Statistical Methods, Multivariate Analysis and Pattern Recognition in HEP

Liliana Teodorescu

Brunel UNIVERSITY WEST LONDON

HEP data analysis

- *more and more complex*
- *open to methods and techniques from other fields*

More and more terminology imported from other fields dealing with analysis of data

- ✓*multivariate analysis*
- ✓*machine/statistical learning*
- ✓*pattern recognition*
- ✓*discriminant analysis*
- ✓*neural networks*
- ✓*boosted decision trees*
- ✓*genetic algorithms/programming*

Do we use them with their actual meaning?
Are these terms completely distinctive?
What do we actually use out of all the methods/techniques designated by these terms?

Outline

"We performed two analyses to separate signal from background
- a cut based analysis
- a multivariate analysis (using neural networks or boosted decision trees etc.)"

Multivariate analysis – very fashionable

Outline of the lecture

Multivariate statistical data analysis
- ✓ *definition and benefits*
- ✓ *preparation of data*
- ✓ *generic data analysis problems*
- ✓ *analysis techniques – examples used in HEP*

Pattern recognition *– links with MVA*

Conclusions

Multivariate analysis

Multivariate analysis (MVA) (broad definition)

❖ *set of statistical analysis methods that simultaneously analyse multiple measurements (variables) on the object studied*

Multiple - number of var >2

Univariate analysis - number of var = 1

Bivariate analysis - number of var = 2

Truly MVA (other definitions)

❖ *methods which analyse variables that must be random and interrelated such that their different effects cannot be meaningfully interpreted separately*

❖*methods which describe, measure and predict the degree of relationships between variate (multivariate character lies in the number of variates not in the number of variables)*

Variate – a linear combination of variables with empirically determined weights

$$Y = \sum_i w_i X_i$$

Definition – IV, DV

Multivariate statistics - *provide analysis when there are many independent variables and/or many dependent variables, all correlated with one another to varying degrees*

Independent variables (IV) *- the different conditions, characteristics of the subject of the analysis*

Dependent variables (DV) *- the response or outcome variables predicted by the independent variables*

IV and DV are defined in the context of a certain research problem
A DV in a research problem can be an IV in another research problem

MVA techniques - classification

1. *Can the variables be divided in IV and DV?*

Dependent techniques – one or more DV are predicted by a set of IV
Interdependence techniques – no variable or set of variables are defined as being IV or DV (involves the symultaneaous analysis of all variavles)

2. *How many variables are treated as dependent in a single analysis?*

3. *Are the IV and DV metric or nonmetric (categorical) variables?*

MVA dependence methods

One or more variables (DV) are predicted by other variables (IV)

One DV

$$Y = \sum_i w_i X_i$$

Analysis of variance – DV metric, IV nonmetric
(analyses the relationship between the set of IV and DV – used to test variace in the group on one DV)

Discriminant analysis – DV nonmetric, IV metric
(analyses group differences based on a DV and predicts the likelihood an entity belongs to a certain class)

Multiple regression analysis – DV metric, IV metric or nonmetric
(predicts the change in DV in response to the change in IV)

Conjoint analysis – both DV and IV are metric or nonmetric
(assesses the importance of the attributes of the entities)

MVA dependence methods

Multiple DV

$$\sum_i Y_i = \sum_j w_j X_j$$

Canonical correlation - both DV and IV can be metric or nonmetric
(develop linear combinations of each set of DV and IV such that to maximise the correlation between the two sets

Multivariate analysis of variance – DV metric, IV nonmetric
(analyses simultaneously the relationship between several IV and more DV- Extension of analysis of variance to more DV)

MVA interdependence methods

All variables are analysed simultaneously to find underlying structure of the variables or of the entities:

Factor analysis – analysis of the structure of the variables (explains a large number of variables in terms of common dimensions called factors – e.g. principal components analysis)

Cluster analysis – analysis of structures of entities (classify a sample of entities into a small number of groups - not predefined groups)

Analyses of the structure of entities based on their attributes (analyses of the similarity of the entities and transformation this similarity into a numerical distance)

- *metric attribute – multidimensional scaling*
- *nonmetric attribute – correspondence analysis*

Garbage in, roses out?

Preliminary analysis *- preparation and understanding of the input data are essential if the results of any MVA is to be believed.*

Check list

1. *Inspect univariate descriptive statistics for accuracy of input*
   - *a. out-of-range values*
   - *b. plausible means and standard deviations*
   - *c. univariate outliers*
2. *Evaluate amount and distribution of the missing data; deal with problems*
3. *Identify and deal with non-normal variables*
   - *a. Check probability plots etc.*
   - *b. Transform variables (if desirable)*
   - *c. Check results after transformation*
5. *Identify and deal with multivariate outliers*
   - *a. identify variables causing multivariate outlires*
   - *b. description of multivariable outliers*
6. *Evaluate multicollinearity and singularity in the correlation matrix*

Missing data

Need to understand the amount and the pattern of the missing data

The pattern of missing data is more important than the amount missing

Missing values scattered randomly through data pose less serious problems than non-randomly missing values (affect the generalisability of the results)

Treatment of missing data

- *Delete cases (data points) – when only a few data cases are missing and they seem to be a random subsample of the whole sample*
- *Delete variables – when missing data are concentrated in a few variables which are not critical for the analysis or are highly correlated with others*
- *Estimate missing data with*
  - ✓ *priori knowledge – make an educated guess of the missing value*
  - ✓ *mean substitution – calculate the mean for the available data and replace the missing value*
  - ✓ *regression*
  - ✓ *Other sophisticated methods*

Normality

Multivariate normality

- needed as the statistical inference less robust for non-normal distibutions
- solution is usually better (generalise better) if the variables are all normally distributed even for methods that do not require explicitly normality

The assumption of multivariate normality can be partially checked by

- ✓ examining the normality, pair wise linearity and homoscedasticity of individual var.
- ✓ examining the normality of the residual in analysis involving prediction.

Homoscedability – the variance of one var is the same at all values of the other var

Transformations of variables

- improve their normality (use unless there is a compelling reason not to)
- not always recommended as the transformed variables are harder to interpret. (e.g. if the scale of the variable is meaningful)

Recommendations

Check the normality after the transformation

Try different transformations

Outliers

Univariate/multivariate outlier – has extreme values of one or more variable such that the statistics is disturbed (outliers detached from the rest of the distribution)

Problems caused

lead to results that do not generalise except to samples with the same kind of outliers

Causes of outliers

- *several different populations are mixed in the same sample*
- *important variables are omitted such that, if included, would attached the outlier to the rest of the distribution*

Describe outliers – determine why they are outliers

- *helps decide if the case/data point is properly part of the sample*
- *need to know which variables to modify, if necessary*
- *provides indications for which cases the results do not generalise*

Reduce the influence of the outliers - consider

- *one variable is responsible for most outliers – delete it (if possible)*
- *modifying the value of some of the outlier's variables*
- *transform a variable for univariable outliers (transform for normality brings outliers close to the distribution)*

Multicollinearity and singularity

Problems in the correlation matrix

Multicollinearity - the variables are very highly correlated (above 0.90)
- occurs form cross products or powers of variables included in the analysis along with the original variables.

