ML Intro

by Ahmet Sacan

Machine Learning

- Assumption:
  - Structure or pattern arises from the fact that we have measurements on similar group of subjects
- Questions:
  - What are these groups ?
  - How many groups are there?
  - Which subject belongs to which group?

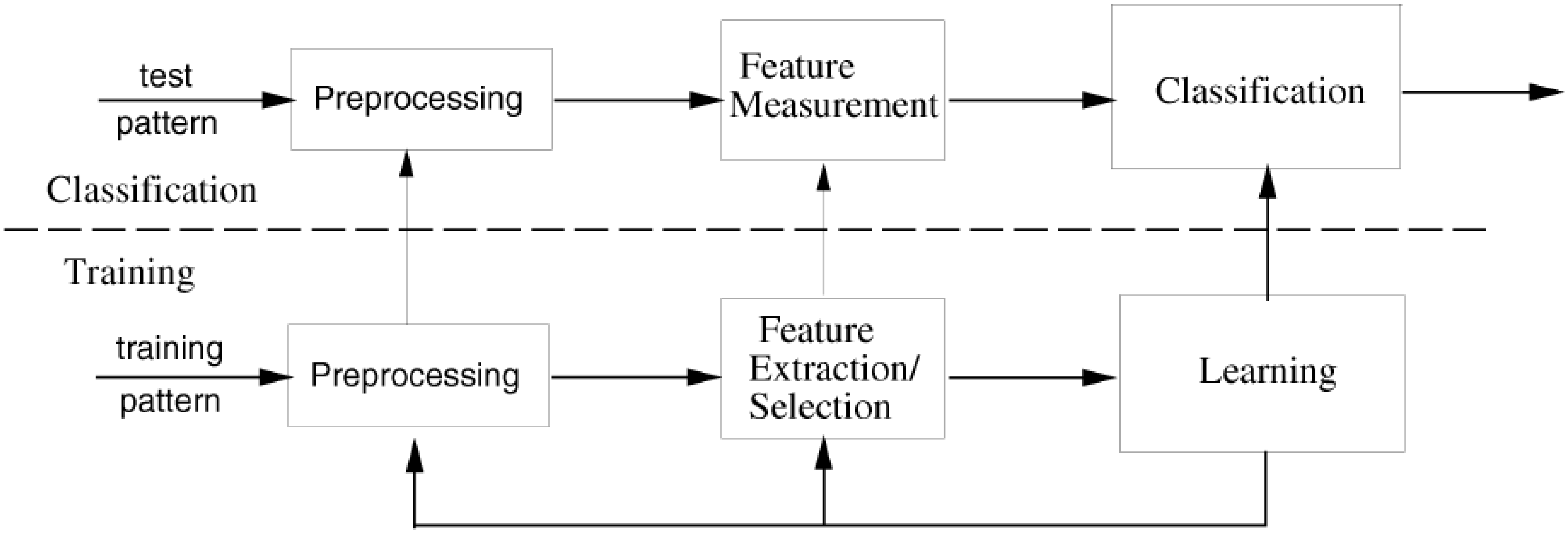
Machine Learning

- Study of computer systems that improve their performance through experience.
  - Learn existing and known structures and rules.
  - Discover new findings and structures.
- Applications:
  - Face recognition
  - OCR
  - Bioinformatics
- Supervised learning vs. unsupervised learning
- Semi-supervised learning

