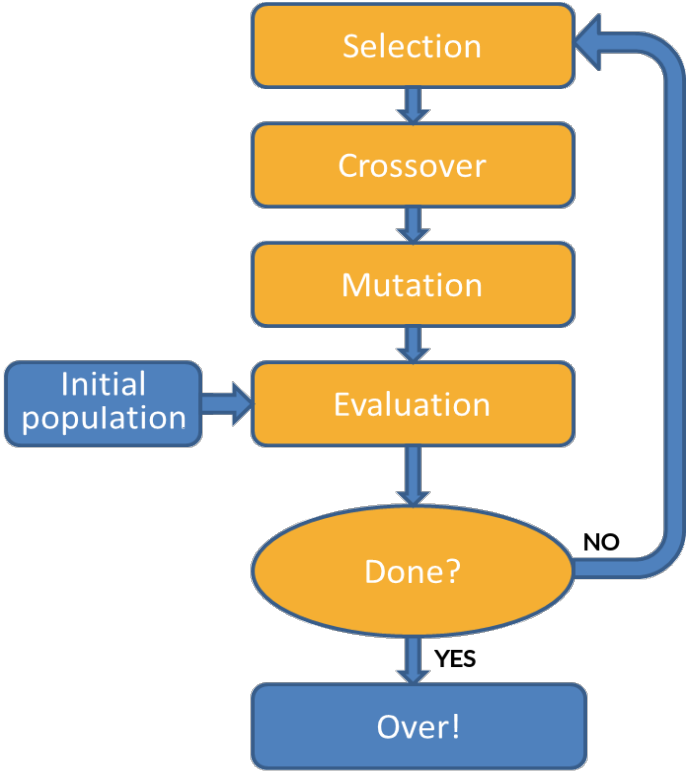
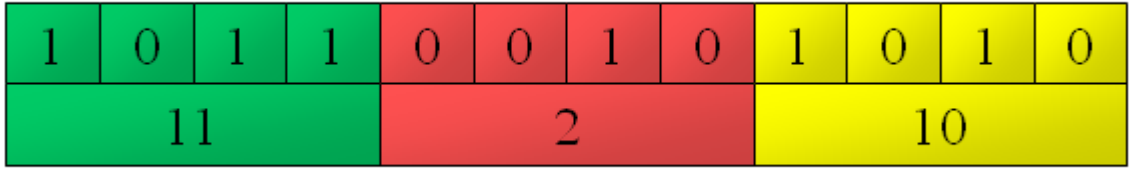
Algoritmos genéticos

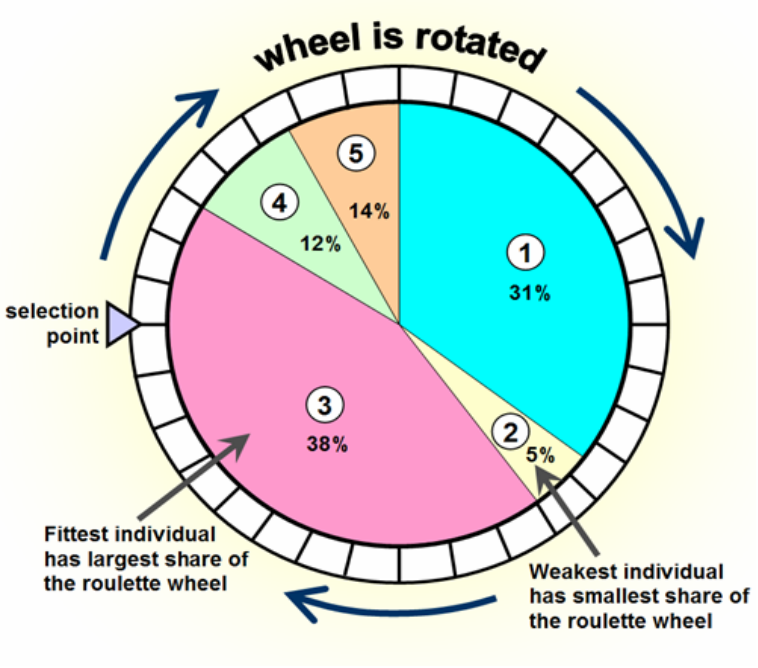


Representación



Operadores: selección

- Ruleta



Probabilidad para el cromosoma m:

$$p_m = \frac{F(c_m)}{\sum_i F(c_i)}$$

Probabilidad acumulada para el cromosoma m:

$$q_m = \sum_{i=1}^{m} p_i$$

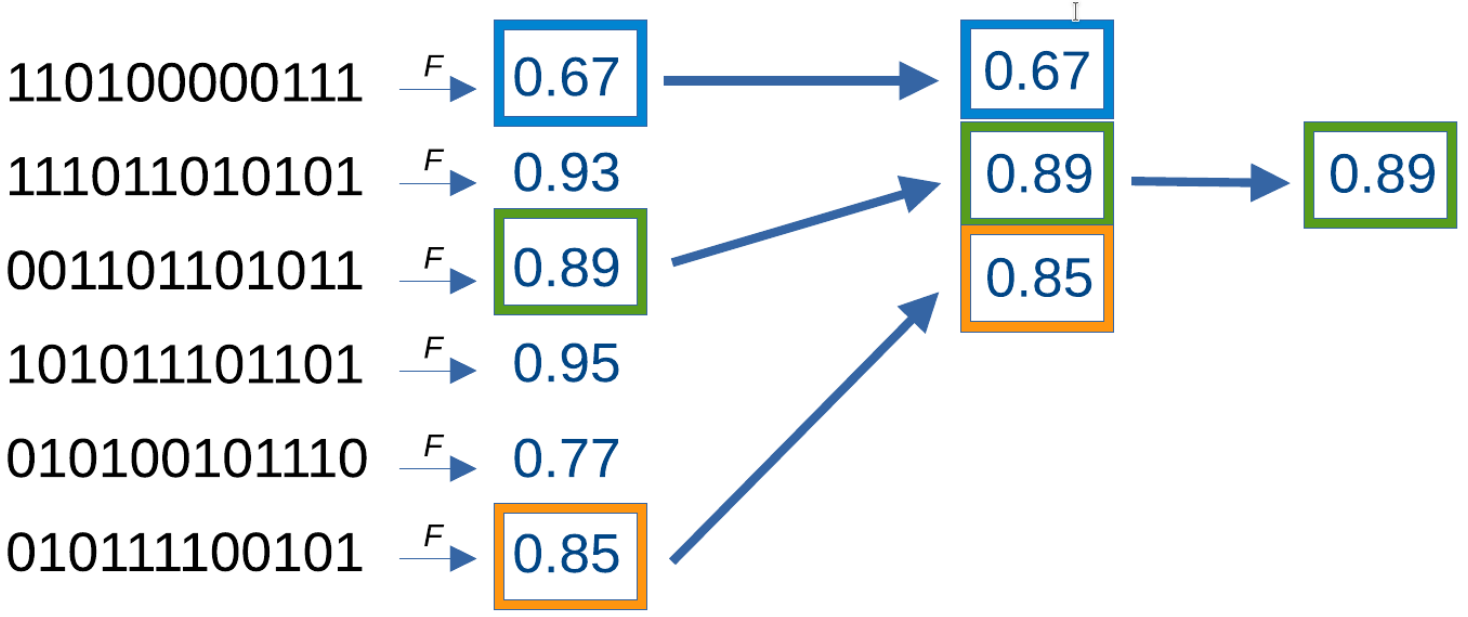
Se genera un r aleatorio entre 0 y 1

Se selecciona c_m que cumpla con :