Reminder - correlation is a measure of the extent to which the values of two variables go up together (positive correlation) or one goes up while the other goes down (negative correlation)

Singularity - variables are redundant (one variable is a combination of two or more others variables)
- redundant variables inflates the size of the error terms and weaken the analysis

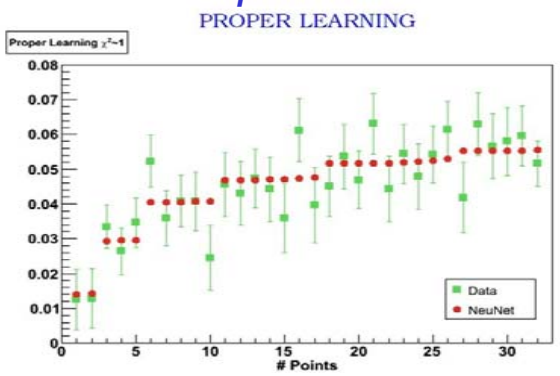
Recommendations

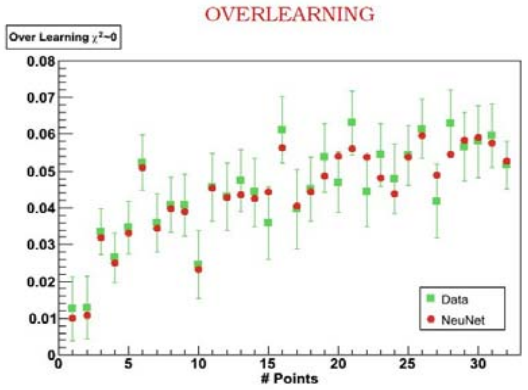
- *avoid including in the analysis two variables with a bivariate correlation higher than 0.70*
- *consider omitting one of the variables or create a composite value for the redundant variables*

Number of variables

Get the best solution with the fewest variables possible for the problem at hand
With more variables the solution improves but only slightly
Overfitting

- *the solution is too good and it will not generalise to a population*
- *occurs when too many variables are included in the analysis comparative with the size of data sample*





NNPDF Coll

Recommendations – include in the analysis

- *only a limited number of uncorrelated variables*
- *reliable variables - many unreliable variables result in unreliable solutions (reflect mainly the measurement errors)*

- *no clear recommendations exist for the effective number of variables adequate for a data sample*

Discriminant analysis

Appropriate when DV are nonmetric and IV are metric
In many cases DV represent two or more groups or classes a priori known (good-bad, high-low, high-medium-low, class1-class2-class3-…

- *two classes involved – two-group/class discriminat analysis*
- *more than 2 classes – multiple discriminant analysis*

Analysis involve deriving a combination of two or more independent variables that will discriminate between the classes

- ✓*Linear combination (variate)*
- ✓*Non linear combinations (e.g. neural networks)*

Y = F(x1, x2, …, Xn)
F – discriminant function
Y – discriminant value

Discriminant function

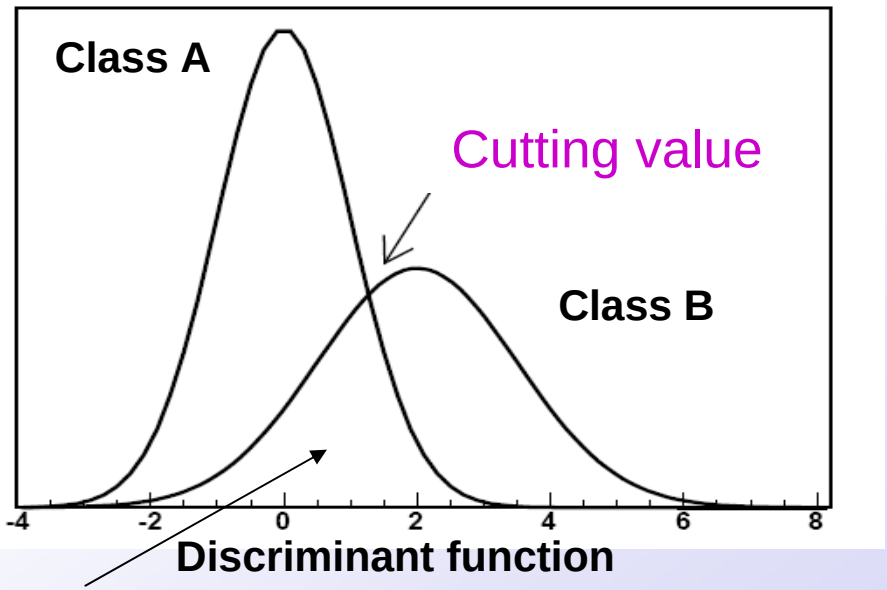
F is developed such that its distribution for the involved classes to have minimum overlap

Assessment of the discrimination power

Cutting value – criterion (value) against which each object's discriminant value is compared to determine into which group the object should be classified

