LØSNINGSSKISSE TIL EKSAMENSOPPGAVE I FAG TDT4300 – JUNI 2013

Oppgave 1

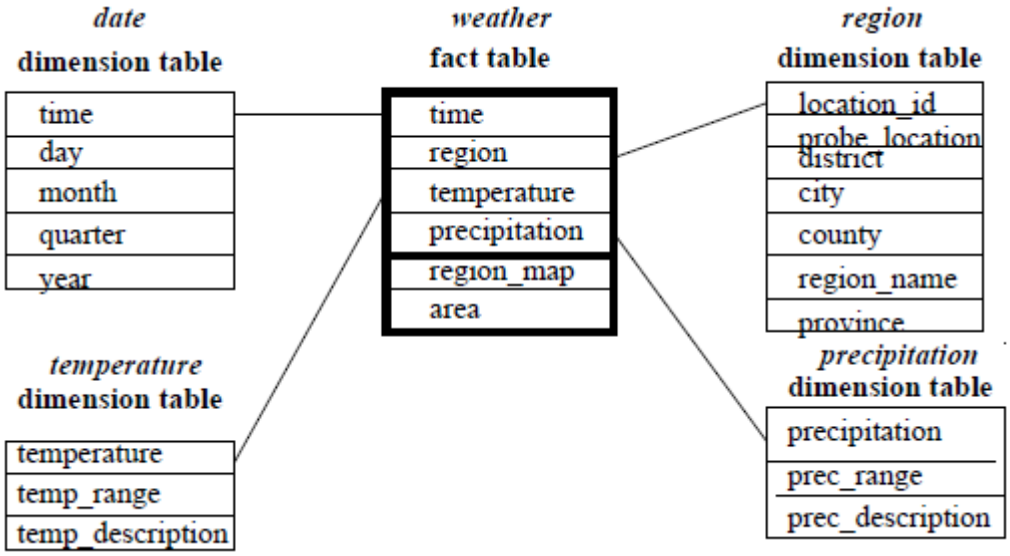
a) 1/(1+1+1)=1/3.
Jaccard egnet i kontekst av asymmetriske attributter (ignorerer attributter der begge er 0). Eksempel: For likhet mellom handlekorger er kjøpte varer det som oftest er interessant. Rett formel men feil på resten: 7p.

b) Sjå boka (Tan kap. 2.3). NB! Ikkje berre opplisting, men også forklaring. Gjev 4p for (rett) oppramsing.
I ein del lærebøker er det som i Tan er omtala som datakvalitet/datavasking presentert som preprosessering, dette gjeld mellom anna outlier removal og duplikatfjerning, så desse vert også godtekne som svar.

c) Prinsipp: Sjå boka (Han kap. 4.4.2)..
Type data egnet for: ” It is especially useful for low-cardinality domains because comparison, join, and aggregation operations are then reduced to bit arithmetic, which substantially reduces the processing time. Bitmap indexing leads to significant reductions in space and input/output (I/O) since a string of characters can be represented by a single bit. For higher-cardinality domains, the method can be adapted using compression techniques.”

Oppgave 2

a) Since the weather bureau has about 1,000 probes scattered throughout various land and ocean locations, we need to construct a spatial data warehouse so that a user can view weather patterns on a map by month, by region, and by different combinations of temperature and precipitation, and can dynamically drill down or roll up along any dimension to explore desired patterns.
The star schema of this weather spatial data warehouse can be constructed as shown below:



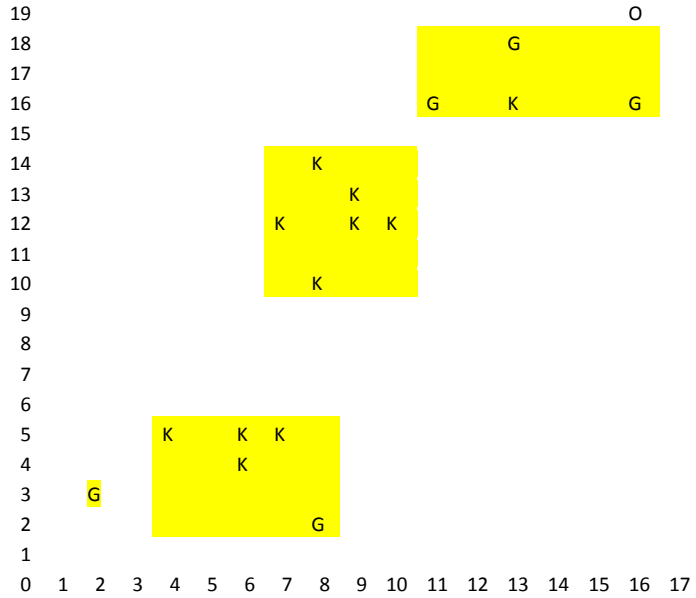
Denne figuren frå Solutions Manual er eignetleg feil utifrå “requirements” men ein del stduentar har tydelegvis sett den og gjer tilsvarande. Uansett, vurderr rett løysing til å vere temperatur og nedbør representert som fakta (attributt men ikkje separat tabell) i fakta-tabell, dette er det som er forelese og gjort på øvingane i faget. Eit anna poeng er at dei ikkje bør ha med avg/max/min sidan dette er noko ein kan rekne ut.

