\omega_1) \int_{R_2} p(\mathbf{x}|\omega_1) d\mathbf{x} + P(\omega_2) \int_{R_1} p(\mathbf{x}|\omega_2) d\mathbf{x}$$



Bayesian Classifier is OPTIMAL with respect to minimising the classification error probability!

Grzegorzek Marcin & Doniec Rafał ,. (2024). *Pattern Recognition*. University: Universität zu Lübeck

Source:https://medium.com/@thommaskevin/tinyml-gaussian-naive-bayes-classifier-31f8d241c67c

- Classification error probability assigns the same importance to all errors, which is wrong for many applications (e. g., $\omega_1 \rightarrow$ "malignant tumour", $\omega_2 \rightarrow$ "benign tumour").
- In such cases a penalty term is assigned to weight each error.
- A modified version of the error probability has to be minimised:

$$r = \lambda_{12} P(\omega_1) \int_{R_2} p(\mathbf{x}|\omega_1) d\mathbf{x} + \lambda_{21} P(\omega_2) \int_{R_1} p(\mathbf{x}|\omega_2) d\mathbf{x}$$

- For the tumour example λ_{12} is much greater than λ_{21}.

If the a priori probabilities are equal: $P(\omega_1) = P(\omega_2)$

If $p(\mathbf{x}|\omega_2) > p(\mathbf{x}|\omega_1)\frac{\lambda_{12}}{\lambda_{21}}$, $\mathbf{x}$ is classified to ω_2

If $p(\mathbf{x}|\omega_2) < p(\mathbf{x}|\omega_1)\frac{\lambda_{12}}{\lambda_{21}}$, $\mathbf{x}$ is classified to ω_1

Discriminant Functions and Decision Surfaces

- Sometimes it is more convenient to work with functions of probabilities instead of probabilities

$$g_i(\mathbf{x}) \equiv f(P(\omega_i|\mathbf{x}))$$

- $f(\cdot)$ is a monotonically increasing function
- $g_i(\mathbf{x})$ is known as discriminant function
- The decision test is now stated as

$$\text{classify } \mathbf{x} \text{ into } \omega_i \quad \text{if} \quad g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \forall j \neq i$$

- The decision surfaces, separating contiguous regions, are described by

$$g_{ij}(\mathbf{x}) \equiv g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0, \quad i, j = 1, 2, \ldots, M \quad i \neq j$$

Grzegorzek Marcin & Doniec Rafał ,. (2024). *Pattern Recognition*. University: Universität zu Lübeck

Bayesian Classification for Normal Distributions

- The likelihood density functions describing the data in each of the classes, are multivariate Gaussian (normal) distributions

$$p(\mathbf{x}|\omega_i) = (2\pi)^{-\frac{l}{2}} |\mathbf{\Sigma_i}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\mu_i)^{\mathrm{T}} \mathbf{\Sigma_i}^{-1}(\mathbf{x}-\mu_i)}$$

- This "monster" will be denoted by

$$p(\mathbf{x}|\omega_i) = \mathcal{N}(\mu_i, \Sigma_i) \qquad i = 1, 2, \ldots, M$$

Grzegorzek Marcin & Doniec Rafał ,. (2024). *Pattern Recognition*. University: Universität zu Lübeck

- Due to the exponential form of the involved densities, the following discriminant function is applied:

$$g_i(\mathbf{x}) = \ln(p(\mathbf{x}|\omega_i)P(\omega_i)) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i)$$

$$\Updownarrow \text{ considering the "monster"}$$

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

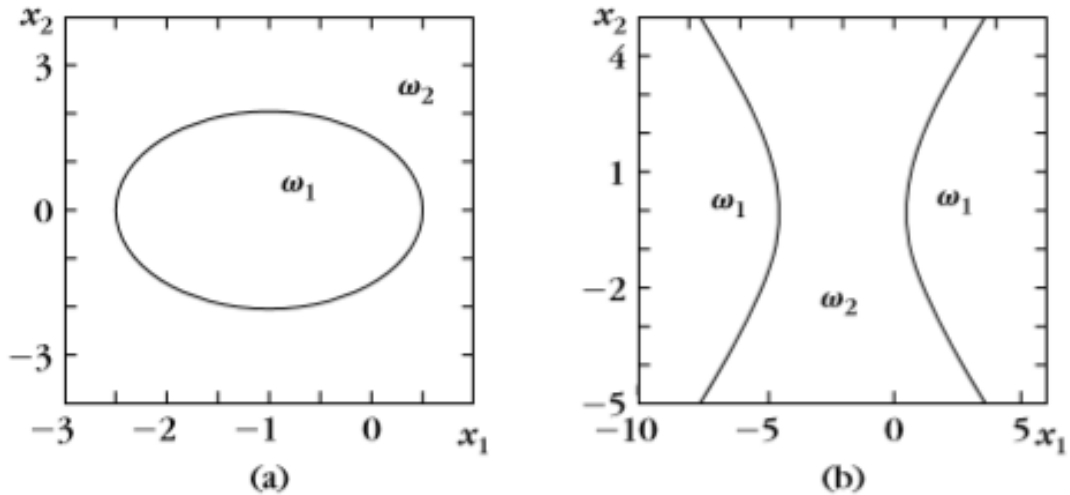
- Where: $c_i = -\frac{l}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i|$

