Machine Learning Techniques in Bio-Genetic Applications

Marco Botta
Dipartimento di Informatica
Università di Torino
www.di.unito.it/~botta/didattica/
botta@di.unito.it

Summary

- Formalization of "learning problems" in Bio-Genetic and Bio-Medical applications.
- Overview of the available learning techniques.
- Multiple strategies integrated approach.
- Some examples of applications.

Three fundamental issues...

Representation

how to represent available information in order to be processed by the learning algorithms?

Definition of the learning problem

which task do we have to solve?

Approach

which algorithm is more suitable to the problem?

Slides by Botta

Representation: *the main problem*

Attribute-value representation

D	Temp	Pres
1	37	125
2	36	120
3	39	195
4	36	140
5	40	180
6	37	135
7	38	170
8	36	130
9	37	120
10	39	135
11	36	115

We want a program able to recognize “anomalous” cases

Multiple instance representation

D	d	Temp	Pres
1	1	37	125
1	2	36	120
1	3	37	118
2	1	38	130
2	2	37	130
2	3	39	170
2	4	38	140
2	5	37	135
3	1	36	115
3	2	37	120
4	1	36	118
4	2	37	120
4	3	39	190
4	4	40	180
4	5	36	115
4	6	37	118

Cases in which there is at least one “anomalous” record

Structured representation

D	d	Temp	Pres
1	1	37	125
1	2	36	120
1	3	37	118
2	1	38	130
2	2	39	133
2	3	39	170
2	4	38	140
2	5	37	135
3	1	36	115
3	2	37	120
4	1	36	118
4	2	37	120
4	3	39	140
4	4	40	180
4	5	36	115
4	6	37	118

Cases in which there are record with similar temperature but different blood pressure

Data Mining Tasks

- Classification
- Functional dependencies / Regression
- Clustering / Segmentation
- Summary/ Characterization
- Association Discovery / Causality
- Anomaly detection
- Analysis of temporal series

Classification and Regression

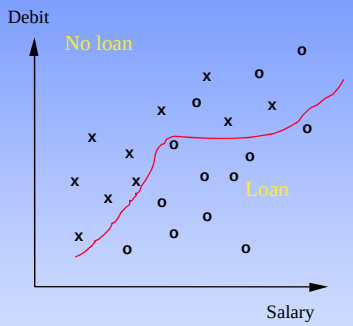
Classification:
forecast the value of
a categoraical attribute

D	Temp	Pres	C
1	37	125	N
2	36	120	N
3	39	195	P
4	36	140	N
5	40	180	P
6	37	135	N
7	38	170	P
8	36	130	N
9	37	120	N
10	39	135	N
11	36	115	N

Regression:
forecast the value of
a numeric attribute

D	Temp	Pres	Fr
1	37	125	65
2	36	120	68
3	39	195	120
4	36	140	80
5	40	180	125
6	37	135	70
7	38	170	195
8	36	130	75
9	37	120	60
10	39	135	85
11	36	115	55

Classification



Tipical problems

Fraud detection
Loan allowance
Fault detection

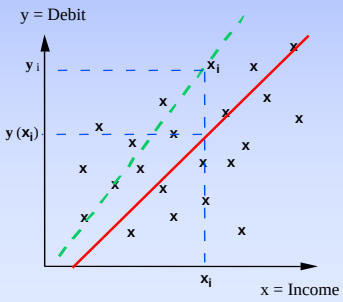
Slides by Botta

Functional dependencies / Regression

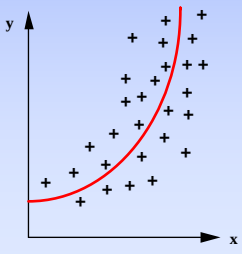
Discovery of functional links among variables in a database

Given a set E = $\{e_1, \ldots, e_n\}$ of elements
described by attributes A = $\{x_1, \ldots, x_k\}$, a regression task
assign to every element e_i of the set E a value of a continuous variable f

$$\forall\, e_i : \quad f = f(x_1, \ldots, x_k)$$

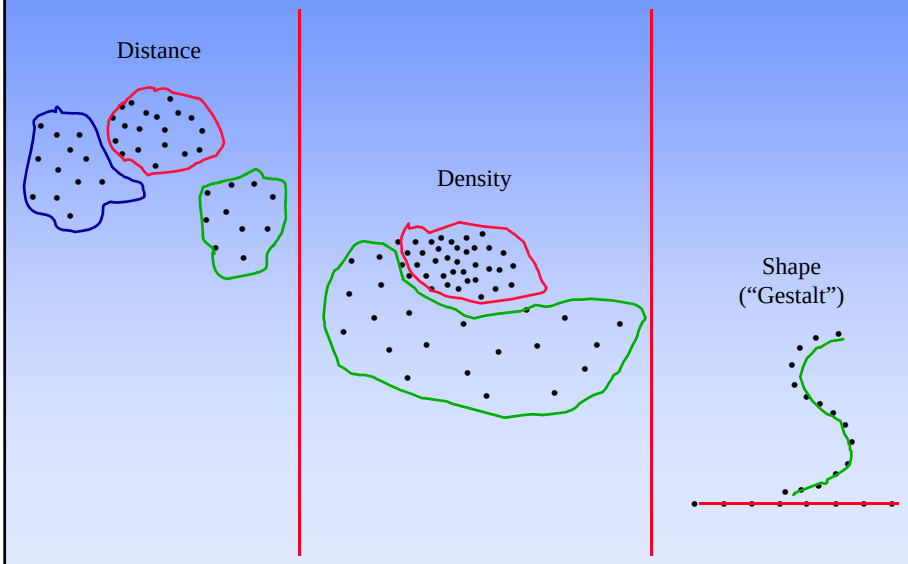


$$\sum_{i=1}^{n} [y_i] - f_{\beta}(x_i)]^2$$



Slides by Botta

Clustering



Slides by Botta

Segmentation

Segmentation is the analysis of actual or potential groups of customers ("segments") with the aim to find features and behaviorsc that can be exploited.

