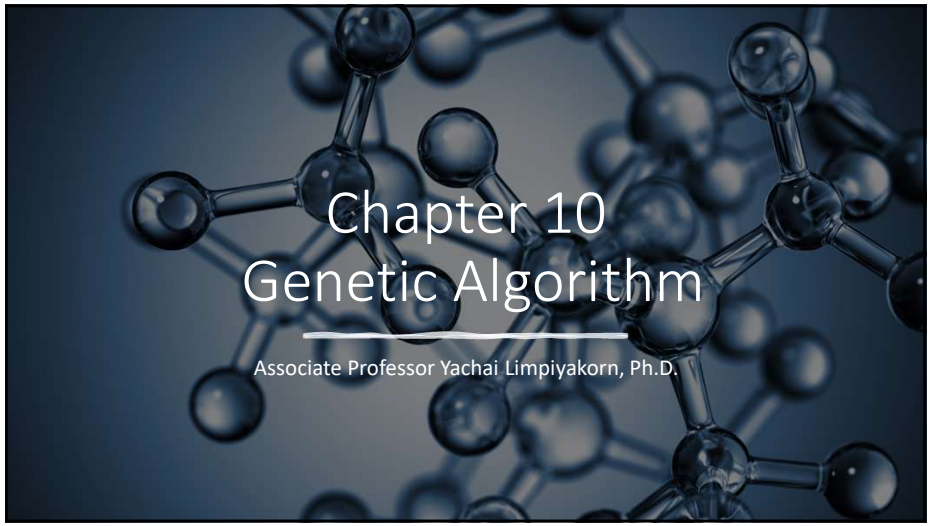


Chapter 10
Genetic Algorithm

Associate Professor Yachai Limpiyakorn, Ph.D.

Principles of Genetics

ในทางชีววิทยา เซลแต่ละเซลในพืชชั้นสูงและสัตว์ประกอบด้วยนิวเคลียส (nucleus) 1 เซล แต่ละนิวเคลียสประกอบด้วยโครโมโซม (chromosome) จำนวนหนึ่ง โดยโครโมโซมจะอยู่กันเป็นคู่มาจากพ่อและแม่อย่างละเส้น โครโมโซมแต่ละเส้นจะมียีน (gene) เป็นตัวกำหนดลักษณะถ่ายทอดทางพันธุกรรมของสิ่งมีชีวิต ในขณะที่มีการจับคู่กันของโครโมโซมอาจเกิด *การไขว้เปลี่ยน (crossover)* ซึ่งเป็นการที่ยีนจากโครโมโซมพ่อแม่สลับเปลี่ยนกัน ทำให้เกิดโครโมโซมใหม่ขึ้น 2 คู่ และในขณะที่เซลแบ่งตัวจะเกิดกระบวนการ *คัดลอกโครโมโซม (chromosome copying)* ซึ่งบางครั้งจะมีการเปลี่ยนแปลงของยีนที่มาจากยีนพ่อแม่ เกิดเป็นยีนที่ไม่เคยมีมาก่อน เราเรียกการเกิดยีนลักษณะนี้ว่า *การกลายพันธุ์ (mutation)*

เราอาจกล่าวได้ว่า การคัดเลือกโดยธรรมชาติ (Natural Selection) เกิดจากการเปลี่ยนแปลงทางพันธุกรรมอันเป็นผลมาจาก *การไขว้เปลี่ยน (crossover)* ของลักษณะทางพันธุกรรม และ *การกลายพันธุ์ (mutation)*