Classification/confusion matrix

		Predicted class A	B
Actual class	A	N_{AA}	N_{AB}
	B	N_{BA}	N_{BB}



F- used for prediction after validation

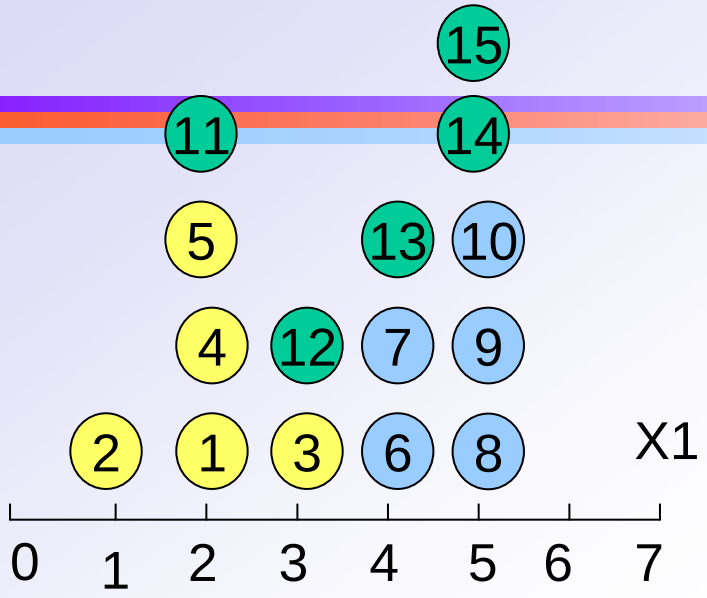
Multi-discriminant analysis

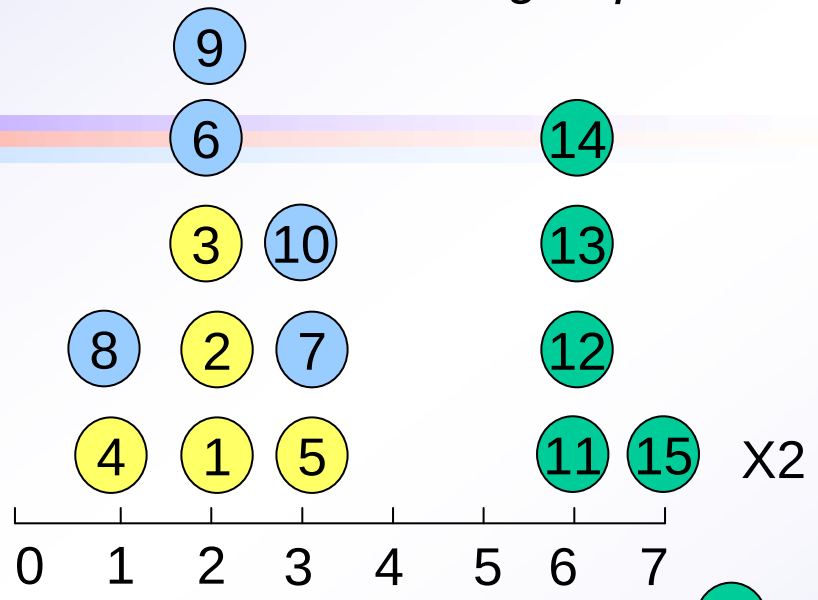
For N classes: N-1 discriminant functions will be calculated

Each discriminant function represents a dimension of the discrimination among classes/groups

Graphs of discrimination dimensions help visualising the discrimination among groups

Number in circle – index of the object; colour – different groups



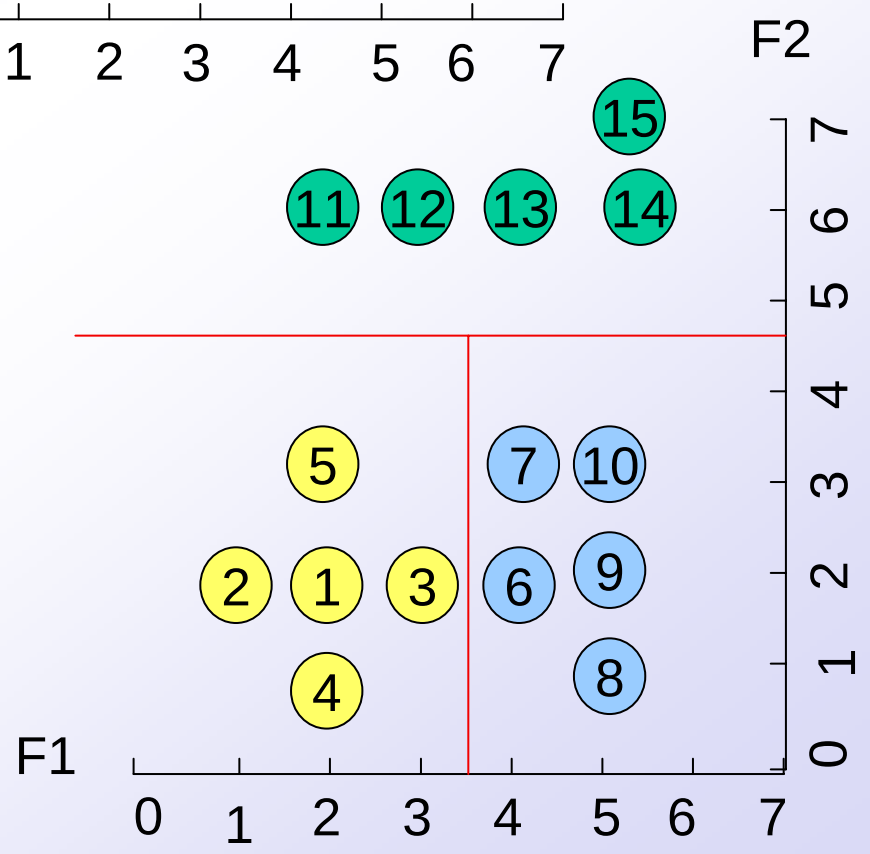


Discriminant functions

*F1 = 1.0*X1 + 0.0*X2*
*F2 = 0.0*X1 + 1.0*X2*

Cut value

F1: 3.5
F2: 4.5



Multi-discriminant analysis

Discriminant functions

$F1 = 1.0*X1 + 0.0*X2$
$F2 = 0.0*X1 + 1.0*X2$

Cut value

F1: 3.5
F2: 2.5

↓

Cuts as in HEP

$X1 < 3.5$
$X2 < 4.5$

NOTE

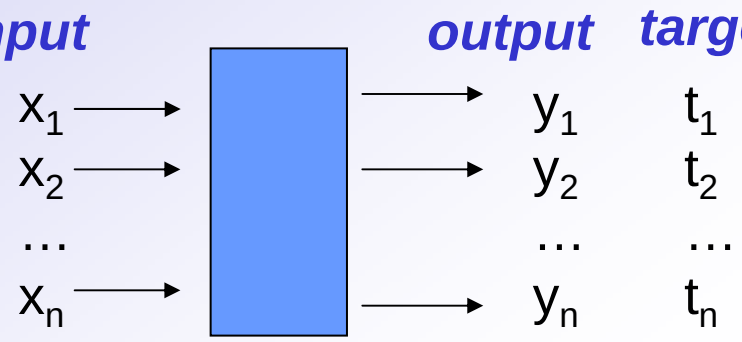
One cut – univariate discriminat

Multiple cuts for 2-class separation

- ✓*make successive univariate discriminant analysis*
- ✓*does not take into account the correlation among variables*

Statistical/machine learning

Supervised learning

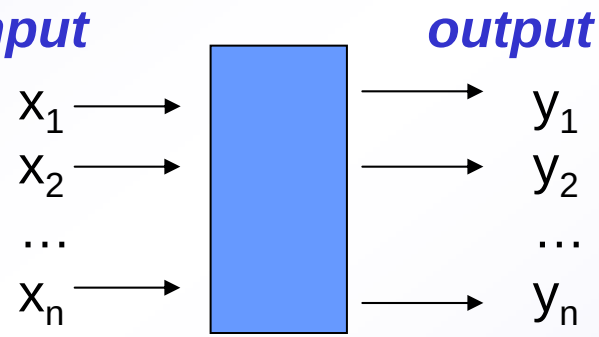


System

- ✓ ***presented with a sequence of inputs and a sequence of outputs (target)***
- ✓ ***develop a model which will minimise the difference between the actual output and the target***
- ✓ ***produce a correct/optimal output when presented with a new input***

e.g. Output is a class label - classification
Output is a real number - regression

