Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

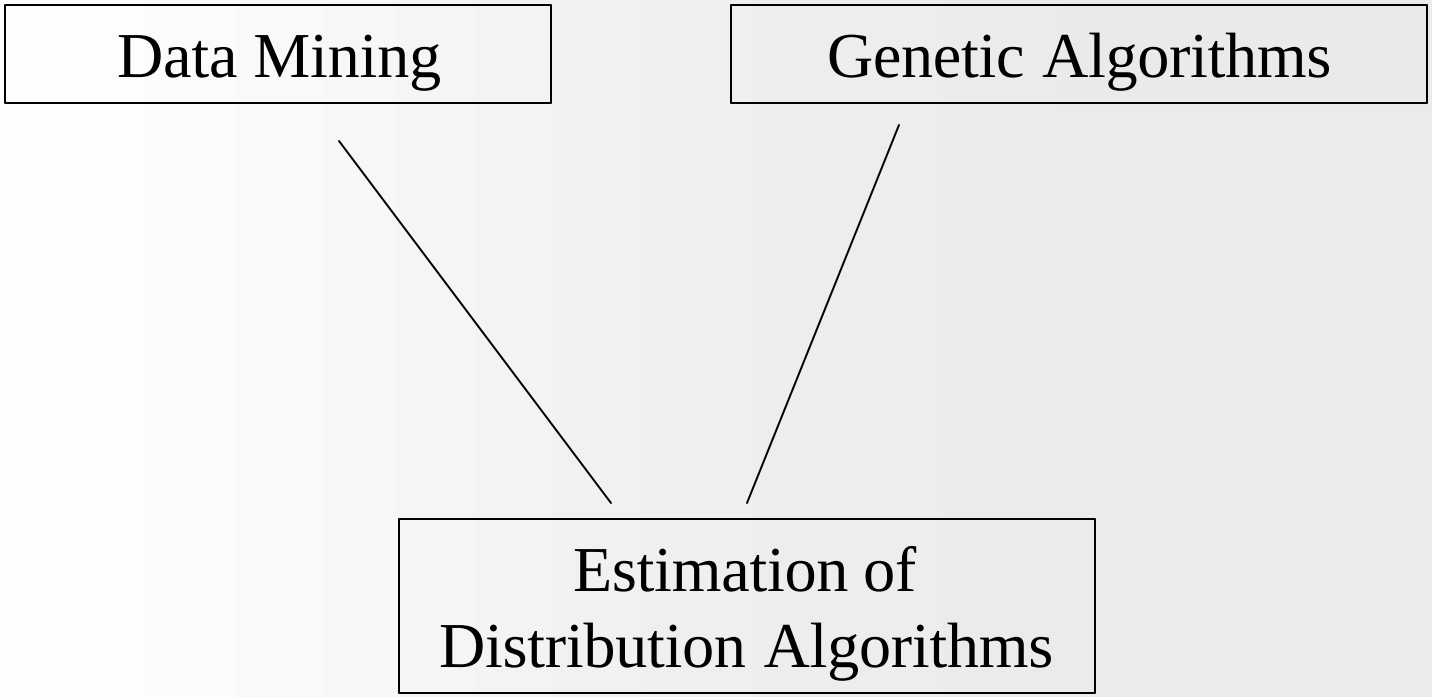
colab
eth zurich

Introduction to Estimation of Distribution Algorithms

Jiri Ocenasek

May 2004

Context

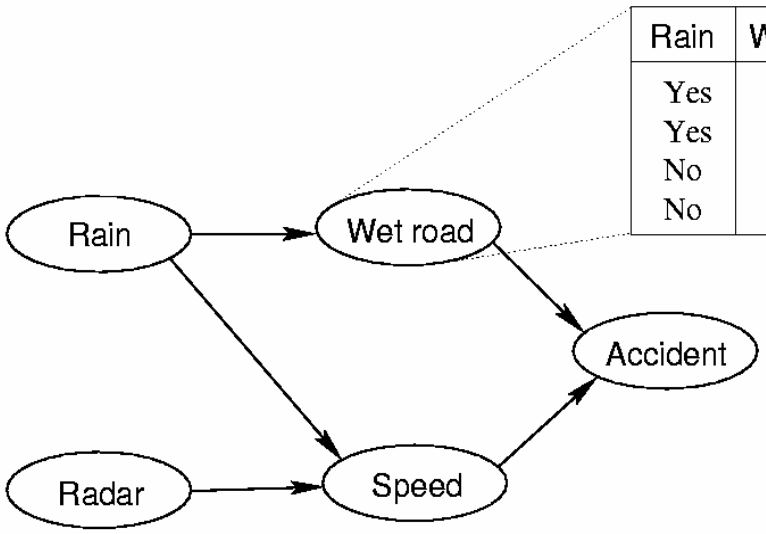

Data Mining
Genetic Algorithms
Estimation of
Distribution Algorithms

Data mining

- Knowledge extraction from databases
- Example: Customer’s cards in shops

Bayesian network

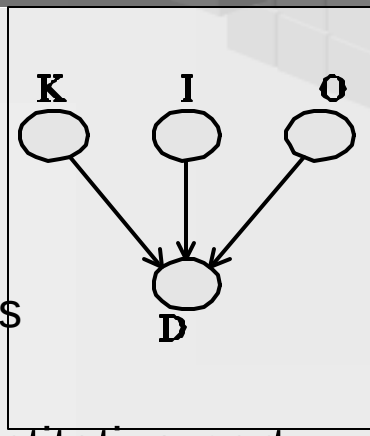
- An Example:



Rain	Wet Road	p(Wet Road | Rain)
Yes	Yes	0.995
Yes	No	0.005
No	Yes	0.060
No	No	0.940

- Another example:

Bayesian network – cont'd.



K I O	p(D|K,I,O)
1 1 1	0.00
1 1 0	0.00
1 0 1	0.95
1 0 0	0.95
0 1 1	0.00
0 1 0	0.00
0 0 1	0.40
0 0 0	0.05

- Formal definition: B = ($\mathcal{G}$, $\mathcal{L}$)

  $\mathcal{G}$ .. Dependency graph (DAG)

  - Vertices correspond to random variables
  - Edges capture dependencies

  $\mathcal{L}$ .. Conditional probabilities $p(X_i|P_i)$ - quantitative part

- Bayesian network encodes a joint probability distribution

$$p(X) = \prod_{i=0}^{n-1} p\left(X_i \mid \Pi_i\right)$$

  $\boldsymbol{X}=(X_0, X_1, \ldots, X_{n-1})$ is a vector of variables, P_i is the set of parents of X_i

- "Each variable X_i is independent of its nondescendants $\boldsymbol{X}\backslash P_i$, given its parents P_i in $\mathcal{G}$ "

Construction of Bayesian networks

- Dependency graph fixed
  - Only conditional probabilities are estimated

$$p(x_i \mid \boldsymbol{p_i}) \approx \frac{m(x_i, \boldsymbol{p_i})}{m(\boldsymbol{p_i})}$$

    - $m(p_i)$ … the number of individuals in the population having variables P_i set to concrete value $\mathrm{p_i}$
    - $m(x_i, p_i)$ … the number of individuals in the population having variables P_i set to concrete value p_i and X_i equal to concrete value x_i
- Dependency graph is not given
  - More complex task:
    - Search the space of possible dependency graphs (NP-hard problem)
    - Each hypothesis about the dependency graph has to be evaluated
    - For each evaluation the conditional probabilities have to be estimated

Incremental construction of a Bayesian network

```
Start with an empty network B;
while any edge can be added do
begin
  for each edge that can be added do
  begin
    compute the metrics of the network B' that
     can be constructed from B by adding this
     edge;
  end
  add the edge giving the highest improvement
     to the network B;
end
```

- Constraints
  - Acyclicity
  - Model complexity penalty

How to evaluate the quality of Bayesian networks?