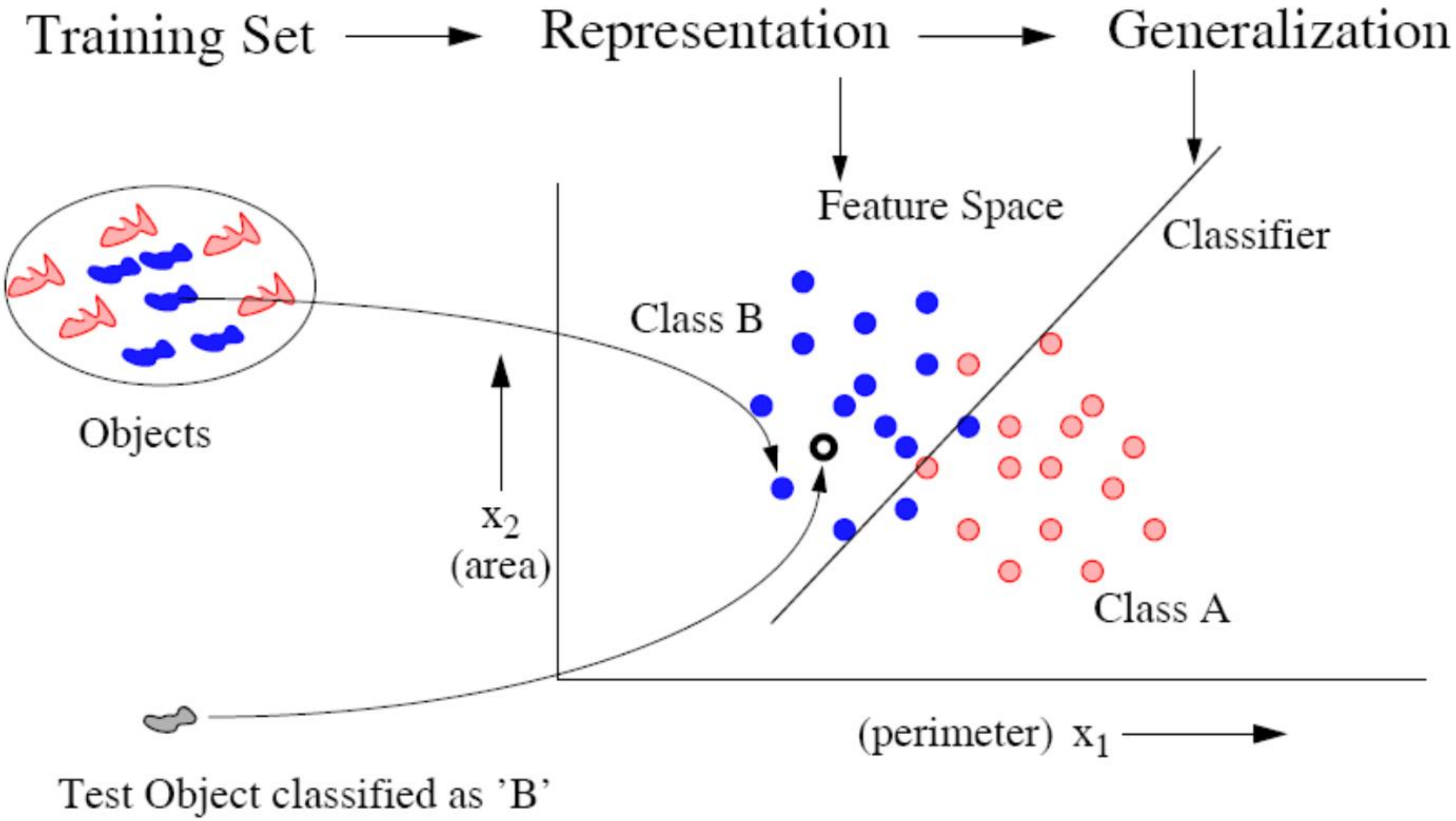
Supervised vs. unsupervised learning

- Why is unsupervised learning useful?
  - Labeling large data sets can be a costly procedure (i.e., speech recognition)
  - Class labels may not be known beforehand (i.e., data mining)
  - Large datasets can be compressed by finding a small set of prototypes (kNN)
- Reinforcement Learning:
  - uses a reward signal (real valued or binary) to tell the learning system how well it is performing
  - the goal of the learning system (or agent) is to learn a mapping from states onto actions (an action policy) that maximizes the total reward

