

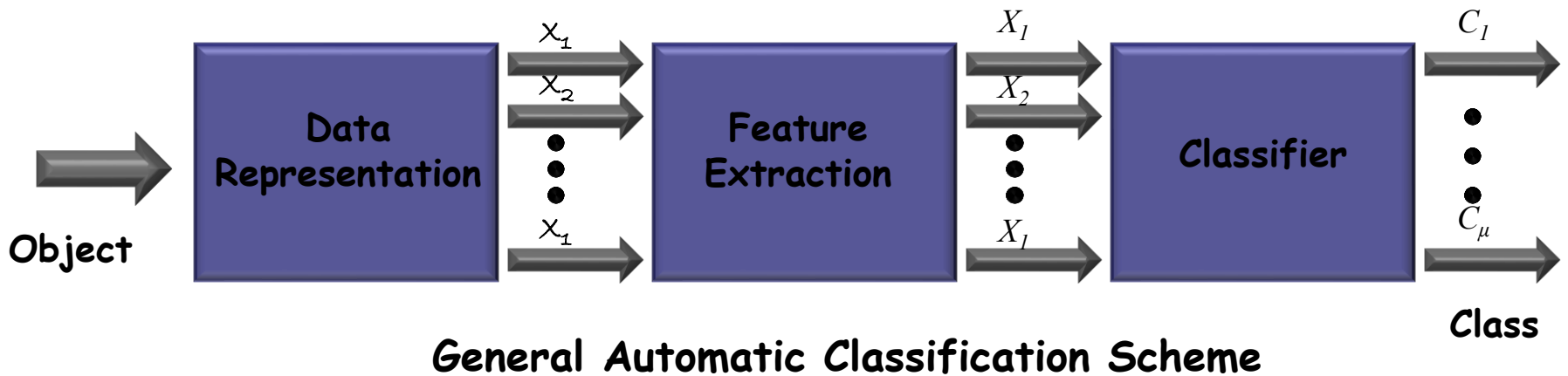


Data Preprocessing & Feature Representation



Dr. Konstantinos Karampidis

Pattern Recognition Systems

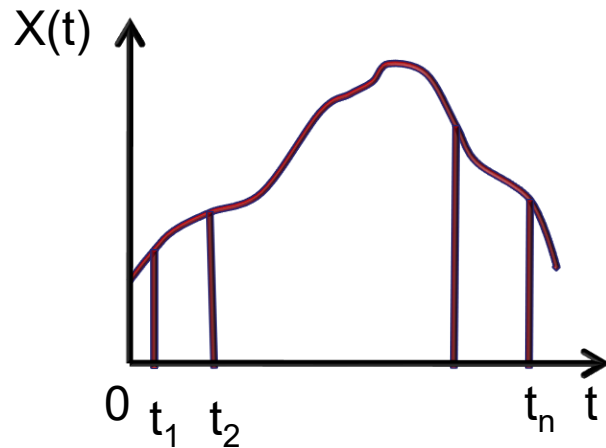




Data Representation (1/2)

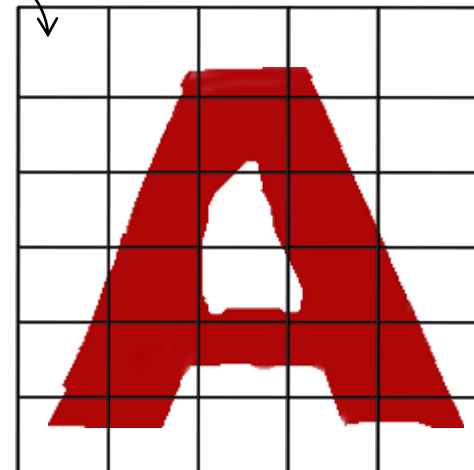
- ❑ **Pattern Representation by Data Collection.**
 - Patterns in time (time series)
 - Signal Sampling N time moments
 - Patterns in space (geometric objects)
 - Intensity values of the pixels from a digital image

Data Representation (2/2)



a. Time Series

Pixel
#1



b. Geometric Object



Feature Vectors

- ❑ **Raw Feature Vector X .**
 - Defines a prototype in N-dimensional space.
 - Elements are random variables
- ❑ **General format**

$$X_{space} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix}$$

$$X_{time} = \begin{pmatrix} X(t_1) \\ X(t_2) \\ \vdots \\ X(t_N) \end{pmatrix}$$

