GENETIC PROGRAMMING CLASSIFIERS and SYMBOLIC REGRESSION

Vic Ciesielski
School of CS&IT
RMIT
vic.ciesielski@rmit.edu.au

SUMMARY

- Evolutionary Algorithms
- Genetic Programming Methodology
- Examples
  - Classification
  - Attribute Selection
  - Symbolic Regression
  - Feature Construction

High Level Evolutionary Algorithm

1. Initialize population of potential solutions
2. Evaluate fitness
3. Select by fitness
4. Crossover & Mutation
5. Generate new population
6. Go to 2 or stop

Evolutionary Algorithm

Current Population

Individual	Fitness
Parent1	0.1
Parent2	0.2
Parent3	0.4
Parent4	0.5

New Population

Individual
Child1
Child2
Child3
Child4

Favour fitter individuals (eg lower error) when selecting parents
The new population becomes the current population

Some Specific Variations

- Genetic Algorithms
- Genetic Programming
- Particle Swarms
- Differential Evolution
- Ant Colony

GENETIC PROGRAMMING

- An individual is a program
- Crossover will combine pieces of parent programs to get children
- Mutation will make a random change to a program

FUNCTIONAL PROGRAMS

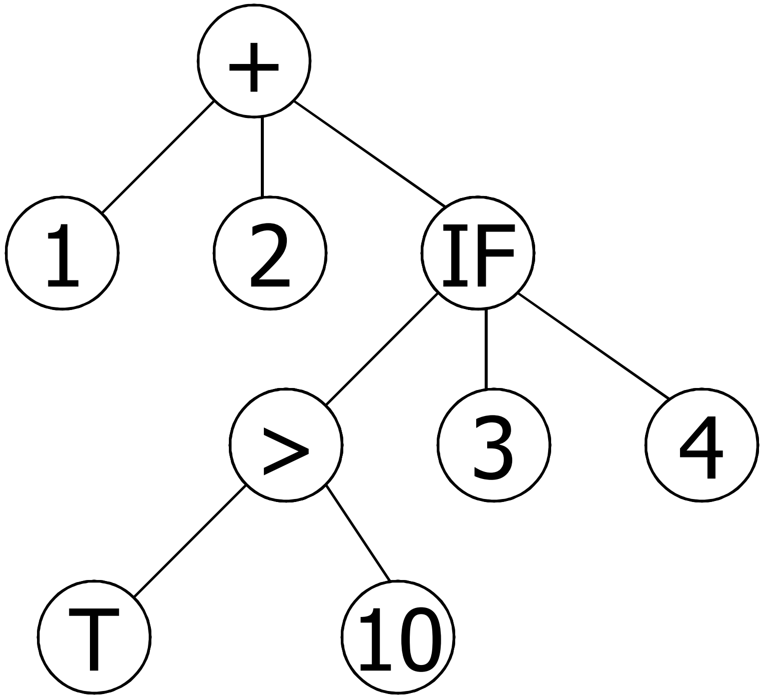
- Form of a function

 (FUNCTION-NAME ARG1 ARG2 …..)

- The arguments are evaluated, the function is applied to the arguments and value returned.
- (+ 1 2 3) evaluates to 6
- (+ (- 3 2) (* 2 4) becomes (+ 1 8) which is 9
- (IF (> TIME 10) 3 4) evaluates to 3 if TIME is 11 or more and to 4 otherwise
- The state of the art in GP does not yet extend to the kinds of programs we are accustomed to writing in C, C++ or Java

