กำหนดส่ง

HW	2018-04-02

กระบวนวิชา 204113

การบ้านปฏิบัติการ11

Artificial Intelligence (ID3)(20 คะแนน)

1) **20คะแนน**(HW11_5XXXXXXXX.py)ให้เขียนโปรแกรมเพื่ออ่านข้อมูลตัวอย่าง (training data) จากไฟล์ input.txt จากนั้นนำข้อมูลที่ได้มาทำการสร้าง Decision tree โดยใช้วิธีคำนวณตามหลักการ ID3 และแสดงผลลัพธ์ดังตัวอย่างที่กำหนดให้

ตัวอย่างข้อมูล training data

Film	Country of origin	Big Star	Genre	Success
1	USA	yes	Science	true
2	USA	no	Comedy	false
3	USA	yes	Comedy	true
4	Europe	no	Comedy	true
5	Europe	yes	Science	false
6	Europe	yes	Romance	false
7	Rest of world	yes	Comedy	false
8	Rest of world	no	Science	false
9	Europe	yes	Comedy	true
10	USA	yes	Comedy	true

training dataอยู่ในไฟล์โดยมีรูปแบบดังนี้	**ผลลัพธ์**ให้แสดงดังนี้
1;USA;yes;Science;true 2;USA;no;Comedy;false 3;USA;yes;Comedy;true	First factor is Country of origin The next factor of USA is Big star The next factor of Europe is Genre

1. ให้เขียนฟังก์ชันinformation_gain(feature)เพื่อคืนค่าinformation gainของ feature แต่ละตัวของข้อมูลตามหลักการของID3โดยสามารถมีการเรียกฟังก์ชันย่อยที่เหมาะสมได้
2. ให้เขียนฟังก์ชัน factor_list()เพื่อแสดงผลลำดับของปัจจัยของ Decision Tree ที่ได้ ตามตัวอย่างด้านบน

หลักการคำนวณโดยใช้ ID3 มีรายละเอียดดังนี้

Entropy = E(S) = - $p_1 \log_2 p_1$ - $p_0 \log_2 p_0$

จากข้อมูล training data จะหา**ปัจจัยตัวแรก**ได้ดังนี้

กรณีปัจจัย Country of originจะหา Entropy ของค่าตัวแปร 3 ค่าได้ดังนี้

E(USA) = -(3/4) $\log_2$ (3/4) – (1/4) $\log_2$ (1/4)

= 0.311+ 0.5 = 0.811

E(Europe) = -(2/4) $\log_2$ (2/4) – (2/4) $\log_2$ (2/4) = 1

E(Rest of world) = - 0 – (2/2) $\log_2$ (2/2) = 0

จากนั้นนำมาคำนวณหาค่า Information gain ของ Country of origin

Country of origin = 1-P(USA)x E(USA)- P(Europe)x E(Europe)- P(Rest of world)x E(Rest of world)

= 1-(0.4x0.811) – (0.4x1) – (0.2x0)

= 0.2756

กรณีปัจจัย Big Starจะหา Entropy ของค่าตัวแปร 2 ค่าได้ดังนี้

E(yes) = -(4/7) $\log_2$ (4/7) – (3/7) $\log_2$ (3/7)

= 0.9852

E(no) = -(1/3) $\log_2$ (1/3) – (2/3) $\log_2$ (2/3) = 0.9151

จากนั้นนำมาคำนวณหาค่า Information gain ของ Big Star

Big Star = 1-P(yes)x E(yes) – P(no) x E(no)

= 1- (0.7x 0.9852) – (0.3 x 0.9151)

= 0.044

กรณีปัจจัย Genreจะหา Entropy ของค่าตัวแปร 3 ค่าได้ดังนี้

E(science) = -(1/3) $\log_2$ (1/3) – (2/3) $\log_2$ (2/3) = 0.9151

E(comedy) = -(4/6) $\log_2$ (4/6) – (2/6) $\log_2$ (2/6) = 0.9151

E(romance) = 0 – 1 $\log_2$ (1) = 0

จากนั้นนำมาคำนวณหาค่า Information gain ของ Genre

Genre = 1-(0.6x0.9151)-(0.3x0.9151)-(0.1x0)

= 0.176

ดังนั้น ปัจจัยตัวแรกของ Decision Tree คือ Country of origin และได้ tree ดังรูป

ขั้นถัดมา หาปัจจัยตัวที่สองที่ต่อจาก USA พิจารณาเฉพาะข้อมูลที่มี Country เป็น USA

Film	Country of origin	Big Star	Genre	Success
1	USA	yes	Science	true
2	USA	no	Comedy	false
3	USA	yes	Comedy	true
4	USA	yes	Comedy	true

กรณีปัจจัย Big Starจะหา Entropy ของค่าตัวแปร 2 ค่าได้ดังนี้

$E(yes) = -1 \log_2 (1) = 0$

$E(no) = 0 - 1 \log_2 (1) = 0$

จากนั้นนำมาคำนวณหาค่า Information gain ของ Big Star

Big Star = 0.811 - P(yes)x E(yes) – P(no) x E(no)

