LØSNINGSSKISSE TIL EKSAMENSOPPGAVE I FAG TDT4300 – MAI 2016

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 enn dei som er gjeve i skissa som vert rekna som korrekt om ein har god grunngjeving eller dei gjev meining utifrå kva det vert spurd om.

Oppgåve 1

a) Unngå å finne mønster i støy (Finne klyngings-tendens til eit datasett, dvs. om ikkje-tilfeldige strukturar faktisk finnest)
For å samanlikne klyngings-algoritmer
For å samanlikne to sett med klynger
For å samanlikne to klynger

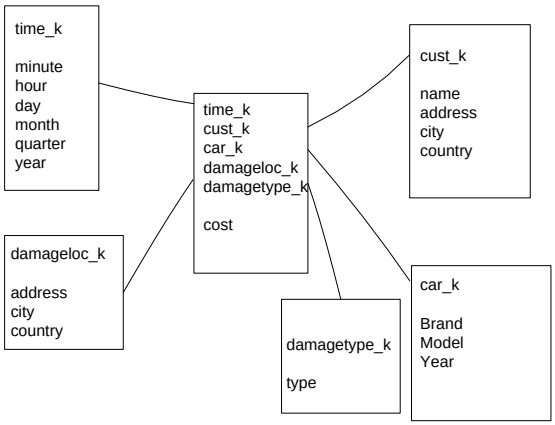
(Også andre er OK, t.d. "finne tal på klynger", "velge riktige parametre", etc.)

b) Fjerne irrelevante referansar og felt i loggar
Fjerne referansar som er resultat av aksessar frå søkerobotar
Fjerne feilaktige side-referansar (404 etc)
Legge til manglande referansar forårsaka av caching (etter sesjonering)

c) M11/(M11+M01+M10)=3/(2+1+3)=3/6=1/2

Oppgåve 2

a) Eit viktig poeng er at fakta-attributt må gje meining utifrå oppgåva, og også gje meining ved aggregering. Det er eigentleg ikkje nødvendig med "Antall" sidan kvar "event" er "ein skade", men sidan det har vore liknande attributt på eksempel i undervisninga trekk vi ikkje for det. Det er ikkje sagt noko om kunde i oppgåva så OK om dette ikkje er med. Med tanke på at oppgåve er relativt triviell er vi tilsvarande strenge, t.d. trekk ved manglande dimensjonstabell (typisk lokasjon), manglande skadetype, om fakta (beløp) er i ein av dimensjonstabellane men ikkje i det heile i faktatabell.



Oppgåve 3

a) 1) {*year, item_name, city*} // Ja (rollup frå city til province_or_state)
2) {*year, brand, country*} // Nei (materialisert, drill-down ikkje mogleg)
3) {*year, brand, province_or_state*} // Nei (materialisert, drill-down ikkje mogleg)
4) {*item_name, province_or_state*} where *year = 2004 // Nei. År 2016 ikkje med i denne*

Det er eit viktig poeng at oppgåva seier "Kven av dei *materialiserte* kuboidane kan brukast til å prosessere spørjinga", dvs. ein kan ikkje bruke drill-down (men ein kan gjere roll-up, dvs. vidare aggregering).

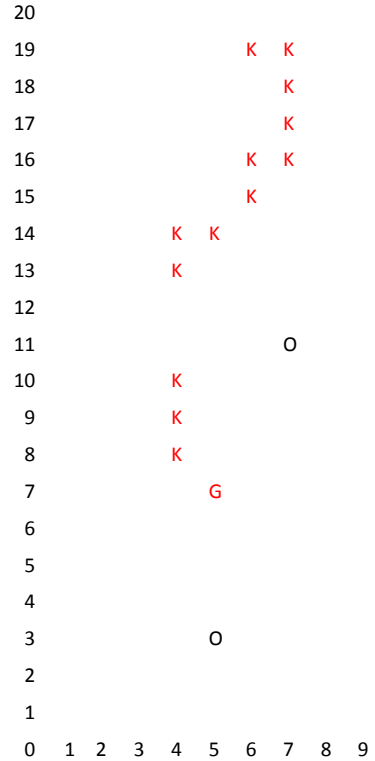
b) Join indexen skal helst vere på tabell-form, ein tuppel for kvar nøkkel/nøkkel. På eksempel på ein av foilane er det bruke "liste" for sekundærattributt så dette vert også godteke. Om tabell er ferdig join-resultat er det teikn på at studenten ikkje veit kva ein join indeks er, og null poeng.

Location/Sales	
LocKey	**TransID**
L1	T1
L1	T3
L1	T6
L1	T7
L2	T2
L3	T4
L3	T5

Item/Sales	
LocKey	**TransID**
I1	T1
I1	T3
I1	T4
I1	T7
I2	T2
I2	T6
I3	T5

Oppgåve 4

		X	**Y**	
A	P1	4	8	K
B	P2	4	9	K
C	P3	4	10	K
D	P4	4	13	K
E	P5	4	14	K
F	P6	5	3	O
G	P7	5	7	G
H	P8	5	14	K
I	P9	6	15	K
J	P10	6	16	K
K	P11	6	19	K
L	P12	7	11	O
M	P13	7	16	K
N	P14	7	17	K
O	P15	7	18	K
P	P16	7	19	K



Det er forventa at ein 1) har klassifisert punkt i K/G/O, og 2) har identifisert en klynge.
C1= P1,P2,P3,P4,P5,P7,P8,P9,P10,P11,P13,P14,P15,P16,
Støypunkt er P6 og P12.