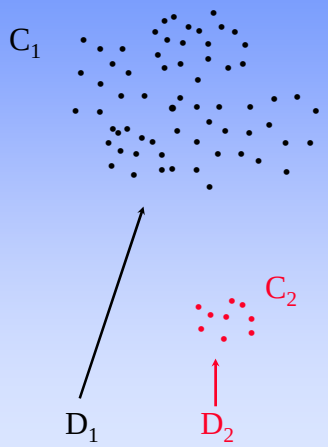
Segmentation allows an organization toconsider, in the limit, each customer as a "segment of one", in order to establish an extremely personalized relation..

There are two basic problems in marketing

- To understand why customers leave (so called "customer attrition")
- To discover new markets ("target marketing" and "cross selling")

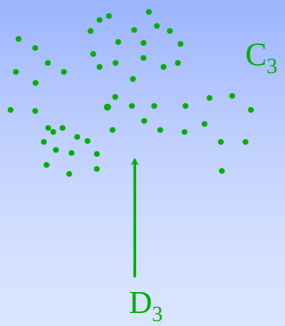
Slides by Botta

Summary / Characterization



Why have they been grouped together?

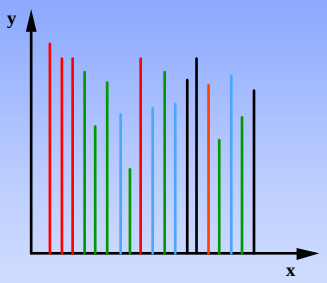
What do they have in common?



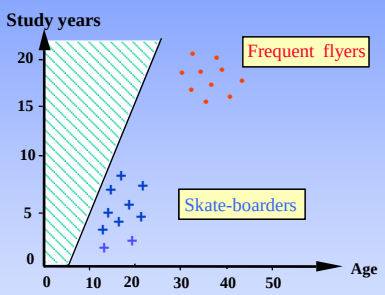
Slides by Botta

Summary / Characterization

Compact Description of a set of data



Mean average
Standard deviation



• Middle Age People with University degrees
+ Youngers with low instruction level

Slides by Botta

Association discovery

Discovery of associations among facts, properties or values of variables (“Link analysis”)

72% of buyers of green salad, also buy a sauce

Tipical Problems

Market Basket Analysis

1/7/99
Bread
Peaches
Eggs
Spaghetti
… …

ticket

2/7/99
Rice
Bread
Meat
Peaches
… …

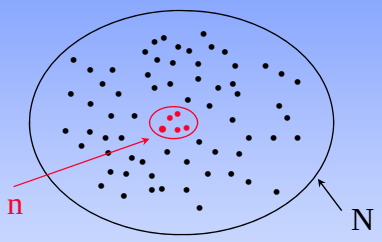
ticket

… … … {Bread, Peaches}

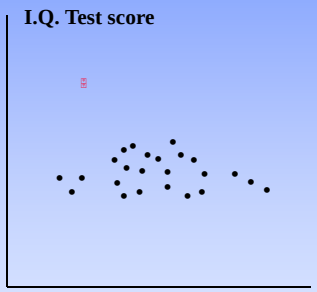
Slides by Botta

Anomaly detection

Detection of values deviating from “normality”
(Exceptions, Rare cases, Errors)



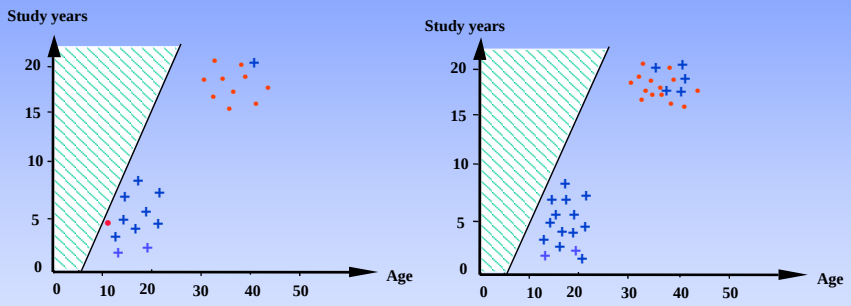
n << N



Slides by Botta

Anomaly detection

Detection of values deviating from “normality”

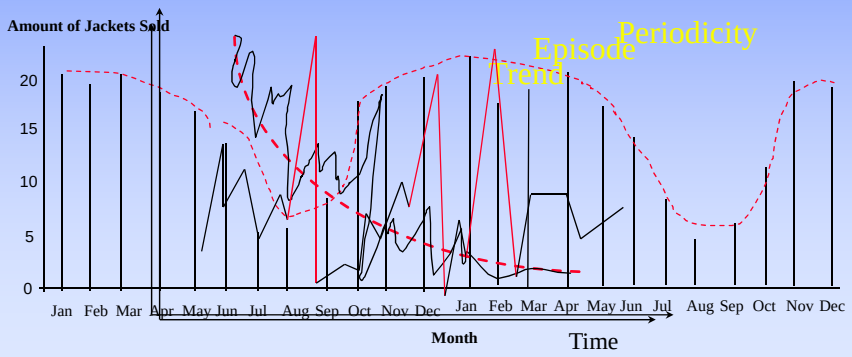


Need a definition of “normality”

Slides by Botta

Temporal series analysis

- Discovery of interesting conformations or episodes
- Trend analysis
- Discovery of periodicity or “seasonal” phenomena



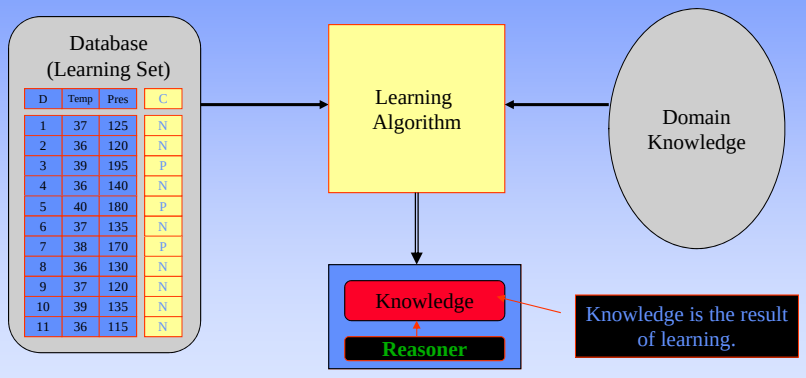
Slides by Botta

Relevant Discipline for Data Mining

- Statistics
- Pattern Recognition
- Artificial Intelligence
  - Machine Learning
  - Bayesian Nets
  - Intelligent Agents
- Databases
  - Query and Reporting
  - "Data Warehousing" → OLAP
- Visualization
  - Graphics
  - Multi-media environments
- Cognitive Sciences