Statistical Pattern Recognition workflow

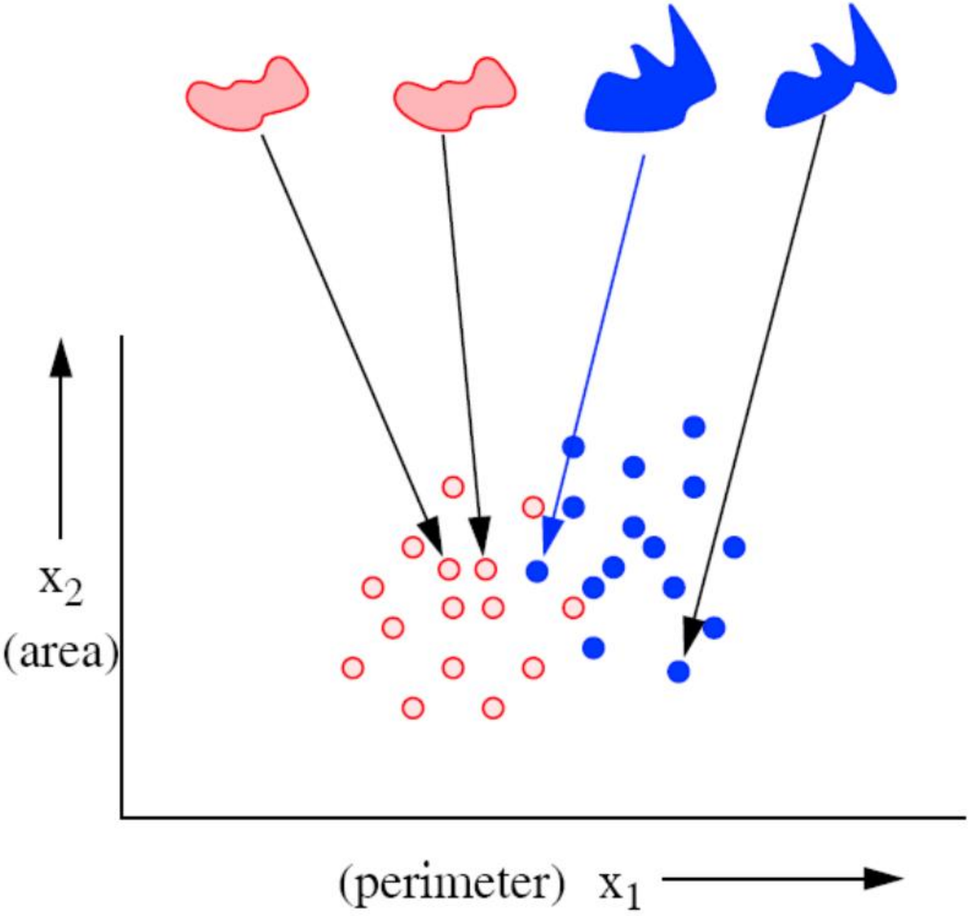


Jain et al. 2000


Training Set
Representation
Generalization
Feature Space
Classifier
Class B
Objects
x_2
(area)
Class A
(perimeter) x_1
Test Object classified as 'B'

Duin et al.

Compactness Hypothesis



Similar objects are close in feature space; Different objects may be close or remote!!