Oppgåve 5

a) *Cross validation:*
   1) Partisjonerer data i *k* disjunkte subsett
   2) *k-fold*: tren på *k-1* partisjonar, teste på den gjenverande
   3) Metrikk er gjennomsnittleg effektivitet

   Brukes til estimering av effektivitet til klassifiseringsmodell.

   Legg merke til at oppgåva inneheld to spørsmål.

b)
Entropi i rotnode:
p(J|Parent) = 6/16 = 0.375, p(N|Parent) = 10/16 = 0.625. GI(J,N) = 1-0.375*0.375-0.625*0.625 = 0.468750

1) **Splitting på alder A1:**
S1=”L”
J1=3, N1=1, GI(J1,N1)=GI(3,1)=0.375

S2=”M”
J2=1, N2=8, GI(J2,N2)=GI(1,8)=0.197

S3=”H”
J3=2, N3=1, GI(J3, N3)=GI(2,1)=0.444

GAIN(A1) = 0.469 – 4/16*0.375- 9/16*0.197 - 3/16*0.444=0.181

2) **Splitting på bonus A2**
S1=”L”
J1=4, N1=2, GI(J1,N1)=GI(4,2)=0.444
S2=”H”
J2=2, N2=8, GI(J2,N2)= GI(2,8)=0.32

GAIN (A2) = 0. 468750-6/16*0.444-10/16*0.32= 0.102

Vi vel attributtet med høgste GAIN, dvs. **alder** vert føretrekt for første splitting av treet.
NB! Viktig å ha med GAIN inkl. p(J|Parent), det kan skje at ein får negativ verdi for begge dei alternative splittingane, som betyr at ein ikkje bør velje nokon av dei.

Oppgåve 6 – Assosiasjonsreglar

NB! I tittelen til oppg. 6 er det referert til ei oppgåve c. Dette er ein feil (kun a og b), det korrekt3 er 10% på a, 5% på b.1, og 10% på b.2. I T5 i a er det ein D for mykje. Studentane vart bede om å slette den eine, men med tanke på dei som allereie hadde gjort oppgåva godtek vi også at dei har rekna med begge D'ane og sett støtte til D til s(D)=6.

a)

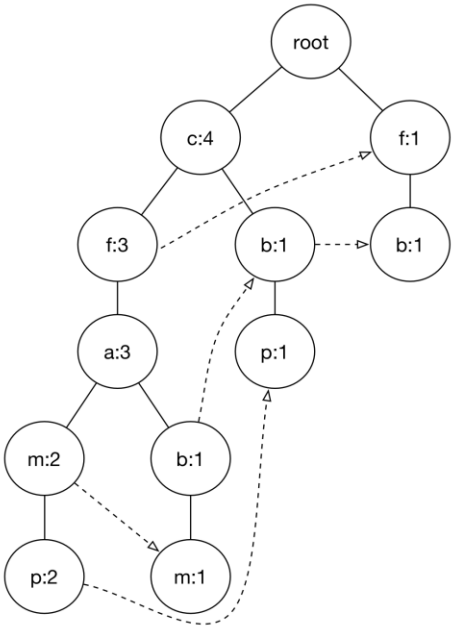
A	4
B	3
C	4
D	5
F	3
G	4
H	4
K	3
M	3
AC	4
AD	3
AG	4
AH	4
CD	3
CG	4
CH	4
DG	3
DH	3
GH	4
ACG	4
ACH	4
AGH	4
CGH	4

Kun eit 4-elementsett mogleg: ACGH | 4

ACGH	4

b)

tid	**Itemset**	**(Ordered) frequent items**
100	*f,a,c,d,g,i,m,p*	*c,f,a,m,p*
200	*a,b,c,f,l,m,*	*c,f,a,b,m*
300	*b,f,h,j,o*	*f,b*
400	*b,c,k,s,p*	*c,b,p*
500	*a,f,c,e,l,p,m,n*	*c,f,a,m,p*



Legg merke til om ein brukar anna sortering på element med same støttetal kan ein få andre (korrekte) tre. Typisk eksempel er cfabmp (”sortert rekkefølgje”).

Item	**Conditional sub-database**	**Conditional FP-tree**	**Frequent itemsets**
p	{(*cfam*:2), (*cb*:1)}	{(*c*:3)}|*p*	p, cp
m	{(*cfa*:2), (*cfab*:1)}	{(*cfa*:3)}|*m*	m, cm, fm, am, cam, fam, cfm, cfam
b	{(*cfa*:1), (*f*:1), (*c*:1)}	∅	b
a	{(*cf*:3)}	{(*cf*:3)}|*a*	a, ca, fa, cfa
f	{(*c*:3)}	{(*c*:3)}|*c*	f, cf
c	∅	∅	c

Legg merke til at andre rekkefølgjer og delvis anna mellom-resultat kan oppstå sidan fleire element har same støttetal (typisk vel mange å bruke abmpcf (”sortert rekkefølgje”)).