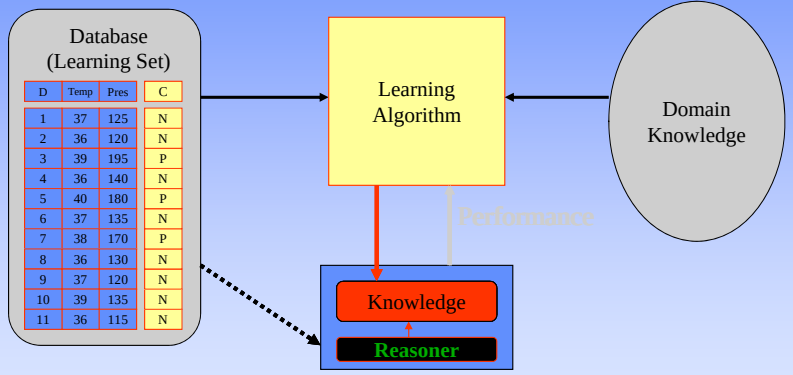
Slides by Botta

Supervised Learning



Slides by Botta

Learning as “Hypothesizing and Testing”

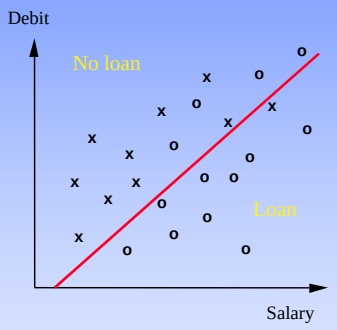


D	Temp	Pres	C
1	37	125	N
2	36	120	N
3	39	195	P
4	36	140	N
5	40	180	P
6	37	135	N
7	38	170	P
8	36	130	N
9	37	120	N
10	39	135	N
11	36	115	N

Slides by Botta

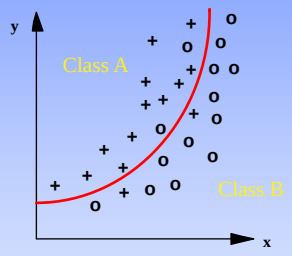
Statistics: Discriminant Analysis

Discriminant Function



Linear

Loan : $y - a x - b < 0$



Non Linear

Class A : $y - a x^2 - b > 0$

Slides by Botta

Statistics: Clustering

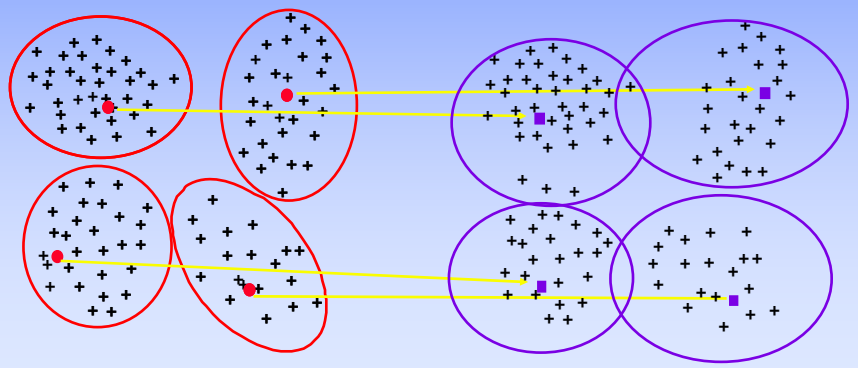
k-Means Algorithm

The number K of clusters is a user choice

distance Function

objective function to optimize:

Maxmize inter-cluster distance and minimize intra-cluster distance



Slides by Botta

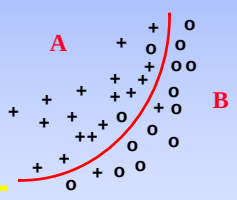
Pattern recognition: Classification

Statistical Approach

Probabilistic function

Bayesian Classifier

Maximum likelihood classifier



$$\Pr\{A \mid \vec{x}\} = \frac{\Pr\{\vec{x} \mid A\}\ P(A)}{\Pr\{\vec{x} \mid A\}\ P(A) + \Pr\{\vec{x} \mid B\}\ P(B)}$$

$$\Pr\{B \mid \vec{x}\} = \frac{\Pr\{\vec{x} \mid B\}\ P(B)}{\Pr\{\vec{x} \mid A\}\ P(A) + \Pr\{\vec{x} \mid B\}\ P(B)}$$

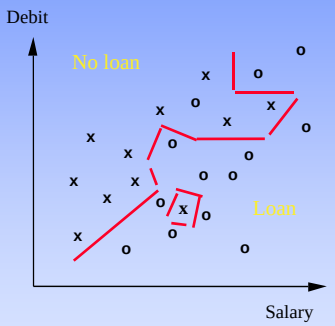
$$\begin{cases} \Pr\{A \mid \vec{x}\} = \Pr\{B \mid \vec{x}\} \\ \Pr\{A \mid \vec{x}\}\, P(A) = \Pr\{B \mid \vec{x}\}\, P(B) \end{cases}$$

