David Goldberg → GA book [1-3]
James Holland → inventor of GA.

- Enumerative search → point by point search. Also called try and error. (but it is random).
- Operations Research → mathematical optimization methods.
- Derivative based → Gradient ascent or descent. Hill climbing is also another word. A slope should always exist. But real life data is not always continuous. Also, the data can be multimodal, the hill climbing gets stuck in local optimum.

In the search for optimal value, the path towards optima is also important.

Where will GA fit in? → Classification.
If we are not able to give unique class labels in D-tree, it is non feasible.

Convergence → how quickly will I reach the leaf node.

Efficiency vs efficacy.

28/1/2019

- Why EVOLUTIONARY?
Can my overall objective keep bettering. GA, Tabu Search etc are search algos that search for optimal.

Fitness of a solution :- how good is a solution.

iteration ≡ Generation.
As we do more generations, the fitness should improve.

Darwin's theory of evolution forms the crux of the GA's.
GA → search algo based on natural selection + Genetics.

The power of ~~it~~ evolutionary algos comes in if the search base is really large and the function is too ~~to~~ complex.

- Objective function :- it is to be identified. This is the first step.
- Initial population :- random collection of solutions

(2, 6, 8) → 3 random search points from search space.
↳ "Roll of a die"
Can be one approach of constructing the initial population.

Using the initial population can we do some Genetics (biological ops).

$f(x) = x^2$.
$[0, 31]$.
initial popn = 2, 6, 8 → max $x^2 = 64$.
using genetics = 5, 9, 11 → max $x^2 = 121$.
we have to reach (31). ⋮ 31^2.

Human "flair" of a flexible search procedure.
Biological system solutions are far more robust than the existing search procedures.

So, can the innovative search brought into GA?

GA
- Selection
- Crossover
- mutation.

] know atleast 1 way of each.

Application specific operators have improved GA over time.

Search space = 10k. (larger the search space, larger can be the initial popn)
Initial popn = 20 pts. (random)
Selection opn :- out of 20 pick only those which are fitter ones.

Allele → attributes → Chromosomes.
Initial popn :- is a collection of chromosomes.
In the probability of fitter solution. Cross parents who have fitter characteristics.

After selection of the fit ones, cross over.
Mutation is hardly recommended. It is done when stuck in the local optima.

Roulette wheel → simple selection (random)
↳ minimizes the owner's loss.

Initial popn = 13, 24, 8, 19.

$f(x) = x^2$

169 576 64 361.
1 2 3 4 ← Chromosomes.
(2) and (4) → better fit

Sum of fitness = 1170. (wheel dimension) (sample space)

Probability of selection = $f_i / \Sigma f_i$.

$= \frac{13}{1170} \quad \frac{24}{1170} \quad \frac{8}{1170} \quad \frac{19}{1170}$. → Sum is 1
1 2 3 4.

$\frac{13}{1170}$ → percent of the wheel is occupied by 13 on the wheel.

0.14	0.49	0.05	0.30.
1	2	3	4.
14%	49%	5%	30%.

Assume that the population size is 4 (fixed).
So cross over a population of size 4.
We don't want any copy of 8.

29/1/2019

- How does GA gell well with classification?
Information gain using entropy, gini index etc. Choosing the optimal path in Dec. Tree. There might more than one DecTree for one classification. We must be able to choose the optimal path from so many trees.

Expert system forms the base for Bucket Brigade classifier.
Expert system :- control is given by domain experts
ELIZA, ALEXA.
Expert systems = Training model.
Expertsystems are basically classifiers. Paths in DecTree is rules in expert systems.

The fitness of the path is important.
GA strikes the balance between efficiency and efficacy.

Efficacy + effective → closer to feasibility.
The solution must be ROBUST too.

Select from existing ↑ generate new fitter points. ↑

(Select → crossover → mutate) → Cumulative in nature

Communique Page No. Date

Efficacy in searching is Relevance. It is qualitative measure.