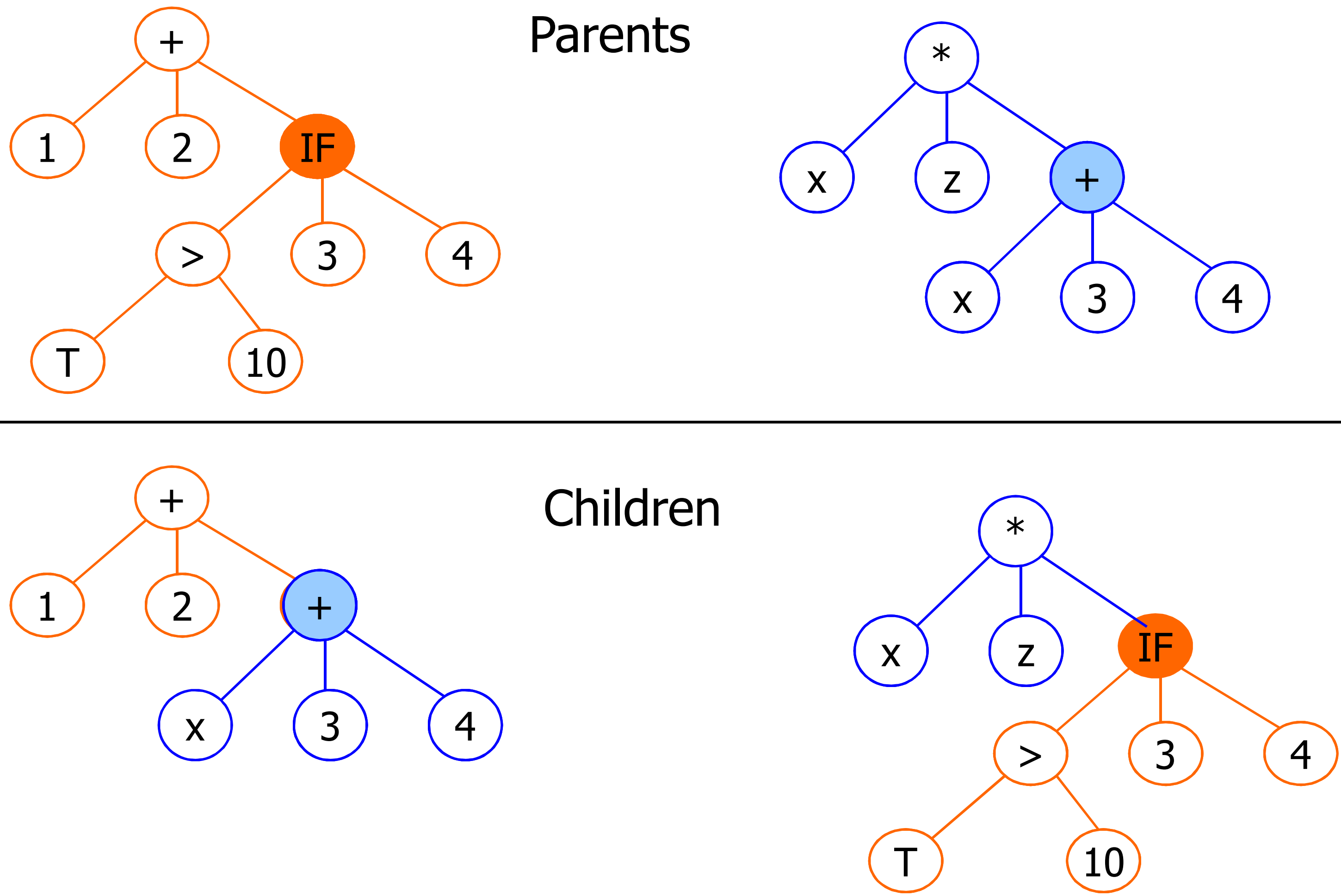
PROGRAMS AS TREES

(+ 1 2 (IF (> T 10) 3 4)