= 0.811 - (0.75x 0) – (0.25 x 0) = 0.811

คิดภายใต้เงื่อนไขของ USA เลยต้องใช้ E(USA)

กรณีปัจจัย Genreจะหา Entropy ของค่าตัวแปร 3 ค่าได้ดังนี้

$E(science) = -1 \log_2 (1) = 0$

$E(comedy) = -(2/3) \log_2 (2/3) - (1/3) \log_2 (1/3) = 0.9151$

E(romance) = 0เพราะไม่มีค่าในส่วนนี้

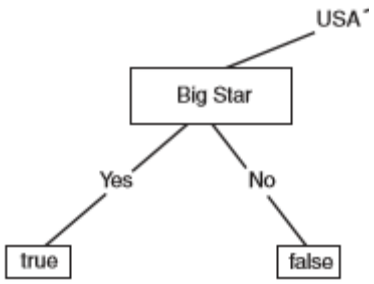
จากนั้นนำมาคำนวณหาค่า Information gain ของ Genre

Genre = **0.811** -(0.25x0)-(0.75x0.9151)-0

= **0.125**

คิดภายใต้เงื่อนไขของ USA เลยต้องใช้ E(USA)

ดังนั้น ปัจจัยตัวถัดมาของ Decision Treeที่ต่อจากUSA คือ Big Starและได้ tree ดังรูป



ขั้นถัดมา หาปัจจัยตัวที่สองที่ต่อจาก Europe พิจารณาเฉพาะข้อมูลที่มี Country เป็น Europe

Film	Country of origin	Big Star	Genre	Success
1	Europe	no	Comedy	true
2	Europe	yes	Science	false
3	Europe	yes	Romance	false
4	Europe	yes	Comedy	true

กรณีปัจจัย Big Starจะหา Entropy ของค่าตัวแปร 2 ค่าได้ดังนี้

$$E(yes) = -(1/3)\log_2(1/3) - (2/3)\log_2(2/3) = 0.9151$$

$E(no) = -1\log_2(1) = 0$

จากนั้นนำมาคำนวณหาค่า Information gain ของ Big Star

Big Star = **1** - P(yes)x E(yes) – P(no) x E(no)

= **1** - (0.75x 0.9151) – (0.25 x 0) = **0.314**

คิดภายใต้เงื่อนไขของ Europe เลยต้องใช้ E(Europe)

กรณีปัจจัย Genreจะหา Entropy ของค่าตัวแปร 3 ค่าได้ดังนี้

E(science) = 0

E(comedy) = 0

E(romance) = 0

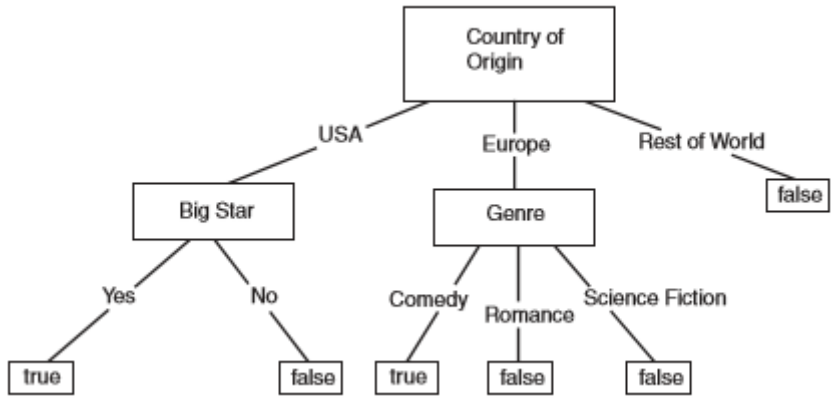
จากนั้นนำมาคำนวณหาค่า Information gain ของ Genre

Genre = **1** - 0

= **1**

คิดภายใต้เงื่อนไขของ Europe เลยต้องใช้ E(Europe)

ดังนั้น ปัจจัยตัวถัดมาของ Decision Tree ที่ต่อจาก Europe คือ Genre และได้ tree ดังรูป



<u>การส่งงาน</u>

1. ลักษณะ/ลำดับข้อความของการรับค่า/แสดงผลจะ**<u>ต้องเป็นไปตามที่ระบุ</u>**ในตัวอย่างการ run
2. ไฟล์งาน/ใบงานที่ส่ง จะต้องมีการแทรก comment/หัวกระดาษ ตามข้อกำหนดใน website รายวิชา
3. ไฟล์งานโปรแกรมที่ส่ง จะต้องมีการ<u>แทรก pseudocode</u> เป็น comment ในแต่ละขั้นตอน
4. Upload ไฟล์ source code ตามที่ระบุในแต่ละข้อ ไปยัง website ที่ใช้ส่งการบ้าน http://hw.cs.science.cmu.ac.th ตาม section ที่นักศึกษาเรียน