Data, Measurements, Features

Middle East Technical University

Dep. of Computer Engineering

2009

compiled by V. Atalay

What do you think of when someone says 'Data'?

- We might abstract the idea that

  data *are information not yet in the form we want it, and therefore needing nontrivial processing.*

- Moreover, the information is *incomplete*, through
  - errors or
  - lack of some measurements,

so probable reconstructions of the incomplete parts are also desired.

In general, **data** consists of *propositions* that reflect reality.

A large class of practically important propositions are *measurements* or *observations* of a *variable*.

Such propositions may comprise *numbers*, *words*, or *images*.

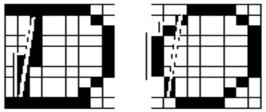
Features

Objects:

- physical entities (images, patients, clients, molecules, cars, signal samples, software pieces), or
- states of physical entities (board states, patient states etc).

Features: measurements or evaluation of some object properties.

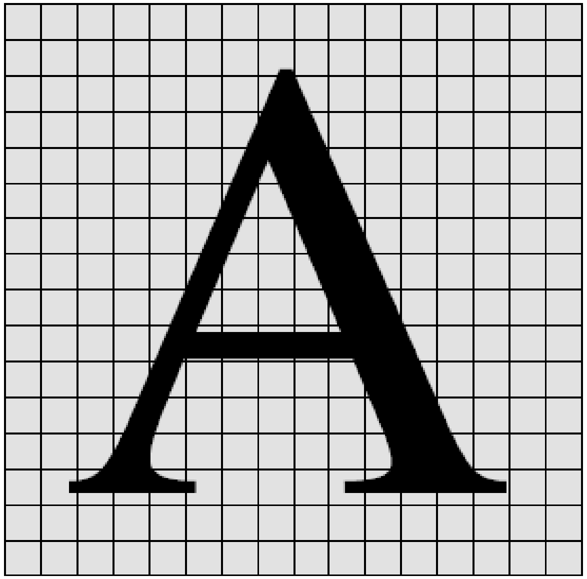
Example: Are pixel intensities good features?
No - not invariant to translation/scaling/rotation.



Better: type of connections, type of lines, number of lines, etc.

Selecting good features, transforming raw measurements that are collected is very important.