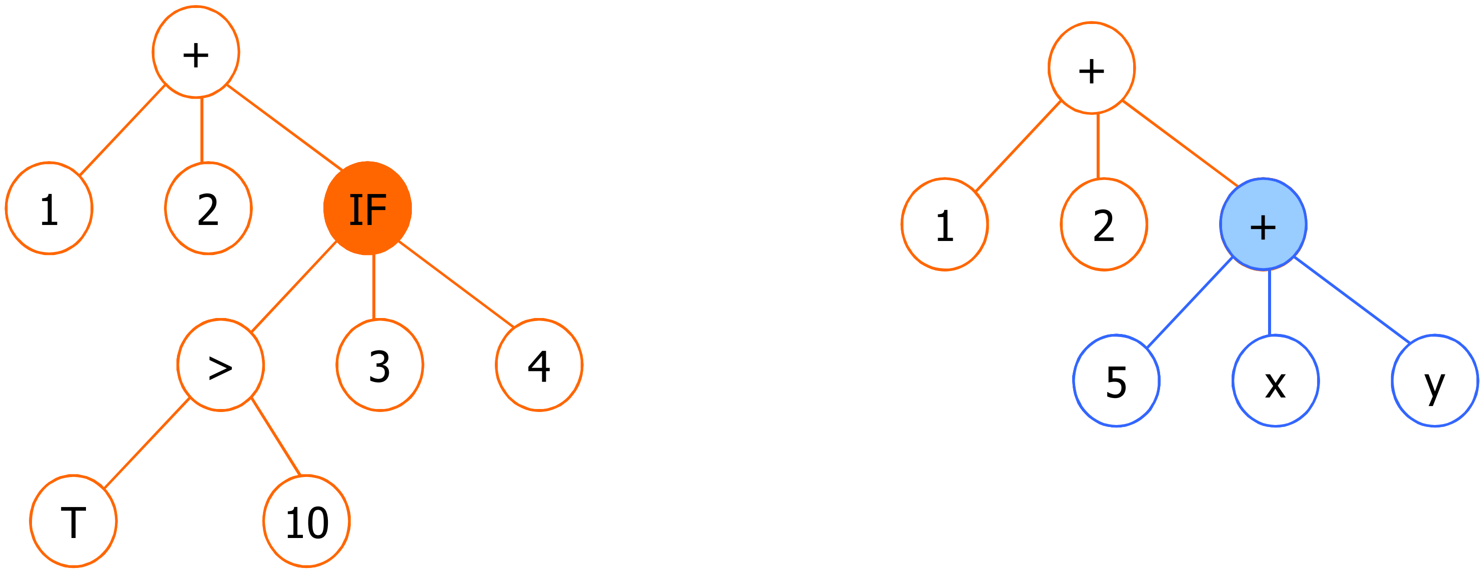


CROSSOVER

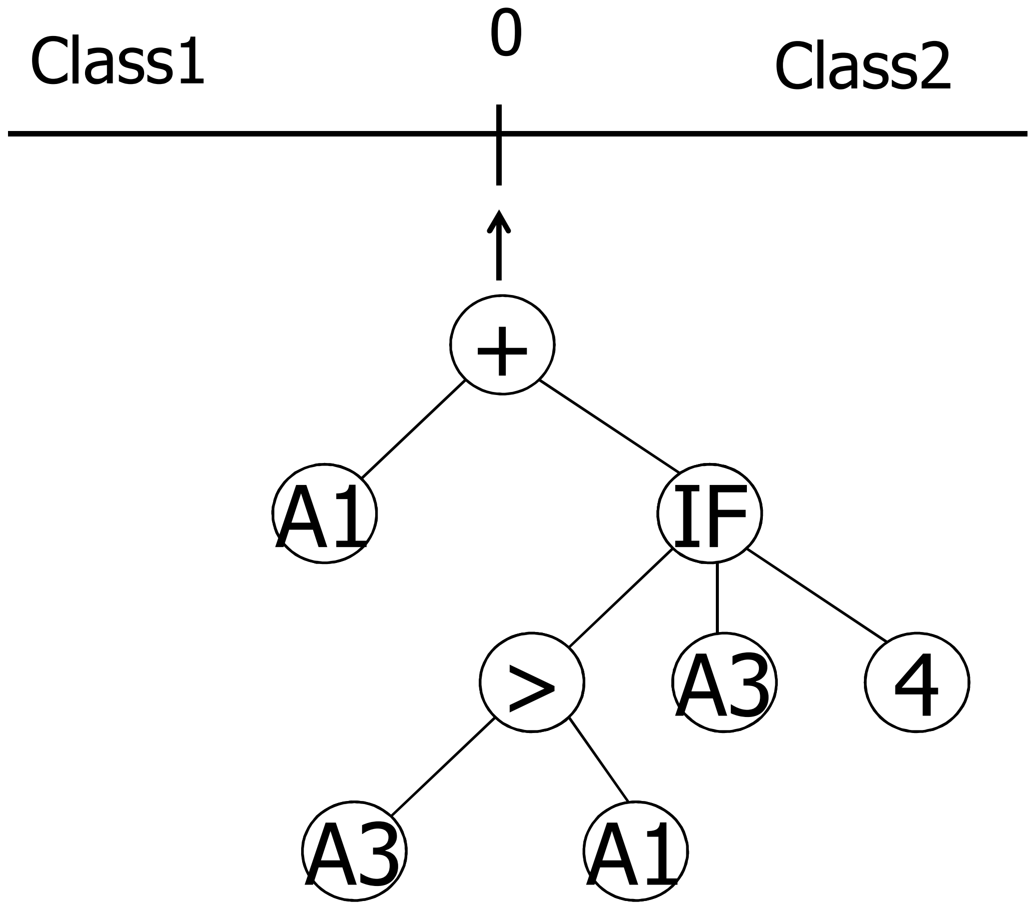
Two parents exchange subtrees



MUTATION

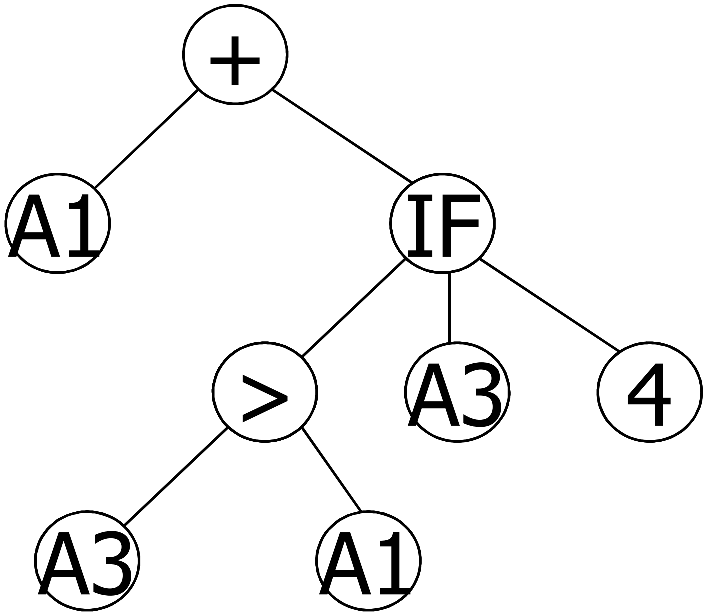


GENETIC PROGRAMMING CLASSIFIERS



GENETIC PROGRAMMING CLASSIFIERS

- If output <= 0 then class1 else class2
- A1 A2 A3 CLASS
- 2 1 3 class2 Correct
- -3 1 2 class1 Incorrect



STEPS IN GENETIC PROGRAMMING

1. Determine the set of terminals
   - Inputs to desired program
2. Determine the set of primitive functions
3. Determine fitness measure
   - Defined for every composition of functions
   - Usually the error between the output of the program and correct result, averaged over a number of inputs

STEPS IN GENETIC PROGRAMMING

4. The parameters for controlling the run
   - Population size
   - Maximum number of generations
   - Crossover, Mutation rates, Max size of a program
5. The method for designating the result and stopping
   - Best so far
   - Error is small enough, max generations reached