Slides by Botta

Pattern recognition: Classification

"Case-Based"/Instance-based approach

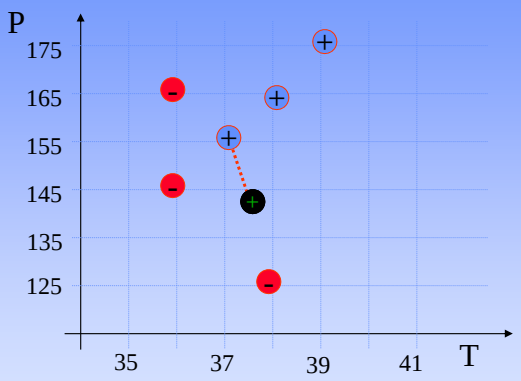
k-Nearest Neighbours



Slides by Botta

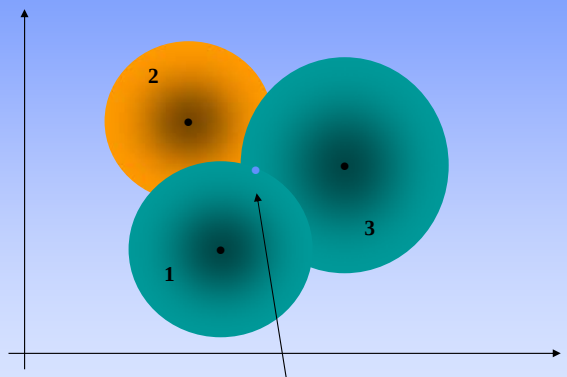
Instance Based Learning

$T=37^0$, P=155	+
$T=38^0$, P=165	+
$T=36^0$, P=145	-
$T=36^0$, P=165	-
$T=39^0$, P=175	+
$T=38^0$, P=125	-



Slides by Botta

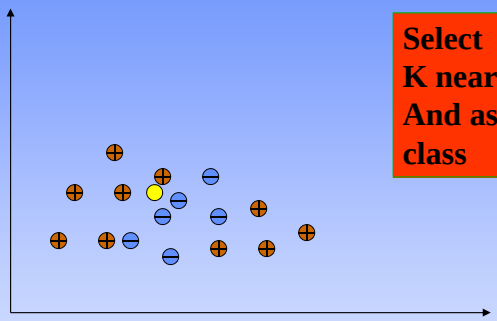
Basic idea



A new instance is influenced by those fields around it and is labeled accordingly

Slides by Botta

k-NN: for classification



Select
K nearest neighbors
And assign the majority class

K = 1 => class = +
K = 3 => class = -

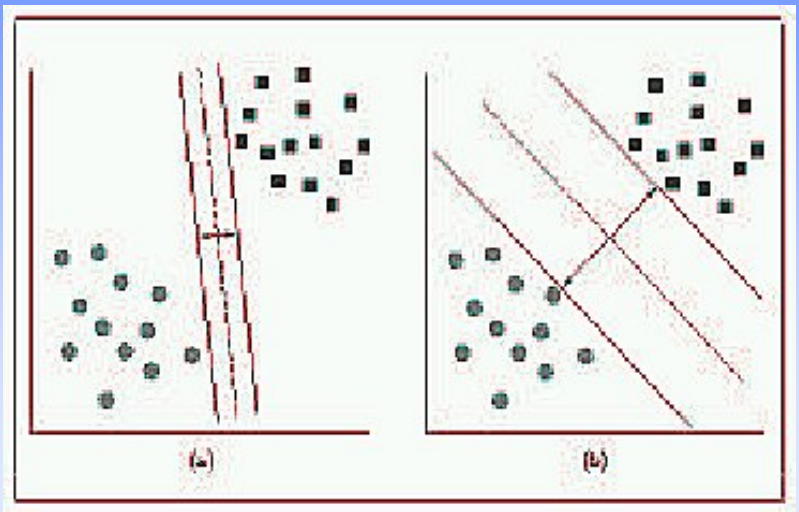
Slides by Botta

Support Vector Machines

- Given a set of data points belonging to 2 classes, an SVM finds the hyperplane that :
  - Keep most data points of the same class in the same semi-space
  - and maximizes the distance among the data points and the hyperplane

Slides by Botta

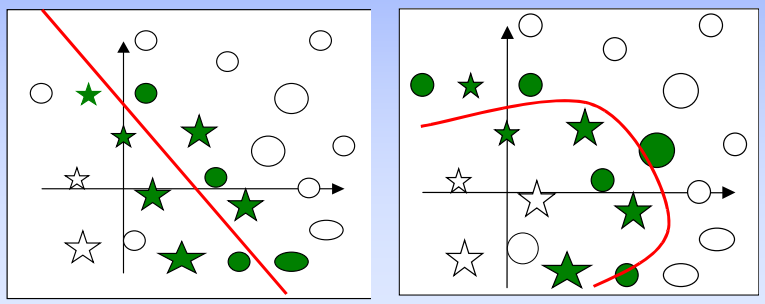
SVM: basic idea



Slides by Botta

Support Vector Machine

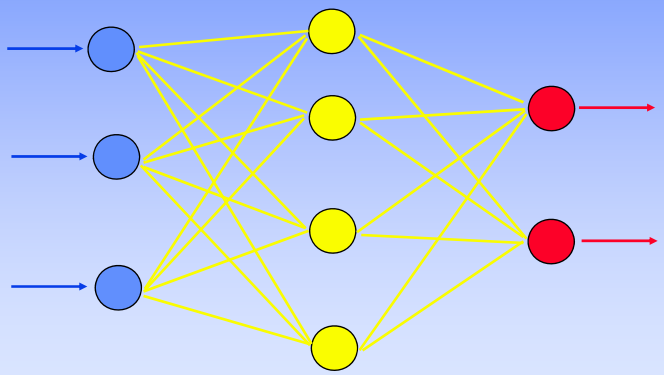
- The data points close to the hyperplane are called Support Vectors



Slides by Botta

Neural Nets

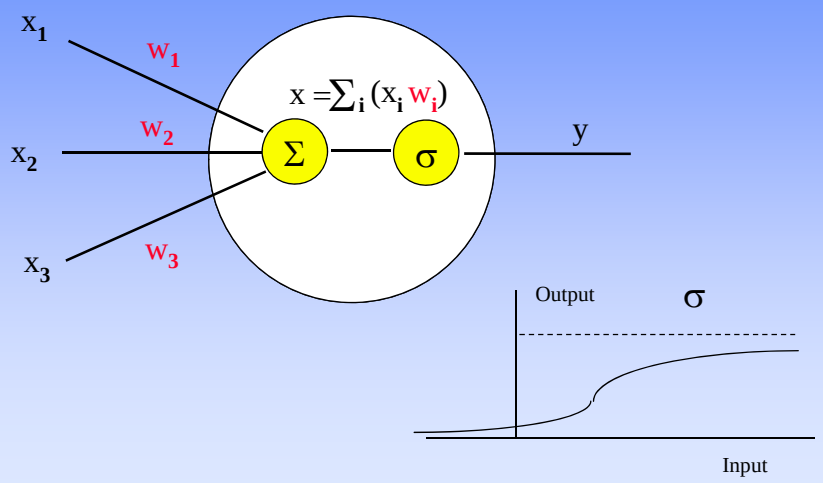
A neural net is a compound structure, formed by simple computational elements, connected according to a multilayer topology => Universal function approximators