$$q_{m-1} < r <= q_m$$

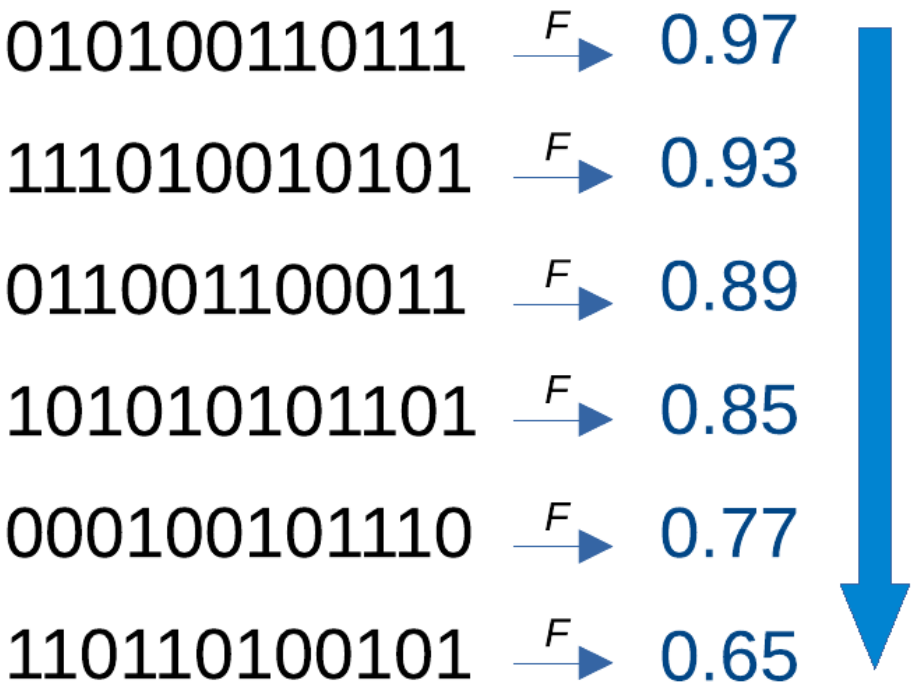
Operadores: selección

- Torneos/Competencias



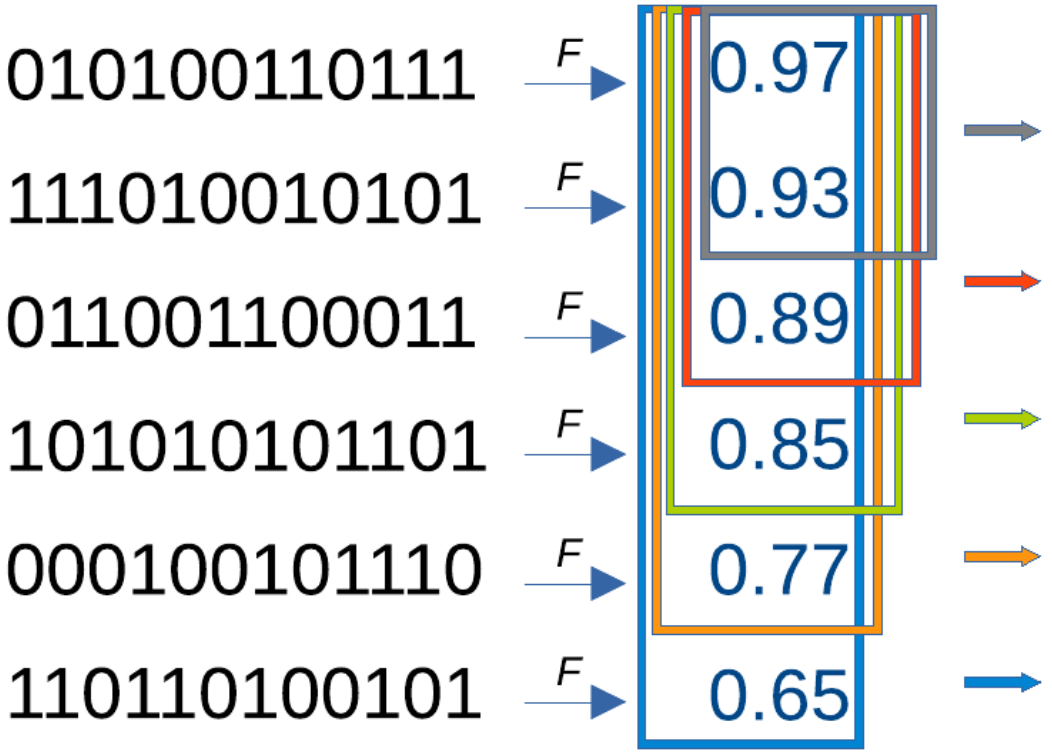
Operadores: selección

- Ventanas



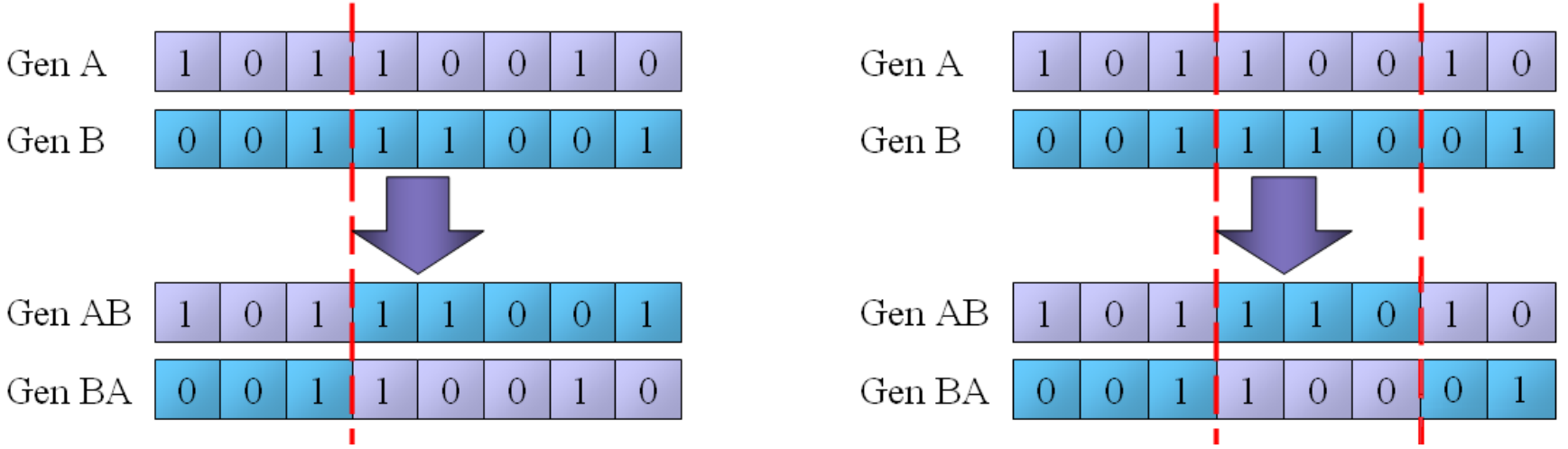
Operadores: selección

- Ventanas



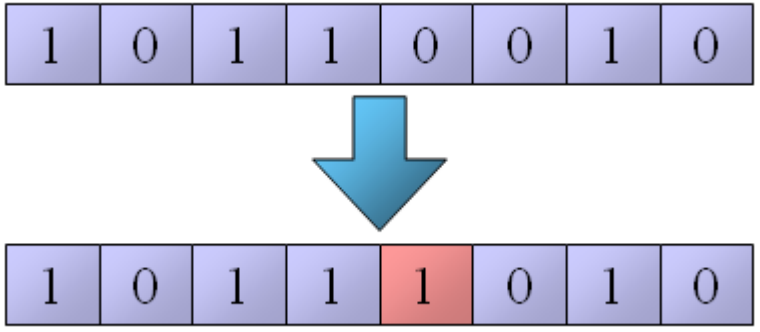
Operadores: Cruza

Probabilidad relativamente alta ~ 0.8-0.9



Operadores: mutación

- Probabilidad relativamente baja ~ 0.1 (a nivel de individuo)



