

LØSNINGSSKISSE TIL EKSAMENSOPPGAVE I FAG TDT4300 – AUGUST 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 (i og med at dette er kontinuasjonseksamen er den heller ikkje fullstendig, for dei to "grafiske" oppgåvene ser vi lett om studentane har rett eller ikkje :). 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 mening utifrå kva det vert spurd om.

Oppgåve 1 – Diverse

- a) "Outliers" er dataobjekter som er signifikant forskjellige fra de fleste andre objekter i datasettet.
- b) **Web-bruk-gruvedrift** (Web usage mining): automatisk oppdaging av mønstre i *clickstreams* og tilhørende data som har vorte samla inn eller generert som følge av brukarar sine interaksjonar med ein eller fleire nettstader
Mål: Analysere åtferdsmønster og profilar av brukarar i interaksjon med ein nettstad
Data i web-bruk-gruvedrift:
- Loggar frå web-tenarar
 - Innhald frå nettstadane
 - Data om brukarar frå eksterne kjelder
 - Andre applikasjonsdata
- c) I hovedsak å bestemme antallet kategorier (n) og finne split-points ($n-1$)
Et sett av intervaller som mappes til kategorier
Unsupervised :
Tar ikke hensyn til klassebetegnelse
Kan være brukerdefinerte intervaller, eller basert på klyngeanalyser (eksempel K-means)
Lik fordeling eller lik frekvens
- d) $M11/(M11+M01+M10)=4/(4+1+2)=4/7=0.57$

Oppgåve 2 – Datavarehus og OLAP

- a) Sjå læreboka.
- b) Enterprise-varehus
Inneholder data om alle emner, for hele virksomheten
Data mart
Hver mart (marked) er et subsett av virksomhetens totale varehus, og inneholder bare det som er av interesse for en subgruppe, underavdeling etc.
Virtuelle varehus
Verktøy som gir datavarehus data/funksjonalitet på toppen av operasjonelle data (relasjons-databaser) vha. views
Kan være begrenset hva som kan gjøres (begrensninger på hvilke views som kan være materialisert)
Avhengig av overskudds-kapasitet på operasjonelle database-tjenere
- c) Sjå læreboka.
- d) Sjå læreboka.
- e) Sjå læreboka.

Oppgåve 3 – Klynging

a) ...

```
1: Initialize the list of clusters to contain the cluster containing all points.
2: repeat
3:   Select a cluster from the list of clusters
4:   for  $i = 1$  to  $number\_of\_iterations$  do
5:     Bisect the selected cluster using basic K-means
6:   end for
7:   Add the two clusters from the bisection with the lowest SSE to the list of clusters.
8: until Until the list of clusters contains  $K$  clusters
```

b) Viktig fordel: Mindre følsom for initialisering

Oppgave 4 – Klassifisering

- a) Sjå læreboka.
- b) Sjå læreboka.

Oppgave 5 – Assosiasjonsreglar

- a) C1: a7,b3,c4,d4,g5,h5,k4
F1: a,c,d,g,h,k
C2: ac4,ad3,ag5,ah5,ak4,cd1,cg2,ch2,ck2,dg2,dh2,dk2, gh5,gk4,hk4
F2:ac,ag,ah,ak,gh,gk,hk
C3:acg*,ach*,ack*,agh5,agk4,ahk4,ghk4
F3:agh,agk,ahk,ghk
C4: aghk4
F4: aghk

*: Inneheld ikkje-frekvente subsett og kan prunast umiddelbart

- b) Forventa at ein har sortert 1-elementsett og transaksjonar med re-ordna element, og deretter sett inn i FP-tre.
- c) A:T1,T2,T3,T4,T5,T6,T8
B: T1,T3,T7
...
G: T1,T2,T5,T6,T8
...

Finn støttetal for AG ved å utføre snitt på listene for A og G som gir T1,T2,T5,T6,T8, resultatet har 5 element dvs. støttetal 5 (støtte 5/8).