Measurements are no Features



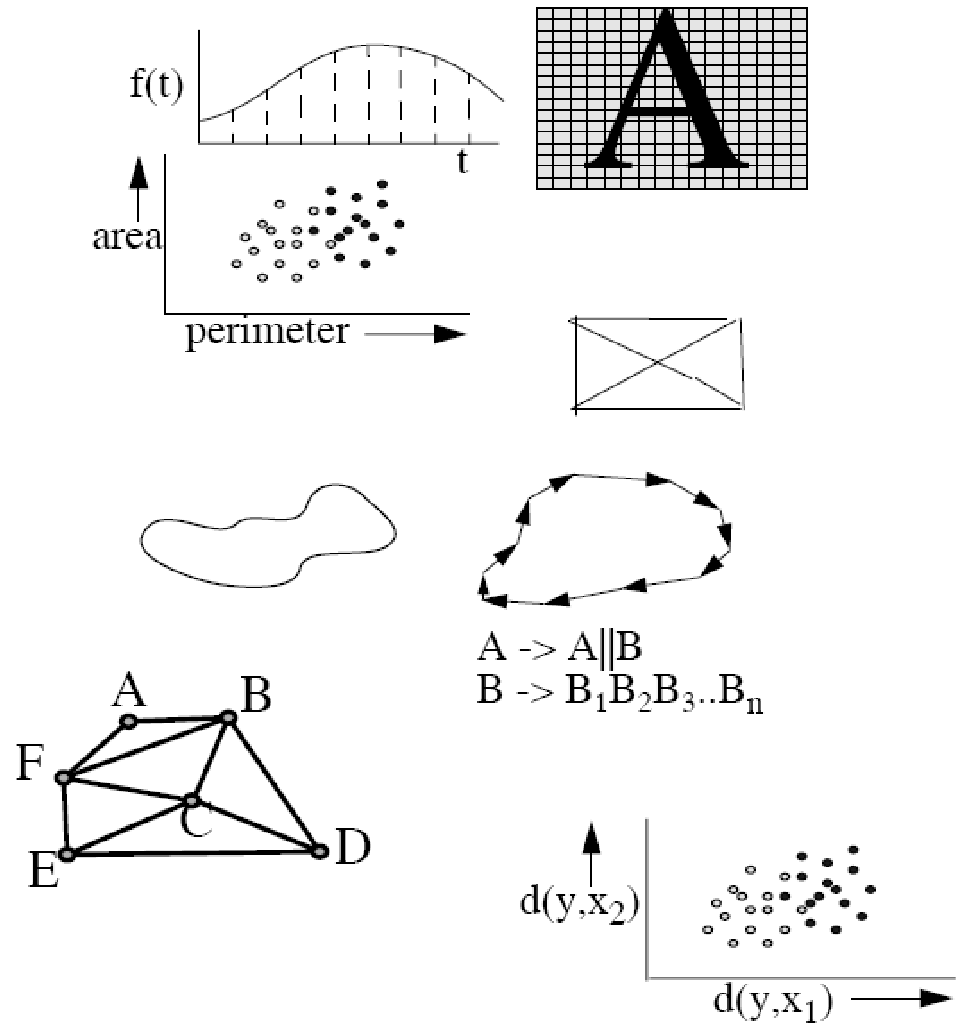
16 x 16 = 256 measurements *describe* the object
Number of *features* (e.g. moments, endpoints, strokes, holes) may be much smaller.
They *represent* the object.
(Duin)

Data representation

- Traditional algorithms work on vectors.
- Images can be represented as matrices or vectors.
- Abstract data
  - Graphs
  - Sequences
  - 3D structures

Possible Object Representations

- Measurement samples
- Feature vector
- Sets of segments or primitives
- Outline samples (shape)
- Symbolic structures
- (Attributed) graphs
- Dissimilarities



Duin etc.

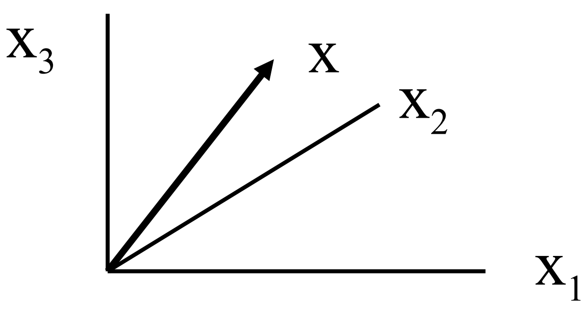
Feature space representation

- Representation: mapping objects into vectors,
  $\{O_i\}$ => $X(O_i)$, with $X_j(O_i)$ being *j*-th attribute of object O_i

- "attribute" and "feature" are used as synonyms

- Types of features.
  Categorical: symbolic or discrete – may be nominal (unordered), like "sweet, salty, sour", or ordinal (can be ordered), like colors or small < medium < large (drink).
  Continuous: numerical values.

Vector $X = (x_1, x_2, x_3 \ldots x_p)$,
or a p-dimensional point
in the feature space.



- Features representing the same entity are combined in the feature vector

  $\boldsymbol{v} = [v_1\ v_2 \ldots v_p]^T$

  p very important since
  - it determines the computational effort
  - has a strong impact on the requirements on the learning set size (also statistical significance)

- ***v*** points to one certain point in the *p*-dimensional space ***V***
- Every point in measurement space ***V*** corresponds to one of the possible constellations of the data

- Different type of features may be arbitrarily mixed.
- Ordering scheme of the components of ***v*** is arbitrary but must be fixed