- The posterior probability of a Bayesian network B given data D can be computed by applying Bayes theorem as :

$$p(B|D) = p(D|B)\, p(B) / p(D).$$

The higher the p(B|D), the more likely the network B is a correct model of the data. Therefore, the value of p(B|D) can be used to score different networks and measure their quality. Since we are interested in comparing different networks (hypotheses) for a fixed data set D, we can eliminate the denominator.

p(B) is a “prior” probability of B

p(D|B) is expressed e.g. by Bayes - Dirichlet metrics

Bayesian Dirichlet metrics

- A closed form of p(D|B) derived by Heckerman (under some additional assumptions about the parameters):

$$p(D \mid B) = \prod_{i=0}^{n-1} \prod_{\mathbf{p_i}} \frac{\Gamma(m'(\mathbf{p_i}))}{\Gamma(m(\mathbf{p_i}) + m'(\mathbf{p_i}))} \prod_{x_i} \frac{\Gamma(m(x_i, \mathbf{p_i}) + m'(x_i, \mathbf{p_i}))}{\Gamma(m'(x_i, \mathbf{p_i}))}$$

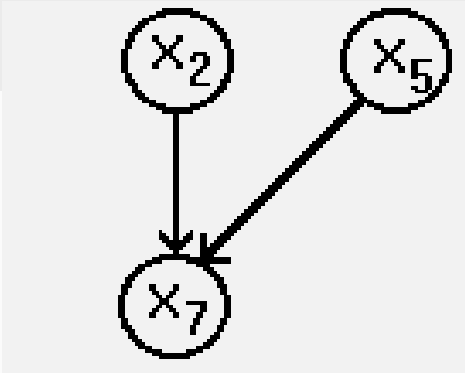
- In the variant K2:

=2

=1

An example of Bayesian Dirichlet metrics usage

- Part of BN:



$X_i \ldots X_7; \quad \Pi_I = \{X_2, X_5\}$

- Population:

X2		X5		X7		
. . .	1	. . .	0	. . .	1	. . .
. . .	0	. . .	0	. . .	0	. . .
. . .	1	. . .	0	. . .	1	. . .
. . .	1	. . .	1	. . .	1	. . .
. . .	0	. . .	0	. . .	0	. . .
. . .	0	. . .	1	. . .	0	. . .
. . .	0	. . .	0	. . .	1	. . .
. . .	1	. . .	1	. . .	0	. . .
. . .	0	. . .	0	. . .	0	. . .
. . .	0	. . .	1	. . .	0	. . .

p_i	x_i	$m(x_i, p_i)$	$m(p_i)$
00	0	3	} 4
	1	1	
01	0	2	} 2
	1	none	
10	0	none	} 2
	1	2	
11	0	1	} 2
	1	1	

..cont'd

$$p(D|B) = \underbrace{\ldots\ldots\ldots\ldots}_{X_0\ part} \cdot \ldots \cdot \underbrace{\underbrace{\frac{2!}{(2+m(X_2=0,X_5=0))!}\prod_{x_7=0}^{1}\frac{(1+m(X_7=x_7,X_2=0,X_5=0))\ !}{1\ !}}_{p_7:\ (X_2=0,X_5=0)} \cdot \underbrace{\frac{2!}{(2+m(X_2=0,X_5=1))!}\prod_{x_7=0}^{1}\frac{\ldots}{1!}}_{p_7:\ (X_2=0,X_5=1)}}_{X_7\ part} \cdot \ldots \cdot \ldots$$

$$p(D|B) = \underbrace{\ldots\ldots\ldots\ldots}_{X_0\ part} \cdot \ldots \cdot \underbrace{\underbrace{\frac{2!}{\left(2+\underbrace{1+3}_{X_2=0,X_5=0}\right)!}\ \underbrace{\frac{(1+1)\ !}{1\ !}}_{X_2=0,X_5=0,X_7=0}\ \underbrace{\frac{(1+3)\ !}{1\ !}}_{X_2=0,X_5=0,X_7=1}}_{p_7:\ (X_2=0,X_5=0)} \cdot \underbrace{\frac{2!}{\left(2+\underbrace{2+0}_{X_2=0,X_5=1}\right)!}\ \underbrace{\frac{(1+2)\ !}{1\ !}}_{X_2=0,X_5=1,X_7=0}\ \underbrace{\frac{(1+0)\ !}{1\ !}}_{X_2=0,X_5=1,X_7=1}}_{p_7:\ (X_2=0,X_5=1)}}_{X_7\ part} \cdot \ldots \cdot \ldots$$

- Incremental adding of edges – only part of the score has to be recomputed

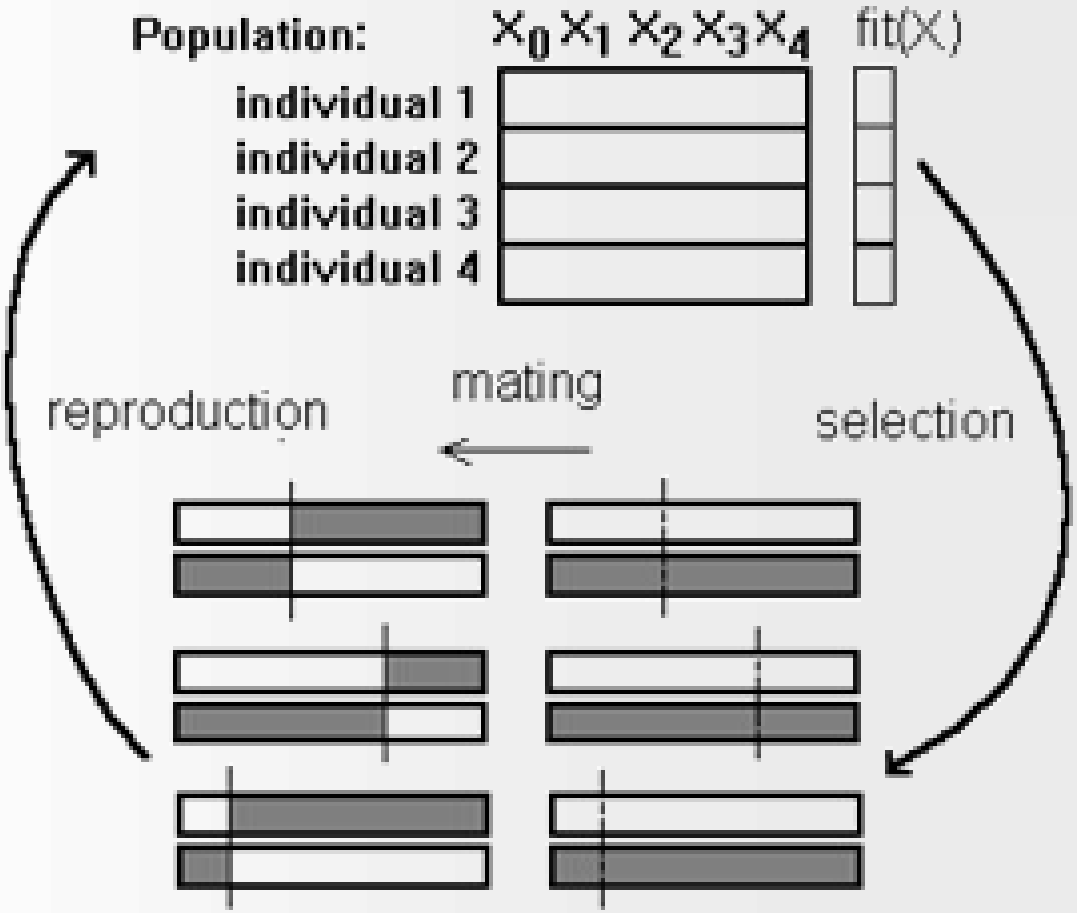
Another alternatives for scoring functions

- Minimum Description Length score (MDL):

$$DL(\mathcal{G}, D) = DL_{graph}(\mathcal{G}) + \sum_{i=0}^{n-1} DL_{tab}(X_i, \mathbf{?}_i) + \underbrace{N \sum_{i=0}^{n-1} H(X_i \mid \mathbf{?}_i)}_{DL_{data}}$$