Criterios de finalización

- Cantidad máxima de generaciones
- Fitness deseado
- Sin mejoras en el fitness por n generaciones

Decodificación

Consideremos la cadena binaria $z = a_1, a_2, a_3, ...a_k$:

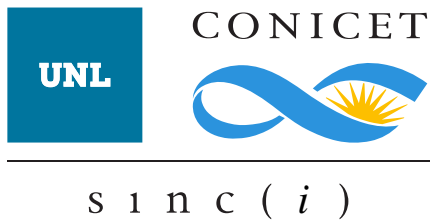
Queremos representar un valor real en $[\alpha, \beta]$:

$$d = \sum_{i=1}^{k} 2^{k-i} a_i$$

$$x = \alpha + d \frac{\beta - \alpha}{2^k - 1}$$

Feature selection techniques

Leandro Vignolo
ldvignolo@sinc.unl.edu.ar



Tópicos Selectos en Aprendizaje Maquinal
Doctorado en Ingeniería,
Mención en Inteligencia Computacional, Señales y Sistemas,
FICH-UNL

October 5, 2023

Introduction

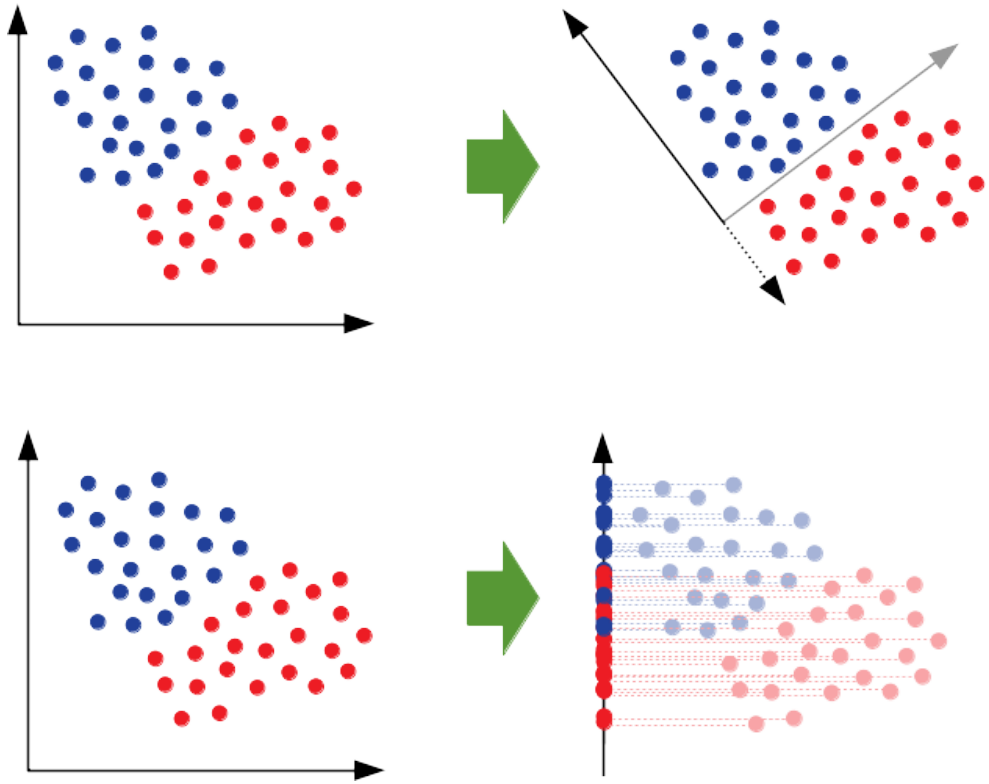
Feature Selection

Problem of selecting some subset of input variables upon which the learning algorithm should focus.

Main objetives of Feature Selection

- Avoid overfitting and improve model performance
- Counteract the curse of dimensionality
- Provide faster and more cost-effective models
- Discriminate between the relevant and irrelevant parts of experience
- ...