One solution is a chromosome in GA terminology. Bucket Brigade is a multiclassifier approach.

Look at the past fit solutions. Discourage unfit solutions, encourage fit solutions.

Index	x	Chromosome (16 8 4 2 1)	f(x) ↷ fitness.
1	13	0 1 1 0 1	169
2	24	1 1 0 0 0	576
3	8	0 1 0 0 0	64.
4	19.	1 0 0 1 1	361

$\max(x^2)$ $[0, 31]$.

Result of 20 coin tosses. 1170 → population fitness.

(PP = 4)

1. Frame the FP.

x can be precision, recall, Fscore, anything

$\bar{f}$ = 293 → avg fitness of popn.

Keep track of current maximum → here 576 (flip of the wheel.)

↷ Prob of selection on Roulette wheel.

Index	$f_i / \Sigma f_i$	Expected Count $(f_i / \bar{f})$	Round
1	.14	.58.	1
2	.49. → draw a wheel.	1.97	2
3	.06	0.22	0
4	.31	1.23.	1

Flipping the wheel is selection.

If Round = 1 then we got it once.

2 then we got it 2ce. (2 copies)

Next generation is the output of selection.

Population size is the same.

0 1 1 0 1

1 1 0 0 0

1 1 0 0 0

1 0 0 1 1

The essence of selection is duplicate fitter ones, Throw away the unfit ones.

1 (single point) crossover. Identify the crossover site.

crossover site

01101 Completely random crossover
11000 is done.
11000 Identify a crossover site at
10011 random

01100
11001
11011
10000

1-1 crossover site possible.
Portion after crossover site gets swapped.

Jumped from 24 to 27.

Index.	x	Chromosome.	f(x).	$\frac{f(x)}{\Sigma f}$	$\frac{f(x)}{\bar{f}}$	Round.
1	12	01100	144.			
2	25	11001	625			
3	(27)	11011	729.			
4	16	10000	256.			
			1754.			

$\bar{f} = 438.5$

You will get the fitter solutions no matter where you cross and whomever you cross.

31/1/2019

Bigdata is not only about the volume.

Collection of databases or datawarehouses. How is the data organised? → Bigdata.

Structured data contains a schema.
Unstructured., Semi structured, Quasi-Structured.
location. weblinks.
Bigdata deals with all the above types of data.

Because of IOT, the amount data generated is increasing.
We have petabytes of data.

The V's.

Volume → how huge the data is
Velocity → streaming data. (weather, tweets).
Variety → structured, unstructured, etc.
Business insights:
- Veracity → business.
- Value → historic data's importance.
- Variability → error in data, how can it be avoided.

* Challenges of Bigdata.

Capturing
Curation
Storage.
:

Hadoop is a framework.
Scale up and scale down(out)
SPOF → single point of failure.
how many machines can be added.

Architectures:- MapReduce, HDFS.

4/2/2019

Recap

- What is the main aim of Roulette wheel?
- Disadvantages of Roulette wheel:-
  - → Monopoly; fitter solutions occupy major part of the wheel.
  - → We may sometimes loose the optimum solution.

If of the population should be increasing to get the single solution.
GA leads to the path of convergence on every iteration.

Based on increase in avg, we will eventually come up with better results. - Mathematical function application.

Schemata is a pattern of chromosomes.

1 0 0 0
1 0 0 1
1 1 0 0.

The pattern in the above data is MSBs are 1.
Pattern is 1***
where * can be either 0 or 1.

Need for schemata - we can reach the optimum solution faster.

Choosing the pattern after two or three results we need to choose the best pattern.

Fundamental Theorem of Schemata :- → single pattern.

Choosing the better pattern may result to the optimum solution - Mathematical proof.

Fitness of schema is average of population.

We are concerned about the increase in fitness from previous generation. (Betterment).

Schema → H

Length of the schema $\delta(H)$:- Difference b/n 1st fixed and the last fixed bits of schema.