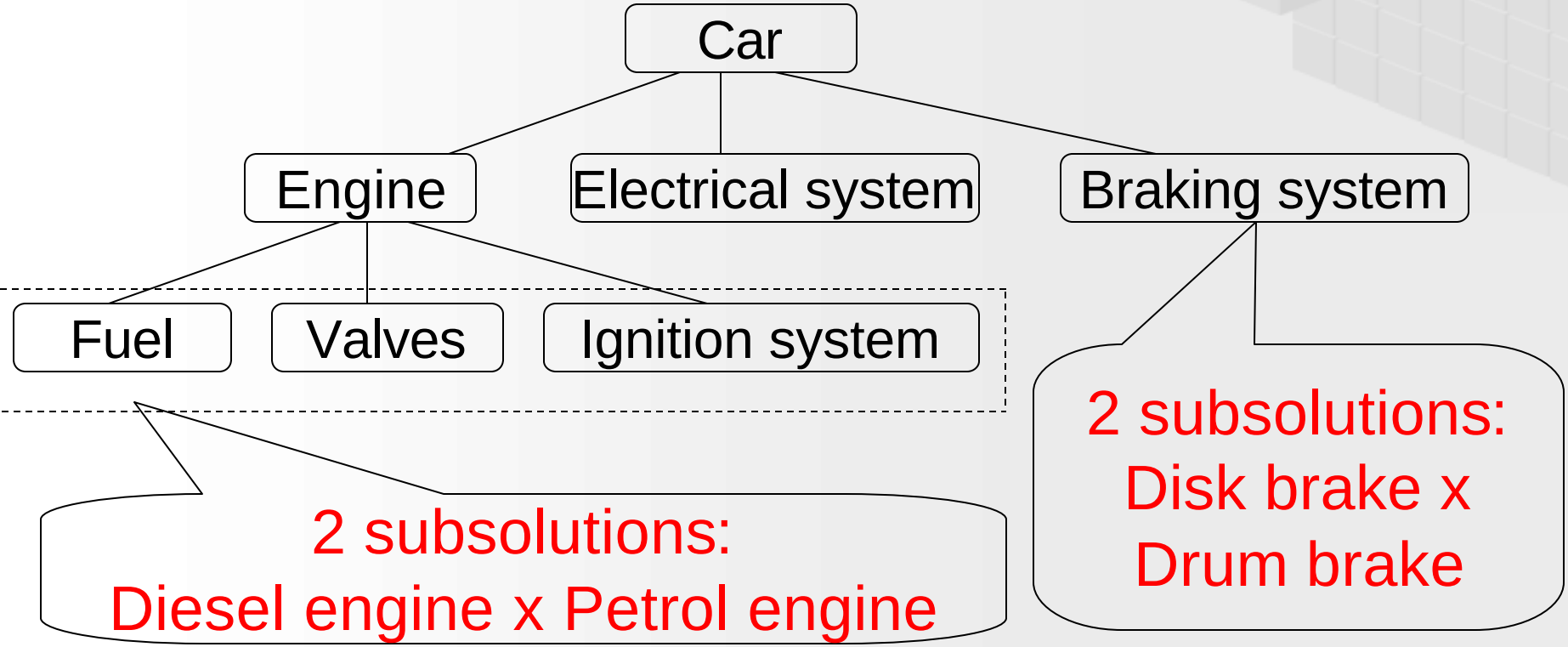
- Bayesian Information Criterion (BIC)
  - Equivalent to MDL when we ignore the description of dependency graph
- Information Gain
- Mutual Information
- …

Recall: Genetic Algorithms



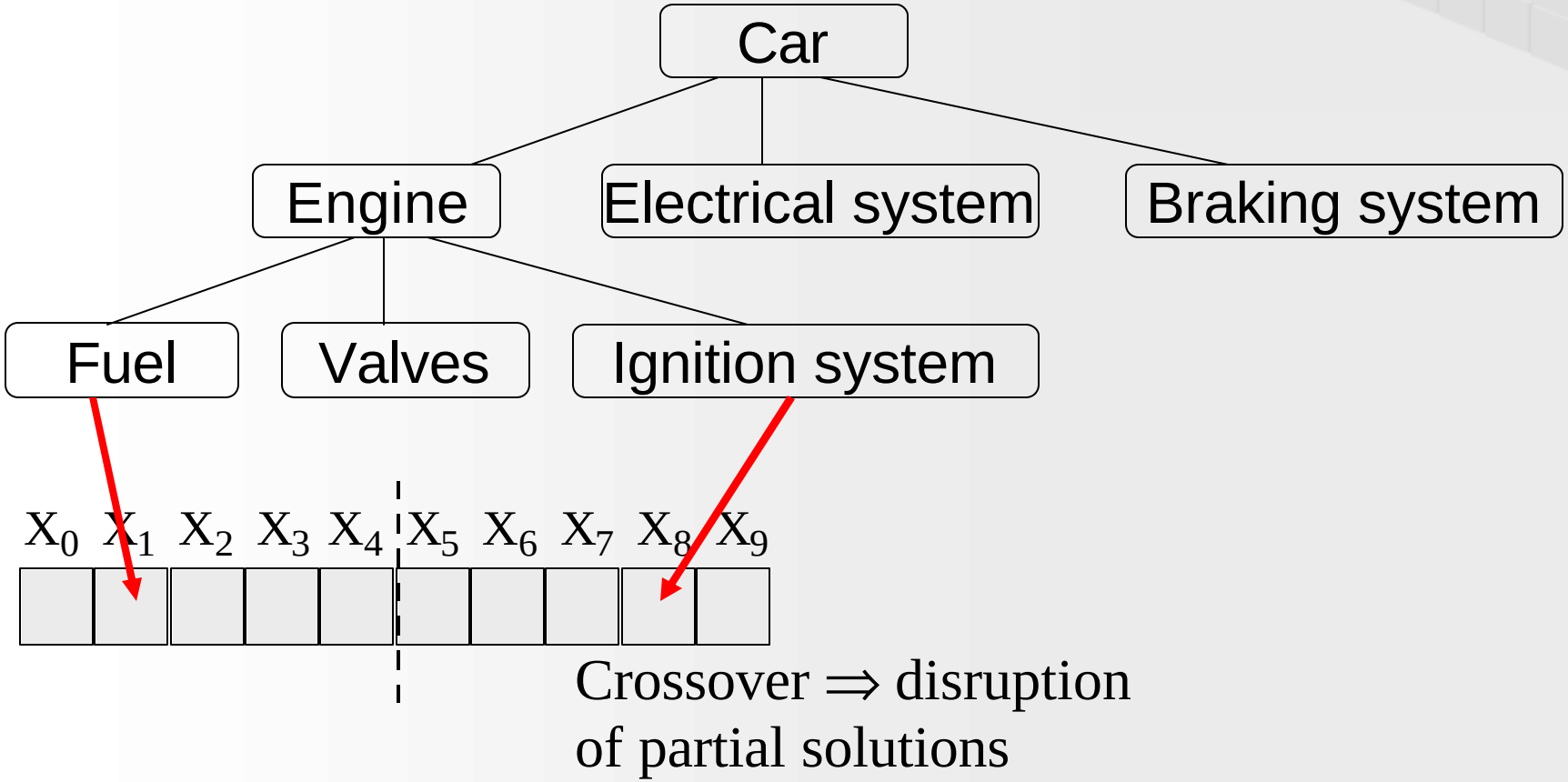
Problem decomposability

- An example: Automotive design



Motivation for Estimation of Distribution Algorithms

- Disruption of subsolutions:



Schema Theorem

- Schema example H: * * * 0 * 1 * * 0 *
  - Chromozome length n=10
  - Schema order o(H)=3
  - Schema defining length δ(H)=5
- Schema Theorem
  - Proportionate selection & one-point crossover with probability p_c and mutation with probability p_m:

$$\langle m(H,t+1)\rangle \geq m(H,t)\frac{f(H,t)}{\bar{f}(t)}\left[1-p_c\frac{d(H)}{n-1}-p_m o(H)\right]$$