OR

$$X_{time} = (X(t_1), \dots, X(t_N))^T$$

$$X_{space} = (X_1, \dots, X_N)^T$$

$()^T$

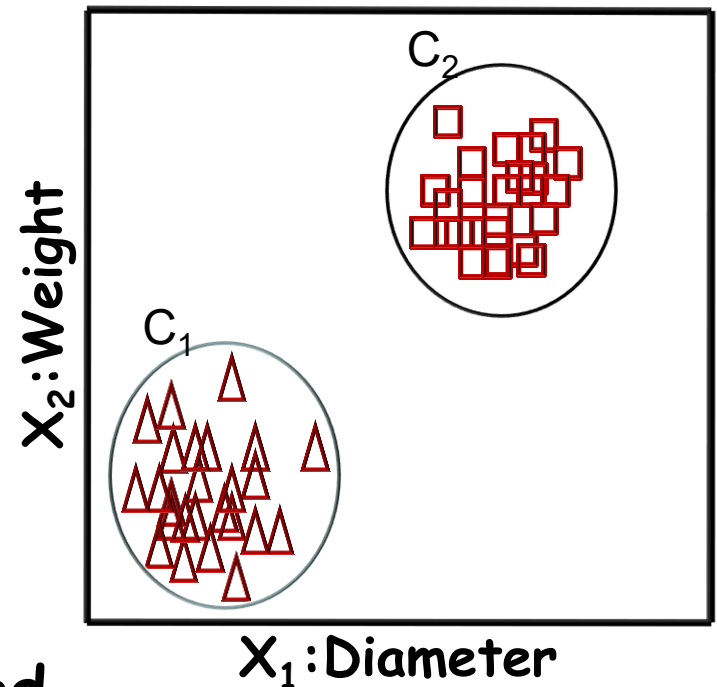
Prototypes (1/2)

□ Two categories of prototypes.

- C_1 : mandarins
- C_2 : watermelons
- Two measurements:
 - X_1 : Diameter
 - X_2 : Weight
- Raw feature vector

$$\mathbf{X} = [X_1, X_2]^T$$

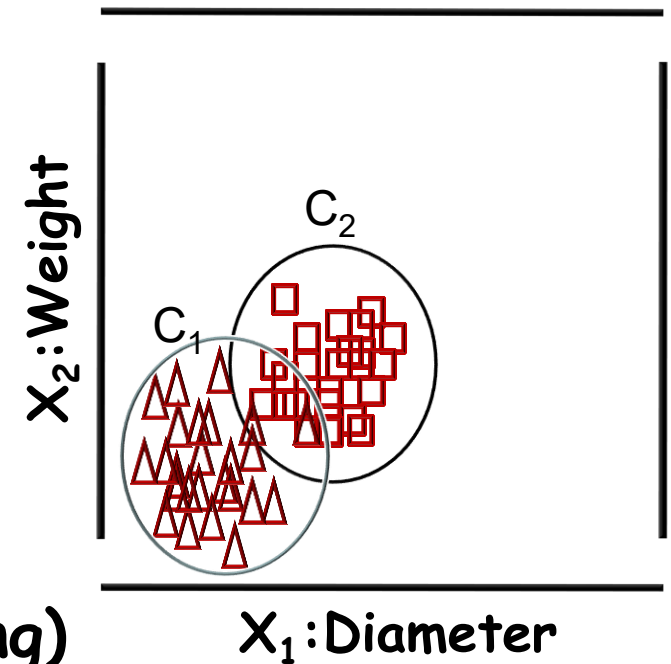
□ 2 separable groups are created



Prototypes (2/2)

□ Two categories of prototypes.