Input units Hidden units Output units

Slides by Botta

Neural Nets: Elementary functions

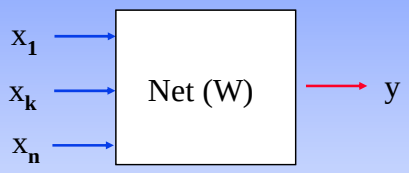


Neural Nets: Learning

"Backpropagation" Algorithm

Minimize the total quadratic error

If the net if multilayered, the error is back-propagated



$$E = \frac{1}{2}\sum_{k=1}^{n}(t_k - y_k)^2 \qquad w_j = -\eta \frac{\partial E}{\partial w_j}$$

η = Learning rate

Artificial Intelligence: Symbolic Learning

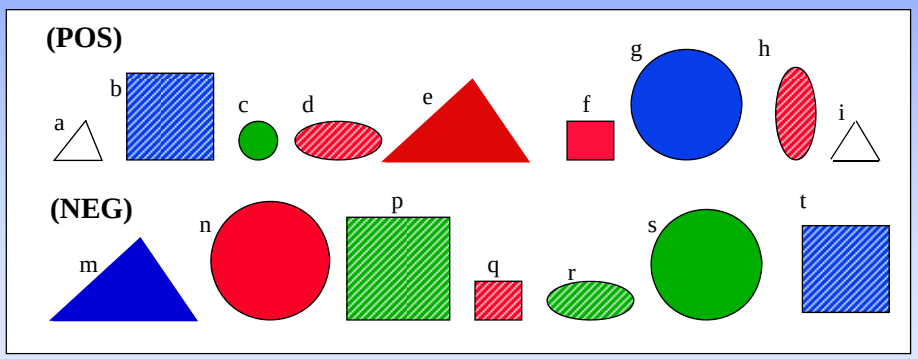
- Decision Trees
- Production Rules
- Bayesian Nets
- Conceptual Clustering

Slides by Botta

Decision trees : an example

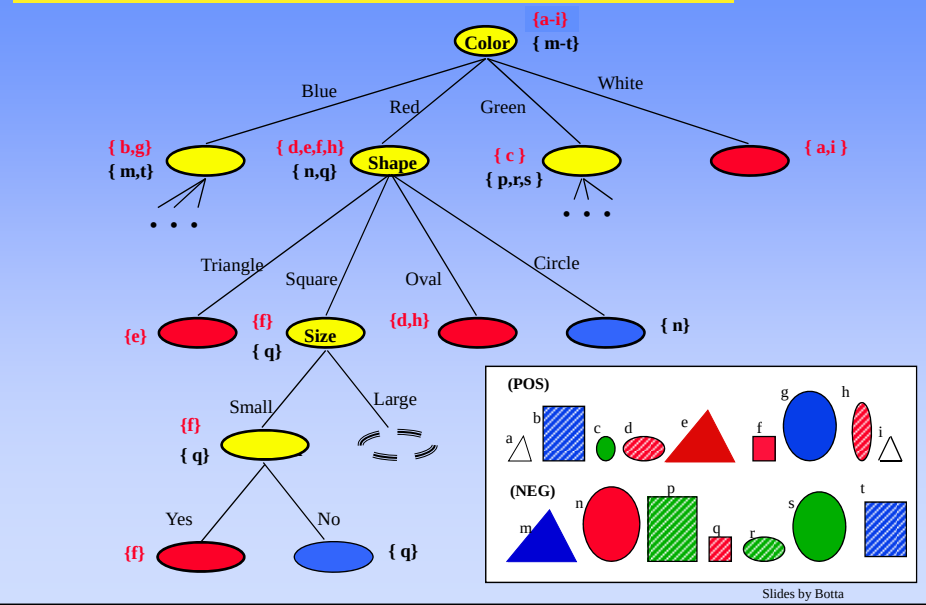
Attributes

Color = {Red, Blue, Green, White}
Shaded = {Yes, No}
Shape = {Square, Triangle, Circle, Oval}
Size = {Small, Large}