Some functions are deceptive for GAs

Additively decomposable functions:

$$f(\mathbf{X}) = \sum_{i=0}^{l-1} f_k(S_i)$$

An example: f_3 deceptive

Triplet	000	001	010	011	100	101	110	111
f_3	0,9	0,8	0,8	0	0,8	0	0	1

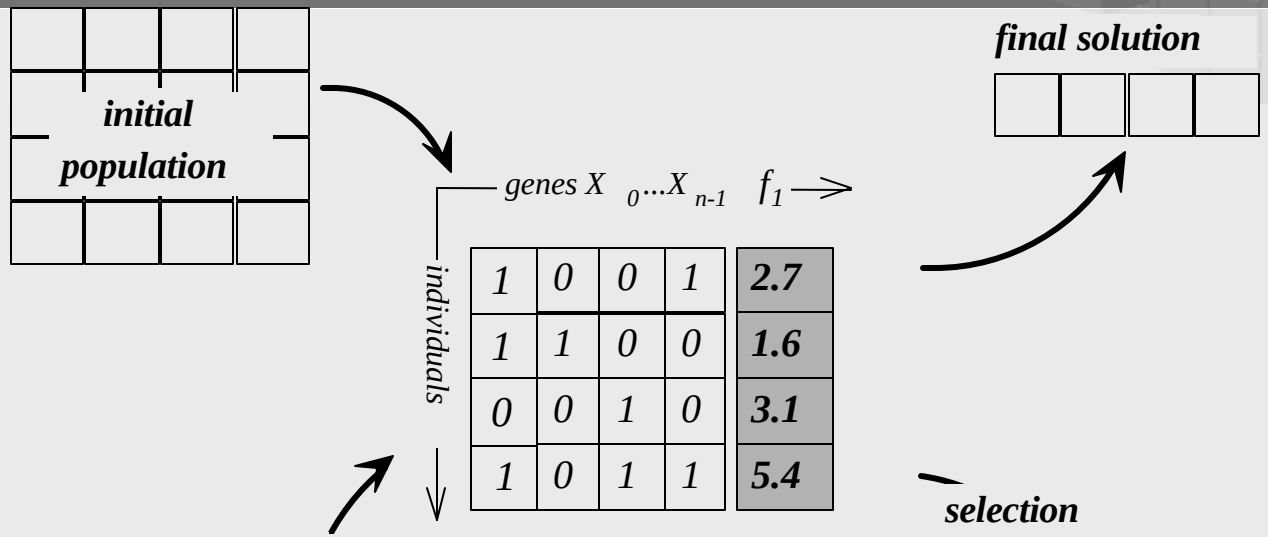
1**: $\frac{0{,}8+0+0+1}{4} = 0{,}45$

0**: $\frac{0{,}9+0{,}8+0{,}8+0}{4} = 0{,}625$

How datamining can help to improve GAs?

- Genes are treated as random variables
- The relationships between genes are captures in the form of probabilistic model

Classical GA Þ Bayesian Optimization Algorithm

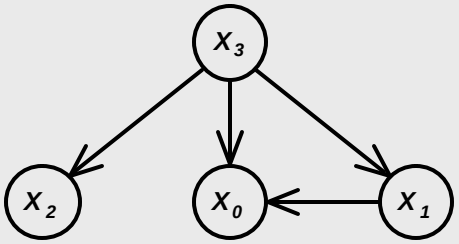


1	0	0	1	2.7
1	1	0	0	1.6
0	0	1	0	3.1
1	0	1	1	5.4

BOA EVOLUTIONARY CYCLE

model sampling (PLS - Probabilistic Logic Sampling)

data-mining techniques



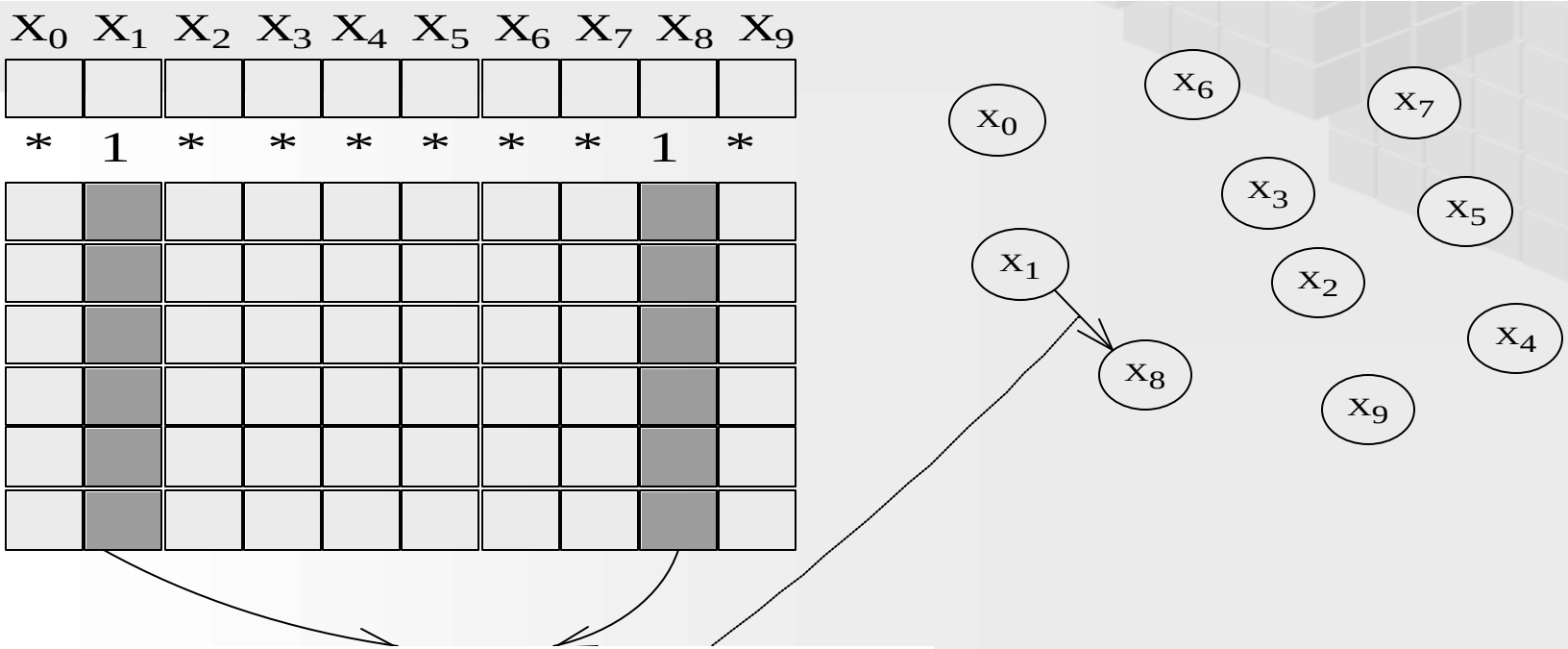
$$p(\mathbf{X})=p(X_3).p(X_1|X_3).p(X_2|X_3).p(X_0|X_1,X_3)$$

Conditional probability distributions (CPDs).

Example of $p(X_0|X_1,X_3)$:

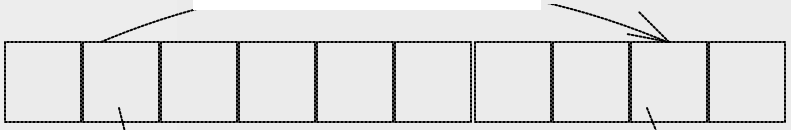
X_1	X_3	$p(X_0=1\|X_1,X_3)$
0	0	0.9
0	1	0.2
1	0	0.7
1	1	0.7

Building block detection & preservation



Bayes-Dirichet metrics

dependency



$p(X_1)=0.7$

$p(X_8| X_1)=0.9$

$p(X_8| \overline{X}_1)=0.3$

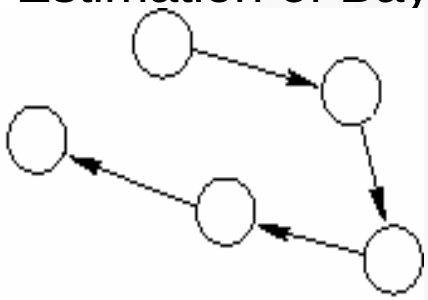
Main Idea

- Probabilistic model preserves significant combinations of parameters
  - Model complexity should be penalized to avoid overtraining (duplication of individuals)
- Selection is the source of progress
- Model complexity issues
  - Too complex model $\Rightarrow$ parent population is nearly reproduced
  - Too coarse model $\Rightarrow$ higher order dependencies are ignored (problem considered nearly separable)