Example:-

$H_1 = *1****0$ (positions 2 and 7)

$\delta(H_1) = 5$

$H_2 = ***10**$

$\delta(H_2) = 1$

Selection :- Selecting the pattern based on schema pattern.

Crossover is a better solution than Mutation.

GA is more probabilistic because of randomness.

Schema are fundamental theorem of GA.

→ Short defined length $\delta(H)$.

→ Low $O(H)$.

→ Above Average.

Selection Schema Theorem :-

$$m(H, t+1) \geq m(H, T) \frac{f(H)}{f} \left\{ 1 - \frac{P_c \delta(H)}{L-1} - P_m O(H) \right\}$$

(survival)

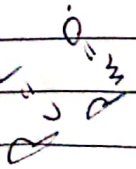
- $m(H, t+1)$ → count function.
- $\frac{f(H)}{f}$ → fitness of schema i.e. Avg.
- $\frac{\delta(H)}{L-1}$ → prob. of crossover. (Destruction probability of crossover)
- $P_m O(H)$ → Destruction of mutation

Schema :-

$H_1 = 1****$

$H_2 = *10***$

$H_3 = 1***0$



Initial population :-

13 : 0 1 1 0 1 → String 1
24 : 1 1 0 0 0 → String 2
8 : 0 1 0 0 0 → String 3
19 : 1 0 0 1 1 → String 4

H_1 schema is the string representation of 2, 4.
H_2 schema is the string representation of 2, 3.
H_3 schema is the string representation of 2.

Why are we keen on string representations

$$M(H_1, t+1) = \frac{2 \times 468.5}{293}$$

$$= 3.19 \{1 - 0\}.$$
$$= 3.19.$$

for H_2

$$M(H_2, t+1) \rightarrow \frac{2 \times 320}{293} = 2.18.$$

∴ There are 2 copies of H_2 in pop.

$$2.18 * \left\{1 - \frac{1}{4}\right\}.$$

$$= 1.63.$$

Benefit of schemabased GA:-
→ It is the parallel processing capability of GA.
Schema based processing is parallel.
∴ processing is fast.

Main objective of schema theorem:-
In selection we are concerned with betterment

fixed positions survival is exponentially increasing.
are near:

fixed positions exponential decay or distinctness
are apart:

11/2/2019

GA classifier : entire concept is ~~bored~~ bucket brigade

efficient path which leads to the o/p faster is
done by GA

Rule + Msg system - in the domain of expert system
Bidding results in
① Penalty.
② Reward.

Credit Apportionment System
good classifier = +ve credit.
bad classifier = -ve credit.

What is encoding in GA context?
Soln:- It is used to extract patterns like MSBs are
1 etc.
Using Decimal numbers extracting patterns becomes
difficult.

In. travelling salesman problem.
The feasible solution is cost of visiting each node.

Each solution is called chromosome in GA.

1 2 3 4 5 6 } swap.
6 3 4 5 1 2
Not feasible.

fitness of classifier means?
Strength of classifier which gives better fitness.
Five alarm example

In GA O(n3) schema is effectively useful where n is the population

GA is inherently parallel.

Among the different paths, the poor paths are not neglected straight away but they are exponentially decreasing in GA.

When m classlabels are there, it is less time consuming than binary class label because getting patterns becomes difficult in binary tree.

18/2/2019

2 aspects of Bucket Brigade :-
1. Rule - Msg system.
2. Apportionment of credit
3. Genetic Algo.

Rule :- <classifier> = <condn> : : <message>
attributes — class labels.
{0, 1, #}* — {0, 1}*

Fitness value = strength.
Increasing or decreasing the strength of classifier is Apportionment of credit.

Table to simulate credit sharing.