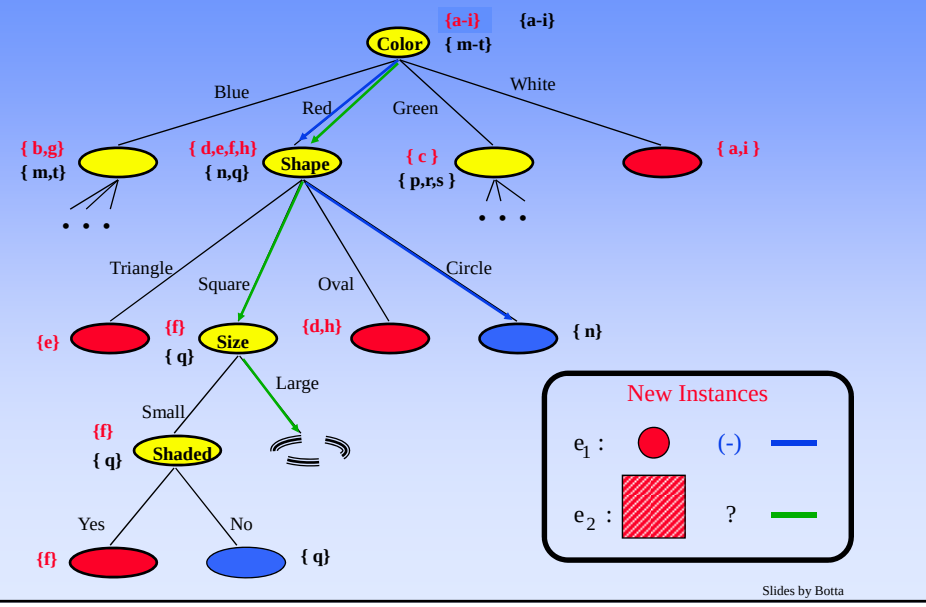
Slides by Botta

Decision trees: learning



Slides by Botta

Decision trees: Classification



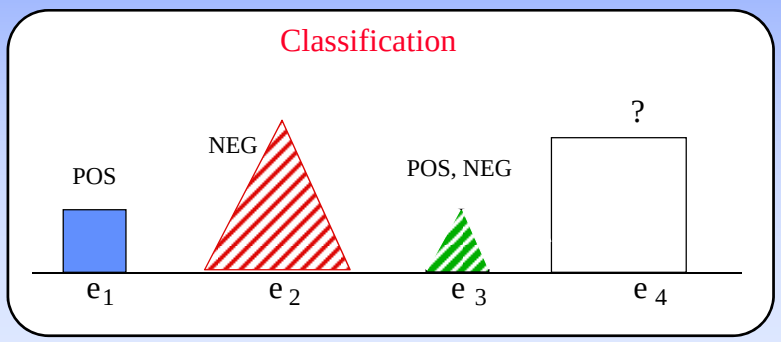
Slides by Botta

Production Rules

Decision rules represented in a logical format:
Propositional Calculus or First Order Logic

$$(\text{shape} = \text{square} \vee \text{triangle}) \wedge (\text{size} = \text{small}) \Rightarrow \text{POS}$$

$$(\text{shape} = \text{triangle}) \wedge (\text{shaded} = \text{YES}) \Rightarrow \text{NEG}$$



Slides by Botta

Genetic Algorithms

- Genetic Algorithms are a general stochastic search method
- Inspired to the Darwinian theory of evolution
- Used both in symbolic learning and neural nets approaches

Ingredients

- Population of solutions (Chromosomes)
- "Fitness" function
- Genetic Operators ("Crossover" e Mutazione)

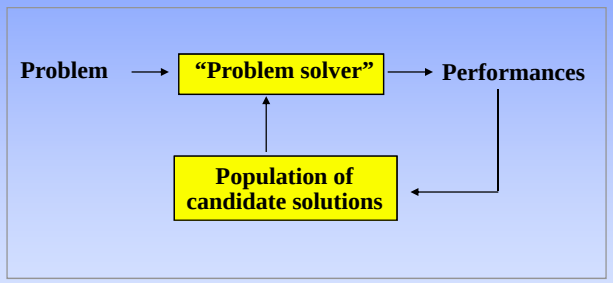
Basic Cycle

- Select from population a set of individuals to be reproduced, proportionally to their fitness
- Selected individuals mate and generate 2 offsprings by applying the crossover operator
- Mutation operator is possibly applied to the offsprings
- The new individuals replace the older population

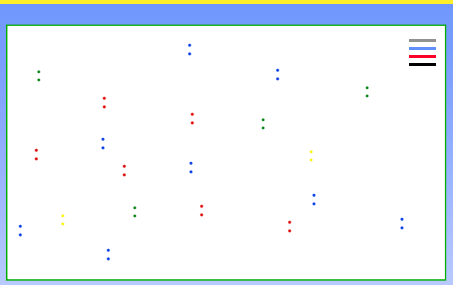
Slides by Botta

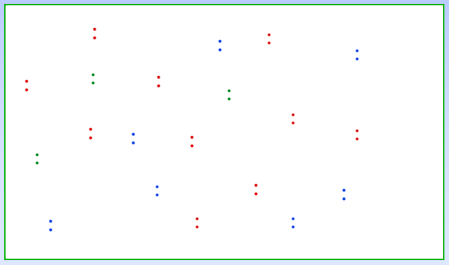
Genetic Algorithms: basic idea

The population of candidate solutions improve at each cycle

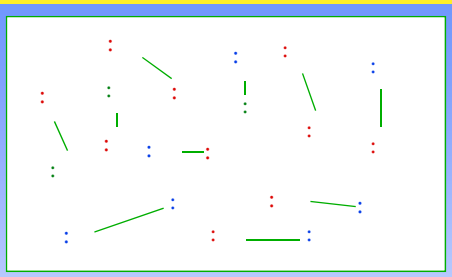


Genetic Algorithms : Selection

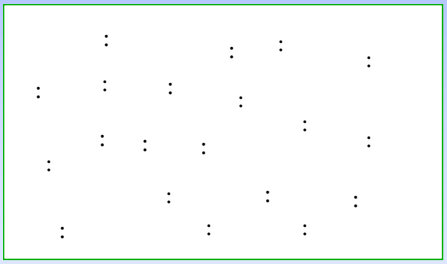




Genetic Algorithms : Reproduction

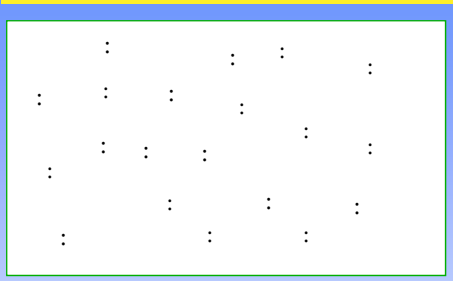


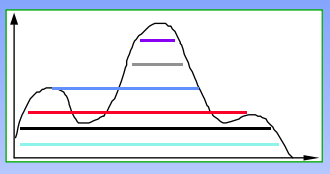
Reproduction



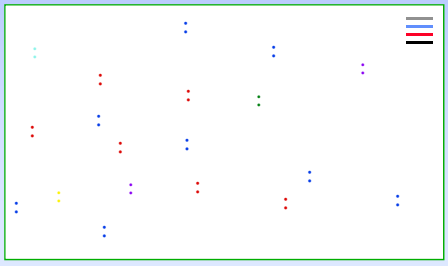
Slides by Botta

Genetic Algorithms : Fitness evaluation



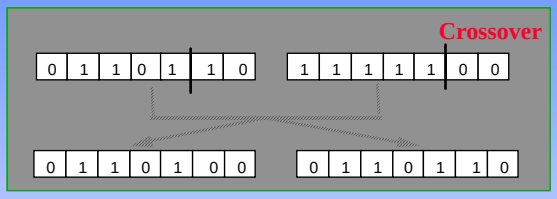


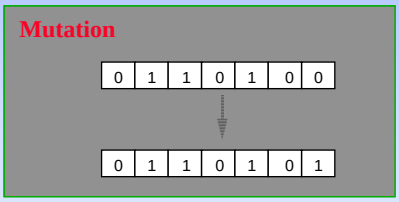
Evaluation



Slides by Botta

Genetic Algorithms : Genetic operators





Slides by Botta

Association Rules

Let I be a set of items

Let D be a set of records, each containing a subset of I

Association Rule:

r: X => Y

X and Y are disjoint subsets of I

Support of a subset Z of I: supp(Z) = |D(Z)|/|D|

Confidence of a rule: conf(r) = supp(X or Y)/supp(X)