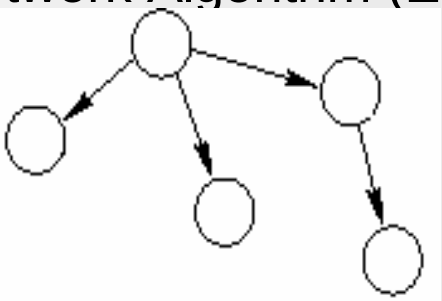
Predecessors of BOA

- The Univariate Marginal Distribution Algorithm (UMDA)
- The Mutual Information Maximization for Input Clustering (MIMIC)
- Combining Optimizers with Mutual Information Trees (COMIT)
- The Bivariate Marginal Distribution Algorithm (BMDA)
- The Learning Factorized Distribution Algorithm (LFDA)
- The Extended compact Genetic Algorithm (EcGA)
- The Estimation of Bayesian Network Algorithm (EBNA)

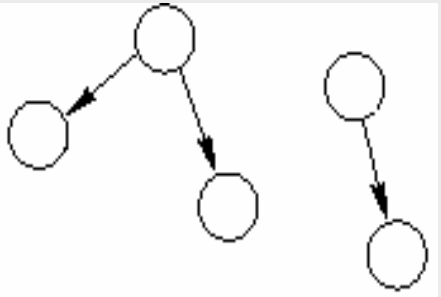
MIMIC



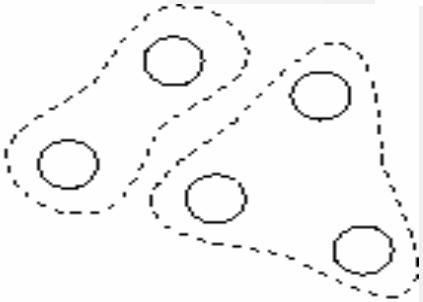
COMIT

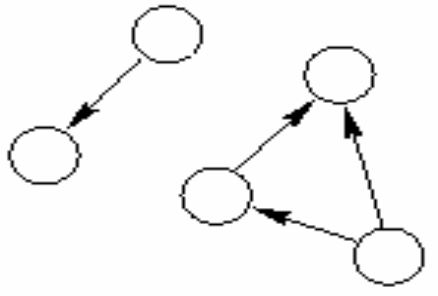


BMDA



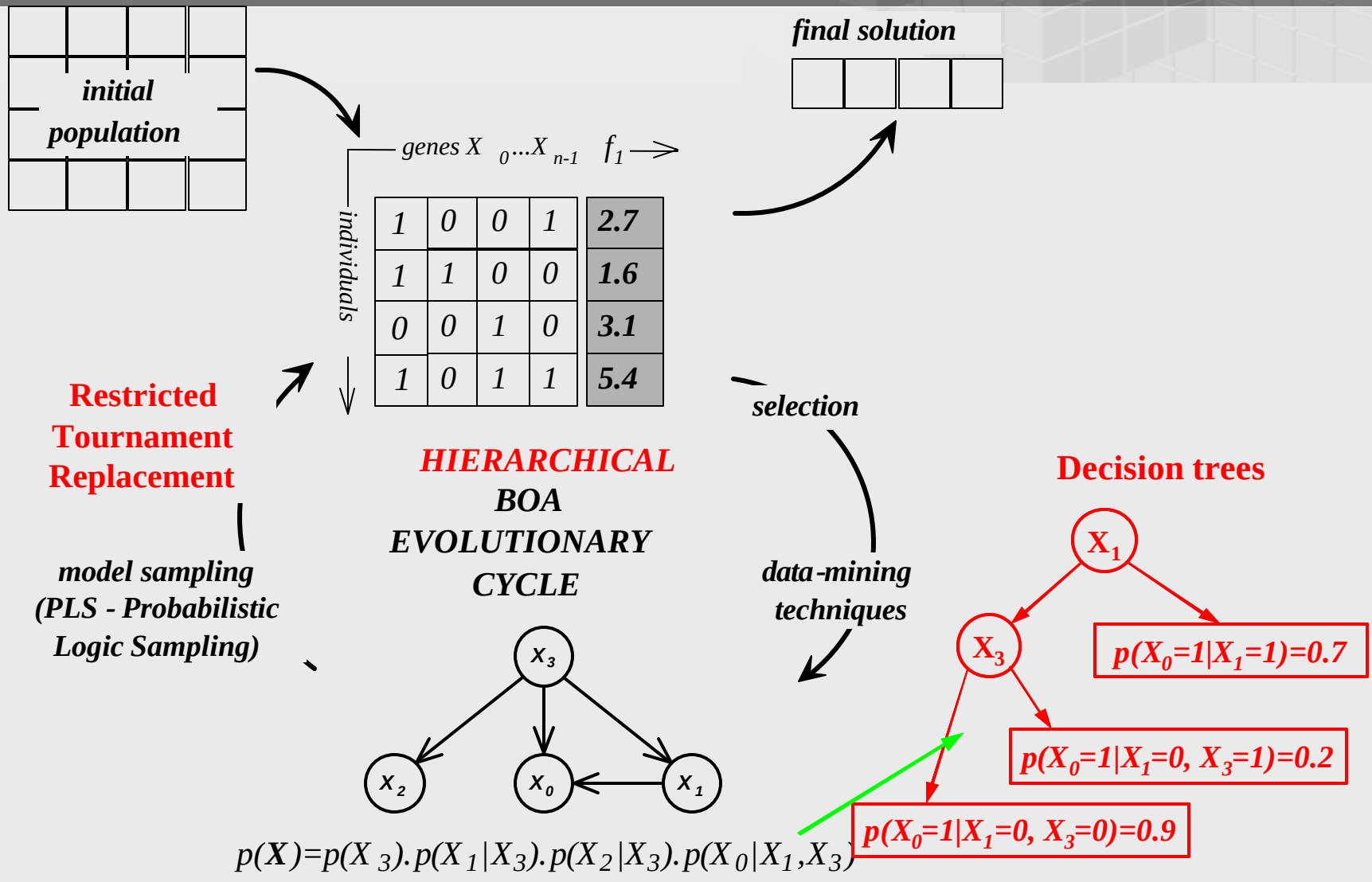
EcGA





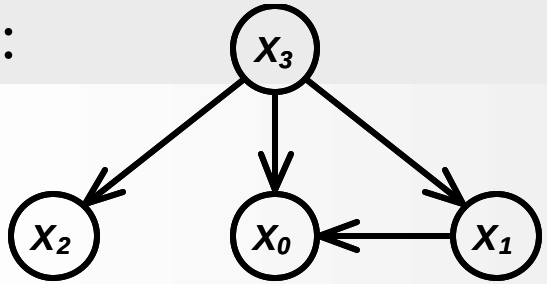
BOA/EBNA/LFDA

Recent advancements – hierarchical BOA



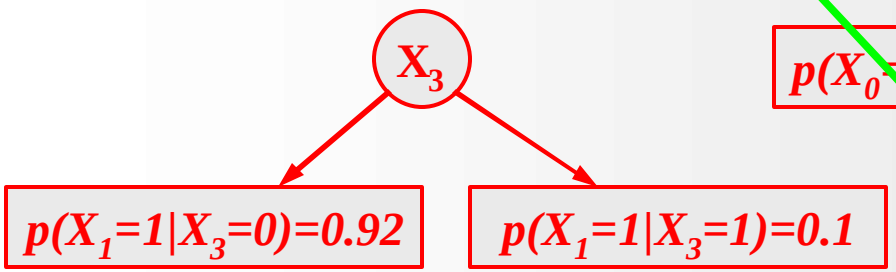
Example of Bayesian network with decision trees