Grzegorzek Marcin & Doniec Rafał ,. (2024). *Pattern Recognition*. University: Universität zu Lübeck

Assuming $l = 2$ and $\sigma_{1,2} = \sigma_{2,1} = 0$, the decision curves

$$g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0$$

are quadrics (i. e., ellipsoids, parabolas, hyperbolas, pairs of lines)



- The only quadric contribution in Equation (2) is $\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\mathbf{x}$
- Assuming that the covariance matrix is the same for all classes $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$ the quadric term will be the same for all discriminant functions
- Thus, the quadric term can be disregarded by decision surface equations. The same is true for the constant c_i
- The simplified version of the discriminant function is just a linear function

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

where

$$\mathbf{w}_i = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i \quad \text{and} \quad w_{i0} = \ln P(\omega_i) - \frac{1}{2}\boldsymbol{\mu}_i{}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i$$

- Assuming equiprobable classes with the same covariance matrix and neglecting the constants Eq. 2 is simplified to:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)$$

- If $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ (diagonal matrix) the maximum $g_i(\mathbf{x})$ implies the minimum Euclidean distance $d_\epsilon = ||\mathbf{x} - \boldsymbol{\mu}_\mathbf{i}||$

- Thus, a feature vector $\mathbf{x}$ is assigned to a class $\widehat{i}$ according to its Euclidean distance to the respective mean points $\boldsymbol{\mu}_i$

$$\widehat{i} = \underset{i}{\operatorname{argmax}}(g_i(\mathbf{x})) = \underset{i}{\operatorname{argmin}}(||\mathbf{x} - \boldsymbol{\mu}_i||)$$

Grzegorzek Marcin & Doniec Rafał ,. (2024). *Pattern Recognition*. University: Universität zu Lübeck

- In practice, it is quite common to assume the Gaussian distribution of the data. In this case, the Bayesian classifier is either linear or quadratic in nature. These approaches are knows as linear discriminant analysis (LDA) or quadratic discriminant analysis (QDA).

- A major problem associated with LDA and QDA is the large number of parameters to be estimated. Thus, l parameters in each mean vector and approximately $\frac{l^2}{2}$ in each covariance matrix. Moreover, a large number of training points N is needed.

- LDA and QDA perform very good for many different applications. However, in many cases the assumed normal distribution is not the right method to statistically model the data.

Estimation of Unknown Probability Density Functions

VIBE PROJECT
Erasmus+
PÉCSI TUDOMÁNYEGYETEM UNIVERSITY OF PÉCS
Silesian University of Technology
U.PORTO
DEX

- So far, we have assumed that the likelihood density functions $p(\mathbf{x}|\omega_i)$ for $i = 1, 2, \ldots, M$ are known.

- This is not the most common case. In many problems, the likelihood density functions have to be estimated from the available training data.

- Here, two estimation methods will be considered, namely
  - Maximum Likelihood Parameter Estimation
  - Maximum a Posteriori Probability Estimation

- Let us consider an M-class problem with feature vectors distributed according to $p(\mathbf{x}|\omega_i)$, $i = 1, 2, \ldots, M$.
- The likelihood functions are assumed to be given in a parametric form. The statistical parameters for the classes ω_i form vectors θ_i which are unknown

$$p(\mathbf{x}|\omega_i) = p(\mathbf{x}|\omega_i; \theta_i)$$

- Goal: to estimate the unknown parameters using a set of known feature vectors in each class.
- Since the estimation process is the same for all classes, the index i will be skipped for further investigations.

- Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be a set of feature vectors describing training samples of a particular class.
- Assuming statistical independence between the different feature vectors, we can form the joint density function

$$p(X;\theta) = p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N;\theta) = \prod_{k=1}^{N} p(\mathbf{x}_k;\theta)$$

- The ML method estimates θ so that the likelihood function takes its maximum value

$$\widehat{\theta}_{\mathrm{ML}} = \underset{\theta}{\operatorname{argmax}} \prod_{k=1}^{N} p(\mathbf{x}_k;\theta)$$

Grzegorzek Marcin & Doniec Rafał ,. (2024). *Pattern Recognition*. University: Universität zu Lübeck

