

ใบงาน Machine Learning

วัตถุประสงค์ หัดแทนค่าเพื่อสร้าง C4.5 decision tree

$$E_{C_k}(S) = - \sum_{j=1}^v p_{j \text{ and } c=k} \log_2(p_{j \text{ and } c=k})$$

$$E_C(S) = \sum_{k=1}^u p_k E_{C_k}(S)$$

โมเดล C4.5 decision tree เพื่อสร้างโมเดลพยากรณ์ เขียนได้ว่า
กล่าวคือ Entropy, E, ของของข้อมูล S เมื่อใช้ Candidate, C_k , มีค่าเป็น
ผลรวมของ $-p_{j \text{ and } c=k} \log_2(p_{j \text{ and } c=k})$ โดย j คือ label ที่สนใจ (เช่น เล่น
หรือไม่เล่น) และ k คือ ค่าของ candidate นั้นๆ และ $p_{j \text{ and } c=k}$ คือความ
น่าจะเป็นที่เหตุการณ์ค่า $c=k$ มีค่าเป็น j
เมื่อนำ E_{C_k} มาผลรวมก็เพียงถ่วงน้ำหนักด้วยสัดส่วนของแต่ละ $c=k$
ก็จะได้ Entropy เมื่อใช้ C เป็น candidate

...จากข้อมูลต่อไปนี้

คำสั่ง

- คำนวณ $E(S)$, $E_{C=\text{สภาพอากาศ}}(S)$, $E_{C=\text{อุณหภูมิ}}(S)$, $E_{C=\text{สภาพลม}}(S)$ ในตาราง
- candidate ไหน ให้ $E_C(S)$ ต่ำที่สุด
- สารสนเทศที่ได้จากข้อ 2 สร้างกฎอะไรได้บ้าง

Weather	Temp	Wind	label
s	h	F	n
s	h	T	n
o	h	F	y
r	m	F	y
r	c	F	y
r	c	T	n
o	c	T	y
s	m	F	n

x	$\log_2(x)$
0	0
1/2	-1
1	0
1/3	-1.585
2/3	-0.585
1/4	-2
3/4	-0.415
1/5	-2.3219
2/5	-1.3219
3/5	-0.737
4/5	-0.3219
1/6	-2.585
5/6	-0.263
1/7	-2.8074
2/7	-1.8074
3/7	-1.2224
4/7	-0.8074
5/7	-0.4854
6/7	-0.2224
1/8	-3
3/8	-1.415
5/8	-0.6781
7/8	-0.1926

$E(S) = \sum_{j=1}^v -p_j \log_2(p_j) = -\frac{4}{8} \log_2(\frac{4}{8}) - \frac{4}{8} \log_2(\frac{4}{8}) = 1$						
		P_j	$P(j=y k)$	$P(j=n k)$	$p_{j,c_k} \log_2(p_{j,c_k})$ /*ติด \log_2 ไว้ได้*/	Remark (P_j)
สภาพอากาศ	k=s	3/8	0/3	3/3	$-(\frac{0}{3}) \log_2(\frac{0}{3}) - (\frac{3}{3}) \log_2(\frac{3}{3}) = 0$	$P_{j=y \text{อากาศ=s}} = \{\}$ $P_{j=n \text{อากาศ=s}} = \{\#1, \#2, \#8\}$
	k=o	2/8	2/2	0/2	$-(\frac{2}{2}) \log_2(\frac{2}{2}) - (\frac{0}{2}) \log_2(\frac{0}{2}) = 0$	$P_y \{ \#3, \#7 \}, P_n \{ \}$
	k=r	3/8	2/3	1/3	$-(\frac{2}{3}) \log_2(\frac{2}{3}) - (\frac{1}{3}) \log_2(\frac{1}{3}) = 0.9182$	$P_y \{ \#4, \#6 \}, P_n \{ \#5 \}$
$E_{C=\text{สภาพอากาศ}}(S) = \frac{1}{8}(0) + \frac{2}{8}(0) + \frac{1}{8}(0.9182) = 0.1148$						
อุณหภูมิ	k=h	3/8	1/3	2/3	$-(\frac{1}{3}) \log_2(\frac{1}{3}) - (\frac{2}{3}) \log_2(\frac{2}{3}) = 0.9182$	$P_y \{ \#3, \#7 \}, P_n \{ \#1, \#2 \}$
	k=m	2/8	1/2	1/2	$-(\frac{1}{2}) \log_2(\frac{1}{2}) - (\frac{1}{2}) \log_2(\frac{1}{2}) = 1$	$P_y \{ \#4 \}, P_n \{ \#5 \}$
	k=c	3/8	2/3	1/3	$-(\frac{2}{3}) \log_2(\frac{2}{3}) - (\frac{1}{3}) \log_2(\frac{1}{3}) = 0.9182$	$P_y \{ \#3, \#7 \}, P_n \{ \#1, \#2 \}$
$E_{C=\text{อุณหภูมิ}}(S) = \frac{1}{8}(0.9182) + \frac{2}{8}(1) + \frac{1}{8}(0.9182) = 0.9365$						
สภาพลม	k=T	3/8	1/3	2/3	$-(\frac{1}{3}) \log_2(\frac{1}{3}) - (\frac{2}{3}) \log_2(\frac{2}{3}) = 0.9182$	$P_y \{ \#3, \#7 \}, P_n \{ \#1, \#2 \}$
	k=F	5/8	3/5	2/5	$-(\frac{3}{5}) \log_2(\frac{3}{5}) - (\frac{2}{5}) \log_2(\frac{2}{5}) = 0.9709$	$P_y \{ \#3, \#4, \#6, \#7 \}, P_n \{ \#1, \#2, \#5 \}$
$E_{C=\text{สภาพลม}}(S) = \frac{1}{8}(0.9182) + \frac{5}{8}(0.9709) = 0.9511$						

- candidate $E_C(S)$ ต่ำสุด คือ สภาพอากาศ
- สร้าง the decision tree algorithm

กำหนดส่ง TBA