LØSNINGSSKISSE TIL EKSAMENSOPPGAVE I FAG TDT4300 – MAI 2018

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.

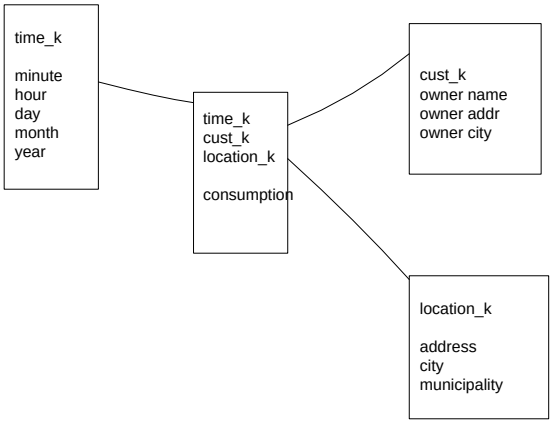
Oppgåve 1

a) Gjere kategoriske attributter til binære.
Ein asymmetrisk binær attributt for kvar moglege verdi av kategorisk attributt

b) Brukast til å evaluere godheita til eit *punkt* i ei klynge, og er metrikk for både kor likt eit punkt er andre punkt i klynga og kor ulikt det er punkt i andre klynger.
a = gjennomsnittleg distanse frå i til punkta i klynga til i
b = min(gjennomsnittleg distanse frå i til punkt i ei anna klynge) Reknar ut for kvar klynge j

c) Minnekrav $O(N^2)$ pga. matrise, og tid $O(N^3)$ (kompleksitet kan reduserast til $O(N^2 \log(N))$ med adekvate datastrukturar).

Oppgåve 2



Det er også mogleg (men ikkje krav) å ha med pris som ekstra attributt i fakta-tabell. Eit viktig poeng er at fakta-attributt må gje meining utifrå oppgåva, og også gje meining ved aggregering. Med tanke på at oppgåve er relativt triviell er vi tilsvarande strenge, t.d. trekk ved manglande dimensjonstabell, om fakta (consumption) er i ein av dimensjonstabellane men ikkje i det heile i faktatabell.

Oppgåve 3

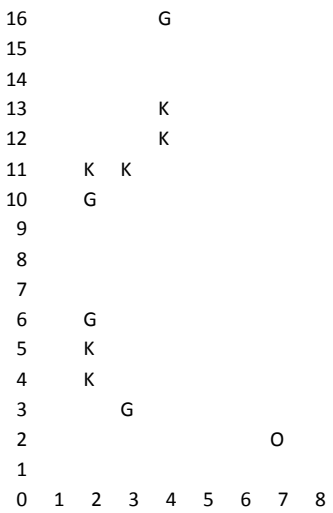
a) Jfr. læreboka (Han) s. 159:
1. No materialization: Do not precompute any of the "nonbase" cuboids. This leads to computing expensive multidimensional aggregates on-the-fly, which can be extremely slow.
2. Full materialization: Precompute all of the cuboids. The resulting lattice of computed cuboids is referred to as the *full cube*. This choice typically requires huge amounts of memory space in order to store all of the precomputed cuboids.
3. Partial materialization: Selectively compute a proper subset of the whole set of possible cuboids. Alternatively, we may compute a subset of the cube, which contains only those cells that satisfy some user-specified criterion, such as where the tuple count of each cell is above some threshold. We will use the term *subcube* to refer to the latter case, where only some of the cells may be precomputed for various cuboids.
For alternativ 3 er problemet å bestemme kva subkuber som skal materialiserast (Han s. 160).

b)
1) {year, brand, city} // Nei, (materialisert drill-down frå brand ikkje mogleg)
2) {year, brand, street} // Nei, (materialisert drill-down frå brand ikkje mogleg)
3) {month, brand, province_or_state} // Nei, kan ikkje ta drilldown frå brand til item_name
4) {item_name, province_or_state} where year = 2006 // Ja, rollup til country (og same år i spørjing og kube)

Oppgåve 4

a) 1) Sorter punkt i høve til distanse til deira *k*'te næraste nabo
2) Plott distansane
3) Høvande Eps er Eps for "knekkpunktet" der avstand aukar drastisk.
k (MinPts): For to dimensjonar er erfaringsvis *k* = 4 høvande, generell tommelfingerregel er k >= D +1 der D er dimensjonalitet
Kan tenkjast at andre strategiar kan gje "litt poeng", men ingen basert på SSE vil gje meining.

b)



Kjernepunkt: Pkt. 1, 2, 5, 7, 8, 9
Grensepunkt: Pkt. 3, 4, 6, 10
Støypunkt: Pkt. 11 (evt. også pkt. 12)
Klynger: 1,2,3,6 og 4,5,7,8,9,10
Punktet (7,2) førekjem to gongar i datasettet. På spørsmål under eksamen fekk studentane valet mellom å enten slette eit av dei (som gjev klynginga på figuren ovanfor), eller tolke dei som to distinkte objekt med same koordinat (som gjev to støypunkt i (7,2)).