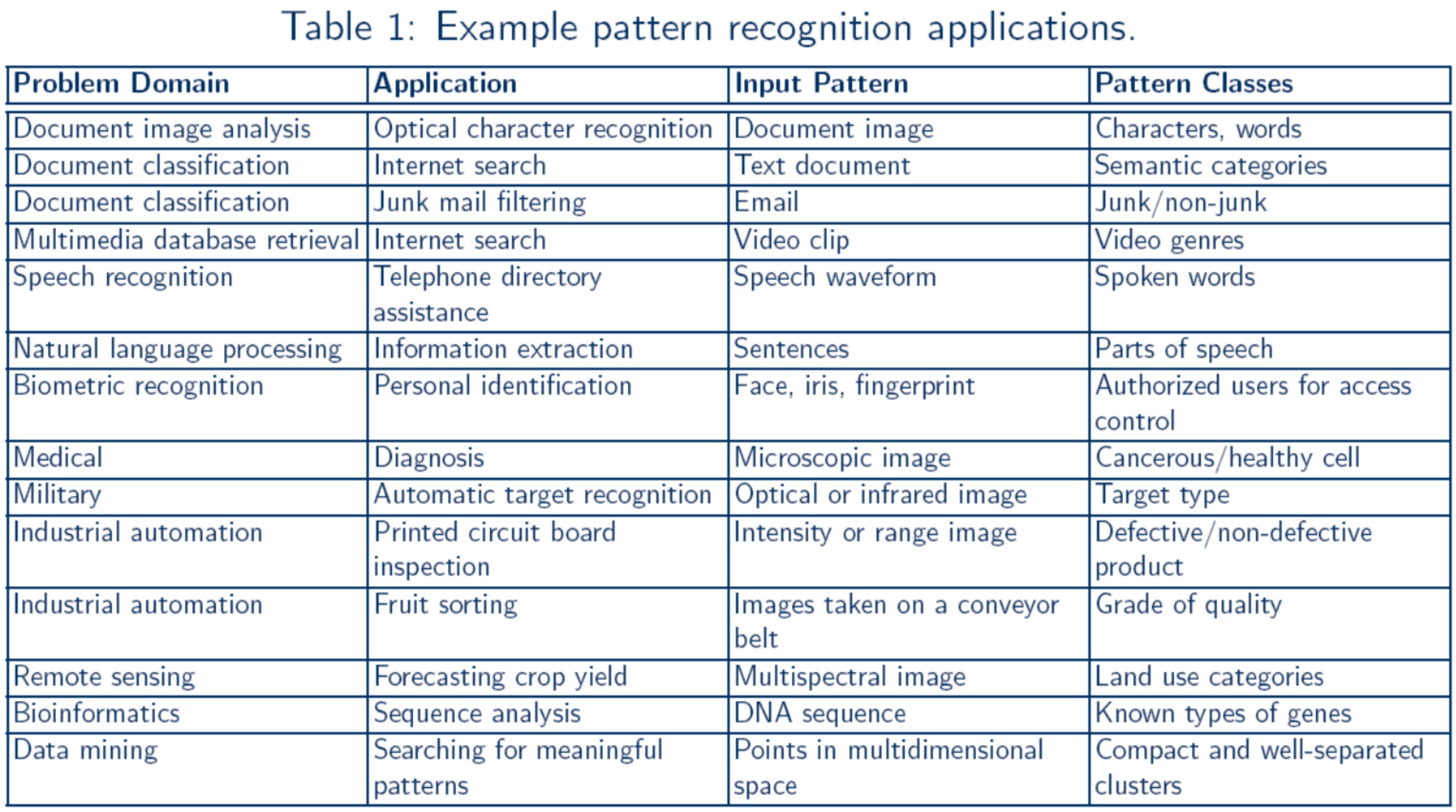


Table 1: Example pattern recognition applications.

Problem Domain	Application	Input Pattern	Pattern Classes
Document image analysis	Optical character recognition	Document image	Characters, words
Document classification	Internet search	Text document	Semantic categories
Document classification	Junk mail filtering	Email	Junk/non-junk
Multimedia database retrieval	Internet search	Video clip	Video genres
Speech recognition	Telephone directory assistance	Speech waveform	Spoken words
Natural language processing	Information extraction	Sentences	Parts of speech
Biometric recognition	Personal identification	Face, iris, fingerprint	Authorized users for access control
Medical	Diagnosis	Microscopic image	Cancerous/healthy cell
Military	Automatic target recognition	Optical or infrared image	Target type
Industrial automation	Printed circuit board inspection	Intensity or range image	Defective/non-defective product
Industrial automation	Fruit sorting	Images taken on a conveyor belt	Grade of quality
Remote sensing	Forecasting crop yield	Multispectral image	Land use categories
Bioinformatics	Sequence analysis	DNA sequence	Known types of genes
Data mining	Searching for meaningful patterns	Points in multidimensional space	Compact and well-separated clusters

Jain et al. 2000

Challenges

- High data dimensionality
- Limited sample size
- Time/memory requirements
- Noisy/incomplete data
- Heterogeneous/inconsistent data
- Sometimes, don't know what we are looking for and just hope we'll get lucky

“Curse of dimensionality“

- Problem in high-dimensional spaces.
- Required sample size grows exponentially with number of dimensions
- Computation complexity
- Estimation accuracy
- Approaches
  - Feature extraction
  - Feature selection
  - Manifold learning
  - Kernel learning

Summary

- Feature space
- Descriptive/inferential statistics
- Exploratory/confirmatory analysis
- Machine Learning
- Supervised/Unsupervised Learning
- Compactness hypothesis
- Challenges:
  - Curse of dimensionality
  - Insufficient/noisy data