Unsupervised learning



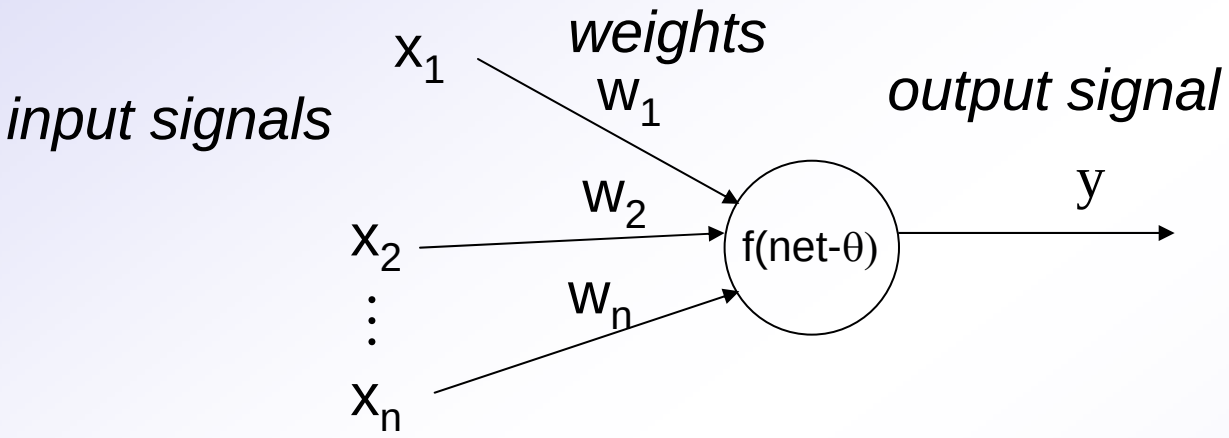
System

- ✓ ***presented with a sequence of inputs (no target provided)***
- ✓ ***discover patterns in the input data***
- ✓ ***output – association of the input to a certain pattern (e.g. association to a cluster)***

e.g. clustering = unsupervised classification

Artificial Neural Networks

Artificial neural networks (NN) – layered networks of artificial neurons (AN)



AN - receives signal from environment or other AN
weights – inhibit (negative values) or excite (positive values) an input signal
AN - collects input signals, calculates a net signal using the activation function f and transmits an output signal

$$net = \sum_{i=1}^{n} w_i x_i$$

AN – summation unit

$$net = \prod_{i=1}^{n} w_i x_i$$

AN – product unit (allow higher order combinations of inputs => increased information capacity)

Neural network types

Multilayer NN

Feedforward NN – receive external signals and propagate them through all the layers producing the output signal (no feedback connection to previous layers)

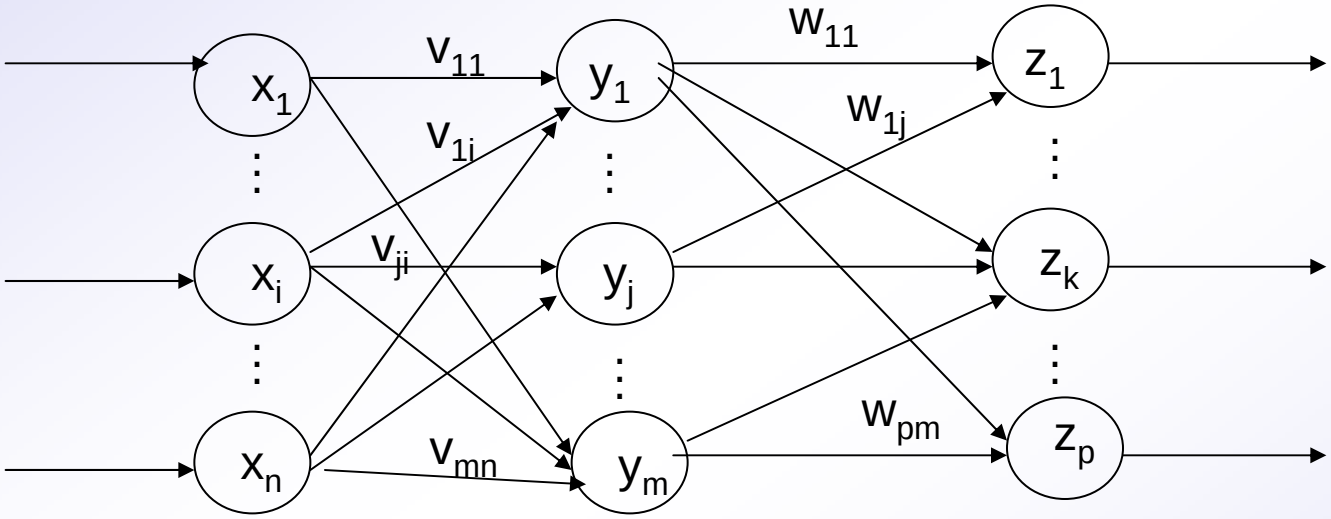
- *standard multilayer NN*
- *functional link NN*
- *product unit NN*

Recurrent NN – have feedback connection to previous layers

Time-delay NN – memorise a window of the previously observed patterns

Feedforward Neural Networks

- *one input layer, one output layer, one or more hidden layers*
- *can have direct (linear) connections between the input and output layers*



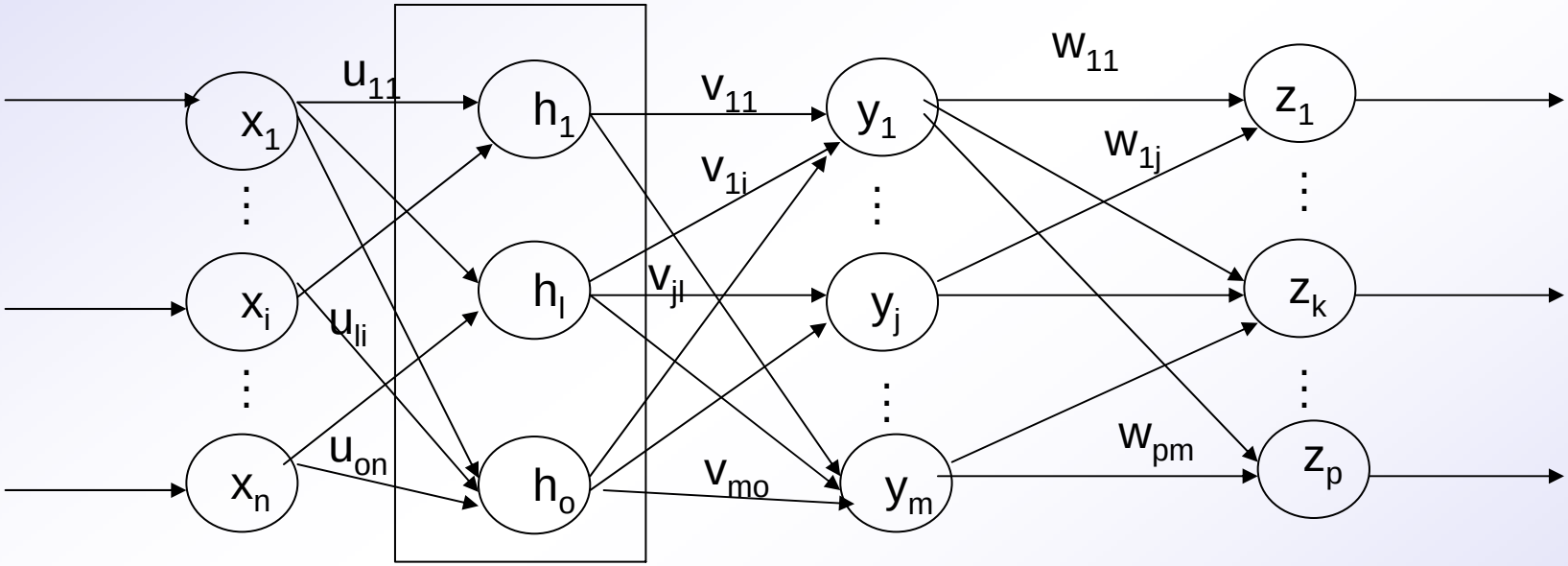
$$z_k = f_{z_k}(net_{z_k}) = f_{z_k}\left(\sum_{j=1}^{m+1} w_{kj}\, f_{y_j}(net_{y_j})\right) = f_{z_k}\left(\sum_{j=1}^{m+1} w_{kj}\, f_{y_j}\left(\sum_{i=1}^{n+1} v_{ji}\, x_i\right)\right)$$