- C_1 : mandarins
- C_3 : oranges
- Two measurements:
 - X_1 : Diameter
 - X_2 : Weight

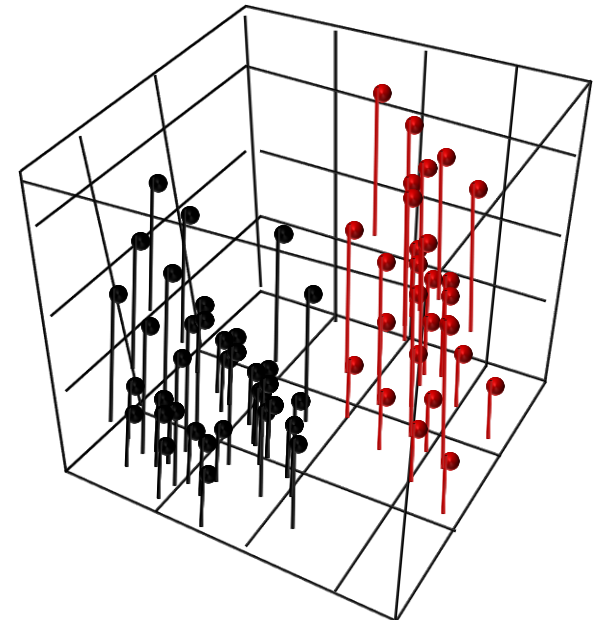


□ Non-Separable groups (overlapping)

Multidimensional Data

- ❑ **Separable Categories in 3D space ($N=3$)**
 - $N>3$: difficult representation
- ❑ **Feature Vector Extraction: Dimensionality Reduction**
- ❑ **Sets of features:**
 - Intraset: Common Properties
 - Interaset: Differences
 - **Intraset are not useful for all sets**

Intraset features may not help distinguish between classes because they describe properties that are common within a set/class and are not necessarily discriminative.





Feature Selection (1/2)

Goal of Feature Selection: Identify the interset/discriminative features that provide separation between different categories.

□ Example:

- Oranges - Intraset: weight, diameter, color
- Mandarins - Intraset: weight, diameter, color
- Mandarins - Oranges - Intraset: color
- Mandarins - Oranges - Interaset: weight, diameter

□ Goal: To find interset (separating) features between categories

□ Difficult from a raw input vector X

- Solution: extraction/selection of discriminative features from X

□ Example: Recognition of printed characters.

$$X = [X_1, X_2, \dots, X_N]^T$$



Feature Selection (2/2)

□ features:

- x_1 : area of the right segment of the character
- x_2 : area of the left segment of the character
- x_3 : perimeter of the character
- z_1 : total area to squared perimeter ratio

$$\frac{(X_1 + X_2)}{X_3^2}$$

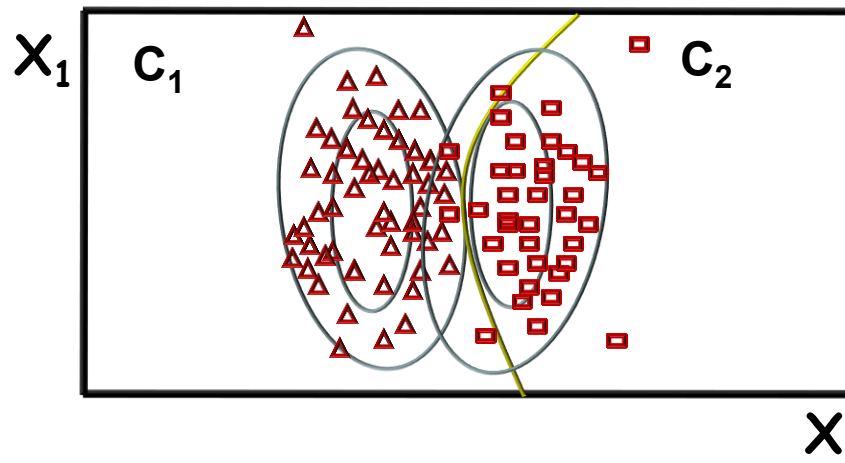
- z_2 : symmetry degree

$$\left(\frac{X_1}{X_2} \right)$$

□ x_1, x_2, x_3 : independent variables, z_1, z_2 : dependent

Decision Boundaries (1/3)