Dimensionality Reduction vs Feature Selection

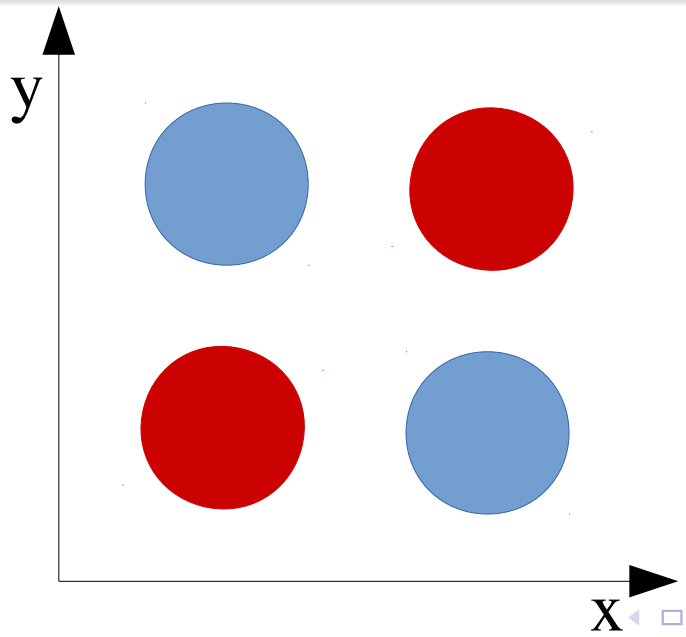


Simple examples

Feature interaction

Can a feature that is useless by itself be useful with others?

- It is tempting to remove the least promising variables (variable ranking methods) before using more complex methods
- Still one may wonder whether some potentially valuable variables could be lost by that filtering process

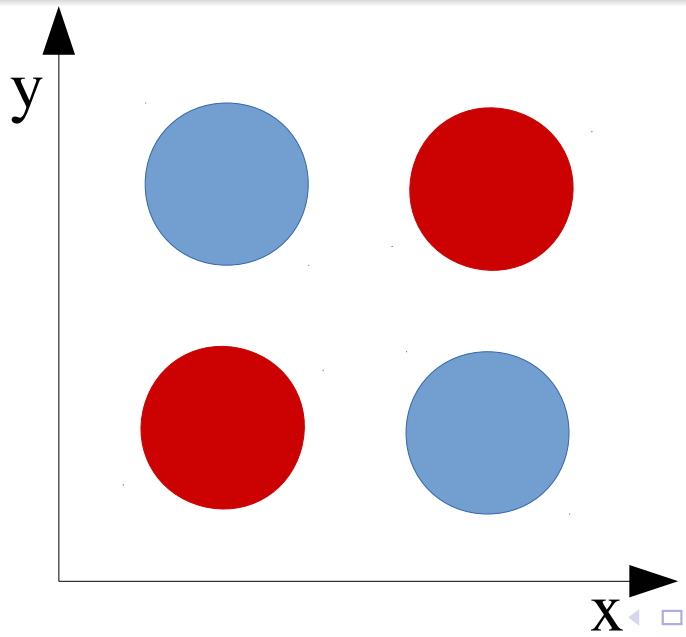


Simple examples

Feature interaction

Can a feature that is useless by itself be useful with others?

- It is tempting to remove the least promising variables (variable ranking methods) before using more complex methods
- Still one may wonder whether some potentially valuable variables could be lost by that filtering process