Sn	Classifier	t=0				t=1				t=2				t=3			
		S	Msg	Mtch	Bid	S	Msg	Mtch	Bid	S	Msg	Mtch	Bid	S	Msg	M	Bid
1	01## : 0000	200		E	20	180	0000			220				220			
2	00#0 : 1100	200				200		1	20	180	1100			218			
3	11## : 1000	200				200				200		2	20	180	1000		
4	##00 : 0001	200				200		1	20	180	0001	2	18	162	0001	3	16.2

Env = 20.
Bid = 10%.
E = 0111 → environment variable
First strength value is random.

Action or Payoff scheme → submit bit.

One classifier reduces it's strength so that the others can get triggered in the subsequent generations.

The Bid is added onto the other classifiers. The same path should not keep getting the priority in each generation.

GA kicks in after one run.

New classifier paths are not explored by bucket brigade.

Bucket Brigade is not a contribution of GA.

$t = 4$.

S	Msg	Match	Bid.
220			
218			
196			
146	0001		

terminal.

Crossover b/w ① & ④.

```
01 | # #
# # | 0 0
```

$= 01\ 0\ 0.$

$\# \# \# \#.$

21/2/2019

TUTORIAL ①

① A decision tree is being error pruned by collapsing of nodes. For the particular node, on the left branch there are 3 training points [5, 7, 9.6] and for the right branch there are 4 training points [8.7, 9.8, 10.5, 11]

(i) What were the original responses along left, right branches?

(ii) what is the new response after collapsing?

(2) Say True / False - Reason. also.

(a) U. can list all fi's & their support if u know the MFI — ~~T~~ (F)

(b) to perform hierarchical clustering., we do not need to know the coordinates of elements, only their mutual distances are enough (F)

(c) If one used cosine measure to measure interesti-ngness of A → B. its value doesn't change if we add a new transaction. that doesn't contain A or B (T)

$$\frac{P(X \cdot Y)}{\sqrt{P(X) \cdot P(Y)}} \qquad \frac{|X \cdot Y|}{\sqrt{|X| \cdot |Y|}}$$

(3) Given T[A .... J].

A B C D E F G H I J

1 → 1110000111 2 → 1010110111

3 → 1001001111 4 → 0110101100.

5 → 0111101000 6 → 1100111010.

7 → 1100000010. 8 → 1110110010.

9 → 1100000011 10 → 0011110000.

(i) Find all fi's with support ≥ 5. {A B C E I}.

(ii) which in (i) are closed?

(4) Given T(A,B,C,D) = {1110, 1110, 0111, 0110, 0111, 1100, 1100 (10 transactions). 1110, 1110, 0011}.

(a) What is the confidence of. (A,B) → ∅

(b) Given itemset X = {A,B,C}, List all rules of form. Y → Z ; Y ≠ ∅ ; Y ⊆ X, Z = X \ Y. with. support ≥ 4 and confidence. = 2/3.

(5)

Classes.	Buys_comp = Yes.	Buys_comp = NO.
Buys_comp = Yes.	6954	46.
Buys_comp = NO	412	2588.

Given above, calculate. Precision, Recall, Sensitivity, Specificity, Accuracy.

(6) Trace A-Close, pincer search for the data labelled X.

X :- T_1 {M, O, N, K, E, Y}. Support = 2
T_2 {D, O, N, K, E, Y}.
T_3 {M, A, K, E}
T_4 {M, U, C, Y}.
T_5 {C, O, O, K, I, E}

4/3/2019

Neural network is a composition of layers.

Input variables are nodes in network topology.

Back propagation of errors.
Neural network is supervised learning. It has training and testing.
What is error in neural network?
How much prediction is deviated from expectation.

Inputs are fed forward. If the error is greater than error threshold, they are fed back.
"Adjust weights".

Neural network is non linear.
One neuron = node in network.
firing of neuron

If we are not happy with the o/p, the add hidden layers.
Activation function: the o/p from hidden layer

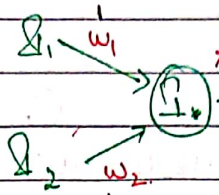
OR Gate.