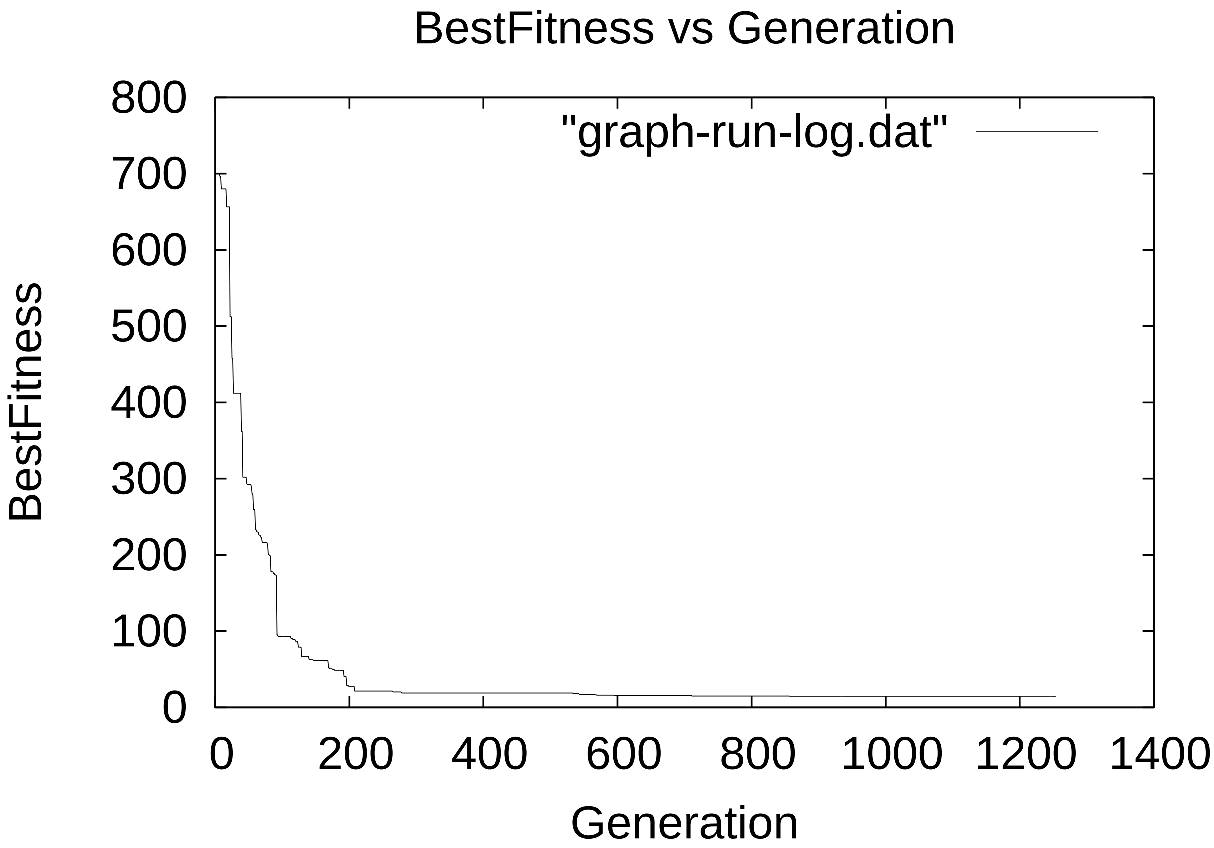
Primitive Functions

- Arithmetic and logical functions
- Protected division (%) returns 0 if denominator is 0
- Random number generator
- Domain specific functions
- A program is a tree composed of functions and terminals

FITNESS

- Use training classification error

A GP Run


BestFitness vs Generation
"graph-run-log.dat"
BestFitness
Generation
0
100
200
300
400
500
600
700
800
0
200
400
600
800
1000
1200
1400

Building a GP Classifier

- Do 10 (or more) training runs
- Select the best evolved individual
- Apply to the test data
- Training is slow
- Evolved classifier is very fast
- Good for constructing ensembles

Attribute Selection

- Evolve a classifier
- Attributes in the evolved program are relevant, others are not
- Repeat n times
- Attributes occurring most often are most relevant

Symbolic Regression

- Fit a formula to some data
  - Data from an experiment
  - Data from a time series
  - Could be several variables
- Example
  - $y = x^3 - 2x^2 + x + 0.5$
  - $z = 0.5\log(x) / (0.8 + y)$

Some Observed Data

y	x
0.382	0.241
0.724	0.616
1.0	1.0
1.524	1.881
5.199	11.855
9.539	29.459

What is the relationship?

$y = ax + b$

$y = ax^2 + bx + c$

$y = ax^3 + bx^2 + cx + d$

$y = sin(x)$

$y = xsin(x)$

$y = x^3$

GENETIC PROGRAMMING CAN BE USED TO FIND IT

POSSIBLE FORMULAS

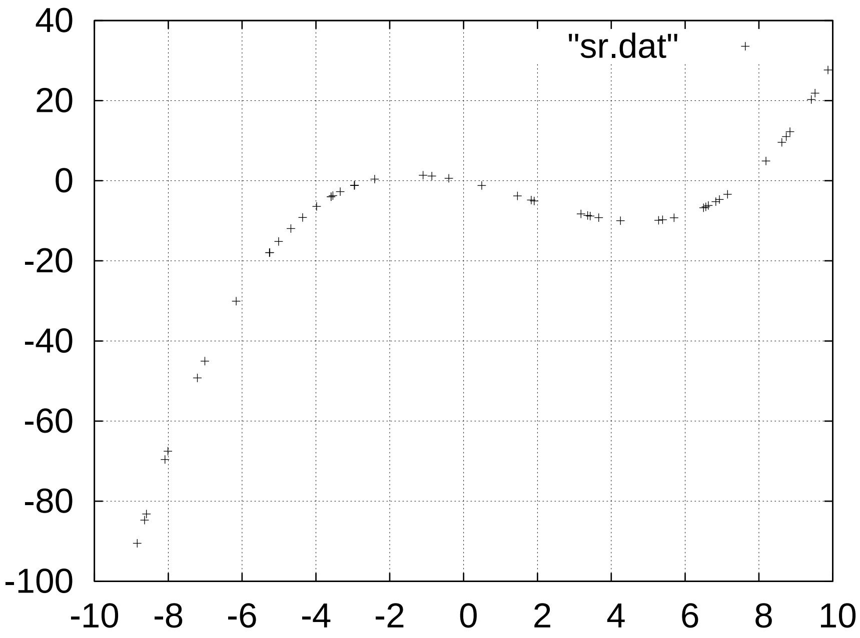
- Y = (+ (* (* -0.534711 (* (+ X -0.038049) (* 0.632555 0.475556))) (* (* (* -0.702996 -0.436195) -0.038049) -0.038049)) X)   [High Error]
- Y = *(+ (* X (+ X (* X X))) (/ (+ X (* X X)) X))*