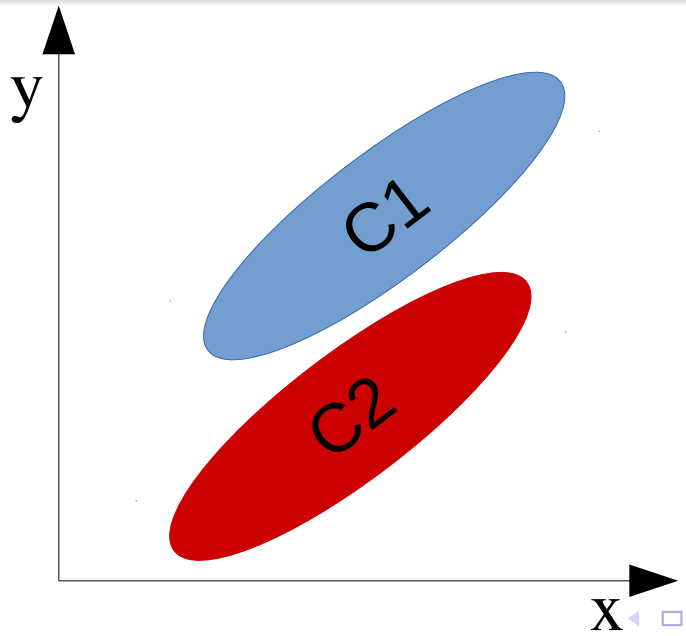


Simple examples

Redundant features

Can presumably redundant variables help each other?

- Several feature selection methods are prone to select non-redundant subsets
- One my wonder wheter considering redundant variables can result in preformance improvement

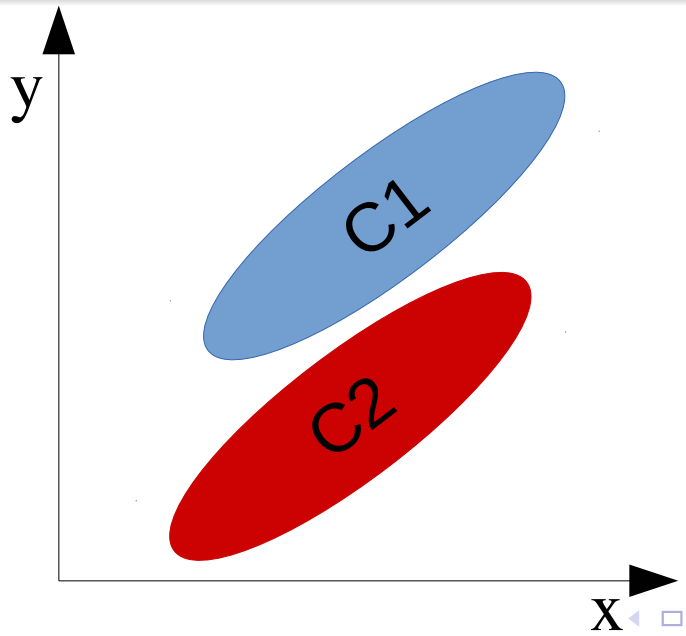


Simple examples

Redundant features

Can presumably redundant variables help each other?

- Several feature selection methods are prone to select non-redundant subsets
- One my wonder wheter considering redundant variables can result in preformance improvement



Feature selection

Multivariate

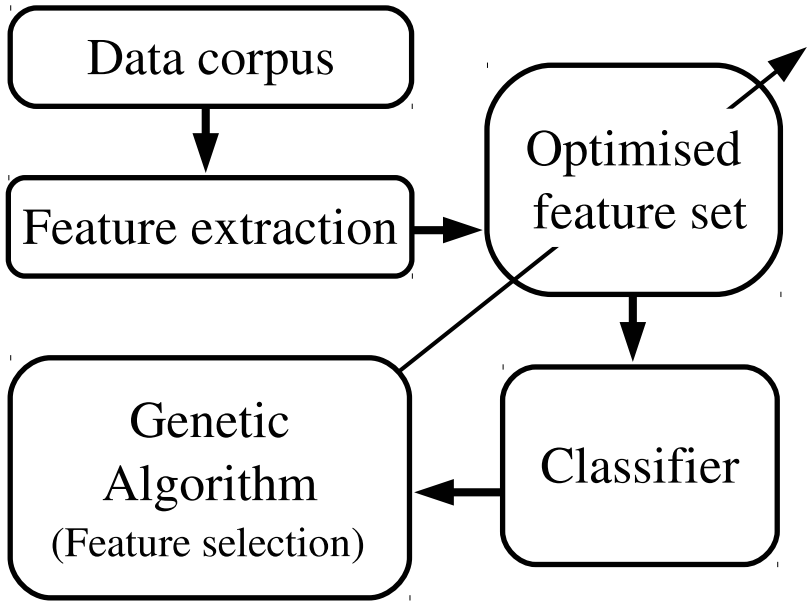
Sub-optimal search

- Supose m features from which we need to select a subset of l.
- All the possible subsets of l out of m features should be considered to guarantee optimality.
- This number of subsets is given by $\frac{m!}{l!(m-l)!}$