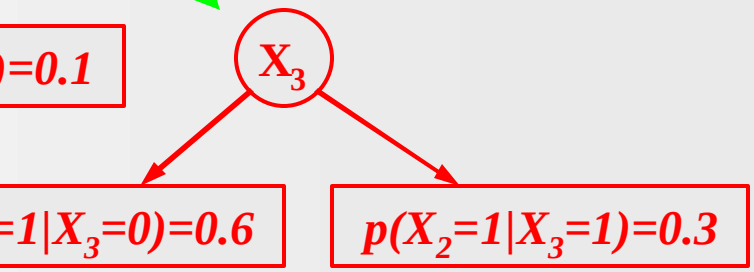
Example:

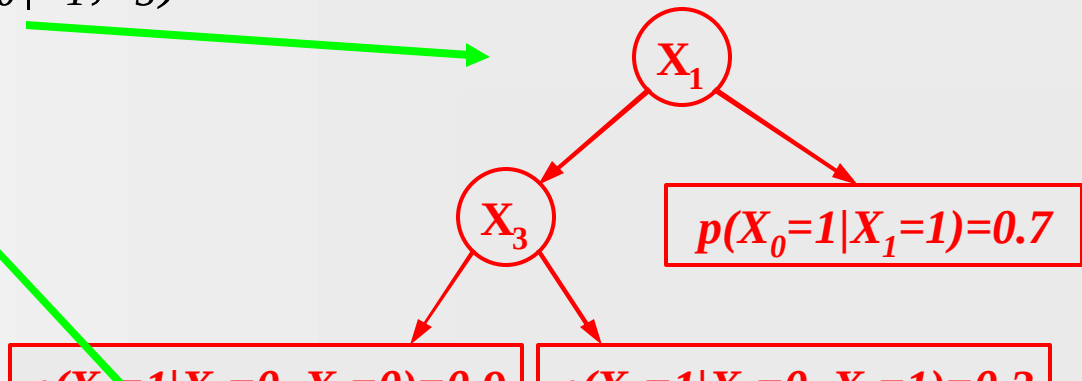


$$p(\mathbf{X})=p(X_3).p(X_1|X_3).p(X_2|X_3).p(X_0|X_1,X_3)$$

$p(X_3=1)=0.8$

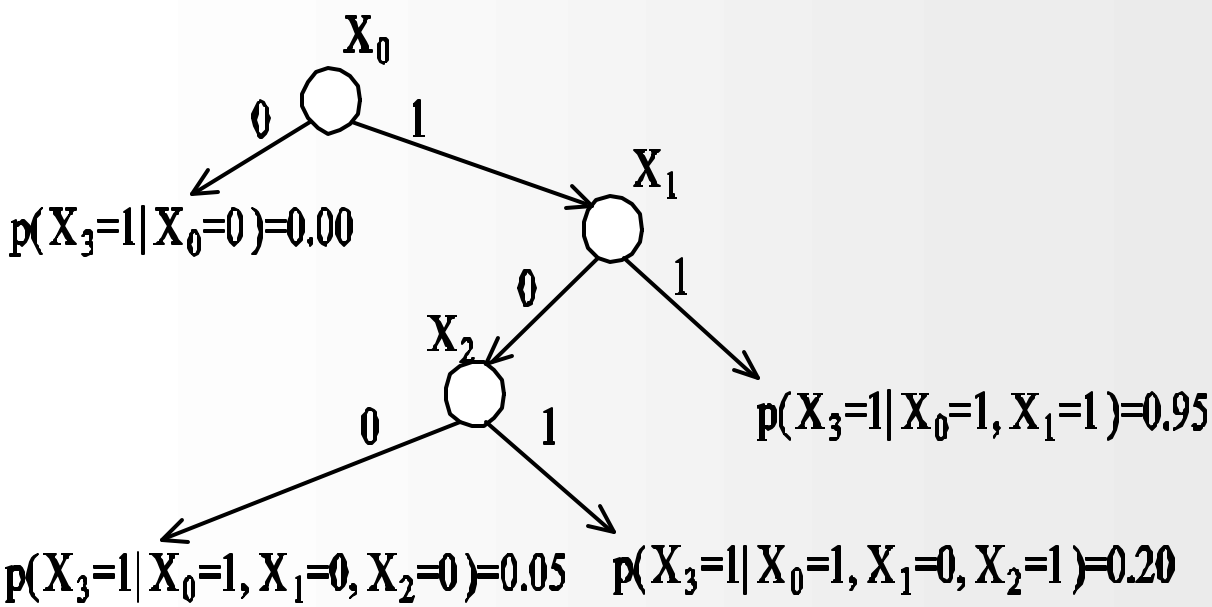






The advantages of decision trees

- Decision trees / graphs are more efficient:
  - Can capture longer building blocks
  - Are more robust



X_0	X_1	X_2	p(X3|X0,X1,X2)
0	0	0	0.00
0	0	1	0.00
0	1	0	0.00
0	1	1	0.00
1	0	0	0.05
1	0	1	0.20
1	1	0	0.95
1	1	1	0.95

Top-down building of decision trees

Population:

x_0	x_1	x_2	x_3
0	0	0	0
0	0	0	0
0	1	0	1
0	1	0	1
1	0	0	1
1	1	0	1
1	1	0	1
1	1	1	0

Split on x_0 :

	x_3=0	x_3=1
x_0 = 0	2	2
x_0 = 1	1	3

gain = 1.40

Split on x_1 :

	x_3=0	x_3=1
x_1 = 0	2	1
x_1 = 1	1	4

gain = 1.92

Split on x_2 :

	x_3=0	x_3=1
x_2 = 0	2	5
x_2 = 1	1	0

gain = 1.65

Sub-population x_1=0:

x_0	x_1	x_2	x_3
0	*	0	0
0	*	0	0
1	*	0	1

Sub-population x_1=1:

x_0	x_1	x_2	x_3
0	*	0	1
0	*	0	1
1	*	1	0
1	*	0	1
1	*	0	1

Split on x_0 :

	x_3=0	x_3=1
x_0 = 0	2	0
x_0 = 1	0	1

gain = 1.54

Split on x_2 :

	x_3=0	x_3=1
x_2 = 0	0	3
x_2 = 1	0	0

gain = 1.07

Split on x_0 :

	x_3=0	x_3=1
x_0 = 0	0	2
x_0 = 1	1	2

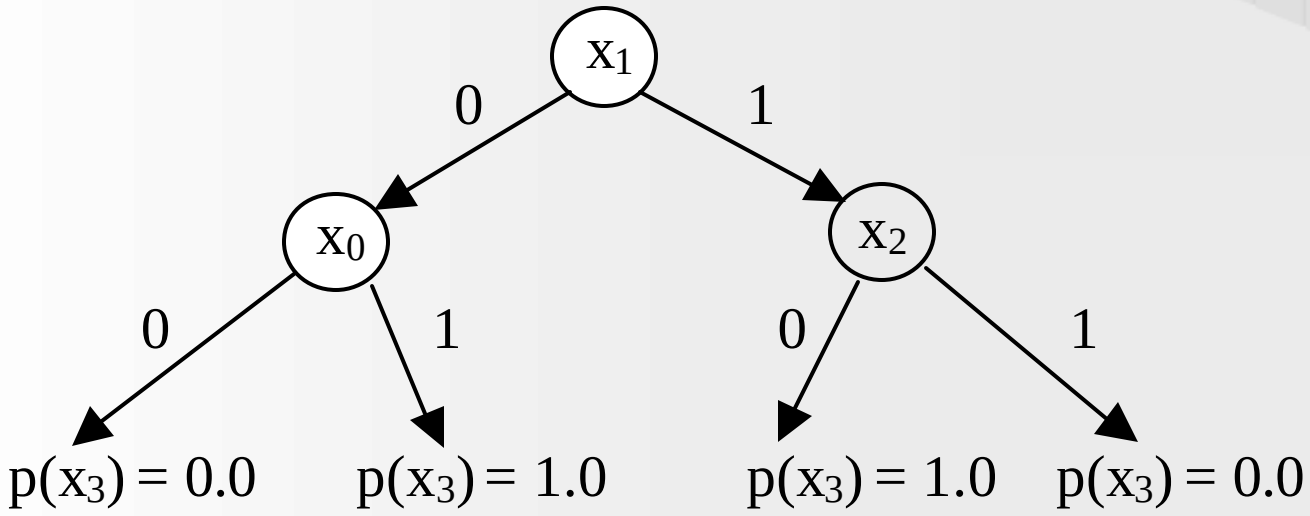
gain = 1.07

Split on x_2 :