Natural Selection

- Charles Darwin ได้อธิบายการสืบทอดของสิ่งมีชีวิตด้วยกฎ Evolution through Natural Selection :-
  - สิ่งมีชีวิตมีแนวโน้มที่จะสืบทอดลักษณะพิเศษให้ลูกหลาน
  - ธรรมชาติจะผลิตสิ่งมีชีวิตที่มีลักษณะพิเศษแตกต่างไปจากเดิม
  - สิ่งมีชีวิตที่เหมาะสมที่สุด (fittest) หรือที่มีลักษณะพิเศษที่ธรรมชาติพอใจมากที่สุด มีแนวโน้มที่จะมีลูกหลานมากกว่าตัวที่ไม่เหมาะสม ดังนั้น ประชากรจะโน้มเอียงไปทางตัวที่เหมาะสม
  - การเปลี่ยนแปลงจะสะสมไปเรื่อยและเกิด species ใหม่ที่เหมาะกับสภาพแวดล้อม เมื่อเวลาผ่านไปนานๆ

Genetic Algorithm (GA)

- Goldberg and Holland, 1988
- GA has its core idea from Darwin's theory of natural evolution "survival of the fittest"
- one of random-based *evolutionary algorithms* (EAs)
- search-based optimization technique GA เน้นการ optimize
- optimization → "how to find the best value for k that maximizes the performance of kNN classifier?"
- in order to find a solution, random changes applied to the current solutions to generate new ones.

How GA works

- GA works on a population consisting of some solutions
- population size is the number of solutions
- each solution is called individual/ hypothesis
- each individual solution has a chromosome
- chromosome is represented as a set of parameters (features) that defines the individual
- each chromosome has a set of genes
- each gene is represented by somehow such as a string of 0s and 1s

Hypothesis/ Individual Representation

Hypothesis: IF (Haircolor =black ∨ red) ∧ (Eyecolor = dark) THEN Sunburn = negative

	Haircolor	Eyecolor	Sunburn
Bit String:	110	10	01

2 ค่าที่เป็นไปได้ = 2 บิต

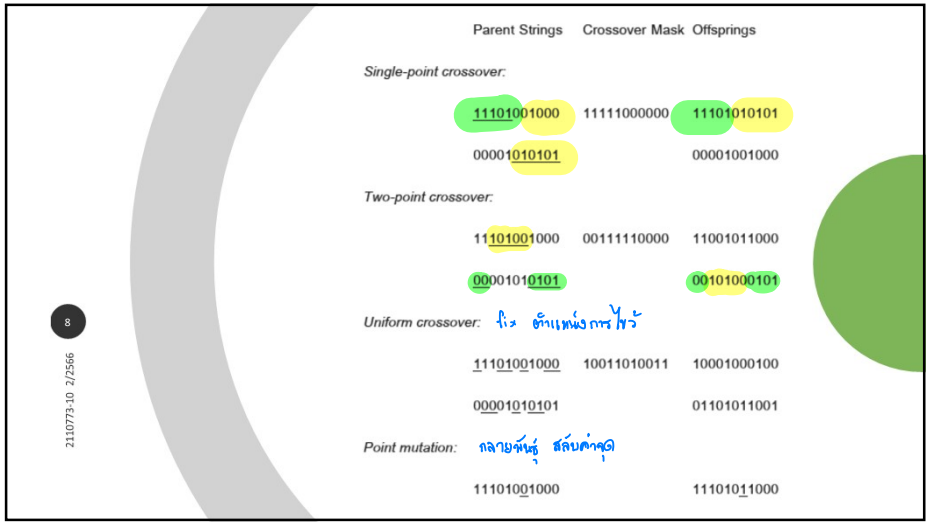
Hypothesis: IF Haircolor =blonde THEN Sunburn = positive

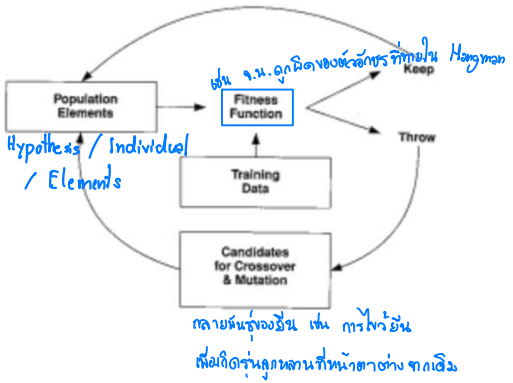
	Haircolor	Eyecolor	Sunburn
Bit String:	001	11	10