- *each activation functions can be different*
- *Input unit can implement activation functions (usually a linear function)*

Functional Neural Networks

- *input layer expended into a layer of functional units* $h_l(x_1,\ldots x_n)$

  $u_{li}=1$ if h_l depends of x_i, $u_{li}=0$ otherwise



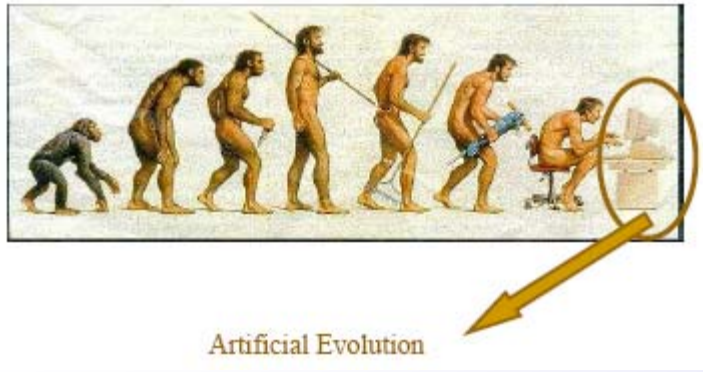
$$z_k = f_{z_k}\left(\sum_{j=1}^{m+1} w_{kj}\, f_{y_j}\left(\sum_{l=1}^{o+1} v_{jl}\, h_l(x_1,\ldots x_n)\right)\right)$$

- *faster training and improved accuracy*

J. Ghosh, Y. Shin, International Journal of Neural Systems, Vol. 3. No4, pp 323-350

Artificial evolution

Artificial evolution - ***simulation of the natural evolution on a computer***

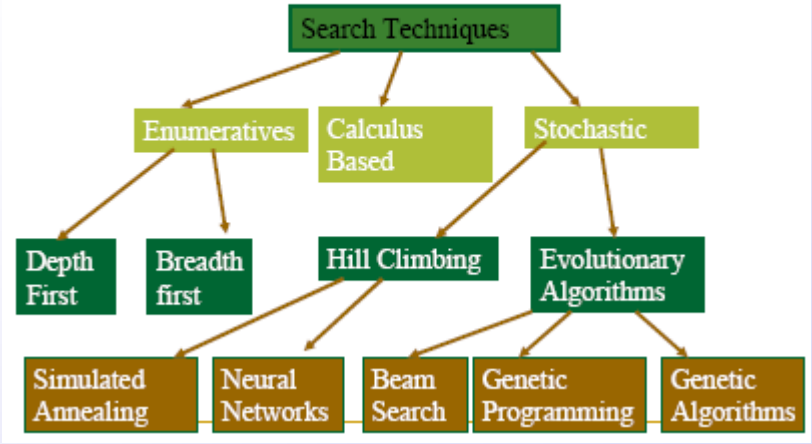


New field - ***Evolutionary computation***
(subfield of Artificial Intelligence)

❖ *Goal of evolutionary computation* - ***to generate a set of solutions to a problem of increasing quality***

Alternative search techniques

e.g. ***Evolutionary Algorithms***



Terminology

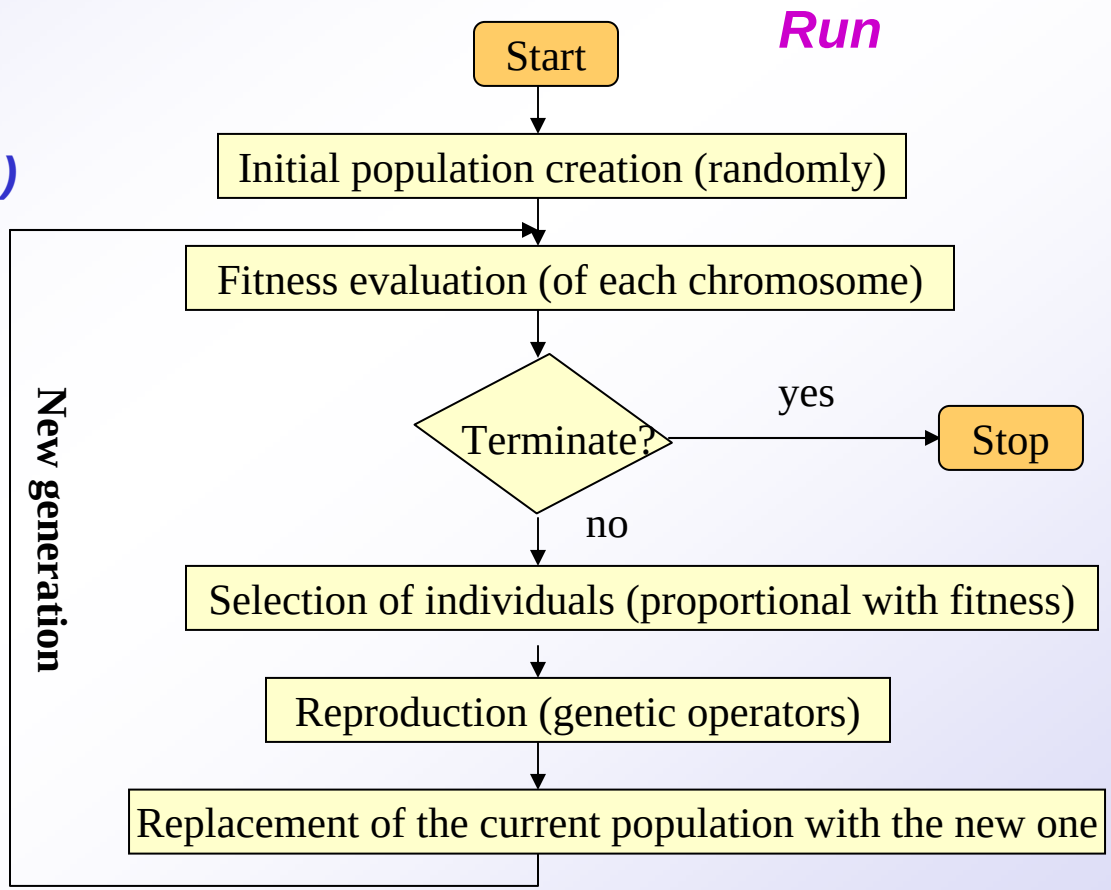
- *Individual* – candidate solution to a problem

decoding ↑ ↓ *encoding*

- *Chromosome* – representation of the candidate solution
- *Gene* – constituent entity of the chromosome
- *Population* – set of individuals/chromosomes
- *Fitness function* – representation of how good a candidate solution is
- *Genetic operators* – operators applied on chromosomes in order to create *genetic variation* (other chromosomes)