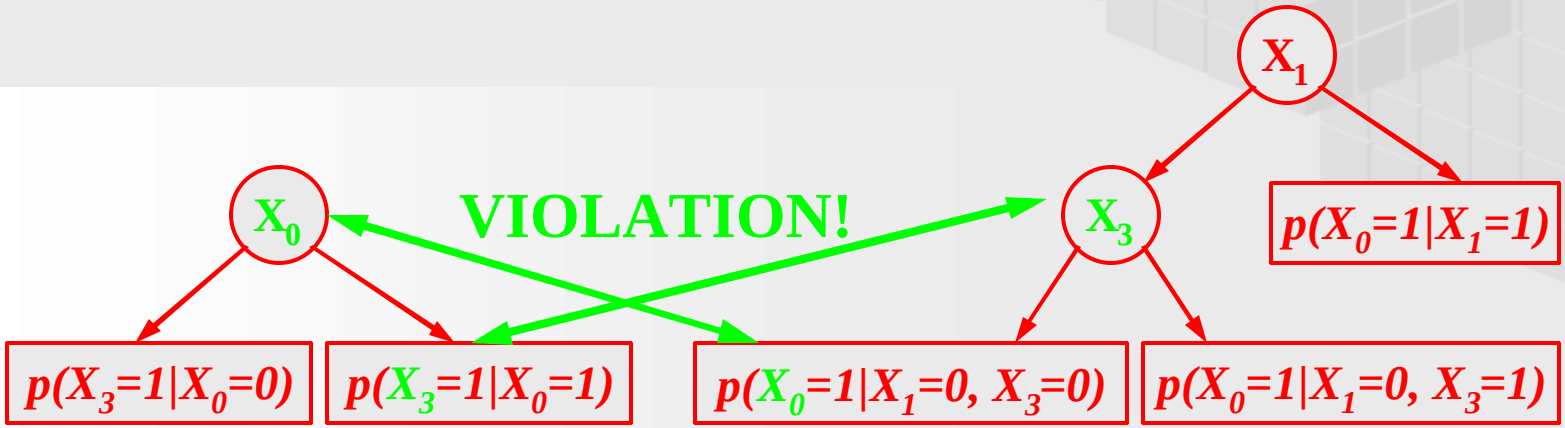
	x_3=0	x_3=1
x_2 = 0	4	0
x_2 = 1	0	1

gain = 2.22

Top-down building of decision trees - cont'd: The resulting decision tree



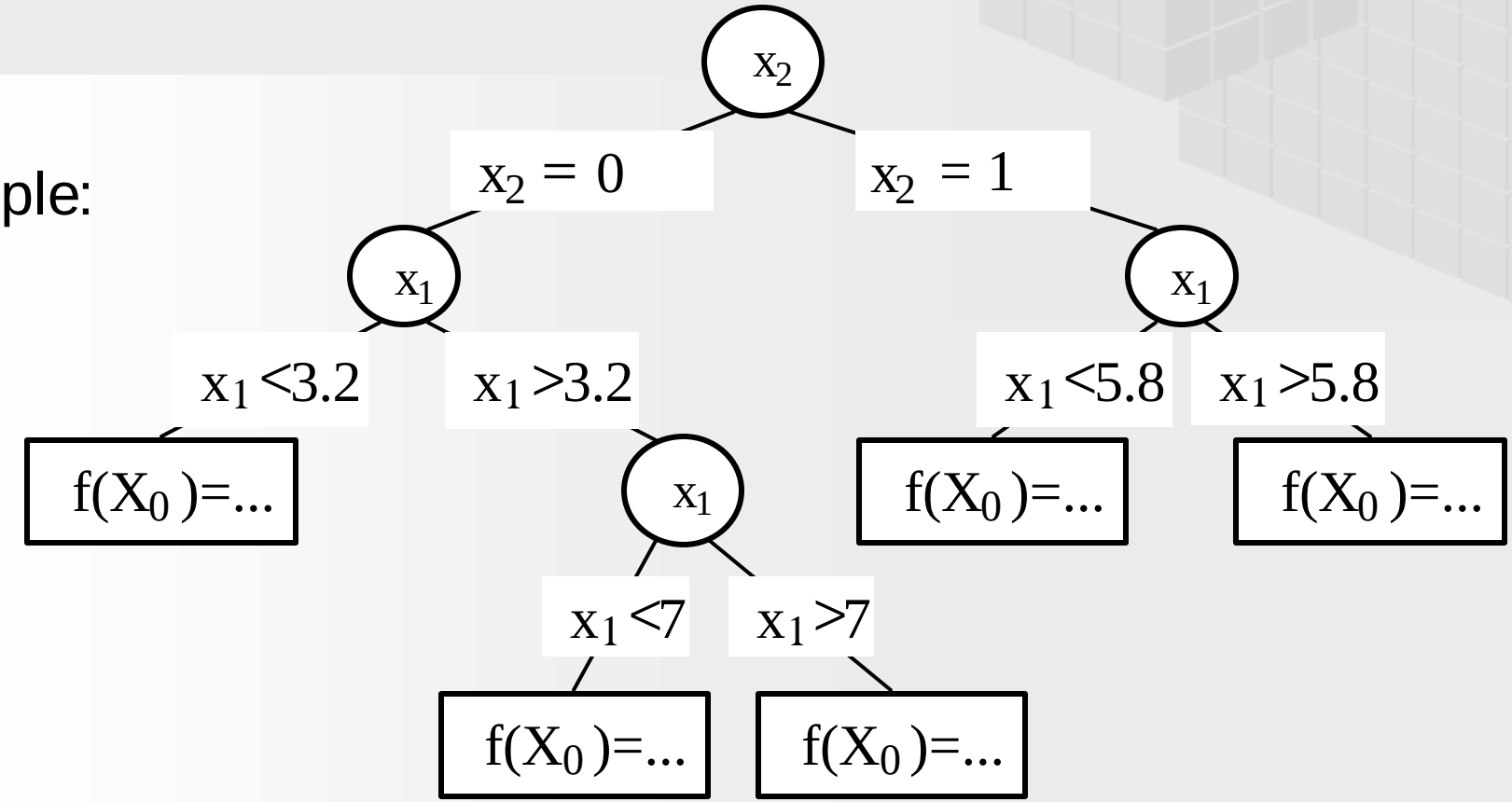
Avoiding cycles



- While constructing all the trees simultaneously, cycles have to be avoided

Mixed decision trees

- Example:



- **Compatible with discrete and continuous domains**

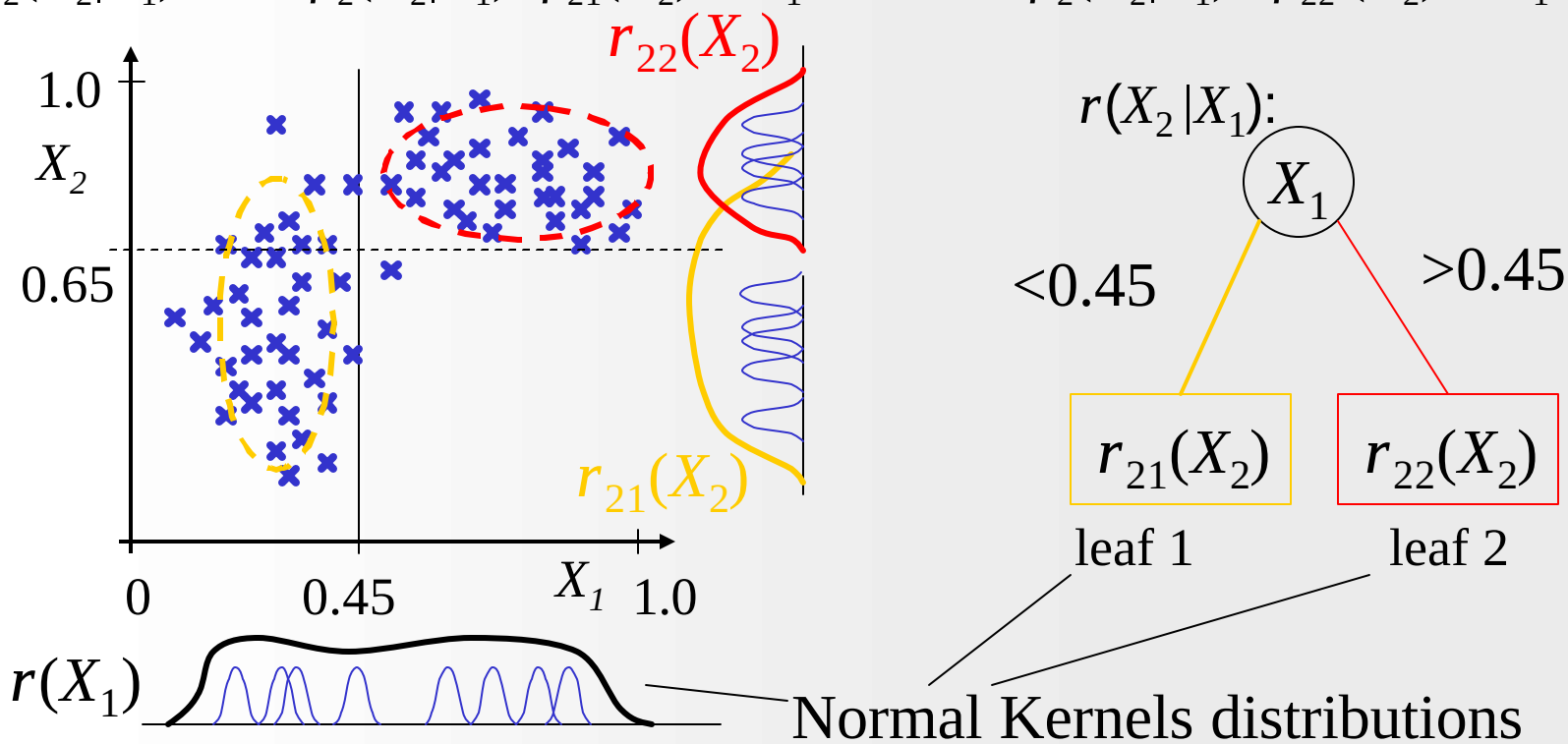
Mixed Bayesian Optimization Algorithm

- Discretization of continuous parameters is an integral part of the learning process.
- 2-level approach:
  - The dependencies are detected using the discretized variables and they are used to partition the search space into subspaces where no correlation appears
  - The Gaussian kernel distribution is used in each subspace to approximate the underlying distribution

Example of probabilistic model in continuous domain (Mixed BOA)

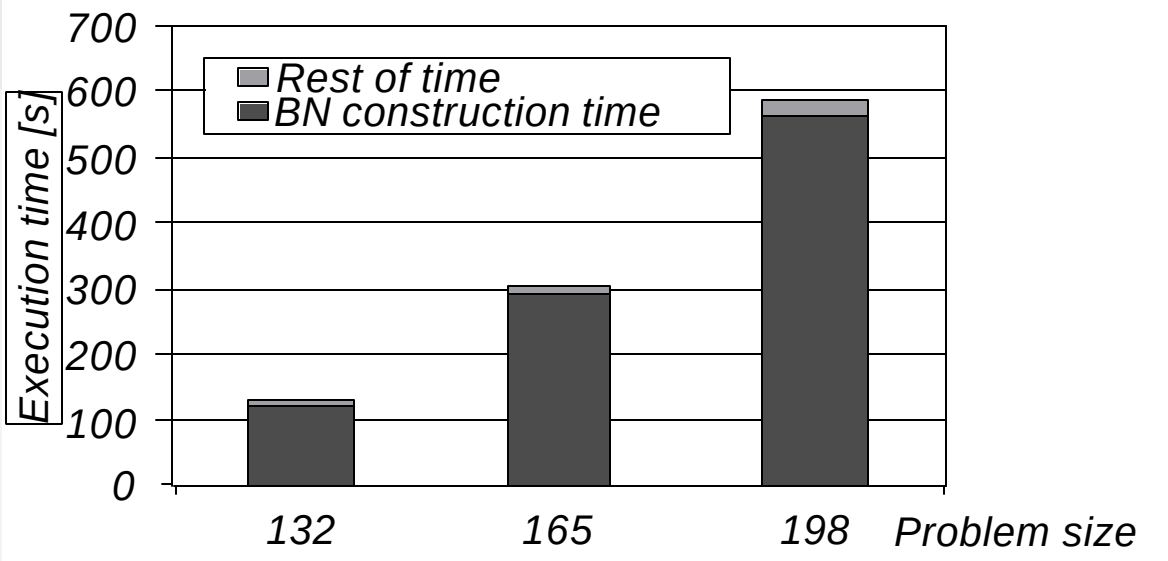
- Decision trees imply conditional factorization
- Example – 2 continuous genes:

$P(X_1,X_2)$ is factorized as $P(X_1,X_2)=P(X_1).P(X_2|X_1)$.
$P(X_1)$ is described by the density function $\rho_1(X_1)$ and $P(X_2|X_1)$ by the density function $\rho_2(X_2|X_1)$, with $\rho_2(X_2|X_1)=\rho_{21}(X_2)$ if $X_1<0.45$ and $\rho_2(X_2|X_1)= \rho_{22}(X_2)$ if $X_1\geq 0.45$

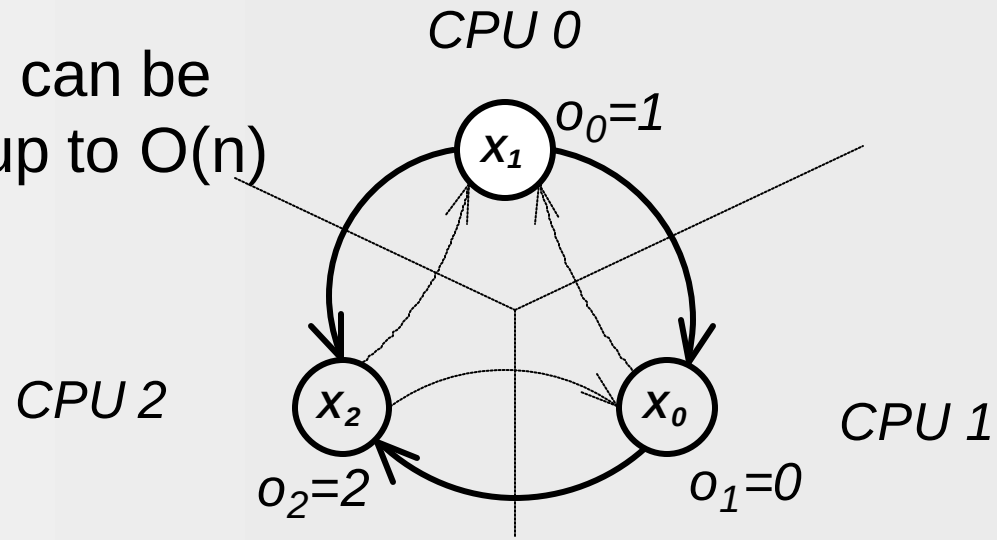


Disadvantages of Model building

- Time profile:



- Solution: model building can be performed in parallel – up to O(n) speedup



GA versus EDA

- Uses the information from individuals in a pairwise manner
- Sensitive to the ordering of parameters
- Recombination is fast

- In EDA the information from all individuals is aggregated in the model
- Not sensitive to the ordering of parameters
- Model building takes time (but can be parallelized)

References

- Holland, J.: Adaptation in Natural and Artificial Systems, Ann Arbor: University of Michigan Press, 1975.
- Goldberg, D. E.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addision-Wesley Publishing Company, 1989.
- Larrañaga, P., Lozano, J. A. (eds.): Estimation of Distribution Algorithms. A new Tool for Evolutionary Computation. Kluwer Academic Publishers, 2002.
- Pelikan, M., Goldberg, D.E., Lobo, F.: A Survey of Optimization by Building and Using Probabilistic Models. IlliGAL Report No. 99018, Illinois Genetic Algorithms Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL, 1999.