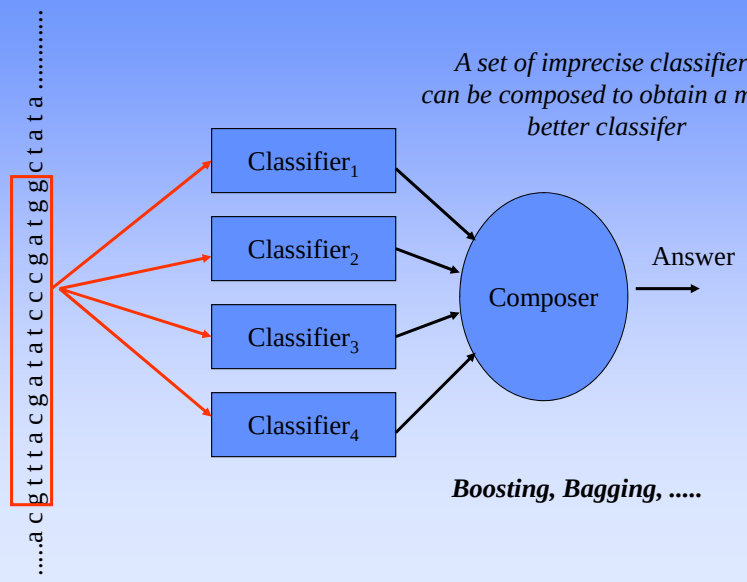
Slides by Botta

Association Rules: *Apriori* Algorithm

- *Apriori* Algorithm
  - Phase 1 => Look for all frequent itemsets
    - Incremental construction starting from cardinality 1
    - Generation of cardinality k candidate itemsets starting from frequent set of cardinality (k-1)
    - Retain only frequent intemsets
  - Phase 2 => Extraction of all possible rules from each frequent itemset

Slides by Botta

Composite Classifiers



Slides by Botta

A biologic problem used as Machine Learning benchmark Splice-Junctions prediction in DNA sequences

- Daa taken from Genbank 64.1 (ftp site: genbank.bio.net) (date back to 1992)
- 3190 sequences in the dataset
- 3 categories:
  - ”ei” (767) and ”ie” (768) include every "split-gene" in Genbank 64.1
  - “n” (1655) non-splice sequences do not include a “splicing site”

Slides by Botta

Splice-Junctions prediction in DNA sequences

- Learning problem defined as: given a position in the middle of a 60 DNA base pairs window decide whether
  - a) it is an "intron -> exon" junction (ie)
  - b) it is an "exon -> intron" junction (ei)
  - c) neither of above (n)

Slides by Botta

Splice-Junctions prediction in DNA sequences

- Propositional Representation with 62 attributes:
  - 1 class {n ei ie} of the sequence
  - 2 sequence name
  - 3-62 60 attributes contain DNA basis, in positions from -30 to position +30 with respect to the splice site.

ATRINS-DONOR-905,
A,G,A,C,C,C,G,C,C,G,G,G,A,G,G,C,G,G,A,G,G,A,C,C,T,G,C,A,G,G,G,T,G,A,G,C,C,C,C,A,C,C,G,C,
C,C,C,T,C,C,G,T,G,C,C,C,C,C,G,C, EI

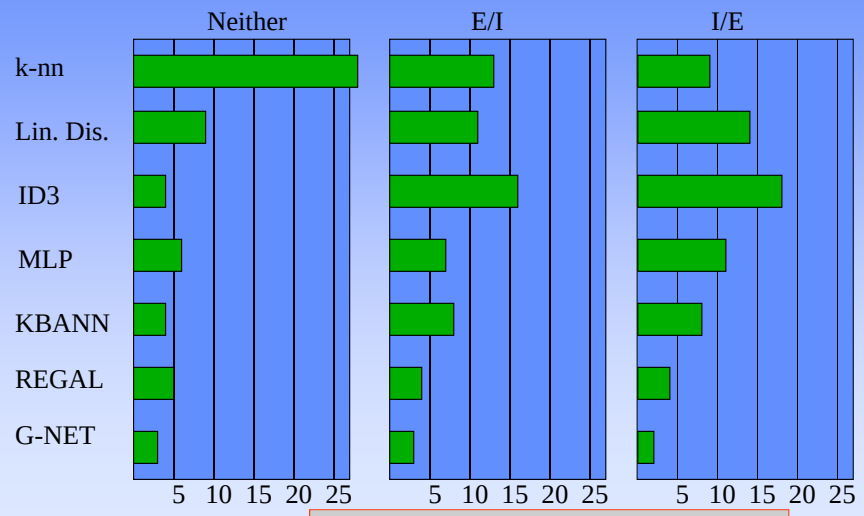
Slides by Botta

Value Distribution

Base	Neither	EI	IE
A	24.984%	22.153%	20.577%
G	25.653%	31.415%	22.383%
T	24.273%	21.771%	26.445%
C	25.077%	24.561%	30.588%
D	0.001%	--	0.002%
N	0.010%	0.010%	--
S	--	--	0.002%
R	--	--	0.002%

Slides by Botta

Splice-Junctions Prediction

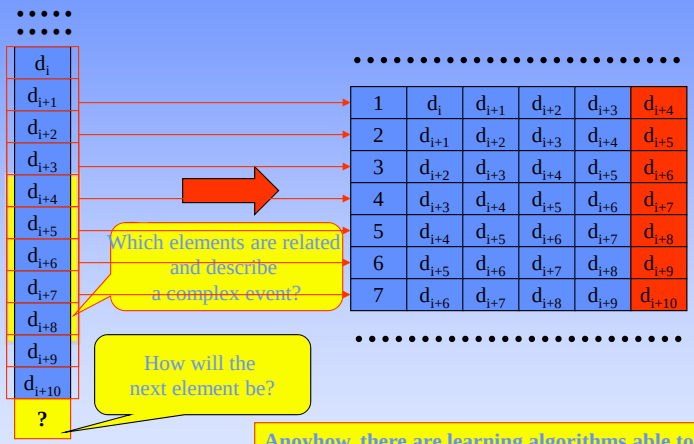


On a dataset of 3600 instanc

On other data taken from Gene-Bank….!?!