- Which features to choose -> design decision
- It should be decided by human insight into the field of application
- Features should have the potential of giving hints as to which class the observed event belongs
- How to extract powerful (?) feature sets from larger sets of feature candidates?
- Discriminative power of a feature set can be improved by properly chosen transformations, e.g. Normalization
- **Remark** measurements are just a set of samples: formed as a result of random selection of some representatives of the set

Methods

Statistics

- **Statistics** is a mathematical science pertaining to the
  - collection,
  - analysis,
  - interpretation or explanation, and
  - presentation of data.
- Statistical methods can be used to summarize or describe a collection of data; this is called **descriptive statistics**.
- Patterns in the data may be modeled in a way that accounts for randomness and uncertainty in the observations, to draw inferences about the process or population being studied; this is called **inferential statistics**.
- Both descriptive and inferential statistics can be considered part of **applied statistics**. There is also a discipline of **mathematical statistics**, which is concerned with the theoretical basis of the subject.
- The word ***statistics*** is also the plural of ***statistic*** (singular), which refers to the result of applying a statistical algorithm to a set of data, as in employment statistics, accident statistics, etc.

Statistics

- In applying statistics to a problem, one begins with a process or population to be studied.
- This might be a population
  - of people in a country,
  - of crystal grains in a rock, or
  - of goods manufactured by a particular factory during a given period.
- It may instead be a process observed at various times; data collected about this kind of "population" constitute what is called a time series.
- For practical reasons, rather than compiling data about an entire population, one usually instead studies a chosen subset of the population, called a sample.
- Data are collected about the sample in an observational or experimental setting.
- The data are then subjected to statistical analysis, which serves two related purposes: description and inference.

Statistics

- Descriptive statistics can be used to summarize the data, either numerically or graphically, to describe the sample.
  - Basic examples of numerical descriptors include the mean and standard deviation.
  - Graphical summarizations include various kinds of charts and graphs.
- Inferential statistics is used to model patterns in the data, accounting for randomness and drawing inferences about the larger population.
- These inferences may take the form of
  - answers to yes/no questions (hypothesis testing),
  - estimates of numerical characteristics (estimation),
  - forecasting of future observations,
  - descriptions of association (correlation), or
  - modeling of relationships (regression).
  - Other modeling techniques include ANOVA, time series, and data mining.

- Univariate
- multivariate

Summary of a multivariate data set

- Summaries for each of the variables separately
- Summaries for the relationships between (pair of) variables

Summaries for each of the variables separately

- Mean
- Variance

Summaries for the relationships between (pair of) variables

- Variance
- Covariance
- Correlation

Analysis of Data

- Exploratory Analysis
  - exploration: attempts to recognize any *non-random* pattern or structure
  - mining: generates possible interesting *hypotheses* for further study
- Confirmatory Analysis
  - After well-defined hypothesis in mind
  - Some type of (well-known) *significance test*

- Search for *structure or pattern* in the data
- If pattern arises from the fact that we have measurements on similar group of subjects
  - *unsupervised pattern recognition or unsupervised learning*
  - But
    - What are these groups ?
    - How many groups are there?
    - Which subject belongs to which group?