b) T.d. år-kvartal-månad-dag og provins-region-county-by

Oppgave 3 – Klynging – 20 %

a) Forklar potensielle ulemper med hierarkisk klynging.
-Når avgjerd er teken om samanslåing av to klynger kan den ikkje gjerast om.
-Høg tidskompleksitet og minneforbruk.
-Dei forskjellige metodane har problem med eit eller fleire aspekt: Sensitivitet mht. støy og outliers, problem med handsaming av klynger med forskjellig størrelse, oppdeling av store klynger

b) 1) DBSCAN-algoritmen: se læreboken/foilene. Forventar også forklaring på kva som er core/border/noise-point, og også sjølve algoritma (mange har berre forklaring på punkttyper og teikna ein sirkel rundt visuelle klynger i del 2).. 2) Gitt et to-dimensjonalt datasett som vist i tabellen til høyre. Utfør klynging ved hjelp av DBSCAN på dette datasettet, gitt MinPts=3 og Eps=3. Det kan forekome andre variantar som er korrekte, avhengig av om ein reknar MinPts som inkludert punktet eller ikkje, og om ein reknar Ept som inklusive eller ikkje (det siste kan føre til at øverste klynge vert støy). Det er også studentar som har brukt Manhattan-distanse for å gjere avstands-utrekning enklare. Kreativt, men rett. :)



1,414	1 opp og 1 til sida
2,236	1 opp og to til side
3,162	1 opp og 3 til side
2,828	2 opp og 2 til side

Oppgave 4 – Assosiasjonsregler

6p på elementsett og 4p på assosiasjonsreglar. 4p max. på elementsett om ein ikkje har 4-elementsettet (det er ikkje noko i oppgåva som tilseier at dei skal stoppe på 3-elementsett!).

A	6
B	4
C	7
D	2
E	4
F	4
G	3

AB	3
AC	5
AE	4
AF	4
BC	4
BE	2
BF	2
CE	4
CF	4
EF	4

ACE	4
ACF	4
AEF	4
CEF	4

Kun eit 4-elementsett mogleg: ACEF | 4

A->C	5/6	0.83		**A->EF**	4/6	.67
C->A	5/7	.71		AE->F	4/4	1*
A->E	4/6	.67		E->AF	4/4	1*
E->A	4/4	1		F->AE	4/4	1*
A->F	4/6	.67		EF->A	4/4	1*
F->A	4/4	1		AF->E	4/4	1*
B->C	4/4	1		**C->EF**	4/7	.57
C->B	4/7	.57		CE->F	4/4	1*
C->E	4/7	.57		E->CF	4/4	1*
E->C	4/4	1		F->CE	4/4	1*
C->F	4/7	.57		EF->C	4/4	1*
F->C	4/4	1		CF->E	4/4	1*
E->F	4/4	1				
F->E	4/4	1				
A->CE	4/6	.67				
AC->E	4/5	.8*				
C->AE	4/7	.57				
CE->A	4/4	1*				
E->AC	4/4	1*				
AE->C	4/4	1*				
A->CF	4/6	.67				
AC->F	4/5	.8*				
C->AF	4/7	.57				
CF->A	4/4	1*				
F->AC	4/4	1*				
AF->C	4/4	1*				

Oppgave 5 – Klassifisering – 25 %

a) Klassifisering er prosessen med å identifisere hvilken av et sett av kategorier (eller klasser) en ny observasjon hører til (predikseringsaspektet er viktig, mange studentar har ikkje med dette) Dette blir ofte gjort på grunnlag av et sett av treningsdata som inneholder observasjoner (eller instanser) hvor kategorimedlemskap er kjent. Basert på denne definisjonen er hensiktene med klassifisering å kunne kategorisere tidligere ukjente data basert på en modell er bygd basert på tidligere observasjoner eller treningsdata. Eksempler på dette er kategorisering av søkeresultater, kategorisering av sykdomsrisiko i en befolking, predikere været basert på tidligere værdata.

b) Kap. 5.2 i Tan. Når det gjelder nærmeste-nabo-klassifisering er det viktig at de viser at de forstår basisideene med avstandsmål mellom dataene og ant. naboer definert av avstandsmålet. For å klassifiser bruker man dette avstandsmålet mot andre data i treningssettet. Deretter kan man identifisere k nærmest nabo innenfor dette målet. For å avgjøre en klasse brukes "etiketten" til naboene som klassen til en ukjent data.

c) **Svar**: NB! Utrekningane under er basert på log_2, i oppgåveteksta er gjeve **log** som gjev andre numeriske verdiar. Log_10-varianten gjeve med tal i kursiv (omtrentlege tal, orka ikkje rekne dei ut :). Også nokon som har bruk ln. Har vore relativt fleksibel med små reknefeil.

Entropy i rootnode:
p(C|Parent) = 11/16
p(A|Parent) = 5/16
I(C, A) = (11, 5) = -11/16*log2(11/16)-5/16*log2(5/16) = 0.896 *0.27*

1) **Splitting på tid A1:**
S1="Morgen"
C1=2, A1=0, I(C1,A1)=I(2,0)=0 *0*

S2="Ettermiddag"
C2=7, A2=4, I(C2,A2)=I(7,4)= 0.946 *0.28*

S3="Kveld"
C3=2, A3=1, I(C3, A3)=I(2,1)=0.918 *0.276*

Dermed GAIN(A1) = 0.896 – (2/16*I(2,0) + 11/16*I(7,4) + 3/16*I(2,1)) = **0.074** ***0.022***

2) **Splitting på kamptype A2**
S1="Master"
C1=3, A1=3, I(C1,A1)=I(3,3)=1 *0.3*
S2="Grand tour"
C2=6, A 2=1, I(C2,A2)=I(6,1)= 0.591 *0.177*
S3="Show"
C3=2, A 3=1, I(C3, A3)=I(2,1)=0.918 *0.276*
Dermed GAIN (A2) = 0.896 – (6/16*I(3,3) + 7/16*I(6,1) + 3/16*I(2,1)) = **0.09** ***0.037 (eller 0.028 :)***

Vi velger attributtet med høyeste GAIN (eller laveste weightet avg.entropy). Derfor foretrekkes A2: Kamptype for første splitting av treet.