Note: Eyecolor = 11 → Don't care

Genetic Operators

- *การไขว้เปลี่ยน (Crossover)* เป็นการสร้างสายอักขระลูกหลาน 2 สาย จากสายอักขระพ่อแม่ 2 สาย โดยการคัดลอกบิตจากสายอักขระพ่อแม่ตามตำแหน่งที่กำหนดโดยหน้ากากไขว้เปลี่ยน (Crossover Mask) การไขว้เปลี่ยนทำได้หลายวิธีดังตัวอย่างที่แสดงข้างล่าง อาทิ การไขว้เปลี่ยน 1 ตำแหน่ง (Single-point Crossover) การไขว้เปลี่ยน 2 ตำแหน่ง (Two-point crossover) การไขว้เปลี่ยนยูนิฟอร์ม (Uniform Crossover) เพื่อให้เกิดความหลากหลาย (Diversity)
- *การกลายพันธุ์ (Mutation)* เป็นการสร้างการเปลี่ยนแปลงต่อสายอักขระลูกหลาน โดยสุ่มเลือกเปลี่ยนค่าบิตใดบิตหนึ่งดังแสดงในตัวอย่างข้างล่าง





GA Overview

การเรียนรู้จำลองวิวัฒนาการของสิ่งมีชีวิต (Biological Evolution)

GA Parameters

- Fitness function for ranking candidate patterns
- Stopping Criteria:
  - Maximum of hypotheses' fitness values >= fitness threshold
  - Max fitness does not change after many generations
  - Reach fixed number of iterations of learning process
- Size of population *p* to be maintained
- Ratio of population *r* to be replaced at each generation
- Mutation rate *m*

GA (Fitness, Fitness_threshold, p, r, m)

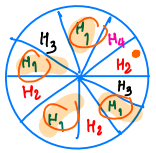
- Initialize: P ← p random hypotheses
- Evaluate: for each h in P compute Fitness(h)
- While [max$_h$ Fitness(h)] < Fitness_threshold
  1. *Select:* Probabilistically select *(1-r)p* individuals of *P* to add to P_s.

     $$\text{Prob}(h_i) = \text{Fitness}(h_i) / \sum_j \text{Fitness}(h_j) \; ; \; j=1,\ldots,p$$

  2. *Crossover:* Probabilistically select *rp/2* pairs of hypotheses from *P*.
     For each pair $< h_1, h_2 >$, produce two offspring by applying Crossover operator. Add all offspring to P_s.
  3. *Mutate:* Invert a randomly selected bit in *mp* random individuals of P_s.
  4. *Update:* $P \leftarrow P_s$.
  5. *Evaluate:* for each *h* in *P* compute *Fitness(h)*
- Return the hypothesis from P that has the highest fitness.

probabilistic selection

roulette wheel selection



H_i	$F(H_i)$	$P(H_i)$
H_1	4	4/10
H_2	3	3/10
H_3	2	2/10
H_4	1	1/10
H_5	0	0
	10	

12

Fitness Function and Selection

โดยทั่วไป โครโมโซมจะถูกเลือกแบบสุ่มเพื่อสร้างประชากรรุ่นใหม่ด้วยค่าความน่าจะเป็นตามสมการที่แสดงใน GA Algorithm ข้างต้น ซึ่งวิธีการเลือกดังกล่าว เรียกว่า Fitness Proportionate selection หรือ Roulette wheel selection เราจะสังเกตได้ว่าฟังก์ชันค่าความเหมาะ (fitness function) จะเป็นตัวกำหนดความน่าจะเป็นที่โครโมโซมจะอยู่รอดในประชากรรุ่น (generation) ถัดไป ซึ่งโครโมโซมที่มีค่าความเหมาะสูงจะมีโอกาสอยู่รอดมากกว่าโครโมโซมเส้นอื่น ๆ แต่ก็ไม่ได้หมายความว่า โครโมโซมที่มีค่าความเหมาะสูงสุดจะถูกเลือกทุกครั้ง ขึ้นอยู่กับการสุ่มค่า