Evolutionary Algorithms

- *Problem definition*
- *Solution representation (encoding the candidate solution)*
- *Fitness definition*
- *Run*
- *Decoding the best fitted chromosome = **solution***



Genetic operators

- ✓ *cross-over – combining genetic material from parents*
- ✓ *mutation - randomly changes the values of genes*
- ✓ *elitism/cloning – copies the best individuals in the next generation*

Genetic Programming

*GP search for the **computer program** to solve the problem, not for the solution to the problem.*

***Computer program** - any computing language (in principle)*
- LISP (List Processor) (in practice)

LISP - highly symbol-oriented

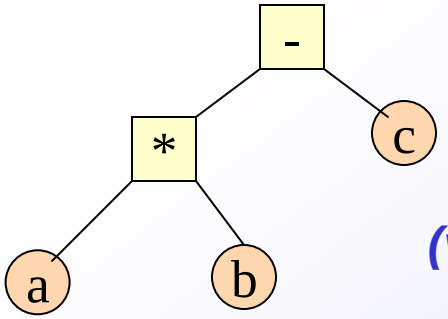
Mathematical expression

a*b-c

S-expression

(-(*ab)c)

Graphical representation of S-expression



functions (+,) and terminals (a,b,c) (variables or constants)*

Solution representation

Chromosome: S-expression - variable length => more flexibility
- sintax constraints => invalid expressions

Reproduction

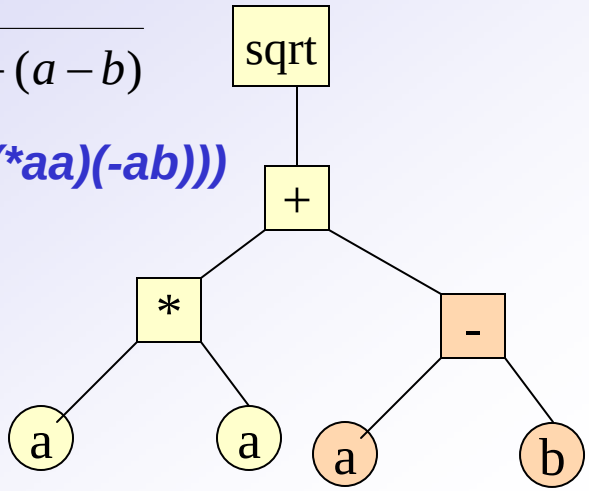
Cross-over (recombination) and Mutation (usualy)

Cross-over operator

Parents

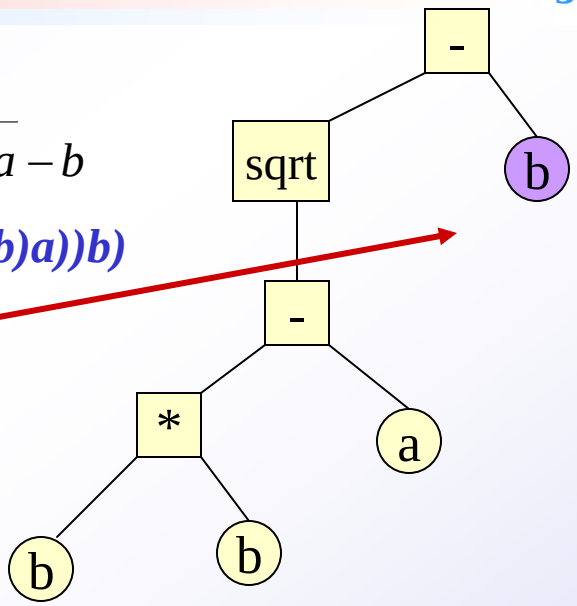
$\sqrt{a^2+(a-b)}$

*(sqrt(+(*aa)(-ab)))*



$\sqrt{b^2-a}-b$

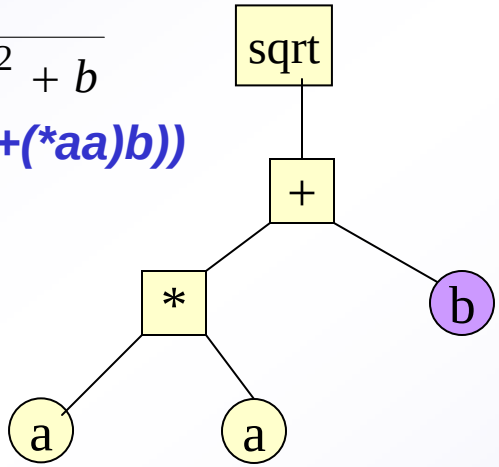
*(-(sqrt(-(*bb)a))b)*



Offspring

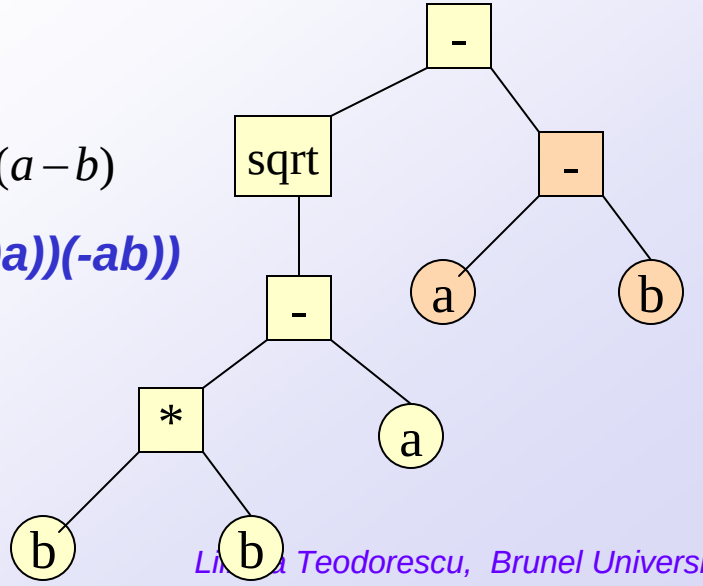
$\sqrt{a^2+b}$

*(sqrt(+(*aa)b))*



$\sqrt{b^2-a}-(a-b)$

*(-sqrt(-(*bb)a))(-ab))*



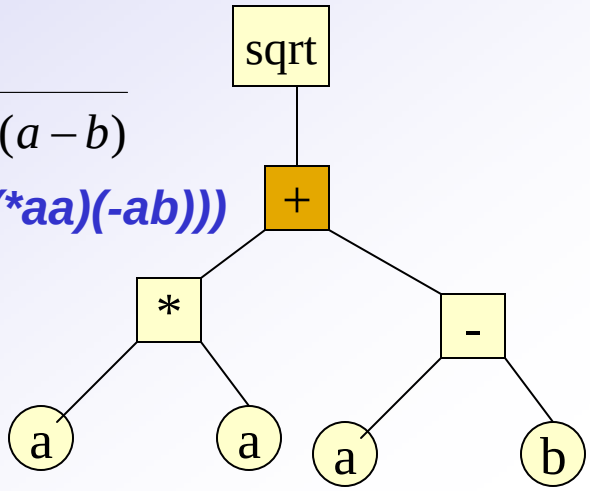
Mutation operator

- *function* replaced by another *function*
- *terminal* replaced by another *terminal*

