LØSNINGSSKISSE TIL EKSAMENSOPPGAVE I FAG TDT4300 – JUNI 2015

NB! Dette er ikkje fullstendige løysingar på oppgåvene, kun skisse med viktige element, hovudsakleg laga for at vi skal ha oversikt over arbeidsmengda på eksamen, og som huskeliste under sensurering. Det er også viktig å være klar over at det også kan vere andre svar ein dei som er gjeve i skissa som vert rekna som korrekt om ein har god grunngjeving.

Oppgåve 1

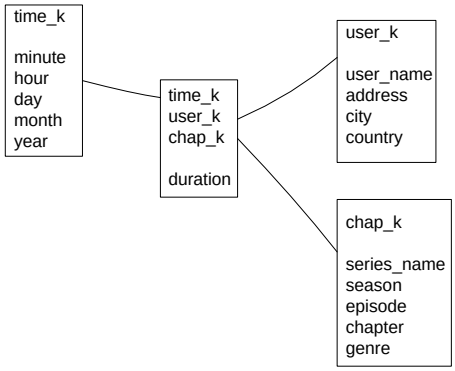
a) Kun tilstedeværelse (attributtverdi ulik 0) viktig (alternativt: nokre verdiar er viktigare enn andre).
Eksempel (forventar forklaring som under, ikkje berre "varer i handlekorg":
Objekt er student, eitt attributt for kvart fag på universitetet
Handlekorg, en attributt for kvar var, kun varer som er kjøpt er interessante.

b) 2/(1+1+2)=2/4=1/2
Poengtrekk der det manglar noko som forklarar talet.

c) (Minst 3) Eliminere (eller overser) objektet, Estimere manglande verdi, Interpolere manglande verdi, Ignorere det aktuelle attributtet.

Oppgåve 2

a) Eit viktig poeng er at fakta-attributt må gje meining utifrå oppgåva, og også gje meining ved aggregering.

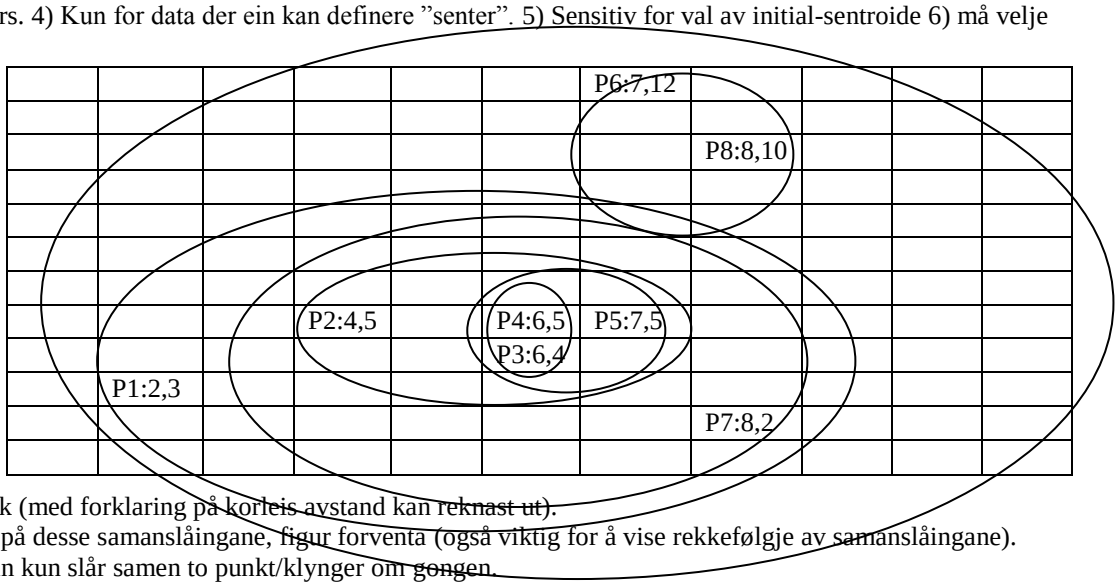


b) Ja, men hierarkiet må i såtilfelle ordnast som eit gitter (lattice), jfr. figure 4.10 i læreboka (Han kap. 4). Godtek også svaret "Nei" med fornuftig forklaring (litt diffust forklart i boka).

Oppgåve 3

a) Pro: Lav kostnad (både mht. utrekning og minne).
Con (minst 4): 1) Finn kun globulære klynger. 2) Problem med klynger med forskjellig størrelse og tettleik. 3) Problem med data med outliers. 4) Kun for data der ein kan definere "senter". 5) Sensitiv for val av initial-sentroide 6) må velje tal på klynger.

b) MIN-link:
P4 med P5 eller P3, deretter vert hhv. P3 eller P5 lagt til. Deretter P2 med P3/P4/P5. P6 og P8. P2/P3/P4/P5 med enten P1 eller P7. Deretter P1 eller P7 til denne. Til slutt, P6/P8 med resten. Er OK om ein løyser den grafisk (med forklaring på korleis avstand kan reknast ut).



Dendrogram: basert på desse samanslåingane, figur forventa (også viktig for å vise rekkefølgje av samanslåingane).
Viktig poeng er at ein kun slår samen to punkt/klynger om gongen.

Oppgåve 4

a) Forvekslingsmatrise er matrise med korrekt klasse-merkelapp (class label) versus estimert klasse-merkelapp for kvar klasse (to rader og to kolonnar som viser tall på falske positive, falske negative, ekte positive og ekte negative, jfr. figuren under). Nøyaktigheit rekna ut som vist i formelen under.

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

a: TP (true positive)
b: FN (false negative)
c: FP (false positive)
d: TN (true negative)

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

b)
Entropi i rotnode:
p(R|Parent) = 12/16 = 0.75, p(V|Parent) = 4/16 = 0.25. GI(C, A) = 1-0.75*0.5-0.25*0.25 = 0.375

1) **Splitting på dag A1:**
S1=”Fredag”
R1=2, V1=0, GI(C1,A1)=GI(2,0)=0

S2=”Laurdag”
R2=2, V2=2, GI(C2,A2)=GI(2,2)=0.5

S3=”Søndag”
R3=8, V3=2, GI(C3, A3)=GI(8,2)= 0.32

GAIN(A1) = 0.375 – 2/16*0 - 4/16*0.5 - 10/16*0.32=0.375-0.125*0-0.25*0.5-0.625*0.32 = 0.05

2) **Splitting på stad A2**
S1=”H”
R1=9, V1=1, GI(C1,A1)=1-0.9*0.9-0.1*0.1=0.18
S2=”B”
R2=3, V2=3, GI(C2,A2)=0.5

GAIN (A2) = 0.375-10/16*0.18-6/16*0.5 = 0.075

Vi vel attributtet med høgste GAIN, dvs. Stad vert føretrekt for første splitting av treet.

Oppgåve 5 – Assosiasjonsreglar

A	6
B	8
C	3
D	4
E	4
F	4
G	2
H	2

AB	6
AD	2
AE	3
AF	2
BD	4
BE	4
BF	4
DE	4
DF	4
EF	4

BDE	4
BDF	4
BEF	4
DEF	4

Kun eit 4-elementsett mogleg: BDEF | 4

B->DE	4/8	0.5	
BD->E	4/4	1.0	*
D->EB	4/4	1.0	*
DE->B	4/4	1.0	*
E->BD	4/4	1.0	*
BE->D	4/4	1.0	*