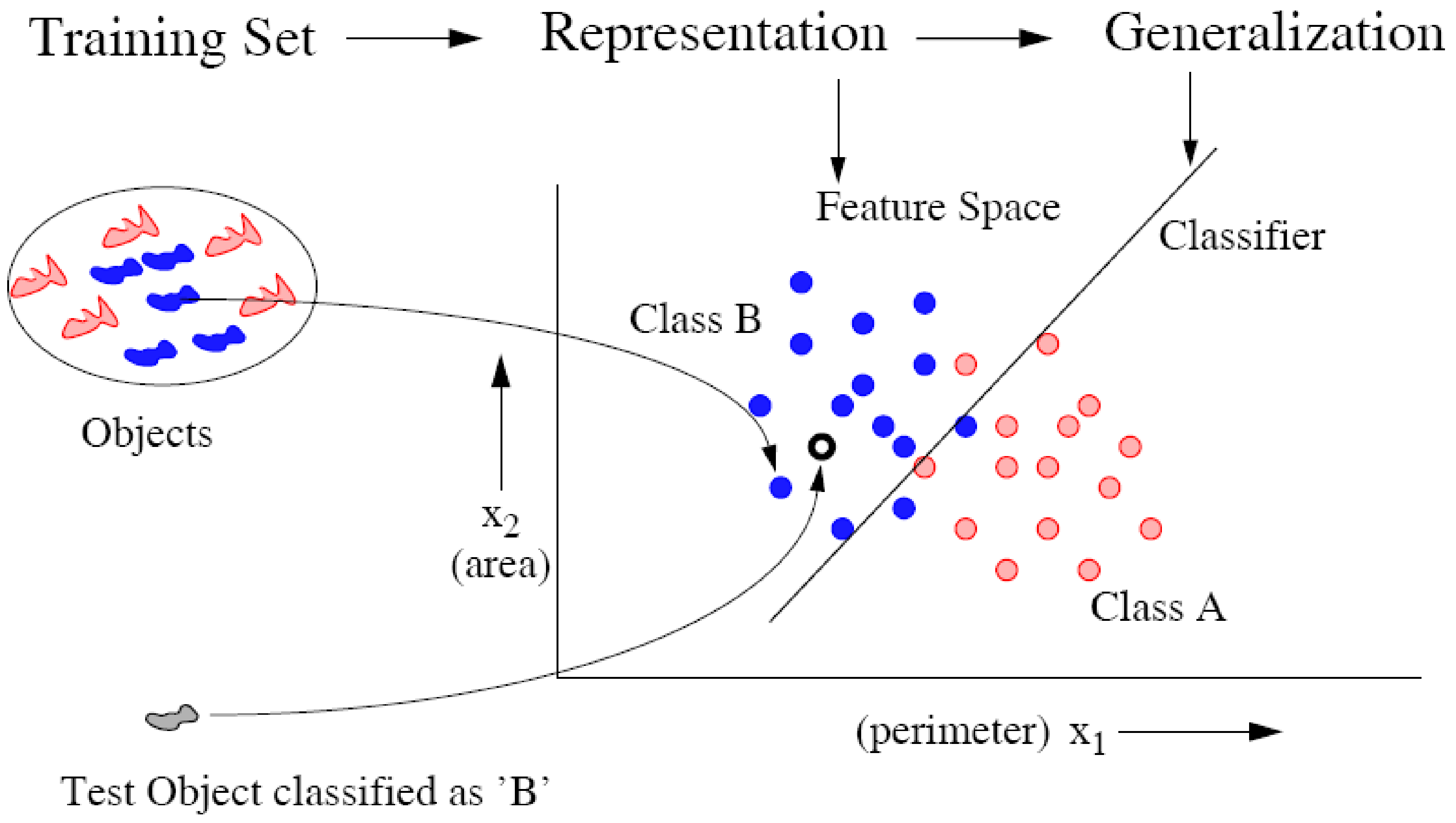
- Other motivations
  - Find latent variables
  - Supervised learning
  - Regression
- But, as usual no systematic methodology

What is machine learning?

- Machine learning is the study of computer systems that improve their performance through experience.
  - Learn existing and known structures and rules.
  - Discover new findings and structures.
    - Face recognition
    - Bioinformatics