Slides by Botta

In General: prediction in a sequence



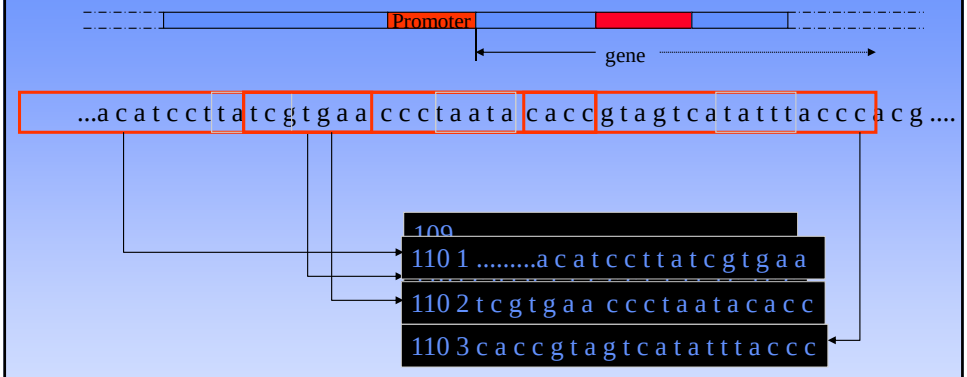
Which elements are related and describe a complex event?

How will the next element be?

Anoyhow, there are learning algorithms able to directly deal with sequences.

Slides by Botta

An Example.....



Slides by Botta

Protein-protein interaction prediction from primary sequence with SVM

...S K I I N F E D L T…

Amminoacidic index

3.8 5.6 4.1 4.1 2.0 5.5 1.0 3.4 3.8 2.2

4.7 4.1 2.2 4.0

Range Standardization

input to **SVM** (Bock e Gough, 2001)

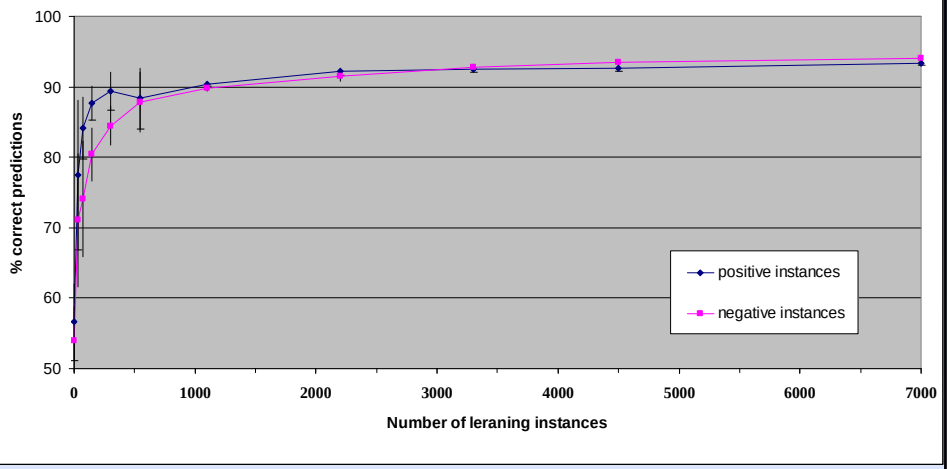
Slides by Botta

Protein Interaction Databases

- BIND contains interactions among proteins, proteins and nucleic acids, and simple molecules
- MINT mainly contains protein-protein interactions
- DIP is the richest database of protein-protein interactions (more than 13500 at 28/6/2002) and is continuously and rapidly growing

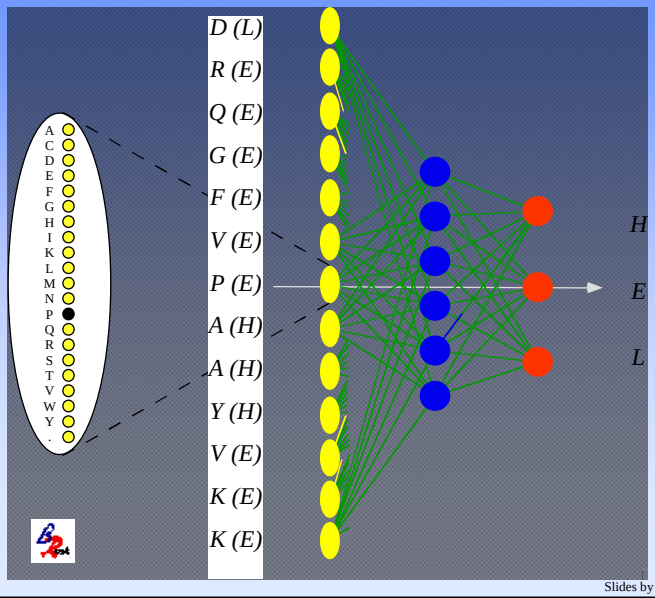
Slides by Botta

SVM Results



Slides by Botta

Neural Network for secondary structure Prediction



Slides by Botta

Available Software

- Symbolic Learning and more: Weka Suite
  - http://www.cs.waikato.ac.nz/~ml/weka/
- Neural Networks:
  - http://www.emsl.pnl.gov:2080/proj/neuron/neural/systems/sharewar e.html
- Genetic Algorithms
  - G-net: http://hermes.mfn.unipmn.it/~attilio/PROJECTS/GNET/gnet.html
  - GALib: http://lancet.mit.edu/ga/

Slides by Botta

Conclusions

Machine Learning can be seen as the integration of several approaches camoing from different disciplines.

Learning from structured data and sequences is an emerging issue that is crucial in biogenetic and biomedical applications nel settore bio-medico.

Two factors are fundamental for having success:

- Work in teams that combine both computer science and biological knowledge.
- Know how to integrate different methodologies in different programs

Slides by Botta