Parents

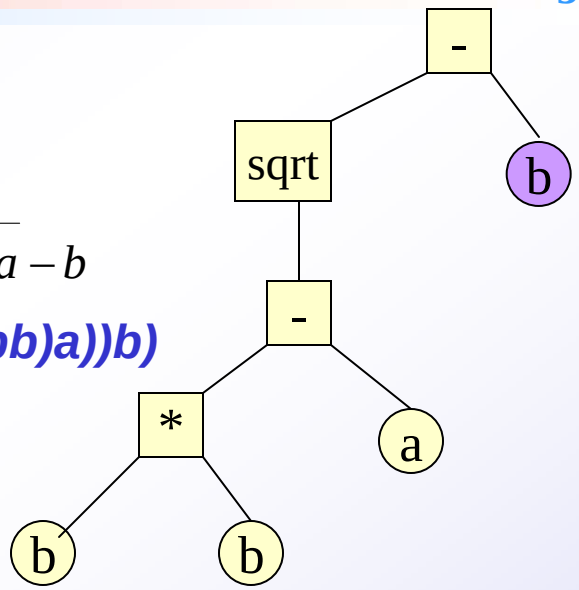
$\sqrt{a^2+(a-b)}$

*(sqrt(+(*aa)(-ab)))*



$\sqrt{b^2-a}-b$

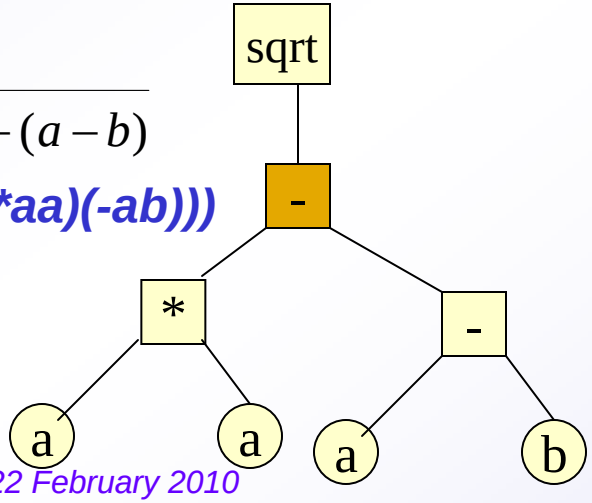
*(-(sqrt(-(*bb)a))b)*



Offspring

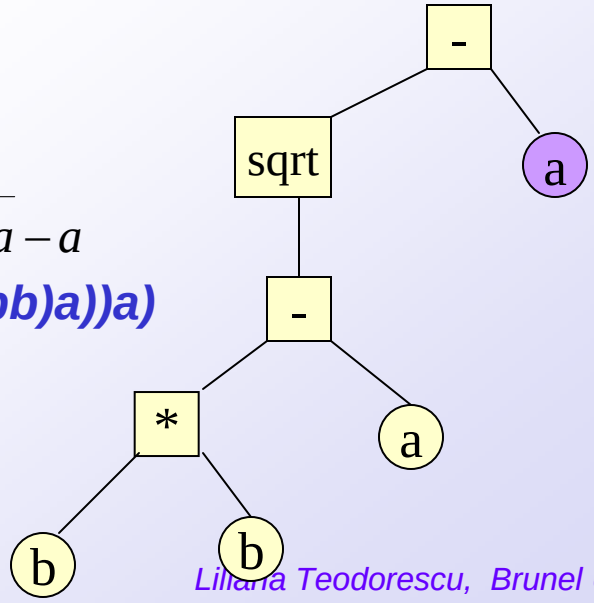
$\sqrt{a^2-(a-b)}$

*(sqrt(-(*aa)(-ab)))*



$\sqrt{b^2-a}-a$

*(-sqrt(-(*bb)a))a)*



GP in PP

Experimental PP - event selection

- ***Higgs search in ATLAS*** *K. Cranmer et.al., Comp. Phys. Com 167, 165 (2005).*
- ***D, D_s and Λ_c decays in FOCUS*** *(J.M. Link et. al., NIM A 551, 504 (2005); PL B624, 166 (2005))*

e.g. Search for $D^+ \to K^+\pi^+\pi^-$ *(FOCUS)*

Chromosome: ***candidate cuts/selection rules - tree of:***

- ***functions: mathematical functions and operators, boolean operators***
- ***variables: vertexing variables, kinematical variables, PID variables***

Fitness function ***(will be minimised)***

$$\frac{S+B}{S^2}\times 10000(1+0.005\times n)$$

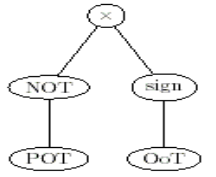
n - number of tree nodes

penalty based on the size of the tree

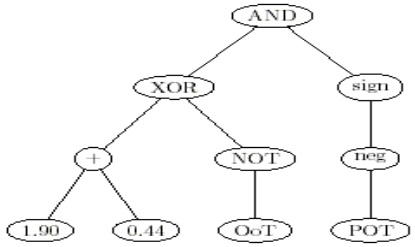
(big trees must make significant contribution to bkg reduction or signal increase)