- Supervised learning vs. unsupervised learning

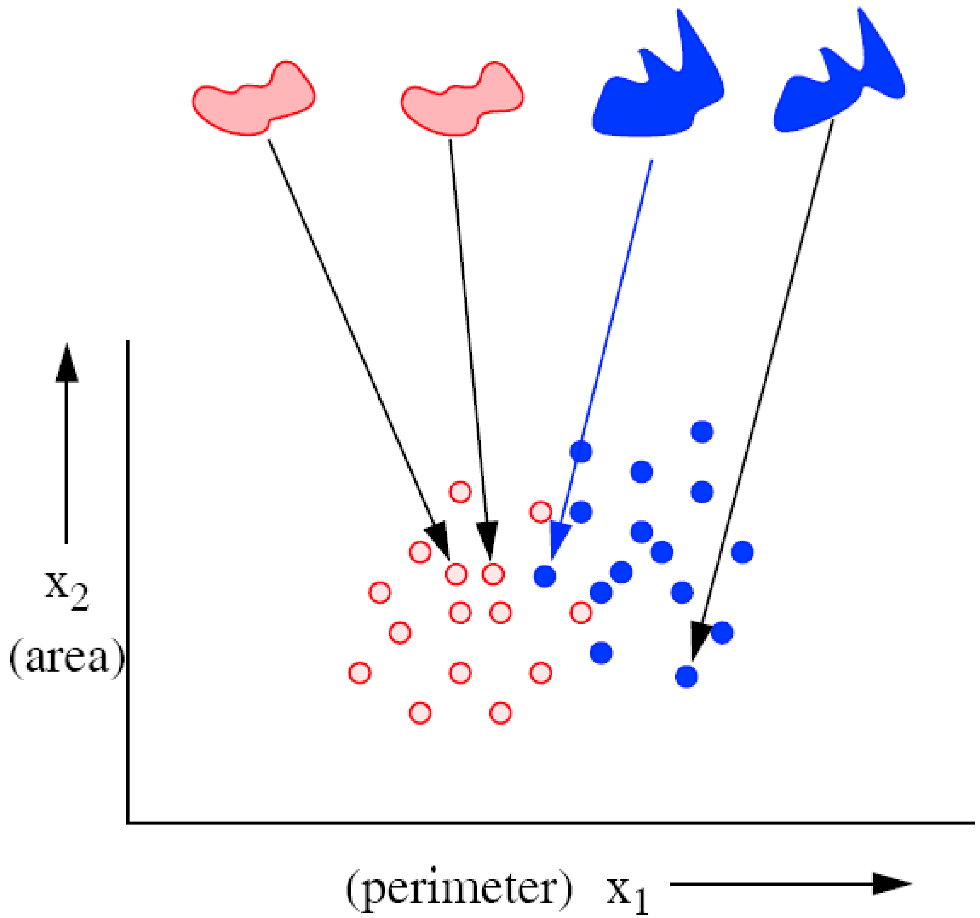
- Semi-supervised learning



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

Duin etc.

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

Machine learning applications

- Bioinformatics: Hugh amount of biological data from the human genome project and human proteomics initiative.
  - Goal: Understanding of biological systems at the molecular level from diverse sources of biological data.
  - Challenge: Scalability, multiple sources, abstract data.
  - Applications: Microarray data analysis, Protein classification, Mass spectrometry data analysis, Protein-protein interaction.

- Others: Computer vision, information retrieval, image processing, text mining, web mining, etc.

Supervised vs. unsupervised learning

- **Although unsupervised learning methods may appear to have limited capabilities, there are several reasons that make them extremely 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)
- **The supervised and unsupervised paradigms comprise the vast majority of pattern recognition problems**
  - A third approach, known as reinforcement learning, uses a reward signal (realvalued or binary) to tell the learning system how well it is performing
    - In reinforcement learning, 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

Curse of dimensionality:

- The problem occurs when searching in or estimating density on high-dimensional spaces.
  1. Computation: The complexity grows exponentially with the dimension, rapidly outstripping the computational and memory storage capabilities of computers.
  2. Estimation: The problem of estimating a density function on a high-dimensional space may be seen as determining the density at each cell in a multidimensional grid. Given a fixed number of K grid lines per dimension, the number of independent cells grows as KP where P is the dimension.

Curse of dimensionality

- Large sample size is required for high-dimensional data.

- Query accuracy and efficiency degrade rapidly as the dimension increases.

- Strategies
  - Feature reduction
  - Feature selection
  - Manifold learning
  - Kernel learning