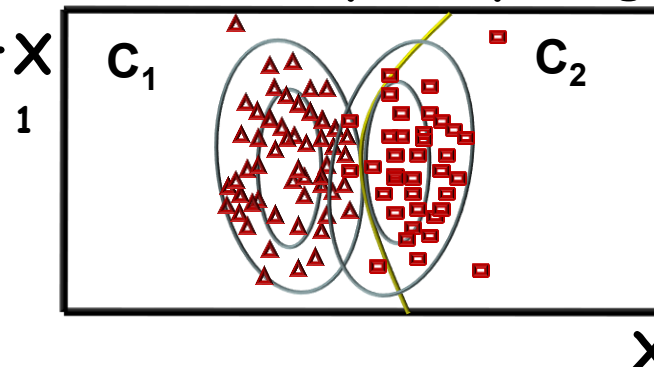
- Recognition & Classification
 - Define optimal decision procedures for recognition and classification.
 - The system decides which category/class a prototype belongs to.
- Classification into M Categories
 - $C_1, C_2, \dots, C_M$
 - Each prototype is assigned to one of the M categories.



$$d(x_1, x_2) = 0$$

Decision Boundaries (2/3)

- Decision Boundaries
 - Decision boundaries separate the feature space into different decision regions, corresponding to the M categories.
 - A decision boundary is the boundary between regions assigned to different classes.
- Probability Density Functions (PDFs)
 - Curves represent the Probability Density Functions of the different categories.
 - Decision boundaries can be determined by comparing the distributions of the classes.



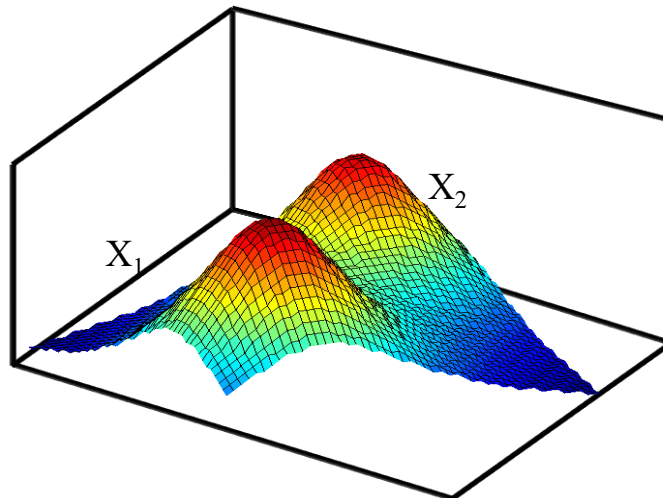
$$d(x_1, x_2) = 0$$

Decision Boundaries (3/3)

For two categories C_1 and C_2 , define a decision function:
$$d(x_1, x_2)$$

The Decision Boundary (Decision Limit) is defined by:
 $d(x_1, x_2) = 0$. Two areas:

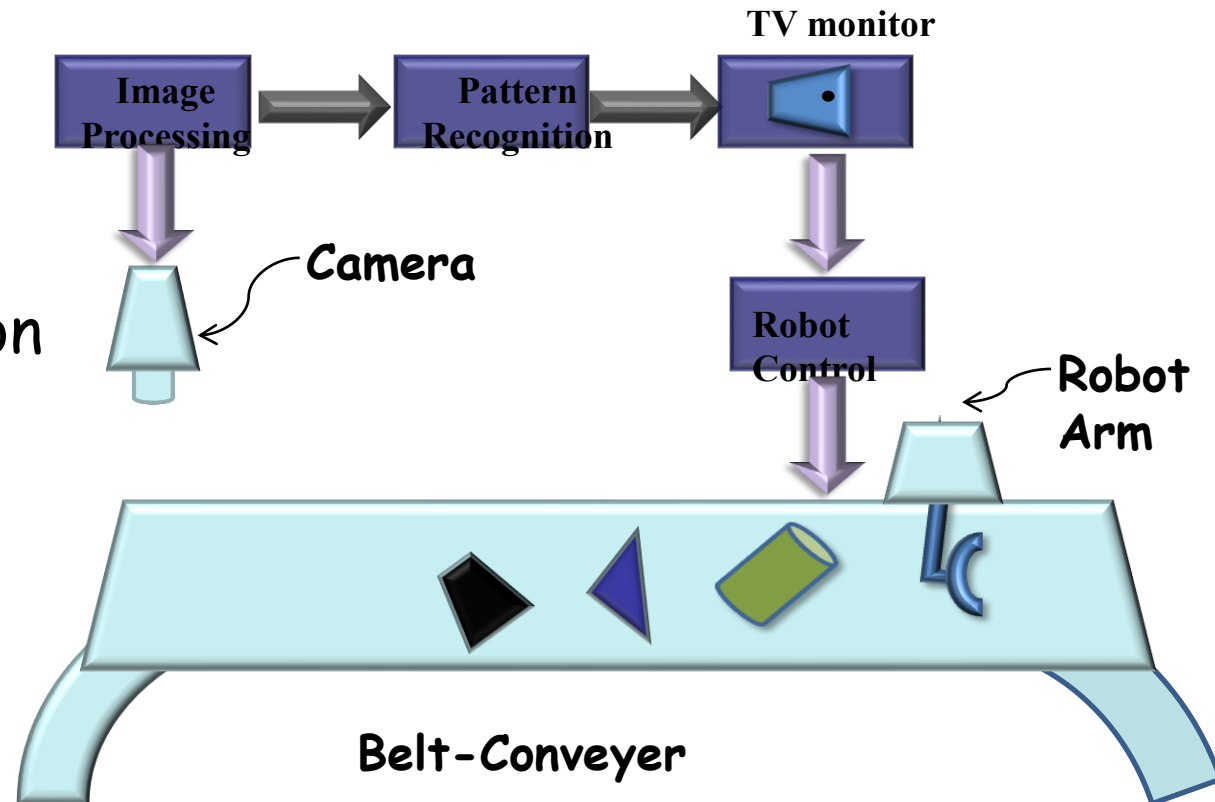
- $d(x_1, x_2) > 0$ category C_2
- $d(x_1, x_2) < 0$ category C_1



■ Probability Density Function for each category

Application: Industrial Robot

- **Conveyer Belt**
- **Robot**
 - Quality inspection
 - Assembly
 - Voice commands





Application: Industrial Robot

- ❑ **Application Analysis – Pattern Recognition Problems:**
 - Recognition of objects (shape based)
 - Camera view independency
 - Quality inspection
 - Voice commands recognition (from the supervisor)
 - Recognition of Supervisors ID (from speech).
 - Recognition of robots improper functioning (self test)