  $=x^3+x^2+x+1$ *[High Error*
- Y = (** (** X 2) (/ 1 3)) [Low Error]

  $y^3 = x^2$
- Discovery of Keplers law of planetary motion

Some Data from an Experiment

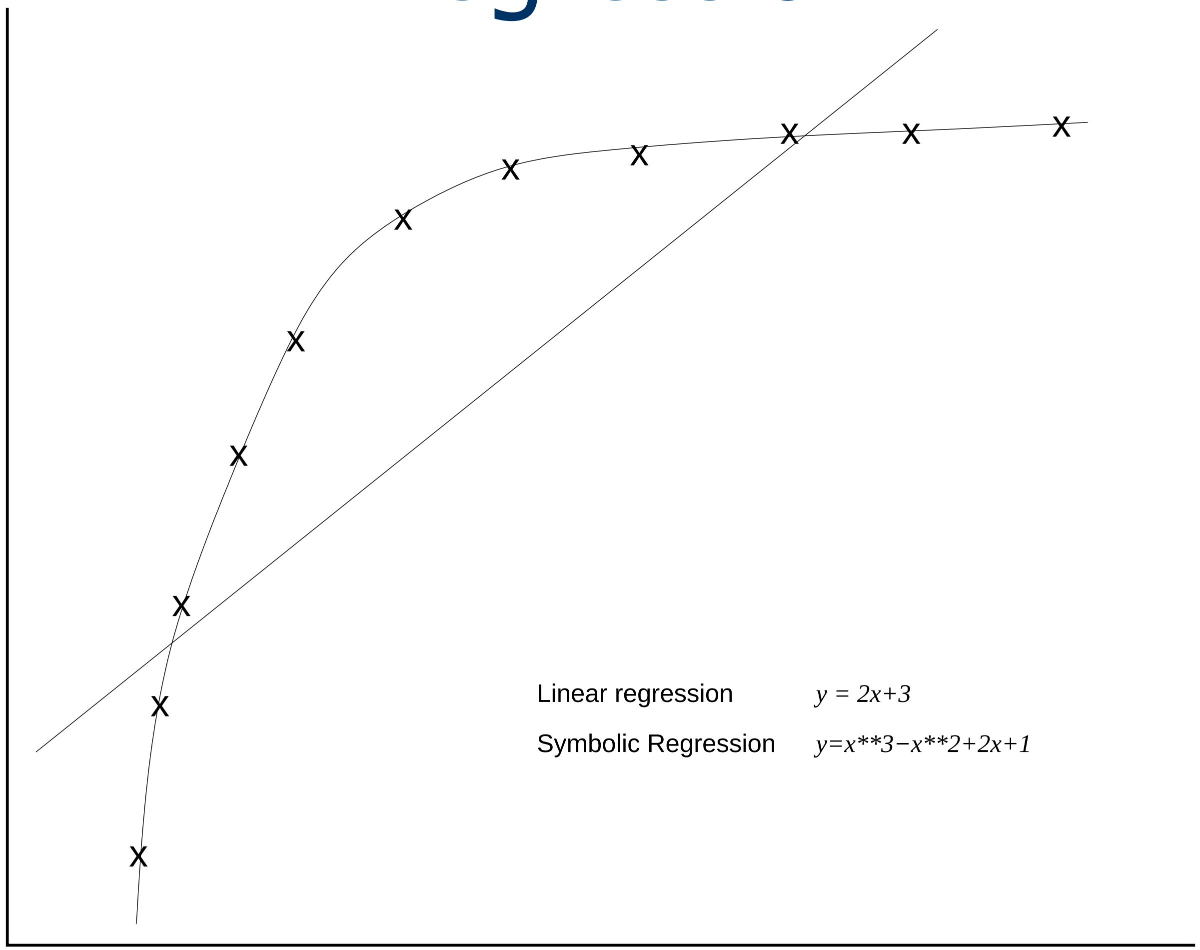
Y	Y
1.46	-3.77
8.84	12.24
-2.96	-1.15
-7.01	-45.03
9.87	27.65
-2.95	-1.11
6.63	-6.19
3.66	-9.22
8.74	11.01
-3.59	-3.99
8.62	9.59
5.70	-9.24
-5.01	-15.17
-3.98	-6.40
6.56	-6.50
-3.54	-3.71
3.43	-8.80
1.83	-4.82
-8.59	-83.20
-5.25	-17.91
-6.16	-30.06
6.83	-5.23
-4.36	-9.18