Oppgåve 5

a) *Pre-pruning* (Early Stopping Rule)
Stopp før ein har eit "fullgrodd" tre
Meir restriktive vilkår:
- Stopp om Gain ved splitting av noverande node er mindre enn ein gjeve brukar-spesifisert terskel
- Stopp om tal på instansar er mindre enn ein gjeve brukar-spesifisert terskel
- Stopp om klasse-distribusjon til instansar er uavhengig av tilgjenglege "features" (t.d. ved bruk av statistisk test)

Post-pruning
- Konstruer det fullstendige treet
- Fjern noder i beslutningstreet botn-opp
- Om generaliseringsfeil vert redusert, erstatt sub-tre med løvnode
- Klasse-etikett (class label) for den nye løvnoda bestemt av kva som er majoritet-klasse til instansane i sub-treet
-

Ein føretrekk vanlegvis post-pruning i staden for pre-pruning pga.:
- Uansett relativt billig å konstruere heile treet
- I praksis vanskeleg å vite parametre for stopp-vilkår i pre-pruning

b)
Gini i rotnode:
p(J|Parent) = 5/15 =0.3333 , p(N|Parent) = 10/15 =0.6667
GI(J, N) = 1-0.333 *0.3333 -0.6667 *0. 6667 = 0.4444

1) **Splitting på A:**
S1="L"
J1=1, N1=9, GI(J1,N1)=GI(1,9)=1-1/10*1/10-9/10*9/10=0.1800

S2=”H”
J2=4, N2=1, GI(J2, N2)=GI(4,1)=1-4/5*4/5-1/5*1/5=0.3200

GAIN(A1) = 0.4444-10/15*0.1800-5/15*0.3200 = 0.2177

2) **Splitting på B**
S1=”T”
J1=2, N1=1, GI()=1-2/3*2/3-1/3*1/3=0.4444
S2=”K”
J2=1, N2=3, GI()=1-1/4*1/4-3/4*3/4=0.375
S3=”F”
J3=2, N3=6, GI()=1-2/8*2/8-6/8*6/8=0.375

GAIN (A2) = 0.4444-3/15*0.4444-4/15*0.375-8/15*0.375 =0.055

Vi vel attributtet med høgast GAIN, dvs. **A** 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

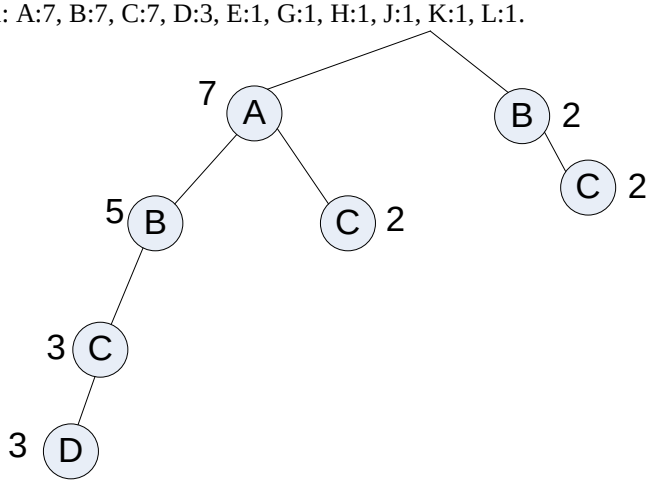
a) Viktig at det går fram at 3-kandidatsett er generert med bruk av $F_{k-1} \times F_{k-1}$

A	4
B	4
C	3
D	7
E	3
F	2
J	4
K	5
AB	1
AD	4
AJ	1
AK	4
BD	4
BJ	4
BK	4
DJ	4
DK	5
JK	4
ADK	4
BDJ	4
BDK	4
BJK	4
DJK	4
BDJK	4

B) BCE:2
AC:2, BE:3, BC:2, CE:2
C:3, A:2, B:3, E:3
Forventa at ein viser korleis ein har kome fram til desse.

Oppgåve 7/8

Støttetal: A:7, B:7, C:7, D:3, E:1, G:1, H:1, J:1, K:1, L:1.



Item	Conditional sub-database	Conditional FP-tree	Frequent itemsets
D	{(ABC:3)}	<A:3, B:3, C:3>	D3, AD:3, BD:3, CD:3, ABD:3, ACD:3, BCD:3, ABCD:3
C	{(AB:3), (A:2), (B:2)}	<A:5, B:3>, <B:2>	C:7, BC5, AC:5
CB	{(A:3)}	A3	ABC:3
(CA	∅	∅)	
B	{(A:5)}	A5	B:7, AB:5
A	∅	∅	A:7

OK om 1-elementsett ikkje er med i tabellen.
Blir godteke om støttetal på frekvente elementsett manglar.