Application: Industrial Robot

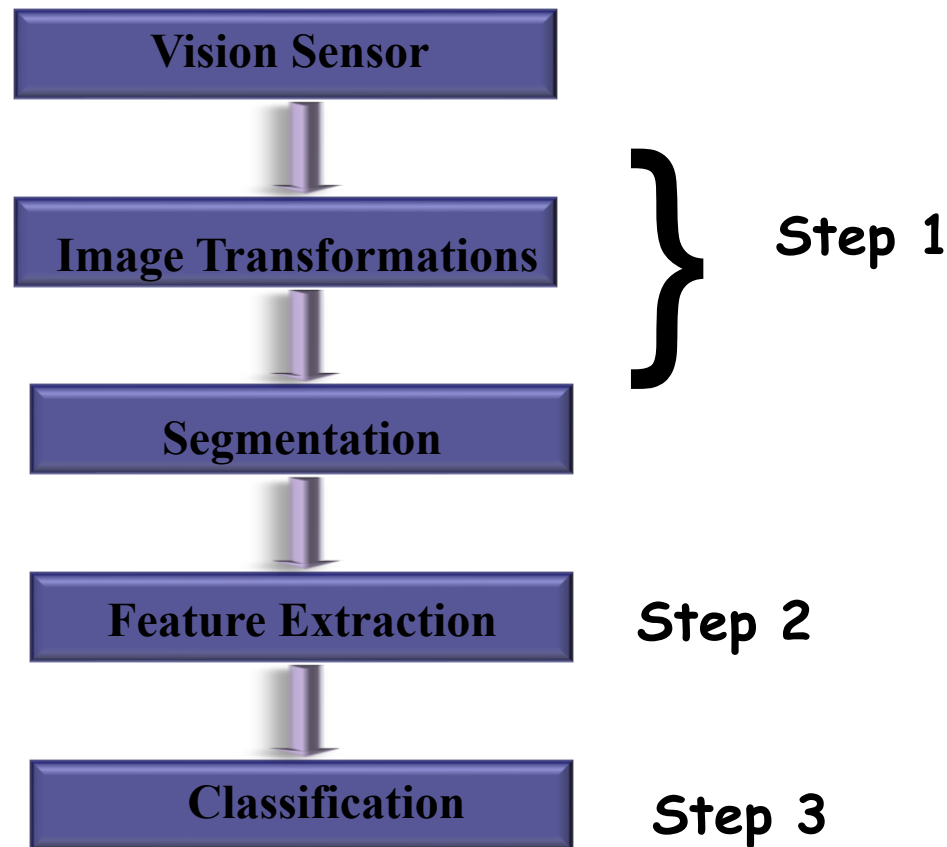


Image based object recognition



Application: Industrial Robot

- ❑ Vision Sensor
 - Sequence of photosensors → simple cases
 - Camera → usual cases
 - Combination of cameras → stereoscopic vision and 3D analysis for complex cases
- ❑ Single Camera Solution: Frame Grabber → captures/acquires the image from the camera.
- ❑ Input Data: digital image

Basic flow: Vision Sensor → Digital Image



Application: Industrial Robot

□ Template Matching:

- Used in some low-budget applications for optical pattern recognition. The method compares the input image with stored templates/samples.

□ Problems:

- ✓ Images may be corrupted by noise.
- ✓ Object orientation not predefined.
 - Store big number of samples for every component (all the angles of view).
 - Digital image: $128 \times 128 \times 8$ (color) = 16 Kbytes

□ Limitation

- A large number of stored templates is required to cover different object orientations. Therefore, Template Matching is not suitable for the specific application.

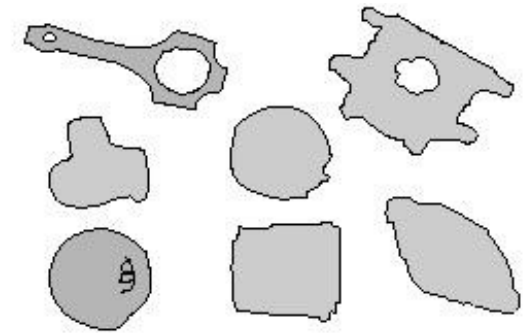
Application: Industrial Robot

□ Feature Extraction

➤ Image transformations

- Enhancement - restoration procedures
- Removal of components' margins,
- Color enhancement and equalization,
- Edge Detection,
- Removal of non-necessary information

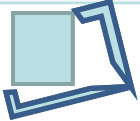
➤ **Final outcome:** binary representation of the component's edges or borderlines



□ Segmentation

- ### ➤ Separates the image into **meaningful regions** corresponding to relevant objects/components.

Image Transformations + Segmentation = Image Processing

Area (no holes)				
Hole Area (Inner Area)				
Biggest Dimension				
Diameter (Ferret)				
Perimeter				
Concave Perimeter				
Maximum Horizontal Chord				

Feature Selection



Pattern Recognition Issues

□ Feature Extraction

- The distinction between feature extraction and classification is somehow arbitrary.
- An ideal feature extractor transforms the input into features that make the classification problem simple.
- Conversely, an ideal classifier would not require a highly specialized feature extractor.



Pattern Recognition Issues

□ Feature Extraction

- Therefore, the distinction between feature extraction and classification is more theoretical than practical.
- **Feature Extraction is application-dependent:** the appropriate features depend on the specific problem and objects being recognized
- **The Classification procedure is more general** and assigns the extracted feature vector to a category.



Pattern Recognition Issues

- ❑ **Noise is unwanted variation/error introduced during data acquisition:**
 - It is related to the **error tolerance of the sensor** and to physical processes at the molecular level
 - Jitter - small random variations in timing/position
 - Distortion- unwanted alteration of the acquired signal/image
 - Salt & Pepper noise - random bright and dark pixels in an image
 - Electromagnetic interference (EMI) - interference caused by electromagnetic sources



Pattern Recognition Issues

□ Noise:

- All practical applications are affected by some form of noise during data acquisition.
- The noise source should be considered as part of the overall system, because it affects the acquired data and consequently the recognition/classification performance



Pattern Recognition Issues

❑ Complexity of a model

Question:

Is it possible to construct a complex pattern recognition system that achieves perfect classification on the training prototypes, but fails to classify real/unseen data?