Linear Regression vs Symbolic Regression

- Linear regression will fit a line that goes through a few points
- Symbolic regression will fit an equation that goes through many points

Linear Regression vs Symbolic Regression



GP Setup for Symbolic Regression

- Terminal set: *{x, rand}. Rand* produces random numbers in [-1.0,1.0]
- Function set: *{+,-,*,%}*
- Fitness cases: 50 random *x* values in [-1, 1] and corresponding *y* values
- Fitness Measure: Sum of errors for 50 cases
- Parameters: Population=4,000, Max generations=51
- Success: Error less than 0.0001

The Best Evolved Formula

Fitness 2.20374
Depth 6
Size 39
Program (/ (+ (- (/ (- X 23.810541) X) (/ (-44.444105 X) 6.228828)) (+ (/ (* X 13.483077) (* 20.075076 13.483077)) X)) (/ 13.483077 (- (+ (/ X 10.382397) (* X X)) (/ (d* X X) (- 5.178991 13.788263)))))

Note the "Bloat"

A Perfect Result

- Generate some ‘experimental’ data from $y=x^3+x^2+x+1$
- Perform the GP run
- After 12 generations the BestFitness is 9.69447e-13
- Best individual is

(+ (X (+ X (* X X))) (/ (+ X (* X X)) X))*

- Which simplifies to above formula

Feature Construction

- Perhaps a new feature which is a combination of the original features will be very good for classification
  - Eg (A1 – A2) *A3
  - Eg (A4 +A5) / (A3 –A2)
- Genetic programming can be used to find such features

Multi-Objective

- Tradeoff between false positives and false negatives
- Unbalanced data, true positive accuracy

Data Mining and Machine Learning Research at RMIT (Vic)

- Genetic Programming for data mining
- Prediction of bi-polar manic episodes from Facebook or Smartphone activity
- Image Mining, Evolution of features for image classification
- Algorithms for Deep Learning (Massive Neural Networks)

Data Mining and Machine Learning Research at RMIT

(Xiaodong Li)

- Personalized journey planning for travellers on public transportation networks
- Data analytics for time series prediction data
- Optimizing Deep learning convolutional neural network architectures using evolutionary algorithms
- Deep learning for solving real-world image classification problems

Data Mining and Machine Learning Research at RMIT

(Andy Song)

- Data driven optimisation
- Text mining
- Time series analysis by evolutionary computation
  Machine vision, image recognition (by EC)

Machine Learning Research at RMIT

Tim Wiley

- Learning Autonomous Robot Behaviours
- Reinforcement learning

Data Mining and Machine Learning Research at RMIT
(Jeffrey Chan)

- Machine Learning
- Itinerary Recommendation
- Social Network Analysis

Data Mining Research at RMIT
(Jenny Zhang)

- Social network data mining
- Event detection on Twitter
- Anomaly detection in information-rich networks
- Information credibility on Twitter
- User behaviour analysis using online newspaper Web server logs

Data Mining Research at RMIT
(Flora Salim)

- Human mobility mining from smartphone and wireless infrastructure data
- Spatiotemporal clustering of urban sensor data
- Semi-supervised learning of user profiles
- Multi-resolution time-series forecasting
- Deep learning of trajectory and sensor data
- Context, intent, and behaviour recognition for intelligent assistants