Crowding เป็นปรากฏการณ์ที่สมาชิกบางตัวในประชากร ซึ่งมีค่าความเหมาะสูง ถูก reproduce อย่างรวดเร็ว ทำให้ความหลากหลาย (diversity) ของประชากรลดลง การแก้ปัญหา *Crowding* สามารถกระทำได้โดย

1. เปลี่ยนวิธีการเลือกความเหมาะ (Altering fitness selection) โดยใช้ Tournament Selection หรือ Rank Selection
2. ใช้วิธีการ Fitness sharing หมายถึง ลดค่าความเหมาะของสมาชิกใน ประชากรรุ่นหนึ่งที่มีความคล้ายกัน (the fitness of a member is reduced by the presence of similar members)

Tournament selection – เลือกสมาชิก 1 คู่จากประชากรรุ่นปัจจุบันแบบสุ่ม แล้วจึงสุ่มเลือกระหว่างสมาชิกคู่นั้น โดยสมาชิกตัวที่มีค่าความเหมาะสูงกว่าจะถูกสุ่มเลือกด้วยค่าความน่าจะเป็นที่กำหนดไว้ล่วงหน้า p ส่วนสมาชิกตัวที่มีค่าความเหมาะต่ำกว่าจะถูกสุ่มเลือกด้วยความน่าจะเป็น (1-p)

Rank selection – เป็นวิธีที่ใช้ควบคุมการเลือกโครโมโซมโดยไม่สนใจค่าความเหมาะของโครโมโซมว่ามีค่าเท่าไร แต่จะใช้ค่าความเหมาะเพียงแค่จัดลำดับเรียงโครโมโซมตามค่าความเหมาะที่มีค่าสูงสุดจนถึงต่ำสุด จากนั้น กำหนดค่าคงที่ p เป็นความน่าจะเป็นที่โครโมโซมลำดับที่ 1 จะถูกเลือก และเป็นความน่าจะเป็นที่โครโมโซมลำดับที่ 2 จะถูกเลือกเมื่อโครโมโซมลำดับที่ 1 ไม่ถูกเลือก และเป็นความน่าจะเป็นที่โครโมโซมลำดับที่ 3 จะถูกเลือกเมื่อลำดับที่ 1 และ 2 ไม่ถูกเลือก เรื่อยไปจนกระทั่งถึงลำดับสุดท้ายซึ่งจะถูกเลือกเมื่อลำดับก่อนหน้าไม่ถูกเลือกเลย ดังนั้น ความน่าจะเป็นที่โครโมโซมลำดับที่ r จะถูกเลือก เท่ากับ $p(1-p)^{r-1}$; r = 1,2,3,...

Rank Selection

Hypothesis	Fitness(h_i)	P(h_i)	Rank	Probability of Rank(h_i)
h_1	4	0.4	1	=p = 0.667
h_2	3	0.3	2	=p(1-p) = 0.667x0.333 = 0.222
h_3	2	0.2	3	$=p(1-p)^2 = 0.667(0.333)^2 = 0.074$
h_4	1	0.1	4	$=p(1-p)^3 = 0.667(0.333)^3 = 0.025$
h_5	0	0	5	$=p(1-p)^4 = 0.667(0.333)^4 = 0.012$