  $m = 20, \;\; l = 5 \Rightarrow 15504$

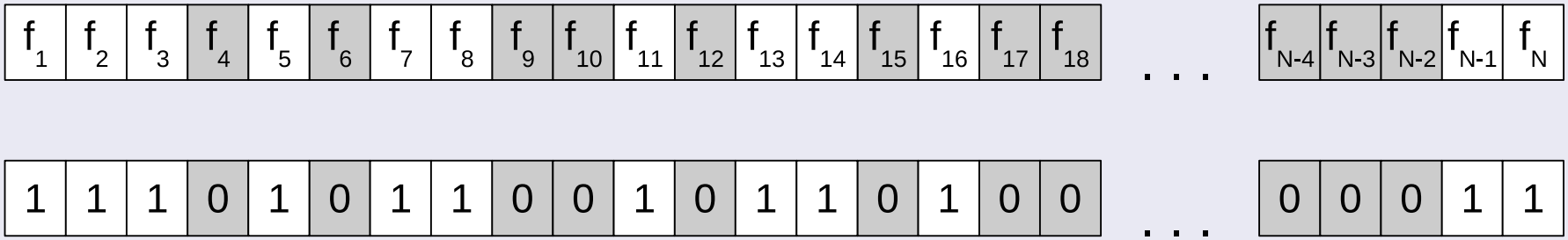
  $m = 100, \;\; l = 50 \Rightarrow 1.01e + 29$
- Notice that here we are considering fixed l.

Genetic algorithms for feature selection



Genetic algorithms for feature selection

Chromosome representation



Binary chromosomes: every individual represents a different selection of features/group of features.

Genetic algorithms for feature selection

Algorithm: Genetic Algorithm

Initialize the population
Evaluate population
repeat
  Parent selection
  Mate selected parents with probability p_c
  Mutate offspring with probability p_m
  Apply population replacement strategy
  Evaluate population
until *stopping criteria is met*

Algorithm: Evaluate population

for *each individual in the population* **do**
  Determine feature subset based on chromosome
  Arrange data samples with feature subset
  Train the classifier on the training set
  Test the classifier on the validation set
  Compute performance to assign fitness

Genetic algorithms for feature selection

Algorithm: Genetic Algorithm

Initialize the population
Evaluate population
repeat
- Parent selection
- Mate selected parents with probability p_c
- Mutate offspring with probability p_m
- Apply population replacement strategy
- **Evaluate population**

until *stopping criteria is met*

Algorithm: Evaluate population

for *each individual in the population* **do**
- Determine feature subset based on chromosome
- Arrange data samples with feature subset
- Train the classifier on the training set
- Test the classifier on the validation set
- Compute performance to assign fitness

Genetic algorithms for feature selection

Objective Functions

- The objective function evaluates the feature subset.
- The classifier is trainined and tested using the selected features.
- Other measures, as dimensionality, can be included included as well.

$$Fitness = Accuracy$$

$$Fitness = \alpha Accuracy - \beta \frac{\#\text{ selected features}}{\#totalfeatures}$$

Dataset Leukemia

Leukemia

- Datos de expresión génica obtenidos a partir de micro-arreglos de ADN para la clasificación de dos tipos de cáncer.
- El objetivo es discriminar dos clases: ALL (Leucemia Linfocítica Aguda) y AML (Leucemia Mielógena Aguda).
- Cada muestra consiste en **7129** variables o características

	Train	Test	UAR
ALL	27	20	
AML	11	14	
Total	38	34	0.82

sinc(i)