LØSNINGSSKISSE TIL EKSAMENSOPPGAVE I FAG TDT4300 – AUGUST 2017

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) 1) Identifisere sesjonar, dvs. *aktivitetar utført av ein brukar frå ho/han først aksesserer nettstaden og til han/ho forlet nettstaden.*
   2) Vanskeleg å få pålitelege data pga.:
   - Proxy-tenarar og anonymisatorar
   - Dynamiske IP adresser
   - Manglande referansar pga. caching
   - Cookies kan koplast ut

   3) Teknikkar: tidsbasert og referent-basert, sjå læreboka.

b) Støy er tilfeldige unøyaktigheter/feil i målingene.
   "Outliers" er dataobjekter som er signifikant forskjellige frå dei fleste andre objekt i datasettet.

c) Forklar minst to av følgjande (jfr. læreboka):
   Dimensjonsreduksjon
   Feature subset selection
   Feature creation

d) Jfr. læreboka.

Oppgåve 2

a) Jfr. læreboka.
b) Jfr. læreboka.
c) Jfr. læreboka.
d) Jfr. læreboka. Eigna for attributtar med få distinkte verdiar.
e) Slicing er seleksjon på ein dimensjon slik at vi får ei subkube.
   Dicing er å velge ut ei subkube med to eller fleire dimensjonar.

Oppgåve 3

Jfr. algoritme 8.4 i læreboka. Kjerne/grense/støy-punkt, Eps og MinPts må forklarast, i tillegg til korleis sjølve klynginga vert gjort med desse som utgangspunkt.

Oppgåve 4

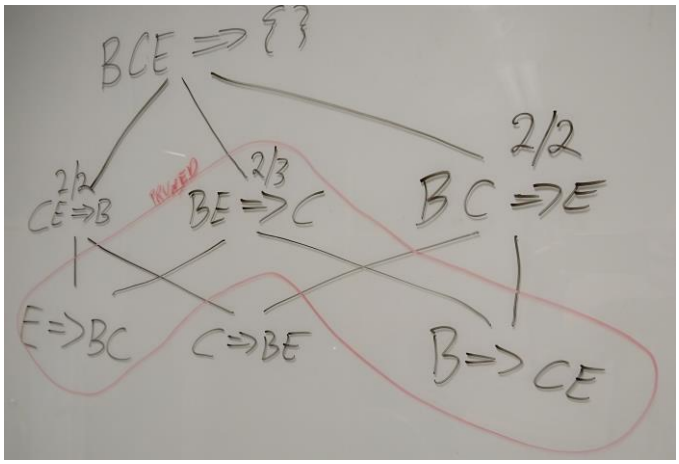
a) Partisjonerer data i k disjunkte
   subsett.
   k-fold: tren på k-1 partisjonar, teste på den gjenverande, etc.
   Metrikk er gjennomsnittleg effektivitet

b) Overtilpasning (overfitting): om modellen er for komplisert (altfor tilpassa treningsdata).
   Pga støy, manglande representative eksempel.
   pre-pruning og post-pruning, jfr. læreboka

Oppgåve 5

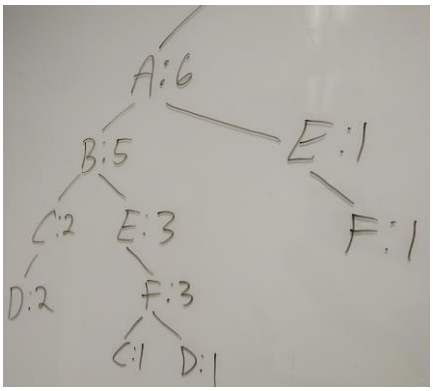
a) Elementsett er maksimalt frekvente om ingen av dei umiddelbare supersetta er frekvente. Eit elementsett er lukka om ingen av dei umiddelbare supersetta har same støtte som elementsettet.
b) Vesentleg poeng at join er brukt for å genere BCA og at ein skal bruke apriori-metoden (som spesifisert i oppgaven), og ikkje berre basere seg på brute-force (teste mot alle moglege variantar).

A	2
B	3
C	3
D	1
E	3
AB	1
AC	2
AE	1
BC	2
BE	3
CE	2
BCE	2



c) Støttetal: A:6, B:5, C:3, D:3, E:4, F:4, G:1
Viktig at G ikkje er med i treet og at rekursivitet er vist/forstått.

tid	**Itemset**	**(Ordered) frequent items**
T1	ABCD	ABCD
T2	ABCD	ABCD
T3	ABCEF	ABEFC
T4	ABEF	ABEF
T5	ABDEFG	ABEFD
T6	AEF	AEF



Item	**Conditional sub-database**	**Conditional FP-tree**	**Frequent itemsets**
D	{(ABC:2), (ABEF:1)}	*A3B3C2*	D:3, AD:3, BD:3, CD:2, ABD:3, ACD:2, BCD:2, ABCD:2
C	{(AB:2), (ABEF:1)}	AB3	C:3, AC:3, BC:3, ABC:3
F	{(AB*E*:3), (AE:1)}	A4(B*E*:3,E:1)	F:4, EF:4, BF:3, AF:4
EF	{(AB:3), (A:1)}	A:4B:3	AFE:4, BFE:3, ABFE:3
BF	A:3	A:3	FBA:3
E	{(AB:3), (A1)}	A:4B:3	E:4, AE:4, BE:3, ABE:3
B	{(A:5)}	A:5	B:5, AB:5
A	{(null)}	{(null)}	A:6

(OK om 1-elementsett ikkje er med i tabellen om sagt tidlegare kva som er frekvente element, eller ved figur har vist at dette er forstått).