- To find a maximum, the gradient has to be zero

$$\frac{\partial \prod_{k=1}^{N} p(\mathbf{x}_k;\theta)}{\partial \theta} = 0$$

- Due to the monotonicity of the logarithmic function, we can use also the log-likelihood function

$$L(\theta) = \ln \prod_{k=1}^{N} p(\mathbf{x}_k;\theta)$$

- Looking for the maximum here, we have

$$\frac{\partial L(\theta)}{\partial \theta} = \sum_{k=1}^{N} \frac{\partial \ln p(\mathbf{x}_k;\theta)}{\partial \theta} = \sum_{k=1}^{N} \frac{1}{p(\mathbf{x}_k;\theta)} \frac{\partial p(\mathbf{x}_k;\theta)}{\partial \theta} = 0$$

- Set of feature vectors $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$
- θ is an unknown random vector
- The starting point is the following density function

$$p(\theta|X) = \frac{p(\theta)p(X|\theta)}{p(X)}$$

- The MAP estimate $\widehat{\theta}_{\mathrm{MAP}}$ is defined as a point where $p(\theta|X)$ becomes maximum

$$\widehat{\theta}_{\mathrm{MAP}} : \frac{\partial}{\partial\theta} p(\theta|X) = 0 \quad \text{or} \quad \frac{\partial}{\partial\theta}(p(\theta)p(X|\theta)) = 0$$

Grzegorzek Marcin & Doniec Rafał ,. (2024). *Pattern Recognition*. University: Universität zu Lübeck

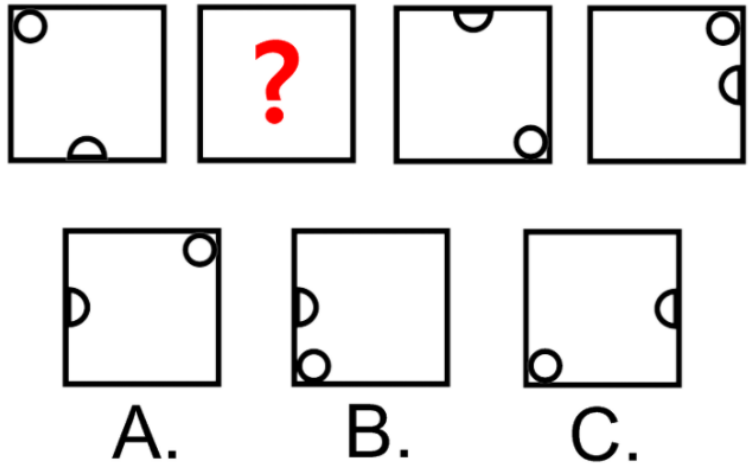
Pattern Recognition Quiz (1)

1. Patterns Using Arithmetic Warmup 18,15,12,9, ?

a) 8
b) 7
c) 6
d) 5

2. Patterns Using Geometry Warmup ?

a) A
b) B
c) C



3 In the generalized formula for Bayes' theorem, what does the Greek letter Sigma in the denominator mean?

a) *Add the results of the terms with subscripts 1 to n*
b) *Multiply the results of the terms with subscripts 1 to n*
c) *It guarantees that the denominator can never be zero*
d) *Only the Secret Order of the Sigma knows*

$$P(A_k|B) = \frac{P(A_k) \times P(B|A_k)}{\sum_{i=1}^{n} P(A_i) \times P(B|A_i)}$$

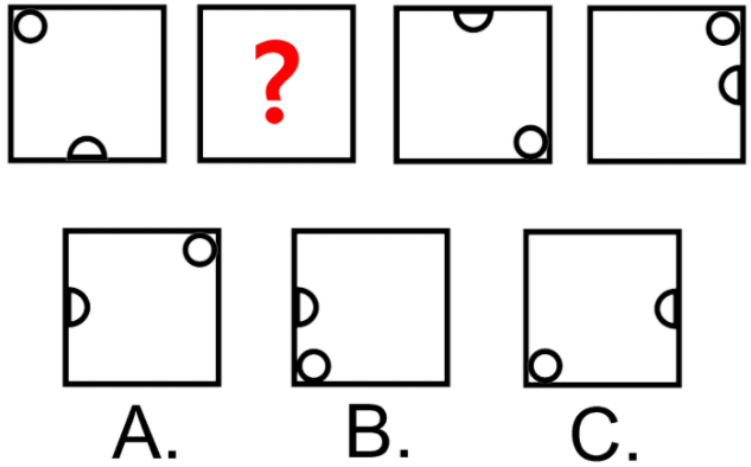
Pattern Recognition Quiz (1)

Thank you for your attention

1. Patterns Using Arithmetic Warmup 18,15,12,9, ?

a) 8
b) 7
c) 6
d) 5

2. Patterns Using Geometry Warmup ?



a) A
b) B
c) C

In the generalized formula for Bayes' theorem, what does the Greek letter Sigma in the denominator mean?

3. **a) Add the results of the terms with subscripts 1 to n**
b) Multiply the results of the terms with subscripts 1 to n