GP in PP

Best fitted chromosomes from generation 0



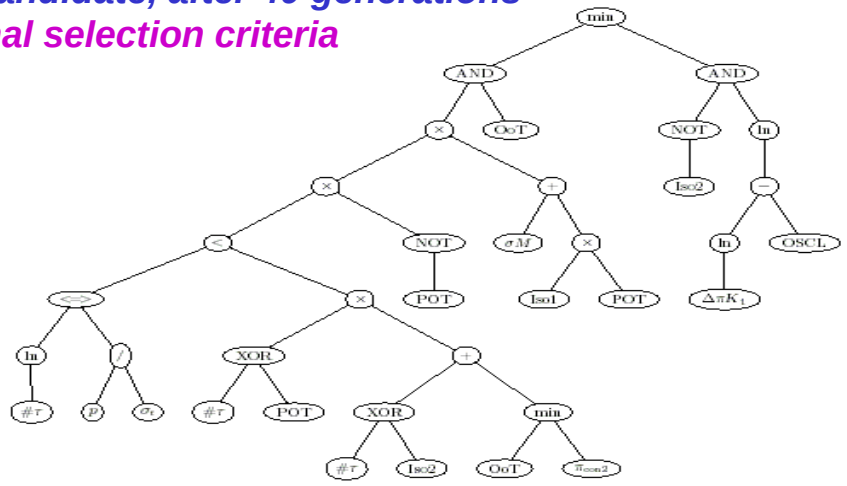
(a) Fitness 0.458



(b) Fitness 0.469

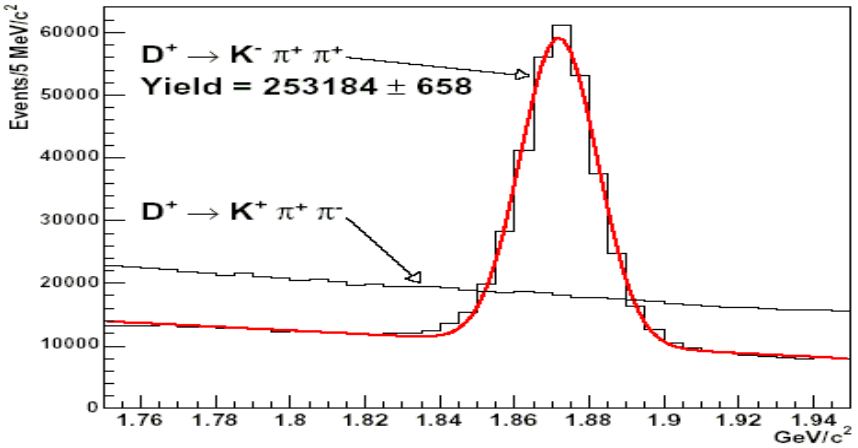
Inter point in target *Decay vertex out of target*

Best candidate, after 40 generations
= final selection criteria

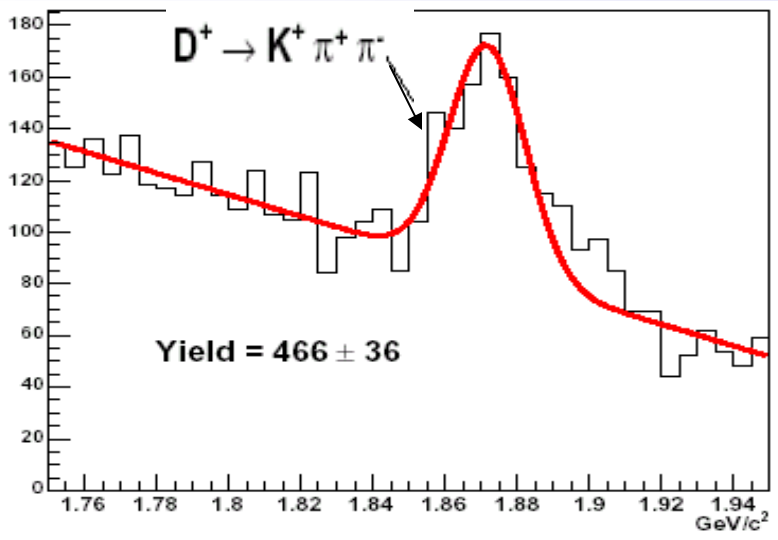


(a) Most fit tree: fitness 0.1234

Initial selection



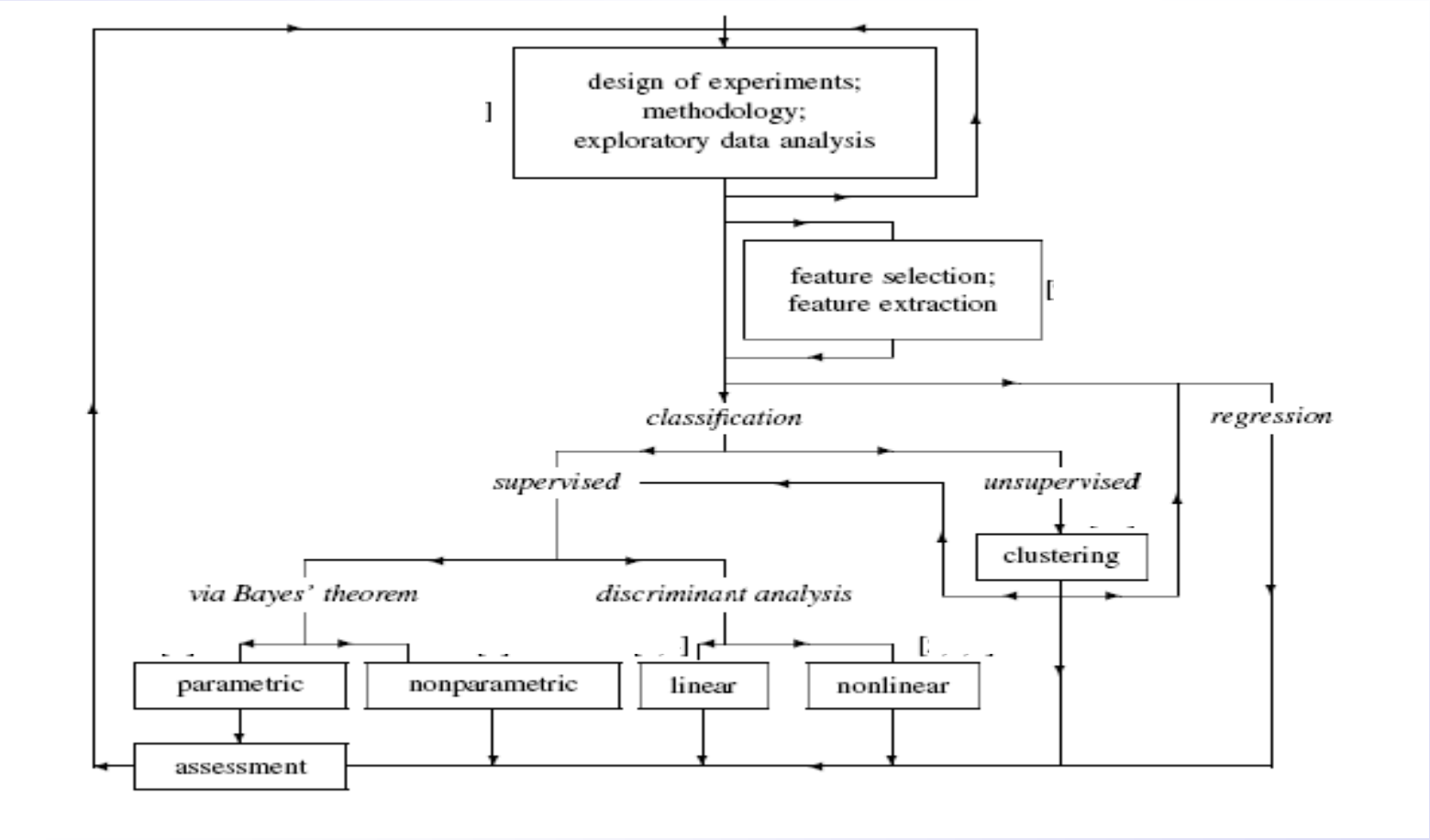
Final selection



Pattern recognition

Pattern – a data vector X=(x1,x2,…,xn)

Pattern recognition – pattern classification (supervised and unsupervised)



Conclusions

Multivariate data analysis

- *Not to be confused/reduce to discriminant analysis*
- *Not to be confused/reduce to statistical/machine learning*
- *Contains both of them plus other analysis techniques*

Pattern recognition - set of multivariate analysis techniques for

- *feature selection and extraction (data preparation)*
- *supervised and unsupervised classification techniques*