Example GA Application

ตัวอย่างชุดข้อมูลสอนการจำแนกประเภทลักษณะลูกค้าที่สนใจ Life Insurance Promotion

Training Instance	Income Range	Life Insurance Promotion	Credit Card Insurance	Sex	Age
1	30–40K	Yes	Yes	Male	30–39
2	30–40K	Yes	No	Female	40–49
3	50–60K	Yes	No	Female	30–39
4	20–30K	No	No	Female	50–59
5	20–30K	No	No	Male	20–29
6	30–40K	No	No	Male	40–49

Fitness function

$$F(E_i) = N / (M+1)$$

where

N คือ จำนวนตัวอย่างในชุดข้อมูลสอนซึ่งอยู่ในคลาสเดียวกับ E_i ที่มีค่าคุณลักษณะตรงกัน

M คือ จำนวนตัวอย่างในชุดข้อมูลสอนซึ่งไม่อยู่ในคลาสเดียวกับ E_i ที่มีค่าคุณลักษณะตรงกัน

หมายเหตุ เพื่อป้องกันการหารด้วยค่าศูนย์ จึงบวกหนึ่งที่ตัวหารในฟังก์ชันความเหมาะ

กฎการคัดเลือกสมาชิกของประชากรรุ่นถัดไป

- สมาชิกที่มีค่าความเหมาะน้อยกว่า ค่าขีดแบ่ง (threshold) ที่กำหนดไว้มีค่าเท่ากับ 1 จะถูกนำไป crossover หรือ mutation
- ถ้าสมาชิกทุกตัวมีค่าความเหมาะไม่น้อยกว่าค่าขีดแบ่ง จะสุ่มเลือกสมาชิกเพื่อไป crossover หรือ mutation
- ในประชากรทุกรุ่น จะต้องมีสมาชิกที่อยู่ในคลาส “Yes” และ “No” อย่างละ 2 ตัว

กำหนดขนาดประชากรแต่ละรุ่นเท่ากับ 4 ตัว

Income Range = ? → don’t care

สมาชิกประชากรรุ่นแรก

- **An Initial Population for Supervised Genetic Learning**

Class

Population Element	Income Range	Life Insurance Promotion	Credit Card Insurance	Sex	Age
1	20–30K	No	Yes	Male	30–39
2	30–40K	Yes	No	Female	50–59
3	?	No	No	Male	40–49
4	30–40K	Yes	Yes	Male	40–49

ตัวอย่างการคำนวณ $F(E_1)$

- Income Range = 20-30K ตรงกับตัวอย่างสอนที่ 4 และ 5
- Credit Card Insurance = Yes ไม่ตรงกับตัวอย่างสอนใดๆ
- Sex = Male ตรงกับตัวอย่างสอนที่ 5 และ 6
- Age = 30-39 ไม่ตรงกับตัวอย่างสอนใดๆ
- จะได้ว่า ค่า N = 2+0+2+0 = 4
- Income Range = 20-30K ไม่ตรงกับตัวอย่างสอนใดๆ
- Credit Card Insurance = Yes ตรงกับตัวอย่างสอนที่ 1
- Sex = Male ตรงกับตัวอย่างสอนที่ 1
- Age = 30-39 ตรงกับตัวอย่างสอนที่ 1 และ 3
- จะได้ว่า ค่า M = 0+1+1+2 = 4

การไขว้เปลี่ยนเพื่อสร้างสมาชิกลักษณะใหม่

- F(E1) = 4/5 =0.80, F(E2) = 6/7 =0.86; F(E1) and F(E2) < 1 → crossover
- F(E3) = 6/5 =1.20, F(E4) = 5/5 =1.00; F(E3) and F(E4) >= 1 → keep

crossover ทั้งแพ่งหลัง class label

Population Element	Income Range	Life Insurance Promotion	Credit Card Insurance	Sex	Age
#1	20-30K	No	Yes	Male	30-39
#2	30-40K	Yes	No	Fem	50-59

Population Element	Income Range	Life Insurance Promotion	Credit Card Insurance	Sex	Age
#2	30-40K	Yes	Yes	Male	30-39
#1	20-30K	No	No	Fem	50-59