$< = 0.5 \Rightarrow 0$

$> = 0.5 \Rightarrow 1$

I_1	I_2	Output	x	O	OR output
0	0	0	0	$0.5 \frac{1}{1+e^{-x}}$	→ 0
0	1	1	1	0.7	→ 1
1	0	1	1	0.7	→ 1
1	1	1	2	0.88	→ 1

OR is non linear combination of inputs due to activation function.



Depending on the weights, the inputs are taken.

$x = I_1 w_1 + I_2 w_2$

if $w_1 = w_2 = 1$.

then $x = I_1 + I_2$.

ACTIVATION FUNCTION.

$$\text{Sigmoid}\phi = \frac{1}{1+e^{-x}}$$

x-T → error threshold.

The quality of NN depends also on the activation function.

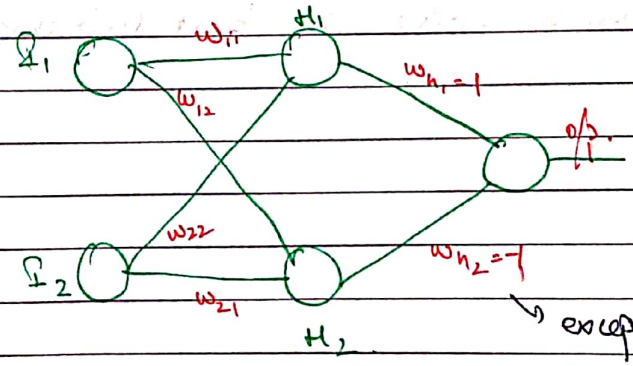
In this network activation function is applied in the output layer.

Where does exclusiveness come in XOR?

XOR needs atleast one hidden layer.

11/3/2019

for XOR ① Hidden Layer ② O/p neuron.



except for this rest of the weights are 1

overall input to the hidden layer H_1

$$Sig\ H_1(x) = \frac{1}{1 + e^{-(x-0.5)}}$$

$$Sig\ H_2(x) = \frac{1}{1 + e^{-(x-1.5)}}$$

$$Sig\ O/p\ (x) = \frac{1}{1 + e^{-(x-0.2)}}$$

These numbers come by trial and error.

$H_1 - H_2$ → put these in → 0.1951, 0.2450, 0.19

	I_1	I_2	X	H_1	H_2	O	out.
4	0	0	0	.3775	.1824	.4998	0
	0	1	1	.6225	.3775	.5712	1
	1	0	1	.6225	.3775	.5712	1
	1	1	2	.8176	.6225	.4998	0

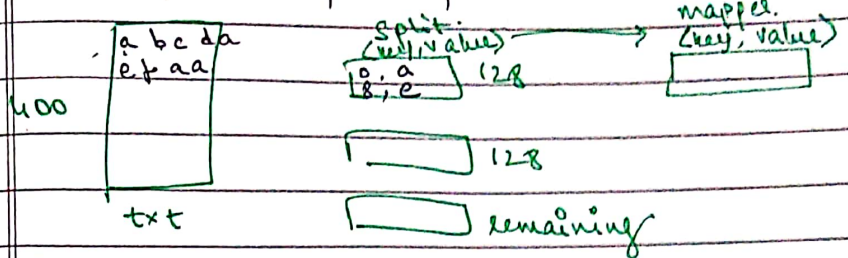
→ <0.5 then 0, >0.5 then 1

You can't have a neural network learning correctly without a hidden layer.

Neural network → supervised.

- Disadvantage:-
  For live inputs there are too many wt adjustments. That means long training times. Interpretability is less than the previous algos.
- ~~Advantages~~
  Non linearity of model.
  Face recognition, etc where human like learning is required. in terms of precision, recall, etc.
  Domains with noise are best suitable for NN.

Aim of Hadoop is parallelism



Before going to mapper, it will go to split. Done in terms of lines.

<key, value>
↓ ↓
pair string.

Each mapper works on a single block.

mapper
<a, 1>
<b, 1>
<c, 1>
<d, 1>
<a, 1>