Answer: Yes.

- ❑ A highly complex model may depend on specific characteristics of the training prototypes rather than on the underlying properties of the real data.
- ❑ As a result, it can perform very well on the training data but poorly on new data.
- ❑ This phenomenon is known as overfitting.



Pattern Recognition Issues

□ Complexity of a model

The goal is to find an appropriate model complexity:

- The model should be complex enough to distinguish differences between categories, but not so complex that it fails to generalize to real/unseen data.

In other words:

Too simple → underfitting

Appropriate complexity → good generalization

Too complex → overfitting



Pattern Recognition Issues

❑ Choosing the Classification Model

- How do we choose the most suitable model for a specific application;

Answer:

By evaluating different candidate models on their ability to classify unseen/real data and selecting the one with the best generalization performance



Pattern Recognition Issues

□ Choosing the Classification Model

- How do we decide whether to reject one model;

Answer:

Reject a model if it has poor classification performance, poor generalization, or unnecessary complexity.



Pattern Recognition Issues

❑ Choosing the Classification Model

- Is there a systematic method for choosing a model, rather than relying only on trial and error;

Answer:

Yes. Models can be systematically evaluated and compared using appropriate performance measures on validation/unseen data



Pattern Recognition Issues

❑ Missing values from feature vectors

In practice, not all feature values are always available. This brings the question:

- How can a classifier operate when some feature values are missing?

A simple solution is to replace missing values with 0, or the average value of the available observations.

However, this may not be the best solution, because the replacement value may not represent the actual missing value and can affect classification performance.



Pattern Recognition Issues

❑ Missing values from training data

Missing values may also occur in the training patterns (prototypes) used to create the recognition system.

❑ This raises another question:

How can the classifier be trained when some training feature values are missing?

The classifier can be trained by using the available features of the training prototypes and appropriately handling the missing values.



Pattern Recognition Issues

- ❑ **Segmentation**
- ❑ **Application of Industrial Robot:**
 - ❑ Components on the conveyor belt are assumed to be non-overlapping.
 - ❑ What happens when components overlap?

Importance of Segmentation

- ❑ The patterns/components must be correctly segmented before their features can be extracted.
- ❑ The components should be separated into meaningful regions so that they can be recognized using the features extracted from their segments.



Pattern Recognition Issues

❑ Segmentation Difficulties

Segmentation can be difficult when objects/components are overlapping or touching.

- ❑ The difficulty depends on the application.
- ❑ Example: Speech Recognition, where it is difficult to segment continuous speech into meaningful units.



Pattern Recognition Issues

❑ Invariable Transoformations

Industrial Robot: object placing on the corridor may vary on
Orientation Translation & scaling independency

- ❑ The recognition system should ideally be independent/invariant to these transformations.
- ❑ Complex transformations can provide invariance, but they are usually application-specific.

Examples

- Optical Recognition of handwritten characters: classifier non-sensitive to line thickness
- Image Recognition: light conditions, shadows.



Pattern Recognition Issues

❑ Invariability / Invariance

Two important questions:

- ❑ How can we determine which variations should be invariant?
- ❑ How can this prior knowledge be incorporated into the recognition system?

Key idea:

- ❑ Changes in the appearance of an object should not necessarily change its class/category.



Questions ?



References

- ❑ Duda, Hart: Pattern Classification and Scene Analysis. J. Wiley & Sons, New York, 1982. (2nd edition 2000).
- ❑ Fukunaga: Introduction to Statistical Pattern Recognition. Academic Press, 1990.
- ❑ Bishop: Neural Networks for Pattern Recognition. Claredon Press, Oxford, 1997.
- ❑ Schlesinger, Hlaváč: Ten lectures on statistical and structural pattern recognition. Kluwer Academic Publisher, 2002.
- ❑ Satoshi Watanabe Pattern Recognition: Human and Mechanical, Wiley, 1985
- ❑ E. Gose, R. Johnsonbaught, S. Jost, Pattern recognition and image analysis, Prentice Hall, 1996.
- ❑ Sergios Thodoridis, Kostantinos Koutroumbas, Pattern recognition, Academiv Press, 1998.